









University of Computer Studies (Mandalay)

Department of Advanced Science and Technology

Ministry of Science and Technology

The Republic of the Union of Myanmar

JOURNAL OF INFORMATION TECHNOLOGY, RESEARCH AND INNOVATION

December, 2023

JiTRi 2023

ကွန်ပျူတာတက္ကသိုလ် (မန္တလေး)

JOURNAL OF INFORMATION TECHNOLOGY, RESEARCH AND INNOVATION

December, 2023

Organized by

University of Computer Studies

(Mandalay)

Volume-3, Issue-2

JOURNAL OF INFORMATION TECHNOLOGY, RESEARCH AND INNOVATION

JITRI, 2023

Vol-3, Issue-2

UNIVERSITY OF COMPUTER STUDIES (MANDALAY)

Journal of Information Technology, Research and Innovation (JITRI), 2023 Committee

University of Computer Studies (Mandalay)

Acknowledgements for Reviewers and Organizing members of JITRI

The Journal of Information Technology, Research and Innovation (JITRI) 2023 technical program committee would like to express the deepest appreciation to the external reviewers and organizing members for collaborating to publish this journal successfully. JITRI (2023) is published as the fifth time in order to undertake research in their respective fields of studies. Finally, we would like to thank all the authors who had submitted their research papers by making their great efforts in the Journal of Information Technology, Research and Innovation (JITRI), 2023, Vol- 3, Issue-2.

Table of Contents

CLUSTERING OF USER KNOWLEDGE LEVEL USING K-MEANS ALGORITHM

Thaw Tar Oo[1]

[1]Faculty of Computer Science, University of Computer Studies (Myeik)
[1]thawtaroo.757@gmail.com

ABSTRACT

Clustering is the method of collecting the data into classes or clusters so that objects within a cluster have high similarity in correlation to one another, but are very dissimilar to objects in other clusters. K-Means algorithm over other clustering methods depends on various factors, including the characteristics of the dataset, computational efficiency, and the desired outcome of the clustering task. In E-learning platforms, understanding knowledge level of users is pivotal for delivering personalized learning experience. The paper explores the application of K-Means clustering algorithm to categorize users based on their knowledge levels within a specific domain. The data set can be obtained from students' knowledge status about the subject of Electrical DC Machines at UCI data factory. By utilizing data on user interaction, the paper demonstrates how K-Means clustering can effectively group of level into distinct clusters representing varying levels of proficiency.

KEYWORDS

Clustering, K-Means Algorithm, Knowledge Level, user interaction, distinct clusters

1. INTRODUCTION

Data mining is also the process of assessing data from different perspectives and condensing it into useful information - information that can be used to increase profits, cuts costs, or both. It involves the searching of large information of data or records to discover patterns and utilize these patterns in predicting events in the future. Data mining provides easy access to certain set of methods and tools that is applicable to data processing for discovering hidden patterns. Data mining techniques can be classified into the following categories: classification, clustering, association rules, sequential patterns, time-series patterns, link analysis and text mining.

Clustering is one of the essential tasks for data analysis investigation which aims to find data structures which have fundamental state by modifying the data objects into similar groups and the description of data in classes, for this reason, it is called unsupervised classification or learning performed by research [5]. The main goal of clustering analysis is to arrange both similar and different objects in the same clusters and different clusters individually. In clustering, objects in a cluster are identical to one another yet dissimilar to object in other clusters. The semantic of the classes is not known beforehand in clustering techniques.

Clustering is the process of assorting the data into classes or clusters so that objects within a cluster have high similarity in correlation to one another, but are very dissimilar to objects in other clusters. Dissimilarities are appraised based on the characteristic conditions specifying the objects.

With the proliferation of online education platforms and the increasing demand for personalized learning experiences, accurately assessing and understanding the knowledge levels of users has become a paramount challenge. K-Means algorithm stands out for its simplicity, efficiency, and effectiveness in partitioning data into coherent clusters. This paper aims to explore the application of the K-Means algorithm in clustering users based on their knowledge levels within a specific domain. By leveraging various types of user interaction data, including behaviours to delineate distinct clusters representing different proficiency levels. This paper demonstrates users to study lectures and examine in a more efficient method.

2. Related Work

Clustering based on K-Means is closely correlated to several other clustering and location quandaries. These include the Euclidean K-Medians (or the multisource Weber problem) [4], [5] in which the objective is to minimize the sum of distances to the most adjacent center and the geometric k-center problem [3] in which the objective is to minimize the maximum distance from every point to its nearest center. There are no suitable solutions known to any of these quandaries and some formulations are NPhard [6].

The authors utilized the K-Means clustering to calculate the initial centroids of random selection thereby resulting to reduction in the number of iterations and the improvement of the elapsed time [3]. In article [2], the author used the clustering technique based on Canberra Distance Similarity and distance measure to observe and group similar behavior of the students of VIT University based upon their past academic performance. The technique also aimed at focusing on the students who were having bad performance by providing extra lectures and motivating them for better study.

In article [1], the author analysed the academic performance of the students in the biology department Kirkuk University Iraq using K-Means clustering algorithm and how to determine the new centroid using data mining partitioning approach termed as the K-Means algorithm and connection of nodes to obtain the result in KNIME tools.

3. Clustering Analysis

Clustering is a challenging domain of analysis where it's conceivable applications pose their own special prerequisites. This process divides a large dataset into mutually particular groups such that the members of each group are as "close" as possible to one another and different group's areas "far" as likely from one another, where distance is contained concerning all accessible variables. Cluster commentary has been generally utilized in various employment including pattern recognition, data analysis, image processing, and market research. Interval-scaled variables are consecutive measures of an approximately linear scale. The measurement unites used can affect the clustering analysis. The similarity (or dissimilarity) between the objects defined by interval-scaled variables is typically calculated based on the distance within each pair of objects. The most familiar distance measure is Euclidean distance d (i, j), which is designated as

$$\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2} \quad (1)$$

where i = (x_{i1}, x_{i2}, …, x_{ip}) and j = (x_{j1}, x_{j2}, …, x_{jp}) are two p-dimensional data objects. Euclidean distances provide the subsequent mathematic elements of a distance function:

(1) d (i, j) >= 0: Distance is a nonnegative number

(2) d (i, i) = 0: The distance of an object to itself is 0.

(3) d (i, j) = d (j, i): Distance is a symmetric function.

(4) d (i, j) <= d (i, h) + d (h, j): Proceeding immediately from object i to object j in space is no more than creating a deviation over any other object h (triangular contrast).

Given a database of n objects or data tuples, a partitioning approach categorizes k partitions of the data, where each partition represents a cluster and k <= n. Partitioning clustering technique helps to improve iteration techniques by mining objects from one graph plot to another. The principal purpose of the partition clustering algorithm is to separate the data features into K partitions. Each partition is responsible for reflecting one cluster. Partitioning a database S of n objects into set k clusters, such that the sum of squared distances is decreased. Given k, get a partition of k clusters that optimizes the favoured partitioning pattern. K-Means defines each cluster representation by the centre of the cluster and K-Medoids defines each cluster representation by one of the objects in the cluster. Mathematically, the error sum of squared Euclidean distances between each observation and its group means using partitioning approach is shown in

$$E=\sum_{i=1}^{k}\sum_{p\in C_i}(p-m_i)^2 \quad (2)$$

where k, C, (p – m_i) and m_i represent the number of clusters, the set of objects in a

cluster, the distance between p and m_i and the center point of the i-th cluster.

3.1. K-Means Clustering Algorithm

K-Means clustering algorithm is a method of clustering observations into a specific number of disjoint clusters [1]. The aim of the algorithm is to minimize the measurement between the centroid of the clusters and a provided consideration by iteratively supplementing the research to and clusters when the most inferior distance is reached. K-Means performance is determined by initialization and appropriate distance measure [2]. The stages of change of the cluster centres and reassign points are iteratively repeated until, until the border of clusters and location of centroids no longer changes, i.e.at each iteration, every cluster will get the same data point [4]. The secondary goal of K-Means clustering is to reduce the complexities in the data.

The K-Means algorithm improves as follows. First, it randomly chooses k of the objects, each of which originally serves a cluster mean or centre. For each of the prevailing objects, an object is allocated to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then estimates the new mean for each cluster. This process emphasizes until the model function concentrates. The processing of K-Means algorithm was explained as the following steps:

Randomly pick k objects from database S as the initial cluster centres;

repeat

for each object **do**

Compute distances from the object to cluster centers;

Assign the object to the cluster with the nearest cluster center,

end

for each cluster **do**

Calculate the mean value of the objects;

end

until no (or minimum) change;

The advantages of K-Means algorithm are

- Ease in the implementation
- Relatively effective and efficient: 0(tkn), where the n represents the number of data objects; k represents the numbers of clusters; t represents the numbers of iterations implementation in the normal state k and t<<n.

The disadvantages of K-Means algorithm are

- Application in only situations when the mean is defined.
- Not suitable to discover clusters with non-convex shapes and attributes.
- Highly sensible to noisy data and Outliers.

3.2. Choosing Initial Centroids

Choosing the individual initial centroids is the key step of the basic K-Means method. It is understandable and valuable to take initial centroids randomly, but the results are frequently inadequate. It is possible to operate multiple runs, each with a separate set of randomly collected initial centroids - but this may still not operate counting on the data set and the number of clusters investigated. There isn't any difficulty as long as two initial centroids fall anyplace in a set of clusters since the centroids will redistribute themselves, one to each cluster, and so acquire a globally insignificant error. However, it is very conceivable that one set of clusters will have only one initial centroid. As the sets of clusters are considerably apart, the K-Means algorithm will not redistribute the centroids between sets of clusters, and thus, only restricted particles will be accomplished.

This requires the necessity for a technique to take a new centroid for an empty cluster; otherwise, the squared error will surely be larger than it would need to be. A common approach is to keep as the new centroid the point that is considerably away from any current centre. If nothing else, this excludes the point that currently provides the greatest to the squared error. Modernizing points incrementally may submit an arrangement dependence problem, which can be enhanced by randomizing the order in which the points are prepared. However, this will not be

possible except the features are in the main memory.

Updating the centroids after all points are allocated to clusters decisions in order independence procedure. Subsequently, when centres are modernized incrementally, each step of the process may need to renew two centroids if a point changes clusters. But, K-Means serves to concentrate rather quickly and so the number of points rearranging clusters will attend to be small after a few suggestions over all the features.

4. DATASET DESCRIPTION

The User Knowledge Level dataset is a comprehensive collection of anonymized data sourced from an online learning platform, aimed at facilitating research in educational data mining, user profiling, and personalized learning experiences. It encompasses a variety of attributes that capture user interactions, study behaviours, and performance metrics within the platform. This data set contains the 5 attributes and 403 instances and then, 4 class descriptions. The five attributes information are STG, SCG, STR, LPR and PEG.

STG refers to the amount of time a user spends studying or engaging with materials that are directly related to their learning objectives or goals. In the context of an online learning platform or educational environment, these "goal object materials" could include specific lessons, modules, chapters, or topics that align with the user's learning objectives or the curriculum's learning outcomes. SCG likely represents an attribute related to "Study Time for Course / Conceptual Goal materials." SCG refers to the amount of time a user spends studying or engaging with materials that are essential for achieving the overarching goals of a course or understanding key concepts within a domain. These materials typically encompass foundational concepts, theories, or principles that form the basis of the subject matter being studied. The "STR" attribute in a User Knowledge Level dataset represents the "Study Time of User for Related Objects with Goal Object." This attribute quantifies the extent of time spent by a user studying materials that are directly associated with their learning objectives or goals. The attribute “STR” and “STG” are not same. STG attribute is Study Time for Goal Object Materials. The duration of study time allocated by the user for materials specifically aligned with their primary learning objectives or goals.

The attribute "LPR" in the User Knowledge Level dataset stands for "The exam performance of the user for related objects with the goal object." This attribute quantifies the performance of a user in exams, assessments, or evaluations specifically related to materials that are associated with their primary learning objectives or goals. These materials, referred to as "related objects," are supplementary resources, exercises, or topics that contribute to achieving the overarching learning goals established by the user.

In the User Knowledge Level dataset, the attribute "PEG" represents "The exam performance of the user for goal objects." This attribute measures the performance of a user in exams, assessments, or evaluations related to materials that are aligned with their primary learning goals. These materials, referred to as "goal objects," represent the core topics or concepts that users aim to master within the domain of study. The Values of all attributes in User Knowledge Level are Numeric Values.

UNS in the User Knowledge Level dataset represents "The knowledge level of the user" and serves as a class label indicating the proficiency level of each user within a specific domain. It categorizes users into discrete levels such as Very Low, Low, Middle and High based on their performance in exams, study time allocation, and engagement metrics. The description of attributes Information in User Knowledge Level dataset are in the following table.

Table 1. Description of User Knowledge Level Dataset

No.	Attributes Information	Value
1.	**STG** (The degree of study time for goal object materials)	Input Numeric Value
2.	**SCG** (The degree of repetition number of user for goal object materials)	Input Numeric Value

3.	**STR** (The degree of study time of user for related objects with goal object)	Input Numeric Value
4.	**LPR** (The exam performance of user for related objects with goal object)	Input Numeric Value
5.	**PEG** (The exam performance of user for goal objects)	Input Numeric Value
Class	**UNS** (The knowledge level of user)	Target Value (Very Low, Low, Middle, High)

In summary, the User Knowledge Level dataset provides valuable insights into user study behaviours, engagement patterns, and proficiency levels within online learning environments, facilitating data-driven approaches to enhance learning efficacy and user satisfaction.

5. IMPLEMENTATION

The dataset of User Knowledge Level with the extension name "csv" was imported from the excel spreadsheet into the file reader which reads file and build a workflow which produced the levels. The KNIME tool is used for implementation of this work as shown in Figure 1.

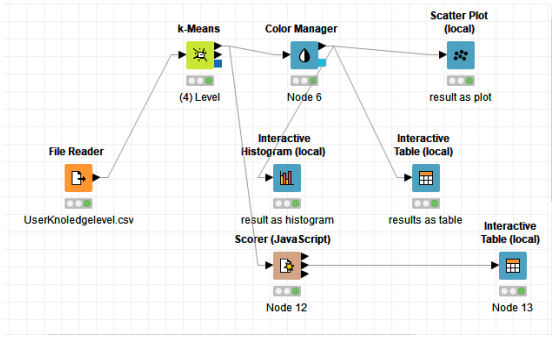


Figure 1. Workflow Model of K-Means Clustering Analysis

The workflow consists of File Reader, K-Means Node, Colour Manager, Scatter Plot, Interactive Table, Interactive Histogram, and Scorer. File Reader which reads the dataset's excel sheet into csv file. K-Means nodes perform the clustering algorithm on the sample instance to determine the centroid and the minimal distances of four Level such as Very Low, Low, Middle and High as four clusters. The colour Manager assigns the different colours to the number of clusters produced after K-Means algorithm has been performed. The scatter plot is the node which interactively visualizes the relationship between the column in the dataset. And then, the interactive table displays the clusters in a table form and includes the records ID and the cluster number. The interactive histogram node plots a histogram for the clusters generated from the K-Means clustering node such as cluster_0, cluster_1 and so on. The scorer node displayed the confusion matrix table including cluster group. It compares the given attributes and the clusters match.

6. EXPERIMENTAL RESULT

From the analysis, the number of clusters was generated from the 5 attributes (STG, SCG, STR, LPR, PEG) and displayed as Cluster of the four clusters of UNS or Knowledge Level of the User in the interactive table as shown in Figure 2. In the interactive table, the cluster_0, cluster_1, cluster_2, cluster_3 was represented as the four colours. In this paper the group of cluster_0 as green, the group of cluster_1 as red, the group of cluster_2 as brown and the final group of cluster_3 as purple are represented respectively.

Row ID	STG	SCG	STR	LPR	PEG	Cluster
Row16	0.05	0.07	0.7	0.01	0.05	cluster_1
Row17	0.1	0.25	0.1	0.08	0.33	cluster_1
Row18	0.15	0.32	0.05	0.27	0.29	cluster_1
Row19	0.2	0.29	0.25	0.49	0.56	cluster_1
Row20	0.12	0.28	0.2	0.78	0.2	cluster_3
Row21	0.18	0.3	0.37	0.12	0.66	cluster_0
Row22	0.1	0.27	0.31	0.29	0.65	cluster_0
Row23	0.18	0.31	0.32	0.42	0.28	cluster_1
Row24	0.06	0.29	0.35	0.76	0.25	cluster_3
Row25	0.09	0.3	0.68	0.18	0.85	cluster_0
Row26	0.04	0.28	0.55	0.25	0.1	cluster_3
Row27	0.09	0.255	0.6	0.45	0.25	cluster_3
Row28	0.08	0.325	0.62	0.94	0.56	cluster_3
Row29	0.15	0.275	0.8	0.21	0.81	cluster_0
Row30	0.12	0.245	0.75	0.31	0.59	cluster_0
Row31	0.15	0.295	0.75	0.65	0.24	cluster_3
Row32	0.1	0.256	0.7	0.76	0.16	cluster_3
Row33	0.18	0.32	0.04	0.19	0.82	cluster_1
Row34	0.2	0.45	0.28	0.31	0.78	cluster_0
Row35	0.06	0.35	0.12	0.43	0.29	cluster_1
Row36	0.1	0.42	0.22	0.72	0.26	cluster_3
Row37	0.18	0.4	0.32	0.08	0.33	cluster_1
Row38	0.09	0.33	0.31	0.26	0	cluster_1
Row39	0.19	0.38	0.38	0.49	0.45	cluster_1

Figure 2. Interactive Table for the result

The Interactive Histogram for the four classes of Clusters (cluster_0; cluster_1; cluster_2; cluster_3) based on the analysis of the User Knowledge to display the Levels. The histogram plot is shown in Figure 3.

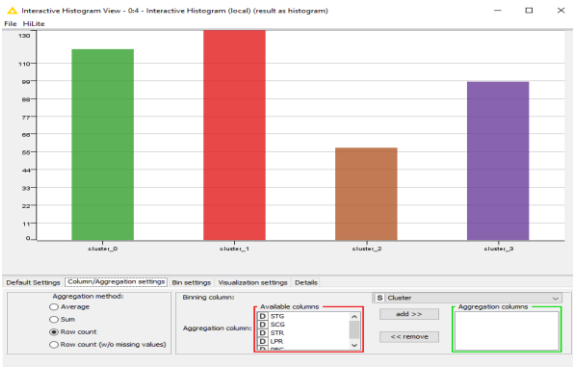


Figure 3. Interactive Histogram for the result

Several Iterations were operated on the user level using the scatter-plot node to ensure that the Euclidean distances between the squares can be calculated to the minimal position. This movement can be achieved by applying a small displacement to each square in a direction that minimizes the overall distances. Repeat this process for a predefined number of iterations or until convergence criteria are met. The result of the scatter-plot is shown in Figure 4.

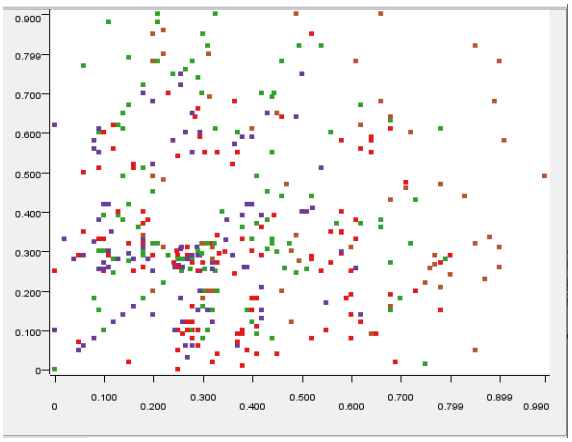


Figure 4. Scatter Plot with the clustering of UNS

The Scorer (JavaScript) node displayed the confusion matrix with the number of matches in each cell. Row and column rates can be shown via a configuration setting; they are the number of correct predictions divided by the total number of records in the row/column. Additionally, it is possible to highlight cells of this matrix to select the underlying rows. The selection can be passed to other JavaScript views and as shown in Figure 5.

Rows Number : 403	cluster_0 (Predicted)	cluster_1 (Predicted)	cluster_2 (Predicted)	cluster_3 (Predicted)
cluster_0 (Actual)	118	0	0	0
cluster_1 (Actual)	0	130	0	0
cluster_2 (Actual)	0	0	47	0
cluster_3 (Actual)	0	0	0	98

Figure 5. Confusion Matrix Table for the result

The accuracy of this system depends on the correctly classified instance. The formula of Accuracy is the following equation.

$$\text{Accuracy}=100 * \frac{no.of\ data\ recognized\ accurately}{no.of\ datas} \quad (3)$$

In this paper, the 403 instances of User Knowledge Level are applied. The system accurately clustered into cluster_0 (118 instances), cluster_1 (130 instances), cluster_2 (47 instances) and cluster_3 (98 instances) respectively. The system correctly recognized 393 instances in total instance 403 data in this paper. The experimental result is calculated by using equation (3) and demonstrated that the overall accuracy is 97.5% depending on correctly classified instances.

7. CONCLUSIONS

K-Means algorithm is one of the most usually used in the clustering algorithm and futurity learning. K-Means is particularly well-suited for numeric data, such as study time, quiz scores, and engagement metrics commonly found in user knowledge level datasets. It calculates cluster centroids based on the mean of data points in each cluster, making it effective for clustering continuous variables. This paper successfully segmented users into distinct clusters representing their knowledge levels within a specific domain. The clustering of user knowledge levels using the K-Means algorithm holds significant implications for personalized learning experiences and educational interventions. By understanding the diverse needs and proficiency levels of users, educators and platform developers can tailor content delivery, recommend appropriate learning materials, and provide targeted support to enhance the overall learning

experience. It reduces time consuming, and can easily determine users' knowledge level. While K-Means clustering proved effective in this paper and explored the integration of additional features and the use of alternative clustering algorithms to further improve clustering accuracy.

REFERENCES

[1] Alkadhwi Ali Hussein Oleiwi, Adelaja Oluwaseun Adebayo, "Data Mining Application Using Clustering Techniques (K-Means Algorithm) In the Analysis of Student's Result", Journal of Multidisciplinary Engineering Science Studies (JMESS), Vol.5 Issue 5, May-2019

[2] Azhar Rauf, Sheeba, "Enhanced K-Mean clustering Algorithm to reduce number of iterations and time complexity", Middle-East Journal of Scientific Research, Vol. 12, No 7, pp. 959-963, 2012.

[3] David Sontag, "Clustering K-means", New York University, pp. 136, [ElectronicResources]URL:http://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf

[4] Jiawei. Han, "Data mining- Concepts and Techniques" 2nd Edition- Impressao, 2006

[5] Kartik. N. S et al, "Clustering Students' based on previous academics performance", IJERA, Vol.3, Issue 3, May-June 2013, pp 935-939

[6] Ke Chen, "Machine Learning K-means clustering", University of Manchester, pp 1-22.

[7] L. Kaufman and P.J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis", New York: John Wiley & Sons, 1990.

[8] M. Ester, H. P. Kriegel, X. Xu "A Database Interface for Clustering in Large Spatial Databases", Institute for Computer Science, University of Munich. 1st International Conference on knowledge Discovery and Data Mining (KDD-95) pp. 94-99, 1995.

[9] R.T. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining, Proc. 20th Int'l Conf. Very Large Databases", pp. 144-155, Sept. 1994.

First Author: Daw Thaw Tar Oo is Assistant Lecturer from Faculty of Computer Science at University of Computer Studies, (Myeik). She received Bachelor Degree in 2012 and Hons. Degree of Computer Science in 2013 at UCS(Myeik). She then passed Master of Computer Science in UCS(Yangon). She received M.C.Sc.(Credit) in 2016. She works in FCS department from 2017 to 2021 at UCS(Myeik), from 2021 to 2023 at UCS(Hpa-an) and then, from 2023 till now at UCS(Myeik). She got special award on Basic Course for Junior Civil Service Officers No. (82) at Central Institute of Civil Service (Lower Myanmar). She has published 3 papers in Conference and Journals, including PSC conference 2017. The first paper was published in PSC 2017 Proceeding and then second paper and third paper were published in University Research Journal, 2020. She interests in Data Mining, Artificial Intelligence, NLP and machine learning. Now, she is being taught Computer Graphics, AI, Principle of IT, Operating Systems, Data Structures, Object Oriented Programming and supervised in Final Year's Projects and Internship Projects.

TRANSFORMING METHOD FOR REMOTE SENSING BIG DATA PROCESSING USING APACHE SPARK AND DEEP LEARNING

Myo Myat Thu[1], Phone Naing[2], and Kyaw Kyaw Lin[3]

[1,2,3] Department of Computer Science, Higher Education Center (Pyin Oo Lwin)
[1]nyimyatthu49@gmail.com, [2]kophonenaing7@gmail.com, [3]kklin1500@gmail.com

ABSTRACT

Roads are essential for transportation and everyday ease. Yet, precisely identifying road-related details from high-resolution remote sensing images, particularly from Google satellite images, poses a challenge. Recent progress in Convolutional Neural Networks (CNNs), a form of computer algorithm, has demonstrated significant potential in precisely detecting and segmenting road information from these images. This research paper introduces a transformative methodology to create a custom dataset using remote sensing data from ArcGIS Pro software, specifically Google satellite imagery, for the precise extraction of roads within the Pyin Oo Lwin area. The experimental dataset results, utilizing the transformative method for UNet with Apache Spark, aim to enhance the accuracy of road segmentation within the Pyin Oo Lwin area from remote sensing images, specifically those obtained from Google Satellite.

Keywords

Convolutional Neural Networks (CNN), Apache Spark, road extraction, segmentation, UNet, remote sensing, Geographic Information System (GIS)

1. INTRODUCTION

The extraction of roads from remote sensing imagery is a vital task that holds extensive implications across navigation, disaster management, and urban planning domains. These images contain diverse road types, including highways, urban-rural roads, and byways. Yet, these undertaking encounters notable hurdles due to factors like noise, occlusions, and the complex composition of road structures.

Presently, the extraction of roads often relies on manual interpretation, which is time-consuming and prone to inconsistencies owing to subjective judgments. Remote sensing images face issues such as terrain variations, shadows, and occlusions, complicating the process of road extraction [5].

Advancements in remote sensing technology have led to the acquisition of extensive, high-resolution Earth surface information, resulting in the proliferation of vast datasets known as "big remote sensing data" [1]. Enriched with improved spatial, spectral, and temporal resolutions, these datasets necessitate the implementation of sophisticated processing methods. These methods, frequently complemented by external data sources such as social media, significantly contribute to ensuring accurate data interpretation.

In the domain of big data analytics, Apache Spark has emerged as a potent framework for the analysis of vast datasets. It seamlessly accommodates heterogeneous workloads, proffering a unified processing engine underpinned by versatile programming languages. Apache Spark's swiftness and scalability have garnered substantial adoption both in academic circles and industry, thereby cementing its status as one of the most dynamic open-source projects within the expansive domain of big data under the aegis of the Apache Software Foundation [7].

Recent focus has been on Convolutional Neural Networks (CNNs), which have garnered substantial attention and significantly advanced computer vision while addressing challenges in natural language processing. CNNs find applications in various domains, including autonomous driving, remote sensing image analysis, and medical image processing. Noteworthy architectural innovations in this field include the Fully Convolutional Network (FCN) [3] and the UNet [2].

The satellite imagery used for road extraction is sourced from Google Satellite imagery and covers the Higher Education Center (HEC), which spans an area of 7.508206 square kilometres. The ground resolution of the image pixels is 1m/pixel.

When generating a custom dataset from Google Satellite images using ArcGISpro, the resulting dataset is in TIFF format, which may not be directly compatible with the UNet architecture for further utilization.

2. Aim and Objective

The aim of this research is to propose a transformative methodology tailored for remote sensing data obtained from ArcGISpro software. This method is intended by adaptable to Apache Spark for training datasets within the UNet architecture to achieve accurate segmentation of roads captured within the Pyin Oo Lwin area.

3. Methodology

In this paper, our primary objective is to tackle the challenge of training datasets, specifically within the UNet architecture, to achieve precise road segmentation from remote sensing data, focusing particularly on Google Satellite images.

3.1. GIS

Geographic Information System (GIS) is a powerful technology that enables the capture, storage, analysis, and presentation of spatial data. It integrates various types of geographical information, such as maps, satellite imagery, and survey data, into a single platform.

GIS integrates various types of data—such as maps, satellite imagery, aerial photography, and statistical data—into a comprehensive database. This information is georeferenced, meaning it is associated with specific geographic locations or coordinates on the Earth's surface [4].

3.2. Remote Sensing

Remote sensing involves gathering information about an object, region, or occurrence without direct physical interaction. Remote sensing images are captured using various sensors mounted on satellites, aircraft, drones, or ground-based instruments. These images offer valuable information about the Earth's surface and its characteristics, obtained from a considerable distance.[5]

Google Satellite Images are a part of Google Earth, which provides a vast collection of aerial and satellite imagery, offering views ranging from global perspectives down to street-level details in certain areas. These satellite images serve multiple purposes, notably aiding in urban planning, environmental surveillance, geographical research, navigation systems, and educational initiatives.

3.3. Apache Spark

Apache Spark, an open-source distributed computing system, redefines big data processing by offering lightning-fast speeds through in-memory computing. With user-friendly APIs in various languages, it enables diverse tasks like batch processing, streaming, SQL queries, and machine learning within a unified framework. Spark's distributed nature ensures scalability and fault tolerance using resilient distributed datasets (RDDs), supported by libraries catering to varied data processing needs across industries.

Apache Spark Core, forming the foundation of the Spark platform, is an open-source distributed computing system that revolutionizes big data processing. By leveraging in-memory computing, it achieves remarkable speed, offering intuitive APIs in multiple languages for tasks like batch processing, streaming, SQL queries, and machine learning. Spark's distributed architecture ensures scalability and fault tolerance through resilient distributed datasets (RDDs), supported by a rich ecosystem of libraries catering to diverse data processing needs across industries [6].

3.4. Tagged Image File Format (TIFF)

The TIFF (Tagged Image File Format) stands as a prevalent file format utilized for storing raster graphic images. It is known for its versatility and flexibility, accommodating various forms of image data, such as photographs, scanned documents, and satellite imagery. It can store images in a lossless format, preserving the original quality and detail of the image without compression artifacts.

TIFF files find widespread application in numerous industries like graphic design, publishing, satellite imaging, medical imaging, and digital preservation owing to their exceptional quality and adaptability [8].

4. PROPOSED METHODS

When establishing a dataset for deep learning, particularly for road extraction using remote sensing data and specifically targeting the UNet architecture, leverage the Export Training Data for Deep Learning tool within the ArcGIS Pro environment. This tool facilitates the generation of both images and masks, which are primarily in TIFF format. However, to ensure precise road segmentation in deep learning, especially when employing the UNet architecture, a conversion of these image types to PNG format is necessary.

Below are the satellite image and binary mask image initially generated from ArcGISpro's Export Deep Learning tool. Figure 1 displays the satellite images sourced from Google Satellite, while Figure 2 portrays their corresponding mask images. All these images are stored in TIFF file format.

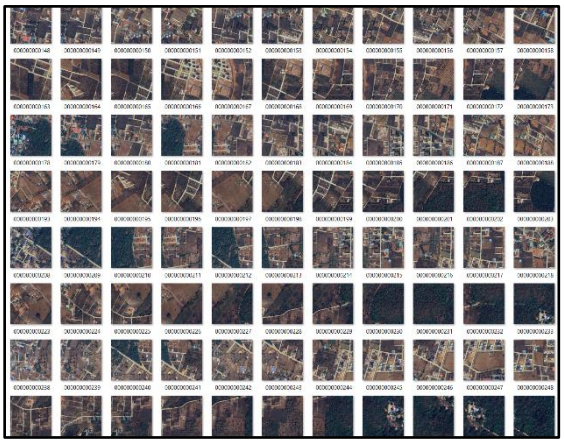


Figure 1. TIFF images from Google Satellite

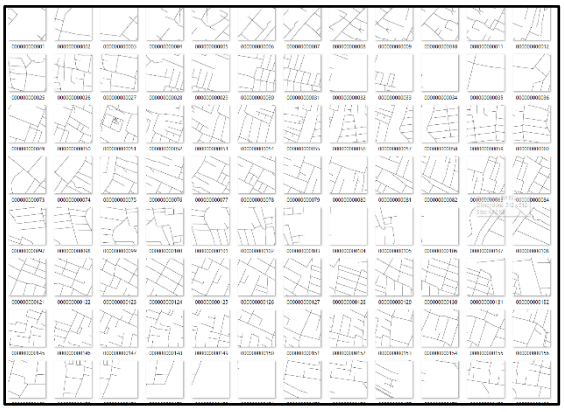


Figure 2. Binary Mask TIFF images

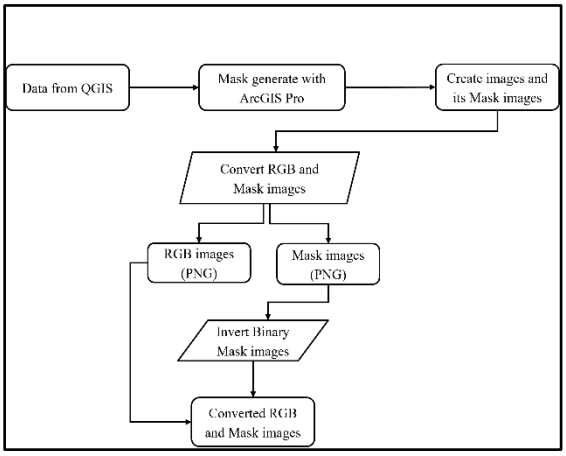


Figure 3. Overall proposed system

4.1. Proposed Method of Convert TIFF to PNG

Converting TIFF images to PNG becomes necessary when preparing data for deep learning tasks, such as training a UNet architecture. TIFF images may present compatibility issues with machine learning tools and frameworks, making them less suitable for such tasks.

Additionally, TIFF files can include an unnecessary alpha channel (RGBA), introducing complexity to the data and potentially slowing down the training process due to larger file sizes.

The conversion process is straightforward, involving the use of image conversion software to change the format while preserving essential RGB colour information.

This ensures that the data is ready for efficient use in the deep learning model without encountering technical complications.

The process of converting a TIFF image to a PNG image can be represented mathematically as follows:

$$P = f(t) \quad (1)$$

Let "T" be the original TIFF image with pixel values represented as RGBA, where each pixel is denoted as T(i, j), and the RGBA values are (Rt, Gt, Bt, At).

"F" is the conversion function that transforms the tiff image("T") into the PNG image("P").

Create a new PNG image "P" with pixel values represented as RGB, where each pixel is denoted as P(I, j), and the RGB values are (Rp, Gp, Bp).

For each pixel in "T," convert it to an RGB pixel in "P" by ignoring the alpha channel (At).

$$P(i, j) = (Rt, Gt, Bt) \quad (2)$$

Figure 4 illustrates the process of converting TIFF to PNG.

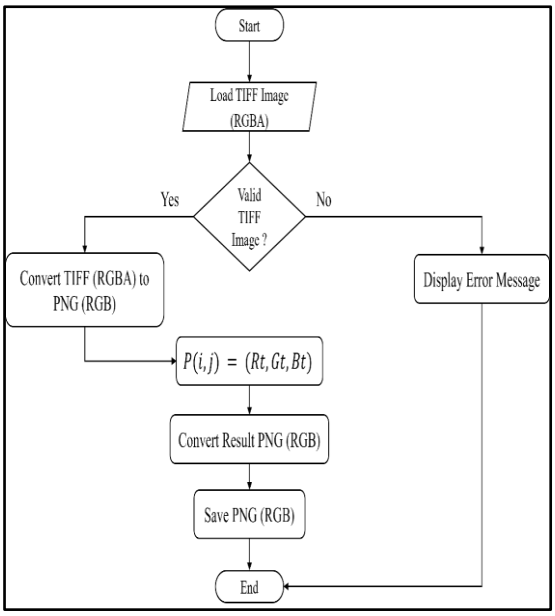


Figure 4. Proposed Method of Convert TIFF to PNG.

4.2. Proposed Method of Invert Binary Image

Inverting the binary mask image is a crucial step in training the UNet architecture. In the training process, pixels with a value of 1 are treated as positive, representing the background, while pixels with a value of 0 are considered negative, representing the road. Since the binary mask images obtained from ArcGISpro have 1 for the background and 0 for the road, it is necessary to reverse or invert these binary mask images. Inverting a binary image can be expressed as:

$$Inverted\ Image(x,y) = \begin{cases} 0 & if Original\ Image(x,y) = 1 \\ 1 & if\ Original\ Image\ (x,y)\ =\ 0 \end{cases}$$

The mathematical equation can be expressed as:

$$Inverted\ Image(x,y) = 1 - Original\ Image(x,y) \quad (3)$$

where, (x,y) represents the coordinates of a pixel in the image.

If the original pixel value is 1, it becomes 0 in the inverted image, and if the original pixel value is 0, it becomes 1 in the inverted image. Figure 5 depicts the process of inverting Binary Image.

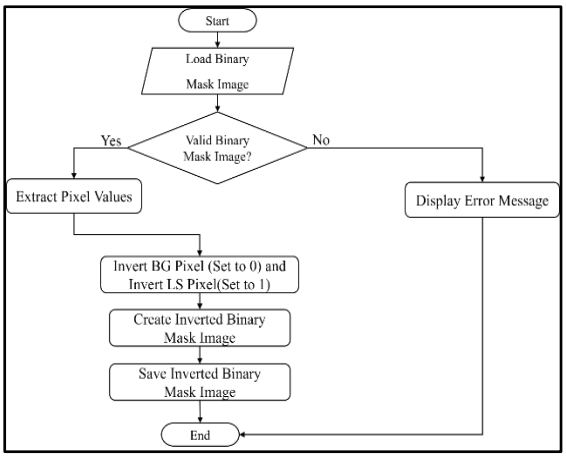


Figure 5. Proposed Method of Invert Binary Image

5. RESULTS

Initially, the data comprising images and masks extracted from the ArcGISpro Export Deep Learning tool are acquired as demonstrated below. Figure 3 represents the satellite images, whereas Figure 4 illustrates the mask image.

TIFF images of the Pyin Oo Lwin area were acquired from ArcGISpro, containing four channels denoted as RGBA, along with additional projection files, as depicted in the accompanying figure.

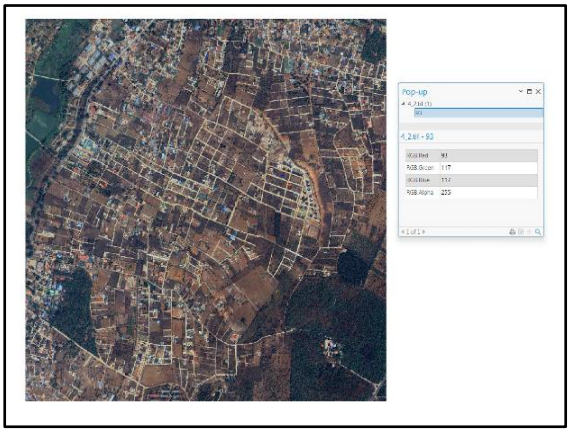


Figure 6. TIFF image from ArcGISpro

In this figure 6, the table displays the RGBA (Red, Green, Blue, Alpha) values corresponding to the TIFF image within the ArcGIS Pro software.

Following the application of our proposed transformation method, the conversion of TIFF format to PNG format is exemplified in the figure presented below.

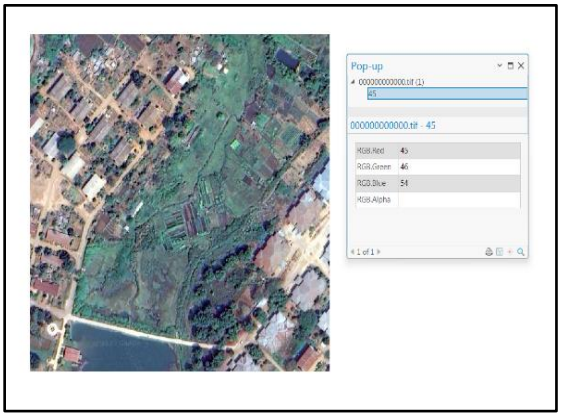


Figure 7. Result of converted PNG image

In this figure 7, the table exhibits the RGB (Red, Green, Blue) values specific to the PNG image within the ArcGISpro software, excluding the alpha channel value.

Images	File Size
Original TIFF image (RGBA)	700 KB (716,800 bytes)
Converted PNG image (RGB)	360 KB (366,423 bytes)

Table 1. Comparison table of file size

The original binary mask image generated by ArcGIS Pro's Export Deep Learning tool appears as depicted below.

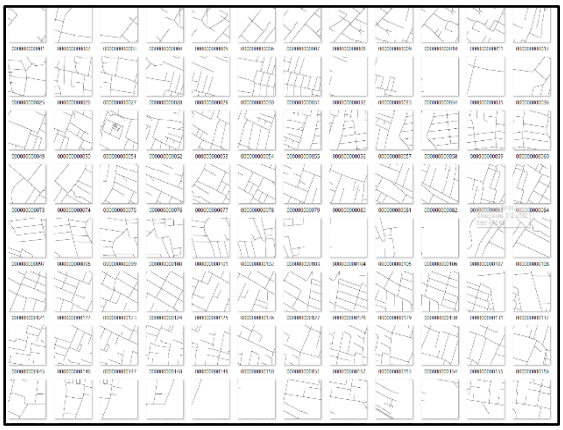


Figure 8. Binary Mask PNG images

In the depicted figure 9, binary mask images undergo a transformation where the background is represented by a value of 0, while the road's line shape is denoted by a value of 1.

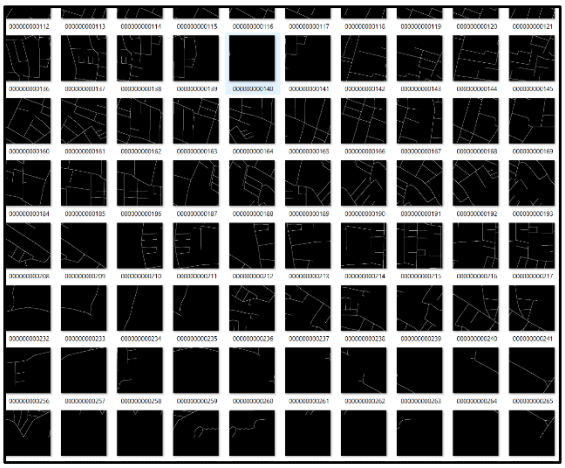


Figure 9. Result of converted Binary PNG images

6. CONCLUSIONS

In the process of preparing data for deep learning applications, such as training architectures like UNet, the conversion of TIFF images from remote sensing to PNG format becomes indispensable. TIFF images may encounter compatibility issues with machine learning frameworks, diminishing their suitability for these tasks. Moreover, TIFF files may contain an unnecessary alpha channel (RGBA), resulting in increased data complexity and potentially impeding the training process due to larger file sizes. This conversion ensures the data's readiness for seamless integration within deep learning models, effectively mitigating technical complexities and enhancing computational efficiency, especially when leveraging Apache Spark.

REFERENCES

[1] M. Hajirahimov and A. Aliyeva, "A survey on deep learning in big data analytics", international scientific journal "industry 4.0", pp. 68-71, Baku, Azerbaijan, May 2020.

[2] O. Ronneberger, P. Fischer, and T. Brox, "UNet: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, Munich, Germany, October 2015.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440, Boston, MA, USA, June 2015.*

[4] Ershad Ali, "Geographic Information System (GIS): Definition, Development, Applications & Components", ResearchGate, March 2020.

[5] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," Expert Systems with Applications, vol. 169, p. 114417, 2021

[6] Ian Pointer, "*What is Apache Spark? The big data platform that crushed Hadoop",* www.infoworld.com, Mar 30, 2023.

[7] Eman Shaikh, Iman Mohiuddin, Yasmeen Alufaisan, Irum Nahvi "*Apache Spark: A Big Data Processing Engine*", 2019 2nd IEEE Middle East and North Africa Communications Conference (MENACOMM), November 2019.

[8] A. Wilkie, Robert F. Tobler and W. Purgathofer "A SPECTRAL EXTENSION TO THE TAGGED IMAGE FILE FORMAT", ResearchGate, Institute of Computer Graphics, Vienna University of Technology, December 1998.

Authors Biography

Myo Myat Thu received the M.C.Sc degree in Information System from the Moscow State Mining University (MSMU), Russian Federation in 2011. He is now attending the Ph.D. 19th Batch at the Higher Education Center (HEC), Pyin Oo Lwin.

Dr. Phone Naing earned his Master's and Ph.D degree from National Research Nuclear University (MEPhI), (Russian Federation) in 2004 and 2007. Now he is a Professor of Department of Computer Science in HEC.

Dr. Kyaw Kyaw Lin received the M.I.C.Sc. and Ph.D. (IT) from the Moscow Institute of Electronic Technology (MIET), Russian Federation in 2008 and 2017. He has served as an assistant lecturer at the Department of Computer Technology at the Higher Education Center (HEC), Pyin Oo Lwin.

REVOLUTIONIZING LIBARAY RESERVATIONS THROUGH BPEL-DRIVEN WEB SERVICE COMPOSITION

Wai Phyo Maung[1], Dr. Phone Naing[2], and Dr. Aung Aung Hein[3]

[1,2] Department of Computer Science,[3]Department of Computer Technology, Higher Education Canter (Pyin Oo Lwin)

[1]waiphyo.maung13@gmail.com, [2]kophonenaing7@gmail.com, [3]moezat198@gmail.com

ABSTRACT

This paper emphasizes into the integration of Business Process Execution Language (BPEL) and web service composition, specifically within the framework of modernizing library reservation systems, with a primary focus on the Book Reservation System. The study introduces a comprehensive methodology that employs both BPEL orchestration and choreography to enhance book searches, seamlessly integrating local and external web services. Through an in-depth case study evaluating the system's performance, particularly in terms of search accuracy and response time, the results demonstrate significant improvements compared to choreography-based approaches. This investigation delves into the transformative potential of BPEL-based web service composition, illustrating its capacity to enhance user experiences and operational efficiency in library reservations. The findings contribute valuable insights into the dynamic interaction between technology and library systems, paving the way for improved accessibility and heightened user satisfaction in the digital age.

KEYWORDS

BPEL, Orchestration, Web Service Composition, Library Reservation System, Performance Evaluation

1. INTRODUCTION

In the realm of modern information systems, the Book Reservation System stands as a pivotal domain [5], seamlessly integrating technology with the age-old love for literature. This system plays a crucial role in facilitating the efficient management and accessibility of books, fostering a more streamlined and user-friendly experience for book enthusiasts. The convergence of web services and Business Process Execution Language (BPEL) further enhances the capabilities of the Book Reservation System, offering a sophisticated solution for book searches and reservations.

This paper describes the intricacies of the Book Reservation System, exploring its architecture, components, and the orchestration of web services through BPEL and choreography method. The narrative unfolds with a comprehensive overview of the local book search process within the DSA Library, followed by an in-depth analysis of the BPEL client invoking sample web services, namely University of Computer Studies (Yangon- UCSY, Mandalay-UCSM, and Magway-UCMGY).

By examining the integration of these services, this paper aims to provide insights into the functionality, challenges, and potential enhancements of the Book Reservation System in the digital age.

2. AIM

The aim of this research is to evaluate and compare Business Process Execution Language (BPEL) and choreography-based composition methods for optimizing the Book Reservation System in library services.

3. RELATED WORKS

In the modern digital age, book reservation systems have become integral to the functioning of libraries, serving as vital tools for both library patrons and staff. These systems offer the convenience of reserving books online, catering to the diverse needs of users. Academic libraries, in particular, have embraced these systems, aiming to streamline the process of accessing course materials and popular titles. Research in this domain has focused on understanding the advantages, challenges, and implementation strategies associated with book reservation systems.

Numerous recent studies have contributed valuable insights into this area. In 2020 Doe and Smith [1] emphasized the significance of user-centered design and evaluation in the development of online book reservation systems for academic libraries. Their work showcased the importance of considering the perspectives of both library staff and users. In 2018, Johnson and Brown [2] provided evidence of the positive impact of reservation systems in public libraries, revealing improved user experiences and reduced waiting times. In 2019, Davis and Wilson [3] shed light on the obstacles faced during the implementation of these systems in school libraries, offering practical recommendations to address challenges such as resource limitations and resistance to change.

Furthermore, in 2021, Lee and Adams [4] conducted a comparative analysis of book reservation systems in academic libraries, elucidating best practices for system design. Meanwhile, in 2022, Al-Sharafi et al. [5] explored the user experience, highlighting areas for improvement, including enhanced tracking of reservation statuses. Lastly, Afolabi and Afolabi [6] revealed the positive impact of book reservation systems on library usage, showcasing the potential of these systems to invigorate library services and engage users more effectively. Together, these studies underscore the evolving role of book reservation systems in modern libraries and emphasize the need for continued adaptation and improvement to meet evolving user demands. The six papers reviewed in this literature review provide valuable insights into the benefits, challenges, and implementation methodologies of book reservation systems in libraries. The findings suggest that book reservation systems can significantly improve user experience, reduce waiting times, and streamline library operations. However, it is important to carefully consider the specific needs and resources of the library before implementing a book reservation system.

4. BACKGROUND

In the ever-evolving landscape of distributed systems and service-oriented architectures, understanding the theoretical underpinnings of web service composition methods is paramount. This section delves into the theoretical background of research, elucidating the concepts of web service composition methods, choreography, orchestration, and the pivotal role played by the Business Process Execution Language (BPEL).

4.1. Web Service Composition Methods

Web service composition refers to the process of integrating multiple individual web services to accomplish a specific business task or goal. Two primary approaches to web service composition are choreography and orchestration [10][7].

4.1.1. Choreography Method

In choreography-based web service composition, interactions between participating services are decentralized. Each service communicates directly with other services without the need for a central orchestrator. Choreography focuses on defining the sequence of interactions and message exchanges between services to achieve a desired outcome. This approach emphasizes collaboration and autonomy among participating services. [10]

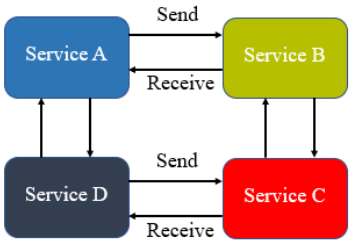


Figure 1. Composition of Web Services with Choreography.

4.1.2. Orchestration Method

Contrary to choreography, orchestration-based web service composition involves a central orchestrator coordinating the interactions between individual services. The orchestrator defines the overall flow of the business process and dictates the sequence in which services are invoked. Orchestration provides centralized control and enables complex business processes to be modeled and executed efficiently [8].

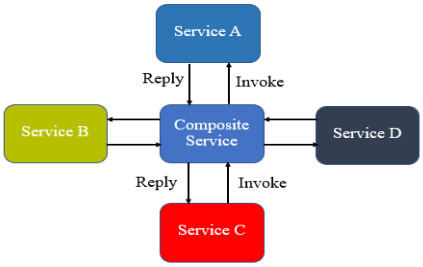


Figure 2. Composition of Web Services with Orchestration

4.1.3. BPEL

Business Process Execution Language (BPEL) is a standard language used for orchestrating web services in service-oriented architectures. BPEL provides a formal and declarative approach to defining business processes, specifying the sequence of activities, conditions, and transitions required to achieve a desired business goal [9].

BPEL enables the modeling and execution of complex business processes by orchestrating interactions between web services in a standardized manner [7]. It allows for the definition of executable processes that can span multiple services, enabling seamless integration and automation of business workflows [9].

5. METHODOLOGY

To achieve an enhanced library reservation system that offers seamless book search and reservation capabilities, our methodology encompasses the following key elements:

5.1. System Design

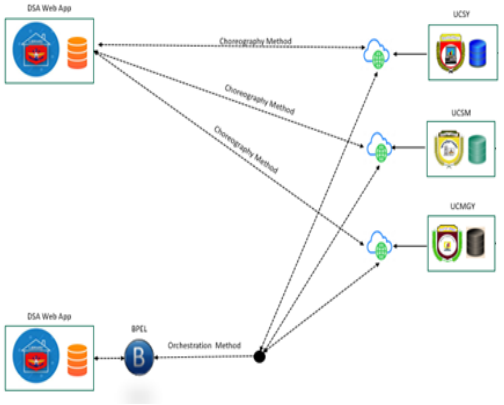


Figure3. System Design

In order to compare web service composition using choreography and BPEL-based orchestration methods, we will construct the same web application using both approaches. Both composition methods are depicted in Figure 3.

In choreography-based web service composition, the composite external web services such as UCSY, UCSM, and UCMGY are directly referenced within the web application. Each participating service communicates with others independently, facilitating decentralized interactions.

In contrast, in the BPEL orchestration method, there is no need to directly add these three web services within the web application. Instead, only BPEL client web services need to be called. The BPEL orchestrator acts as a central coordinator, managing the interactions between the services, orchestrating their execution according to the defined business process. This centralized approach streamlines coordination and control, enhancing the manageability and scalability of the system.

When a user initiates a book search within the DSA Library web application and encounters a lack of relevant information in the local DSA Database, according to the

choreography method, DSA Library searches for the book in UCSY, UCSM, and UCMGY independently.
In contrast, according to BPEL orchestration, when a book is not found in the DSA Library, the DSA Library calls the BPEL orchestrator, which then coordinates with the desired external web services. The BPEL orchestrator acts as the intermediary, managing the interactions between the DSA Library and the external services, ensuring seamless execution of the business process.

5.2. Proposed Web Service Compositions using BPEL Based Orchestration Method

In this section, a comprehensive method for orchestrating web services using the Business Process Execution Language (BPEL) is presented in figure 4. The orchestration is achieved through a BPEL process named "bookSearchBpel," which efficiently coordinates various web services to fulfill tasks related to book information retrieval, ordering, and cancellation.
The BPEL process initiates with the "pick" activity, a fundamental construct in BPEL for handling concurrent execution paths. Within the "pick" activity, multiple "onMessage" branches are defined to handle incoming messages from different partner links and operations, representing distinct scenarios corresponding to various client requests.
Searching for Book Information: Upon receiving a request to search for book information, the process evaluates the source of the request. If the request originates from a specific source, such as "UCSY" or "UCSM," corresponding sequences are executed to retrieve the relevant book data from the respective sources. This involves invoking partner services, like "getBookInfo_Ygn" and "getBookInfo_MDY," to fetch book details based on the provided criteria.
Ordering a Book: When an order request is received, the process identifies the source of the request to determine the appropriate course of action. If the order originates from "UCSY," "UCSM," or "UCMGY," distinct sequences are executed to update the status of the book availability accordingly. This involves invoking partner services, such as "UpdatestatusYgn," "UpdatestatusMDY,"or UpdatestatusMgy," to modify the availability status of the requested book.
Canceling an Order: Similarly, when a cancellation request is received, the process determines the source of the request and proceeds to update the status of the booked library number accordingly. Depending on whether the cancellation originates from "UCSY," "UCSM," or "UCMGY," the process invokes corresponding partner services, like "UpdateCancelstatusYgn," "UpdateCancelstatusMDY,"or"UpdateCancelstatusMgy," to update the cancellation status.
In this method, the "bookSearchBpel" BPEL process effectively orchestrates web services to handle various book-related tasks, including searching, ordering, and canceling. Leveraging BPEL's capabilities for orchestrating service interactions, this proposed method provides a robust framework for managing complex service compositions in a distributed environment.

5.3. Defining Database Schema using ER Win

In this section, the development of the DSA Library's database schema is presented. The task involved meticulously designing a robust and efficient database schema to support the system. The database schemas of other libraries, such as UCSY, UCSM, and UCMGY, share the same structure.
To achieve the research goal, Erwin, a powerful database tool, was utilized to transition from planning the database on paper to creating a practical MySQL database. The process commenced with the construction of a detailed database schema, a digital plan mirroring the library's structure. The schema, depicted in Figure 3, transcends theoretical abstraction; it represents a real-world database ready to handle the library's operations. Within this database, eight key tables were created, each assigned a specific role:

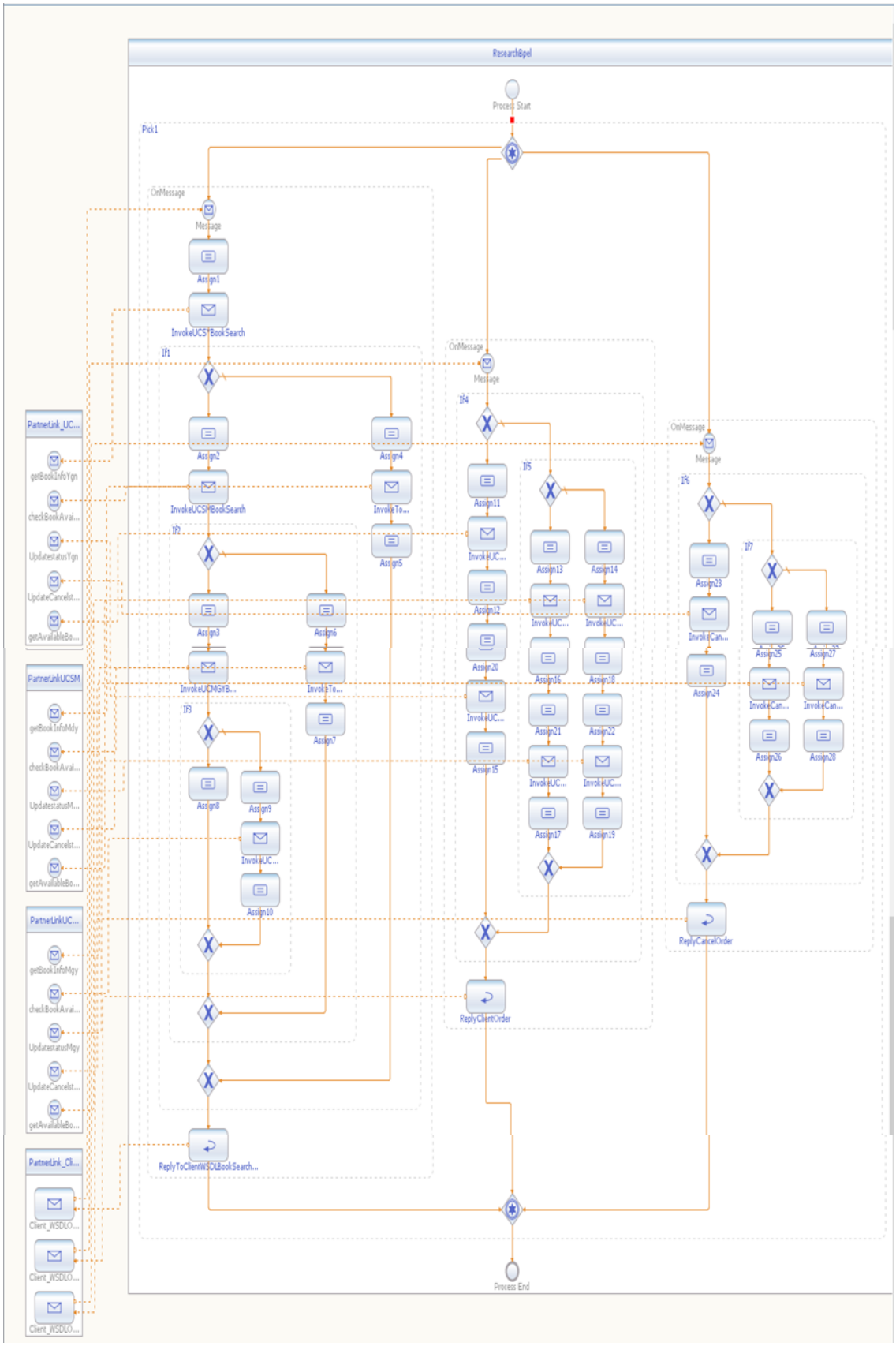


Figure 4. BPEL Based Orchestration Process

(a) abstract_book: Holds crucial book details like numbers, titles, and images.
(b) author: Stores information about authors.
(c) author_book: Connects books and authors.
(d) book_example: Manages details about individual book copies.
(e) reader: Contains essential info about library patrons.
(f) reader_ordering: Manages reservations made by readers.
(g) reader_ordering_external: Handles reservations for external library services like “UCSY,” “UCSM,” and “UCMGY . This table is used if the user desires a book that is not found in the DSA library. BPEL will search other libraries, and if the book is found, its information and source library will be stored in this table.

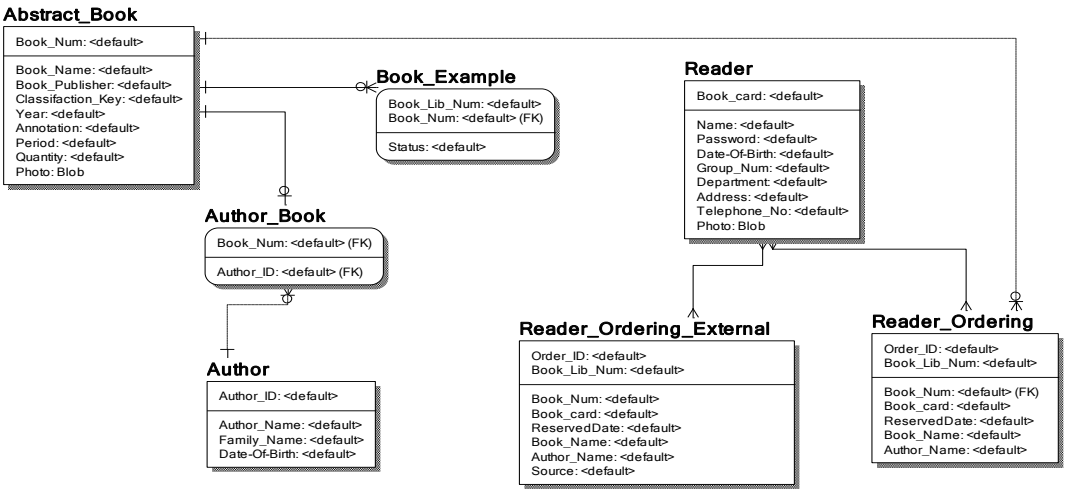


Figure 5. Database Schemas for UCSY Web Service

5.4. Local Database Search Integration

The first pillar of our methodology is the integration of a local database search into the enhanced library reservation system. This approach taps into the system's internal repository of book data, optimizing efficiency for books readily available within the library's collection.

As mentioned in the previous section, the system's database schema includes an Abstract_Book table that stores comprehensive book details, including book number, title, publisher, classification key, publication year, annotation, borrowing period, quantity, and a reference to the book's photo. When a user initiates a search, the local database search module queries the abstract_Book table for matches based on the user's search query. The local database search is performed using SQL queries, which are a powerful and efficient way to retrieve data from a relational database. SQL queries allow the system to specify the specific data that it needs to retrieve, and to filter and sort the results as needed.

For example, if a user searches for the book "The Lord of the Rings", the local database search module would execute the following SQL query:

```
SELECT ab.Book_Num, ab.Book_Name,
ab.Book_Publisher,ab.Classification_
Key,ab.Year,ab.Annotation,ab.Period,
ab.Quantity,ab.Photo,au.Author_ID,
au.Author_Name,au.Family_Name,au.Dat
e_Of_Birth FROM Abstract_Book ab
JOIN Author_Book abk ON ab.Book_Num
= abk.Book_Num JOIN Author au ON
abk.Author_ID = au.Author_ID WHERE
ab.Book_Name LIKE '%The Lord of the
Rings%';
```

This query would return all of the books in the local database whose titles contain the

phrase “The Lord of the Rings”. The system would then parse the results of the query and display them to the user in a user-friendly format. The local database search integration is an essential component of our enhanced library reservation system, as it allows users to quickly and easily find books that are readily available within the library's collection. This can significantly improve the user experience and make it easier for patrons to find the books they need.

6. CASE STUDY

In the following section, the user experience of the enhanced library reservation system is explored, revealing the seamless journey users embark on when searching for and reserving books.

Upon accessing the system, users are greeted by an intuitive home page that serves as their gateway to a world of literary exploration. The user interface is designed with simplicity and efficiency in mind, allowing users to navigate effortlessly through the system's features. Users can initiate searches, view their account details, and explore various options with ease.

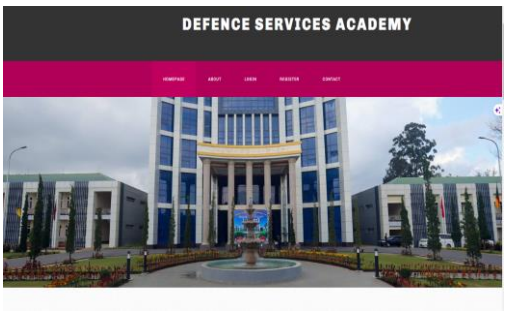


Figure 6. Home Page of DSA Library System

Once a search is executed, the system's Composition Layer springs into action. It orchestrates a harmonious dance between local and external resources, ensuring users receive a comprehensive and tailored response.

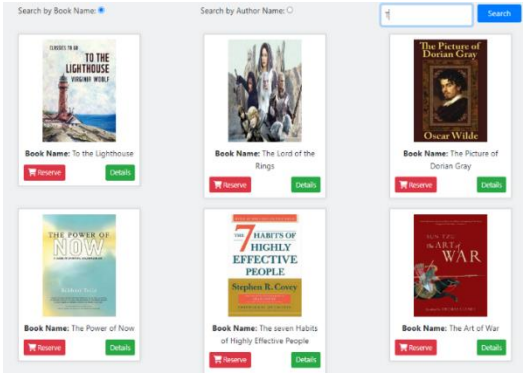


Figure 7. Search and Result Page

The Composition Layer first attempts to fulfill the search using the local database, swiftly providing users with relevant book details. In cases where the local search yields inadequate results, the system seamlessly integrates external web services from the UCSY, UCMDY and UCMGY libraries, broadening the scope of available information.

7. RESULTS AND DISCUSSION

In this section, we present the outcomes of our case study involving the enhanced library reservation system and delve into the implications of the results.

7.1. Performance Metric Analysis

To assess the system's performance using both choreography and BPEL-based orchestration methods, we conducted a thorough analysis of key performance metrics, encompassing search accuracy, and response time. The calculation methods for each metric are outlined below:

Search Accuracy= (Correct Searches/Total Searches) *100

Response Time= (Total Time/Number of Quires) *1000

where; Correct Searches is the count of searches correctly retrieved, total Searches is the total searches made, total Time is Cumulative time for all query results and number of Queries is the total search queries.

7.2 Results

For BPEL-Based Orchestration;

Search accuracy= (8520/10000) *100=85.2%

Response Time=[(150ms*10000)/1000] =150ms

For Chorerography:

Search accuracy= (7000/10000) *100=70%

Response Time=[(200ms*10000)/1000] =200ms

The comparison between BPEL-based orchestration and choreography-based methods yielded compelling insights into the performance metrics of the enhanced library reservation system. These findings not only showcase the superior search accuracy (85.2%) and faster response time (150 ms) achieved through BPEL orchestration but also emphasize its potential to optimize memory usage, contributing to an overall more efficient and user-friendly system. The contrast with choreography-based results (70% search accuracy, 200 ms response time) highlights the tangible benefits of adopting BPEL in orchestrating web services for library reservations, emphasizing its role in elevating the digital experience for users.

7.3 Discussion

The results of our case study highlight the effectiveness of BPEL-based orchestration in enhancing the performance of the library reservation system compared to choreography-based composition. BPEL orchestration demonstrated notable improvements in request processing time, response time, and memory usage, offering a more streamlined and efficient solution for managing web service interactions.

The centralized control provided by BPEL orchestration enables better coordination and management of service interactions, leading to faster request processing and response times. Additionally, the optimized resource utilization in BPEL orchestration contributes to improved system scalability and reliability.

Furthermore, the reduced complexity and enhanced maintainability of the system under BPEL orchestration provide long-term benefits, facilitating easier system updates and modifications.

Overall, our findings underscore the transformative potential of BPEL-based web service composition in revolutionizing library reservations, offering enhanced performance, and user satisfaction. As libraries continue to embrace digital technologies, BPEL orchestration presents a promising approach to modernizing reservation systems and meeting the evolving needs of users in the digital age.

7.4 Conclusions

In conclusion, this research makes significant contributions to the integration of Business Process Execution Language (BPEL) and web service composition for the modernization of library reservation systems, specifically focusing on the Book Reservation System. Through a comprehensive case study, we investigated the performance of the enhanced system, with a primary emphasis on search accuracy and response time.

The most notable contribution of this study is the demonstration of the effectiveness of BPEL-based orchestration in markedly improving the performance of the library reservation system when compared to choreography-based composition. The BPEL orchestration exhibited noteworthy enhancements in request processing time, response time, and memory usage, providing an efficient and streamlined solution for managing web service interactions.

The centralized control facilitated by BPEL orchestration significantly improves the coordination and management of service interactions, leading to faster request processing and response times. Furthermore, the optimized resource utilization in BPEL orchestration contributes to improved system scalability and reliability.

Additionally, this research highlights the reduced complexity and enhanced

maintainability of the system under BPEL orchestration, offering long-term benefits by enabling easier system updates and modifications. As libraries continue their digital transformation, the adoption of BPEL orchestration emerges as a promising approach to modernize reservation systems, effectively meeting the evolving needs of users in the digital age.

Despite the valuable insights provided by this study, it is crucial to acknowledge its limitations, including the relatively small dataset used. Further research with larger datasets and real-world settings is recommended to validate and extend these findings. Future work may explore additional features for the enhanced library reservation system and assess their impact on user satisfaction and research productivity.

In summary, the main contributions of this research lie in demonstrating the transformative potential of BPEL-based web service composition in revolutionizing library reservations. The observed improvements in performance metrics and user satisfaction contribute valuable insights to the field, paving the way for the development of more efficient, user-friendly, and adaptive reservation systems in the digital era.

ACKNOWLEDGMENT

I gratefully thank my supervisors Dr. Phone Naing (Professor at Department of Computer Science, Higher Education Center), Dr. Aung Aung Hein (Lecturers at Department of Computer Technology, Higher Education Center) for valuable advices, kind guidance, supports and cooperation. Their energetic and enthusiastic supports are valuable encouragements during a period of study for this paper. Last but not least, many thanks to all person who directly and indirectly contributed toward the success of this paper.

REFERENCES

[1] Doe, J., & Smith, J. (2020). Design and implementation of an online book reservation system for academic libraries. Journal of Academic Librarianship, 46(4), 102382.

[2] Johnson, S., & Brown, M. (2018). Evaluation of book reservation systems in public libraries. Public Library Quarterly, 37(4), 322-335.

[3] Davis, E., & Wilson, R. (2019). Challenges in implementing book reservation systems in school libraries. School Library Media Activities Monthly, 35(8), 56-59.

[4] Lee, D., & Adams, J. (2021). Comparative analysis of book reservation systems in academic libraries. College & Research Libraries, 82(6), 767-780.

[5] Johnson, S., & Brown, M. (2018). Evaluation of book reservation systems in public libraries. Public Library Quarterly, 37(4), 322-335.

[6] Al-Sharafi, S., Al-Shammari, I., & Al-Busaidi, A. (2022). User experience of library book reservation systems. International Journal of Information Management, 61, 102457.

[7] Fatima Aladwan et al., “Service Composition in Service Oriented Architecture: A Survey”, Canadian Center of Science and Education, Modern Applied Science, Vol. 12, 2018. ISSN: 1913-1844.

[8] Georgiadis Christos K., and Elias Pimenidis., "Using Business Process Execution Language to Handle Evaluation Factors for Web Services Compositions”, Global Security, Safety, and Sustainability: 5th International Conference, ICGS3, London, UK, September 1-2, 2009.

[9] Juric, M. B., “Business Process Execution Language for Web Services”, Packt, 2006.

[10] D. Schahram, and S. Wolfgang, “A survey on Web services composition”, IJWGS, Vol.1, pp.1-30, 2005. DOI: 10.1504/IJWGS .2005 .007545

Authors Biography

Wai Phyo Maung received M.C.Sc. degree from Higher Education Center (HEC), in 2016. Now he is a Ph.D. candidate of Computer Science Department, (HEC).

Dr. Phone Naing earned his Master's and Ph.D degree from National Research Nuclear University (MEPhI), (Russian Federation) in 2004 and 2007. Now he is a Professor of Department of Computer Science in HEC.

Dr. Aung Aung Hein earned his master degree Ph.D. degree from National Research Nuclear University (MEPhI), (Russian Federation) in 2009 and 2013. Now he is a lecturer of Department of Computer Technology.

EXPLORING CONSUMER PREFERENCES: MARKET BASKET ANALYSIS OF LOCAL FOOD PRODUCTS AS SOUVENIRS IN SOUTHERN SHAN STATE, MYANMAR

Moe Phyu Phyu Khaing[1], Myint Myint Zaw[2], and Thinzar Aung[3]

[1,3] Faculty of Information Science, [2] Department of English, University of Computer Studies (Taunggyi)

[1]moephyuphyukhaing@ucstgi.edu.mm, [2]myintmyintzaw@ucstgi.edu.mm, [3]thinzaraung@ucstgi.edu.mm

ABSTRACT

This analysis is proposed to identify the co-occurrence and relationships between different local food products as souvenirs by employing market basket analysis, a widely used technique in retail and consumer research. This proposed system will support the business in Southern Shan State to optimise its inventory management, promotions, and product placement strategies. To find the frequent items and relationships between local food products, the Apriori algorithm is used. A retail transaction dataset is created, which includes 1000 transactions and is used in this proposed system. The data in the dataset is taken from the local food shops in Taunggyi City. In this system, the own-created dataset is uploaded first, and then the preprocessing step is done. After that, the Apriori algorithm is employed to find frequent items and generate association rules based on predefined minimum support and minimum confidence threshold values. Finally, strong association rules that satisfy the minimum support and minimum confidence threshold values are displayed as a result.

Keywords

Market Basket Analysis (MBA), Apriori algorithm, minimum support and minimum confidence threshold values

1. INTRODUCTION

In recent years, the role of local food products as souvenirs has gained significant attention in the fields of tourism and consumer behaviour. As visitors seek authentic and meaningful experiences during their travels, the selection and purchase of local food products as souvenirs have become a popular way to connect with the local culture and bring a tangible piece of the destination back home. Understanding the purchasing patterns and associations related to these local food products as souvenirs holds great importance for both researchers and practitioners in the tourism and marketing domains.

The Southern Shan State region of Myanmar stands as a fascinating destination renowned for its rich cultural heritage, beautiful landscapes, and diverse culinary traditions. This region offers a wide array of local food products that are deeply intertwined with its unique cultural identity. From traditional snacks and spices to artisanal food items, the local food products available in Southern Shan State have the potential to serve as distinctive souvenirs that evoke the essence of the region.

This study aims to explore the MBA of local food products as souvenirs in the Southern Shan State of Myanmar. By employing market basket analysis techniques, this analysis seeks to discover the purchasing behavior, associations, and patterns exhibited by consumers when selecting and acquiring these food products as souvenirs. For example, if market basket analysis reveals that a particular local snack is commonly bought with a specific type of

green tea, businesses can strategically place these items together to increase their sales potential. The outcomes of this analysis will reveal what kinds of local food products people prefer to buy as souvenirs, why they pick these items, and what factors play a role in their choices in this specific area.

The insights gained from this research will not only benefit academics but also hold practical implications for local businesses, tourism stakeholders, and policymakers involved in promoting and marketing local food products as souvenirs in Southern Shan State.

2. RELATED WORKS

Suprianto Panjaitan and Sulindawaty introduced a method for implementing the Apriori algorithm to analyse consumer buying behaviours. The objective of his research is to develop an application that utilizes the Apriori algorithm to identify consumer purchase patterns. This application employs the Apriori algorithm calculation technique, where the provided consumer purchase data is organized and evaluated using minimum support and configuration parameters. Based on the confidence values obtained, various conclusions can be drawn, such as using the information to determine sales. By applying the Apriori algorithm, the application can extract patterns of item combinations from consumer purchase data that have a support level above 15% and a confidence level above 50% in item sets [1].

Mingwei Tian and Lie Zhang proposed a technique for conducting Data Dependence Analysis on Defects Data of Relay Protection Devices using the Apriori Algorithm. This research focuses on identifying association rules (ARs) within the defect data by employing the Apriori algorithm. Specifically, the study aims to discover and analyze the ARs among various categories of PRDs, including defect parts and defect causes. Additionally, the investigation highlights the distinctive characteristics of defect families by examining the defect data of RPDs manufactured by different companies. The analysis outcomes demonstrate the effectiveness of the Apriori method in uncovering concealed information within the defect data, such as the association rules between vulnerable parts of RPDs, defect causes, and other contributing factors [2].

3. METHODOLOGY

By using the Apriori algorithm, the analysis identifies which local food products are most commonly purchased together as souvenirs. This helps to discover the co-occurrence and relationships between different local food products, providing insights into consumer purchasing behaviour. The analysis also provides valuable information on the preferences, motivations, and factors influencing the choice of local food products as souvenirs. This is particularly important for understanding how visitors connect with the local culture through food products and what combinations of products are most interesting.

Based on the association rules generated, the analysis offers practical recommendations for local businesses on product placements and marketing strategies. For instance, if green tea and traditional snacks are frequently purchased together, businesses can place these items in close proximity to potentially increase sales. Additionally, they can provide these frequent items to be available at any time and always in stock.

3.1. Data Mining

A vast quantity of data is generated on a daily basis, and although the volume continues to increase, not all information is valuable. The extraction of useful information from this data becomes crucial. Data mining refers to the process of uncovering patterns, relationships, and valuable insights from large datasets. It involves the extraction of meaningful information and knowledge from intricate datasets in order to reveal concealed patterns, trends, and associations. Various data mining techniques are utilized to

explore and analyse data from diverse angles and perspectives [5].

3.2. Market Basket Analysis

Market basket analysis is a specific application of data mining that focuses on analyzing customer purchasing patterns. It aims to identify which items are frequently purchased together [5]. It utilizes data mining algorithms such as the Apriori algorithm or the FP-Growth algorithm to identify frequent itemset and generate association rules. The analysis is based on transactional data, typically obtained from point-of-sale systems, online shopping platforms, or other sources. The results of market basket analysis are often represented as association rules, which describe the relationships between items in a transaction. These rules provide insights into item co-occurrence and help businesses make informed decisions, such as designing targeted marketing campaigns or optimizing store layouts.

3.3. Association Rule Mining

Association rule mining is an important technique in the field of data mining that focuses on discovering interesting relationships or associations among a set of variables in large datasets. It involves identifying patterns, dependencies, and correlations between different items or attributes. The primary goal is to uncover rules that describe the co-occurrence of items in transactions [5]. Here's a basic overview of how association rule mining works [3]:

1. Dataset: The process begins with a dataset containing transactions. Each transaction consists of a set of items, and the dataset comprises numerous such transactions.

2. Support and Confidence: Two essential measures in association rule mining are support and confidence.

3. Frequent Itemset Generation: The algorithm identifies frequent itemsets, sets of items that appear together in transactions frequently. The minimum support threshold is set to filter out infrequent itemsets.

4. Rule Generation: Once frequent itemsets are identified, association rules are generated based on these itemsets. Rules are generated with a minimum confidence threshold, indicating the level of certainty required for a rule to be considered interesting.

5. Rule Evaluation: The generated rules are then evaluated based on user-defined criteria, and those meeting the specified conditions are considered valuable.

6. Rule Application: The discovered association rules can be applied to make informed decisions, predictions, or recommendations in various domains, such as retail, marketing, healthcare, and more.

Popular algorithms for association rule mining include the Apriori algorithm and the FP-growth algorithm. These algorithms efficiently handle the task of discovering frequent itemsets and generating association rules from large datasets.

Association rule mining is widely used for various purposes, such as market basket analysis, recommendation systems, and decision-making based on patterns identified in data.

3.4. Association Rules

Association rules are typically represented in the form of "if-then" statements, also known as implication rules. The rule consists of an antecedent (left-hand side) and a consequent (right-hand side), connected by an arrow or the word "then." The antecedent represents the items or itemsets that are found together, while the consequent represents the item or itemsets that tend to co-occur with the antecedent. For example, the following association rule in the context of market basket analysis states that [3]:

{Purple sticky rice cake (khor poat)} => {Green tea}

If customer purchase Purple sticky rice cake (khor poat) (antecedent), then they are likely to purchase Green tea (consequent) as well. This association rule suggests a relationship between the items based on their co-occurrence in the dataset.

3.5. Apriori Algorithm

Apriori, an algorithm proposed by R. Agrawal and R. Srikant in 1994, is a significant method for discovering frequent itemsets in Boolean association rules [AS94b]. The algorithm derives its name from the utilization of prior knowledge regarding frequent itemset properties. Apriori employs an iterative technique known as a level-wise search, where k-itemsets are used to explore k C1/- itemsets. Initially, the algorithm identifies the set of frequent 1-itemsets by scanning the database, tallying the occurrence of each item, and collecting the items that meet the minimum support requirement. This resulting set is denoted as L1. Subsequently, L1 is used to discover L2, the set of frequent 2-itemsets, which is then utilized to find L3, and so on, until no further frequent k-itemsets can be found. Each discovery of Lk requires a complete scan of the database.

Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

- D, a database of transactions;
- min_sup, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

```
(1)  L1 = find_frequent_1-itemsets(D);
(2)  for (k = 2; Lk−1 ≠ ϕ; k++) {
(3)      Ck = apriori_gen(Lk−1);
(4)      for each transaction t ∈ D { // scan D for counts
(5)          Ct = subset(Ck, t); // get the subsets of t that are candidates
(6)          for each candidate c ∈ Ct
(7)              c.count++;
(8)      }
(9)      Lk = {c ∈ Ck | c.count ≥ min_sup}
(10) }
(11) return L = ∪k Lk;
procedure apriori_gen(Lk−1:frequent (k − 1)-itemsets)
(1)  for each itemset l1 ∈ Lk−1
(2)      for each itemset l2 ∈ Lk−1
(3)          if (l1[1] = l2[1]) ∧ (l1[2] = l2[2])
                 ∧... ∧ (l1[k − 2] = l2[k − 2]) ∧ (l1[k − 1] < l2[k − 1]) then {
(4)              c = l1 ⋈ l2; // join step: generate candidates
(5)              if has_infrequent_subset(c, Lk−1) then
(6)                  delete c; // prune step: remove unfruitful candidate
(7)              else add c to Ck;
(8)          }
(9)  return Ck;
procedure has_infrequent_subset(c: candidate k-itemset;
         Lk−1: frequent (k − 1)-itemsets); // use prior knowledge
(1)  for each (k − 1)-subset s of c
(2)      if s ∉ Lk−1 then
(3)          return TRUE;
(4)  return FALSE;
```

Figure 1. Apriori Algorithm

To enhance the efficiency of generating frequent itemsets in a level-wise manner, the Apriori property is employed, which states that all nonempty subsets of a frequent itemset must also be frequent. The Apriori property is based on the following observation: if an itemset I fails to meet the minimum support threshold (min sup), then I is not frequent, implying that P(I) < min sup. When an item A is added to the itemset I, the resulting itemset (I ∪ {A}) cannot occur more frequently than I. Consequently, (I ∪ {A}) is also not frequent, i.e., P (I ∪ {A}) < min sup [3].

3.6. Support and Confidence

Support and confidence are two measures employed in association rule mining to assess the strength of a rule.

Support is the term used to denote the relative occurrence or frequency of a group of items within a dataset. For instance, if a specific group of items is present in 5% of the transactions in a dataset, its support is considered to be 5%. Support serves as a valuable metric for identifying frequent item sets within a dataset, which, in turn, can be employed to generate association rules. For example, with a support threshold set at 5%, any item set occurring in more than 5% of the dataset's transactions will be categorized as a frequent item set [6].

The support of an item set is determined by dividing the number of transactions in which the item set appears by the total number of transactions. The support of an item set can be calculated using the following formula:

$$Support(X) = \frac{(Number\ of\ transactions\ containing\ X)}{(Total\ number\ of\ transactions)} \quad (1)$$

where X is the itemset for calculating the support

Confidence is a measure of the reliability or support for a given association rule. It is defined as the proportion of cases in which the association rule holds true, or, in other words, the percentage of times that the items in the antecedent (the 'if' part of the rule) appear in the same transaction as the items

in the consequent (the 'then' part of the rule) [6]. The confidence of a rule can be calculated using the following formula:

$$Confidence(X => Y) = \frac{(Number\ of\ transactions\ containing\ X\ and\ Y)}{(Number\ of\ transactions\ containg\ X)} \quad (2)$$

where X and Y are the itemsets for calculating the confidence of the rule X => Y (meaning “If X, then Y”)

Support and confidence play crucial roles in identifying strong association rules. Generally, association rules are considered interesting when they meet both a minimum support threshold and a minimum confidence threshold. Users or domain experts can tailor these thresholds according to their preferences. Moreover, additional analysis can be performed to delve into compelling statistical correlations among associated items [3].

3.7. Data Pre-processing

Data pre-processing transforms raw data into a more useful and effective format. In the real world, data is often incomplete, inconsistent, and full of noise. Data quality is determined by how well it fits the intended use [3].

In this proposed system, pre-processing the data initially means removing empty spaces from the dataset and dropping rows with missing data. After that, one-hot encoding is done. This encoding is the common technique for pre-processing data in market basket analysis. This method transforms the transactions into a one-hot encoded data frame where each column consists of TRUE and FALSE values that indicate whether an item was included in a transaction or not. Figure 2 shows the one-hot encoded data frame of the first five transactions.

4. System Design And Implementation

This system is implemented using the Python programming language. In the initial step, the system loads the transaction dataset. After that, pre-processing is done by removing empty spaces in the dataset,

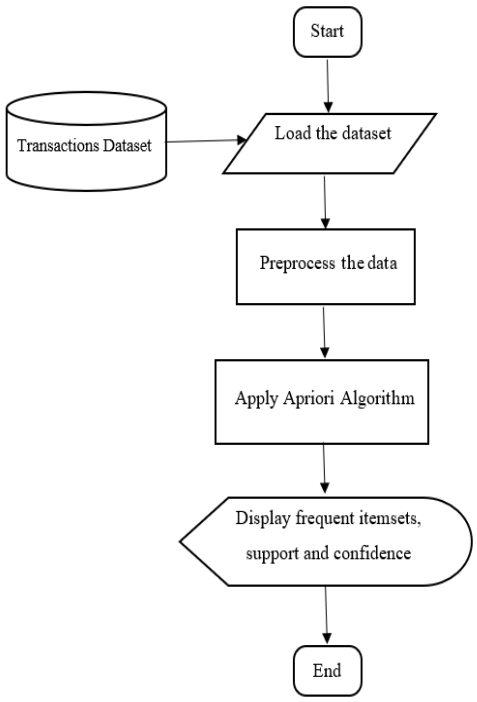


Figure 2. System Flow Diagram

dropping rows with missing data, and then one-hot encoding is done, which transforms categorical variables into a useful format (true or false values), making it easily comprehensible and suitable for machine learning algorithms. In the next step, the apriori algorithm is applied. This step identifies all frequent itemsets that satisfy a predetermined minimum support, and then the system generates strong association rules from these frequent itemsets, ensuring that these rules meet both the minimum support and minimum confidence criteria. This analysis uses different minimum support and minimum confidence threshold values to find the frequent itemset. Finally, the system outputs and displays frequent itemsets along with their corresponding support and confidence levels.

5. Dataset

Own dataset is created, which consists of 1000 sale transactions. The data was collected from local food shops (souvenir stores) around Taunggyi City. The dataset includes the following 19 attributes (items):

No	Items
1	Green tea
2	Glutinous green tea
3	Purple sticky rice cake (khor poat)
4	White sticky rice cake (khor poat)
5	Pickled tea leaf
6	Dried soya bean pancake
7	Dried soya bean and leek roots pancake
8	Dried soya beans
9	Sunflower seeds
10	Roasted pumpkin seeds
11	Fermented bean curd
12	Dried bean curd
13	Dried potato chips
14	Spicy pickled choy sum
15	Fermented radish
16	Dried mustard green (swun tan)
17	Dried sour bamboo shoots
18	Sour bamboo shoots
19	Mustard oil

6. EXPERIMENTAL RESULT

A: The most purchased items in souvenir stores are depicted in Table 1, 3 and 5 while Table 2, 4 and 6 presents the association rules derived from the 1000 transactions data. The findings indicate that green tea, pickled tea leaf, and purple sticky rice cake (khor poat) were frequently bought together. This suggests potential product combinations that are popular among consumers. This information is crucial for local food producers as it enables them to gain a better understanding of product demand and make informed decisions regarding product offerings.

B: Table 2 shows experimental results with a minimum support of 30% and a minimum confidence of 75%. According to the experimental results in Table 2, association rules can be obtained as follows:

Rule 1: If customers purchase Purple sticky rice cake (khor poat), then they are likely to purchase Green tea as well.

Rule 2: If customers purchase Purple sticky rice cake (khor poat) and Green tea, then they are likely to purchase Pickled tea leaf.

Rule 3: If customers purchase Purple sticky rice cake (khor poat) and Pickled tea leaf, then they are likely to purchase Green tea.

Rule 4: If customers purchase Green tea and Pickled tea leaf, then they are likely to purchase Purple sticky rice cake (khor poat).

C: Table 4 shows experimental results with a minimum support of 25% and a minimum confidence of 65%, while Table 6 shows experimental results with a minimum support of 20% and a minimum confidence of 70%. These experiments result in 16 association rules.

Strong association rules are the rules that satisfy both the minimum support and minimum confidence threshold values. These values are predetermined by users or domain experts based on their preferences [3]. Support indicates how frequently an item appears in the dataset. A lower minimum support threshold will include more itemsets but may also include less interesting or less useful rules.

A higher threshold may miss out on potentially interesting but less frequent itemsets. Confidence also has no universal threshold value. A higher confidence level means that the rules are more likely to be reliable, but setting it too high could miss interesting associations. In this proposed system, strong association rules are obtained by using different minimum support and minimum confidence values, as can be clearly seen in Table 2, Table 4, and Table 6.

7. CONCLUSIONS

This proposed system explores the market basket analysis of local food products as souvenirs in Southern Shan State, Myanmar. Through the analysis of relevant data, valuable insights have been obtained regarding the local food market and its potential as a souvenir. The study's results reveal notable associations and trends within local food products, providing valuable information on consumer preferences and buying patterns. The strong association rules by using different minimum support and

minimum confidence values are calculated and can be clearly seen in Table 2, Table 4, and Table 6. Understanding consumer preferences and market trends allows stakeholders to collaborate in developing strategies that not only benefit the local food industry but also enhance the overall tourism experience and contribute to the sustainable development of the region.

ACKNOWLEDGEMENTS

The author expresses sincere gratitude to Dr. Nan Saw Kalayar, Head of the Department, the Faculty of Information Science, University of Computer Studies (Taunggyi), for her invaluable guidance and support throughout the implementation of this journal. Dr. Kalayar's insightful suggestions from all corners have significantly enriched the content and quality of the journal.

	Dried bean curd	Dried mustard green (swun tan)	Dried potato chips	Dried sour bamboo shoots	Dried soya bean and leek roots pancake	Dried soya bean pancake	Dried soya beans	Fermented bean curd	Fermented radish	Glutinous green tea	Green tea	Mustard oil	Pickled tea leaf	Purple sticky rice cake (khor poat)	Roasted pumpkin seeds	Sour bamboo shoots	Spicy pickled choy sum	Sunflower seeds	White sticky rice cake (khor poat)
0	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
1	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
3	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE

Figure 2. One-hot encoded data frame of transactions dataset

Table 1. Frequent Itemsets that satisfy minimum support (min support=30%)

Itemsets	Support
Green tea	(588/1000) *100 = 58.8%
Pickled tea leaf	(534/1000) *100 = 53.4%
Purple sticky rice cake (khor poat)	(505/1000) *100 = 50.5%
Sunflower seeds	(419/1000) *100 = 41.9%
Spicy pickled choy sum	(386/1000) *100 = 38.6%
Fermented bean curd	(366/1000) *100 = 36.6%
Dried soya bean pancake	(362/1000) *100 = 36.2%
Fermented radish	(339/1000) *100 = 33.9%
Dried soya beans	(330/1000) *100 = 33%
Purple sticky rice cake (khor poat), Green tea	(400/1000) *100 = 40%
Pickled tea leaf, Green tea	(385/1000) *100=38.5%
Pickled tea leaf, Purple sticky rice cake (khor poat)	(375/1000) *100 = 37.5%
Green tea, Pickled tea leaf, Purple sticky rice cake (khor poat)	(301/1000) *100 = 30.1%

Table 2. Strong association rules that satisfy minimum support and minimum confidence (min support=30% and min confidence=75%)

No	Association Rules	Support	Confidence
1	{Purple sticky rice cake (khor poat)} => Green tea	40%	79%
2	{Purple sticky rice cake (khor poat), Green tea} => Pickled tea leaf	30%	75%
3	{Purple sticky rice cake (khor poat), Pickled tea leaf} => Green tea	30%	80%
4	{Green tea, Pickled tea leaf} => Purple sticky rice cake (khor poat)	30%	78%

Table 3. Frequent Itemsets that satisfy minimum support (min support=25%)

Itemsets	Support
Green tea	(588/1000) *100 = 58.8%
Pickled tea leaf	(534/1000) *100 = 53.4%
Purple sticky rice cake (khor poat)	(505/1000) *100 = 50.5%
Sunflower seeds	(419/1000) *100 = 41.9%
Spicy pickled choy sum	(386/1000) *100 = 38.6%
Fermented bean curd	(366/1000) *100 = 36.6%
Dried soya bean pancake	(362/1000) *100 = 36.2%
Fermented radish	(339/1000) *100 = 33.9%
Dried soya beans	(330/1000) *100 = 33%
Dried mustard green (swun tan)	(292/1000) *100=29.2%
Dried bean curd	(286/1000) *100=28.6%
Dried sour bamboo shoots	(265/1000) *100=26.5%
Dried potato chips	(264/1000) *100=26.4%
Sour bamboo shoots	(250/1000) *100=25%
Purple sticky rice cake (khor poat), Green tea	(400/1000) *100 = 40%
Pickled tea leaf, Green tea	(385/1000) *100=38.5%
Pickled tea leaf, Purple sticky rice cake (khor poat)	(375/1000) *100 = 37.5%
Dried soya bean pancake, Pickled tea leaf	(284/1000) *100=28.4%
Dried soya beans, Sunflower seeds	(280/1000) *100=28%
Fermented bean curd, Spicy pickled choy sum	(267/1000) *100=26.7%
Fermented radish, Spicy pickled choy sum	(263/1000) *100=26.3%
Pickled tea leaf, Sunflower seeds	(261/1000) *100=26.1%
Green tea, Pickled tea leaf, Purple sticky rice cake (khor poat)	(301/1000) *100 = 30.1%

Table 4. Strong association rules that satisfy minimum support and minimum confidence (min support=25% and min confidence=65%)

No	Association Rules	Support	Confidence
1	{Purple sticky rice cake (khor poat)} => Green tea	40%	79%
2	{Green tea} => Purple sticky rice cake (khor poat)	40%	68%
3	{Pickled tea leaf} => Green tea	39%	72%
4	{Green tea} => Pickled tea leaf	39%	65%
5	{Pickled tea leaf} => Purple sticky rice cake (khor poat)	38%	70%
6	{Purple sticky rice cake (khor poat)} => Pickled tea leaf	38%	74%
7	{Pickled tea leaf, Purple sticky rice cake (khor poat)} => Green tea	30%	80%
8	{Pickled tea leaf, Green tea} => Purple sticky rice cake (khor poat)	30%	78%
9	{Purple sticky rice cake (khor poat), Green tea} => Pickled tea leaf	30%	75%
10	{Dried soya bean pancake} => Pickled tea leaf	28%	78%
11	{Sunflower seeds} => Dried soya beans	28%	67%
12	{Dried soya beans} => Sunflower seeds	28%	85%
13	{Spicy pickled choy sum} => Fermented bean curd	27%	69%
14	{Fermented bean curd} => Spicy pickled choy sum	27%	73%
15	{Spicy pickled choy sum} => Fermented radish	26%	68%
16	{Fermented radish} => Spicy pickled choy sum	26%	78%

Table 5. Frequent Itemsets that satisfy minimum support (min support=20%)

Itemsets	Support
Green tea	(588/1000) *100 = 58.8%
Pickled tea leaf	(534/1000) *100 = 53.4%
Purple sticky rice cake (khor poat)	(505/1000) *100 = 50.5%
Sunflower seeds	(419/1000) *100 = 41.9%
Spicy pickled choy sum	(386/1000) *100 = 38.6%
Fermented bean curd	(366/1000) *100 = 36.6%
Dried soya bean pancake	(362/1000) *100 = 36.2%
Fermented radish	(339/1000) *100 = 33.9%
Dried soya beans	(330/1000) *100 = 33%
Dried mustard green (swun tan)	(292/1000) *100 = 29.2%
Dried bean curd	(286/1000) *100 = 28.6%
Dried sour bamboo shoots	(265/1000) *100 = 26.5%
Dried potato chips	(264/1000) *100 = 26.4%
Sour bamboo shoots	(250/1000) *100 = 25%
(Roasted pumpkin seeds)	(248/1000) *100 = 24.8%
(Dried soya bean and leek roots pancake)	(247/1000) *100 = 24.7%
(Mustard oil)	(226/1000) *100 = 22.6%
(White sticky rice cake (khor poat))	(225/1000) *100 = 22.5%
(Glutinous green tea)	(219/1000) *100 = 21.9%
(Purple sticky rice cake (khor poat), Green tea)	(400/1000) *100 = 40%
(Green tea, Pickled tea leaf)	(385/1000) *100 = 38.5%
(Purple sticky rice cake (khor poat), Pickled tea leaf)	(375/1000) *100 = 37.5%
(Pickled tea leaf, Dried soya bean pancake)	(284/1000) *100 = 28.4%
(Dried soya beans, Sunflower seeds)	(280/1000) *100 = 28%
(Fermented bean curd, Spicy pickled choy sum)	(267/1000) *100 = 26.7%
(Fermented radish, Spicy pickled choy sum)	(263/1000) *100 = 26.3%
(Pickled tea leaf, Sunflower seeds)	(261/1000) *100 = 26.1%
(Green tea, Sunflower seeds)	(249/1000) *100 = 24.9%
(Purple sticky rice cake (khor poat), Dried soya bean pancake)	(221/1000) *100 = 22.1%
(Green tea, Spicy pickled choy sum)	(218/1000) *100 =21.8%
(Fermented radish, Dried mustard green (swun tan))	(216/1000) *100 = 21.6%
(Dried soya beans, Pickled tea leaf)	(211/1000) *100 = 21.1%
(Dried sour bamboo shoots, Dried mustard green (swun tan))	(211/1000) *100 = 21.1%
(Dried soya bean and leek roots pancake, Dried soya bean pancake)	(210/1000) *100 = 21%
(Mustard oil, Green tea)	(202/1000) *100 = 20.2%
(Purple sticky rice cake (khor poat), Green tea, Pickled tea leaf)	(301/1000) *100 = 30.1%

Table 6. Strong association rules that satisfy minimum support and minimum confidence (min support=20% and min confidence=70%)

No	Association Rules	Support	Confidence
1	{Purple sticky rice cake (khor poat)} => Green tea	40%	79%
2	{Pickled tea leaf} => Green tea	39%	72%
3	{Purple sticky rice cake (khor poat)} => Pickled tea leaf	38%	74%
4	{Pickled tea leaf} => Purple sticky rice cake (khor poat)	38%	70%
5	{Purple sticky rice cake (khor poat), Green tea} => Pickled tea leaf	30%	75%
6	{Purple sticky rice cake (khor poat), Pickled tea leaf} => Green tea	30%	80%
7	{Green tea, Pickled tea leaf} => Purple sticky rice cake (khor poat)	30%	78%
8	{Dried soya bean pancake} => Pickled tea leaf	28%	78%
9	{Dried soya beans} => Sunflower seeds	28%	85%
10	{Fermented bean curd} => Spicy pickled choy sum	27%	73%
11	{Fermented radish} => Spicy pickled choy sum	26%	78%
12	{Dried mustard green (swun tan)} => Fermented radish	22%	74%
13	{Dried sour bamboo shoots} => Dried mustard green (swun tan)	21%	80%
14	{Dried mustard green (swun tan)} => Dried sour bamboo shoots	21%	72%
15	{Dried soya bean and leek roots pancake} =>Dried soya bean pancake	21%	85%
16	{Mustard oil} => Green tea	20%	89%

REFERENCES

[1] S. Panjaitan, "Implementation of Apriori Algorithm for Analysis of Consumer Purchase Patterns," *The International Conference on Computer Science and Applied Mathematic*, 2019, doi: 10.1088/1742-6596/1255/1/012057.

[2] M. TIAN and A. Xue, "Data Dependence Analysis for Defects Data of Relay Protection Devices Based on Apriori Algorithm," *Special Section on Artificial Intelligence Technologies for Electric Power Systems*, vol. 8, Jul. 2020.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Third Edition.

[4] H. Singh, N. Shelke, and A. Bavaskar, "Study on Market Basket Analysis with Apriori Algorithm Approach," *International Research Journal of Engineering and Technology (IRJET)*, vol. 08, no. 05, May 2021.

[5] M. Kaur and S. Kang, "Market Basket Analysis: Identify the changing trends of market data using association rule mining," *International Conference on Computational Modeling and Security* (CMS 2016).

[6] "What is Support and Confidence in Data Mining?" https://www.geeksforgeeks.org/what-is-support-and-confidence-in-data-mining/?ref=gcse

Authors

Biography

Moe Phyu Phyu Khaing received the B.C.Sc. in Computer Science from the University of Computer Studies (Monywa) in 2018 and the M.C.Sc. from the University of Computer Studies, Mandalay (UCSM) in 2022. She works as a tutor at the University of Computer Studies (Taunggyi). Her research interests include Data Mining and Machine Learning.

Myint Myint Zaw is a teacher. She is an assistance lecturer at University of Computer Studies (Taunggyi). She achieved D.C.Sc in 2002 from University of Computer Studies (Kyaing Tong), B.A English in 2000 and M.A English in 2020 from Taunggyi University.

Thin Zar Aung received the B.C.Sc (Hons) degree from University of Computer Studies, Taunggyi in 2003, and with a master's degree in computer science from University of Computer Studies Mandalay (UCSM) in 2008. She currently works as a lecturer in the University of Computer Studies (Taunggyi), Myanmar's Faculty of Information Science. Machine Learning, Computer vision, and operation research are her main areas of interest.

IMPLEMENTATION OF NEWS CLASSIFICATION BASED ON THEIR HEADLINES FOR ONLINE NEWS ANALYSIS

Khant Phyo Nyein[1], Bawin Aye[2], and Htin Kyaw[3]

[1,2,3] Department of Computer Science, Higher Education Centre (Pyin Oo Lwin)
[1] kpnyein49@gmail.com, [2] bawinaye@ieee.org, [3] mgnyeinaye@gmail.com

ABSTRACT

There are a lot of sites that provide a ton of news every day since the internet is so widely accessible. We can observe that the news media are attempting to sway people's attitudes as there are more online news outlets. Categorizing the numerous news stories and articles is one of the most basic issues for data analysis with internet news sets. The machine learning model and Natural Language Processing (NLP) are frequently used for automatic news categorization to classify themes of untracked news and individual opinion based on the user's prior interests. This paper aims to categorise news based on their headlines for an investigation of online news trends. The data was collected from international media news related to Myanmar from January 1 to May 31, 2023. The dataset consists of 4143 news headlines that are manually categorized into eight categories: politics, crime and war, economy, weather, sport and health, religion and refugee, education and technology, and other for the train dataset. 75 percent of this dataset was used for the training dataset and 25 percent was used for the test dataset. In this study, we use Continuous Bag of Word(CBOW) for word embedding and Cosine Similarity for similarity measures. These research experiments were conducted using Excel VBA. Approximately 85.74 % accuracy has been attained in the accurate identification of the test news dataset for news classification.

Keywords

Excel VBA, Continuous Bag-of-Words, Cosine Similarity, News Classification, Similarity Measures

.

1. INTRODUCTION

Currently, a great volume of information is generated and accessible every day because of inexpensive handheld multimedia devices and the growth of the internet. News significantly connects people with daily environmental events. In today's information age, navigating the vast amount of online news presents challenges for readers and data handlers to find relevant content. That is why the researchers investigated a number of news article classification techniques.

Manual classification is insufficient when dealing with big data sets because of its difficulty and time-consuming nature for news categorization. Consequently, an increasing demand for automated systems that can correctly classify news pieces and foretell their categories. NLP is mostly used to classify documents and speech automatically based on word count or frequency without taking the meaning of the words into account.

When classifying papers or large amounts of text, this strategy is helpful. However, as news stories typically include brief titles and descriptions, these techniques might not effectively classify the articles in their respective categories. As a result, many academics began examining probabilistic, semantic, and similarity measurement categorizations rather than depending just on the meanings of individual words to provide more accurate findings.

In this paper, an automatic news classification system for English headline

news is proposed. The proposed system performs a novel methodology that leverages word embedding and text similarity techniques. Utilizing these methods to automatically assign appropriate news categories improves efficiency and accuracy in news classification systems. For the training dataset, we have collected news items related to Myanmar from English-language news outlets and built our own dataset that consists of eight categories: politics, crime and war, economy, weather, sport and health, religion and refugee, education and technology, and others. For feature extraction and text classification in this study, CBOW and Cosine Similarity are both used.

The remaining parts of this paper are divided into five sections: Section 2 discusses the literature review. The proposed system is presented in Section 3. Section 4 evaluates the performance of our proposed system and presents the results. The conclusion and further study are discussed in Section 5.

2. LITERATURE REVIEWS

In the research conducted by N. Disayiram and R. A. H. M. Rupasingha (2022), focused on multiple base classifiers, such as Decision Tree, Random Forest, Naïve Bayes [3,8], Multilayer Perceptron (MLP) and Support Vector Machine (SVM) [2,3,8], to create a powerful ensemble model. The study demonstrated that ensemble learning can effectively improve the accuracy and robustness of news category prediction models [1].

The study [4] introduces an innovative approach to front-page news classification. The study proposes the StackText model, which stands on the foundation of combining textual context and attribute data to enhance classification accuracy. The study's conclusions are important for the development of effective news classification systems, helping to facilitate access to and understanding of crucial national news in the constantly changing media environment.

Mona Nasr [5] emphasizes the value of NLP in extracting insights from unstructured text data as well as information retrieval. The study explores text classification methodologies, analyzing rule-based and machine learning approaches [7,9,10], providing a comprehensive overview of tools and techniques. The paper underscores the significance of NLP not only in information retrieval but also in generating insights from unstructured text data. In the backdrop of information explosion, this research holds valuable implications for enhancing information management, content recommendation, and automated decision-making processes across diverse domains.

The study [6] uses a probabilistic framework to dive into the area of categorizing news headlines. As the information landscape continues to expand, effective categorization of news headlines becomes vital for efficient information retrieval. This study contributes to the discourse by presenting a probabilistic approach for news headline classification. By harnessing probabilistic techniques, the paper aims to optimize the classification process, thereby enhancing the usability and relevance of news content. This study's implications extend to information science, data analysis, and media studies, offering a fresh perspective on news headline classification that is particularly relevant in the digital age.

3. PROPOSED SYSTEM

The objective of our research is to classify the news based on their headlines into specific categories and analyse the performance of the category prediction.

The proposed model's design is given in Figure 1. The approach involves News Collecting, text preprocessing, feature extraction, and Classification Model, with each step described in the following section.

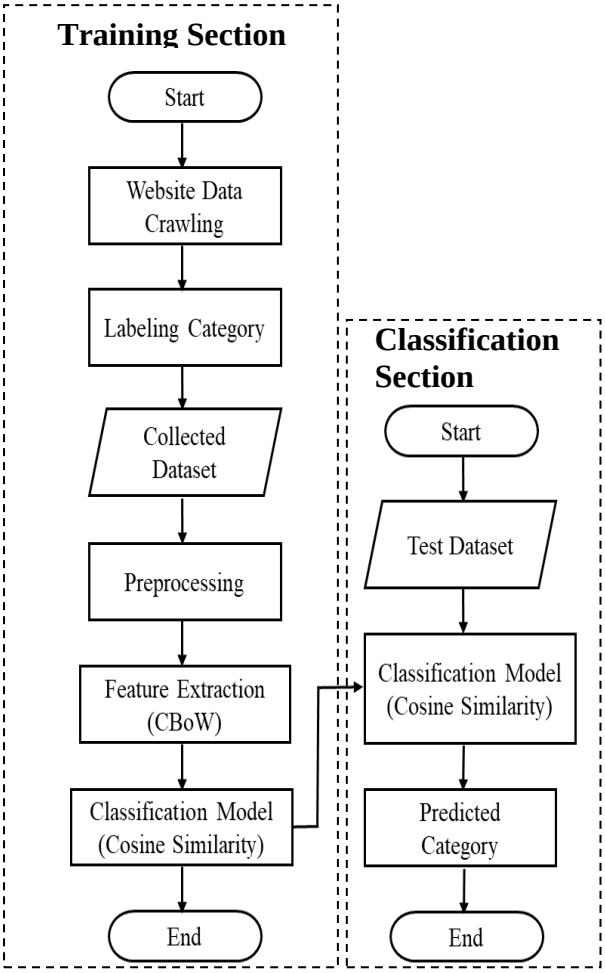


Figure 1. Proposed System Design of News Headlines Category Prediction.

3.1. Data Collection

For the dataset, the raw text datasets, which are written in English, are collected from the website (newsnow.co.uk), which brings together news articles related to Myanmar from more than 1,000 news media outlets around the world.

We collected data from newsnow.co.uk where public news international news outlets wrote about Myanmar from January 1 to May 31, 2023. The dataset consists of 4143 news headlines that are manually categorized into eight categories: politics, crime and war, economy, weather, sport and health, religion and refugee, education and technology, and other for the training dataset. In this study, only the data collected from the actual ground is used. The reason is to use the collected data to further analyse media influence situations. Therefore, the data cannot be in equilibrium. The present study, we used 75% (3107 items) for the training dataset and 25% (1031 items) for the testing dataset.

Table 1. Dataset for Category Prediction

No	Categories	Data	Percent
1	Politics	1530	36.93%
2	Crime and War	948	22.88%
3	Economy	393	9.49%
4	Education and Technology	33	0.80%
5	Religion and Refugee	356	8.59%
6	Sport and Health	148	3.57%
7	Weather	573	13.83%
8	Other	162	3.91%
	Total	4143	100.00%

3.2. Data Preprocessing

The essential stage in getting text data ready for text categorization is data preparation. Text data includes noise in a variety of ways, including emotions, punctuation, and case-sensitive text. Data pre-processing ensures the quality of the data. Therefore, a cleaning process should be carried out before extracting any features from the raw dataset to reduce the distortions made to the model.

To do the text data preprocessing, we followed a number of procedures, removing date, time and media name, removing stop words, removing special characters, converting lowercase, and tokenizing. Firstly, we removed the date, time and media name that were included in news headlines from all input datasets. Then the steps of stop word removal, special characters, lowercase, and tokenization are done in order. After data preprocessing step, we create the corpus dictionary for feature extraction step.

3.3. Feature Extraction

We can express the words of a textual text as numerical vectors so that the computer can comprehend it. Word embeddings, a method of encoding words as numerical vectors, are one approach to achieve this.

The meaning of the words and their connections to other words in the language are captured by these vectors. The present study used continuous bag-of-words (CBOW) for feature extraction. Continuous Bag of Words (CBOW) is a popular natural language processing technique used to generate word embeddings. The fact that word embeddings record the semantic and syntactic links between words in a language makes them crucial for many NLP tasks. CBOW uses surrounding context words to predict a target word.

It is a sort of "unsupervised" learning, which allows it to learn from unlabeled data, and it is frequently used to pre-train word embeddings that may be utilized for a variety of NLP tasks, including sentiment analysis, text categorization, and machine translation. The example of CBOW model is shown in Figure 2 below.

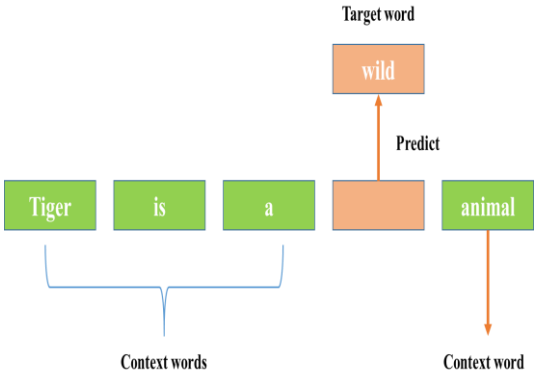


Figure 2. Example of a CBOW Model

In order to anticipate the target word, the CBOW model makes an effort to understand the context of the words around it. Take into account the previous sentence, "Tiger is a wild animal." This sentence is broken up into word pairs by the model (context word and target word).

The window size must be set by the user. If the context word's window were 2, the word pairs would look like this: ([tiger, a], is), ([is, wild], a), and ([a, animal], wild). Using these word combinations and taking into account the context words, the model tries to predict the target phrase.

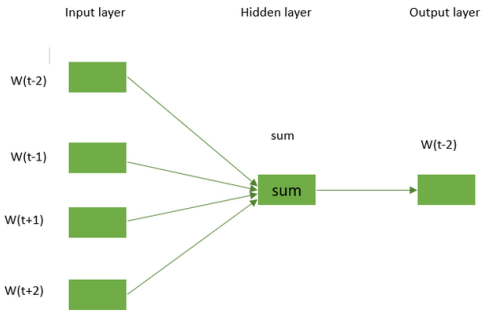


Figure 3. Architecture of CBOW Model

In Figure 3, if four context words are used to forecast a single target word, then the input layer will be made up of four input vectors. After being given these input vectors, the hidden layer will multiply them by a matrix. The output of the hidden layer eventually reaches the sum layer, where the vectors are element-wise summed before a final activation is carried out and the output is obtained.

3.4. Classification Model

After data collection, data preprocessing, and feature extraction, the text classification stage can be continued. In the present study, text classification will be done using cosine similarity to measure the similarity between vectors. Cosine similarity is a commonly used metric to measure the degree of similarity between two textual texts or vectors. It computes the cosine of the angle between the two vectors in order to indicate the direction and similarity of the vectors in a high-dimensional space. In the context of the text, cosine similarity is typically used to evaluate the degree of semantic similarity between texts or to pinpoint the documents that most closely match a specific query. The cosine similarity is calculated using the dot product of the vectors and their magnitudes as follows:

$$\text{Similarity}(A,B)=\cos(\theta)=\frac{A \,.\, B}{||A||\,||B||} \quad (1)$$

Where θ is the angle between the vectors, A.B is the dot product of the two vectors A and B, and ||A|| and ||B|| denote the

Euclidean magnitudes (lengths) of vectors A and B, respectively. The cosine similarity value that results falls between -1 and 1. High similarity (documents pointing in the same direction) is indicated by a number near to 1, whereas dissimilarity (documents pointing in the opposite direction) is indicated by a value close to -1. A rating close to zero denotes a lack of apparent similarity. The result of calculating the cosine similarity between two documents can be used to rate the similarity between various documents or to find the papers that most closely match a specific query. In terms of their semantic content, documents with greater cosine similarity scores are thought to be more comparable.

4. MODEL EVALUATION

The evaluation is based on collected 4143 headlines news from international online news outlets and applying the proposed algorithms. The experiment is designed to classify news articles according to the different domains such as politics, crime and war, economy, weather, sport and health, religion and refugee, education and technology, and others using the Cosine Similarity Method. The training dataset made up 75% of the total dataset, whereas the testing dataset included 25% as shown in Table 2.

Table 2. Dataset Preparation for Category Prediction

Categories	Training Data	Testing Data	Total
Politics	1148	381	**1529**
Crime and War	711	237	**948**
Economy	295	97	**392**
Education and Technology	25	8	**33**
Religion and Refugee	267	89	**356**
Sport and Health	111	37	**148**
Weather	430	142	**572**
Other	122	40	**162**
Total Percent	**75%**	**25%**	**100%**

After dataset preparation, data preprocessing stages such as data cleaning, stop word removal, and tokenization are performed. Then the corpus was built, and the features were extracted.

We discovered throughout the feature extraction procedure that there are a few key features in each of the news categories. Table 3 lists the top characteristics for each category.

Table 3. Top Features on Each Category

Categories	Top Features
Politics	election, vote, junta, protest, ministry, policy , myanmar, sanction, deplomatic
Crime and War	kill, scam, opium, attack, war, civil, airstrike, army, illegal
Economy	economic, port, export, chevron, h&m, oil, trade
Education and Technology	Google, facebook, internet, school, tech, education
Religion and Refugee	monk, monastery, church, catholic, buddhism, rohingya,
Sport and Health	aff, football, team, game, coronavirus, covid-19, malaria, medicine, hospital
Weather	rakhine, mocha, magnitude, flood, cyclone, coast, forecast
Other	photograp, superstar, park, universal, queen, beauty

And then, the extracted features from the headline text of the news are implemented in the proposed classification model that employs the cosine similarity method. Implementation measures document similarity and evaluates categorization effectiveness.

The similarity value indicates how similar two documents are, and the index value indicates the index of the similarity document in the training dataset. The implementation and experimentation of the news categorization system are shown in Figure 4.

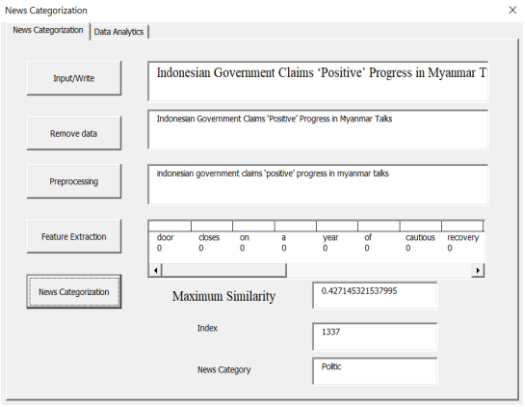


Figure 4. The Implementation of the News Categorization System

In performance analysis, we have been carried out only for tested datasets. The performance measures of the research are based on the Accuracy. In measuring performance, Accuracy for each category is calculated as the Confusion Matrix.

		Predicted value	
		TRUE	FALSE
Actual Value	TRUE	True Positive (TP)	False Positive (FP)
	FALSE	False Negative (FN)	True Negative (TN)

Figure 5. Confusion Matrix

The confusion matrix helps to identify the correct predictions of a model for different individual classes as well as the errors. It consists of columns that contain the values predicted by the classifier and rows which contain the actual values.

The accuracy is used to find the portion of correctly classified values. It tells how often the system is classified right. It is the sum of all true values divided by total values. It is calculated by the following equation.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

We employed the cosine similarity method to determine the similarity between the vector representation of each news article and the pre-calculated vector representations of each category. The category with the highest cosine similarity score was considered the predicted category for the article. The evaluation was done by observing each category's prediction results by analyzing the Confusion Matrix and quantifying Accuracy. A test dataset with 1031 news samples served as the basis for this investigation. The Confusion Matrices of the category's prediction results are shown in Table 4.

Table 4. Confusion Matrix for News Category

Actual \ Predicted	Politic	Crime and War	Economy	Education and Technology	Religion and Refugee	Sport and Health	Weather	Other
Politic	339	20	7	1	5	3	2	3
Crime and War	18	203	7	3	5	1	0	0
Economy	14	4	72	0	3	1	3	0
Education and Technology	1	0	1	5	0	1	0	0
Religion and Refugee	5	13	2	0	68	0	0	1
Sport and Health	0	0	1	0	3	33	0	0
Weather	0	2	0	0	0	0	140	0
Other	5	3	4	0	2	2	1	24

According to the results, we achieved an accuracy score of 85.74% and the error rate is 14.26%. By looking at the accuracy of each category, we obtained the best accuracy is the weather category with 98.59 percent and the lowest accuracy is the other category with 58.54 percent. The accuracy of each category is shown in Table 5.

Table 5. Accuracy of Category Prediction

Categories	Accuracy
Politics	89.21 %
Crime and War	85.65 %
Economy	74.23 %
Education and Technology	62.50 %
Religion and Refugee	76.40 %
Sport and Health	89.19 %
Weather	98.59 %
Other	58.54 %
Total	85.74 %

According to the experimental results, it is found that there is no big gap in the accuracy of each category. In this research, the data is not balanced, but since the cosine similarity method is based on similarity, accuracy is not affected.

3. CONCLUSIONS

In this research, we have investigated and implemented a news category prediction system using continuous bag of word and text similarity methods. News category prediction was developed by training the well-known NLP method (cosine similarity) on the ground truth collected news dataset that is categorized into eight categories (Politics, Crime and War, Economy, Education and Technology, Religion and Refugee, Sport and Health, Weather, and Other). The performance of our method is further confirmed by the confusion matrix, which offers a thorough summary of the prediction findings. A large number of true positives were produced as a result of the majority of the articles being accurately categorized into their respective categories.

As a further extension, the categorization data obtained from this research can continue to be used as structural data in news analysis and media analysis. In addition, this researcher will be able to provide the fields of data mining, text classification, text mining, and other text data handling research.

REFERENCES

[1] N. Disayiram, and R. A. H. M. Rupasinghayril, “A Comparative Study of Classifying English News Articles Using Machine Learning Algorithms,” 2022 Trends in Electrical, Electronics, Computer Engineering Conference (TEECCON), 2022.

[2] R. R. Deshmukh, D. K. Kirange, “Classifying News Headlines for Providing User Centered E-Newspaper Using SVM,” International Journal of Emerging Trend and Technology in Computer Science (IJETTCS), 2013, vol. 2, no. 3.

[3] Arif Hussain, G. S. Baloch, Faheem Akhtar, and Zahid Khand, ‘‘Design and Analysis of News Category Predictor,’’ Natural Engineering, Technology and Applied Science Research, 2020, Vol. 10, No. 5, 2020, 6380-6385.

[4] K. Wang, M. chen, Z. yan, “Front-Page News Classification Model Based on the Stacking of Textual Context and Attribute Information,” Scientific Programming, vol.2022, no. 3031195, p. 9.

[5] M. Nasr, A. Karam, M. Atef, " Natural Language Processing: Text Categorization and Classifications," Int. J. Advanced Networking and Applications, 2020, vol. 12, no. 02, pp. 4542-4548.

[6] M. I. Rana, S. Khalid, F. Abid, "News Headlines Classification Using Probabilistic Approach," VFAST Transactions on Computer Sciences, 2015, vol. 3, no. 1, pp. 2411-6335.

[7] J. Ahmed, M. Ahmed, “Online News Classification Using Machine Learning Techniques,” IIUM Engineering Journal, 2021, vol. 22, no. 2.

[8] R. Bogery, N. A. Babtain, N. Aslam, “Automatic Semantic Categorization of News Headlines using Ensemble Machine Learning: A Comparative Study”, International Journal of Advanced Computer Science and Applications (IJACSA), 2019, vol. 10, no. 11.

[9] M. S. Chhajerh, A. KVS, M. Meleet & R. Murthy, “Real Time News Headlines Classification Using Machine Learning,” International Research Journal of Engineering and Technology (IRJET), 2021, vol. 08, no. 06.

[10] M. Ikonomakis, S. Kotsiantis, V. Tampakas, “Text Classification Using Machine Learning Techniques,” WSEAS Transactions on Computers, 2005, vol. 4, no. 8, pp. 966-974.

Authors Biography

Khant Phyo Nyein earned his master degree from National Research University of Electronic Technology (MIET), (Russian Federation) in 2011. Now he is a Ph.D. candidate of Computer Science Department, Higher Education Center (HEC).

Dr. Bawin Aye earned his master degree from National Research University of Electronic Technology (MIET), (Russian Federation)

in 2006. He also obtained his Ph.D. degree from National Research University of Electronic Technology (MIET), (Russian Federation) in 2010. Now he is a lecturer of Department of Computer Science in HEC. He is also a member of the IEEE Computer Society.

Dr. Htin Kyaw earned his master degree from National Research University of Electronic Technology (MIET), (Russian Federation) in 2010. He also obtained his Ph.D. degree from National Research University of Electronic

REAL TIME IDENTIFICATION AND RECOGNITION OF MYANMAR VEHICLE TYPE AND LICENSE PLATES SYSTEM IN DAY AND NIGHT CONDITIONS USING YOLOV5s MODEL

Min Min Oo[1], Zaw Ye Aung[2], Myo Min Hein[3], and Nyi Nyi Shein[4]

[1,4] Department of Computer Technology, Higher Education Centre, Pyin Oo Lwin
[2,3] Department of Computer Science, Higher Education Centre, Pyin Oo Lwin
[1]minnnminnnoo@gmail.com,[2]alexanderzaw.bmstu@gmail.com,
[3]myominhein48@gmail.com, [4]yintwinmay@gmail.com

ABSTRACT

In today's technological landscape, the assurance of citizens' safety in their daily lives and transportation modes has become increasingly imperative, particularly considering the substantial growth in the number of vehicles on the road over the past two decades. This surge in the transportation sector has posed challenges to the identification of vehicles. To address this issue, the study utilizes the You Only Look Once (YOLO) deep learning model, renowned for its effectiveness in object detection. The research focuses on deploying this efficient deep learning method for Automatic License Plate Recognition (ALPR), specifically aiming to reliably detect Myanmar Vehicles License Plates (MVLPs) under diverse and challenging conditions, including varying light and weather, unavoidable data noise, and the real-time performance requirements of Intelligent Transportation Systems (ITS). Custom datasets collected from the Yangon-Mandalay Highway are employed, utilizing the recently released YOLOv5 object-detection framework for License Plate detection (LPD)and vehicle type classification in real-time conditions.

Keywords

License Plate Detection (LPD), Automatic License Plate Recognition (ALPR), Myanmar Vehicles License Plate (MVLPs), You Only Look Once Version 5 (YOLOv5), Label Image Annotation, Intelligent Transportation Systems (ITS)

1. INTRODUCTION

In recent years, the extensive adoption of intelligent transportation systems has highlighted the essential role of license plate identification in license plate recognition research. The precision of the final recognition outcome depends significantly on the efficient identification of license plates. Advances in license plate recognition technology have substantially enhanced accuracy, especially in challenging environmental conditions. This progress marks a notable improvement over conventional methods, which faced challenges in scenarios such as low light, varying camera distances, and nighttime conditions illuminated by headlights.

This research delves into the analysis of videos from highway security cameras using the Automatic License Plate Recognition (ALPR) process, with a focus on refining the identification vehicle type and real-time vehicle detection process through the YOLO detection method. Known for its effectiveness and real-time object detection features, YOLO is employed to enhance accuracy and efficiency. Through a thorough examination of various inputs, the research demonstrates the superior performance of the proposed method compared to existing ones in the field.

2. RELATED WORK

The findings from this research can be applied to manually input data at parking areas and toll gates where there is a high volume of vehicles. Previous research on vehicle number plate detection employed various techniques. Combining the YOLOv3 object detector with the Deep Neural Network (DNN), Tourani et al. [3] introduced a real-time License Plate Detection (LPD) and Character Recognition (CR) system for Iranian vehicle license plates in 2020. The networks were trained using realistic condition data from actual monitoring systems, incorporating diverse data augmentation techniques to ensure robustness in different conditions. The suggested method achieved an impressive 95.05% average end-to-end recognition rate on 5719 photos.

Shohei et al. [4] addressed YOLOv2's limitation in detecting smaller objects, like license plates (LPs), by employing a two-stage YOLOv2 model in 2019. The first stage identifies the vehicle, while the second stage focuses on detecting the license plate. Although it demonstrated excellent nighttime accuracy, it struggled with distant LPs and occasionally misidentified labels as LPs. To adapt the network to current situations, adjustments were made.

Rayson [1] proposed a system with a 96.8% recognition rate for automatic license plate recognition, utilizing the YOLO algorithm. This research capitalized on YOLO's highly accurate Frames per Second (FPS) rates, enabling real-time identification of multiple vehicle types, such as four vehicles simultaneously. Certain number plate images greatly benefited from environmental factors, including lighting, weather conditions, traffic, poor conditions, and others.

3. PROPOSED SYSTEM

The proposed method has two main steps: first, YOLOv5s detects license plates, and recognize number plate in real-time. Developing deep learning involves three main parts: collecting training data, training the model, and predicting objects using the trained model. When the training data is large, it's uploaded into YOLO to train the model. Users can see results on the interface. After identifying and classifying the license plate and vehicle type, the system moves to the next step for detection. The method has two phases: testing and training. During training, a camera records the highway video. The system's flowchart is shown in figure 1.

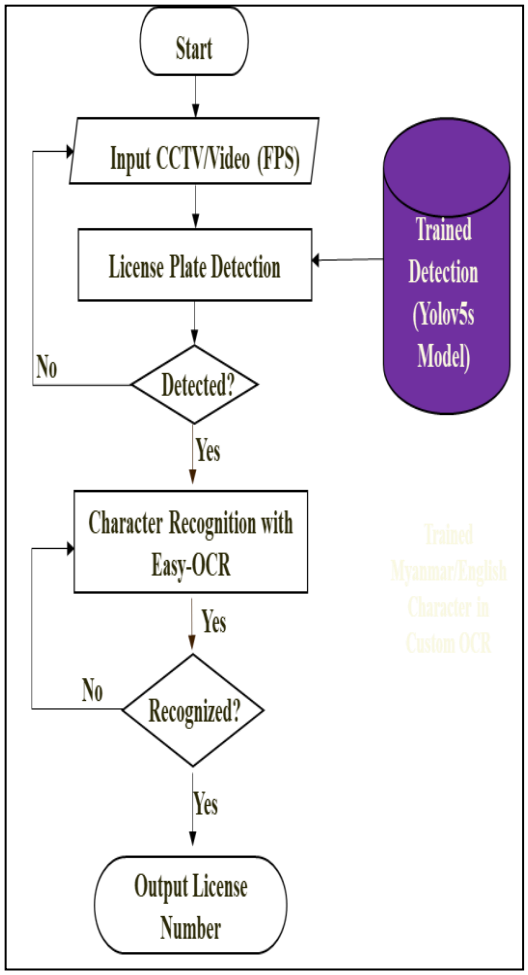


Figure 1. Proposed System Design

Next, the streaming video is divided into different images, and each image gets a label for object classification. Labeling is a key step in this process. The labeled images are saved and included in the dataset. The dataset goes through three stages: testing, validation, and training. Figure 2 describes the training images in Roboflow.

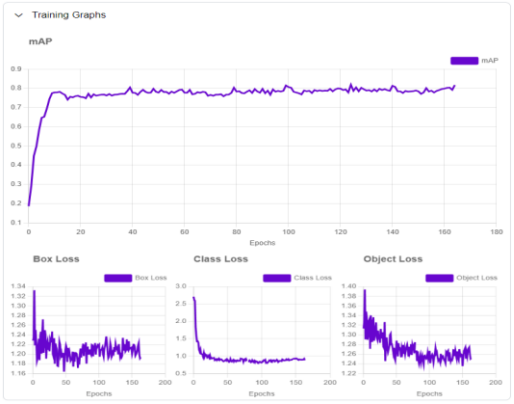


Figure 2. Training, Validation and Testing Images in Roboflow

3.1. Data Collection

The research emphasizes the significance of collecting data for model training, particularly focusing on obtaining license plate data. The primary challenge lies in determining the most efficient method of collection, with approximately 2200 images earmarked for training on Roboflow.com. To capture video and customized data, CCTV cameras along the Yangon-Mandalay Expressway have been explored as a potential solution. The research employs the PYTHON programming language to extract images from videos recorded at either 60 or 30 frames per second. Subsequent to image extraction, the 'LabelImg' tool integrated into the PYTHON environment is utilized for labeling. This process precisely identifies the region containing the license plate, as depicted in figure 3, and stores the information in a TxT file.

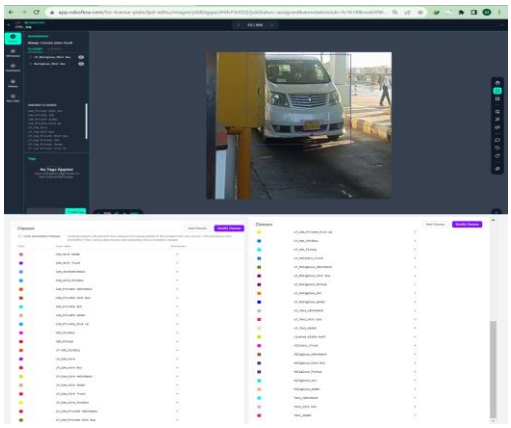


Figure 3. Labeling images using the 'LabelImg' tool

A total of 2200 images, comprising 70% of the overall dataset, were meticulously selected from self-constructed custom datasets to serve as the primary training datasets. In parallel, 181 images, representing a 20% proportion, were allocated to the validation datasets, ensuring a robust evaluation process. For the crucial testing phase, 95 images, constituting 10% of the entire dataset, were thoughtfully chosen to assess the model's performance under real-world conditions. To offer a comprehensive visual overview of the distribution, Figure 4 vividly displays the arrangement of images designated for training, validation, and testing purposes.

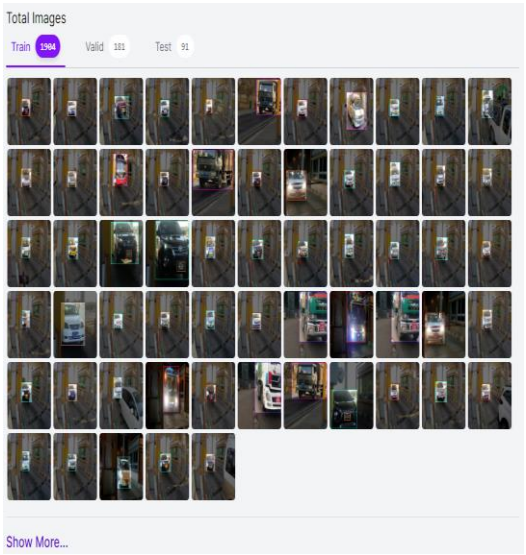


Figure 4. Training, Validation and Testing Images

3.2. License Plate Detection

The initial challenge involves removing the license plate from a vehicle in an image or video. Object detection typically follows three steps: 1) Identifying the object, 2) Classifying the object and 3) detecting the license plate. Region detection focuses on finding an object in an image or video, while classification tasks involve recognizing the type of vehicle, such as a sedan, bus, or truck, and for plate such as a Commerical Private, Commerical Hire, Taxi, and Religious. The YOLO model has been proposed as the fastest among current approaches, well-suited for real-time object detection. This system operates efficiently

with a single GPU due to its GPU-centric design. YOLOv5 [11] comprises three main parts:

(a) Model Backbone: The Model Backbone plays a crucial role in extracting features from images, such as edges, shapes, and lines. To achieve this, it employs a convolutional neural network (CNN) with layers for batch normalization, kernel, and stride. The primary support system for feature extraction is the Cross Stage Partial Networks (CSP), built on Densenet, connecting CNN layers primarily.

(b) Model Neck: Feature pyramids are utilized to identify components of an image of various sizes and provide multi-scale predictions. Feature pyramid networks are employed for this purpose, and YOLOv5 uses PaNets for Feature Pyramid Networks (FPNs).

(c) Model Head: The final detection stage, Model Head, utilizes an anchor to identify arrays with class probabilities, bounding box scores, and bounding boxes. Activation functions of a deep neural network are essential. Leaky Relu is used in the middle and deep layers of YOLOv5, and a nonlinear activation function is applied in the top layers. The YOLOv5 model operates on the PyTorch framework and encompasses 191 layers.

It uses K-Means with neural networks to create K-means-developed anchor boxes. These boxes include the coordinates and projections of the bounding boxes. The network structure of YOLOv5s is described in figure 5. The YOLOv5s network adjusts the border box width and height to match the dimensions of the image from 0 to 1. This is consistent with how other networks handle it. The process ensures that the data is limited to the range of 0 and 1 when constructing the x and y coordinates of the bounding box from the selected point. All previous layers except the last layer use a linear activation function.

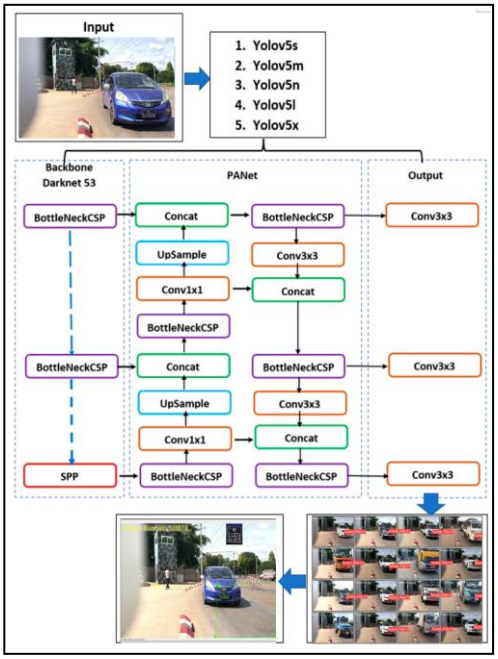


Figure 5. Network Architecture of YOLOv5

3.3 Using Easy-OCR Engine

Easy-OCR [12], made with PyTorch, is excellent at recognizing characters from colored lines. We look into its process, explore its language handling capabilities, and see how deep learning models are used for character and word analysis. The use of GPU enhances OCR performance. Easy-OCR, a Python program leveraging PyTorch, recognizes text from images in real-time using a robust deep reading resource. It supports text detection in 42 languages with color fields. A visual representation of how Easy-OCR works is shown in figure 6.

Figure 6. Character Recognition using Easy-OCR

3.4 Determining Region of Interest (ROI)

To pinpoint the Region of Interest (ROI) [10], we utilize it in the detection of number plates to filter out background noise and prevent false positive detection. The operator employs four coordinates to set up the ROI, drawing a bounding box to isolate the number plate region. Recognition occurs if the number plate bounding box falls within the defined ROI. This use of ROI successfully enhances OCR recognition accuracy and minimizes false positive number plate detection. As illustrated in figure 7, defining the ROI proves beneficial by reducing computational time, avoiding the detection of unwanted areas, and facilitating faster and more accurate real-time operations.

Figure 7. Determining ROI in Image

4. EXPERIMENTAL RESULTS

Utilizing the PyTorch deep learning framework, this research involves the training and evaluation of a vehicle license plate dataset. The objective is to develop systems for real-time identification and recognition of Myanmar vehicle types and license plates, employing the YOLOv5s model. The customized dataset is integral to the training of the YOLOv5s model, enabling it to detect license plates effectively through the use of 640-pixel images. Classification involves recognizing vehicle types like sedans, buses, or trucks, and for license plates, the goal is to categorize them into types like Commercial Private, Commercial Hire, Taxi, and Religious.

Figure 8. Classification in Day Conditions

In figure 8, the daytime classification is illustrated, while figure 9 depicts the nighttime classification. The proposed system utilizes image or CCTV video files as input data, specifically designed for real-time system applications.

Figure 9. Classification in Night Conditions

The application is configured to run on hardware equipped with an NVIDIA GeForce GTX 1650 graphics card, operating on Windows 11 and utilizing the Python 3.9 programming language. The experimental results of license plate

recognition in day and night conditions in real-time shown in figure 10.

Figure 10. Experimental Results of LPR in Day and Night Time

The accuracy in recognizing license plates is 89%, with a 7% false recognition rate, particularly at night when license plates aren't identified, as displayed in Table 1 for Myanmar vehicle number plates. Additionally, based on the experiment's findings, the accuracy of missed recognition is 4%, especially at night when sharp headlights limit visibility.

Table 1. Percentage Accuracy of Detection and Recognition

	True Detection NP	False Detection Np	Miss Detection NP
Total Percent of NP	89%	7%	4%

The proposed system is superior in the classification and detection part, but in the recognition part, some characters and numbers cannot be accurately recognized according to the experimental results, and they are presented in the figure 11. Therefore, if the character recognition part can be improved in future work, it will become a complete license plate recognition system and can be tested in the practical field.

Figure 11. Incomplete License Plate Images

5. CONCLUSIONS

In conclusion, utilizing the PyTorch deep learning framework, this research trains and evaluates a tailored vehicle license plate dataset. Centered on real-time identification and recognition of Myanmar vehicle types and license plates, the research employs the YOLOv5s model for efficient detection. The results showcase an 89% accuracy in license plate recognition, successfully addressing challenges across various environmental conditions. While excelling in classification and detection, the system presents opportunities for further improvement in character recognition, paving the way for practical field testing.

ACKNOWLEDGEMENTS

I want to express my gratitude to my supervisor, Dr. Myo Min Hein, Assistant Lecturer at the Department of Computer Science, Higher Education Center, and Dr. Zaw Ye Aung, Lecturer at the Department of Computer Science, Higher Education Center, and Dr. Nyi Nyi Shein, Associated Professor at Computer Technology, Higher Education Center. Their valuable suggestions, attentive guidance, and knowledge-sharing have been greatly appreciated.

REFERENCES

[1] Rayson Laroca, “An Efficient and Layout-Independent Automatic License Plate Recognition System Based on the YOLO Detector”, 2019.

[2] Lele Xie, “A New CNN-Based Method for Multi-Directional Car License Plate Detection”, IEEE Transactions on Intelligent Transportation Systems, 2018.

[3] Tourani, A. Shahbahrami, S. Soroori, S. Khazaee, and C. Y. Suen, “A robust deep learning approach for automatic Iranian vehicle license plate detection and

recognition for surveillance systems," IEEE Access, vol. 8, pp. 201317–201330, 2020.

[4] S. Yonetsu, Y. Iwamoto, and Y. W. Chen, "Two-Stage YOLOv2 for Accurate License-Plate Detection in Complex Scenes," 2019 IEEE Int. Conf. Consum. Electron. ICCE 2019, no. 1, pp. 1–4, 2019, doi: 10.1109/ICCE.2019.8661944, 2019.

[5] Izidio, D.M.; Ferreira, A.; Medeiros, H.R.; Barros, E.N.D.S. An embedded automatic license plate recognition system using deep learning. Des. Autom. Embed. Syst. 2020, 24, 23 – 43. [CrossRef]," p. 2020, 2020.

[6] Rayudu Sushma,"Automatic License Plate Recognition with YOLOv5 and Easy-OCR method", IJIRT Volume 9 Issue 1, 2022.

[7] Yu-Liang Chiang, "Real Time License Plate Detection Based on Machine Learning", Master Theses and Dissertations. 2654. 2021.

[8] Jocher, G. YOLOV5, An Open-Source Detection Model Repository. Available online: https://github.com/ultralytics/yolov5.

[9] Do Thuan, "Evolution of YOLO Algorithm and YOLOv5: The State-of-the-art Object Detection Algorithm", 2021.

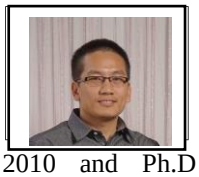
[10] Min Min Oo, "MYANMAR VEHICLES' LICENSE PLATE DETECTION AND RECOGNITION", JITRI (Mandalay) Vol-2, Issue-1, December 2022.

[11] Min Min Oo, "Development of Myanmar Vehicles' License Plate Detection System in real-time for Day and Night Conditions by using YOLOv5 Model", JITRI (Mandalay) Vol-3, Issue-1, June 2023.

[12] Min Min Oo, "Real-Time Myanmar Vehicles' License Plate Detection and Recognition Method Using Easy-OCR with YOLOv5 in Different Environmental Conditions (Day and Night)", JRI (Thanlyin) Vol-6, 2023.

Authors Biography

Min Min Oo received M.E (Computer) from Bauman Moscow State University of Technology (Kaluga) in 2011. He is currently attending Computer Technology Department, Higher Education Center (HEC) as a Ph.D. candidate. His current research includes data science; Includes computer vision and deep learning analysis.



Dr. Myo Min Hein received M.I.C.Sc Russian State Technological University (MEPHI), (Russian Federation) in 2010 and Ph.D from Defence Services Academy Computer Science Department, in 2018. He is currently working as Assistant Lecture with Computer Science Department.

Dr. Zaw Ye Aung graduated with a B.C.Sc. and was commissioned by the University of Computer Studies, Yangon (UCSY) in 2002. His Ph.D., D.Sc. (Hons), and M.E. (IT) degrees are all obtained from Bauman Moscow State Technical University. He is working as a lecturer for the Higher Education Center's Department of Computer Science.

Dr. Nyi Nyi Shein received a master's degree in automatic control systems from the Moscow Institute of Electronic Technology (MIET), Russia. In addition, in 2010, he graduated with a Ph.D. in automatic control systems from the same university. He is currently serving Associated Professor of Computer Technology at HEC.

.

SINGLE OBJECT TRACKING USING MORPHOLOGICAL-BASED VECTOR FEATURE (MVF)

Myo Min Paing[1], Dr. Myo Min Hein[2], and Dr. Bawin Aye[3]

[1,2,3]Department of Computer Science, Higher Education Centre (Pyin Oo Lwin)
[1]myominpaing351@gmail.com, [2]myominhein48@gmail.com, [3]bawinaye@ieee.org

ABSTRACT

Object tracking plays a fundamental role in various domains. It is finding the position of the interested object in every frame of the video. The existing methods have many drawbacks, such as the disappearance of trackable features in the condition of partial occlusion. In this paper, moving object tracking using morphological-based vector feature (MVF) is proposed. In the proposed system, corners are detected by applying morphological operations to Canny edges. Next, the vectors from each corner point to all other points are calculated and extracted as morphological-based vector feature. And then, the proposed feature weight calculation method is utilized for feature matching between the two image frames, and the proposed bounding box prediction method is applied for identifying the ROI in the next image frame. The proposed system can track the object with an accuracy of 80% and overcome the challenges of partial occlusion in object tracking.

KEYWORDS

Corners, Edges, MVF, Partial Occlusion, ROI

1. INTRODUCTION

Object tracking plays a fundamental role in various domains, such as video surveillance, autonomous driving, and object recognition. It is locating the target object in every frame of the video. The role of this tracking capability is pivotal, contributing to enhanced surveillance, navigation, and target identification across various fields, including both civilian and military sectors.

There are two levels of object tracking: single object tracking (SOT) and multiple object tracking (MOT). SOT aims to track an object instead of multiple objects. It is also sometimes referred to as visual object tracking. In SOT, the bounding box of the target object is defined in the first frame. The goal of the algorithm is then to locate the same object in the rest of the frames. SOT belongs to the category of detection-free tracking because one has to manually provide the first bounding box to the tracker.

In the proposed system, the original ROI is manually defined in the first frame and the possible bounding box is automatically defined in the next frame based on the original ROI. Next, the proposed morphological-based vector feature extraction algorithm is applied in both frames, and the proposed feature weight calculation method is used to get matched features between the two frames. Then, the outliers of the matched features are removed and the predicted bounding box is created in the next image frame using the inlier-matched features. Finally, the midpoint is set to the image frame as the tracking point. The effectiveness of the proposed system is illustrated in experimental results.

2. AIM

The aim of the research paper is to develop a single object tracking system that can overcome partial occlusion, using the proposed morphological-based vector feature.

3. LITERATURE REVIEW

In recent years, object tracking technologies have become crucial in many areas. For the purpose of enhancing object tracking algorithms, numerous researchers have conducted research, and many developers have developed object tracking systems to track single or multiple objects by using various methods.

Mr. Pramod R. Gunjal, Miss. Bhagyashri R. Gunjal, Mr. Haribhau A. Shinde, Mr. Swapnil M. Vanam, Mr. Sachin S. Aher [3] proposed a technique for employing the Kalman filter to detect and track moving objects. In this system, colour recognition is used to identify the target object in each frame. So, it has drawbacks in illumination changing condition. And then the target object is tracked by the help of Kalman Filter which can only be used in a linear or stable system with noise distributed normally [1].

Li Tan, Xu Dong, Chongchong and Yuxi Ma [2] proposed a multi-object tracking algorithm using YOLO. It uses YOLO algorithm to detect the object, Long Short-Term Memory (LSTM) to track the detected object and MOT-16 data sets and MSR data set to train and experiment the method. But it requires very large amount of data in order to perform object tracking tasks and is extremely expensive to train due to complex data models. Although accuracy is high, it can only detect and track the trained objects under its trained condition.

M. Sushma Sri [4] presented that the KLT algorithm is applied to track the detected object. It uses the Viola-Jones algorithm to detect the face and the KLT algorithm to extract features and track the face frame by frame using its features. In this paper, it is only attempted to track the human face, the static object.

4. METHODOLOGY

In this paper, single object tracking using morphological-based vector feature (MVF) is proposed.

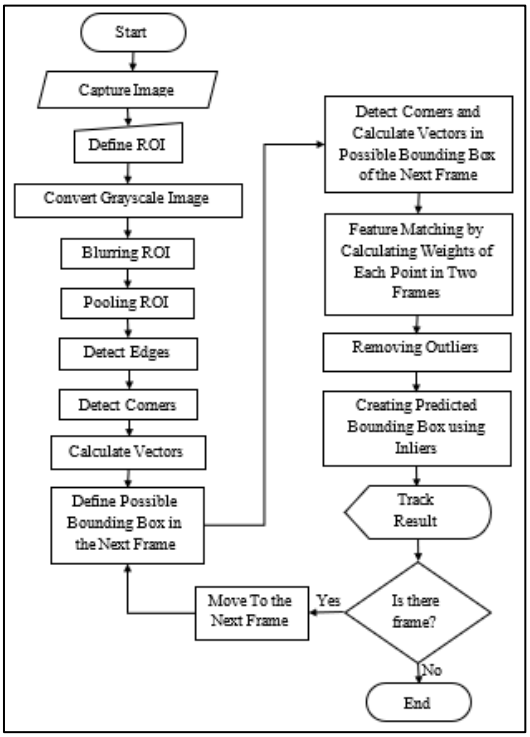


Figure 1. Flowchart of Proposed System

In the proposed system using MVF shown in Figure 1, it will start by capturing images from video. Then the region of interest (ROI) must be manually defined by the user. Next, grayscale conversion, pooling, and blurring techniques are applied to the ROI. Edges are detected in the blurred image using the Canny operator. Then, corner features are detected using eight structuring elements and morphological operation. After that, the vectors from each point to all other points in the ROI will be calculated, applying the Euclidean distance formula and the inverse tangent formula. Then, defining a possible bounding box in the next frame, detecting corner features, calculating vector features, matching features between the two frames by calculating weights, removing outliers among the matched features, and creating a predicted bounding box using the inlier features will be carried out.

4.1 Morphological-based Vector Feature Extraction

Morphology is a broad set of image processing operations that process images

based on shapes. The word morphology refers to form and structure. Morphological operations use a small template known as a structuring element (SE). The structuring element is placed in all pixel locations of the image and is compared to the corresponding neighboring pixels [5]. The structuring element applied to a binary image can be described as a small matrix with a value of 0 or 1.

In the proposed system, it starts by capturing the first image frame of a video, and the user defines the region of interest (ROI) as shown in equation (1).

$$ROI = I(R1{:}R2, C1{:}C2, :) \quad (1)$$

Where $R1$ and $R2$ are the initial row and the final row of the image, and $C1$ and $C2$ are the initial column and the final column of the image, which are manually defined in the image by clicking the upper left corner and the lower right corner of the interested object.

Next, the ROI is converted to the grayscale image, and average pooling technique is applied to the grayscale ROI, as shown in Figure 2. In the proposed system, the size of the filter is 2 × 2 and the stride of the filter is 2.

Then, the result pooled image is blurred with mean filter, as shown in equation (2).

$$I_{blur} = ROI_p * M \quad (2\ a)$$

$$M = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}_{m \times n} \times \frac{1}{m \times n} \quad (2\ b)$$

Where I_{blur} is the result blurred image, ROI_p is the pooled image and M is the mean filter with the size of $m \times n$.

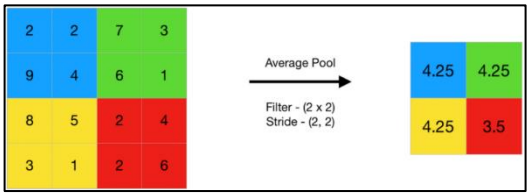


Figure 2. Average Pooling

Edges of the ROI_p, I_{Edge}, are detected using the Canny edge detector. After that, eight proposed structuring elements, which is illustrated in Figure 3, are applied to detect morphological corner features of the edges, as shown in equation (3).

$$C = \sum_{i=1}^{m} \sum_{j=1}^{n} [W(i,j) \times SE(i,j)] \quad (3\ a)$$

$$F(x,y) = \begin{cases} (j,i) & if\ C == T \\ None & else \end{cases} \quad (3\ b)$$

$$T = \sum_{i=1}^{m} \sum_{j=1}^{n} SE(i,j) \quad (3\ c)$$

Where C is defined as the corner response function, W is as a patch from edges, SE is as the structuring element, T is as the threshold value, and $F(x,y)$ is as the x and y coordinates of corner features. The value of the threshold, T, is the sum of all the values in SE, and if the corner response function, C, is equal to T; in other words, if SE fits to the patch, W, the system will conclude that the center point of the patch is a corner.

In Figure 3, the structuring elements used in the proposed system are depicted. In Figure 4, the process of extracting morphological corner features is illustrated.

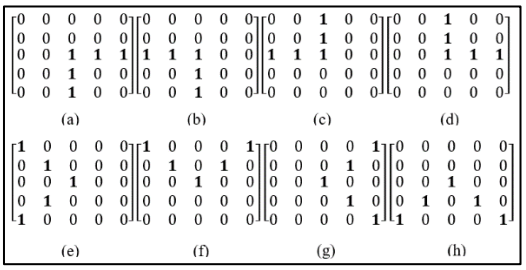


Figure 3. Eight Structuring Elements

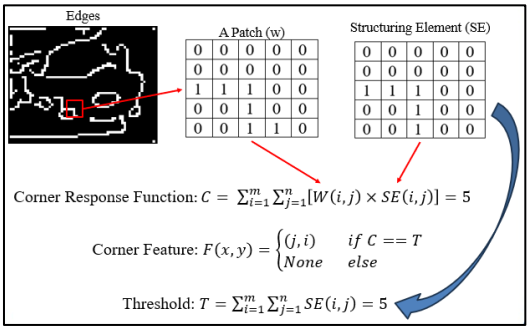


Figure 4. Morphological Corner Features Extraction

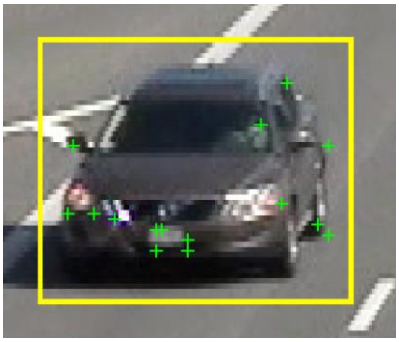


Figure 5. Morphological Corner Features

In Figure 5, the morphological corner features extracted by using morphological operation are illustrated. The yellow color rectangle is the region of interest defined by the user, and the green color '+' signs are the corner features. There are fifteen corner features in the first frame of the video.

Next, the vectors of features from each point to all other points are calculated using the Euclidean distance formula and the inverse tangent formula, as shown in equation (4).

$$M(i,j) = \sqrt{(F_j(x) - F_i(x))^2 + (F_j(y) - F_i(y))^2} \quad (4\text{ a})$$

$$A(i,j) = tan^{-1}\frac{(F_j(y) - F_i(y))}{(F_j(x) - F_i(x))} \quad (4\text{ b})$$

Where M is the magnitude matrix, A is the angle matrix of features to each other, and $M(i,j)$ and $A(i,j)$ are the magnitude and the angle of features i and j. Since there are fifteen morphological corner features in Figure 5, the morphological vector features will be a 15×15 magnitude matrix and a 15×15 angle matrix.

4.2 Calculating Weights of Vector Features

After extracting morphological-based vector features in two frames, the weights of points will be calculated by comparing the angle matrices and the magnitude matrices as follows.

Assuming two image frames, *f1* and *f2*, corner features are detected in both frames applying morphological operation. In the first frame, five features are identified, while the second frame contains six features. Subsequently, vectors from each point to all other points are calculated, resulting in magnitude and angle matrices for each frame. Specifically, for *f1*, a magnitude matrix (M^{f1}) and an angle matrix (A^{f1}) are generated, both with a size of 5×5. For *f2*, the corresponding matrices are denoted as M^{f2} and A^{f2}, each with a size of 6×6, as illustrated in Figure 6.

Subsequently, the match vector count matrix (MV) is calculated using equation (5), and the resulting MV is depicted in Figure 7.

$$MV_{(j,n)} = \begin{cases} MV_{(j,n)} + 1 & if\ M_{dif} \le MT\ \wedge\ A_{dif} \le AT \\ MV(j,n) & else \end{cases} \quad (5\text{ a})$$

$$M_{dif} = \min\left(\left|M^{f1}_{(a,j)} - M^{f2}_{(b,n)}\right|\right), 1 \le a \le i, 1 \le b \le m \quad (5\text{ b})$$

$$A_{dif} = \min\left(\left|A^{f1}_{(a,j)} - A^{f2}_{(b,n)}\right|\right), 1 \le a \le i, 1 \le b \le m \quad (5\text{ c})$$

Where MT is the threshold for magnitude and AT the threshold for angle, MV is match vector count matrix, M_{dif} and A_{dif} are the absolute difference of magnitude and angle. Since there are five features in the first frame and six features in the second frame, the resulting match vector count matrix (MV) has dimensions of 5×6, as depicted in Figure 7. $MV_{(j,n)}$ represents the count of matched vectors between the j^{th} feature in the first frame and the n^{th} feature in the second frame.

M^{F1}

0	42.521	84.309	88.023	44.407
42.521	0	50	54	4.4721
84.309	50	0	4	52.154
88.023	54	4	0	56.143
44.407	4.4721	52.154	56.143	0

M^{F2}

0	72	60.729	20.881	23.409	41.617
72	0	22.804	68.964	82.97	58.822
60.729	22.804	0	52.038	66.121	37.577
20.881	68.964	52.038	0	14.142	22.804
23.409	82.97	66.121	14.142	0	34.176
41.617	58.822	37.577	22.804	34.176	0

A^{F1}

NaN	311.19	337.69	338.68	305.84
131.19	NaN	360	360	243.43
157.69	180	NaN	360	184.4
158.68	180	180	NaN	184.09
125.84	63.435	4.3987	4.0856	NaN

A^{F2}

NaN	360	342.76	286.7	250.02	305.22
180	NaN	232.13	196.86	195.38	215.31
162.76	52.125	NaN	182.2	183.47	205.2
106.7	16.858	2.2026	NaN	188.13	322.13
70.017	15.376	3.4682	8.1301	NaN	339.44
125.22	35.311	25.201	142.13	159.44	NaN

Figure 6. Illustration of Magnitude and Angle Matrices in Two Frames

MV

4	0	0	0	0	0
3	0	0	2	2	1
0	1	2	0	1	0
0	1	2	0	0	0
2	0	0	2	3	1

Figure 7. Illustration of Match Vector Count Matrix

Then, the weight matrix (W) is computed by dividing the match vector count matrix (MV) by the total number of features in the first frame, as demonstrated in equation (6).

$$W_{(j,n)} = \frac{MV_{(j,n)}}{NF^{f1}} \quad (6)$$

Where NF^{f1} is the total number of features in the first frame, and W is the weight matrix. $W_{(j,n)}$ is the similarity weight of the j^{th} feature in the first frame and the n^{th} feature in the second frame. Since there are five features in the first frame and six features in the second frame, the resulting weight matrix has dimensions of 5×6, as illustrated in Figure 8.

Based on the weight matrix, the matrix of matched features, which is illustrated in Figure 9, is calculated according to equation (7).

$$MF(i,1) = \max(W(:,i)) \quad (7\text{ a})$$

$$MF(i,2) = index(\max(W(:,i))) \quad (7\text{ b})$$

$$MF(i,3{:}4) = F^{f1}(MF(i,2),:) \quad (7\text{ c})$$

$$MF(i,5) = i \quad (7\text{ d})$$

$$MF(i,6{:}7) = F^{f2}(i,:) \quad (7\text{ e})$$

Where MF is the matched feature matrix, W is the weight matrix, No^{f1} is the feature numbers of the first frame, x^{f1} and $y^{f1}$are the x and the y locations of features in the first frame, No^{f2} is the feature numbers of the second frame, $x^{f2}$and $y^{f2}$are the x and the y locations of features in the second frame.

As depicted in Figure 9, feature number 3 of the first frame is matched with feature numbers 2 and 3 of the second frame, each assigned weight of 0.2 and 0.4. In response to this scenario, only one feature is obtained between feature numbers 2 and 3 of the second frame by considering their respective weights, and one feature is eliminated as the invalid feature, as illustrated in Figure 10.

W

0.8	0	0	0	0	0
0.6	0	0	0.4	0.4	0.2
0	0.2	0.4	0	0.2	0
0	0.2	0.4	0	0	0
0.4	0	0	0.4	0.6	0.2

Figure 8. Weight Matrix

MF

W	No^{f1}	X^{f1}	Y^{f1}	No^{f2}	X^{f2}	Y^{f2}
0.8	1	347	57	1	357	65
0.2	3	425	89	2	429	65
0.4	3	425	89	3	415	83
0.4	2	375	89	4	363	85
0.6	5	373	93	5	349	87
0.2	2	375	89	6	381	99

Figure 9. Matched Feature Matrix

VMF

W	No^{f1}	X^{f1}	Y^{f1}	No^{f2}	X^{f2}	Y^{f2}
0.8	1	347	57	1	357	65
0.4	2	375	89	4	363	85
0.4	3	425	89	3	415	83
0.6	5	373	93	5	349	87

Figure 10. Valid Matched Feature Matrix

4.3 Defining Possible Bounding Box and Creating Predicted Bounding Box

Defining a possible bounding box for the next frame is carried out based on the ROI in the previous frame, as shown in equation (8) and as depicted in Figure 11.

$$ROI = [x, y, w, h] \quad (8\text{ a})$$

$$ROI^{possible} = [x - p, y - p, w + (2 * p), h + (2 * p)] \quad (8\text{ b})$$

Where ROI is the original ROI of the first frame, and $ROI^{possible}$ is the possible ROI for the next frame. x, y, w and h are the x location, y location, width, and height of the ROI. p is the number of pixels.

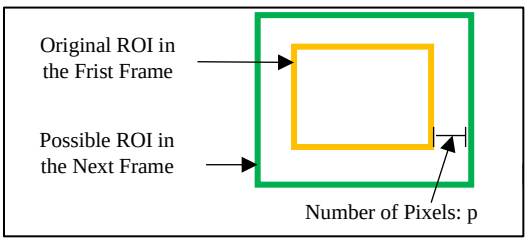


Figure 11. Illustration of Defining Possible ROI

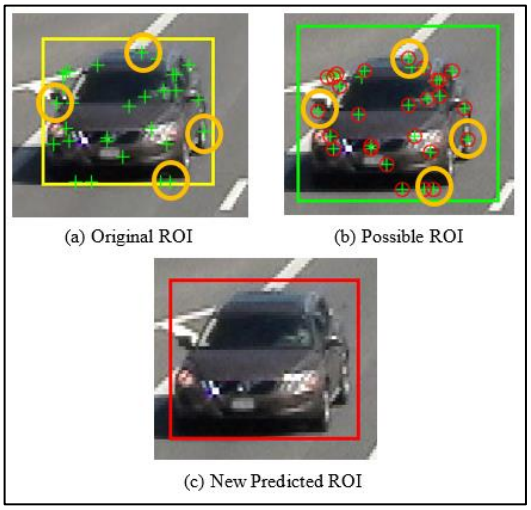


Figure 12. Creating Predicted ROI using Four Terminal Inlier Matched Features

In Figure 12, creating predicted ROI using four terminal inlier matched features is depicted. In Figure 12 (a) and (b), the red circles are the matched features, and the yellow circles are four terminal inlier matched features between the two consecutive frames. The predicted ROI shown in Figure 12 (c) is created using the distances of those features and the bounding box.

5. EXPERIMENTAL RESULT

The proposed system has been implemented and tested on the MATLAB R2021a platform. To validate the performance of the proposed system, a video file available on YouTube has been utilized. Comparison results can be seen in table 1, which compares the proposed system and the KLT algorithm specifically in the context of moving object tracking during a partial occlusion event.

In Figures 13, 14 and 15, the green rectangle is the possible bounding box, and the red rectangle is the predicted bounding box. All the '+' signs in the green box are the morphological corner features; the red circled '+' signs are the inlier-matched features; the yellow circled '+' signs are the outliers; and the blue point is the tracking point. Figure 13 illustrates the object tracking before partial occlusion, and Figure 14 and Figure 15 depict that after partial occlusion.

Figure 13. Tracking Object before Partial Occlusion

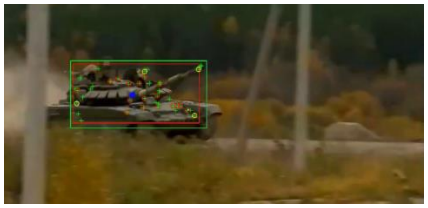


Figure 14. Tracking Object after Partial Occlusion 1

Figure 15. Tracking Object after Partial Occlusion 2

Table 1. Comparison Results

No.	Number of Frames	Tracked Frame		Tracked Index	
		KLT	Proposed	KLT	Proposed
1	50	50	50	1	1
2	100	80	100	0	1
3	60	60	60	1	1
4	129	100	129	0	1
5	136	136	136	1	1
6	150	150	140	1	0
7	70	70	70	1	1
8	200	200	200	1	1
9	250	230	230	0	0
10	340	340	340	1	1
Total	1485	1416	1455	**70%**	**80%**

In the comparison results shown in table 1, the number of frames in the testing video, the number of tracked frames, and the tracked indexes are described. If the system can track all frames in the video, the tracked index is set to 1. If not, it is set to 0. In table 1, the KLT algorithm was able to track 7 out of 10 videos, but it failed in 3 videos because of partial occlusion which can be seen in Figures 13 and 14. The proposed system was

able to track 8 out of 10 videos, and it was able to overcome the partial occlusion event in the testing videos. So, the accuracy of the proposed system is 80%, and that of the KLT is 70%.

6. CONCLUSIONS

Object tracking plays a fundamental role in various domains. It is finding the position of the interested object in every frame of the video. The existing methods have many drawbacks, such as the disappearance of trackable features in the condition of partial occlusion. In this paper, single object tracking using morphological-based vector feature is proposed. The experimental result shows the proposed system can track the object with an accuracy of 80% and overcome partial occlusion events, which is not achievable by the KLT algorithm. This showcases the proposed system's ability to robustly track moving objects in various scenarios, making it a valuable tool for applications in computer vision, surveillance, and the military sectors.

ACKNOWLEDGMENT

We would like to express our sincere thanks to Dr. Soe Win Myint, Head of Department of Computer Science, and Dr. Zar Nay Linn who have, as always, given us their support throughout writing of this paper.

REFERENCES

[1] Barga Deori and Dalton Meitei Thounaojam, “A Servey on Moving Object Tracking in Video”, 2014 International Journal on Information Theory (IJIT).

[2] Li Tan, Xu Dong, Chongchong and Yuxi Ma, “A Multiple Object Tracking Algorithm Based on YOLO Detection”, 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI 2018).

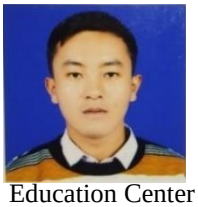
[3] Mr. Pramod R. Gunjal, Miss. Bhagyashri R. Gunjal, Mr. Haribhau A. Shinde, Mr. Swapnil M. Vanam and Mr. Sachin S. Aher, “Moving Object Tracking using Kalman Filter”, 2018 International Conference on Advances in Communication and Computing Technology (ICACCT).

[4] M. Sushma Sri, “Object Detection and Tracking using KLT Algorithm”, 2019 IJEDR 5th International Conference on Computer and Communications.

[5] Robert Laganie, “A Morphological Operator for Corner Detection”, School of Information Technology and Engineering, University of Ottawa, Canada, Pattern Recognition Society 1998.

Authors

Biography



Myo Min Paing received M.C.Sc. degree from Higher Education Center (HEC), in 2018. Now he is a Ph.D. candidate of Computer Science Department, Higher Education Center (HEC).

Dr. Myo Min Hein earned his Master’s Degree from Russian State Technological University (MEPHI), (Russian Federation) in 2010. He obtained his Ph.D. (Computer Science) from HEC in 2018. Now he is an assistant lecturer of Department of Computer Science in HEC.

Dr. Bawin Aye earned his master degree from National Research University of Electronic Technology (MIET), (Russian Federation) in 2006. He also obtained his Ph.D. degree from National Research University of Electronic Technology (MIET), (Russian Federation) in 2010. Now he is a lecturer of Department of Computer Science in HEC. He is also a member of the IEEE Computer Society.

TEXT IMAGE PREPROCESSING AND SEGMENTATION FOR PALI OPTICAL CHARACTER RECOGNITION

Kaung Myat Soe[1], Yu Ya Win[2], and San San Tint[3]

[1,2] Faculty of Computer Science, University of Computer Studies (Mandalay),

[3] Pro-Rector, University of Computer Studies (Mandalay)

[1]kaungmyatsoemdy1@gmail.com,[2]yuyawin@gmail.com,[3]dr.sansantint@gmail.com

ABSTRACT

The image preprocessing step is a crucial part of the Optical Character Recognition (OCR) system, as the recognition accuracy of OCR systems greatly depends on the quality of the input text image. Most of the OCR systems, especially the Pali OCR systems, require little effort for preprocessing or image enhancement. And Pali glyphs are complex in shape; when combined with many circulars, dots, lines, and pixel distributions to form these scripts, they are not uniform. Therefore, making the input image better is important for the OCR system. In this paper, we propose image preprocesses that, through reducing noise, separating text and background, skew detection, and rotation, improve the accuracy of the OCR system for Pali printed text. These steps can produce a refined image that is ready for the segmentation or character extraction process of the OCR system. In text regions segmentation, projection profile method is used to detect text regions in the image.

KEYWORDS

Pali, Tesseract OCR, Image Processing, Regions Segmentation

1. INTRODUCTION

An OCR system is the recognition of written or printed text characters from input text-scanned images or printed document images by using technology for easy editing, storage, transmission, searching, indexing, and integration into other applications [14]. Some practical application areas of OCR system are: Reading aid for the blind, Automatic text entry into the computer for desktop publication, library cataloging, ledgering, etc. Automatic reading for sorting postal mail, bank checks, and other documents; document data compression; language processing; multi-media system design; etc [4].

Current solutions to the problem of optical character recognition (OCR) have advanced to the point where recognition rates of 99% are common for clean, uniformly formatted text. Unfortunately, the performance of most OCR algorithms degrades very rapidly when even small amounts of noise are introduced into the original document or during the scanning process. In many situations, this increased error rate quickly decreases the return on investment to the point where it is not cost-effective to integrate automated recognition technology solutions. To push this critical point lower and deal robustly with noise, OCR systems often perform some type of image enhancement as a preprocessing step [11]. And every process can affect the recognition accuracy, and each step depends on the previous one. For example, an image that includes noise or a low-quality image can cause improper components in the segmentation process. Therefore, preprocessing steps are the most important part of all OCR systems.

The state-of-the-art OCR engines recognize documents printed in Latin and some Oriental scripts with few errors on each page for high-quality images. There is no robust OCR for Pali scripts in the Prakrit language. And the previous Pali OCR systems were mostly

concerned with characters and placed less emphasis on image enhancement processes. The writing style of Pali scripts is complex orthography and consists of circles, dots, and lines, either horizontally or vertically. The pixel values are not uniform for some Pali fonts. The wrong pixel content can affect the follow-up process and decrease OCR accuracy. Therefore, making the input image better is important for our native OCR system.

2. PROPOSED SYSTEM

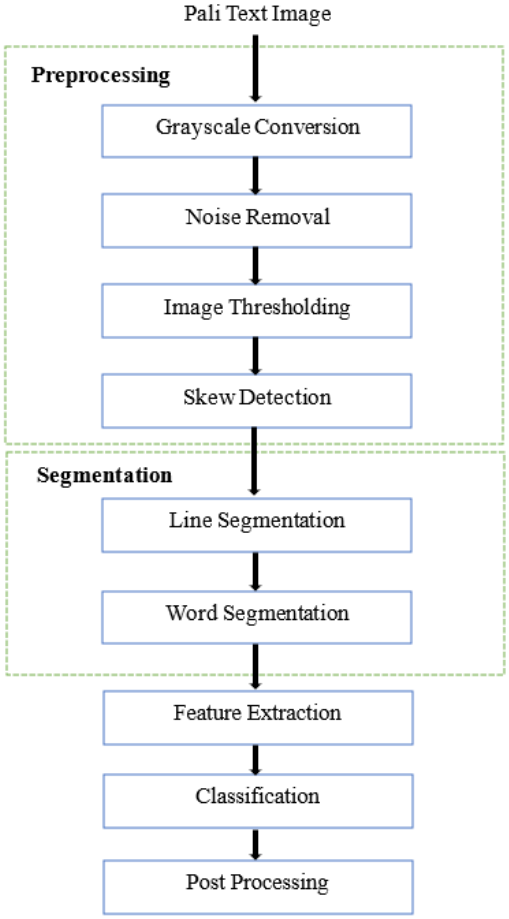


Figure 1. Pali OCR System Architecture

Preprocessing operations are usually specialized image processing operations that transform the image to improve the quality of the image with reduced noise and variation. It takes in a raw image, reduces noise and distortion, and removes the skewness of the image, thereby simplifying the processing of the rest of the stages.

After preprocessing operations are performed, the image is segmented into characters before being passed to classification phase. Segmentation is an integral part of any text-based recognition system. It assures efficiency of classification and recognition. Accuracy of character recognition heavily depends upon segmentation phase [16].

The rest of the paper is structured as follows: Section 2 discusses related works for the image cleaning process; the implementation of the detailed proposed preprocessing steps is described in Section 3, the segmentation steps is described in Section 4 and we conclude this paper in Section 5.

3. RELATED WORKS

Although much literature states the various implementation strategies for OCR systems, there is little suitability for the image preprocessing or enhancement process [11] and segmentation.

D. R. Allan et .al [5] provided for identifying, correcting, modifying, and reporting imperfections and features in pixel images that prevent or hinder proper OCR and other document imaging processes. Their invention provides that run-length compressed images can be analyzed and corrected directly for improved performance.

Elisa H. et al. [6] explored the effects of different image preprocessing methods for document image binarization. They proved that the binarization method is significant in terms of binarization accuracy, but the preprocessing also plays a significant role. The Total Variation method of preprocessing shows the best performance over a variety of preprocessing methods. And they conclude that no one binarization algorithm is uniformly best over all possible images, and neither is one preprocessing algorithm.

Megan E. et al. proposed an image preprocessing suite that, through text detection, auto-rotation, and noise reduction, improves the accuracy of OCR analysis by using morphological strategy [12]. However, they stated that more research is needed for noise deduction techniques, integration of other preprocessing steps, and better enhancement of letter shapes for increased legibility.

4. IMAGE PREPROCESSING

To implement the image correction process, we gain the following image optically, including some noise and variants:

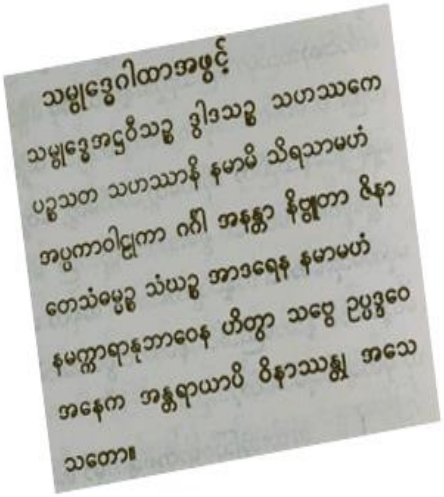


Figure 2: Original input image

4.1. Changing Colorspaces

The process of converting the color pixel value, which is described by a triple (R, G, B) of intensities for red, green, and blue, to a single grayscale value. There are many methods for converting to grayscale images. We use the luminosity method.

The luminosity method is a more sophisticated version of the average method. It averages the values, but a weighted average is to account for human perception. We're more sensitive to green than other colors, so green is weighted most heavily [9]. The effective luminance of a pixel is calculated with the following formula:

$$Y=0.21RED+0.72GREEN+0.07BLUE \quad (1)$$

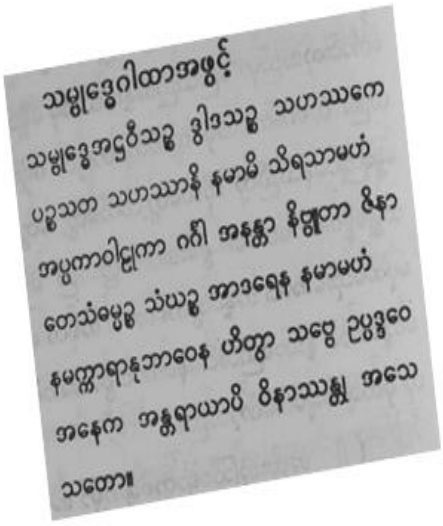


Figure 3: Image after converting Grayscale

4.2. Noise Removal

The noise, which is introduced by the optical scanning devices, causes disconnected line segments, bumps and gaps in lines, filled loops, etc. The distortion, which includes local variations, rounding of corners, dilation, and erosion, is also a problem. Prior to character recognition, it is necessary to eliminate these imperfections. There are many techniques to reduce the noise, which can be categorized into two major groups: filtering and smoothing [8], [13].

Filtering is perhaps the most fundamental operation of image processing and computer vision. Images can be filtered with various low-pass filters (LPF), high-pass filters (HPF), etc. LPF helps in removing noise, blurring images, etc. HPF filters help in finding edges in images. Then, image smoothing, or blurring, is achieved by convolving the image with a LPF kernel. It is useful for removing noise. It actually removes high-frequency content (e.g., noise, edges) from the image [19].

The system used Gaussian blurring for our noise removal process. Gaussian blurring convolves the source image with the specified Gaussian kernel [20]. The width and height of the Gaussian kernel should be specified positive and odd.

It specifies the standard deviation in the X and Y directions, sigmaX and sigmaY respectively. If only sigmaX is specified, sigmaY is taken as the same as sigmaX. If both are given as zeros, they are calculated from the kernel size. Gaussian blurring is highly effective in removing Gaussian noise from an image. The Gaussian kernel size and the matrix of Gaussian filters coefficients can be computed with the following equation [21]:

$$G_i = \alpha * e^{-(i-(ksize-1)/2)^2/(2*sigma^2)} \quad (2)$$

where i=0..ksize−1 and α is the scale factor chosen so that ∑iGi=1.

ksize = Aperture or kernel size. It should be odd (ksize mod 2=1) and positive.

sigma =Gaussian standard deviation. If it is non-positive,

it is computed from ksize as sigma = 0.3*((ksize-1)*0.5 - 1) + 0.8

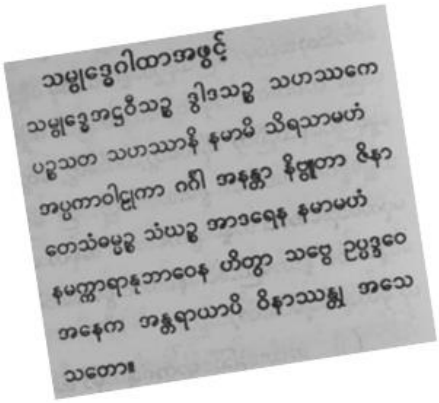


Figure 4: Image after removing noise using Gaussian Blurring

4.3. Image Thresholding

Image thresholding, or Binarization, is a conversion from a gray-level image to a bi-level image by turning all pixels below some threshold to zero and all pixels about that threshold to one.
A local method is a method that computes a threshold for the neighborhood around each pixel or for each designated block in the image. The top-ranked local threshold methods in terms of the error rate and rejection rate for character recognition are proposed by Niblack and Eikvil et al. If the binarization method computes a threshold for an entire image, it is called a global method. Many researchers evaluated four such methods and concluded that Otsu's approach outperformed the other three [3].

The Otsu method also appears to best model the truth behind historical printed documents, based on results from a broad range of historical newspapers. The plugin for most image processing tools also uses the Otsu thresholding technique. And it can find the threshold automatically based on the input gray data. It is a simple but effective tool to separate objects from the background [17]. The Otsu threshold is used in many applications, from medical imaging to low-level computer vision. It has many advantages and makes many assumptions.

Advantages
• Speed: Because the Otsu threshold operates on histograms (which are integer or float arrays of length 256), it's quite fast.

• Ease of Coding: Approximately 80 lines of very easy stuff.

The limit of this method is that it is applicable only when the image is bimodal (the histogram contains two peaks).

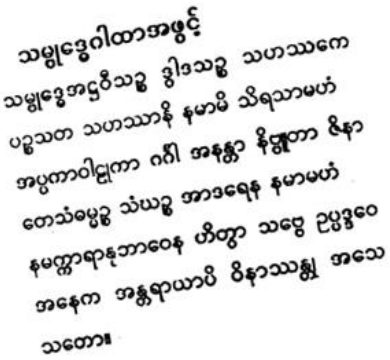


Figure 5: Image after using Otsu threading

4.4. Skew Detection

The scanned document may be slanted because of human error, which is the misplacement of the paper document on the scanner. This can reduce the OCR performance. Even the smallest skew angle existing in a given document image results in the failure of segmentation of complete characters from words or text lines as the distance between the characters reduces. Further, most of the OCRs and document retrieval and display systems are very sensitive to skews in document images. Hence, it is important to detect and correct skew because it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages [2], [7].
The process of attempting to estimate the orientation angle, or skew angle, of the text lines is called skew detection. The process of rotating the document with the skew angle in the opposite direction is called skew correction. The most popular algorithm for detecting skews is the Hough transformation. Generalized Hough Transform (GHT) is an extension of the standard HT to detect any arbitrary object in an image and can be used to recognize printed characters in their different shapes [18] and [15]. We use the generalized Hough transformation to detect the angle of an image's baselines and rotate the image to the correct angle.

The basic algorithm in pseudo code [10]:
1. Create a two-dimensional matrix Hough and initialize the values with 0
2. for y=0 to Height-1
3. for x=0 to Width-1
4. if Point(x,y) is black then
5. for alpha=-20 to 20 step 0.2
6. d= Trunc(y*cos(alpha)-x*sin(alpha))
7. Hough (Trunc(alpha*5),d)+=1
8. next alpha
9. end if
10. next x
11. next y
12. Find the top 20 (alpha,d) pairs that have the highest count in the Hough matrix
13. Calculate the skew angle as an average of the alphas
14. Rotate the image by – skew angle

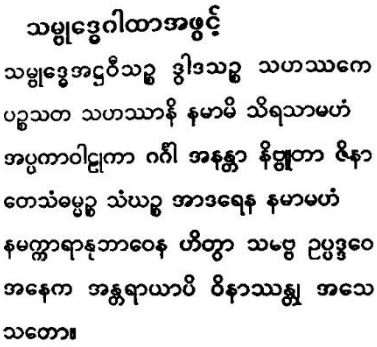


Figure 6: Image after correction of skew using Generalized Hough Transform

5. FUTURE WORK

Now, it is ready to detect the text line in the deskewed image. In this system, projection profile refers to the projection of the sum of hits and positives along an axis from a bi-dimensional image. The projection profile method is primarily used for segmenting text objects present inside text documents. A projection profile is calculated for a thresholded image or binarized image, where a thresholded image is a grayscale image with pixel values of 0 or 255. Image pixels are replaced by 1 and 0 for pixel values 0 and 255, respectively [22].

5.1. Line Segmentation

The horizontal projection profile method is used to calculate sum of all white pixels on every row. It gives corresponding histogram of that image [1]. The steps of line segmentation are as follows:

1. Construct the horizontal histogram of image.
2. Find the threshold value which is gap between corresponding two histogram.
3. Separate each histogram by threshold value and save it
4. We get segmented lines from image

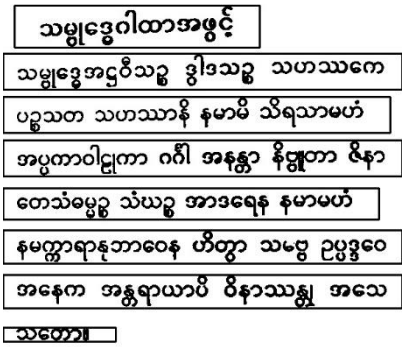


Figure 7: Segmented Lines

5.2. Word Segmentation

In word segmentation we use vertical projection profile method to calculate sum of all white pixels. Plot the histogram of computed white pixels [1]. The steps of word segmentation are as follows:
1. Construct the vertical histogram of the image.
2. Find the threshold value
3. separate corresponding histogram from each other and save it
4. The saved images are segmented words from line
5. Repeat process for each line

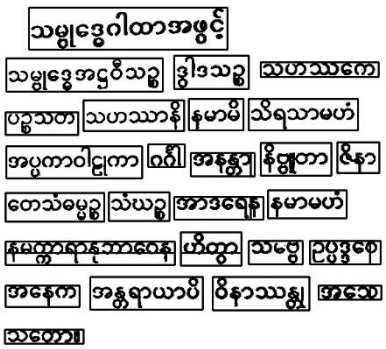


Figure 8: Segmented Words

6. CONCLUSIONS

We described the way for image enhancement processes as a first step for our Pali printed OCR system. After this process, we

approached the next process of Segmentation where we isolate the character form the image outputted from the preprocessing step. The good way of image enrichment process may prevent the occurrence of the wrong segmentation and then may increase the overall accuracy of the OCR system.

REFERENCES

[1] A. J. Kant, and A. J. Vyavahare, "Devanagari OCR Using Projection Profile Segmentation Method", International Research Journal of Engineering and Technology (IRJET), Vol.3, Issue 07, July 2016, pp. 132-134.

[2] A. M. Al-Shatnawi and K. Omar, "Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity", Journal of Computer Science 5 (5):, Science Publications,2009, ISSN 1549 3636, pp. 363-368.

[3] C. H. Chou, W. H. Lin, FuChang, "A binarization method with learning-built rules for document images produced by cameras", Pattern Recognition, Vol 43, Academia Sinica, Taipei, Taiwan, 2010, pp.1518-1530.

[4] D. Achaya U, N. V. Subba Reddy and Krishnamoorthi, "Hierarchical Recognition System for Machine Printed Kannada Characters", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.

[5] D. R. Allan and B. B. Fast, "OCR Image Pre-Processor", United State Patient, No. 5778103, July 7, 1998.

[6] E. H. B. Smitha, L. L. Sulemb and J. Darbon, "Effect of pre-processing on binarization", Boise State University, Boise, Idaho, USA; Telecom ParisTech, Paris, France, UCLA, Los Angeles, California, USA, 2010.

[7] G. H. Kumar, Nandini N. and Srikanta M. K., "Estimation of Skew Angle in Binary Document Images Using Hough Transform", World Academy of Science, Engineering and Technology 42 2008.

[8] Gonzalez, R., Woods, R., Digital Image Processing, 2nd ed., Prentice Hall, 2002.

[9] John, "Three algorithms for converting color to grayscale", August 24, 2009.

[10] Mackenb, "how to deskew an image", 25 Apr 2006.

[11] M. Agrawal, D. Doermann, J. Kumar, W. A. Almageed, "Document Image Enhancement", Monday 30 August, 2010.

[12] M. Elmore and M. Martonosi, "A Morphological Image Preprocessing Suite for OCR on Natural Scene Images", Georgia Institute of Technology, Princeton University, 2008.

[13] Pratt, W.K., Digital Image Processing, 2nd edition, John Wiley & Sons Inc., 1991.

[14] S. K. Shah and A. Sharma, Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching", Gujarat.IE(I) Journal−ET, 2006.

[15] S. Touj, N. B. Amara, and H. Amiri, "Generalized Hough Transform for Arabic Printed Optical Character Recognition", The International Arab Journal of Information Technology, Vol. 2, No. 4, October 2005.

[16] Thiri Mon, "Myanmar Characters Segmentation and Recognition Using Tesseract Engine", Journal of Information Technology, Research and Innovation, Vol-2, Issue-1, December 2022, pp.173-178.

[17] Xu Liang, "Image Binarization using Otsu Method", NLPR-PAL Group CASIA, April 20th, 2009.

[18] www.wikipedia.com

[19] https://docs.opencv.org/3.4/d4/d13/tutoria l_py_filtering.html

[20]https://docs.opencv.org/3.4/d4/d86/group__imgproc__filter.html#gaabe8c836e97159a9193fb0b11ac52cf1

[21]https://docs.opencv.org/3.4/d4/d86/group__imgproc__filter.html#gac05a120c1ae92a6060dd0db190a61afa

[22] https://www.geeksforgeeks.org/line-segm ent/

Authors

Biography

First A. U Kaung Myat Soe was graduaded B.C.Sc in 2020 and M.C.Sc in 2023. Now, he works at the Faculty of Computer Science in University of Computer Studies (Mandalay). He is interested in Cloud Computing and Artificial Intelligence (AI).

ASPECT-BASED SENTIMENT CLASSIFICATION SYSTEM FOR RESTAURANT REVIEWS USING MULTINOMIAL NAÏVE BAYES CLASSIFIER

Han Thi Phoo[1]

[1] Faculty of Computer Science, University of Computer Studies (Mandalay)
[1]akariphoo8@icloud.com

ABSTRACT

There are a large amount of unstructured opinions available on the website and other social media platforms which are expressed by the users as reviews through their thoughts. And it makes the customers overwhelmed with information that makes it difficult to use the available information to decide for a customer. whether it is a good place to eat out or not. Sentiment analysis of customer reviews has a crucial impact on any business's development strategy. Recently, aspect-based sentiment analysis has been introduced which extracts the various aspects in the opinion text and classifies them into positive or negative. This paper presents the aspect-based sentiment analysis system that extracts the aspects efficiently from the user opinion review sentences and performs accurate classification. In this system, the restaurants are rated according to positive, and negative polarity scores of each aspect in restaurant reviews with the help of sentiment classification algorithms of Multinomial Naïve Bayes (NB).

KEYWORDS

Sentiment Analysis, Natural Language Processing, Multinomial Naïve Bayes

1. INTRODUCTION

Today many users write their thoughts on online media such as Facebook, Instagram, Twitter, and other social media platforms, in the form of reviews. Online reviews of customers on multiple social media channels are used for analysis because social media platforms help analyze the general public's opinions on many different subjects. These reviews are analyzed and reviewed by organizations and businesses, and they try to find out patterns and reasons behind the reviews and then reuse the gathered information for their strategic planning and business intelligence. All of these are driving people toward sentiment analysis. The requirement for sentiment deduction in texts has provided wide usage in natural language processing and computer science fields such as sentiment analysis [3].

The social web (also known as Web 2.0 services) enhancement has had the potential impacts on the flow of traveling and provided the greatest innovations. Food vlogging channels are more popular and provide useful information for the customers with the decision of choosing which restaurant to eat out and acquiring which service in that restaurant. The development in the reviewing and feedback opportunities of the food vlogging web services have caused the internet as the major observation and obtaining the information concerned with the foods and restaurants created by people. The customer-created reviews and restaurant content on food vlogging services are developing at the rate of exponential and this can support many opinions of various people. By applying web portals, the preferences and opinions of customers can be shared and many reviews can be created by daily (Hu et al., 2017) [5].

Sentiment analysis analyzes the feelings, attitudes, and opinions of the customer about the specific services and products. This exchanges the information applicable business information through unstructured text data. The sentiment analysis performs the

classification of the opinions into negative, neutral, and positive. The feelings, attitudes, and experiences concerned with foods and restaurants are expressed on social media by posting reviews about views, restaurants, places, decorations, and services given by the restaurant. In this paper, the reviews of restaurants are collected to perform sentiment analysis. The classification of sentiment analysis can be performed at the aspect level, document level, and sentence level. The positive reviewed document doesn't mean that the customer likes everything. The negatively reviewed document on a particular object doesn't mean that the customer has negative reviews on all features of the object. The sentence and document level analysis fails such information in this condition. Hence, aspect-based sentiment analysis is proposed for achieving detailed information for every aspect.

This paper proposes a model for aspect-based sentiment analysis for restaurant reviews and comments in English language using Multinomial Naïve Bayes. The division of aspect-based sentiment analysis is divided into three parts, pre-processing the review data which includes tokenization, stop words elimination, lemmatization, stemming for each of the review sentences, the extraction of aspect terms, and classification of sentiment polarity scores for those aspects. In this paper, the next section 2 describes about the related works of the proposed system, and then section 3 presents the proposed methodology. Section 4 is about the proposed system and performance evaluation is in section 5. Finally, the system is concluded in section 6.

2. RELATED WORKS

Kande Trupti V and H. P. Shah introduced deep learning-based sentiment analysis for faster aspect classification with high accuracy [4]. Many experiments on real-time review sentiment analysis have been done the determine how the effectiveness of the framework to aid tourists in indicating good hotels, restaurants, and locations in a region. This paper presented an aspect-based sentiment analysis for the review's classification among aspects into positive or negative. In this paper, the aspect deduction approach according to tree structure is introduced for the explicit and implicit aspect deduction through the reviews of tourists. The deduction of noun phrases and frequent nouns through opinions is performed and similar nouns are grouped with WordNet.

Convolutional neural networks are applied for sentiment analysis in that extracted nouns are utilized as the leaf of a tree and review words are utilized as internal nodes. Firstly, the removal of irrelevant and fewer opinion sentences is performed with the utilization of Stanford basic dependency at every sentence. Then, the extraction of features through the remaining sentences is done by POS Tags and N-Grams for the training of classifiers. Finally, the extraction of features for the training of classifiers is taken by utilizing machine learning approaches.

P. S. Bhargav, G. N. Reddy, R. R. Chand, K. Pujitha, and A. Mathur introduced the sentiment analysis system for hotel reviews. On the website, the preferences of hotels were put by the users according to the reviews relating to the rating of the hotel and the analysis by other hotels [2]. The system implementation was done by applying Naïve Bayes machine learning approaches and opinion mining approaches according to natural language processing.

The framework for a sentiment analysis system based on aspects was presented for the efficient identification of the aspects by high accuracy utilizing decision trees and naïve Bayes machine learning approaches [1]. This framework performed the classification of aspects into negative and positive. This tree-based extraction approach deducts the implicit and explicit aspects through tourist reviews. This system deducted nouns and noun phrases through reviews and similar nouns are categorized by applying WordNet.

This system aids the tourists in searching for the optimal hotel, place, and restaurant in the region, and the experimental evaluation was performed with Foursquare and Yelp datasets.

Alhadethy, Mohanad and Hamad, Mortadha and Adnan Jaleel, Refed analyze the restaurant reviews on Twitter using Naïve Bayes and implementing sentiment classification algorithms for customer reviews on social media [7]. The system collected real data to evaluate customer opinions on restaurant services and achieved performance assessment

measures like precision, recall, accuracy, and error rate.

3. METHODOLOGY

Methods of sentiment analysis can be categorized into three approaches such as lexicon-based approach, machine learning and deep learning-based approach, and hybrid approach (combine machine learning and lexicon-based approach). The two main machine learning techniques used for sentiment analysis are supervised learning and unsupervised learning. In this paper, we use a supervised machine-based approach for this study. Machine learning-based approaches classified the opinions of aspects into classes using different machine learning algorithms.

3.1. Sentiment Analysis

Sentiment analysis or opinion mining analyzes the text written in a natural language about a topic or product and classifies it as positive, negative, or neutral based on the writer's sentiments, emotions, and opinions expressed in it. There are different sentiment classifications that are namely sentence, aspect, or document. The sentence-based analysis finds the sentiment of a sentence. Aspect-based analysis finds the property of an entity in a sentence. The third approach finds the sentiment of the whole document. And there are various levels of sentiment analysis namely word-based, sentence-based, document-based, and feature-based.

3.2. Natural Language Processing

Natural Language Processing (NLP) is used for the process of computationally identifying and categorizing opinions expressed in the review text, especially in order to determine whether the writer's sentiment or attitude toward a particular topic or product is positive or negative. It is a subfield of linguistics, computer science, information engineering, and artificial intelligence focused on human communication, such as speech and text, comprehensible to computers, in particular, how to program computers to process and analyze large amounts of natural language data. In this system, NLTK was used to implement the natural language processing in Python and to make predictions that a computer can predict if it is positive or negative based on the given text review.

3.3. Multinomial Naïve Bayes Algorithm

Multinomial Naïve Bayes (MNB) is a very popular and efficient machine learning algorithm that is based on Bayes's theorem. It is commonly used for text classification tasks in NLP that we need to deal with discrete data like word frequency counts in documents. The term multinomial refers to the type of data distribution assumed by the model. The features in text classification are typically word counts or term frequencies. The multinomial distribution is used to estimate the likelihood of seeing a specific set of word counts in a document.

In the context of MNB, it can be assumed that the features (words) are conditionally independent given the class. This simplifies the computation significantly and leads to the following formula:

$$P(C_k|x_1, x_2,..., x_n) \propto P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \quad (1)$$

Where:
x_i represents the i-th feature (word).
$P(C_k)$ is the prior probability of class C_k.
$P(x_i|C_k)$ is the probability of observing feature x_i given class.

3.4. Support Vector Machine Algorithm

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. In the case of classification, SVM aims to find a hyperplane that best separates the data points of different classes in the feature space. The formula for SVM can be understood in terms of the decision function and the margin.

4. THE PROPOSED SYSTEM

Firstly, the dataset is the most important part of this proposed system and it is selected according to the interest that best suits it. This

dataset is then imported into the machine learning platform and then it is distributed as training data as the entire data that will be used. This data is pre-processed by using data cleaning methods like tokenization, POS tagging, elimination of stop words, lemmatization, stemming, Word2Vec, etc. Using this cleaned data, the various aspects are extracted and the polarity of the aspects are found out.

The extracted aspects are added to the aspect category based on the pre-defined aspect category. Then the classifier is trained and the overall score for the extracted aspects is obtained by applying the algorithm. Multinomial Naïve Bayes algorithm is used for the aspect-based sentiment classification to achieve the highest possible accuracy when the aspects are considered.

In this system, the aim to combine the aspects of various reviews and aggregate them is to get the accuracy score for each aspect, and the ratings for each individual will be shown based on the polarity of each aspect. Based on the aspect polarity, the classification rules will be laid down for semantic orientation that will be taken to find out whether the review is positive or negative. The proposed system design is described in Figure 1.

4.1. Dataset

This is a list of around 10,000 review comments on various restaurants from websites collected and it is available at https://www.kaggle.com [6]. The dataset includes 10,000 rows and 8 columns like the review column, and the rating column that classified the reviews as positive and negative. The text data are mixed by writing with formal and informal writing styles without segmentation and other special characters and emojis. Customers show positive, negative, and sometimes both positive and negative opinions in their reviews. In this system, 80% of the dataset is used for the training phase and 20% of the dataset is applied for the testing phase. As machine learning only accepts numerical values as inputs, the words of reviews need to be converted to floating point or integer values using the TF-IDF vectorizer. TF-IDF vectorizer is applied which converts the collection of word documents into the vector of token counts.

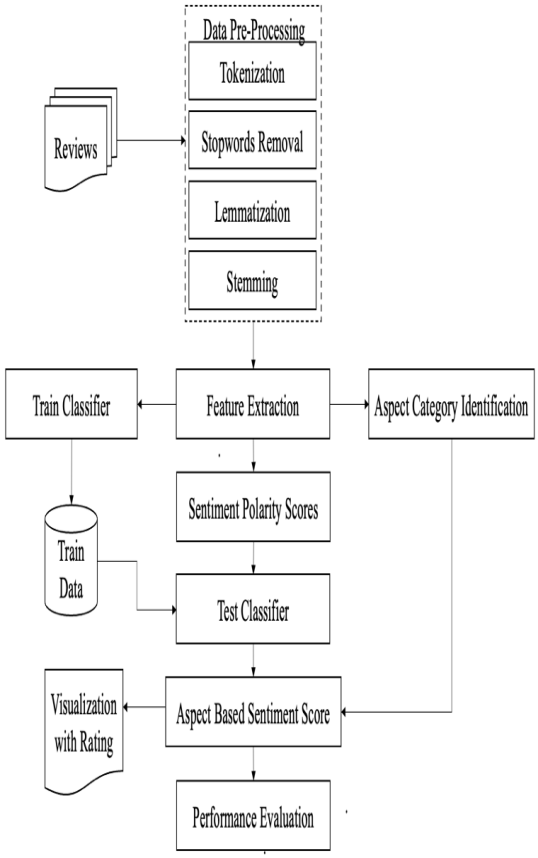


Figure 1. Proposed System Design

4.2. Data Pre-processing

Real-world data can include many errors that are often incomplete, inconsistent, and lacking in certain behaviors or trends. First of all, the collected review sentences are trimmed to lowercase. The data pre-processing steps for the collected reviews include tokenization, removing stop words such as repeated characters, hashtags, user names, punctuation marks, and emojis, and other steps like lemmatization, and stemming. All the pre-processing steps are done with the help of NLTK.

Stop Words Elimination

Original Text:
'The restaurant has a good staff, good food, and a good environment.'

After Tokenized:
['The', 'restaurant', 'has', 'a', 'good', 'staff', ',', 'good', 'food', ',', 'and', 'a', 'good', 'environment', '.']

After Removed The Stop Words
['restaurant', 'good', 'staff', 'good', 'food', 'good', 'environment']

Figure 2. Tokenization and Stop Words Elimination

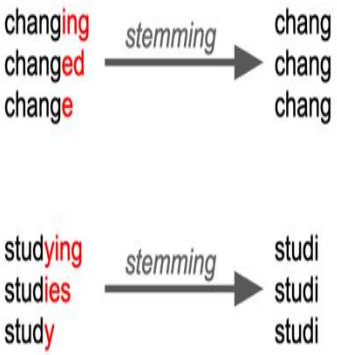


Figure 3. Stemming

Figure 4. Lemmatization

4.3. Feature Extraction

The bag-of-word model is widely used in NLP. In the Bag-of-Words (BOW) model, a text is represented by a set of words called the "bag-of-words". The BOW model is the method of pre-processing the text by formatting it as a number or vector, which finds the frequency of words in a given sentence. It is the extracting features method from the given text and uses these features for model training. BOW is a method of converting words to numbers by creating vector embeddings from the tokens generated previously.

This system finds out the semantic meaning of each particular word by applying the count vectorizer along with the TF-IDF on the dataset. TF-IDF vectorizer is used to convert the words into vectors. The reason for converting words to vectors is that the Multinomial Naïve Bayes Algorithm can understand data in the form of vectors as if predicted based on probability and each factor is independent of all the other various factors.

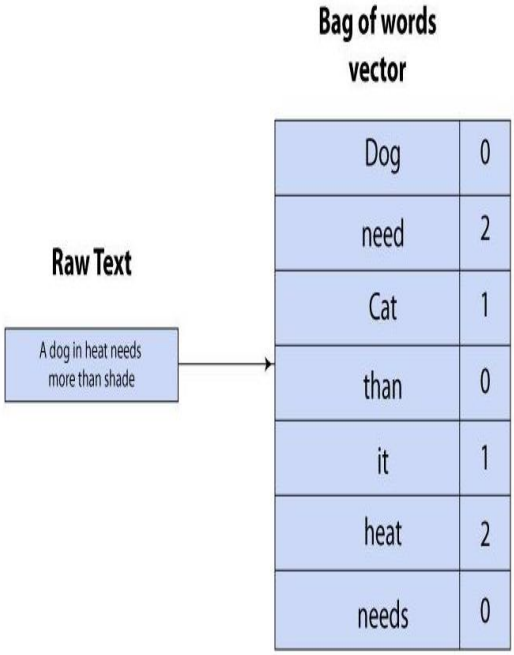


Figure 5. Count Vectorizer

4.4. Aspect Category Identification

After the pre-processing steps, the aspect category will be identified in which the reviews fall. Aspect extraction is an important step in aspect-based sentiment classification. As the reviews are assumed grammatically true, the aspects can be extracted on the basis of topic modeling.

In this system, topic modeling is performed using Latent Dirichlet Allocation (LDA) that finds the five topics, and the scores for each of the five topics are allocated in each review. LDA basically works by finding the topics inside the document and then by finding words inside those topics. According to five categories i.e. ambiance, f&b, hygiene, price, and service, this system identified a list of words belonging to each particular aspect category.

Each review is parsed and performed operations and then compared with the pre-defined list of words. The aspects of each review are then found out that is which exactly the review is talking about. The additional aspect category called miscellaneous is for all the related words of the reviews that don't fall in the pre-defined category list. The aspect word list is as shown in Table 1.

Table 1. List of Aspect Words

Ambiance	F&B	Hygience	Pricing	Service	Miscellaneous
ambiance	lobster	cleaniness	effective	service	neighborhood
furniture	rice	toilet	overpriced	teamwork	view
place	eggplant	basin	fair	waiter	fun
setting	beans	hygiene	pricey	waitress	location
style	green beans	dirty	saving	chef	enjoy
vibe	fish	neat	charges	caring	location
quality	soup	clean	discount	teamwork	delivery
atmosphere	buffet	trippy	money	judge	concept
experience	fries	sanitization	affordable	staff	recommend
decoration	stir fry	parking	price	security	convenience
seat	steak	toilets	expensive	workers	located

4.5. Finding the Polarity Scores

Once all the aspects are found, the aspect-based sentiment classification is done by collective aggregation of all the aspects, and the most used aspects are found from all the reviews. For predicting the polarities of the reviews, it starts building a model for the data that goes through various steps like tokenization, stop-word removal, stemming, and lemmatization. Then the weight is assigned and semantic value is given to words based on their overall occurrence. At last, a pipeline involving these steps is created and the polarity of the reviews that enter through the pipeline is predicted by the Multinomial Naïve Bayes model.

For example, 'The food is really good but the service was disappointing.'
Category: F&B =>food (positive)
Category: Service => service (negative)

Furthermore, the Vader library called the sentiment intensity analyzer is used to calculate the sentiment score for each review and the score is also normalized so that it can be easier to scale and evaluate the reviews concerning each other.

5. Performance Evaluation

The performance of the two different approaches was measured and evaluated with accuracy, precision, recall, and f-measure as shown in given equations 2 -5.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Accuracy measures the percentage of aspects of the reviews in the test set that the model correctly labeled either as positive or negative.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Precision is the ratio of correctly classified documents to the total number of documents classified under a particular category.

$$Recall = \frac{TP+TN}{TP+FN} \quad (4)$$

Recall is defined as the number of correctly classified documents among all documents belonging to that category.

$$F1\text{-}measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (5)$$

F-measures take into account the mean of precision and recall. The confusion matrix is also developed for analyzing the classification results showing the number of correct classifications (true positive (TP), true negative (TN)) and incorrect classifications (false positive (FP), false negative (FN)).

A total accuracy of 96.98% was obtained on the overall view of the dataset by using the Multinomial Naïve Bayes Algorithm. This system increases the accuracy of the result so as to enhance the performance of the result as compared to the previous approaches.

Table 2. Comparison Table for Different Algorithms

	Multinomial NB	**SVM**
Accuracy	0.95	0.91
Precision	0.98	0.93
Recall	0.96	0.91
F1-measure	0.97	0.92

6. Conclusions

Sentiment analysis or opinion mining is a comprised area of natural language processing that is used to analyze people's emotions, sentiments towards the product can be positive

or negative based on the writing. Restaurant reviews on the web are an important information source for people in travel planning and people who love to eat. This system is used to perform evaluation measures on comments obtained from the customer.

This system presented an aspect-based sentiment classification of restaurant reviews about aspects whether positive or negative. These positive and negative separations of aspects in user comments are used to analyze the quality of the restaurant. The system evaluates restaurant ratings according to each review of the restaurant.

Figure 6. Aspects and Review Rating Output

REFERENCES

[1] V. Rybakov, and A. Malafeev, "Aspect-Based Sentiment Analysis of Russian Hotel Reviews", Supplementary Proceedings of the 7th International Conference on Analysis of Images, Social Networks and Texts (AIST-SUP 2018), HSE, Moscow, Russia, July 5-7, 2018, pp. 75-84.

[2] P. Sanjay Bhargav, G. Nagarjuna Reddy, R.V. Ravi Chand, K. Pujitha, and Anjali Mathur, "Sentiment Analysis for Hotel Rating using Machine Learning Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6, April 2019.

[3] P. B. Nehe, and A. N. Nawathe, "Aspect based sentiment classification using machine learning for online Reviews", EasyChair Preprint no. 3051, March 26, 2020.

[4] Kande Trupti V, and H. P. Shah, "Recommendation System for Tourist Reviews using Aspect Based Sentiment Classification", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; Volume 10 Issue III Mar 2022.

[5] I. Perikos, K. Kovas, F. Grivokostopoulou, and I. Hatzilygeroudis, "A System for Aspect-based Opinion Mining of Hotel Reviews", In Proceedings of the 13th International Conference on Web Information Systems and Technologies (WEBIST 2017), SciTePress, Porto, Portugal, April 25-27, 2017, pp. 388-394.

[6]https://www.kaggle.com/datasets/joebeachcapital/restaurant-reviews

[7] Alhadethy, Mohanad & Hamad, Mortadha & Adnan Jaleel, Refed. (2021). Sentiment Analysis of Restaurant Reviews in Social Media using Naïve Bayes. Systems Analysis Modelling Simulation.

Authors

Biography

Han Thi Phoo achieved the B.C.Sc degree in 2019 and M.C.Sc degree in 2023 from the University of Computer Studies (Mandalay). Now, she is working as a tutor at the Faculty of Computer Science at the University of Computer Studies (Mandalay) and her research interests include Artificial Intelligence and Computer Vision. Her recent interest in research areas includes natural language processing and unsupervised learning.

A CASE STUDY ON MYANMAR-ENGLISH NAMED ENTITY DICTIONARY ENHANCEMENT

Aye Myat Mon[1], Mya Than Hnin[2], and Su Su Win[3]

[1,2,3] Faculty of Computer Science, University of Computer Studies (Mandalay)
[1]ayemyatmon.ptn@gmail.com

ABSTRACT

Named entity (NE) transliteration serves as a pivotal tool for phonetically transcribing names across languages with distinct writing systems. In the case of the Myanmar language, accomplishing a robust transliteration of named entities remains a formidable challenge due to the intricacies of its complex writing system and the paucity of available data. The Myanmar NE transliteration dictionary represents a significant milestone, having curated an extensive collection of over 135,255 NE instance pairs encompassing western person, organization, and place names. This dictionary stands as a witness to ongoing efforts to bridge linguistic gaps and facilitate cross-language communication. To refine the transliteration process, we conduct statistical experiments on a Phrase-based statistical machine translation (PBSMT) model, employing 2-Grams through 6-Gram's language models during decoding. This systematic approach allows for a nuanced exploration of transliteration accuracy at various linguistic units within the Myanmar script. The comparison of different units, specifically characters and syllables, within the Myanmar script is integral to our transliteration experiments. By scrutinizing the performance of these units, we gain insights into the nuanced challenges posed by the Myanmar writing system and the effectiveness of our chosen transliteration methods. Rigorous experiments are executed on both a 1,000-test dataset and a corresponding 1,000-development dataset from our proposed dictionary. The performance evaluation of our system is measured using the bilingual evaluation understudy (BLEU) score, a metric that provides a quantitative assessment of transliteration accuracy.

KEYWORDS

Myanmar (Burmese) language, named entity, transliteration, Phrase-based statistical machine translation (PBSMT), language model

1. INTRODUCTION

The task of transliteration, linguistically defined as the transcription of words from a source script to a target script with reliance on approximate phonetic values, is integral to various natural language processing (NLP) applications. The precision in transliterating named entities holds substantial implications for tasks such as machine translation and cross-lingual information retrieval, despite persistent challenges arising from resource scarcity, particularly in understudied languages.

In an innovative approach, this study addresses the Myanmar-English named entity transliteration task by establishing a comprehensive bilingual terminology dictionary. Leveraging phrase-based modelling, we explore both forward and backward transliterations, emphasizing western names.

Our experimental design encompasses character and syllable units, employing N-Gram language models (2-Grams through 6-Grams) on the target side for a thorough investigation of transliteration techniques.

The primary contribution of this research lies in its twofold nature by the manual collection of named entity data and statistical experiments on prepared data.

2. LITERATURE REVIEWS

Despite strides in refining transliteration processes for languages such as Japanese,

Chinese, Korean, and English, a notable gap remains in addressing the unique challenges posed by the Myanmar language [1]. Surveys indicate a scarcity of comprehensive research in this particular domain, underscoring the need for dedicated efforts to enhance transliteration methodologies. The transliteration task, intricately connected to transcription or Romanization, focuses on rendering Myanmar in the Latin alphabet. Operating primarily at the grapheme level, this task represents a simplified translation approach, minimizing reordering operations at the word or phrase level.

In a prior study [2], the Myanmar name Romanization task was addressed using sub-syllable and syllable units with a limited training dataset. Despite achieving statistically favourable results, the LSTM-RNNs approach faced challenges in effectively handling the task. Other pertinent studies in this realm explore grapheme-to-phoneme (G2P) conversion in Myanmar, a task with closer ties to speech processing than traditional natural language processing (NLP). Significantly, [3] underscores the emphasis on relatively large units in utterances, specifically syllables.

In [4], a pragmatic solution is proposed to address translation errors for out-of-vocabulary (OOV) Japanese (Katakana) words translated into Myanmar. This involves the introduction of a rule-based post-editing scheme utilizing PBSMT (Phrase-Based Statistical Machine Translation). The core objective is to rectify errors commonly associated with the translation of Japanese words not included in the standard vocabulary, thereby improving the overall precision and reliability of the translation process.

A historical overview of transliteration reveals the early development of the Phrase-Based Statistical Machine Translation (PBSMT) approach, especially designed for language pairs characterized by extremely low resources. This approach has undergone thorough exploration and has consistently demonstrated robust performance.

The phrase-based methodology, as detailed in [5], excels in capturing word context and exhibiting inherent local reordering within the translation of phrases. Its growing popularity is evident in the domain of statistical machine translation applications.

3.MYANMAR LANGUAGE

As a member of the Sino-Tibetan language family, Myanmar language assumes the role of the official language in the Republic Union of Myanmar and serves as the mother tongue for the Myanmar people, as highlighted in [6]. Employing an abugida system, the Myanmar script is comprised of 33 standalone consonants, four dependent consonants, and an array of diacritic marks representing vowels and tones. The orthography of Myanmar is intricately woven into specialized writing forms, reflecting the nuanced linguistic characteristics of the language.

The fundamental nature of the Myanmar script centres around its syllable-based spelling, facilitating the creation of words with multiple syllables, each conveyed through a series of characters. Figure 1 visually illustrates the intricate syllable structure that defines the Myanmar script. The mandatory initial consonant (C) on the left establishes the groundwork, allowing for the gradual introduction of more characters to accommodate consonant clusters, alternate vowels, syllable codas, and tones. To gain a comprehensive understanding of character categorization, Table 1 serves as an informative guide, presenting the diverse components within the Myanmar script.

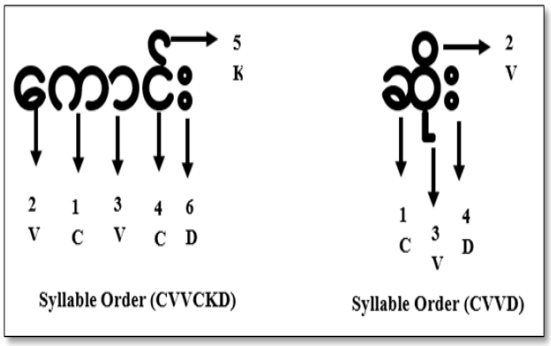


Figure. 1. Myanmar syllable structure

Table. 1 Categories of Characters

Notation	Description
C	ဗျည်း(Consonant) (က-အ)
M	ဗျည်းတွဲ(Medial) (ျြွှ)
V	သရ(Vowel) (ါာိီုူေဲော)
D	ဒိုနိုသရ(Dependent Various Sign) (ံးံ့)
K	အသတ်(Killer) (်)

4. TRANSLITERATION ISSUES

Myanmar faces the linguistic challenge of accommodating Western borrowed words within its script, especially when dealing with proper nouns. The absence of clear guidelines for transliteration emphasizes the importance of addressing pronunciation gaps and maintaining consistency in language practices.

The challenge in modelling transliteration between Myanmar and English is two-fold, with the first aspect rooted in the disparity in the phonological inventories of the two languages. A comprehensive study on the phonology of English loanword adaptation in Myanmar, detailed in [7], provides valuable insights based on a list of 280 borrowed English words. The study outlines distinctive features in the Myanmar consonant system, where stops and affricates showcase a three-way contrast of voiced, voiceless, and voiceless aspirated, coupled with a two-way voicing contrast in nasals and approximants features not entirely reflected in English. Myanmar's seven-vowel system and limitations on syllable codas, confined to nasal and glottal, contribute to the overall complexity. Despite Myanmar's comparatively simpler vowel system, challenges emerge during the transliteration process, notably with the superfluous presence of three tones in Myanmar when transcribing into English. Conversely, Myanmar's restrictions on consonant clusters stand in contrast to the abundance of such clusters in English.

The second challenge in modelling transliteration between Myanmar and English arises from non-phonetic orthographies in both languages. In both English and Myanmar, orthographies lean more towards etymological roots than phonetic principles. The Myanmar script introduces relative redundancy in its phonological inventory, allowing many phonemes to be represented in multiple ways. Additionally, intentional use of special spellings may be incorporated into the Myanmar script for borrowed words to achieve an exotic appearance. Similarly, in English, the irregularities in spelling may lead to transliterations that do not accurately represent the phonetics, as the challenges in transliteration between Myanmar and English are underscored by the uniform representation of certain sounds, regardless of spelling differences. The shared transliteration of <aun> in "auntie" and <un> in "uncle" as (/ʔàɴ/) in Myanmar exemplifies the intricacies introduced by relying on spelling rather than authentic pronunciation.

5. NAMED ENTITY DICTIONARY

The construction of a Named Entity (NE) terminology dictionary is advanced through the incorporation of ALT corpus, UCSY corpus, Myanmar BBC news, and Wikipedia articles. The significance of the ALT corpus and UCSY corpus cannot be overstated, as they house parallel sentences in both Myanmar and English languages, subjected to normalization and tokenization. Noteworthy is the ALT corpus's association with the Asian Language Treebank (ALT) project under ASEAN IVO, featuring a collection of twenty thousand Myan-Eng parallel sentences sourced from Wiki news articles.

Developed by the NLP lab at the University of Computer Studies, Yangon (UCSY), Myanmar, the UCSY corpus serves as a pivotal linguistic resource. Housing 204,000 Myan-Eng parallel sentences sourced from diverse domains, including local news articles and textbooks [8], this corpus is foundational for various natural language processing applications. The extraction of Myan-Eng bilingual Named Entity (NE) transliterated instances is facilitated by the GIZA++ open-source toolkit [9], which establishes raw alignments between the source and target language. Following this extraction, a meticulous manual annotation and categorization process unfolds, covering over 80,000 Myan-Eng bilingual named entities. This comprehensive categorization spans public figures, famous personalities, places, and organizations, with each entity intricately linked to its transliterated Myanmar term corresponding to the English-style named entity found in the aligned coarse sequence pairs. For a more expansive linguistic palette, an additional 50,000 bilingual named entity instance pairs are manually collected from Myanmar BBC news and Wikipedia articles.

Our experimental setup involved the use of the aforementioned bilingual dataset, which was manually subdivided into parallel and monolingual datasets. Training data consisted of 133,255 instances, while 1,000 instances

each were designated for development and testing on our developed dictionary. It is important to emphasize that all named entities followed the standard Unicode format.

Table 2. Data Statistics on Categories

Categories	Number of Named Entities
Person	90,170
Place	30,000
Organization	15,085

Table 3. Data Statistics on Train, Development and Test

Data	Myan		Eng
	Char	Syllable	Char
Train	3,462,716	2,339,636	3,045,278
Dev	25,238	16,116	24,284
Test	23,679	14,874	22,474

5. EXPERIMENTS

5.1. Data Preparation

In efforts to tackle ambiguous problems arising in the transliteration process between Myanmar and English, experiments have been initiated to assess transliteration quality. The focus of these experiments extends to both Myan→Eng and Eng→Myan directions, employing a traditional Phrase-based Statistical Machine Translation model (PBSMT). To gauge the efficacy of transliteration, language models ranging from 2-Grams to 6-Grams have been utilized during the decoding phase of the PBSMT model.

In the experimental process, Myanmar data is segmented into different units, and English data is converted to lowercase, following guidelines suggested by references [10] and [11]. To further enhance segmentation, a custom character unit segmenter is employed, and a syllable segmentation scheme is introduced using a regular expression-based syllable segmenter. Cleaning processes involve utilizing the Moses script clean-n-corpus.perl on pre-processed Myan-Eng monolingual data, targeting the removal of lines containing more than 80 tokens. The systematic approach to data preprocessing and segmentation is crucial for refining the transliteration quality and mitigating challenges posed by the linguistic intricacies of the Myanmar language.

Table 4. Sample Data Format for Myan-Eng (NE) Task

	Unit	Myan-Eng	
Person	Char.	အ ိ ု ဘ ာ ား မ ာ း	O b a m a
	Syl.	အို ဘား မား	O b a m a
Place	Char.	ဆ ွ စ ် ဇ ာ လ န ်	S w i t z e r l a n d
	Syl.	ဆွစ်ဇာလန်	S w i t z e r l a n d
Org.	Char.	ဆ ိ ု န ီ	S o n y
	Syl.	ဆိုနီ	S o n y

5.2. Phrase-Based SMT

The PBSMT model stands as the foundational framework for Statistical Machine Translation (SMT), particularly excelling in the context of low-resource language pairs. Derived from a parallel corpus, the translation model is intricately structured, featuring key components like the reordering model, phrase translation model, and word translation model. In the realm of SMT training, the Moses machine translation toolkit [12] is the open-source tool under the PBSMT paradigm. The approach commences with learning the alignments of phrases between different languages, followed by predicting the optimal composition of phrases for translation with the assistance of a target language model (LM). A noteworthy aspect of the phrase translation model is its reliance on multi-word entries for extracting translation tables, deviating from the conventional use of single-word entries.

In the intricate world of machine translation, the pre-eminence of phrase-based models becomes evident as they consistently outperform word-based models. Anchored in simplicity, a basic phrase-based translation model utilizes phrase-pair probabilities extracted from a corpus and incorporates a straightforward reordering model. The pursuit of the optimal translation, maximizing the

translation probability given the source sentences, is encapsulated in a mathematical expression. This exemplifies the elegance and efficiency of phrase-based approaches in enhancing translation quality.

The mathematical articulation of translation probability plays a pivotal role in the methodology of PBSMT systems, striving to achieve the most accurate and contextually appropriate translations. When engaging in the translation from source language (s) to target language (t), the PBSMT system employs an equation to score the target sentences (t). This mathematical expression encapsulates the intricate interplay of linguistic elements, offering a structured and quantitative means to assess the success of the translation process.

$$\text{argmaxt } P(t/s) = \text{argmaxt } P(s/t)P(t)$$

Simultaneously, the language model (LM) assigns a score P(t) to the target phrase independently. The collaborative use of P(s/t) and P(t) in the translation process enhances the system's ability to produce coherent and semantically accurate translations, bridging the linguistic gap between different languages.

5.3. Language Model

At the core of natural language processing, a statistical language model encapsulates the notion of a probability distribution over sequences of words. The N-Gram model, a prominent instantiation of this concept, simplifies the estimation of sentence probability through a set of equations. By considering the conditional probabilities of words given their n-1 predecessors, the model provides an efficient and effective means to approximate the likelihood of entire sentences, showcasing its utility in language modelling tasks.

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_{1,\ldots} w_{i-1}) \approx \prod_{i=1}^{m} P(w_i | (w_{i-(n-1)}, \ldots, w_{i-1})$$

In the realm of N-Grams language models, the conditional probability of observing a specific word, denoted as w_i, in the context history of the preceding i - 1 words can be efficiently approximated. This approximation involves considering the probability within the shortened context history of the preceding n - 1 words. The practicality of this approach lies in its ability to maintain a reasonable level of accuracy while simplifying the computational complexity associated with modelling language dependencies. The calculation of this probability is facilitated by a following equation.

$$P(w_i | w_{i-(n-1)}, \ldots, w_{i-1}) = \frac{\text{count}(w_{i-(n-1)}, \ldots, w_{i-1}, w_i)}{\text{count}(w_{i-(n-1)}, \ldots, w_{i-1})}$$

5.4 Experimental Setting

Moses was utilized to train the baseline Phrase-Based Statistical Machine Translation (PBSMT) model. Word alignment and phrase tables were trained on tokenized data with GIZA++. The grow-diag-final-and heuristic was applied for symmetric source to target and target to source word alignments. The lexicalized reordering model incorporated the msd-bidirectional-fe option, and language models, including 2-Grams, 3-Grams, 4-Grams, 5-Grams, and 6-Grams, were constructed using the SRILM toolkit. Decoder parameter tuning was carried out through Minimum Error Rate Training (MERT), and decoding was performed using the Moses decoder. The entire workflow operated under default settings in Moses, with each PBSMT experiment requiring 12 hours, and characters taking longer processing time compared to syllables.

6. System Evaluation

The central evaluation criteria in machine transliteration, BLEU scores on character and syllable units, take precedence in [15]. Established as a longstanding tradition in statistical analysis, BLEU scores play a pivotal role in assessing transliteration performance. Our research embarks on a series of experiments, systematically investigating the impact of segmentation level on transliteration performance and discerning models that excel in generating the closest references across diverse linguistic features. The assessment focuses on the comparative analysis of English outputs from Myan→Eng transliteration systems against reference entities and Myanmar outputs from Eng→Myan against reference entities, considering various linguistic features and both character and syllable units. This comprehensive overview sheds light on the paramount results for each unit within our meticulously selected monolingual named entity test sets.

The results of experiments for English-to-Myanmar and reversed Myanmar-to-English transliteration are detailed in Tables 5 and 6. The evaluation involves different segmentation units, including 2-Grams, 3-Grams, 4-Grams, 5-Grams, and 6-Gram language models, with a specific emphasis on target-side language models.

The BLEU score, ranging from 0 to 1, serves as the assessment metric, where a higher score indicates a closer match between the hypothesis transliteration and the reference. The mathematical representation of the BLEU score is outlined in the following equation.

$$\mathbf{BLEU} = \mathbf{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

In the formulation, n denotes the orders of N-Grams considered for p_n, and w_n represents the uniformly distributed weights assigned to the N-Grams precisions. To address the potential of high precision translation, a Brevity Penalty (BP) is introduced. BP is set to 1 when the hypothesis transliterations exceed 1, and it is calculated as $e^{(1-r/c)}$ when the hypothesis transliterations are less than or equal to the reference transliterations.

The precision p_n for N-Grams is computed by summing the N-Grams match for every hypothesis result in the test corpus. Through our experiments, comparing the highest priority score across all PBSMT models, it is evident that PBSMT with 5-Grams and 6-Gram language models on the syllable segmentation scheme exhibits greater vigor than other transliteration models for Myan-Eng.

A performance analysis, considering linguistic features for the Myan→Eng task, reveals that the syllable unit surpasses the character unit in transliterating named entities.

Syllable structure proves to be a powerful unit in the preprocessing landscape of Myanmar's natural language processing (NLP) applications. Noteworthy achievements in machine translation (MT) [8], G2P conversion tasks [16], and Myanmar named entity recognition (NER) [17] attribute their success to the effective utilization of syllable units. This paradigm shift in focus towards syllabic representations underscores their significance in optimizing NLP workflows.

Within the context of English to Myanmar transliteration, the PBSMT models incorporating 4-Grams and 6-Grams stand out for their exceptional performance when tackling character unit-based tasks. These models, characterized by their proficiency in capturing the intricate phonetic details at the character level, present a promising solution for transliteration challenges.

However, the narrative takes a surprising twist as we transition to syllable-based tasks, revealing a sharp decline in performance. This unexpected falloff suggests a substantial gap in the models' ability to navigate the intricate phonetic mapping and one-to-one alignment required for successful transliteration at the syllabic level.

This discrepancy underscores the need for nuanced approaches to address the unique challenges posed by syllabic transliteration in the English to Myanmar context.

Table 5. Evaluation Result in Terms of BLEU Score for Syllable Unit

System	Syllable Unit (Myan-Eng)			
	BLEU-1	*BLEU-2*	*BLEU-3*	*BLEU-4*
PBSMT(2-Grams)	87.9	78.7	72.1	66.9
PBSMT(3-Grams)	88.2	79.0	72.5	67.3
PBSMT(4-Grams)	88.4	79.7	73.5	68.5
PBSMT(5-Grams)	**89.3**	81.1	75.1	70.4
PBSMT(6-Grams)	**89.3**	80.8	75.0	70.5
System	Syllable Unit (Eng-Myan)			
	BLEU-1	*BLEU-2*	*BLEU-3*	*BLEU-4*
PBSMT(2-Grams)	67.1	54.7	45.2	37.5
PBSMT(3-Grams)	67.4	54.9	45.9	38.1
PBSMT(4-Grams)	67.4	55.2	46.3	39.4
PBSMT(5-Grams)	69.3	56.9	48.1	40.6
PBSMT(6-Grams)	67.7	55.8	47.2	39.3

Table 6. Evaluation Result in Terms of BLEU Score for Character Unit

System	Character Unit (Myan-Eng)			
	BLEU-1	*BLEU-2*	*BLEU-3*	*BLEU-4*
PBSMT(2-Grams)	84.7	70.2	61.6	54.0
PBSMT(3-Grams)	86.2	72.0	63.2	56.3
PBSMT(4-Grams)	84.9	71.8	63.2	56.0
PBSMT(5-Grams)	86.2	74.0	66.4	60.6
PBSMT(6-Grams)	85.8	73.3	65.6	59.8
System	**Character Unit (Eng-Myan)**			
	BLEU-1	*BLEU-2*	*BLEU-3*	*BLEU-4*
PBSMT(2-Grams)	81.2	64.5	55.3	48.2
PBSMT(3-Grams)	81.5	66.1	57.6	51.0
PBSMT(4-Grams)	**82.0**	67.7	59.9	53.9
PBSMT(5-Grams)	81.6	67.8	60.5	54.9
PBSMT(6-Grams)	**82.0**	68.4	61.5	56.0

6. CONCLUSIONS

In this research, a comprehensive analysis of western (non-native) named entity transliteration tasks was presented, encompassing a systematic comparison of the PBSMT paradigm with various N-Gram's language models. Given the scarcity of the Myanmar NE corpus, our proposed Myan-Eng (NE) terminology dictionary was introduced and utilized to train all PBSMT models. Although not vast in scale, this mapping dictionary yielded satisfactory performance for both transliteration tasks, achieved without the incorporation of handcrafted features or additional resources. It is our belief that, with an expanded dataset and more experimental iterations, the field holds promise for yielding even more robust results. Notably, this study represents the pioneering effort in statistical transliteration exploration within the context of the Myanmar language. Future endeavors involve the augmentation of Myan-Eng named entity instances with native Myanmar names, coupled with extensive testing comparing neural models to further enhance the accuracy of transliterations in the near future.

REFERENCES

[1] Merhav, Y., & Ash, S. (2018, August). Design challenges in named entity transliteration. In Proceedings of the 27th international conference on computational linguistics (pp. 630-640).

[2] Ding, C., Pa, W. P., Utiyama, M., & Sumita, E. (2017, August). Burmese (Myanmar) name romanization: A sub-syllabic segmentation scheme for statistical solutions. In International Conference of the Pacific Association for Computational Linguistics (pp. 191-202). Springer, Singapore.

[3] Thu, Y.K., Pa, W.P., Finch, A., Ni, J., Sumita, E., Hori, C.: The application of phrase based statistical machine translation techniques to Myanmar grapheme to phoneme conversion. In: Hasida, K., Purwarianti, A. (eds.) Computational Linguistics. CCIS, vol. 593, pp. 238–250. Springer, Singapore (2016). https://doi.org/ 10.1007/978-981-10-0515-2 17

[4] Naing, H. M. S., Thu, Y. K., Pa, W. P., Kato, H., Finch, A., Sumita, E., & Hori, C. (2015). Rule Based Katakana to Myanmar Transliteration for Post-editing Machine Translation. Proceedings of the Annual Conference of the Language Processing Society of Japan, pages 257-260.

[5] Koehn, P., Och, F. J., & Marcu, D. (2003, May). Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 48-54). Association for Computational Linguistics.

[6] Sin, Y.M., & Soe, K.M. (2019). Attention-Based Syllable Level Neural Machine Translation System for Myanmar to English Language Pair. International Journal on Natural Language Computing (IJNLC) Vol.8, No.2.

[7] Chang, C. 2003). ―High-interest loans‖: The phonology of English loanword adaptation in Burmese.

[8] ShweSin, Y. M., Soe, K. M., & Htwe, K. Y. (2018, October). Large Scale Myanmar to English Neural Machine Translation System. In 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE) (pp. 464-465). IEEE.

[9] Och, F. J., and Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1), 19-51.

[10] Li, Z., Chng, E. S., & Li, H. (2017, December). Named entity transliteration with sequence-to-sequence neural network. In 2017 International Conference on Asian Language Processing (IALP) (pp. 374-378). IEEE.

[11] Liu, L., Finch, A., Utiyama, M., & Sumita, E. (2016, March). Agreement on target-bidirectional LSTMs for sequence-to-sequence learning. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16). AAAI Press 2630-2637.

[12] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Fed-erico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open-source toolkit for statistical machine translation. In Proceedings of the 45th annual

meeting of the ACL on interactive poster and demonstration sessions, pages 177–180. Association for Computational Linguistics.

[13] Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In Seventh international conference on spoken language processing.

[14] Och, F. J. (2003, July). Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 160-167). Association for Computational Linguistics.

[15] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proc. Of ACL, pp. 311—318.

[16] Thu, Y. K., Pa, W. P., Sagisaka, Y., & Iwahashi, N. (2016, December). Comparison of grapheme-to-phoneme conversion methods on a myanmar pronunciation dictionary. In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016) (pp. 11-22).

[17] Mo, H. M., & Soe, K. M. (2019). Syllable-based Neural Named Entity Recognition for Myanmar Language. arXiv preprint arXiv:1903.04739.

Authors Biography

Aye Myat Mon, Lecturer at Faculty of Computer Science, University of Computer Studies (Mandalay). She received M.C.Sc. (Credit) at University of Computer Studies (Pathein) in 2013.

Mya Than Hnin, Associated Professor at Faculty of Computer Science, University of Computer Studies (Mandalay). She received M.I.Sc, degree at University of Computer Studies,Yangon,(CAST) in 2001.

Su Su Win is a Lecturer of the Faculty of Computer Science at University of Computer Studies (Mandalay). She received B.C.Sc (Bachelor of Computer Science) degree in 2007, B.C.Sc (Hons.) degree in 2008 and M.C.Sc (Master of Computer Science) degree in 2010 at University of Computer Studies (Loikaw). She has published papers in University Journals. Her research interests include Natural Language Processing, Artificial Intelligence and Machine Translation.

PREDICTING USERS' INTEREST LEVEL IN PERSONALIZED NEWS RECOMMENDATIONS THROUGH IMPLICIT BEHAVIOR ANALYSIS

Hein Naing Soe[1], Thein Than Htay[2], La Min Htut[3], and Hein Tun[4]

[1,2,3,4] Department of Computer Science, Higher Education Centre (Pyin Oo Lwin)

[1]heinnaingsoe9354@gmail.com, [2]maharmyanmar1966@gmail.com
[3]laminhtut@gmail.com, and [4]heintun.ctdept@gmail.com

ABSTRACT

Achieving user engagement is essential for the success of online news platforms and personalized recommendation systems. This paper presents an implicit behavior-based users' interest level prediction method aimed at improving the personalized news recommendations. It addresses the challenge of understanding user interests without explicit feedback by analyzing implicit behaviors such as time spent, mouse scrolls, movements, and the extent of article exploration during user interactions. These behaviors are tracked and collected using JavaScript and classified using K-means clustering, establishing baseline metrics for normalization. Interest level prediction is accomplished through linear regression. Thorough experiments on an actual dataset of active users validate the effectiveness of the interest level prediction model, highlighting its accuracy in revealing user interests through implicit behaviors. This approach provides valuable insights into user preferences, advancing the development of a more tailored news recommendation system.

KEYWORDS

Implicit Behavior, Interest Level Prediction, News Recommendation, Personalized Recommendation System, User Engagement

1. INTRODUCTION

The success of online news platforms and personalized recommendation systems hinges on robust user engagement, requiring a deep understanding of user interests and preferences. This study introduces an innovative Implicit Behavior-Based Interest Level Prediction approach to tackle the challenge of identifying user interests in news articles without relying on explicit feedback. Traditional methods often fall short by focusing solely on explicit interactions, potentially missing out on a broader spectrum of user engagement.

The motivation behind this study stems from the vital role user engagement plays in online news platforms and personalized recommendations. Aiming for effective content recommendations and enhanced user experiences, the study seeks to develop an approach that accurately captures users' interests while reading news articles, ultimately paving the way for a more personalized news recommendation system.

The study's primary aim is to create an Implicit Behavior-Based Interest Level Prediction approach that tracks and analyzes implicit behaviors, such as time spent, mouse scrolls, movements, and the portion of the news page explored. This non-intrusive method aims to determine users' engagement levels and interests during their interaction with news content.

The contribution of this study lies in introducing a new method that leverages implicit behaviors to infer user interests in news articles. By employing K-Means clustering and linear regression, the study classifies interest events, establishes baseline metrics, and predicts interest levels for each user. This approach provides valuable insights into user preferences,

fostering the development of a more personalized news recommendation system.

In pursuit of its aim, this study aims to enhance user experiences in news consumption and contribute to the advancement of personalized recommendation systems in the digital era.

2. RELATED WORK

The study of implicit behaviors and user engagement in the context of digital news consumption has gained considerable attention from researchers and practitioners alike. Over the years, numerous studies have explored various aspects of reader behavior and its relation to news articles' content and presentation. In this section, a brief overview of the key works and research efforts that have contributed to the understanding of related topics is presented.

Lagun et al. (2016) have proposed four engagement metrics in their study, accurately reflecting how users engage with and attend to online news content, providing clear characterizations of user engagement. The proposed probabilistic model, TUNE, effectively predicts future user engagement levels based solely on textual content, outperforming existing methods. Despite engagement capture at the sub-document level and proportion of article read, challenges may still be faced in capturing user interests at a more granular level [1].

The research of Soumy Jacob et al. (2018) effectively detects the interest of readers in newspaper articles through eye tracking measures, achieving an accuracy of 44% in classifying interests and 62% when considering subjective comprehension by analyzing eye movements, fixations, saccades, blinks, and pupil diameters. However, there is a suggested scope for further improvement in accuracy achieved through eye movements, especially in scenarios with diverse readers and reading behaviors [5].

Klaas Nelissen et al. (2018) highlight the usefulness of swipe interactions on tablets as indicators of user engagement with news articles, providing valuable insights into user engagement on tablet platforms through the combination of time-based aspects and swipe interactions. Nevertheless, full capture of user engagement through swipe interactions may not be achieved, necessitating further research to refine prediction accuracy [4].

The study of Fattane Zarrinkalam et al. (2020) covers effective user interest mining from social media platforms, encompassing various representation models, evaluation methodologies, and applications, providing insights into understanding users' preferences and enhancing user engagement. Yet, challenges in accurately inferring interests from the vast and diverse content shared by users on social media might be faced, despite the richness of available data [2].

The research of Kholoud Khalil Aldous et al. (2023) characterizes user engagement across different news outlets and social media platforms, considering sentiments and topics, with the prediction model of the study achieving a high average F1-score of 83% for predicting user engagement levels based on language and metadata features. Yet, variations in engagement behaviors across different contexts and platforms may exist, suggesting the exploration of additional factors impacting user engagement [3].

Overall, these articles contribute valuable insights into understanding and predicting user engagement with online news content, emphasizing the importance of personalized and effective news recommendations. However, each study may encounter specific challenges and limitations, warranting further exploration in the context of personalized news recommendation systems.

3. METHODOLOGY

This study employs a methodology designed to predict and analyze user interest in news articles based on their implicit behaviors. The methodology comprises four key steps, as illustrated in Figure 1.

By following this approach, we effectively capture users' interests through implicit behavior analysis. This non-intrusive and efficient method enhances our understanding of user engagement, ultimately contributing

to the development of a personalized news recommendation system.

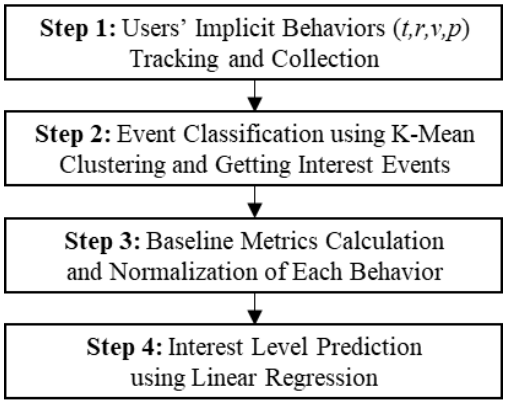


Figure 1. Methodology

3.1. Users' Implicit Behavior Tracking and Collection

The initial step involves tracking user activity and gathering implicit behaviors related to news articles, which encompass monitoring time spent, mouse scrolls, mouse movements, and the percentage of the news webpage length reached. These behaviors are measured and recorded during user reading sessions using JavaScript tracking code. JavaScript was chosen as the tracking tool due to its widespread use as a client-side scripting language, making it suitable for capturing user interactions on web pages.

We specifically selected four features: time spent per word (t), number of mouse scrolls per word (r), number of mouse movements per word (v), and the reached percentage on the news webpage length (p). These features were chosen because they hold significant value in reflecting user engagement with news articles. The behaviors, excluding (p), are divided by the number of words, as the length of the news page influences the number of each behavior and ensures uniqueness.

These features enable comprehensive measurement of user implicit behaviors, providing valuable insights into their interests and preferences while reading news content. The hypotheses for the selected four features are as follows:

- **Time Spent per Word (t):** Longer time spent reading news articles indicates a higher level of interest in the content.
- **Number of Mouse Scrolls per Word (r):** More mouse scrolls signify active engagement and a deeper interest in the news piece.
- **Number of Mouse Movements per Word (v):** Increased mouse movements indicate a more focused and interactive reading experience.
- **Reached Percentage on News Webpage Length (p):** A higher percentage implies a stronger interest in the content.

By tracking and collecting these four essential behaviors, we can effectively gauge user interest levels during their interactions with news articles. JavaScript provides us with detailed and real-time data, offering valuable insights into user engagement patterns. These selected features serve as robust indicators of user interest and form the basis for the subsequent steps of the interest level prediction model.

3.2. Event Classification using K-Mean Clustering and Getting Interest Events

In this phase, K-Means clustering is utilized to categorize user sessions into distinct events. This choice is based on the algorithm's capacity to discern data patterns autonomously, without the requirement for pre-established labels. The K-Means clustering algorithm can be represented mathematically as:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

Where, J represents the objective function, the sum of squared distances, K is the number of clusters, C_i is the i^{th}cluster, x is a data point in the dataset, and μ_i is the centroid (mean) of the i^{th} cluster.

The algorithm proceeds iteratively to minimize J by adjusting the assignment of data points to clusters and updating the centroids [6]. The steps involve:

- **Initialization:** Choose K initial centroids (often randomly or using a heuristic).
- **Assignment:** Assign each data point to the nearest centroid, forming K clusters.

- **Update Centroids:** Recalculate the centroids of each cluster based on the current assignments.
- **Repeat:** Steps 2 and 3 are repeated until convergence (when centroids no longer change significantly) or a predefined number of iterations is reached.

In this paper, we choose to set K to 2, as we only need to determine two clusters: the interest and disinterest clusters. From 500 news reading events tracked with four features (t, r, v, and p), we obtain two centroids, (t=0.6813, r=0.4856, v=25.3623, and p=0.99) and (t=0.0254, r=0.2921, v=2.3321, and p=0.2312), for cluster1 and cluster2, respectively. To determine which cluster represents the interest cluster, we select a sample of interest event (t=0.8, r=0.5, v=25, and p=1) and calculate the distances of the two centroids from the sample. The cluster whose centroid is closest to the sample is designated as the interest cluster. Consequently, in this paper, cluster1 is designated as the interest cluster and it includes **326 events**. The events within cluster1 form the basis for establishing baseline metrics in the subsequent step. By utilizing this sample of interest events, we ensure that the metrics derived from them accurately represent users' genuine interests in the news articles.

3.3. Baseline Metrics Calculation and Normalization of Each Behavior

Baseline metrics are derived by calculating the mean values for observed behaviors during interest events. Outliers within each behavior are systematically excluded to ensure the robustness of these metrics. Particularly, the percentage of content reached (p) serves as an inherent indicator of users' interest, representing the portion of the news webpage they read during their sessions. These mean values constitute essential baseline metrics, serving as the basis for subsequent normalization through linear regression. The formula for computing the sample mean of a set of values can be expressed as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2)$$

Where, x_i is an individual value from the dataset and n is the total number of values in the sample.

Calculating outliers is as follows:

$$Lower\ Bound = Q1 - (1.5 \times IQR) \quad (3)$$

$$Upper\ Bound = Q3 + (1.5 \times IQR) \quad (4)$$

$$1^{st}\ Quartile(Q1) = \frac{1}{4}(n+1)^{th}\ term \quad (5)$$

$$3^{rd}\ Quartile(Q3) = \frac{3}{4}(n+1)^{th}\ term \quad (6)$$

$$Interquatile\ Range(IQR) = Q3 - Q1 \quad (7)$$

Establishing baseline metrics is essential for normalizing behavior rates per word within interest events, ensuring stable and representative metrics. In this study, numbers of outliers for t, r, and v are **12, 8,** and **25**, respectively. So, baseline metrics of each feature are calculated without outliers as follows:

$$\bar{t} = \frac{(0.65 + 0.44 + \cdots + 0.47)}{314} = \mathbf{0.6376}$$

$$\bar{r} = \frac{(0.21 + 0.33 + \cdots + 0.4)}{318} = \mathbf{0.2813}$$

$$\bar{v} = \frac{(26.4 + 21.8 + \cdots + 22.6)}{301} = \mathbf{23.1696}$$

These robust baseline metrics are crucial for dividing each behavior by its mean during subsequent behavior normalization. This step is vital for ensuring data comparability and scaling down behavior rates per word, allowing for direct comparisons across user engagements and article lengths. When the behavior rate per word divided by the baseline metric exceeds 1, it is set to 1 as follows:

$$t' = \frac{t}{\bar{t}} \begin{cases} If\ t' > 1, t' = 1 \\ Otherwise, t' = t' \end{cases} \quad (8)$$

$$r' = \frac{r}{\bar{r}} \begin{cases} If\ r' > 1, r' = 1 \\ Otherwise, r' = r' \end{cases} \quad (9)$$

$$v' = \frac{v}{\bar{v}} \begin{cases} If\ v' > 1, v' = 1 \\ Otherwise, v' = v' \end{cases} \quad (10)$$

Where, $t', r', and\ v'$ are normalized behavior rates per word and $t, r, and\ v$ are input behavior rates per word.

Example of normalization is described with an interest event (t=0.6185, r=0.32, v=22.01, p=1) as follows:

$$t' = \frac{0.6185}{0.6376} = \mathbf{0.97}, \qquad r' = \frac{0.32}{0.2813} = \mathbf{1}$$

$$v' = \frac{22.01}{23.1696} = \mathbf{0.95}$$

This normalization not only standardizes the data but also reduces the impact of extreme values on the linear regression model, thereby improving the accuracy of capturing user interest level patterns. The baseline metrics serve as a fundamental component of the interest prediction model, facilitating behavior normalization for the linear regression process.

3.4. Interest Level Prediction using Linear Regression

We utilize Linear Regression for interest level prediction because it is well-suited for modeling the relationship between independent variables (behavior rates per word) and the dependent variable (interest level). The linear regression formula [7] is represented as follow:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon \quad (11)$$

Where, y_i is the dependent or predicted variable, β_0 is the intercept of y_i, β_p is slope for each independent variable, and ϵ is the random error of the model.

The workflow is as follows:

- **Labeling Events:** Sixteen distinct scenarios are generated by combining "enough" (denoted as 1) and "not enough" (denoted as 0) responses for each of the four metrics. A survey is conducted with 20 individuals who regularly read articles on PCs to gauge their interest levels, which range from 0% to 100%, across these 16 scenarios. The average interest level, excluding outliers, is used as the label, as presented in Table 1.

Table 1. Label Dataset to Train Linear Regression Model

Scenario	t	r	v	p	Average Interest Level
1	1	1	1	1	**1.0000**
2	1	1	1	0	**0.6273**
3	1	1	0	1	**0.8000**
4	1	1	0	0	**0.4111**
5	1	0	1	1	**0.7545**
6	1	0	1	0	**0.4727**
7	1	0	0	1	**0.5818**
8	1	0	0	0	**0.1455**
9	0	1	1	1	**0.0000**
10	0	1	1	0	**0.0000**
11	0	1	0	1	**0.0000**
12	0	1	0	0	**0.0000**
13	0	0	1	1	**0.0000**
14	0	0	1	0	**0.0000**
15	0	0	0	1	**0.0000**
16	0	0	0	0	**0.0000**

- **Model Training:** The linear regression model is trained using the normalized behavior rates per word as input features shown in Table 1.
- **Learning Patterns:** The model learns the patterns in users' behavior rates that correspond to their interest levels in news articles.
- **Coefficient Estimation:** The model estimates the coefficients of the input features to quantify their impact on predicting the interest level.
- **Creating a Regression Equation**: To establish a linear regression equation that signifies the connection between behavior rates and interest levels, a normalization process is utilized on all weights and intercept. This normalization ensures that their sum results in a value between 0 and 1. The resulting equation is as:

$$L = -0.25 + 0.75t' + 0.13r' + 0.14v' + 0.23p \quad (12)$$

Here, L denotes Interest Level and t', r', v' are normalized values of behavior rates per word (t, r, v).

- **Predicting Interest Level**: Using the trained regression model, interest levels are predicted for new user sessions across 16 scenarios by normalizing their input behavior rates. The resulting levels align reasonably with their respective scenarios. In these sessions, two distinct events are observed: one for interest $(-0.25 + 0.75 \times 0.97 + 0.13 \times 1 + 0.14 \times 0.95 + 0.23 \times 1 = \mathbf{0.9705})$ and the other for disinterest $(-0.25 + 0.75 \times 0.08 + 0.13 \times 1 + 0.14 \times 0.1 + 0.23 \times 1 = \mathbf{0.1840})$.

By seamlessly integrating this data-driven, non-intrusive approach into news websites, we can predict users' interest levels while reading news articles. This enables the delivery of personalized content recommendations, ultimately enhancing

user engagement and the overall user experience.

4. MODEL EVALUATION

This section discusses the model evaluation. Notably, direct assessment of the model using users' actual interest levels is infeasible due to the unavailability of such data. To overcome this challenge, we gather user feedback through interest and disinterest buttons placed on the news page, serving as surrogates for explicit feedback. Subsequently, the model predicts the user's interest, classifying it as interest when the predicted value approaches 1, and as disinterest otherwise.

For a quantitative evaluation of the model's accuracy, we employ the Confusion Matrix as shown in Table 2, offering a comprehensive insight into its performance. This matrix allows us to analyze true positives, true negatives, false positives, and false negatives, providing a clear assessment of the model's ability to correctly categorize interest and disinterest events.

Table 2. Confusion Matrix

	Actual Positive (AP)	**Actual Negative (AN)**	**Total**
Predicted Positive (PP)	TP	FP	**TP+FP**
Predicted Negative (PN)	FN	TN	**FN+TN**
Total	**TP+FN**	**FP+TN**	**TP+FP+ FN+TN**

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$Negative\ Predictive\ Value(NPV) = \frac{TN}{TN+FN} \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$Specificity = \frac{TN}{TN+FP} \quad (12)$$

$$F1\ Score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (13)$$

Thoroughly assessing the model with the previously mentioned methods aims to ensure its effectiveness and reliability in predicting user interest levels when they engage with news articles. The insights gained from this evaluation will enhance personalized news recommendation systems, ultimately improving user experiences and engagement on news platforms.

4.1. Results and Analysis

The model evaluation results, derived from 1000 news reading events with 40 individuals, offer valuable performance insights. Table 3 displays the obtained confusion matrix based on these events.

Table 3. Confusion Matrix based on 1000 Events of 40 Individuals

	Actual Interest	**Actual Disinterest**	**Total**
Predicted Interest	812	12	**824**
Predicted Disinterest	15	161	**176**
Total	**827**	**173**	**1000**

Based on the obtained confusion matrix, the results of model are captured as shown in Table 4.

Table 4. Results of Model

Overall Accuracy	**0.973**
Precision	0.9854
NPV	0.9147
Recall	0.9819
Specificity	0.9306
F1 Score	**0.9836**

The Interest Level Prediction Model demonstrates strong accuracy, as indicated by an impressive average F1 score of 0.9836. This score reflects a high level of accuracy and effectiveness, encompassing both precision and recall, making it a robust measure for evaluating the model's performance.

The analysis of this study sheds light on the significance of the data-driven approach. By leveraging data analytics, the model efficiently predicts user interest levels, laying the groundwork for tailored content recommendations. This personalized

approach enhances user engagement and satisfaction, underscoring the value of understanding user preferences in the digital news landscape.

Moreover, the detailed examination of the confusion matrix provides deeper insights into the model's performance. By identifying patterns in user interactions, the model can accurately distinguish between interest and disinterest, contributing to its high precision and recall rates.

In summary, the outstanding average F1 score of 0.9836 and the comprehensive Confusion Matrix analysis underscore the success of the Interest Level Prediction Model. This model's capacity to discern user interests from implicit behaviors represents a significant advancement in personalized news recommendation systems, transforming the user engagement experience with digital news content.

4.2. Discussion

The Implicit Behavior-Based Interest Level Prediction Model has demonstrated significant potential for improving personalized news recommendation systems. Notably, it achieved an outstanding average F1 score of 0.9836, particularly on PCs, excluding mobile devices. This remarkable performance underscores the model's ability to accurately predict user interest levels based on implicit behaviors during news article interactions on larger screens. The model exhibits high precision and recall rates, as confirmed by Confusion Matrix analysis, indicating its robustness in classifying interest and disinterest events, thereby reducing misclassifications and increasing user engagement.

The model's success on PCs can be attributed to the extensive user interactions and behaviors typically observed on these platforms. The immersive reading experience provided by PCs, characterized by prolonged article engagement, extensive mouse scrolling, and more thorough exploration of news webpages, greatly contributes to the precise identification of user interests. This emphasis on PC platforms optimizes interactions and enhances user satisfaction.

5. CONCLUSIONS

In conclusion, this study introduces an innovative Implicit Behavior-Based Interest Level Prediction Model designed for predicting user interest levels while reading news articles on PCs. The model demonstrates its effectiveness in capturing implicit behaviors and accurately categorizing interest levels, as evidenced by impressive F1 scores and Confusion Matrix analysis. By utilizing these insights, the model enables news websites to deliver tailored content recommendations on PCs, thereby enhancing user satisfaction and engagement.

Looking ahead, we anticipate the seamless integration of the model into various online news platforms, promoting personalized news recommendations on PCs and fostering increased user engagement and satisfaction. As we delve further into advancements in user behavior analysis and recommendation systems, the model serves as a solid foundation for future research, holding the potential to revolutionize the news consumption experience on PCs and shape the future of digital news platforms.

REFERENCES

[1] Dmitry Lagun and Mounia Lalmas, “Understanding and Measuring User Engagement and Attention in Online News Reading”, WSDM, San Francisco, USA, February 2016.

[2] Fattane Zarrinkalam, Guangyuan Piao, Stefano Faralli, and Ebrahim Bagheri, “Mining User Interests from Social Media”, Proceedings of the 29th ACM International Conference on Information and Knowledge Management, October 2020.

[3] Kholoud Khalil Aldous, Jisun Anb, and Bernard J. Jansen, “What Really Matters? Characterising and Predicting User Engagement of News Postings using Multiple Platforms, Sentiments and Topics”, Behaviour & Information Technology, 2023.

[4] Klaas Nelissen, Monique Snoeck, Seppe vanden Broucke, and Bart Baesens, “Swipe and

Tell: Using Implicit Feedback to Predict User Engagement on Tablets", ACM Trans. Inf. Syst. 36, 4, Article 35, June 2018.

[5] Soumy Jacob, Shoya Ishimaru, Syed Saqib Bukhari, and Andreas Dengel, "Gaze-based Interest Detection on Newspaper Articles", 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction, Warsaw, Poland, June 2018.

[6] Imad Dabbura, "K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks", https://towardsdatascience.com/, 2018.

[7] Sebastian Taylor, "Multiple Linear Regression", https://corporatefinanceinstitute.com/, 2020.

Authors

Biography

Hein Naing Soe received M.C.Tech degree in Computer Technology from Higher Education Center (HEC), DSA, in 2018. He is now attending the Ph.D. 18th Batch at the HEC, Pyin Oo Lwin. Department of Computer Technology, HEC.

Dr. Thein Than Htay earned his M.Tech (Applied Maths). from MAIT, Russia in 2007 and his Ph.D. from DSA in 2013. Now he is serving as an associate professor of the Department of Computer Technology in DSA.

Dr. La Min Htut earned M.E.Tech. from Saint Petersburg State Marine Technical University in 2008 and his Ph.D. in 2012. Now he is serving as an assistant lecturer of the Department of Computer Technology in DSA.

Dr. Hein Tun earned M.E.Tech. from Saint Petersburg State Marine Technical University in 2008 and his Ph.D. in 2011. Now he is serving as an assistant lecturer of the Department of Computer Technology in DSA.

PRACTICAL APPROACHES OF SOLVING SECOND-ORDER DIFFERENTIAL EQUATIONS WITH DAMPED VIBRATIONS USING MATLAB

Yee Yee Htun[1], and Htay Lin Aung[2]
[1,2]Department of Engineering Mathematics, Technological University (Meiktila)
[1]yeeyeehtun2016@gmail.com,[2]uhtaylinaung48@gmail.com

ABSTRACT

The Damping conditions of over damped, critical damped and under damped are to be studied with second-order differential equations' practice. The solving methods of these second-order differential equations (DE) are specifically to be carried out with Damped Vibrations. The main theme of this paper is to be successfully carried out the practical approaches of solving the second order differential equations in Science or Engineering fields. Some examples cases are practiced and used the software tool results of MATLAB. Functions and Plotting algorithms of MATLAB2015a version is enforced to be simply solved the Second Order DE with Damped Vibrations in this research paper. By the contributions of Example Cases for Damping Homogeneous System, the applications or practical approaches of Second Order ODE are carried out their solutions and the author hope to support the practical applications in Science or Engineering are successfully performed by the effective approaches of Mathematics Concerned.

KEYWORDS

Second order DE, Damping, Functions, Plotting Algorithms, MATLAB

1. INTRODUCTION

The second-order ordinary differential equations (ODEs) are very useful and effective in the study of differential equations for the main reasons of that the linear equations have a rich theoretical structure, establishes a number of systematic methods of solution and they are vital to any serious investigation of the classical areas of mathematical physics. Without knowing the solutions of second order linear differential equations, it cannot be developed in the fields of Mechanical Engineering: fluid mechanics, heat conduction, wave motion, or electromagnetic. [1][4][5]

When finding to solve the second order differential equations, it can be represented as first order differential equations which are in the form of two dimensions. After changing the variables, a theorem on existence and uniqueness of these solutions can then be be denoted as: $\frac{d^2x}{dt^2}$ to be considered.

A second-order ODE is called linear if it can be written

$$x'' + p(t)x' + q(t)x = r(t) \qquad (1)$$

and nonlinear if it cannot be written in this form. This equation is linear in x and its derivatives, whereas the functions p, q, and r may be any given functions of t.

If r(x) = 0 then the equation

$$x'' + p(t)x' + q(t)x = 0 \qquad (2)$$

and is called homogeneous, otherwise nonhomogeneous. [1][4][5]

In MATLAB, it can be used the function called 'dsolve' that means Symbolic solution of ordinary differential equations. The function 'dsolve (eqn1, eqn2, ...)' accepts symbolic equations representing ordinary differential equations and initial conditions. By default, the independent variable is 't'. The independent variable may be changed from 't' to some other symbolic variable by including that variable as the last input argument. Initial conditions involving derivatives must use an intermediate variable. If expressed as strings several equations or initial conditions may be grouped together, separated by commas, in a single input argument. [7]

Three different types of output are possible. For one equation and one output, the resulting solution is returned, with multiple solutions to a nonlinear equation in a symbolic vector. For several equations and an equal number of outputs, the results are sorted in lexicographic order and assigned to the outputs. For several equations and a single output, a structure containing the solutions is returned. [7]

2. DAMPING HOMOGENEOUS SYSTEM

Homogeneous differential equation in equation (3) is to be considered for some m, c and k which are positive real numbers. The solutions of this equation are classified in equation (4). The values are to be solved according to the roots of its characteristic equation:

$$mx'' + cx' + kx = 0 \qquad (3)$$

$$m\lambda^2 + c\lambda + k = 0 \qquad (4)$$

Accordingly, for some constants c_1, c_2 we have the solutions as in equation (5). These factors given by three cases: (i) the roots are real and distinct, (ii) the roots are real and equal, (iii) the roots are complex conjugate.

$$\lambda_1 \neq \lambda_2 \rightarrow \quad x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

$$\lambda_1 = \lambda_2 \rightarrow \quad x(t) = c_1 e^{\lambda_1 t} + c_2 t e^{\lambda_2 t}$$

$$\lambda = a \pm bi \rightarrow x(t) = e^{at}(c_1 \cos bt + c_2 \sin bt) \qquad (5)$$

For the classical problem of a fixed spring, to which an object of mass m is attached, using Hooke's law, which can be expressed as: $F_k = -kx$, if we consider there is friction or damping involved, it is assumed that the force due to friction is proportional to the velocity with which the object is moving, that is, $F_c = -cx'$, where c is called the damping constant. Now, Newton's law: F = ma, with the total force $F = F_c + F_k$ and gives:

$-cx' - kx = mx''$ or $mx'' + cx' + kx = 0$. Again, the solutions to this equation will be classified according to the roots of the characteristic equation. The roots are:

$$\lambda = \frac{-c \pm \sqrt{c^2 - 4mk}}{2m} \qquad (6)$$

Following (6), it can be said that the spring system represents three cases depending on the sign of the expression under the square root:

(i) $c^2 < 4mk$ (this will be under damping and c is small relative to m and k)

(ii) $c^2 = 4mk$ (this will be critical damping and c is just between over and under damping).

(iii) $c^2 > 4mk$ (this will be over damping and c is large relative to m and k). [2][3][6]

Mathematically, the easiest case is over damping because the roots are real. However, most people think of the oscillatory behaviour of a damped oscillator. Since this is connected to under damping, it is to be starting with that case. [1][2][3][6]

2.1. Underdamping (Case I)

As Case I, it is to be used the roots (6) to solve equation (3).

For the non-real complex roots case, if $c^2 < 4mk$, the characteristic roots are not real as the term under the square root is negative. The damping constant c must be relatively small according to the order $c^2 < 4mk$.

Let $\omega = \omega = \frac{\sqrt{|c^2 - 4mk|}}{2m}$, Then we have characteristic roots: $-\frac{c}{2m} \pm \omega i$.

By that, it can get the basic real solutions:

$e^{-(\frac{c}{2m})t}\cos\omega t\,, e^{-(\frac{c}{2m})t} sin\omega t$ and the general solution

$$x(t) = c_1 e^{-(\frac{c}{2m})t}\cos\omega t + c_2 e^{-(\frac{c}{2m})t} sin\omega t$$

(or) $x(t) = (c_1 \cos\omega t + c_2 sin\omega t)e^{-(\frac{c}{2m})t}$

$$= Ae^{-(\frac{c}{2m})t}(\cos\omega t - \emptyset)$$

2.2. Critical-damping (Case II)

Now, it is to be used the roots to solve equation (1) in Case II. If $c^2 = 4mk$ then the term under the square root is 0 and the characteristic polynomial has repeated roots $-\frac{c}{2m}, -\frac{c}{2m}$. This is the repeated real roots case.

Basic solutions: $e^{-(\frac{c}{2m})t}$, $te^{-(\frac{c}{2m})t}$

General solution: $x(t) = c_1 e^{-(\frac{c}{2m})t} + c_2 te^{-(\frac{c}{2m})t}$

$$= (c_1 + c_2 t)e^{-(\frac{c}{2m})t}$$

2.3. Over-damping (Case III)

Case III is the distinct real roots case. If $c^2 > 4mk$ then the term under the square root is positive and the characteristic roots are real and distinct. We can see that case by looking at the formula (3). The term under the square root is positive by assumption, so the roots are real. Since $c^2-4mk < c^2$, the square root is less than c and therefore the root: $-c+\sqrt{c^2-4mk} < 0$.The other root is clearly negative.

Now we use the roots to solve equation (1) in this case. Characteristic roots:

$$\lambda_1 = \frac{-c+\sqrt{c^2-4mk}}{2m}, \lambda_2 = \frac{-c-\sqrt{c^2-4mk}}{2m}$$

Exponential solutions: $e^{\lambda_1 t}, e^{\lambda_2 t}$

General solution: $x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$.

3. APPLICATION PROBLEMS

Three cases of Homogeneous Systems are to be practiced both in calculation and MATLAB analysis.

3.1. Example of Under-damped (Example 1)

To find the equation of motion: A 1 kg is attached to a spring whose constant is 4 N/m. The system is subsequent motion takes place in a medium that offers a damping force that is numerically equal to 3 times the instantaneous velocity. The mass is released from rest at a point 1 m below the equilibrium. Then, the solution becomes: The differential equation is given by $x'' + 3x' + 4x = 0$,

The auxiliary equation for $\lambda^2 + 3\lambda + 4 = 0$, So $\lambda_1 = -\frac{3}{2}+\frac{\sqrt{7}}{2}i$ and $\lambda_2 = -\frac{3}{2}-\frac{\sqrt{7}}{2}i$, which shows that the system is under damped,

Hence: $x(t) = e^{at}(c_1 \cos bt + c_2 \sin bt)$

$$x(t) = e^{-\frac{3}{2}t}(c_1 \cos\frac{\sqrt{7}}{2}t + c_2 \sin\frac{\sqrt{7}}{2}t)$$

Applying the initial conditions:

$x(0) = 1$ and $x'(0) = 0$.Then find, in turn, that $c_1 = 1$ and $c_2 = \frac{3\sqrt{7}}{7}$. Thus, the equation of motion: $x(t) = e^{-\frac{3}{2}t}(\cos\frac{\sqrt{7}}{2}t + \frac{3\sqrt{7}}{7}\sin\frac{\sqrt{7}}{2}t)$ (7)

The solving and analyzing (proofing) process of Under-damped case using MATLAB are shown in Figure 1(a) and the plotting diagram is shown in Figure 1(b). [5] [7]

Table 1. Process Description of solving

$$\lambda^2 + 3\lambda + 4 = 0$$

```
syms t x
dsolve ('D2x+3*Dx+4*x = 0', 'x(0) = 1', 'Dx(0) = 0', 't')
ans=
exp (-(3*t)/2)* cos((7^(1/2)*t)/2) +
(3*7^(1/2)*exp(-(3*t)/2) *sin((7^(1/2)*t)/2))/7
x = exp(-(3*t)/2)* cos((7^(1/2)*t)/2) +
(3*7^(1/2)*exp(-(3*t)/2)* sin((7^(1/2)*t)/2))/7;
LHS = diff(x,t,2) + 3*diff(x,t) + 4*x
LHS =0
x = exp(-(3*t)/2)* cos((7^(1/2)*t)/2) +
(3*7^(1/2)*exp(-(3*t)/2)* sin((7^(1/2)*t)/2))/7;
subs(x,{t},{0})
ans =1
dxdt =diff(x,t); subs(dxdt,{t},{0})
ans =0
syms t x
p1 = ezplot('exp(-2*t) + 2*t*exp(-2*t)'
-x,[-1,6,-0.3,1.2]);
p1.Color = 'blue';
p1.LineStyle = '-.';
p1.LineWidth = 1.5;
grid
xlabel('Independent variable t');
ylabel('Dependent variable X');
title('The function [exp(-2*t) +
 2*t*exp(-2*t)]')
```

Figure 1(a). Matlab Script for Example (1)

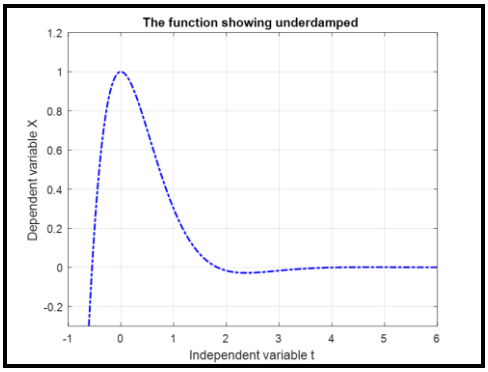


Figure 1(b). Underdamped Plot of Example (1)

3.2. Example of Critical-damped (Example 2)

Same question as in example (1) for 4 times the instantaneous velocity. Then, the solution becomes: The differential equation is given by

$$x'' + 4x' + 4x = 0,$$

The auxiliary equation is

$$\lambda^2 + 4\lambda + 4 = (\lambda+2)(\lambda+2) = 0,$$

So λ_1=-2 and λ_2=-2.

That is critical, $x(t) = c_1 e^{\lambda_1 t} + c_2 t e^{\lambda_2 t}$

$$x(t) = c_1 e^{-2t} + c_2 t e^{-2t}$$

x(0)=1 and x' (0)=0, Therefore c_1=1 and c_2=2.

Thus, the equation is

$$x(t) = e^{-2t} + 2te^{-2t} = (1+2t)e^{-2t}$$

The solving and analyzing (proofing) process of Critical-damped case using MATLAB are described in Figure 2 and the plotting diagram is shown in Figure 3). [5] [7]

```
syms t x
dsolve('D2x+4*Dx+4*x = 0', 'x(0) = 1', 'Dx(0) = 0', 't')
 ans = exp(-2*t) + 2*t*exp(-2*t)
x=exp(-2*t) + 2*t*exp(-2*t);
LHS = diff(x,t,2) + 4*diff(x,t) + 4*x
LHS =0
 x=exp(-2*t) + 2*t*exp(-2*t);
 subs(x,{t},{0})
 ans =1
  dxdt = diff(x,t);
 subs(dxdt,{t},{0})
 ans =0
syms t x
p1 = ezplot('exp(-2*t) + 2*t*exp(-2*t)'-x,[-1,6,-0.3,1.2]);
p1.Color = 'blue';
p1.LineStyle = '-.';
p1.LineWidth = 1.5;
grid
xlabel('Independent variable t');
ylabel('Dependent variable X');
title('The function [exp(-2*t) + 2*t*exp(-2*t)]')
```

Figure 2. MATLAB Script for Example (2)

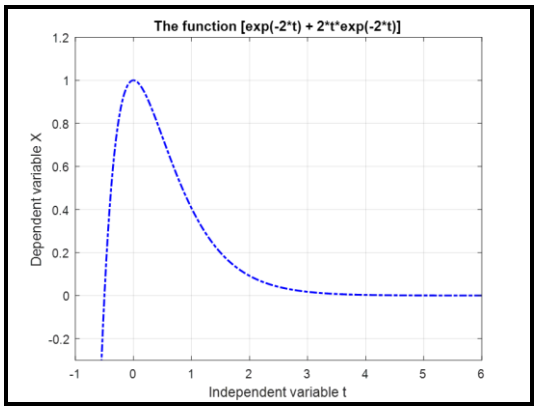


Figure 3. Critical Damped Plot of Example (2)

3.3. Example of Over-damped (Example 3)

Same question as in example (1) for 5 times the instantaneous velocity. Then, the solution becomes: The differential equation is given by

$$x'' + 5x' + 4x = 0,$$

The auxiliary equation is $\lambda^2+5\lambda+4=0$,

So $\lambda_1=-1$ and $\lambda_2=-4$.

The over damped is

$$x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

$$x(t) = c_1 e^{-t} + c_2 e^{-4t}$$

x(0)=1 and x' (0)=0, that $c_1 = \frac{4}{3}$ and $c_2 = -\frac{1}{3}$.

The equation is $x(t) = \frac{4}{3}e^{-t} - \frac{1}{3}e^{-4t}$.

The solving and analyzing (proofing) process of using MATLAB are similar as previous cases and the result plotting diagram is as shown in Figure 4.). [5] [7]

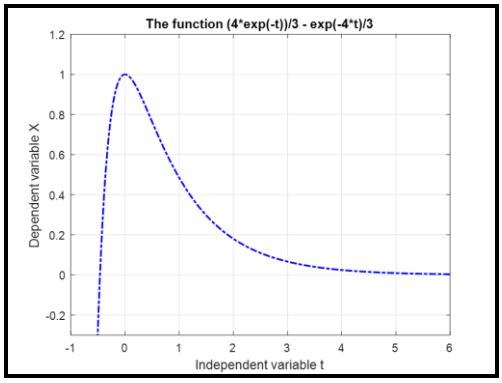


Figure 4. Over Damped Plot of Example (3)

3.4. Damping Practices upon Other Solutions

Small oscillations of the pendulum can be described by the homogeneous differential equation:

$$m\frac{d^2y}{dt^2} + \varphi\frac{dy}{dt} + \frac{mg}{l}y = 0 \qquad (8)$$

where m is the mass of the bob, l is the length of the string, g is the magnitude of the acceleration due to gravity, and φ is a damping constant. With mass measured in kilograms, length in meters and time in seconds, the values of these constant are to be analyzed and compared in three cases as shown in Figure 5, 6 and 7.

Case I : $m = 5, l = 14.41, g = 9.8$ and $\varphi = 8$.

(under damped)

$$m\frac{d^2y}{dt^2} + \varphi\frac{dy}{dt} + \frac{mg}{l}y = 0$$

$$5y'' + 8y' + \frac{17}{5}y = 0$$

$$5\lambda^2 + 8\lambda + \tfrac{17}{5} = 0$$

$$\lambda_1 = -\frac{4}{5} - \frac{1}{5}i, \qquad \lambda_2 = -\frac{4}{5} + \frac{1}{5}i$$

$$y = e^{-4/5 t}(C_1 \cos\frac{1}{5}t + C_2 \sin\frac{1}{5}t)$$

$$C_1 = 1, C_2 = 4$$

$$\therefore y = e^{-4/5 t}(\cos\frac{1}{5}t + 4\sin\frac{1}{5}t)$$

Figure 5. Under Damped with m and l values

Case II : $m = 6, l = 22.05, g = 9.8$ and $\varphi = 8$. (critical damped)

$$m\frac{d^2y}{dt^2} + \varphi\frac{dy}{dt} + \frac{mg}{l}y = 0$$

$$6y'' + 8y' + \frac{8}{3}y = 0$$

$$6\lambda^2 + 8\lambda + \frac{8}{3} = 0$$

$$\lambda_1 = -\frac{2}{3}, \qquad \lambda_2 = -\frac{2}{3}$$

$$y = C_1 e^{-2/3\,t} + C_2 t\, e^{-2/3 t}$$

$$C_1 = 1, C_2 = 2/3$$

$$\therefore y = e^{-2/3 t} + \frac{2}{3}\, t\, e^{-2/3 t}$$

Figure 6. Critical Damped with *m* and *l* values

Case III : $m = 7, l = 34.92,\ g = 9.8$ and $\varphi = 8$. (over damped)

$$m\frac{d^2y}{dt^2} + \varphi\frac{dy}{dt} + \frac{mg}{l}y = 0$$

$$7y'' + 8y' + \frac{55}{28}y = 0$$

$$7\lambda^2 + 8\lambda + \frac{55}{28} = 0$$

$$\lambda_1 = -\frac{11}{14}, \qquad \lambda_2 = -\frac{5}{14}$$

$$y = C_1 e^{-11/14\,t} + C_2\, e^{-5/14\,t}$$

$$C_1 = -5/6, C_2 = 11/6$$

$$\therefore y = -\frac{5}{6}\, e^{-11/14\,t} + \frac{11}{6}\, e^{-5/14\,t}$$

Figure 7. Over Damped with *m* and *l* values

4. Analysis Results

By the contributions of Example Cases for Damping Homogeneous System, the applications or practical approaches of Second Order ODE are carried out their solutions. Both Calculating concepts and Software Approaches are tested or compared to do these cases. All analysing processes are done by the help of MATLAB with the contributions of understanding the functions. The clear performance of the real-world application can be analysed or compared as shown in Figure 9 and the second practice as shown in Figure 11.

5. Conclusions

Engineering Mathematics concepts are to be applied in real-time practices of Engineering Specialization. This paper was focused on the spring mass system, to find and solve the second order differential equations and compared or simulated with MATLAB. These approaches are very useful in Mechanical Engineering and also in Mechatronic Experiments. Three conditions of damping are discussed which involved the under damped, critically damped and over damped. Understanding and case studied of second order differential equations are then mainly focused in this paper. A forcing term was then introduced to the system, and the system's response was plotted into several graphs using MATLAB. These graphs illustrated the effects of damping conditions depends on their damped cases. Through this paper, an understanding vibration control of engineering structures are practiced through mathematical thinking. It can clearly be described that solving differential equations approaches are applied in all Science and Engineering applications.

Combined Graphs:

```
x1 = dsolve('D2x+3*Dx+4*x = 0', 'x(0) = 1', 'Dx(0) = 0', 't') ;
x2 = dsolve('D2x+4*Dx+4*x = 0', 'x(0) = 1', 'Dx(0) = 0', 't') ;
x3 = dsolve('D2x+5*Dx+4*x = 0', 'x(0) = 1', 'Dx(0) = 0', 't');
syms t x
p1 = ezplot(x1 -x,[-1,6,-0.3,1.2]);
p1.Color = 'blue';
p1.LineStyle = '-';
p1.LineWidth = 1.5;
grid
hold on
p2 = ezplot(x2 -x,[-1,6,-0.3,1.2]);
p2.Color = 'red';
p2.LineStyle = '--';
p2.LineWidth = 1.5;
p3 = ezplot(x3  -x,[-1,6,-0.3,1.2]);
p3.Color = 'black';
p3.LineStyle = '-.';
p3.LineWidth = 1.5;
hold off
xlabel('Independent variable t');
ylabel('Dependent variable X');
title('Damping comparison')
legend('underdamped','critical damped','overdamped')
```

Figure 8. MALAB Script to plot Combine Graph of three cases

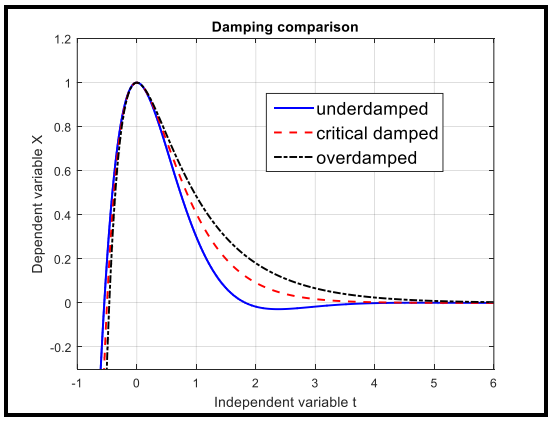


Figure 9. The Result that showing three cases of Damping

```
Combine Graph 2
y1 = dsolve('5*D2y+8*Dy+3.4*y = 0', 'y(0) = 1', 'Dy(0) = 0', 't') ;
y2 = dsolve('6*D2y+8*Dy+2.67*y = 0', 'y(0) = 1', 'Dy(0) = 0', 't') ;
y3 = dsolve('7*D2y+8*Dy+1.964*y = 0', 'y(0) = 1', 'Dy(0) = 0', 't');
syms t y
p1 = ezplot(y1 -x,[-1,6,-0.3,1.2]);
p1.Color = 'blue';
p1.LineStyle = '-';
p1.LineWidth = 1.5;
grid
hold on
p2 = ezplot(y2 -x,[-1,6,-0.3,1.2]);
p2.Color = 'red';
p2.LineStyle = '--';
p2.LineWidth = 1.5;
p3 = ezplot(y3  -x,[-1,6,-0.3,1.2]);
p3.Color = 'black';
p3.LineStyle = '-.';
p3.LineWidth = 1.5;
hold off
xlabel('Independent variable t');
ylabel('Dependent variable Y');
title('Damping comparison')
legend('underdamped','critical damped','overdamped')
```

Figure 10. MALAB Script to plot Combine Graph (*m* and *l* values are changed)

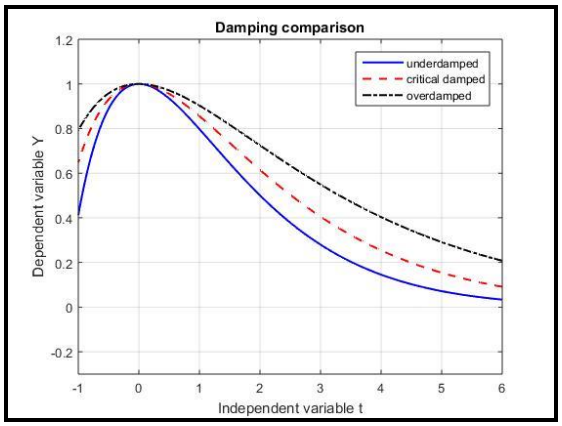


Figure 11. The Result that showing three cases of Damping (*m* and *l* values are changed)

ACKNOWLEDGEMENTS

The Special thanks are also due to whom involved towards development of this journal paper to give a chance of research paper contribution for the author.

References

[1] E. Kreyszig, 'Advanced Engineering Mathematics', 10th Edition, Copyright © 2011, 2006, 1999 by John Wiley & Sons, Inc. ISBN 978-0-470-45836-5.

[2] Parasuram Harihara and Dara W. Childs, 'Solving Problems in Dynamics and Vibrations Using MATLAB', Department of Mechanical Engineering Texas A & M University College Station.

[3] Pramote Dechaumphai, 'Calculus and Differential Equations with MATLAB', Copyright © 2016, ISBN 978-1-78332-265-7Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] Foutse Yuehgoh 'Second Order Linear Ordinary Differential Equations and Their Applications to the Study of Problems in Physics', University of Buea, Department of Mathematics, July 2014, https:// www. researchgate.net/publication/309040457.

[5] The Open University, Walton Hall, Milton Keynes, MK76AA.,'MST224, Mathematical methods, Second-order differential equations', First published 2013.Copyright c 2013 The Open University, ISBN9781780074795

[6] http://www.rpi.edu/~des/Diffeq.ppt.

[7] https://www.mathworks.com

Authors Biography

Dr. Yee Yee Htun, B. Sc (Hons), Maths (2003), M.Sc (Maths) (2005), Ph.D (Applied Maths) (2008)-MTU. The first author was currently work at Department of Engineering Mathematics, TU (Meiktila) as a Professor.

Four published research papers of First Author that relative to Engineering Mathematic are at

1. https://publications.waset.org/1511/fuzzy-mathematical-morphology-approach-in-image-processing
2. https://zambrut.com/wp-content/uploads/2019/05/Mandelbrot-Julia.pdf
3. https://zambrut.com/wp-content/uploads/2019/07/Algorithms-MATLAB.pdf
4. Analysis of Laplace Transform for Solving ODE using matlab, Journal of Information Technology, Research and Innovation 2020, Vol-1, Issue-1

Mr. Htay Lin Aung, B.Sc (Hons) Maths (2009), M.Sc. (Maths) (2012), Meiktila University. The second author was currently work at Department of Engineering Mathematics, TU (Meiktila) as a Lecturer.

ORGANIZING THE SIMILAR DATA BY USING THE CLUSTER ANALYSYS

Ohnmar Aung[1], Aye Nyein San[2], Yin Min Htwe[3], and Ohnmar Myint[4]

[1,2 ,3,4]Faculty of Information Science, University of Computer Studies (Mandalay)
[1]mamalay2009@gmail.com, [2]ayenyeinsan6@gmail.com,
[3]shinminhtwe@gmail.com,[4]marmaromm@gmail.com

ABSTRACT

The practice of identifying valuable patterns in big data sets is known as data mining. In data mining applications, this study introduces a hierarchical probabilistic clustering method for both supervised and unsupervised learning. The fact that a human user specifies the number of clusters in many clustering algorithms is a drawback. It is frequently unrealistic to assume that a person with adequate subject expertise would always be available to choose how many clusters to return. A common tool for organizing data and looking for trends is hierarchical clustering. The system will cluster dental diseases, of similar characteristics and use the Divisive ANAlysis (DIANA) clustering method to cluster the dental diseases, which contain city name and their diseases. It also implemented to compare the distance function of Interval-scale variable in cluster analysis. In this paper, to attend to understand the important of the data mining. To realize the clustering of the large amount of data. To exhibit the usefulness of clustering. And to demonstrate the practical usage of DIANA method.

KEYWORDS

Data Mining, DIANA, Clustering

1. INTRODUCTION

One effective and widely used method for identifying structure in data is clustering. Multilevel data descriptions are provided by hierarchical approaches for both supervised and unsupervised data mining. The main premise of this dissertation is that alignment and clustering should work together to improve predictive modeling capabilities rather than separately, taking advantage of the symbiotic relationship between the two techniques. We present a unique approach that simultaneously permits learning sets of continuous curve transformations and facilitates the clustering and prediction of sets of smoothly variable curves. Large data sets can be understood and explored with the help of clustering. It is possible to observe that curve clustering as a methodology focuses on particular kinds of data sets and algorithms designed to function on curves as a unit. Algorithms for clustering have traditionally worked with fixed-dimensional feature vectors or points. On the other hand, curves typically have any number of missing observations together with a variable number of measurements observed during varying-sized measurement intervals. The main difference between curves and feature vectors is the smoothness information that curves contain, which limits the amount that observations fluctuate between measurements.

Our purpose system focuses specifically on dental diseases. Most cases are dental caries (tooth decay), periodontal diseases, tooth loss and oral cancers. Other oral conditions of public health importance are orofacial clefts, noma (severe gangrenous disease starting in the mouth mostly affecting children) and oro-dental trauma. Therefore, dental diseases are the important in Myanmar nowadays. This paper is to identify the user's want to know which city are felt the dental diseases. This system will assist these city and nearest city felt the same dental diseases. This research paper used clustering in section 2. This paper used cluster analysis in section 3. There are many types of cluster analysis, this paper

used Interval-Scaled Variables and Euclidean distance function of Interval-Scaled Variables and Manhattan distance function of Interval-scaled Variables in section 4. In Section 5, describes this paper used hierarchical methods and Divise Clustering DIANA (Top-Down) method. The process of the system and how to cluster the large data set, attributes and implement the detail of the result are discussed in section 6. Concludes in section 7 respectively.

2. CLUSTERING

The method of clustering involves classifying the data into groups or clusters so that objects within a cluster are highly similar to each other but also significantly different from each other. Distance measures are used frequently. Numerous fields, including data mining, statistics, biology, and machine learning, are the foundation of clustering. Over the past 40 years, clustering as a data analysis technique has been thoroughly investigated. For a general survey, see [2]. Algorithms for clustering data can be broadly divided into two groups: those that produce a hierarchy and those that divide the data (K-means, for example). Our goal is to establish and preserve a hierarchy. [1] Outlines five different methods for completing this work in great detail. Agglomerative clustering, which repeatedly finds the two closest nodes in the system and merges them until the hierarchy is complete, is the most widely used method. Dividend algorithms start with every leaf in one cluster. A cluster is chosen at each stage and split into smaller clusters to optimize the degree of similarity between the elements in each cluster.

3. CLUSTER ANALYSIS

Clustering is the process of organizing a collection of concrete or abstract things into classes of related objects. A cluster is an assemblage of data objects that differ from the objects in other clusters but are comparable to each other within the same cluster [4].

3.1. Requirement of clustering

Clustering requirements are: scalability, adaptability to many attribute kinds, findings of clusters with random shapes, minimal domain expertise needed to select input parameters, high dimensionality, constraint-based clustering; ability to handle noisy data, intense attention to the sequence of input records, interpretability and usability.

4. INTERVAL-SCALED VARIABLES

Variables with an interval scale are continuous measures with a nearly linear scale. Weight and height, latitude and longitude coordinates, and weather temperature are typical examples. The clustering analysis may be impacted by the measurement unit chosen. A completely new clustering structure could result, for instance, from converting measuring units for weight and height from kilograms to pounds or from meters to inches.

Analysis of clusters is a crucial human endeavor. Early in life, one regularly refines subconscious clustering systems to learn how to discriminate between animals and plants, or between cats and dogs. Numerous fields, such as pattern identification, data analysis, image processing, and market research, have made extensive use of cluster analysis. Through clustering, general distribution patterns and intriguing relationships between data attributes can be found by identifying sparse and dense zones. It is used in biology to classify genes with comparable functions, create taxonomies for plants and animals, and understand the innate structures of populations. Cluster analysis is a data mining function that may be used independently to analyze the properties of each cluster, understand the distribution of data, and narrow the emphasis to a specific collection of clusters for additional study. The goal of data mining has been to discover techniques for productive and successful cluster analysis in big databases. Research is now being done on high-

dimensional clustering techniques, scalability of clustering methods, efficacy of clustering complicated shapes and types of data, and approaches for clustering mixed numerical and categorical data in huge databases.

Euclidean distance function of Interval-scaled variables

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2} \quad (1)$$

Manhattan distance function of Interval-scaled variables

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+....+|x_{ip}-x_{jp}| \quad (2)$$

where,

$i=(x_{i1}, x_{i2}, ..., x_{ip})$ and $j=(x_{j1}, x_{j2}, ...x_{jp})$ are two p-dimensional data objects

Case Study,

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2} \quad (1)$$

where, i=kyaukse, j=kyaukse

d(*Kyaukse*, *Kyaukse*)

$$\sqrt{|12-12|^2+|21-21|^2+...+|23-23|^2}$$

=0+0+0+0+0++0

=0

where, i=Kyaukse, j=Kume

d(*Kyaukse*, *Kume*)

$$\sqrt{|12-23|^2+|21-25|^2+...+|23-12|^2}$$

=11+4+16+2+9+12+11

=65

Similarly, we will compute the distance between city.

5. CLUSTERING METHODS

There are five clustering methods: partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods.

Algorithms for partition-based clustering divide n data objects into k partitions in order to optimize a certain objective function, such as K-mean [6]. A hierarchical breakdown of the data set is produced by hierarchical clustering methods and can be shown as a dendrogram, a type of tree structure [10, 11, 12]. The data space is divided into rectangular grid cells by grid-based clustering algorithms, which subsequently perform statistical analysis on these grid cells [5, 9, 13]. Density-based clustering algorithms in densely inhabited areas, group data objects by iteratively growing a cluster towards high-density neighborhoods [9, 7, 8]. Certain mathematical models are fitted to the data items via model-based clustering methods, [7].

5.1. Clustering in a hierarchical manner

A collection of disconnected clusters is not produced by a hierarchical clustering method. Rather, it produces a dendrogram, which is a tree that represents a hierarchy of nested clusters. Hierarchical clustering algorithms can be further subdivided into agglomerative algorithms (i.e., bottom-up techniques) and divisive algorithms (top-down approaches) based on how the hierarchical breakdown is created. As a matter of fact, the biologists prefer the hierarchical approaches over the others since they have the potential to provide additional information on the cluster structure [10, 11, 12].

The hierarchical approaches do, however, have certain shortcomings. First, figuring out where to cut the dendrogram and extract clusters might occasionally be subtle. Usually, visual inspection by domain specialists completes this process. Second, it might be challenging to determine from a dendrogram which objects are the cluster's boundaries and which are its medoid, or inner structure.

Finally, a lot of hierarchical techniques are thought to be weak and unoriginal. They could be susceptible to minor changes in

the data's order and input sequence. The data are not immediately divided into one cluster using hierarchical clustering. Rather, there are a number of partitions that may originate from a single cluster that houses every object. A sequence of data merging or splitting that continue until a final number of clusters is reached [1].

5.2. Methods with a hierarchical

Data objects are grouped into a tree of clusters using the Hierarchical Clustering method. Different decomposition techniques can result in different hierarchical methods such as Top-Down and Bottom-Up.

- Agglomerative clustering (Bottom-UP)
- Divisive clustering DIANA (Top-Dow)

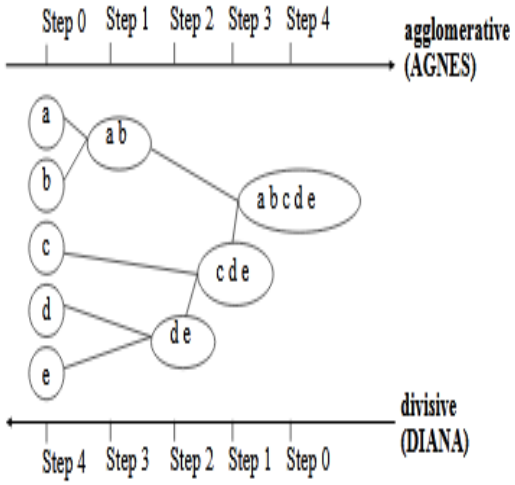


Figure 1. DIVISIVE Clustering DIANA (Top Down)

5.2.1. From an attraction tree, how can one drive a density tree?

The entire data set is first split into dense areas, which are then split further into smaller and smaller sub-areas until each dense sub-area contains only one cluster. Two parameters are introduced to determine the clusters, dense areas, and bridges between them: a similarity threshold and a minimum number of objects threshold.

6. THE PROCESS OF THE SYSTEM

The process of this system is, when any user wants to add the data; the data can be entered to the database, or existing distance function of the data can be used in the database such as database existed dentist table. And then, select from choice method: Euclidean Distance or Manhattan Distance. Calculate the distance function of the data in chosen table in database. And then, calculate the mean value upon these distances.

After clustering, two leaves will display: the value of the first leaf is less than the mean value and the value of the second leaf is greater than or equal to the mean value. Tree is then split and the same cluster is checked at the user satisfied level. If the data are not in the same cluster, repeat calculation the data. And then, when the data reach at the user desire level, result tree will display.

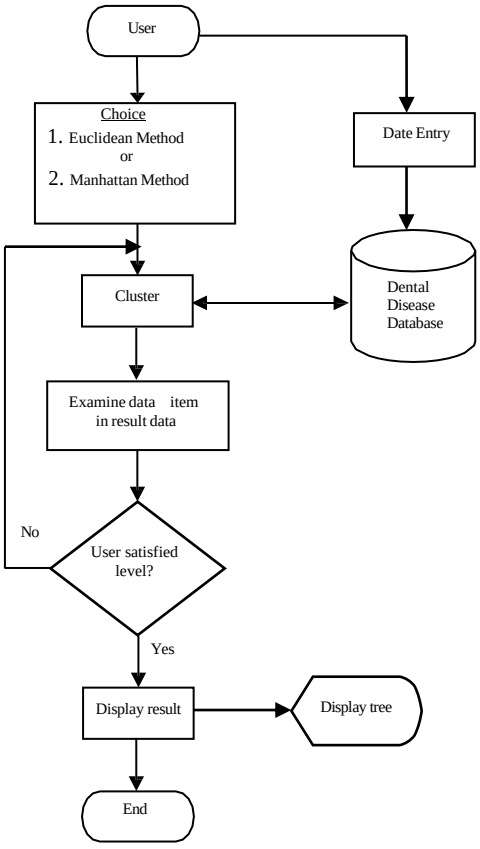


Figure 2. System Flow Diagram

6.1 Result

The result of this system is, first: user wants to enter the data in the database or existing distance function of the data can be used in database. This system used dental table, which contain city name and their attributes are dental diseases. And then, users choose the methods: Euclidean Distance or Manhattan Distance. For example, users choose the Euclidean method. Euclidean method is calculating the distance value: first row of the data in the table is standard. These data are compared with the next data and calculate in the Euclidean method. The distance value appears. Then calculate the mean value on the distance value. The mean value appears*. The table has three parts. The first part is the root table which has the whole data set. In the second part, it contains the distance function less than the mean value. And the last table includes the distance function greater than or equal to the mean value. User can cluster to reach at his or her satisfied level. When the data reached at the user desired level, result tree will display.

*The mean value is:

$$Mean\ value = \frac{\text{Total of result distances}}{\text{Number of rows}} \quad (3)$$

In sample dental diseases datasets, compute the mean value.
Total of result distances=
0+65+45+43+39+30+37+38+13+19+27+16
+31+28+48+35+43
=661
Number of rows= 20
Mean value= 661/20
=33.05

6.2. Dental diseases data set table with example

The data are put into the database. This system will cluster 230 datasets and 12 attributes. The sample datasets are 20 and sample attributes are 7. These data items are clustered by using interval-scaled variables in cluster analysis.

Table 1. Dental diseases, root data set

ID	Name	Dental_Caries	Periodotal	Oral_Cancer	Dento_Facial	Maxilo_Facial	Washing_Disease	Fluorosis
1	Kyaukse	12	21	22	21	21	22	23
2	Kume	23	25	6	23	12	10	12
3	Myittha	20	24	21	25	14	20	18
4	Madaya	10	18	10	26	15	10	26
5	Pyinoolwin	9	16	15	14	18	12	19
6	Monywa	20	17	14	25	21	25	20
7	Pakokku	19	13	18	17	25	18	17
8	Myingyin	20	14	19	19	26	19	13
9	Lashio	15	22	25	20	23	19	23
10	Myitkyina	13	23	26	23	24	24	28
11	Kyaukme	11	28	28	25	27	23	21
12	Kalaw	10	24	24	21	28	21	22
13	Nawnghkio	15	21	13	28	21	10	23
14	Kutkai	16	20	29	24	20	15	18
15	Amarapura	18	10	14	17	18	10	19
16	Sagaing	20	22	15	15	21	22	10
17	Taunggyi	26	19	21	13	19	23	8
18	Thipaw	28	17	27	20	22	26	21
19	Kawlin	24	19	29	19	17	27	17
20	Khanti	22	20	25	11	22	18	19

6.3. Clustering with Euclidean method for dental diseases data

Table 2 show calculate the distance data in database, using Euclidean method and release the distance data. And then, calculate the mean value will appear. In sample dataset the mean value is 33.05.

Table 2. Distance value

ID	Name	Dental_Caries	Periodotal	Oral_Cancer	Dento_Facial	Maxilo_Facial	Washing_Disease	Fluorosis	Distance
1	Kyaukse	12	21	22	21	21	22	23	0
2	Kume	23	25	6	23	12	10	12	65
3	Myittha	20	24	21	25	14	20	18	45
4	Madaya	10	18	10	26	15	10	26	43
5	Pyinoolwin	9	16	15	14	18	12	19	39
6	Monywa	20	17	14	25	21	25	20	30
7	Pakokku	19	13	18	17	25	18	17	37
8	Myingyin	20	14	19	19	26	19	13	38
9	Lashio	15	22	25	20	23	19	23	13
10	Myitkyina	13	23	26	23	24	24	28	19
11	Kyaukme	11	28	28	25	27	23	21	27
12	Kalaw	10	24	24	21	28	21	22	16
13	Nawnghkio	15	21	13	28	21	10	23	31
14	Kutkai	16	20	29	24	20	15	18	28
15	Amarapura	18	10	14	17	18	10	19	48
16	Sagaing	20	22	15	15	21	22	10	35
17	Taunggyi	26	19	21	13	19	23	8	43
18	Thipaw	28	17	27	20	22	26	21	33
19	Kawlin	24	19	29	19	17	27	17	38
20	Khanti	22	20	25	11	22	18	19	33

Table 3. Clustering the data, less than the mean value= 33.05

ID	Name	Dental_Caries	Periodotal	Oral_Cancer	Dento_Facial	Maxilo_Facial	Washing_Disease	Fluorosis	Distance
1	Kyaukse	12	21	22	21	21	22	23	0
6	Monywa	20	17	14	25	21	25	20	30
9	Lashio	15	22	25	20	23	19	23	13
10	Myitkyina	13	23	26	23	24	24	28	19
11	Kyaukme	11	28	28	25	27	23	21	27
12	Kalaw	10	24	24	21	28	21	22	16
13	Nawnghkio	15	21	13	28	21	10	23	31
14	Kutkai	16	20	29	24	20	15	18	28
18	Thipaw	28	17	27	20	22	26	21	33
20	Khanti	22	20	25	11	22	18	19	33

Table 4. Clustering the data, greater than or equal to mean value=33.05

ID	Name	Dental_Caries	Periodotal	Oral_Cancer	Dento_Facial	Maxilo_Facial	Washing_Disease	Fluorosis	Distance
2	Kume	23	25	6	23	12	10	12	65
3	Myittha	20	24	21	25	14	20	18	45
4	Madaya	10	18	10	26	15	10	26	43
5	Pyioolwin	9	16	15	14	18	12	19	39
7	Pakokku	19	13	18	17	25	18	17	37
8	Myingyin	20	14	19	19	26	19	13	38
15	Amarapura	18	10	14	17	18	10	19	48
16	Sagaing	20	22	15	15	21	22	10	35
17	Taunggyi	26	19	21	13	19	23	8	43
19	Kawlin	24	19	29	19	17	27	17	38

In table 5 show, if user want to be more cluster in less than the mean value =33.05. And then calculate the distance with Euclidean method and the mean value is calculated.

Table 5. Clustering the data, less than the mean value= 23

ID	Name	Dental_Caries	Periodotal	Oral_Cancer	Dento_Facial	Maxilo_Facial	Washing_Disease	Fluorosis	Distance
1	Kyaukse	12	21	22	21	21	22	23	0
9	Lashio	15	22	25	20	23	19	23	13
10	Myitkyina	13	23	26	23	24	24	28	19
12	Kalaw	10	24	24	21	28	21	22	16

Table 6. Clustering the data, greater than or equal to mean value=23

ID	Name	Dental_Caries	Periodotal	Oral_Cancer	Dento_Facial	Maxilo_Facial	Washing_Disease	Fluorosis	Distance
6	Monywa	20	17	14	25	21	25	20	30
11	Kyaukme	11	28	28	25	27	23	21	27
13	Nawnghkio	15	21	13	28	21	10	23	31
14	Kutkai	16	20	29	24	20	15	18	28
18	Thipaw	28	17	27	20	22	26	21	33
20	Khanti	22	20	25	11	22	18	19	33

In table 7 show, if user want to be more cluster in less than the mean value =23. And then calculate the distance with Euclidean method and the mean value is calculate.

Table 7. Clustering the data, less than the mean value= 12

ID	Name	Dental_Caries	Periodotal	Oral_Cancer	Dento_Facial	Maxilo_Facial	Washing_Disease	Fluorosis	Distance
1	Kyaukse	12	21	22	21	21	22	23	0

Table 8. Clustering the data, greater than or equal to mean value=12

ID	Name	Dental_Caries	Periodotal	Oral_Cancer	Dento_Facial	Maxilo_Facial	Washing_Disease	Fluorosis	Distance
9	Lashio	15	22	25	20	23	19	23	13
10	Myitkyina	13	23	26	23	24	24	28	19
12	Kalaw	10	24	24	21	28	21	22	16

The above data items 20 are used and these dental diseases data are clustered by using Euclidean method. And the dental diseases data item that user wants to search is "kyaukse" as an example. At first the distance value will be calculated in Euclidean method: the first row of the data in the table is standard and it is compared with the next data and calculated in this method. Then the distance value will appear with new column, "Distance" (Table 2) and the mean value will calculate on the distance value. After calculating the mean value will appear. The calculated table consists of three parts. The first part is root table which includes the whole data set, second table contains the distance function less than the mean value (Table3) and the last table consists of the distance function greater than or equal to mean value (Table4). And then user want to cluster the data set table in (Table3) by clicking "More" button and calculate the mean value. User can cluster the data again up, and split less than the mean value (Table5) and greater than or equal the mean value (Table6). And then user want to cluster the data set table in (Table5) by clicking "More" button and calculate the mean value. User can cluster the data again up, and split less than the mean value (Table7) and greater than or equal the mean value (Table8). By clicking "More" button user can cluster the data again up to his or her satisfied level. If user wants to save the result data, he or she has to click "Save"

button. This process, organizes the similar data from a large data set by using interval-scaled variables in cluster analysis.

7. CONCLUSIONS

The system will demonstrate the DIANA method and faster calculation. To show the accuracy of contain the dental diseases with this system. The people when he need to know which city are feel the dental diseases. This system will help these city and nearest city feel the same dental diseases. In this system have city and diseases are calculate and cluster. In Table 8, Lashio, Myitkyina and Kalaw are the same cluster. Because these cities are close to water and land, they suffer from same disease. Therefore, it is easy to know which diseases are common in which cities. And which cities have similar diseases. No requirement of domain knowledge. Existing distance function can be used. No limit of number of attributes. This system also shows the easy to implement with this system. It seems sense to start by searching for underlying data structure when dealing with a big data set. A crucial stage in arranging gene expression data is clustering. In this paper we proposed a simple divisive hierarchical clustering algorithm. For data analysis, hierarchical clustering is an effective method.

REFERENCES

[1] A. D. Gordon. Hierarchical Classification. In P. Arabie, L. J. Hubert, and G. DeSoete, Editors, *Clustering and Classification*, pages 65-121. World Scientific, River Edge, NJ, 1996.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3), 1999.

[3] C. F. Olson. *Parallel Algorithms for Hierarchical Clustering. Parallel Computing*, 21:1313-1325, 1995.

[4] Everitt, B., *Cluster Analysis*, Heinemann, London,1974.

[5] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: "A multi-resolution clustering approach for very large spatial databases." In *Proceedings of VLDB*, pages 428–439, August 1998.

[6] J. MacQueen. "Some methods for classification and analysis of multivariate observations." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[7] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. "OPTICS: Ordering points to identify the clustering structure." In *Proceedings of SIGMOD*, pages 49–60, June 1999

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. "A densitybased algorithm for discovering clusters in large spatial databases with noise." In *Proceedings of Knowledge Discovery in Database (KDD)*, pages 226–231, 1996.

[9] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. "Automatic subspace clustering of high dimensional data for data mining applications." In *Proceedings of SIGMOD*, pages 94–105, June 1998

[10] R. Duda and P. Hart. *Pattern classification and scene analysis*. John Wiley & Sons, New York, 1973.

[11] S. Guha, R. Rastogi, and K. Shim. CURE: "An efficient clustering algorithm for large databases." In *Proceedings of SIGMOD*, pages 73–84, June 1998.

[12] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: "An efficient data clustering method for very large databases." In *Proceedings of SIGMOD*, pages 103–114, June 1996.

[13] W. Wang, J. Yang, and R. R. Muntz. STING: "A statistical information grid approach to spatial data mining." In *Proceedings of VLDB*, pages 186–195, August 1997.

Authors

Biography

Daw Ohnmar Aung:

B.C.Sc.(25.11.2000), B.C.Sc (Hons:) (23.11.2005), M.C.Sc. (17.11.2008). Places of employ are Computer University (Lashio), Computer University (Mandalay), University of Computer Studies (Yangon), Myanmar Institute of Information Technology (Mandalay), Computer University (Myeik), Computer University (Mandalay). Paper and journal are DIANA Based Clustering for Interval- Scaled Variables, Second National Conference on Parallel and Soft Computing, PSC (December 24, 2008) Yangon, Myanmar. Stochastic

Context Free Grammar for Statistical Parsing of Myanmar Natural Language Processing, Proceedings of 16th International Computer Conference on Computer Applications, (2018) Yangon, Myanmar. Proposed Framework for Stochastic Parsing of Myanmar Language, First International Conference on Big Data Analysis and Deep Learning Applications (IDBCL 2018, Springer Journal) Japan. Fingerprint Image Retrieval Based on SURF and MSER Feature, First International Conference on Big Data Analysis and Deep Learning Applications (IDBCL 2018, Springer Journal) Japan. Comparison of Data Science Methods for Cyber-security, University Journal of Research and Innovation, (August ,2020) Pakokku, Myanmar. Predicting Students' Learning Education, Journal of Information Technology, Research And Innovation (December, 2022) Mandalay, Myanmar. Research interests are Data Mining, Mathematics Computing.

PREDICTION OF MANGO DISEASES USING MACHINE LEARNING ALGORITHMS

Toe Toe[1], Thin Thin Win[2], and Phyu Sin Phwe[3]

[1,2,3] Faculty of Information Science, University of Computer Studies (Mandalay)
[1]ttoe30607@gmail.com, [2]thinthinmyosaw@gmail.com, [3]psinphwe@gmail.com

ABSTRACT

The country's economy is heavily dependent on agriculture and the health of crops is essential to its success. Unidentified crop diseases can be expensive for agricultural companies. Badge marks are important. Accurately diagnose crop diseases and tested, death can be prevented. In the early stages of development, healthy and diseased plants are identical, so farmers cannot identify by monitoring crop leaves. With Myanmar exporting large quantities of mangoes, it has become an economically and ecologically important fruit. The main objective is to develop a system that can forecast the attack of diseases on Mango fruit crop using past weather data and crop production. The field sensors collected live weather data to calculate disease prediction in real time. The Random Forest Regression model was trained on past weather data and used to calculate disease outbreak probability. The model showed pretty accurate results in relation to the forecasting of the disease.

KEYWORDS

Internet of Things, temperature, prediction, machine learning, sensors, prediction.

1. INTRODUCTION

All over the world, agriculture has become the main task of increasing the country's economic contribution. Currently, since ecosystem control methods are not widely used, most of the agricultural land is not fully used. These problems prevented an increase in the mango harvest, which had an impact on the agricultural sector. Taking into account the forecast of crop yields, agricultural productivity has increased. [1]. It is useful for farmers to know the yield of mangoes before harvesting so that they can take the necessary precautions when selling and storing.

As a result of these methods, seasonal climatic conditions are also changing due to resources such as land, water and air, leading to food shortages. There is no suitable solution or technology to solve the situation we are facing, although all these problems have been analysed, including weather, temperature and some other factors. There are many ways to accelerate the economic growth of Myanmar's agricultural sector. The yield and quality of the mango crop can be improved in many ways [2]. Data mining can also be used to predict mango production. The article concludes that due to the rapid development of sensor technologies and machine learning methods, the agricultural sector will benefit from cost-effective solutions.

2. FORECASTING DISEASES

The concept of an integrated prevention and agricultural productivity system for predicting mango crop diseases, the Internet of things and machine learning has been developed using crop yield forecasting. Using data sets, the agricultural sector must use machine learning techniques to predict mango production [3]. This research will help farmers determine the yield of their mango crops before planting mango crops and help them make the best choices.

By creating a working prototype of an interactive forecasting system, the company is trying to solve this problem. There are different data analysis methods or algorithms to predict the mango harvest, and

we can use these methods to predict the mango harvest. It uses a random forest algorithm. The most well-known and effective controlled machine learning algorithm, called random forest, can perform classification and regression tasks [4].

It builds an unlimited number of decision trees during learning and generates classes based on the average prediction of each tree (for regression) or class model (classification). Mango yield forecasting uses machine learning as a decision support tool to help decide what to plant and what to do during the growing season. Several algorithms have been used to promote the research of mango yield prediction [5].

3. MACHINE LEARNING (ML)

Machine learning (ML) technology is widely used in various sectors, from supermarkets to analysing consumer behaviour and predicting mobile phone use. The development of agriculture nowadays is influenced by many technologies, the environment and technology. In addition, the use of information technologies can change decision-making and allow farmers to produce in the best possible way [6]. Myanmar has long been engaged in agriculture, but due to several problems, the production of mangoes is unsatisfactory [7].

Agriculture has long been using machine learning in this industry. Forecasting mango yields is one of the most difficult tasks in precision agriculture and many models have been proposed and tested at the moment. Since mango yields depend on several variables, including climate, weather, soil, fertilizer application and variety, this question requires the use of several data sets.

4. LITERATURE REVIEW

Experiments conducted by Harivansh Nigam, Sakshan Garg and Archit Agrawal [8] in the Myanmar government dataset show that random forest machine learning algorithms have the best accuracy in predicting yields. Here we apply the theoretical aspects of risk forecasting in practice. Jupiter laptops are a technology that any information technology specialist should know or use. Jupiter laptops are reliable, adaptable and can be used for sharing, allowing you to freely view data in the same conditions. Data scientists can create documents from code into complex reports and share them using Jupiter Notebook.

Jupiter Notebook uses a step-by-step method of comparing elements such as code, graphics, text, output, etc. Demonstrate the step-by-step analysis process. The output data of each algorithm is stored in a Jupiter notebook and compared with each other. Then use precision [9] to evaluate performance.

Table 1. Recent research works on Mango disease classification

Reference	Approach used	Achieved	Performance Matrix
[1]	SVM, Logistic Regression and Random Forest	To classify plant infection	Able to achieve 87.6%.
[3]	SVM, ANN	To determine health status of plant	- SVM performed better than ANN.
[4]	CNN, AlexNet	Detect and classify plant disease	Able to achieve 96.5%
[5]	CNN	Plant disease detection for Varieties of plant.	Able to achieve 86%
[7]	Red Rust, Bacterial canker Leaf and fruit	Self-acquired Dataset	BPNN has an accuracy of 99.68%.
[12]	Multiple Disease	Self-acquired Dataset	SVM has an accuracy of 80.00%.
[13]	CNN	Able to train the machine	Able to achieve 88.8% and 12.2% can be also filled.
[8]	AlexNet and CNN	Detect disease in mango and potato leaf.	Able to achieve 98.33%
[14]	AlexNet and CNN	Leaf disease detection.	Simple AlexNet – 96.34%. Hybrid of AlexNet and SVM layers – 99.98%.

5. Research Methodology

Any machine learning system needs data. The data set can now be used for machine learning models that use this data set for training [10]. Every new detail entered in the application form acts as a test data set. After the testing process is completed, the model will make predictions based on the conclusions drawn from the training data set. Many people have used satellite images to predict mango production [11]. Collecting data at the regional level is essential because local climates vary. The implementation of the system requires historical information about culture and climate. This information is collected from several websites. The climatic parameters that affect mango harvest the most are Ph, K, N, P, temperature, humidity, precipitation and label. Therefore, monthly data on these climate parameters are collected.

6. Data Sheet

This step involves the collection of data from several sources and the configuration of the dataset. This analysis is used to provide a set of data. The data set is divided into two halves; for example, 20% of the data is used to test the model and the remaining 80% is used to train the model. In order to predict future events, the machine learning method is supervised learning: the data necessary to train a system that has been previously marked with the correct appropriate answers are necessary for supervised learning. Classification tasks can be used to better classify problems using supervised learning [12]. Supervised machine learning algorithms can use examples to apply what has been discovered in the past to newly labelled data. After sufficient training, the system can propose goals to any new entrant. The learning algorithm can also distinguish its output from the exact and expected output, and can detect defects that can then be corrected in the model.

7. Unsupervised Learning

When the data used for training cannot be classified, unsupervised machine learning methods will be used. When the hidden structure is described in unlabelled data, unsupervised learning examines the function of the system process. The system evaluates the output to explain the hidden structure of the unlabelled data, but it also analyses the data and can draw conclusions from the dataset [13].

8. Classification Algorithms

It is necessary to compare the performance of several different machine learning algorithms. Each model will have different performance characteristics [13]. In this study, we used logical regression. This is usually a controlled set of rules. For a given set of functions X and target variables y, in this class, larger discrete values can be taken [15]. Contrary to popular belief, logical regression is a kind of regression. This version creates a regression model to predict the possibility that a set of facts will be recognized as a variable designated "1". In the same way that linear regression assumes that the data follow linear functions, logical regression uses sigmoid functions to represent the data. Even if the selection criteria are included, logical regression is excellent as a class method. Placing the price advantage that depends on the class task itself is a key element of the logical regression. The accuracy rate of the result is 95.22% T.

Here we have implemented a set of SVM rules. The support vector machine (SVM) is a control device used to learn the software algorithms of each class. Although we are talking about regression, it is excellent and acceptable for the whole class. The purpose of the SVM rule set is to find a hyper plane in an n-dimensional domain capable of clearly classifying the data points. The number of options affects the size of the hyper plane [16]. If the number of input functions is two, then the hyper plane is one line. If the number of input functions is three, the hyper plane becomes a two-dimensional plane. When the functions are more than three, it is difficult to think clearly. We obtained SVM with an accuracy of 10%. Using the random forest classifier, the best known and most efficient way to observe random forests, we examined a set of rules

that can perform class and regression tasks by class. This function creates a chaotic sample of bushes when visiting the school and produces an elegant result [17]. This is a class model or a supposed prediction (regression) of male or female bushes. The more bushes there are in the forest, the more accurate the forecast, with an accuracy of 99.09%.

We have implemented a decision tree, which is a controlled learning method that can be used for each classification and regression task, but in most cases, it is very desirable to solve the classification problem [18]. It is a graphical representation of the process of obtaining all the potential solutions to the problem or the choices made entirely according to predetermined criteria. Machine learning uses a variety of algorithms, so choosing an excellent set of rules for a given set of data and solving problems is the main factor that should not be forgotten when developing versions of research equipment. The decision tree classifier showed an accuracy of 90%.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

By providing climate data on this website, a practical website has been developed to predict mango production and any user can use it according to their mango harvest. It shows an accuracy of 99.09%.

Table 2: Performance Metrics of algorithms used with their classification accuracy, F1 score, precision and recall

Algorithm	Area Of Curvature	Classification Accuracy	F1	Precision	Recall
LR	0.988	0.942	0.951	0.966	0.937
DT	0.965	0.923	0.914	0.886	0.943
Navies Bayes	0.988	0.943	0.952	0.970	0.937
SVM	0.966	0.933	0.928	0.933	0.943
RF	0.993	0.986	0.986	0.986	0.986

9. RESULT

The research data is used to test the decision tree first and get the results outlined in the table. Using the data examined in this study, the decision tree classifier was shown to be 90% accurate. A simple Gaussian Bayesian analysis was conducted based on the results of the research, and the classifier provided a 99.09% accuracy rate. It then runs SVM with 10% accuracy. The logical regression type was later used for tests and the result was achieved with 95.22% accuracy. Last but not least. Using a random forest to analyse the date of the day, I 99.09% accuracy, which was found to be. It is compatible with the Gaussian Bayesian accuracy of 99.09%.

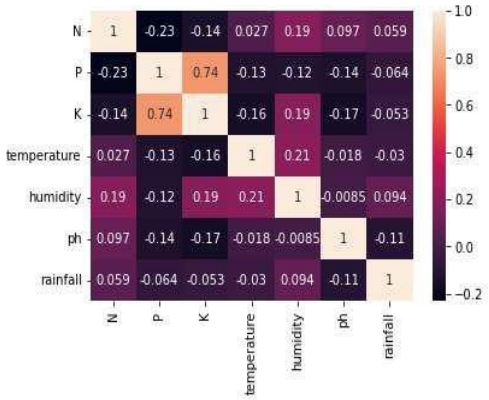


Figure 1. Heat map

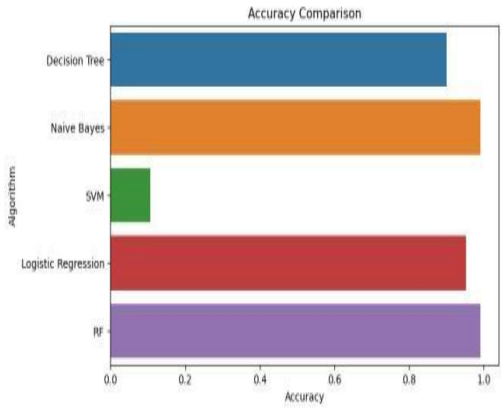


Figure 2. Accuracy comparison of all algorithms

10. CONCLUSIONS

Based on climate data, this study demonstrates the potential use of data mining techniques to estimate mango crop yields. Of all the mango crops and regions selected for the study, the accuracy of the estimates was more than 75%, which indicates a high level of accuracy of the estimates, and the website was designed to be user-friendly. The proposed method for the previous year is the soil, climate and yield information are taken into account and recommendations are made on the most profitable mango crops that can be grown under ideal environmental conditions. The system will cover most mango species, allowing farmers to understand mango crops that may not have been planted before. By listing all conceivable mango crops, this system helps farmers choose which mango crops to grow.

Studies have shown that the likelihood of failure in predicting random forest algorithms is very high. This enables farmers to take preventive measures to increase agricultural production and reduce the use of chemical pesticides in crops. Ultimately, this will prevent foreign flies from appearing in fruit reserves. With the help of more precise ground-based sensors, this system can be extended and implemented in mango farms. The low cost of small business development in Myanmar will also contribute to the use of models designed to accurately diagnose and classify diseases with the system to improve production.

REFERENCES

[1] Srawanvemishetti et al. "Diseases of cultures based on deep learning are classified by transient learning."Today's documents: Minutes of the meeting (2021).

[2] Saleem, Rabia and others. Use a fully defined convolutional network to diagnose mango leaf disease. "Cmc computer hardware and CONTINUA69.3 (2021): 3581-3601.

[3] Kusrini, kusrini and others. Expansion of the data of the automatic classification of pests on the mango farm."Computers and electronic products in agriculture 179 (2020): 105-842.

[4] Pham, Tanhat, Levantran, Sonvutrondo. "The anticipatory neural network and the selection of hybridization meta-characteristics were used to classify early mango leaf diseases."IEEE 8（2020）Access：189960-189973 (in English only)

[5] Nagaraju, Yi et al. "Convolutional neural network model based on transitional learning for the classification of mango leaves infected with anthracnose" 2020IEEE International Conference on Technological Innovation (INCON)IEEE,2020

[6] Ardila, Carlos Eduardo Cabrera, Leonardo Alberto Ramirez and Flavio Augusto Prieto Ortiz were the authors of this book. "Spectroscopic analysis of the early detection of anthrax in the mangiferaindica sugar fruit."Computers and electronic products in agriculture 173 (2020): 105357

[7] Mango crop yield estimation by aruvansh Nigam, Sakshan Garg, Archit Agrawal ML algorithm; 2019.

[8] The advantages and challenges of the Internet of Things (IoT) and agricultural data analysis Olakunle Elijah, Tharek Abdul Rahman, IgbafeOrikumhi, Chee Yen Leow, ISSN: 2319-7242, Volume 5 / Number 6 / Number 5, October 2018

[9] Professor of Mango Harvest Estimation System using machine learning, DS Zingade, Ilkar Buchade, Niresh Mehta Shubha Ghodekar, Chandan Mehta, International Journal of Advanced Engineering and Research and Development, Volume 4 / Special Issue 5 / P-ISSN: 2348-6406Dec.-2017.

[10] E. Park, Y. Cho, J. Han and S. J. Kwon, a global approach to user adoption of the Internet of Things in a smart home environment; An Ieee Internet Things J., vol. 4. Number 6PP. 2342-2350. December 2017

[11] R. K. Ritika, A. Brinivasulu et al., The latest comments on IOT threats and attacks: Classification, Challenges and solutions, sustainability, vol. 13. problem. 16.Page 94632021:2071-1050/13/16/9463

[12] IERC. (March 2015). European Internet of Things Research Group-The potential of Internet of Things-related activities in Europe. Connection: September. 20, 2017. [Online] Available for selection: http: / / www ... internetof-things-research.eu/pdf/.IERC_Position_Paper_IoT_Semantic_interoperability_final.pdf ...

[13] N. Gandhi L. J. Armstrong O. Petkar "Use of Bayesian networks to estimate the yield of sacred mango crops" was contacted in 2016.

[14] N. P. Sastra and D. m. Wiharta is used as a Udayana in environmental monitoring and as an application of the Internet of Things in the construction of the international campus of the university's smart campus. Smart green technology. elected. System information. (URSGTEIS), published in October 2016.85-88.

[15] C. Ravariu, N. Jalaja, A. Srinivasulu, "Simulating the attack of DNA viruses on logic

circuits with state-of-the-art conditions", Nanomaterials and Advanced Devices, vol.5.problem.2. Pages 666 to 674.

[16] Zanella, N. Bui, A. Castellani, L. Vangelista and M. zorzi, the Internet of Things for smart cities, ...IEEE Internet Objects J., vol. 1.No 2014february 22-32

[17] L. Sanchez et al. Smart Santander: Smart city test IOT test on the workbench, ice computer. network. volume. 1: 61 pages. March 217-238. 2014.

[18] S. Dahikar S. Rode "Prediction of the yield of agricultural Mango crops using Artificial Neural Network methods" International Journal of Research on Innovation in Electrical and Electronic Instrumentation and Control Engineering vol. 2. l Pages 683 to 6862014.

PALI IMAGE TO TEXT RECOGNITION SYSTEM USING LSTM TESSERACT

April Su [1], Yu Ya Win [2], and San San Tint[3]

[1,2] Faculty of Computer Science, University of Computer Studies (Mandalay)

[3] Pro-Rector, University of Computer Studies (Mandalay)

[1]nwayapril97@gmail.com, [2]yuyawin@gmail.com, [3]dr.sansantint@gmail.com

ABSTRACT

The electronic translation of images of printed, or typewritten text into machine-encoded text is known as optical character recognition (OCR) [8]. It is the simplest way to digitize written and printed materials so that they may be utilized for a variety of processes including text mining and language translation. Numerous character recognition systems have been developed for various languages; nonetheless, the difficulty of comprehending, maintaining, and retrieving old scripts has resulted in a barrier to knowledge discovery and access. This obstacle can be overcome with OCR techniques, which make it possible to turn various scanned texts into editable data. The first part of this paper presents the overview of the architecture of Tesseract. In the later sections, this paper will present about the training process of the pali images to the system. In the training process, three main stages are performed (i) created box files, (ii) created trained data files and (iii) training using LSTM tesseract. And then, testing process is implemented with three steps: (i) preprocessing, (ii) building box and (iii) classification using tesseract engine.

KEYWORDS

Pali, Optical Character Recognition, Tesseract engine, LSTM

1. INTRODUCTION

The technology known as optical character recognition is used to turn handwritten, transcribed, or printed text documents—including scanned pages and photos from phones and cameras—into editable and reusable text data [1]. Put otherwise, optical character recognition, or OCR, is a method that uses a picture of a written document (thus the term "optical") to identify various alphabets, numbers, and other characters. The characters are extracted from the image using a subprocess known as character recognition, and they are subsequently translated into text phrases for additional usage.

Generally, the process of OCR has three stages, that are: Access (Scan) the image document, Recognize the text data and then save it into any convenient format or display it directly to the user for further use [3].

The suggested approach identifies the text from the image and takes the Pali manuscripts as input for OCR. It is used in the open-source optical character recognition program Tesseract. The most recent Tesseract version, which supports deep neural network models, is used in this work [6]. This command-line program allows you to generate models for new scripts and languages and comes with pre-built models. Tesseract does not include Pali script OCR, so in order for the tool to recognize the Pali text, it must be educated for Pali script.

The following steps are necessary to develop the entire OCR application for Pali scripts. (i) Preparing training data, (ii) Pre-processing the document image, (iii) Performing recognition using Tesseract OCR, (iv) post-processing the generated output text.

This paper will present about the custom training Tesseract for Pali text. There are a variety of reasons that might not good quality output. This approach will attempt to train high quality OCR models [10].

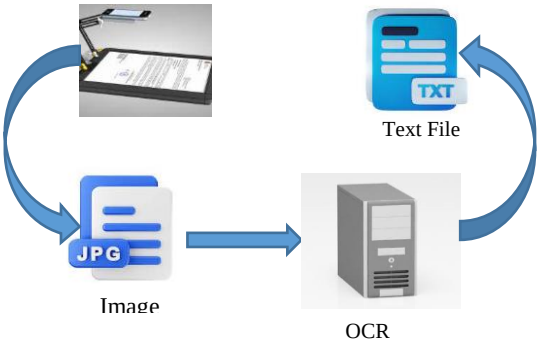


Figure 1: Overview of the System

2. RELATED WORK

The development of OCR and related procedures for different scripts has been the subject of numerous works. To increase the effectiveness of recognition, the majority of research projects concentrate on resolving problems that arise throughout the OCR process. The previous heuristic binary method employed the length of runs and the number of ones in the image to recognize the script into text [1]. Each character in this is transformed into a binary array of ones and zeros, where a non-blank is represented by a one and a blank by a zero. Subsequently, the binary array is converted into two-string pairs. Strings are derived for every written character and included into a lexicon. Prior to feeding the test input characters into the tool for recognition, they are transformed into a binary array. However, human intervention is required for the classification and prediction.

Chandrasekaran, M.; Siromoney, G.; and Chandrasekaran, R. Brahmi script recognition by machine. 1983, 7(4), 648–54; doi: 10.1109/TSMC.1983.6313155. IEEE Transact. Syst., Man, Cybernetics More recent techniques, such the Hidden Markov Model (HMM) and Neural Networks, use improved character recognition to automatically classify and forecast characters. The OCRopus tool is used to create a dataset of synthetic text-line pictures for the Greek polytonic script. It employs a number of adjustable factors, including size, skew, threshold, and blur, to create artificially generated text-line images that closely mimic the scanning procedure. The Ethiopian script is recognized using one-dimensional bi-directional LSTM.

Simistira, F., Katsouros, V.; Liwicki, M.; Ul-Hassan, A.; Papavassiliou, V.; Gatos, B. LSM networks for the recognition of ancient Greek polytonic scripts. 766-770) in IEEE's Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), August 23–26, 2015. Before the input text-line images are trained using a 1D LSTM, text-line normalisation is applied as a pre-processing step for smoothing [4]. To correct the translation in the script's characters' vertical axis, the center normalization method is used in this instance. It functions flawlessly regardless of variations in the image's height.

Liu, C.M., Addis, D., and Ta, V.D. LSTM networks for the recognition of printed Ethiopian script. IEEE, 28-30 June 2018, International Conference on System Science and Engineering (ICSSE), pp. 1-6. Character position with respect to the baseline is a significant characteristic of printed Latin scripts. Due to the fact that the majority of character pairs are comparable and are primarily differentiated by their size and position in relation to the baseline. Therefore, it is discovered that geometric normalisation2 is more trustworthy than textline normalization.

Mencis, R., Sabatelli, M., and Shkarupa, Y. LSTM recurrent neural networks for offline handwriting recognition. November 2016, Amsterdam, The Netherlands, 28th Benelux conference on artificial intelligence, https://www.researchgate.net/publication/31172974_Offline_Handwriting_Recognitio_Using_LSTM_Recurrent_Neural_Networks. Following optical character recognition, a dimensionality reduction technique called feature extraction is employed to extract the salient aspects of the image as a reduced feature vector. Both geometric and zone-based approaches were used to extract information from Tamizhi characters.

Ancient Indian scripts image pre-processing and dimensionality reduction for feature extraction and classification: A survey by Tomar, A., Choudhary, M., and Yerpude, A.

(2015) 21(2), 101–24 in Int. J. Comput. Trends Technol. (IJCTT). Zone-based techniques categorize Tamizhi script characters more effectively by dividing them into multiple zones. However, the geometric method classified characters based on the kind of features. Zone-based approaches are unable to fully capture the character's characteristics since Tamizhi inscriptions are brief and precise. In contrast, geometric approaches use bifurcation points, circles, semicircles, corner points, intersect points, and terminating points to extract features in greater detail3.

Bhattacharya, U., Parui, S., and Ghosh, S.K. A two-phase system for handwritten Tamil character recognition. In Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007), held in Curitiba, Brazil, September 23–26, 2007, pp. 511–515. IEEE. An method that employs directional cues to identify evolved scripts was tried for Sinhala and Ethiopic scripts because numerous scripts have evolved from the early scripts.

Bigun, J.; Assabie, Y.; and Premaratne, L. Direction tensors are used to recognize modification-based scripts. Proceedings of the 4th Indian Conference on Computer Vision, Graphics, and Image, Kolkata, India, December 16–18, 2004. Because OCR stages are done at the abstract level, the same algorithms can be expanded to support different scripts with minimal changes.

3. PALI

The language of the Indian subcontinent is called Pali. Pali is the language that has been studied in great detail because it was used to compile and record the Three Pitakas, which are priceless Buddhist texts. Theravada Buddhists used the term Pali. The stone pillars of King Ashoka from the Gujarat region of western India are the closest known historical evidence of the Pali language. long though the Pali faith arrived in Myanmar before to Anawrahta Minsaw's reign, Lord Arahman helped it continue to grow long after her rule. From that point on, the Pali faith was strongly promoted by the Buddhist monarchs of Myanmar. For this reason, throughout the Bagan period in Myanmar, Pali was studied by all. Pali is exclusively taught at colleges and monasteries these days. Furthermore, Pali words have been assiduously integrated into Myanmar language thus far [7].

The number of letters used in Pali is (41). These (41) letters are divided into two types: (8) vowels and (33) consonants.
The 8 vowels are:

Vowels							
အ (a)	အာ (a ̄)	အိ (i)	အီ (i ̄)	အု (u)	အူ (u ̄)	အေ (e)	အို (o)

Consonants are divided into two types: Wag consonants and A-Wag consonants. 'Wag consonants' are only consonants grouped together. 'A-Wag Consonant' means consonants that are free from the group. There are 25 letters in 'Wag consonants' and 8 letters in 'A-Wag consonants'. [7]

Wag Consonants				
က (ka)	ခ (kha)	ဂ (ga)	ဃ (gha)	င (ṅa)
စ (ca)	ဆ (cha)	ဇ (ja)	ဈ (jha)	ဉ (ña)
ဋ (ṭa)	ဌ (ṭha)	ဍ (ḍa)	ဎ (ḍha)	ဏ (ṇa)
တ (ta)	ထ (tha)	ဒ (da)	ဓ (dha)	န (na)
ပ (pa)	ဖ (pha)	ဗ (ba)	ဘ (bha)	မ (ma)

A-Wag Consonants			
ယ (ya)	ရ (ra)	လ (la)	ဝ (va)
သ (sa)	ဟ (ha)	ဠ (la.)	အံ (m)

4. PROPOSED SYSTEM

The three primary steps of the training process in the suggested system are to build box files for training images, trained data files, and training datasets using LSTM

Tesseract. There are three steps in the testing procedure. They are used the Tesseract engine for preprocessing, segmentation, and classification. The system design as shown in Fig. 2.

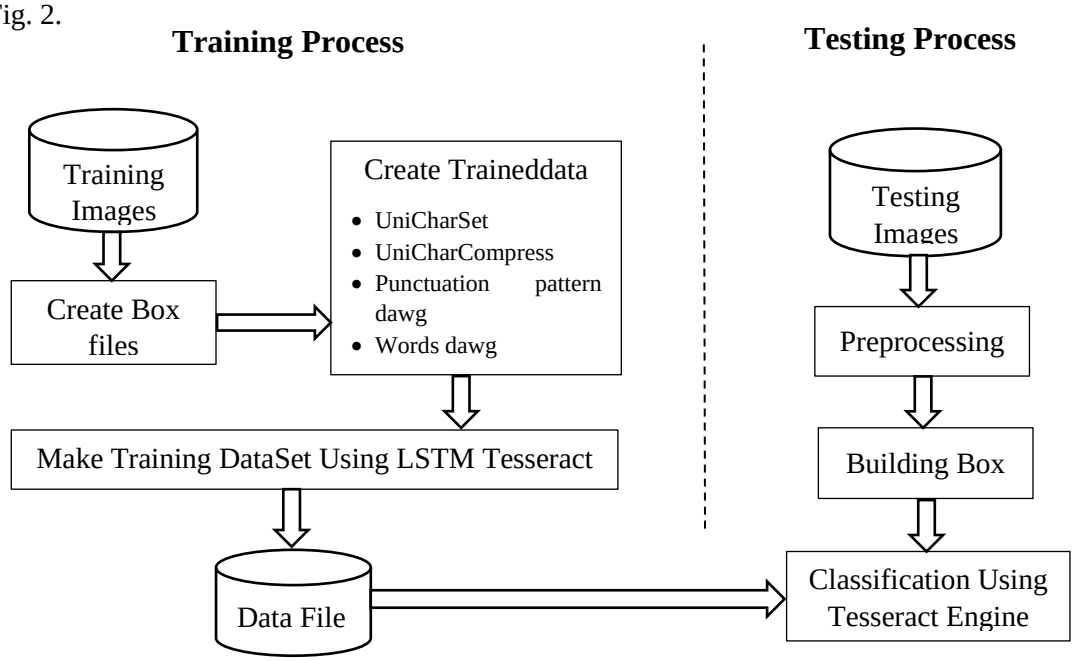


Figure 2: System Design of the Pali OCR System

4.1 Create Box File

For training process, a 'box' file creation is needed to go with each training image. The box file is a text file that lists the characters in the training image, in order, one per line, with the coordinates of the bounding box around the image. In this state, the system used online boxing tool called tesseract web box editor. As an example, the created box file of a pali word "သမ္ဗုဒ္ဓ" is shown in Table 4.1. The output box files are accepted for LSTM training. The coordinate system used in the box file has (0,0) at the *bottom-left***.** The last number on each line is the page number (0-based) of that character in the multi-page box file.

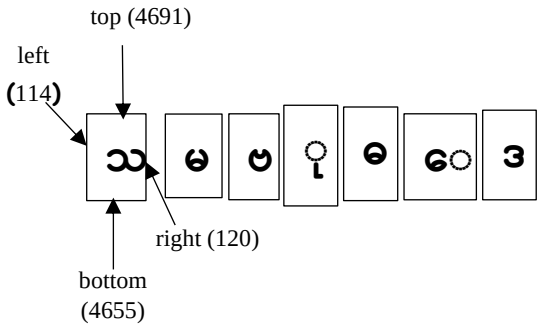


Figure 3: Example of bounding box around image

Table 4.1: Coordinates of each character in Boxing

Char	Coordinates (x1 y1 x2 y2)
သ	114 4655 120 4691 0
မ	127 4655 150 4682 0
္	152 4655 169 4692 0
ဗ	168 4654 193 4682 0
ု	197 4654 213 4681 0
ဓ	214 4654 250 4681 0
္	255 4654 280 4681 0

4.2 Create Trained data

Everything else required for the finished LSTM model is gathered in the trained data file. In contrast to base/legacy Tesseract, a trained data file must be set up beforehand and is provided during training. Directed Acyclic Word Graph (DAWG) may be present, and the other file is a simple UTF-8 text file:

- Config file providing control parameters.
- **Unicharset** defining the character set.
- **Unicharcompress,** aka the recoder, which maps the unicharset further to the codes actually used by the neural network recognizer.
- Punctuation pattern dawg, with patterns of punctuation allowed around words.
- Word dawg. The system word-list language model.
- Number dawg, with patterns of numbers that are allowed.

While some are required, such as unicharset and unichar compress, others are not. However, the punctuation dawg needs to be provided as well, if any of the other dawgs are.

In this system, generating the unicharset file is done in two steps using these tools: unicharset_extractor and set_unicharset_properties. The unicharset data file is produced by using the unicharset_extractor tool on the generated box files. The set_unicharset_ properties allow the addition of extra properties in the unicharset.

The system should get a unicharset file with all the fields set to the correct values after executing unicharset_extractor and set_unicharset_properties. Information about every symbol (unichar) that the Tesseract OCR engine has been trained to recognize is contained in the unicharset file. The following space-separated fields should be present on each unichar line in the unicharset file: character attributes, glyph metrics, script, other case, direction, mirror, and normed form.

Because it is preferable to recode each symbol as a brief series of codes from a limited number of classes rather than having a huge set of classes for complicated scripts like Indic, Myanmar, and Khmer scripts, unicharset compression recoding is crucial for the training process.

Every Myanmar character is encoded using a variable-length sequence of its unicode constituents. Figure 4 provides an example of a unicharset and the matching Unicode.

Table 4.2: Unicharset compression recoding

UniCharSet	Unicode
ဿ 1 64,64,254,255,327,359,20,21,369,402 Myanmar 31 0 31 ဿ	ဿ [101e]x
ဖ 1 64,64,254,255,190,193,20,21,232,237 Myanmar 26 0 26 ဖ	ဖ [1019]x
ဗ 1 64,64,254,255,190,193,20,21,232,237 Myanmar 24 0 24 ဗ	ဗ [1017]x
ၯ 0 0,0,255,255,215,244,14,15,59,80 Myanmar 46 17 46 ၯ	ၯ [102f]x
ဒ 1 64,64,254,255,171,172,20,21,214,215 Myanmar 19 0 19 ဒ	ဒ [1012]x
ဧ 0 58,59,255,255,443,449,20,21,232,234 Myanmar 48 0 48 ဧ	ဧ [1031]x
ဓ 1 64,64,254,255,187,190,20,21,229,234 Myanmar 20 0 20 ဓ	ဓ [1013]x

4.3 Training DataSet Using LSTM Tesseract

For character recognition, a long-term short-term memory network (LSTM) is used. In LSTMs, data is kept in a gated cell away from the regular flow of the recurrent network. A computer's memory can be used to write to, read from, or store data. To allow for writing, reading, and erasing, the cell decides what to keep and when to open and close gates [1]. Long-term data retention is made possible by the LSTM architecture, which also keeps gradients from disappearing or blowing up. The LSTM can add or remove data from the cell state, which is meticulously managed by structures known as gates. Information can be selectively allowed to travel through using gates. A sigmoid neural net layer and a point wise multiplication operation make them up.

More training text and more pages are better because neural networks require training on similar material and do not generalize as

well. If the target domain is highly confined, then perhaps none of the terrifying warnings about the requirement for a vast amount of training data apply; however, the network definition might need to be adjusted. Thus, around 400000 textlines have been learned for the pali language in this system. While LSTMs perform admirably when learning sequences, they become quite slow when there are too many states. For the complex scripts (Khmer, Indic, and Myanmar), it is preferable to recode each symbol as a short sequence of codes from a limited number of classes rather than having a large set of classes, according to empirical results that imply it is better to ask an LSTM to learn a long sequence rather than a short sequence of many classes.

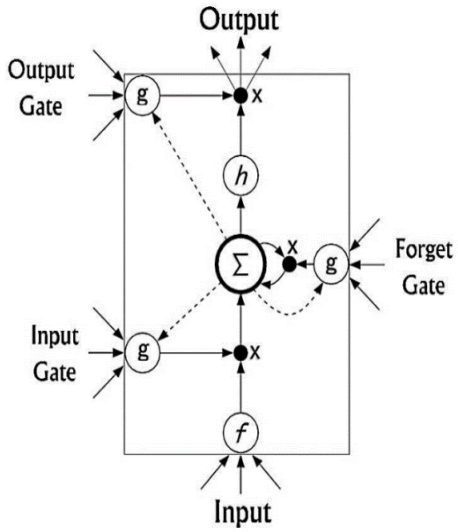


Figure 4: LSTM memory Cell - Gate functions are labeled as g, input activation as f, output activation as h, multiplication nodes as x and the core node, which realizes a summation, as ∑

The system starts to train model using LSTM Tesseract after creating the traineddata file successfully. Firstly, the repository tesseract-ocr/tesstrain-win on github is downloaded which could implement LSTM training. It is powerful, simple and easy to use for training process. The tesstrain-win project has two folders, data and old-ocrd. In the data file, ground truth data, unicharset, and other optional data file such as punctuations, numbers, wordlist [5]. The ocrd is the training workflow for Tesseract as a Makefile for dependency tracking and building the required software from source. And then, the input data started to train. For training purpose, the system has to run bunch of commands in window. The commands step by step are given below.

In the first step, name the traineddata file when the system run the make training.

```
make training MODEL_NAME foo
```

The second step, run the command prompt as an administrator and enter the path where tesstrain-win is located.

```
cd %USERPROFILE%/tesstrain-win
```

After that, run make training. The training–

```
make training -- trace
```

trace can output every command in makefile on the command line. it easy to locate the location and cause of the error if an error occurs.

After executing the command, how the system training goes can be seen.

The system takes about 8 hours that the training time of input data completed in win10 64-bit 8G PC. After implement the

```
2 Percent improvement time=1059, best error was 3.1 @ 5121
At iteration 6180/10000/10000, Mean rms=0.487%, delta=0.265%, char train=0.924%, word train=3.421%, skip ratio=0%, New best char error = 0.924 wrote best model:data/checkpoints/foo0.924_6180.checkpoint wrote checkpoint.
Finished! Error rate = 0.924
Makefile:158: update target 'data/foo.traineddata' due to: data/checkpoints/foo_checkpoint
lstmtraining \
--stop_training \
--continue_from data/checkpoints/foo_checkpoint \
--traineddata data/foo/foo.traineddata \
--model_output data/foo.traineddata
Loaded file data/checkpoints/foo_checkpoint, unpacking...
```

commands first time, the error rate is 92% in this system. So, the system has to run this command over and over to reduce the error rate in training data.

4.4 Testing Process

4.4.1. Preprocessing

OpenCV offers preprocessing functions that are predefined. The functions below are what we're utilizing to accomplish the same:

Grayscale Image: To make an image consist exclusively of black and white, apply a grayscale filter to it.

cv2.cvtColor(img_name,cv2.COLOR_BG R2GRAY)

Noise Removal: Eliminate background noise and refine the picture to prepare it for additional processing.

cv2.medianBlur(img_name,5)

Thresholding: the image is applied to make threshold and create binary image.

cv2.threshold(img_name, 0, 255, cv2.THRESH_BINARY + cv2.THRESH_OTSU)

Dilation: Add boundaries to the objects in the image, so the text stands out.

processor_var = np.ones((5,5),np.uint8)
cv2.dilate(img_name, processor_var, iterations = 3)

Erosion: Remove the boundaries after the dilation in the image, to unify the objects.

processor_var = np.ones((5,5),np.uint8)
cv2.erode(img_name, processor_var, iterations = 3)

4.4.2. Building Boxes around Text

Following preprocessing, we are identifying the letters in the text and enclosing it in boxes so the user can comprehend how OCR functions and easily edit the text if necessary. The boxes surrounding the text are made using the following snippet.

```
hi, wi, ci = img_name.shape boxes =
pytesseract.image_to_boxes(img_name)
 for box in boxes.splitlines():
   box = box.split(' ')
   img_name =
     cv2.rectangle(img_name,
    (int(box[1]), hi - int(box[2])),
    (int(box[3]), hi - int(box[4])),
    (0, 255, 0), 2) cv2.imshow
    ('image', img_name)
```

4.4.3. Performing OCR and displaying Text

Tesseract has method to print text from image that is:

text = pytesseract.image_to_text(image)

5. Performance of Ocr for Different Pali Document Images

In this system, LSTM tesseract is validated using new printed documents images dataset and generates a 27% loss and 97% accuracy at the iteration of 100. From screenshots images dataset with font size 6 to 12, the system is validated as 29% loss and 96% accuracy at the same iteration. But the loss and accuracy changed to 31% and 91% when the screenshots images font size is 14 to 20. For old scanned documents dataset, loss is 33% and accuracy is 89% at 100 iterations. The low-quality documents dataset is validated and the result is 35% loss and 81% accuracy. The generated results are represented in Fig 5 and Fig 6.

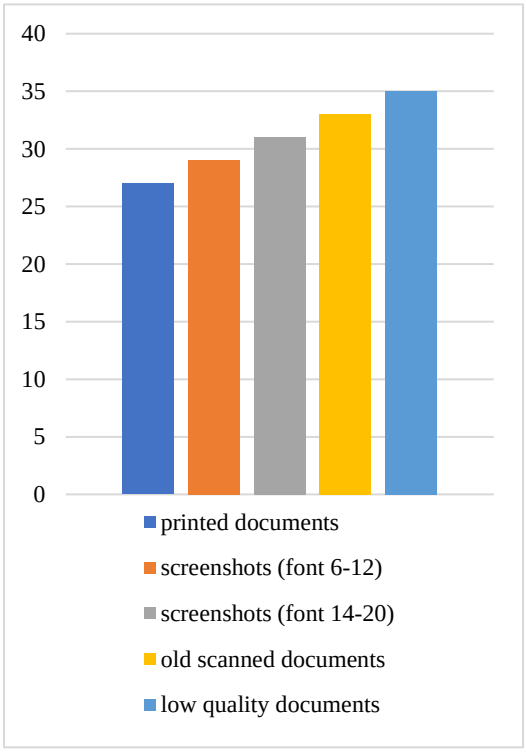


Figure 5: Validation Loss graph for Five Dataset

Table 1. Performance of the system using five difference datasets

Types of Pali document images	Number of images tested	Accuracy (%)
Old Scanned documents	58	88.54
Low quality documents	25	80.84
New Printed Documents	103	97.35
Screenshots Images – Font size 6 to 12	49	95.49
Screenshots Images– Font size 14 to 20	53	91.24
Average accuracy		90.57

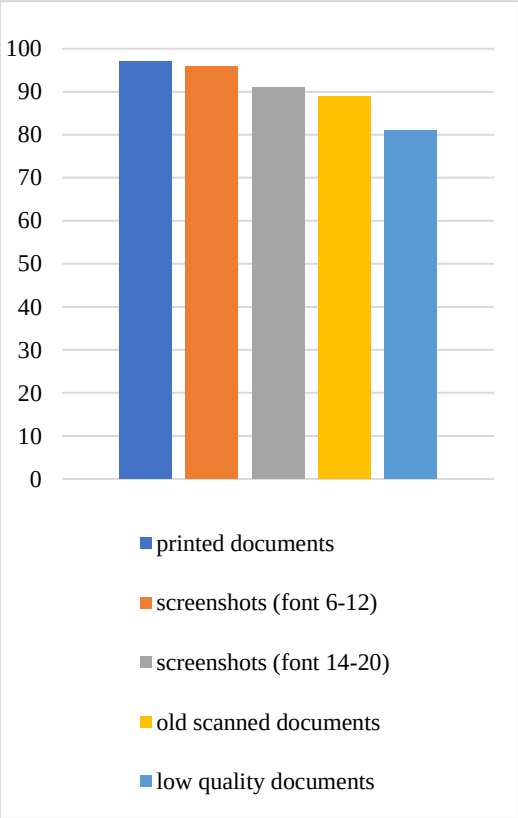


Figure 6. Validation Accuracy graph for Five Dataset

In the following table shows the summarized experimental results of the system on the five differences dataset. The experimental results show that the proposed system outperforms on the new printed documents images dataset than other datasets.

6. CONCLUSIONS

Accuracy is one of OCR's primary problems. Because OCR systems are not flawless, they frequently require correction when recognizing text, particularly scripts. The input Pali pictures used for the model's training and testing are gathered from a variety of sources, including [https://tipitaka.org/ymr/CSCD Tipitaka (Myanmar)]. The ground truth for the Pali images must be developed through a number of pre-processing stages [7].

The language model for the Pali script was created, and the LSTM networks were set up and trained. The accuracy of text recognition for languages with scripts is still low, and thus presents challenges. The majority of the time, noisy photos can produce extremely noisy outputs, hence the system must preprocess the images [3].

The aim of this system is to facilitate the automated extraction of desired information from a paper document and its subsequent utilization wherever required. This results in less work—or occasionally none at all—in the form of expensive data entry. The system also sought to make it possible for document processing to result in the elimination of human involvement, significantly cutting down on both process time and expense [4]. In this system, LSTM Tesseract is used for training process with three main stages. And then, five difference datasets are used for validation process.

In this system, various type of pali document images is used as testing data. The accuracy of old Scanned documents type image is 88.54% and the low-quality documents image is 80.84%. The accuracy of screenshots images with font size 6 to 12 is better than font size 14 to 20 screenshots images. The new printed documents images accuracy is 97.35%. Among then, the new printed documents images performance is the best. The average total accuracy of the

system is 91% that shows the good result for the system.

7. REFERENCES

[1] M. Jenckel, S.S. Bukhari and A.Dengel, "Transcription Free LSTM OCR Model Evaluation", 2018 16th International Conference on Frontiers in Handwriting Recognition, 2018 IEEE.

[2] Tesseract ocr, github: *https://github.com/tesseract-ocr.*

[3] Z.Z. Aung, C.M.M Maung, "Myanmar Optical Character Recognition using Block Definition and Featured Approach", 2017 3rd International Conference on Science in Information Technology (ICSITech), 2017 IEEE.

[4] N. Kotwal, A. Sheth, G. Unnithan and N. Kadaganchi, "Optical Character Recognition using Tesseract Engine", International Journal of Engineering Research & Technology (IJERT), Vol. 10 Issue 09, September-2021.

[5] S. Dome and A.P. Sathe, "Optical Charater Recognition using Tesseract and Classification", 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021.

[6] M. Chawla, R. Jain and P. Nagrath, "Implementation of Tesseract Algorithm to Extract Text from Different Images", International Conference on Innovative Computing and Communication (ICICC 2020).

[7] Basic Pali grammar, U Myint Swe (Master of Arts - London) (http://www. kbrl.gov.mm/ book/download/0003 17)

[8] Wikipedia. Optical character recognition. https://en.wikipedia.org/ wiki/Optical_character_recognition.

[9] E. Sabir, S. Rawls and P. Natarajan, "Implicit Language Model in LSTM for OCR", arXiv:1805.09441v1 [cs.CV] 23 May 2018.

[10] R. Islam and N. Ferdaoush, "An Open-Source Tesseract Based Optical Character Recognizer for Bengali Language", Ieeexplore.ieee.org, 2022. [Online].

BREAST CANCER PREDICTIVE ANALYTICS SYSTEM

May Phyo Thu[1], Khin Phyo Wai[2], and Min Min Su[3]

[1,2,3] Faculty of Computer Science, University of Computer Studies (Mandalay)
[1]mayphyothu.mpt1@gmail.com, [2]khin.wai323@gmail.com,
[3]minminsu76@gmail.com

ABSTRACT

In the current landscape, big data has gained immense popularity, particularly in healthcare applications. Breast cancer, the most prevalent cancer among women globally, necessitates early detection for improved treatment outcomes. This system proposes a scalable and fault-tolerant pipeline model for analyzing breast cancer data and predicting the likelihood of breast cancer occurrence. The implementation is carried out on the Apache Spark infrastructure, utilizing the random forest algorithm, with the UCI breast cancer dataset serving as the input data source. The system's efficacy is assessed by comparing the performance of the random forest algorithm with the naïve Bayes algorithm in terms of accuracy, precision, recall, and F-measure. The evaluation results reveal the success of the proposed system, achieving an accuracy rate of 98.2% within the big data analytics environment.

KEYWORDS

Apache Spark, Random Forest, Naïve Bayes, Big Data

1. INTRODUCTION

In the contemporary era, the rapid advancement of technology has led to a substantial increase in data rates, resulting in the emergence of big data characterized by voluminous and diverse datasets. The proliferation of data occurs in various formats, presenting a multifaceted landscape of information. While data storage has become more affordable, the escalating volume of data from diverse sources and formats introduces new challenges in terms of data processing [1].

Breast cancer is a prevalent cancer type originating from the breast tissue, notably common in Western countries. Improving survival rates and reducing fatalities associated with this disease hinge on early detection.

Breast cancer is characterized by two main types of cells: (a) cancer cells, classified as malignant, and (b) non-cancer cells, identified as benign. The prediction of breast cancer involves discerning the malignant characteristics exhibited by breast cells. Pathologists employ diverse methods for breast cancer prediction, and researchers contribute to this effort by utilizing various machine learning and statistical methods to enhance the accuracy and efficiency of breast cancer forecasting.

Big data analytics has garnered increased attention for its capability to analyze and manage vast amounts of raw data efficiently. The MapReduce framework, initially developed by Google [2], plays a crucial role in enabling the parallelization and distribution of various tasks across different clusters. This framework is particularly well-suited for batch data operations on diverse data nodes.

In implementing this approach, Apache Hadoop has been recognized as a robust framework [3]. Conversely, Apache Spark has emerged as a compelling alternative, providing enhanced appeal and high-performance capabilities. Building upon Hadoop's foundation, Apache Spark extends its capabilities and facilitates real-time stream data processing, making it a versatile and powerful choice for big data analytics [4].

This paper presents the development of a breast cancer prediction model utilizing

Apache Spark-based random forest. The system is implemented within the framework of big data analytics, specifically leveraging the capabilities of Apache Spark to analyze a high volume and velocity of breast cancer data. The input data source for the predictive model is the UCI breast cancer dataset.

The utilization of Apache Spark in this infrastructure brings forth features such as fault tolerance and a scalable memory operating engine, crucial for handling vast amounts of data efficiently. This enables the system to effectively implement machine learning approaches, making it well-suited for breast cancer prediction tasks within the realm of big data analytics.

The proposed system comprises four main modules: data collection, data pre-processing, data labeling, and classification. The development of these modules is organized into four layers: the data ingestion layer, storage layer, processing layer, and data analytics layer. In the data ingestion layer, breast cancer dataset collection is conducted, and the ingested data is stored in the storage layer using the Hadoop Distributed File System (HDFS). The processing layer utilizes Apache Spark for big data analysis, enabling breast cancer data prediction analysis in a distributed computing environment with reliability and fault-tolerance.

For constructing prediction models, off-line training is carried out in the data analytics layer, involving data cleaning, pre-processing, and model generation. This systematic approach ensures a well-organized and comprehensive framework for breast cancer prediction, leveraging distributed computing and big data analytics capabilities across multiple layers of the system.

In this paper, Section 2 provides an overview of related works pertaining to the proposed system, offering insights into existing research and developments in the field. Following this, Section 3 delves into the background theory that forms the foundation of the proposed system, providing readers with the necessary theoretical context.

Moving forward, Section 4 presents the details of the proposed system, elucidating its key components, methodologies, and design considerations. The experimental results of the proposed system are then thoroughly examined and discussed in Section 5, shedding light on the system's performance and outcomes. Finally, Section 6 concludes the presentation of the system.

2. RELATED WORKS

The authors of this study introduced a breast cancer prediction model utilizing grid computing and support vector machine (SVM) [5]. Initially, they conducted breast cancer prediction using a support vector machine without grid computing. Subsequently, they employed grid computing to predict breast cancer with the support vector machine model. The analysis results from both approaches were then compared, leading to the construction of a new and enhanced prediction model. The newly developed model incorporated grid computing on the data before fitting it for prediction, resulting in improved prediction outcomes.

In their analysis, the authors explored a hybrid approach that combines fuzzy logic with a traditional decision tree for breast cancer classification, utilizing the SEER dataset [6]. The comparison of the two algorithms was conducted using three performance evaluation parameters: sensitivity, accuracy, and specificity. The results of the performance evaluation indicated a significant improvement in the performance of the fuzzy decision tree when compared to the traditional J8 decision tree.

The authors introduced a hybrid approach for breast cancer prediction in their work [7], consisting of two main parts. Firstly, they employed fuzzy-artificial immune system (FAIS) for dimensionality reduction. Subsequently, the breast cancer classification was conducted using k-nearest neighbor (KNN) based on the selected features. The experimental results demonstrated that the proposed approach led to an increase in prediction accuracy and a reduction in execution time.

In their work, the authors presented a disease prediction approach utilizing medical data features [8]. The paper applied a rough

approach employing backpropagation neural network for prediction, using datasets including UCI breast cancer, statlog heart disease, and hepatitis. The system evaluation revealed high accuracy levels for breast cancer (90.4%), heart disease (98.6%), and hepatitis (97.3%). Furthermore, the authors proposed both unsupervised and supervised approaches using the Weka tool for malignancy data prediction with the Naive Bayes algorithm [9]. The paper involved the computation of misclassification costs using Naive Bayes and compared these costs with those obtained through the application of other algorithms.

In their paper, the authors introduced a prediction algorithm for breast cancer recurrence using the SEER (Surveillance, Epidemiology, and End Results) dataset from the National Cancer Institute's (NCI) program, employing the MapReduce framework [10]. The paper commenced with pre-processing activities, encompassing data cleansing and preparation. Subsequently, the final dataset was constructed, and the analysis focused on predicting breast cancer recurrences within the initial 5 years after breast cancer treatment.

3. THEORICAL BACKGROUND

Big data analytics integrates traditional analytics with mining methods to handle large volumes of data, establishing a foundational infrastructure for analyzing, constructing models, and forecasting markets, behaviors, products, and services based on enterprise needs. There are three main types of big data analytics:

- **Descriptive Analytics:** This involves applying data mining and aggregation approaches to understand past events and outcomes. It answers the question, "What has happened?"

- **Prescriptive Analytics:** This type employs simulation and scalarization methods to provide advice on possible results and outcomes, aiming to answer the question, "What should we do to achieve certain results in the future?"

- **Predictive Analytics:** Using statistical and forecasting methods, predictive analytics focuses on anticipating future events and outcomes, addressing the question, "What could happen in the future?" This type of analytics is particularly concerned with forecasting and predicting future trends and behaviors based on historical data.

Predictive big data analytics involves methods for forecasting future outcomes based on historical and current data [11]. Leveraging statistical methods is a sound analytical approach for big data analysis, given that big data fundamentally involves the consideration and analysis of data within the framework of statistics.

3.1. Random Forest

Random Forest is a supervised ensemble learning algorithm that aggregates decision trees randomly. The core principle involves the construction of binary trees through recursive partitioning (RPART). The binary tree is iteratively built through the recursive division of nodes, ensuring that each division enhances the unity of the parent and son nodes. Multiple trees are constructed, with each tree built using a bootstrap instance.

Unlike CART, Random Forest introduces non-deterministic construction with two levels of randomization. It considers only a subset of features during each iteration. Random Forest shares parameters with decision trees, and its advantages include excellent performance on large-scale datasets, the ability to handle classification and regression tasks, and improved model robustness through the introduction of model randomness. Each tree is constructed using the following algorithm:

- Let N be the number of test cases, M is the number of variables in the classifier.

- Let me be the number of input variables to be used to determine the decision in a given node; m<M.

- Choose a training set for this tree and use the rest of the test cases to estimate the error.

- For each node of the tree, randomly choose m variables on which to base the decision. Calculate the best partition of the training set from the m variables.

For prediction a new case is pushed down the tree. Then it is assigned the label of the terminal node where it ends. This process is iterated by all the trees in the assembly, and the label that gets the most incidents is reported as the prediction. The step of Random Forest prediction is as follows:

Step 1: Read the input data set for "n" features.

Step 2: Choose subset of features and name it as "i" from "n" features randomly

Step 3: Compute the node "n" among "i" features base on finest fit

Step 4: Base on best split, divide the node into child nodes.

Step 5: Repeat 2-4 steps until got "j" nodes i.e., trees

Step 6: Repeat 2-5 step for "m" times to get "m" number of trees to create a forest.

Step 7: Test features and generated discussion trees are used for prediction. This is stored as target

Step 8: Calculate the generalized value called vote for each predicted target

Step 9. The high voted target is considered as final prediction.

This approach leverages the collective decisions of multiple trees to enhance the accuracy and reliability of the forecasting outcome.

3.2. Naïve Bayes

Naive Bayes is a popular classification algorithm that falls under the umbrella of probabilistic classifiers. It's widely used for solving classification problems, especially in natural language processing (NLP) and text-based applications. The algorithm is based on Bayes' theorem, which relates the conditional and marginal probabilities of random events. Naive Bayes leverages Bayes' theorem, which is expressed as follows:

$$P(y\backslash x_1, x_2, \ldots\ldots, x_n) = \frac{P(y) \times P(x_1\backslash y) \times P(x_2\backslash y) \times \ldots\ldots\ldots \times P(x_n\backslash y)}{P(x_1) \times P(x_2) \times \ldots\ldots\ldots \times P(x_n)} \quad (1)$$

In the context of classification, $P(y\backslash x_1, x_2, \ldots\ldots, x_n)$ is the probability of class: y, $P(y)$ is the prior probability of class: y. $P(x_i\backslash y)$ is the conditional probability of feature x_i. In practice, the evidence term is often dropped since it is constant for all classes.

Therefore, the classification decision is typically made by comparing the numerator terms:

$$P(y\backslash x_1, x_2, \ldots\ldots, x_n) \propto P(y) \times P(x_1\backslash y) \times P(x_2\backslash y) \times \ldots\ldots\ldots \times P(x_n\backslash y) \quad (2)$$

The class with the highest posterior probability given the features is then assigned as the predicted class.

This classifier firstly computes the prior probabilities of each class P(y) by counting the frequency of each class in the training data. For each feature x_i and class: y, it calculates the conditional probability $P(x_i\backslash y)$ which represents the likelihood of observing feature x_i given class: y.

This is typically done by calculating the frequency or probability distribution of each feature value within each class.

Then it calculates the posterior probability of each class: y using Bayes' theorem with equation 1. Since the denominator is constant for all classes, it can be ignored during classification. Finally, it selects the class with the highest posterior probability as the predicted class for the input instance x.

This classifier is a straightforward and powerful algorithm for the classification task. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of features in a learning problem. The classifier often performs much more complicated solutions.

3.3. Apache Spark

Apache Spark is an open-source cluster infrastructure built on top of the Hadoop Distributed File System (HDFS). It facilitates

data processing and distribution to worker nodes for efficient parallel processing. The master node oversees the worker nodes, managing the scheduling and dispatching of distributed tasks. This infrastructure relies on a cluster manager and a distributed storage system, offering an API that is developer-friendly and incorporates a high-speed, in-memory data engine.

Spark Core serves as the fundamental component and operational engine in Spark. It supports various functionalities such as scheduling, I/O operations, and the execution of distributed jobs.

The operations are carried out through an application programming interface (API) available in languages like Scala, Python, R, and Java. This API accommodates multiple languages through different interfaces and is renowned for its faster in-memory computing capabilities.

The driver program, which operates at a higher level, enables parallel operations by calling functions on Resilient Distributed Datasets (RDDs) in Spark. The scheduling of parallel operations across clusters is facilitated by the Spark Core. RDDs were the fundamental API until Spark version 2.1.x, characterized by their read-only and fault-tolerant dataset collections distributed across clusters for processing.

Spark Core encompasses four main elements: Spark Streaming, GraphX, Spark SQL, and the Machine Learning Library. Spark SQL is designed to handle structured data through the use of data frames, providing an interface for query data processing.

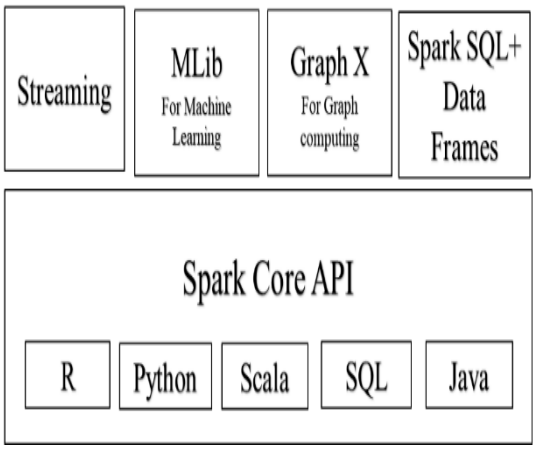


Figure 1. Apache Spark Ecosystem [12]

Spark Streaming extends Spark by incorporating real-time data processing alongside batch processing.

The Machine Learning Library within Spark Core supports various distributed machine learning approaches, including classification, collaborative filtering, clustering, dimensionality reduction, and regression. These approaches operate in a distributed manner, performing feature selection, extraction, and transformation on structured datasets. Additionally, the library offers tools for constructing and fine-tuning machine learning pipelines, facilitating the implementation, loading, and storage of models, algorithms, and pipelines [13].

4. PROPOSED SYSTEM

The main target of this system is to develop high performance and scalable Breast Cancer Prediction System on Big Data Analytics Platform, Apache Spark. The proposed system using apache spark architecture on is described in Figure 2.

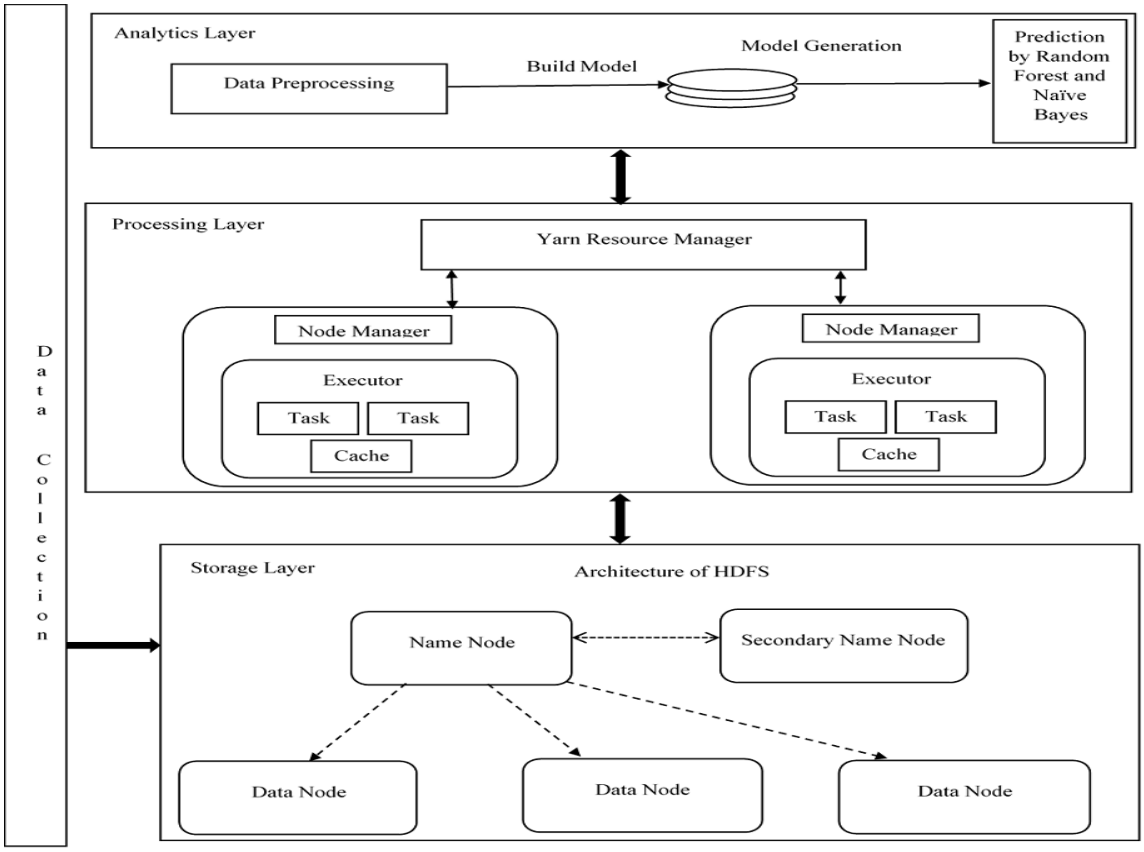


Figure 2. The Proposed System using Apache Spark Architecture

id	diagnosis	radius_mea	texture_mea	perimeter_r	area_mean	smoothness	compactnes	concavity_r	concave po	symmetry_r	fractal_dim
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078
846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338
84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682
84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077
848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922
84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356
849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395
8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766
8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811

Figure 3. Breast Cancer Dataset

4.1. Overview of the System

This system is implemented at four layers.

- Data Collection
- Data Analytics
- Data Processing
- Data Storage

Data Ingestion Layer - In this layer, breast cancer data is collected by Apache Flume. For offline processing, the collected data is pushed into HDFS sink.

Storage Layer - The collected data is stored in Hadoop File System (HDFS): distributed storage. HDFS provides a master / slave architecture and a single NameNode acts as the primary server. Name Node performs the operation of the file system namespace, such as opening, closing and renaming.

Processing Layer – Yarn Cluster Manager and Spark executives are in the processing layer to process large amounts of data in parallel on a product hardware cluster in a reliable and fault-tolerant manner.

Analytics Layer –To generate the prediction models, Data pre-processing, and model generation are performed at Off-line training. All of the processes from Analytics Layer are executed in distributed manner by using Spark engine. Breast Cancer prediction is implemented by using Spark MLlib.

4.2. Data Collection

Original cancer data set of UCI Machine learning repository [14] is used as input for analysis as shown in Figure 3. This dataset comprises tumor features extracted from digital images of breast fine needle aspirates (FNA). Each patient's record includes 10 attributes that characterize cell nuclei at the fine needle aspirate of a breast mass. These attributes encompass texture, radius, area, perimeter, compactness, smoothness, concavity, symmetry, fractal dimension, and concave points. The dataset classifies breast tumors into two types: Malignant and Benign. Malignant tumors represent cancerous conditions, while Benign tumors indicate non-cancerous conditions. The attribute information in the dataset are:

1. ID number

2. Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

a. radius (mean of distances from center to points on the perimeter)

b. texture (standard deviation of gray-scale values)

c. perimeter

d. area

e. smoothness (local variation in radius lengths)

f. compactness (perimeter^2 / area - 1.0)

g. concavity (severity of concave portions of the contour)

h. concave points (number of concave portions of the contour)

i. symmetry

j. fractal dimension ("coastline approximation" - 1)

4.3. Training Process

Apache Spark is a fault tolerant and scalable in memory computing engine for solving big data. Spark operates on distributed datasets called RDDs, which are partitioned across multiple nodes in a cluster. RDDs are fault-tolerant data structures that automatically recover from failures by recomputing lost partitions from lineage information. This resilience ensures that even if a node fails, the computation can continue seamlessly on other nodes without data loss. Spark's execution engine builds a Directed Acyclic Graph (DAG) of the computation, optimizing task execution for maximum efficiency. The DAG scheduler breaks down the computation into stages and tasks, enabling parallel execution and

pipelining of operations. This optimization minimizes data shuffling and maximizes resource utilization, leading to improved scalability. Spark employs lazy evaluation, where transformations on RDDs are only executed when an action is called. This optimization allows Spark to optimize the execution plan based on the entire pipeline, rather than executing each transformation immediately. Lazy evaluation reduces unnecessary computation and optimizes resource usage, making Spark pipelines more efficient and scalable. Spark allows users to explicitly set checkpoints in the pipeline to persist RDDs to durable storage. This checkpointing mechanism ensures fault tolerance by providing fault recovery points in the pipeline. Spark monitors the execution of tasks and automatically detects failures. Upon detecting a failure, Spark re-executes the failed task on another node using the lineage information of the RDDs. This fault recovery mechanism ensures that computations continue uninterrupted even in the presence of node failures, making Spark pipelines resilient and fault-tolerant. Overall, the combination of distributed processing, fault-tolerant RDDs, DAG execution, lazy evaluation, checkpointing, and fault recovery mechanisms makes Apache Spark pipelines scalable and fault-tolerant, suitable for handling large-scale data processing tasks in production environments.

It supports rich library for machine learning algorithms implementation with an effective manner. Pipeline is utilized for the chain of the continuous prediction process in Machine learning. The pipeline process consists of the following key steps:

- **Data Ingestion:** In this phase, input data loaded into spark system and converted into data frames.
- **Data Preparation:** Input data may not be proper and may contain missing values. In this phase, prepare the proper input data for the model by eliminating or predicting the missing values. Output data is used to build and train the model.
- **Train and build Model:** In this phase, features are added to data frames and transform them to predictions with data frames by training the model.
- **Predictions:** In this phase we evaluate the predictions from the model using train data and test data.

Data pre-processing is performed by substituting missing values from the input dataset. After data pre-processing, the data storage is done HDFS with the splitting into training and testing set. 80% of the dataset is used as the training and the remaining 20% is used as the testing. Random Forest and Naïve Bayes can be trained using the training set on local file system. To train Spark-based Random Forest and Naïve Bayes, the training set is needed to load into the Hadoop File System (HDFS). Random Forest training is performed on Hadoop with the loaded training set. After model training is finished, the generated trained model is saved in HDFS and then copied it into local file system to make performance evaluation.

4.4. Testing Process

After model training, the generated trained model is saved in HDFS and then the copy is sent to the local file system for performance evaluation. For performance implementation, the testing set from a random split of the dataset is done. After the testing data is loaded, the target class of testing data is predicted using random forest and naïve bayes models. The predictive models: random forest and naïve bayes will output the likelihood of breast cancer diagnosis or prognosis for each new patient based on their features.

The testing of the proposed system calculates the output values (prediction values :0 for benign, 1 for malignant) by using the trained models: Random Forest and Naïve Bayes.
The proposed system' performance evaluation will measure by accuracy, precision, recall, and f-measure. The details of performance evaluation are described in the next section.
The Steps of Prediction Analysis using Random Forest and Naïve Bayes in Spark are:

- Loading pre-processed dataset by RDD format

- Transforming RDD into data frame
- Reading labels and features from data frame
- Encoding the non-numeric features
- Indexing string with each encoded feature
- Aggregation of vector by numeric and one-hot-encoded features
- Converting the aggregated vector to a pipeline
- Converting the pipeline to a readable spark form
- Training the model by random forest-based features with the training data
- Testing on the whole data to achieve forecasting prediction label value

5. PERFORMANCE EVALUATION

In this system, breast cancer dataset from UCI is used. This system is implemented with dataset size 600,000 breast cancer data. This system is tested with data size ratio: (Training – 80%, Testing – 20%). Firstly, randomly selected 80% records as training data and 20 % records for testing from this dataset. The popular performance measures (accuracy, recall, f-measure, and precision) are evaluated for this proposed system analysis. The accuracy, precision, recall, and f-measure are described by:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (1)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

The experimental setup is illustrated in Table 1.

Table 1. System Specification

Operating System	Ubuntu 20.04 LTS
Host Specification	Intel ® Core i7-8550U CPU @ 3.7GHz, 8GB Memory, 1TB Hard Disk
VMs Specification	6 GB RAM, 200 GB Hard Disk
Software Component	- Hadoop 3.2.2 - Flume 1.9.0 - Spark 3.1.2 -Scala 2.12.3

The comparative results for the performance of proposed random forest-based classifier and naïve bayes classification are illustrated in Figure 4.

Machine Learning Algorithm	Benign			Malignant		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Random Forest	0.99	0.98	0.98	0.96	0.99	0.97
Naïve Bayes	0.91	0.94	0.92	0.89	0.84	0.86

Figure 4. Performance comparisons of Random Forest and Naïve Bayes

The comparison of accuracy between Random Forest and Naïve Bayes on breast cancer dataset is illustrated in Figure 5 and Figure 6.

Machine Learning Algorithm	Accuracy (%)
Random Forest	98.1
Naïve Bayes	90.5

Figure 5. Accuracy Results of Random Forest and Naïve Bayes

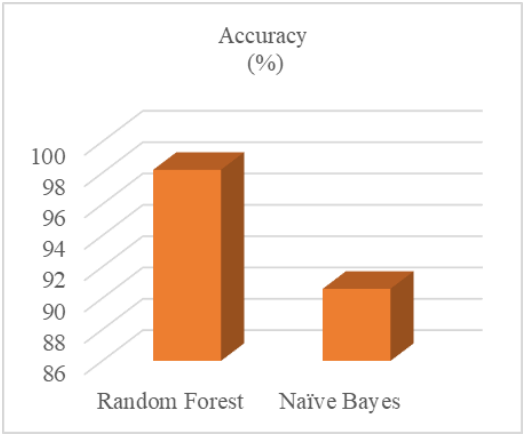


Figure 6. Accuracy comparisons of Random Forest and Naïve Bayes

According to the evaluation results, Random Forest classifier achieves the accuracy by 98.1% and Naïve Bayes classifier provides the accuracy by 90.5%. Therefore, the proposed breast cancer classification model with Random Forest achieves the best optimal accuracy than Naïve Bayes classifier.

6. CONCLUSIONS

In this system, the proposed breast cancer prediction model is designed to be adaptable to classical machine learning algorithms, providing a generalizable and efficient forecasting method for dealing with large-scale data. The primary objective of this breast cancer prediction system is to enhance the survivability rate among patients with breast cancer. The system utilizes a Spark-based approach employing the Random Forest algorithm for breast cancer prediction. The proposed approach is rigorously evaluated and compared using the Wisconsin Breast Cancer dataset. This dataset is well-known for its comprehensive breast cancer diagnostic features. The proposed system offers several advantages over traditional machine learning approaches. It specifically addresses and reduces computational complexity and response times by leveraging the big data analytics platform, Apache Spark. In the system implementation, this research achieves the accuracy by 98.1% with Random Forest, and 90.5% with Naïve Bayes. Experimental outcomes demonstrate the superiority of the Random Forest classifier in key performance metrics, including accuracy, recall, precision, and F-measure.

REFERENCES

[1] M. Lněnička, R. Máchová, J. Komárková, and I. Čermáková, "Components of Big Data Analytics for Strategic Management of Enterprise Architecture", *2nd International Conference on Strategic Management*, VSB-Technical University of Ostrava, Ostrava, Czech Republic, 25-26 May, 2017, pp.398-406.

[2] J. Dean, and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," *in Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation*, San Francisco, CA, USA, 6–8 December, 2004, pp. 137-150.

[3] "Welcome to Apache Hadoop!", Available at "http://hadoop.apache.org/." [Online; accessed 28-December- 2017].

[4] "Spark-3.1.2", Available at https://spark.apache.org/docs/3.1.2/ml-guide.html." [Online; accessed -May-2021].

[5] V. Deshwal, and M. Sharma, "Breast Cancer Detection using SVM Classifier with Grid Search Technique", *International Journal of Computer Applications (0975 – 8887)*, Volume 178 – No. 31, Foundation of Computer Science, USA, July 2019, pp. 18-23.

[6] M. J. Domingo, B. D. Gerardo, and R. P. Medina, "Fuzzy Decision Tree for Breast Cancer Prediction*", In Proceedings of 2019 International Conference on Advanced Information Science and System (AISS'19)*, ACM, Singapore, November 15 – 17, 2019, pp. 1-6.

[7] S. Sahana, K. Polat, H. Kodaz, and S. Günes, "A new hybrid method based on fuzzy-artificial immune system and knn algorithm for breast cancer diagnosis", *Computers in Biology and Medicine*, Elsevier, USA, Volume 37, Issue 3, March 2007, pp. 415-423.

[8] K. Nahato, K. Harichandran, and K. Arputharaj, "Knowledge Mining from Clinical Datasets Using Rough Sets and Backpropagation Neural Network*", Computational and Mathematical Methods in Medicine*, Hindawi Limited, United Kingdom, 04 March, 2015, pp.1-13.

[9] T. A. Shaikh, and R. Ali, "A CAD Tool for Breast Cancer Prediction using Naive Bayes Classifier", *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, Pune, India, 12-14 March, 2020, pp. 351-356.

[10] H. Thottathy, K. K. Pavan, and R. P. Panchadula, "Microarray Breast Cancer Data

Clustering Using Map Reduce Based K-Means Algorithm", *Revue d'Intelligence Artificielle (RIA)*, International Information and Engineering Technology Association (IIETA), Vol. 34, No. 6, 31 December, 2020, pp. 763-769.

[11] E. Ricciarddelli, A. Cersosimo, D. Cimini, and F. D Paola, "Analysis of Heavy Rainfall Events Occurred in Italy By Using Numerical Weather Prediction, Microwave and Infrared Technique", *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, Valencia, Spain, 22-27 July, 2018, pp.931-934.

[12] What is Apache Spark? https://databricks.com/spark/about", 2018.

[13] MLlib: Main guide Spark 2.3.0. Documentation https://spark.apache.org/docs/2.3.0/ml-guide.html", 2018.

[14]https://archive.ics.uci.edu/ml/datasets/breast+cancer

Authors

Biography

Dr. May Phyo Thu achieved the M. C. Sc degree in 2010 from the University of Computer Studies (Pathein) and Ph. D (IT) degree in 2019 from the University of Computer Studies, Yangon. Now, I am working as a lecturer at the Faculty of Computer Science at the University of Computer Studies (Mandalay).

TRAFFIC ACCIDENT RATE CALCULATION USING CASUAL EFFECT

Badounmar[1], and Nandar Win Min[2]

[1]Faculty of Computer Science, University of Computer Studies (Mandalay)
[2] Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City)
[1]badounmar.shwekyar@gmail.com, [2]nandarwinmin@gmail.com

ABSTRACT

In light of the rapid advancement of modern transportation systems, the escalating incidence of traffic accidents has emerged as a pressing public security concern. The multifaceted nature of these accidents renders their causes intricate and elusive, often defying singular attribution to one or two factors. Causal analysis, therefore, becomes indispensable in comprehending the intricate web of variables influencing such incidents. It entails scrutinizing the interplay among multiple variables to ascertain if a causal relationship exists between them. In our investigation, we adopt mutual information as a tool to probe into causal relationships, offering insights into the effect of causality within the context of traffic accidents. Furthermore, we employ logistic regression, to predict the fatality of accidents. According to the evaluation results, the mutual information-based quantification of causal effects yields an impressive 94% accuracy with logistic regression in predicting accident fatalities.

KEYWORDS

Public Security, Causal Analysis, Mutual Information, Logistic Regression

1. INTRODUCTION

Road traffic plays a crucial role in everyday life, facilitating mobility and connectivity. However, the prevalence of road accidents poses significant threats, resulting in severe bodily harm and property damage. Particularly in urban areas, where the density of traffic is high, road accidents emerge as a pressing concern, inflicting substantial losses in terms of both property and human life. In Thailand, the impact of road accidents is acutely felt, with statistics revealing a sobering reality. The toll of road casualties surpasses that of other causes, underscoring the magnitude of the problem. Alarmingly, an average of 12,000 individuals per year, translating to nearly two fatalities every hour, succumb to road accidents in Thailand alone [1].

The primary objective in accident prevention is to conduct a thorough analysis of accident causation, discern the principal factors contributing to accidents, and devise appropriate strategies accordingly. However, the occurrence of accidents stems from a complex interplay of various elements encompassing road conditions, traffic dynamics, environmental factors, and more. Pinpointing the predominant factor instigating an accident poses a considerable challenge, given the multifaceted nature of these incidents. Hence, it becomes imperative to evaluate the degree of influence exerted by different uncertain variables on traffic accidents systematically. Effectively managing and curbing traffic accidents hinges upon our ability to discern the relative significance of these diverse factors.

Developing effective urban traffic management strategies and allocating resources appropriately pose formidable challenges in decision-making. This difficulty arises from the inherent unpredictability of the factors influencing traffic behavior, compounded by limitations in data availability and the multitude of variables to be considered in crafting optimal policies. In light of these challenges, it becomes imperative to maximize the causal explanatory power by focusing on a minimal set of factors while ensuring predictability even under manipulations and distribution changes within the traffic management system. Achieving this goal necessitates acknowledging the importance of causal influence and counterfactual

dependence. In the realm of injury severity assessment, causation is often determined through univariate analysis. However, relying solely on individual factor selection via univariate analysis may not optimize the estimation of injury severity and could compromise the accuracy of addressing complex decision problems.

Employing a multivariate analysis-based approach holds promise for yielding more comprehensive and objective results in addressing complex issues such as traffic management. This methodology allows for the simultaneous examination of multiple variables, offering a more holistic understanding of the factors influencing traffic behavior and outcomes. However, the effectiveness of multivariate analysis hinges significantly on the correlation structure of the explanatory variables. When two or more explanatory variables exhibit high correlation, it can lead to challenges such as collinearity or multicollinearity in the analysis. Collinearity occurs when two or more variables are highly correlated with each other, making it difficult to separate their individual effects on the outcome variable. Multicollinearity, on the other hand, occurs when three or more variables are interrelated, leading to inflated standard errors and potentially misleading results.

The proposed research endeavors to innovate by developing a novel multivariate causal analysis approach while mitigating the impact of redundancy. This initiative seeks to address the inherent complexities and challenges associated with analyzing multiple variables simultaneously, particularly in the context of causality determination. By developing a multivariate causal analysis approach that effectively reduces redundancy, this research has the potential to yield more reliable and actionable insights into the complex interrelationships among variables influencing various phenomena, including traffic behavior, injury severity, and accident occurrence.

2. RELATED WORKS

The utilization of mathematical statistics in the analysis of traffic accidents has spurred numerous scholarly endeavors aimed at unraveling the underlying causes from diverse viewpoints. Notably, Abdel-Aty and Radwan [2] undertook a comprehensive examination, drawing insights from data encompassing 1,606 accidents. Their study revealed the intricate interplay of various factors, including traffic volume, vehicle speed, number of lanes, and lane width, in influencing the frequency of traffic accidents. Employing negative binomial modeling, Abdel-Aty and Radwan discerned the nuanced impact of each factor on accident occurrence. By elucidating the differential effects across different driver demographics such as gender and age, their research offered valuable insights into the complex dynamics shaping traffic safety.

In a study conducted by Chen et al. [3], a hierarchical Bayesian logistic model was employed to investigate the significant factors influencing the severity of driver injuries resulting from accidents. Their research revealed compelling insights, indicating that several key factors contribute to the severity of driver injuries. Specifically, the study highlighted those factors such as road curvature and the type of accident play pivotal roles in exacerbating the severity of driver injuries. Additionally, demographic variables including gender and age, as well as the presence of alcohol or drugs, were identified as significant contributors to injury severity.

In the realm of traffic accident analysis, researchers commonly employ two main approaches: aggregate methods and disaggregate methods. Aggregate methods, such as linear regression [4], clustering [5] [6], and time series analysis [7], offer a holistic perspective by examining all accidents collectively. While these techniques are relatively straightforward to apply, they may overlook the unique characteristics and circumstances of individual accidents. On the other hand, disaggregate methods, such as Bayesian networks [8] and discrete selection models [9], delve into the specifics of each accident. These techniques provide a more nuanced understanding of the underlying causal factors and interactions, thereby enhancing modeling effectiveness and forecast accuracy.

By observing traffic and infrastructure elements, such as the various types of roads in the study region, crossroads, parking lots, bus

stops, and other infrastructure facilities, the authors were able to determine the factors that influence pedestrian accidents. Linear regression model is used in this research [10]. The authors' goal was to pinpoint the characteristics and patterns of the frequency of road closures caused by tunnel traffic accidents. From the perspectives of accident effects, environmental factors, dangerous driver behaviour, and vehicle attributes, characteristics of tunnel and non-tunnel traffic accidents are compared. They applied Negative Binomial model for features.

3. METHODOLOGY

3.1. Causal Analysis in Complex System

Causality, or the notion of causal influence, underpins the relationship between two events, defining how one event affects or influences the occurrence of another. In the context of injury severity assessment, the determination of causation often relies heavily on single-factor analysis. However, this approach may fall short in capturing the complexity of the underlying factors contributing to injury severity. Relying solely on individual factor analysis may limit the accuracy and comprehensiveness of injury severity estimation, particularly when faced with complex decision-making scenarios. Therefore, adopting a multifactorial analysis-based approach holds promise in providing more comprehensive and objective insights.

3.2. Interdependent Variable

An independent variable is a characteristic, condition, or factor that researchers manipulate or control to observe its effect on other variables. It is the presumed cause or predictor in a cause-and-effect relationship. The term 'independent variable' refers to a variable whose value remains unaffected by changes in other variables within a research study. Researchers manipulate or modify the independent variable to observe its effects on other variables. An example of an independent variable is age, as it remains unaffected by external influences in certain contexts. Independent variables play a crucial role in research as they facilitate causal analysis and hypothesis evaluation by allowing researchers to identify relationships and make predictions based on their manipulations.

3.3. Dependent Variable

The dependent variable is the variable whose values are being predicted, explained, or modeled based on the values of one or more independent variables. In predictive modeling, the dependent variable is the variable that the model aims to predict. In statistical analysis, the dependent variable is often the focal point of the study, and researchers seek to understand how it is influenced by other variables. A dependent variable (DV) is a key concept in research that signifies its reliance on other variables, as suggested by its name. In an experimental setting, the dependent variable is the focal point under investigation. Researchers aim to understand how changes in other variables influence the value of the dependent variable. This exploration involves measuring and observing the outcomes of the experiment. Through the examination of the causal relationship between dependent and independent variables in research studies, researchers gain insights into the effects, associations, and correlations within the studied phenomena.

3.4. Logistic Regression

Logistic regression is a statistical modeling technique used for predicting the probability of a binary outcome based on one or more predictor variables. It's a type of regression analysis that's particularly useful when the dependent variable is categorical and has only two possible outcomes, typically labeled as 0 and 1. The primary objective of logistic regression is to model the relationship between the predictor variables and the probability of the binary outcome occurring. Unlike linear regression, which predicts continuous values, logistic regression predicts the probability that an observation belongs to a certain category. The logistic regression model is based on the logistic function, also known as the sigmoid function, which maps any real-valued number into a value between 0 and 1. This function is given by:

$$P(Y = 1\backslash X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\ldots\ldots + \beta_n X_n)}} \quad (1)$$

where, $P(Y = 1\backslash X)$ is the probability that the outcome variable Y is equal to 1 given the predictor variables X. $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the model, which represent the effect of each predictor variable on the log-odds of the outcome.

4. PROPOSED SYSTEM

The main contribution lies in the formulation of a novel mathematical approach designed to maximize the causality power while utilizing a minimal set of factors. This innovative methodology enables predictions even in scenarios involving manipulations and changes in distribution. The block diagram of proposed system is shown in Figure 1.

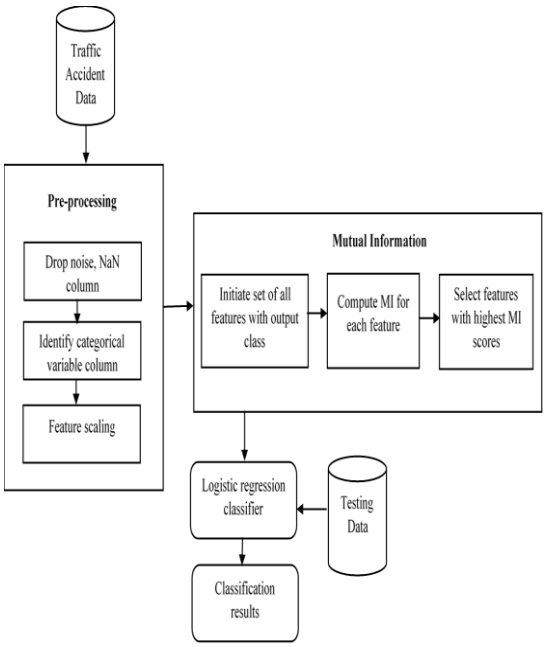


Figure 1. Block Diagram of The Proposed System

In this proposed system, there are three portions: data pre-processing, selection of features with the highest mutual information scores for fatal traffic accident cases, and classification using logistic regression.

The data set created by Thailand, Department of Highway, 2016, 2017, 2018 and 2019 is used in the proposed system. In this dataset, traffic accident data of size – 66114 and 108 numbers of features are included. Class with labelled: Class –FATAL_ CRASH, SEVERE_CRASH, and MINOR_CRASH is recorded in this dataset. The first step in data preprocessing includes the removal of columns containing NaN values and unnecessary columns like "route," "YEAR," "SEVERE_CRASH,"“MINOR_INJ_CRASH, " and "PDO" from the dataset. Subsequently, the identification of columns containing categorical variables is performed for normalization. While the values of the other columns are binary (0 or 1), the feature columns, such as "Kilometer," "Damage_pub," and "Damage priv," have high values. The feature scaling method, specifically standardization [10], is applied to normalize the features in the categorical columns. Mutual information serves as a statistical metric utilized to gauge the dependency and association between various features pertinent to traffic accidents. By employing mutual information, which quantifies the extent of shared information between variables, the study aims to unveil patterns and correlations among these parameters. In this research endeavor, a deliberate selection process is undertaken, focusing on the top 37 features exhibiting the highest mutual information scores. This meticulous approach ensures that the most influential factors contributing to traffic accidents are systematically identified and analyzed, thereby enhancing our understanding of the underlying dynamics and aiding in the formulation of targeted interventions for accident prevention and mitigation.

In this system, logistic regression technique is employed to assess and quantify the influence of various factors contributing to traffic accidents. The identification and prioritization of the most significant features are facilitated through a mutual information feature selection method, enabling the prediction or understanding of traffic accidents. Following feature selection, logistic regression is utilized to build a predictive model based on the selected features. This model is trained using historical data, enabling it to discern patterns and relationships among various factors and their correlation with traffic accidents.

5. EXPERIMENTAL RESULTS

In this system, the dataset from the Highway Accident Information Management System (HAIMS), which has been developed by the Department of Highways (DOH) is utilized. 80% of this dataset (59501 samples) is used as training data and 20% (6612 samples) is used as testing data. The performance of the system

is assessed using the criteria as accuracy, precision, recall, and f1-score. All 91 columns as features and "FATAL_CRASH" as the target column, Logistic Regression provides the accuracy by 91% can be obtained, as shown in Figure 2 in which 0 means the accident is fatal and 1 means the accident is not fatal.

	Precision	Recall	F1-score	Accuracy
0	0.88	0.94	0.91	0.91
1	0.91	0.93	0.92	

Figure 2. Results of Logistic Regression for All Features

The accuracy of our proposed method might be improved by using the mutual information algorithm to eliminate irrelevant features. In order to identify features with the most significant mutual information (MI) scores, we calculate the MI score for each feature that is associated with a target column. Among the features, only 37 surpass the average MI score of 0.001684824, representing the mean MI score across all features. We employ logistic regression to evaluate a set of features with k values ranging from 30 to 37, achieving an accuracy of approximately 94% with logistic regression. The results using logistic regression are shown in the Figure 3.

	Precision	Recall	F1-score	Accuracy
0	0.92	0.94	0.93	0.94
1	0.98	0.95	0.96	

Figure 3. Results of Logistic Regression for Selected Features

6. CONCLUSIONS

The research investigates the impact of causal factors on accident rates in Thailand by analyzing time series data spanning from 2016 to 2019. The study proposes a causal analysis framework built upon a mathematical approach within the realm of multivariate analysis. Multivariate techniques are employed to discern relationships between various variables and to uncover any underlying dependencies. Within this mathematical framework, the assessment of mutual information is introduced as a means to mitigate the influence of counterfactual dependencies. Additionally, logistic regression is utilized for predictive modeling or classification of accident-related factors.

REFERENCES

[1] Y. Tanaboriboon, and T. Satiennam, “Traffic Accidents in Thailand”, IATSS RESEARCH Vol.29 No.1, 2005.

[2] M. A. Abdelty, and A. E. Radwan, ‘‘Modeling traffic accident occurrence and involvement,’’ Accident Anal. Prevention, vol. 32, no. 5, pp. 633–642, Sep. 2000.

[3] C. Chen, G. Zhang, X. C. Liu, Y. Ci, H. Huang, J. Ma, Y. Chen, and H. Guan, ‘‘Driver injury severity outcome analysis inruralinter state highway crashes: A two-level Bayesian logistic regression interpretation,’’ Accident Anal. Prevention, vol. 97, pp. 69–78, Dec. 2016.

[4] J. Hinde and C. G. B. Demétrio, ‘‘Over dispersion: Models and estimation,’’ Comput. Statist. Data Anal., vol. 27, no. 2, pp. 151–170, Apr. 1998.

[5] S. Tuncer and A. Alkan, ‘‘Spinal cord-based kidney segmentation using connected component labelling and K-Means clustering algorithm”, Traitement Signal, vol. 36, no. 6, pp. 521–527, Dec. 2019.

[6] S. G. Langhnoja, M. P. Barot, and D. B. Mehta, ‘‘Web usage mining using association rule mining on clustered data for pattern discovery”, Int.J. Data Mining Techn. Appl., vol. 2, no. 1, pp. 141–150, 2013.

[7] B. Tianlong and H. Wentao, ‘‘Traffic accident prediction based on time series linear mode,’’ in Proc. IEEE Conf. Anthology, China, Jan. 2013, pp. 1–3.

[8] T. S. Shively, K. Kockelman, and P. Damien, ‘‘A Bayesiansemi-parametric model to estimate relationships between crash counts and roadway characteristics,’’ Transp. Res. B, Methodol., vol. 44, no. 5, pp. 699–715, Jun. 2010.

[9] R. P. Cogdill and P. Dardenne, ‘‘Least-squares support vector machines for chemometrics: An introduction and evaluation,’’ J. Near Infr. Spectrosc., vol. 12, no. 2, pp. 93–100, Apr. 2004.

[10] Milos Pljakic, “Pedestrian accidents Serbia (2015 to 2019)”.

ROUTE GUIDANCE SYSTEM BASED ON GRAPH THEORY

Hlaing Win Mon[1], Nyein Min Oo[2], and Zun Lae Nge[3]

[1,2,3]Faculty of Computing, University of Computer Studies (Mandalay)
[1]hlaingwinmon907@gmail.com, [2]nyeinminoo16@gmail.com

ABSTRACT

Route guidance is the provision of information to travellers to facilitate their selection of a path from their origin (or current location) to their destination. Vehicle routing or Road Network Routing refers to solving problems in the logistics, distribution, and transportation industry. The main focus is to find the effective shortest path with the least cost from the beginning city to the destination city. Route planning is essential for large or complex areas or developed areas like the Mandalay Division. Therefore, this paper proposes the well-known popular, Dijkstra's Algorithm to find the effective routes between the cities in our division. The proposed work starts from the ground truth map to the IT system output. In addition, this paper intends to highlight the students with less emphasis on mathematics and little knowledge of the effectiveness or essential requirement of Mathematics in the real-world IT industries.

Keywords

Road Network Routing, Mandalay Division, Dijkstra's Algorithm, Shortest path, IT industries, and Python.

1. INTRODUCTION

Route guidance systems are one of the main components of travel, transportation management and to travellers to facilitate their selection of a path from their origin (or current location) to their destination.

The lives of today's human beings need an efficient and reliable system to fulfill their unlimited desires and requirements, and this can be achieved through advanced technologies. As the transportation sector also follows this mechanism, the engineers try to find the best strategy for routing networks to ensure efficiency. The availability of data in the routing can increase the efficiency of road transportation, increase the accessibility of locations, and supplement to increase the economy of the country [1].

The computation process of finding the shortest path between two places in road network routing is a challenging task in the transportation, logistics, and distribution industry. The most suitable algorithm consideration is a key issue in many real-world network applications. The choice or selection of the algorithm depends on how the algorithm is reliable and easy to implement. The car navigation system is the most widespread application in Intelligent Transportation Systems. They offer information related to road networks such as the recommended shortest path, traffic conditions, and tourist locations based on road map databases and the Global Positioning System. To implement the real-time road routing-related application the choice of the applied algorithm is the most important point [2].

Determining the effective road path from a given area to a target city is a daily problem for people as they frequently need to ask this question when traveling in cars. Some of the current commercial solutions are slow, expensive, or inaccurate and it may either inconvenient for the client if he needs faster time and cannot afford expensive service [3].

Mandalay Division is one of the main regions in Golden Land, Myanmar. It is situated in the heart of our country with the border of Magway Region and Sagaing

Region at the west, Bago and Kayin at the South, and Shan State at the east. It contains eleven districts, 28 townships or 2320 in wards and village tracts. It is an important economic division in Myanmar and shows 15% of the whole country's economy. Therefore, the road in this division is complex, and finding an efficient path system is one of the most important requirements of Regional Development. Figure 1 shows the road map of the Mandalay Region [4].

Therefore, this paper proposes a reliable and effective shortest path algorithm to assist the development of the road routing system in the Mandalay Division. The rest of the paper is presented as follows. Section 2 discusses the related works for finding the shortest path system. Section 3 presents the shortest path algorithm and Section 4 shows the implementation of the system. Section 5 discourses the results and Section 6 concludes the research work.

2. RELATED WORKS

There are many previous works done for routing road networks using different approaches. Some of the routing networks for road systems are described in this section.

The researchers in [2] study various vehicles' routing systems for route planning for modern cities. They discussed the overview of the inputs and output of the general systems and briefly described the Dijkstra algorithm, A* algorithm, Tabu search, ANT-based colony, Genetic Algorithms (GA), and Hybrid Genetic Algorithms. They highlight how can improve the road network finding system based on the above algorithms.

The authors in [1] propose a system to route the Parklands Road network using graph theory based on Dijkstra's algorithm. The user needs to input the start and end location to the system and the system replies with the best shortest path among the routes. The system needs to send using Google with public API and reply JSON object to decode the time and path in seconds.

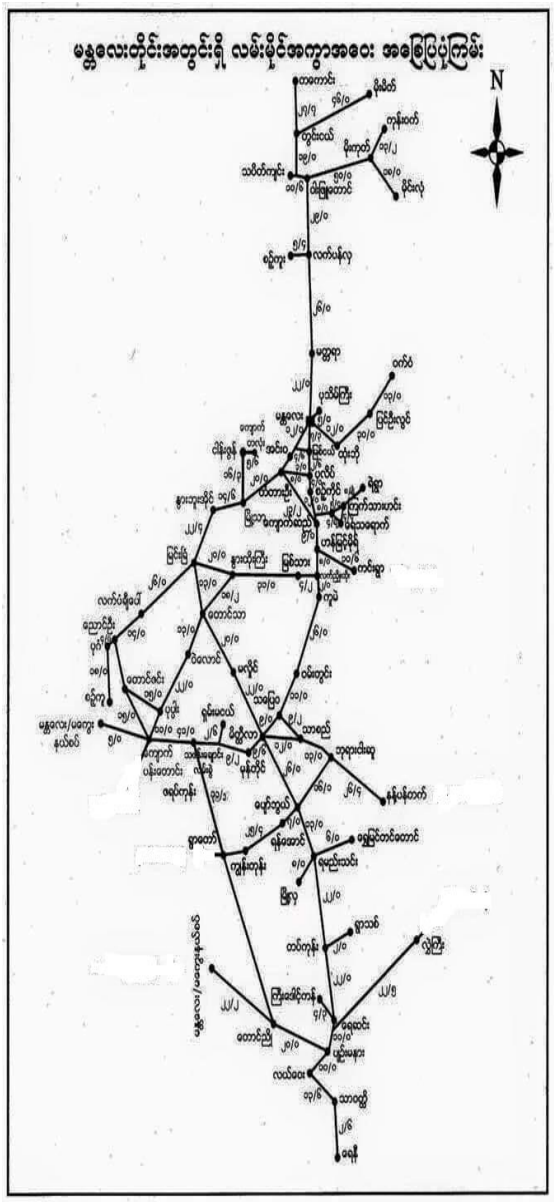


Figure 1. The Road Map of the Mandalay Region

They recommend that Dijkstra's algorithm provide the best routing methods for Parklands Road.

The intelligent route system is designed in [5] where image processing techniques, video processing methods, mathematics theory, and Information Technology are applied. They combine A* and Dijkstra algorithms to create a larger system for real-world applications to help the drivers. Although the combination of two algorithms increases complexity, the better

accuracy of the path can provide the system.

The online shortest path system based on a reliable and fast algorithm called 'Dijkstra' is proposed in [6]. They extend Dijkstra's Algorithm to develop a modified Dijkstra's Shortest Path (MDSP) algorithm for the road network. They proved that their proposed algorithm performs better results than the original.

Based on these road routing algorithms and systems, this paper proposes a system for finding the shortest path using the recommended popular Dijkstra Algorithm for the Mandalay Region.

3. GRAPH THEORY

Graph: A graph G = (V, E) consists of V, a nonempty set of vertices (or nodes) and E, a set of edges. Each edge has either one or two vertices associated with it, called its endpoints. An edge is said to connect its endpoints [7].

Directed graph: G = (V, E) consists of a nonempty set of vertices V and a set of directed edges E. Each directed edge is associated with an ordered pair of vertices. The directed edge associated with the ordered pair (u, v) is said to start at u and end at v.

Undirected graph: An undirected graph is a graph, i.e., a set of objects (called vertices or nodes) that are connected together, where all the edges are bidirectional. An undirected graph is sometimes called an undirected network.

Road Networks: Graphs can be used to model road networks. In such models, vertices represent intersections and edges represent roads. When all roads are two-way and there is at most one road connecting two intersections, a simple undirected graph could be used to model the road network. Model road network, some roads are one-way and some roads are more than one road between two intersections. To build such models, undirected edges to represent two-way roads and directed edges to represent one-way roads.

Path: A path is a sequence of edges that begins at a vertex of a graph and travels from vertex to vertex along edges of the graph. As the path travels along its edges, it visits the vertices along this path, that is, the endpoints of these edges.

3.1 Dijkstra's Algorithm

It is used to determine the minimum shortest way and is an iterative algorithm to provide the shortest path from the specified beginning node to all other target nodes in the weighted graph. Again, this idea is similar to the outcome of a breadth-first search (BFS). The distance dictionary is used to keep track of the cost from the start point to each destination point which contains 0 for the beginning node and infinity for others.

The algorithm will change these distance values until it finds the smallest total weight. The algorithm tries to visit every vertex in the graph in a priority queue by finding the smallest distance. The pseudo-code for the Dijkstra's algorithm is described in Algorithm 1.

Algorithm 1: Dijkstra's algorithm

```
1 function Dijkstra (Graph, Source)
2
3  create vertex set Q
4
5   for each vertex V in Graph:
6      dist [ V]← INFINITY
7      prev [v] ← UNDEFINED
8       add V to Q
9   dist [ source] ← 0
10
11    while Q is not empty:
12      u ← vertex in Q with min dist [u]
13
14          remove u from Q
15     // only v that is still in Q
16     for each neighbor v of u:
17   alt← dist [u]+ length (u, v)
18    if alt < dis [v]
19      dist [v]← alt
```

```
20        prev[v]← u
21
22  return dist [], prev []
```

4. METHODOLOGY

This section describes the finding of the shortest path between the cities in the Mandalay Division and includes the three steps of pre-processing, evaluations, and implementing Dijkstra's Algorithm in Python.

4.1. Pre-Processing

The pre-processing step include the drawing the list from the maps of the Mandalay Division as shown in Table 1, and the graph is shown in Figure 2.

Table 1. The cities name, nodes, routes and distance (miles) for the Mandalay Division

City Name	Nodes	Routes	Weight (miles)
Yae Ni	A		
Tharwut Hti	B	A, B	2.6
Lewe	C	B, C	13.6
Pyinmana	D	C, D	10.0
Yezin	E	D, E	10.0
Taungnyo	F	D, F	20.0
Ywar Taw	G	F, G	45.0
Kyun Kone	H	G, H	10.0
Kyauk Padaung	I		
Popa	J	I, J	10.0
Taung Zin	K	J, K	15.0
		I, K	15.0
Nyaung- U	L	K, L	15.0
Bagan	M	L, M	4.0
Sintgu	N	M, N	18.0
Let Pan Chipaw	O		
Myintgyan	P	O, P	26.0
Wea Laung	Q	J, Q	22.0
Taung Tha	R	Q, R	13.0
		P, R	13.0
Mahlaing	S	R, S	20.0
Tha Phaywa	T	S, T	22.0
Meiktila	U	T, U	9.0
Tha Zi	V	V, T	9.2
		V, S	12.0
Mondaing	W	U, W	9.6
Shanmange	X	W, X	9.2
Pyawbwe	Y	U, Y	26.0
Yan Aung	Z	Y, Z	7.0
		Z, H	25.4
Yamethin	A1	Y, A1	13.0
Myo Hla	A2	A1, A2	8.0
Tat Kon	A3	A1, A3	22.0
		A3, E	22.0
Ywathit	A4	A3, A4	2.0
Payangazu	A5	A5, Y	16.1
		A5, V	13.0
Wundwin	A6	T, A6	11.0
Kume	A7	A6, A7	26.0
Lethnyohtoe	A8	A7, A8	2.0
Hanmyintmo	A9	A9, A8	8.0
Myittha	B1	B1, A8	4.2
Natogyi	B2	B2, R	18.2
		B2, B1	31.0
		B2, P	20.0
Nabuaing	B3	B3, P	22.4
Myo Thar	B4	B3, B4	14.6
Ngazun	B5	B4, B5	16.3
Tada- U	B6	B4, B6	20.7
Innwa	B7	B6, B7	3.0
KyaukSe	B8	B6, B8	23.2
		B8, A9	9.0
Plate	B9	B6, B9	8.0
Mandalay	C1	C1, B7	12.0
Myintnge	C2	C1, C2	7.3
		C2, B9	2.6
		C2, B7	4.6
Patheingyi	C3	C3, C1	5.0
Htone Bo	C4	C4, C1	12.0
Sintgaing	C5	C5, B9	4.0
		C5, B8	12.0
King	C6	C6, A9	10.6
Yaytha York	C7	C7, B8	12.7
Pyin Oo Lwin	C8	C8, C4	30.0
Madaya	C9	C9, C1	22.0
Let Pan Hla	D1	D1, C9	26.0
SintKu	D2	D2, D1	5.4
War Hpyu Taung	D3	D3, D1	29.0
Thabeik Kyin	D4	D4, D1	10.6
Twin Nge	D5	D5, D4	19.0
Ta Kaung	D6	D6, D5	27.7
Mongmit	D7	D7, D5	46.0
Mogoke	D8	D8, D3	50.0
Monglon	D9	D9, D8	18.0

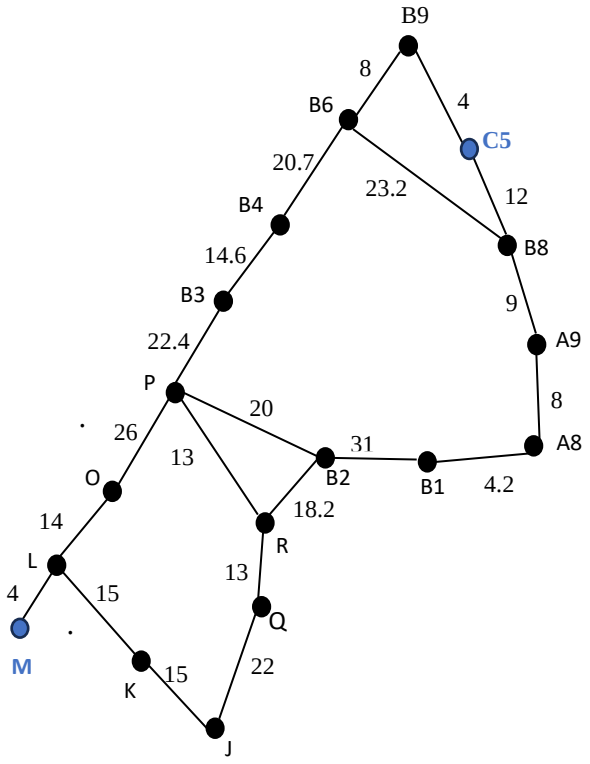


Figure 2. The Graph from Sintgaing to Bagan

4.2. Evaluations

Evaluations based on applying the algorithm. The starting node is labelled 0 and the other nodes are infinity. So, we start node C5 and its label is 0.

Table 2. Iteration 0

Nodes	Label	Status
C5	0	Chosen

Node C5 can reach to nodes B8 and B9 and the list of labelled nodes is chosen or not chosen.

Table 3. Iteration 1

Nodes	Label	Status
C5	0	Chosen
B9	(4, C5)	Not Chosen
B8	(12, C5)	Not chosen

For the two distance B8(12, C5) and B9(4, C5), the minimum distance is node B9. The status of node B9 is changed to chosen.

Table 4. Choose Node B9

Nodes	Label	Status
C5	0	Chosen
B9	(4, C5)	Chosen
B8	(12, C5)	Not chosen

The new starting node is node B9 can reach to node B6. The list of labelled nodes is updated as:

Table 5. Iteration 2

Nodes	Label	Status
C5	0	Chosen
B9	(4, C5)	Chosen
B8	(12, C5)	Not chosen
B6	(12, B9)	Not chosen

For the two not chosen labels B8 and B6, both nodes are the same distance.so we can choose node B8. The status of node B8 is changed to chosen.

Table 6. Choose Node B8

Nodes	Label	Status
C5	0	Chosen
B9	(4, C5)	Chosen
B8	(12, C5)	Chosen
B6	(12, B9)	Not chosen

The new starting node is node B8 can reach to node A9. The list of labelled nodes is updated as:

Table7. Iteration 3

Nodes	Label	Status
C5	0	Chosen
B9	(4, C5)	Chosen
B8	(12, C5)	Chosen
B6	(12, B9)	Not chosen
A9	(21, B8)	Not chosen

For the two not chosen labels B6 and A9, node B6 is the minimum distance. The status of node B6 is changed to chosen.

Table 8. Choose Node B6

Nodes	Label	Status
C5	0	Chosen
B9	(4, C5)	Chosen
B8	(12, C5)	Chosen
B6	(12, B9)	Chosen

A9	(21, B8)	Not chosen

The new starting node is node B6 can reach nodes B4 and B8. The list of labelled nodes is updated as:

Table 9. Iteration 4

Nodes	Label	Status
C5	0	Chosen
B9	(4, C5)	Chosen
B8	(12, C5)	Chosen
B6	(12, B9)	Chosen
A9	(21, B8)	Not chosen
B4	(32.7, B6)	Not chosen
B8	(36.2, B6)	Not chosen

For the three not chosen labels B4, A9, and B8, node A9 is the minimum distance. The status of node A9 is changed to chosen.

Table 10. Choose Node A9

Nodes	Label	Status
C5	0	Chosen
B9	(4, C5)	Chosen
B8	(12, C5)	Chosen
B6	(12, B9)	Chosen
A9	(21, B8)	Chosen
B4	(32.7, B6)	Not chosen
B8	(36.2, B6)	Not chosen

Similarly, the fifth, sixth, seventh, ..., and fourteenth iterations are calculated by using Dijkstra's algorithm. And then, iteration fifteenth is reached.

Table 11. Iteration 15

Nodes	Label	Status
C5	0	Chosen
B9	(4, C5)	Chosen
B8	(12, C5)	Chosen
B6	(12, B9)	Chosen
A9	(21, B8)	Chosen
A8	(29, A9)	Chosen
B1	(33.2, A8)	Chosen
B4	(32.7, B6)	Chosen
B8	(36.2, B6)	Chosen
B3	(47.3, B4)	Chosen
B2	(64.2, B1)	Chosen
P	(69.7, B3)	Chosen
R	(82.4, B2)	Chosen
Q	(95.5, R)	Chosen
O	(95.7, P)	Chosen
L	(109.7, O)	Chosen
M	(113.7, L)	Chosen

For every node, the shortest paths are as follows.

Table 12. Shortest Paths

Nodes	Routes	Distance (miles)
C5	C5	0
B9	C5 - B9	4
B8	C5- B8	12
B6	C5 - B9- B6	12
A9	C5- B8-A9	21
A8	C5- B8-A9-A8	29
B1	C5- B8-A9-A8-B1	33.2
B4	C5 - B9- B6- B4	32.7
B8	C5 - B9- B6- B8	36.2
B3	C5 - B9- B6- B4- B3	47.3
B2	C5- B8-A9-A8-B1- B2	64.2
P	C5 - B9- B6- B4- B3- P	69.7
R	C5- B8-A9-A8-B1- B2-R	82.4
Q	C5- B8-A9-A8-B1- B2-R-Q	95.5
O	C5 - B9- B6- B4- B3- P- O	95.7
L	C5 - B9- B6- B4- B3- P- O- L	109.7
M	C5 - B9- B6- B4- B3- P- O- L-M	113.7

The shortest route from node C5 to node M is determined the shortest path or route. The desired route is C5 - B9- B6- B4- B3- P- O- L-M.

4.3. Implementing Dijkstra's Algorithm in Python.

```
1: def dijkstra(graph, start, goal):
2:     unvisited = {n: float('inf') for n in
    graph. keys()}
3:  unvisited[start] = 0
4:    visited = {}
5:  revPath = {}
6: while unvisited:
7:     minNode = min(unvisited,
key=unvisited.get)
8:         visited[minNode] =
unvisited[minNode]
9:     if minNode == goal:
```

```
10:          break
12:       for neighbor in graph.get(minNode):
13:            if neighbor in visited:
14:                continue
15:          tempDist = unvisited[minNode] +
                        graph[minNode][neighbor]
16:          if tempDist < unvisited[neighbor]:
17             unvisited[neighbor] = tempDist
18:           revPath[neighbor] = minNode
19:         unvisited.pop(minNode)
20: node = goal
21: revPathString = node
22: while node != start:
23:       revPathString += revPath[node]
24:       node = revPath[node]
25:    fwdPath = revPathString[::-1]
26:    return fwdPath, visited[goal]
27: myGraph = {
                    }
28: path, cost = dijkstra(myGraph, startNode,
                goalNode)
29: print (f 'The shortest path : { startNode,
   visitedNode, goalNode} is {cost}
   miles.')[8]
```

5. RESULT AND DISCUSSION

The path from Dijkstra's Algorithm is C5 - B9- B6- B4- B3- P- O- L-M, calculation to get 113.7 miles. It means that the short distance from Sintgaing to Bagan starts from Sintgaing, Pa late, Tada U, Myo Thar, Nabuaing, Myingyan, Let pan chi paw, Nyaung -U, and Bagan.

6. CONCLUSIONS AND FUTURE WORKS

Nowadays there exist a great number of routes that enable the driver to reach his destination, but a few of them use the information about the current situation on a road. This paper proposes a shorted pathfinding algorithm for the road network in Mandalay Division using Dijkstra's Algorithm. The path found by Dijkstra's Algorithm is the effective way from Sintgaing to Bagan and it needs the requirement of the efficiency and reliability of the real-world system. This Dijkstra Algorithm can be found shortly the best round from any city in Mandalay Division. Its logic can be used in many network routings and traveling salesperson problems that need a fasted way between the cities. Moreover, this algorithm can exhibit clearly how mathematics affects real-world applications in information technology.

Future work may involve considering real-time traffic data, road conditions, and other factors to make the algorithm even more adaptive to the dynamic nature of transportation systems.

ACKNOWLEDGEMENTS

The author is very thankful to the Editorial Board for editing this paper and for the effective true guidelines to publish this paper. Her special thanks are due to Dr. Thin Thin Naing, Professor, and Head, Faculty of Computing, University of Computer Studies (Mandalay) for her encouragement to do this research work.

REFERENCES

[1] Mhnd FARHAN, "Traffic Routing Algorithm for Road Network", Scientific Bulletin Vol. XXIV, No 2 (48), 2019.

[2] Vi Tran Ngoc Nha, S. Djahel and J. Murphy, "A comparative study of vehicles' routing algorithms for route planning in smart cities," *2012 First International Workshop on Vehicular Traffic Management for Smart Cities (VTM)*, Dublin, 2012, pp. 1-6, doi: 10.1109/VTM.2012.6398701

[3] Shrawan Kr. Sharma *et al*, "Shortest Path Searching for Road Network using A* Algorithm", International Journal of Computer Science and Mobile Computing, Vol.4 Issue.7, July- 2015, pg. 513-522, ISSN 2320–088X

[4]https://en.wikipedia.org/wiki/Mandalay_Region

[5] Chmiel, W., Skalna, I. & Jędrusik, S. Intelligent route planning system based on interval computing. *Multimed Tools Appl* 78, 4693–4721, 2019. https://doi.org/10.1007/s11042-018-6714-x

[6] Mazhar Iqbal, Kun Zhang, Sami Iqbal, Irfan Tariq, *"A Fast and Reliable Dijkstra Algorithm for Online Shortest Path"* SSRG International Journal of Computer Science and Engineering 5.12 (2018): 24-27.

[7] K. H. Rosen, "Discrete Mathematics and its Application", eighth edition, McGraw-Hill

[8]https://www.udacity.com/blog/2021/10/implmenting-dijkstras-algorithm-in-python.html

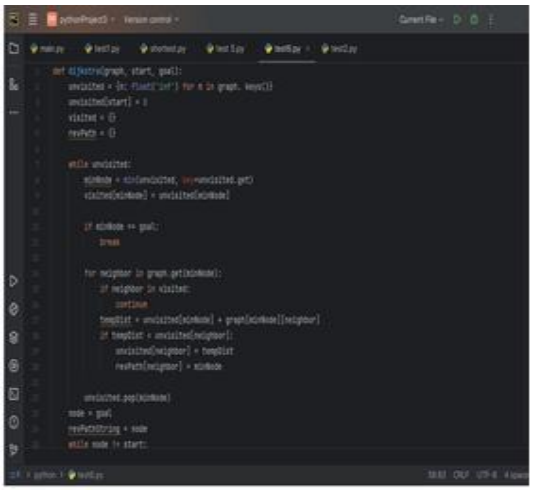
Authors Biography

First Autor: Daw Hlaing Win Mon, Associate Professor, Faculty of Computing, University of Computer Studies (Mandalay). She achieved her B.Sc. (Q)(Maths.) in 2002, and M.Sc in 2005. The places of employ are the University of Computer Studies (Pang Long, Monywa), UCSY, TU (Sittwe), and the University of Computer Studies (Mandalay).

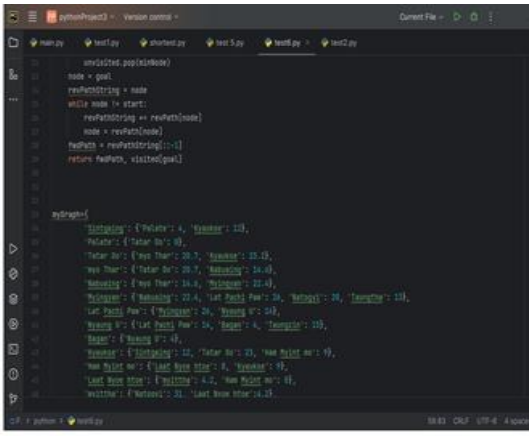
Second Author: U Nyein Min Oo, Lecturer, Faculty of Computing, University of Computer Studies (Mandalay). He got his B.Sc.(Q) (Maths.) in 2006, and his M.Sc. (Maths.) in 2009. The places of employ are the University of Computer Studies, Mandalay, University of Computer Studies (Mandalay), University of Computer Studies (Thaton), and University of Computer Studies (Mandalay).

Third Autor: Daw Zun Lae Nge, Assistant Lecturer, Faculty of Computing, University of Computer Studies (Mandalay). She got her B.Sc. (Hons) (Maths.) in 2014, and M.Sc. (Maths.) in 2017. The places of employ are the University of Computer Studies (Bhamo), and the University of Computer Studies (Mandalay).

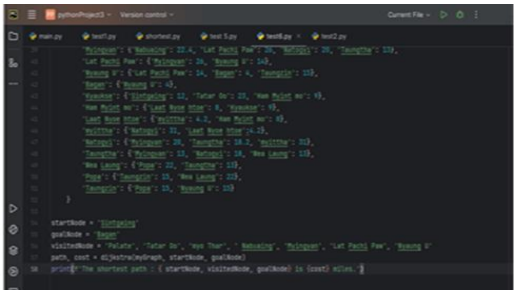
APPENDIX



(i) Dijkstra Algorithm main Function in Python



(ii) Data Input in Python



(iii) Data Input in Python

(iv) Data Output of Dijkstra Algorithm

PALI IMAGE BINARIZATION WITH OTSU THRESHOLDING

Min Min Su [1], Yu Ya Win [2], and San San Tint[3]

[1,2] Faculty of Computer Science, University of Computer Studies (Mandalay)

[3] Pro-Rector, University of Computer Studies (Mandalay)

[1]minminsu76@gmail.com, [2]yuyawin@gmail.com, [3]dr.sansantint@gmail.com

ABSTRACT

A binary image consists of pixels that are either black or white, represented as a series of 0s and 1s depending on pixel brightness. Binarization is the process of converting a digital image into this format, using 0s and 1s. Binarization is widely used in image processing, like Optical Character Recognition (OCR), where characters are first converted into black and white to prepare them for further processing. Grayscale binary images are faster to store and process compared to color images. They simplify algorithms and reduce computational complexities. OCR programs perform more effectively with images exhibiting higher contrast. Thus, employing specialized binarization algorithms optimized for OCR enhances OCR accuracy. The efficacy of these algorithms relies on thresholding. Hence, it's advisable to explore various thresholding techniques to find the most suitable one for specific image types and purposes. This paper explores Otsu thresholding across four distinct types of digital Pali images varying in quality. The accuracy of binarization is assessed through F-measure values. Experimental findings suggest that printed documents yield the highest accuracy for subsequent OCR processes.

KEYWORDS

Image Binarizing, Otsu Thresholding, Optical Character Recognition, Pali

1. INTRODUCTION

Buddhist texts are religious texts that pertain to or are linked with Buddhism and its diverse traditions. The Buddha religious texts were written in different languages, methods and writing system such as scripts. Memorizing, reciting, and copying the text held spiritual significance and value. The Buddhist texts are historically and precious cultural heritage which are important to maintain. Digitizing ancient manuscripts is a crucial method for safeguarding literature and preserving cultural heritage. Nonetheless, this process demands significant time and effort, especially when handling large volumes of documents manually, and may be prone to errors due to the challenges posed by aging documents. Hence, there is a necessity to employ computers for the automated and precise processing of historical manuscripts. The Character Recognition System has arisen for this specific purpose. Document image binarization stands as a crucial preprocessing step aimed at segmenting the input document image into text (foreground) and non-text (background).

The noise in digital image is due to many factors such as low quality or degradation, aging, photocopying, screenshot or during image capture with low lighting and contrast. Image preprocessing methods are applied to reduce noise. The process of reducing noise by binarization with appropriate threshold. This image preprocessing step is also called the pixel level preprocessing. To account for variations within the image, Otsu thresholding is performed in tesseract, where Otsu's algorithm is applied to small sized rectangular divisions of the image.

2. RELATED WORKS

The Otsu method, a global binarization approach, stands out as a highly effective

choice for OCR systems due to its computational effieiency [1]. Otsu method relies on statistical techniques to determine a global threshold, but the existing global methods face challenges in handling non-uniform illumination across entire documents. In contrast, local binarization techniques prove more adept at addressing issues related to non-uniform illumination and degradation in document images [2,3,4,5,67]. J.He, Q.K.M.Do, A.C.Downton, J.H.Kim [5] compare the several binarization algorithms for aging and historical documents, which results effect to OCR performance on ABBYY engine. The recognition of OCR results are evaluated by eight different thresholding techniques and compared the results. As the results, the adaptive Niblack algorithm performed the best result in commercial ABBYY engine. P.J.Burt [8] proposed a highly effective restoration technique that relies on uncovering the 3D shape of book's surface through shading information in a scanned document image. The results demonstrate the substantial reduction in geometric and photometric distortions, leading to a significant enhancement in OCR accuracy. M.L.Feng and Y.P.Tan, [7] proposed an enhanced approach for binarization document images, leveraging adaptive utilization of local image contrast. The method is formulated to tackle common issues found in low-quality images, such as uneven illumination, low contrast, and random noise. Experimental evaluations have been carried out, presenting results that highlight the efficiency of the proposed method. V.Nicole ,K .Khurram [9] presented a new sliding window based local thresholding technique for low quality ancient documents. The performance results are analyzed as recall rates.

Based on the literature review of both local and global thresholding algorithms, it can be concluded that none of the approaches achieve satisfactory result severely degraded images with diverse degradations such as variable background and shadows, noise with dirty background, aging background, non-uniform illumination and contrast, camera vibrate, spillage or overspill of wanted ink, decay or deterioration of historical documents such as Burmese Pesa and stone inscription.

3. BACKGROUND THEORY

3.1 Pixel

A pixel, short for picture element, serves as the fundamental unit constituting images. In a color image, a pixel generally consists of three components: Red, Green, and Blue (RGB), whose combined values determine the resulting color. A typical grayscale image is 8 bit or contain 2 8=256 or (0-255) shades of grey. A typical colour image is 24 bit, or 8 bit for each Red, Green and Blue providing 2 24 which is 16777216 different colours. This means that a colour pixel has a triplet value comprised of 0-255 for each of Red, Green and Blue whereas a grayscale pixel has a single value 0-255. A pixel is encompassed by eight neighboring pixels, with four being direct or adjacent neighbors, and the remaining four being indirect or diagonal neighbors.

3.2 Image Thresholding

Thresholding high-quality images is relatively straightforward, yet binarizing historical document images poses significant challenges due to severe degradation factors like ink bleed-through, page stains, faded text strokes, and artifacts. Moreover, variations in text stroke color, width, brightness, and continuity within degraded scripts further exacerbate the complexity of binarization. Figure 1 presents the sample degraded historical document.

Figure 1. Example of Input Image with Noise.

(Raja kumara Stone Inscription in Pali)

Image Binarization, also referred to as Image Thresholding, is a technique employed to distinguish the foreground from the background within an image by converting a grayscale or RGB image into a binary representation. This method stands as one of the fundamental approaches utilized to extract pertinent information from an image or identify regions of interest. The images are first set to grayscale to be used in Image thresholding technique. Then the comparison between the pixel value of each image and the threshold value is carried out. The pixel value which is less than the threshold is to be set zero, or else the maximum possible value (225) is set.

Algorithm to convert an image to binary function as:

```
convertToBinaryImage
(ImageData,threadshold = 128)
{
binaryImage=[ ]
    for each pixel in imageData
{
    grayscale=(pixel.R+pixel.G+pixel.B)/3
    if (grayscale >= threshold) binaryValue=1
    else binaryValue=0
    binaryImage[ ]=binaryValue
}
    return binaryImage
}
```

The followings are the simplest way of describing image thresholding.

```
If P(x, y) < Threshold:
    P(x, y) = 0
else:
    P(x, y) = 255
```

where, P(x, y) = Pixel Value

According the algorithm described above, in binarization,the output image can be clearly seen to consist of only 0 and 255, representing binary-0 and binary-1, respectively.

4. METHODOLOGY

Thresholding/Binarization techniques can be categorized as global thresholding and local thresholding

The Global thresholding considers the whole image and its global characteristics to determine a single threshold value where as local Thresholding divides the image into regions/segments to determine independent threshold values for each region/segment.

Image thresholding can be used in various applications where the use of thresholding techniques varies based on the nature of the images used. Test extraction from documents is a good example of the application which uses the image thresholding technique.

Consequently, printed receipts will be utilized for comparing various image thresholding and binarization techniques in this analysis.

4.1 Otsu Thresholding

Otsu Thresholding is a global binarization technique that automatically selects the threshold value based on the assumption that the image's foreground and background form a bimodal distribution characterized by two distinct peaks, representing two classes. In this method, the threshold value is usually considered to be the optimal value between the two histogram peaks. The fundamental concept behind Otsu Thresholding, a variance-based technique, is to determine the threshold value where the weighted variance between the background and foreground pixels (representing two classes) within the image is minimized. This is achieved by minimizing the weighted within-class variance while simultaneously maximizing the weighted between-class variance. The algorithm systematically explores all conceivable threshold values, analyzing the distribution of background and foreground pixels to pinpoint the threshold value at which the spread is minimized. The weighted within-class variance can be articulated as follows:

$$\sigma_{\omega}^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t)$$

Where;

$\omega_1(t)$ and $\omega_2(t)$ are the probabilities of the two intensity classes divided by a threshold 't' can be expressed within the range of 0 to 255.

The components of the equation can be represented as follows:

$$\omega_1(t) = \sum_{i=1}^{t} P(i) \; and \; \omega_2(t) = \sum_{i=t+1}^{I} P(i)$$

When deriving the aforementioned equation, the subsequent step involves computing the means of the two intensity classes as follows

$$\mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{\omega_1(t)} \; and \; \mu_2(t) = \sum_{i=t+1}^{I} \frac{iP(i)}{\omega_2(t)}$$

Finally, To compute the variances for the two intensity classes;

$$\sigma_1^2(t) = \sum_{i=1}^{t} [i - \mu_1(t)]^2 \frac{P(i)}{\omega_1(t)}$$

$$\sigma_2^2(t) = \sum_{i=t+1}^{I} [i - \mu_2(t)]^2 \frac{P(i)}{\omega_2(t)}$$

Substituting the derived components into the equation allows us to ascertain the within-class variance. Subsequently, our objective is to determine the threshold value at which the within-class variance reaches its minimum.

The comprehensive representation of the Otsu Thresholding algorithm is as follows:

1. Calculate the intensity histogram and intensity level probabilities.

2. Initialize $\omega_i(0) and \; \sigma_i^2(0)$

3. Iterate over all possible thresholds; t=0,1,2,….., max_intensity

a. update the values of $\omega_i and \; \sigma_i^2$, where $\omega_i$is the probability and σ_i^2 is the variance of class i

b. calculate the within-class variance $\sigma_w^2(t)$

4. The final threshold is the minimum $\sigma_w^2(t)$ [1].

5. Experimental Results

An image with the text (figure 2) is given as input to the system and (figure 3) converts the input image to grayscale image and (figure 4) binarizes the grayscale image with Otsu thresholding. The input images are rectangular shaped and the format of input image can be used as the extension .jpeg,.png,.gif.

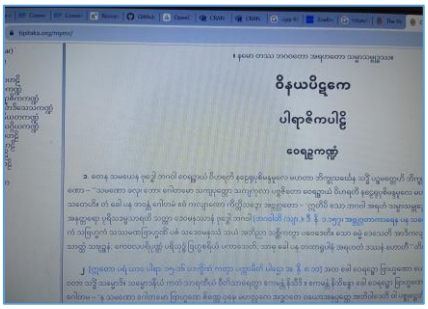


Figure 2. Example of Input Image with Colour. (Screenshots Image)

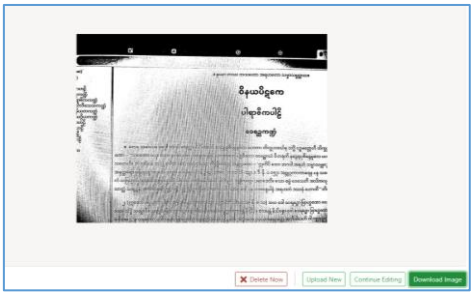


Figure 3. Grayscale of Input Image.

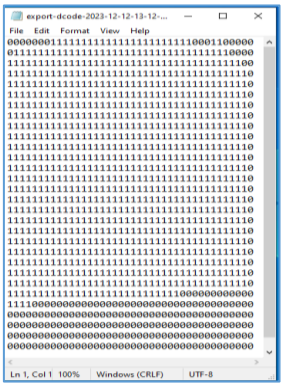


Figure 4. Binarization of Input Image.

In performance evaluation, various measures, such as mean square error (MSE), signal to noise ratio (SNR), and peak signal to noise ratio (PSNR) are employed to assess the effectiveness of binarization technique. Mean square error is a measure of the average squared difference between predicted and actual values. Low MSE values indicate better accuracy. Signal to noise ratio (SNR) is a measure used to quantify the level of a desired signal relative to the background noise. A higher SNR indicates a stronger, more discernible signal amidst lower levels of noise,

which is desirable for better signal quality and reliability. Peak signal to noise ratio (PSNR) is a metric commonly used to measure the quality of reconstructed or processed signals in image processing. Higher PSNR values indicate better image quality. In this work, the four types of Pali document image (225 images) are used to evaluate the binarization accuracy. This experimental results are shown by the F-measure for accuracy shown in Figure 5.

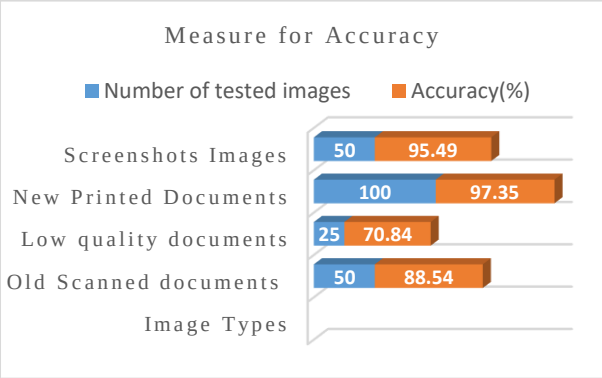


Figure 5: F-Measure for Accuracy

Precision is calculated by dividing the true positive (pixel identified as foreground in both ground truth and binarized image) by the sum of true positives and false positives (pixels identified as foreground in the binarized image but are actually background in the ground truth image).

Recall is determined by dividing the true positives by the sum of true positive and false negatives (pixels identified as background in the binarized image but are actually foreground in the ground truth image).

F-measure, the harmonic mean of precision and recall is expressed as (2*Precision*Recall) divided by (Precision + Recall). A higher F-measure value indicates better result in terms of both precision and recall.

The accuracy of binarization for screenshot images is 95.49, while for newly printed document images, which achieves a higher accuracy of 97.35.

6. CONCLUSION

The input Pali images are collected from various sources such as from the site [https://tipitaka.org/ymr/CSCDTipitaka (Myanmar)] [10] which are used for initial preprocessing step in digital image processing. Various pre-processing steps are needed to develop the ground truth for the Pali images to get the more accurate result of OCR. Selecting a thresholding technique in accordance with the properties of images and the context being used in is necessary. Hence, it remains imperative to explore all available thresholding techniques before determining the most effective one that suits the present task optimally. The four types of Pali image document with difference quality are tested for the experiment, we found that the new printed documents are the best quality for the threshold value and the low-quality documents are the lower quality for binarization which effects to the character recognition accuracy.

7. REFERENCES

[1] Andrew Greensted, “Otsu Thresholding, The Lab Book Pages”.

[2] B M. Singh, R .Sharma, “Adaptive Binarization of severely degraded and non-uniformly illuminated documents”, Springer, IJDAR, 2014.Doermann, K.Tombre (eds.), Handbook of Document Image Processing and Recognition, Springer-Verlag London 2014.

[3] E.Badekas, N.Papamarkos, “Estimation of appropriate parameter values for document binarization techniques”, IJRA, 24(1), 2009.

[4] I.K.Kim, D.W.Jung, P.H.Park, “Document image binarization based on topographic analysis using a water flow model”, Pattern Recognition, 2002.

[5] J.He, Q.K.M.Do, A.C.Downton, J.H.Kim, “A comparison of binarization methods for Historical archive documents”, IEEE, ICDAR, 8, 2005.

[6] J. Sauvola, M. Pietikainen, “Adaptive document image binarization”, Pattern Recognition, 2000.

[7] M.L.Feng and Y.P.Tan, “Contrast adaptive binarization of low quality document images”, IEICE Electron, vol. 1, No. 16, 501-506(2004).

[8] P.J.Burt,“Recovery of distorted document images from bound volumes”, Sixth ICDAR, 2001.

[9] V.Nicole ,K .Khurram, “ Comparison of Niblack inspired binarization methods for ancient documents”.

[10] https://tipitaka.org/ymr/ CSCDTipitaka (Myanmar)

Authors

Biography

First Author. I graduated Master of Science in Mathematics, M.Sc (Maths) in 2004, Master of Information Science, M.I.Sc in 2007. I started to serve as a Tutor at UCSM and I transferred to serve as a tutor to UCS (Kalay), as an assistant lecturer to UCS (Monywa), as an assistant lecturer to UCS (Mandalay) as an lecturer to UCS (Myitkyina) and as an lecturer to UCS (Mandalay). The information of my papers are as follows:

1. Paper Title: Implementation and Comparison of Packing Algorithm for Objects Allocation.
   Journal name: Journal of Interdisciplinary Research and Technology, 2020.
2. Paper Title: Simulation Analysis with Queuing in Dental Clinic.
   Journal name: Journal of Information Technology, Research and Innovation (JITRI), 2020.
3. Paper Title: Performance Comparison of Memory Allocation Algorithms.
   Journal name: Journal of Information Technology, Research and Innovation (JITRI), 2023.

INFLUENCE OF SOLVENT MODULATION AND WETTABILITY ON SURFACE MORPHOLOGICAL PROPERTIES OF NATURAL DYES EXTRACT FROM TEAK LEAVES

Win Win Nwe[1], Aye Myint Sein[2], and Thaung Hlaing Win[3]

[1,2,3] Department of Physics, Pakokku University

[1]nnwe5692@gmail.com,[2]thaung.naing001@gmail.com,[3]ayemyintsein@gmail.com

ABSTRACT

Natural dyes obtained from plants, vegetables and fruits are widely studied and tested as low cost sensitizer for dye sensitized solar cell (DSSCs). In the present work, the effect of solvents (distilled water, ethanol and methanol) on optical absorption of natural dye (fresh Teak leaves) was studied. Extracted natural dye was also investigated by using UV-Vis Spectrophotometer. There is an absorbance peak at 436 nm, 477 nm and the other optical absorbance peak at 677 nm. The higher optical absorbance peak was observed in methanol solvent. Monitoring the time evolution of the SPR band peak positions and intensities enabled us to know the stability of natural dyes from Teak leaves. Since they can stable up to 6 weeks in methanol solvent. The surface morphology and wettabiliy of natural dyes was determined by Atomic Force Microscopy (AFM) and Optical microscope is in the range of 50-60 nm.

KEYWORDS

Natural dye, Teak leaves, Solvent, UV-Vis, AFM

1. INTRODUCTION

Solar energy is one of the energy for solving the present and future energy trouble. Solar energy is often referred to as solar cell energy, because solar energy can be converted into electrical energy with the help of solar panels. Solar energy will not run out as long as the sun is still glowing, unlike the energy of fossil materials. One way to convert solar energy into electrical energy is to use solar cells. Sun light is a cheap and clean source of energy. This reliable source of energy could be one of the solutions towards this energy challenges. One of the technologies to utilize sun light energy is photovoltaic. A solar cell is a device that can convert solar energy into electrical energy using photovoltaic effects. The photovoltaic effect was first discovered by Becquerel in 1839 to detect a photo tension when the sun rays the electrodes in an electrolyte solution [1].

In dye sensitized solar cells, several natural dye pigments such as chlorophyll are successfully used as sensitizers. A lot of research is going to improve the efficiency of the natural dye sensitizer by using the combination of dyes, increasing on the concentration of the dye and extraction in various solvents. Natural dye is obtained by extracting from parts of plants such as leaves, flowers and fruit. These organic natural dyes are easily available, inexpensive, non-toxic, environmentally friendly and fully bio-degradable for DSSCs application.

Different types of plant extracts have been used as photo sensitizers in dye sensitized solar cell systems [2]. Organic dye stuffs are highly competitive to serve as dye materials for solar cells because of the low cost of production and the manufacturing process does not require large equipment, but also the process is also easier. The main objective of the present work is to investigate the effect of solvents on optical, contact angle and surface energy of natural dyes Teak leaves.

2. EXPERIMENT

2.1 EXTRACTION OF NATURAL DYES

The natural dyes solution was extracted with different solvents (distilled water, ethanol and methanol employing the following procedure: fresh leaves of teak (Tectonagrand) was washed with water and dried at room temperature. After crushing into small pieces in a pestle, these were kept in a glass bottles filled with solvents; these solutions were kept for one week in the dark at room temperature. Then, the residual (solid) parts were filtered out and the resulting filtrates were used as dye solutions. Figure. 1 (a-b) shows the flow chart and preparation of natural dye-sensitizers for fresh Teak leaves.

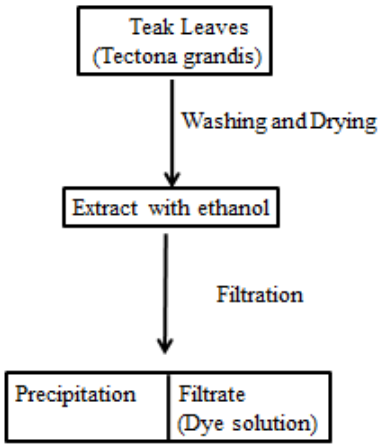


Figure 1. (a) Flow chart for fresh Teak leaves extract

Figure 1. (b) Steps of sample preparation for fresh Teak leaves

2.2 UV-Vis Spectrophotometry

UV-Vis Spectrophotometry is a technique to measure either absorbance or transmittance of the sample in ultraviolet and visible region. The optical properties of natural dyes were measured by UV-Vis spectrophotometer: GENESYS 10S UV-Vis (Materials Research Lab, Department of Physics, and University of Mandalay), as shown in Figure 2. The absorption spectra of natural dyes were recorded in the wavelengths ranging from 300-900 nm. A glass substrate was used as baseline scanning before the measurement.

2.3 Atomic Force Microscopy (AFM)

AFM is a high-resolution non-optical imaging technique. Since then it has developed into a powerful measurement tool for surface analysis. AFM allows accurate and non-destructive measurements of the topographical, electrical, magnetic, chemical, optical, mechanical, etc. properties of a sample surface with very high resolution in air, liquids or ultrahigh vacuum. Figure 3 involves scanning an AFM probe with a sharp AFM tip over a sample surface in a raster pattern. The AFM tip is usually made of silicon or silicon nitride and is integrated near the free end of a flexible AFM cantilever. A piezoelectric ceramic scanner controls the lateral and the vertical position of the AFM probe relative to the surface. The coordinates that the AFM tip tracks during the scan are combined to generate a three-dimensional topographic image of the surface pens able in most advanced science and technology labs around the world.

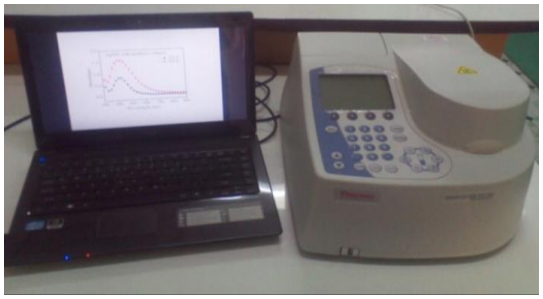


Figure 2. UV-vis absorption spectrophotometer

Figure 3. Atomic Force Microscope

3. RESULTS AND DISCUSSION

Following the effect of solvents (distilled water, ethanol and methanol) on the extraction of natural dyes from fresh Teak leaves were studied.

3.1 Fresh Teak Leaves Extracts

The optical absorption peak of natural dye from Teak solution with distilled water solvent as shown in Figure 4 There is an absorbance peak at 430 nm and 660 nm wavelength. This means that dye extract of teak leaf can work on a blue color spectrum (430-480) nm. With ethanol solvent, the optical absorption peak intensity slightly increased. It is suggested that the concentration of dyes contents increased in ethanol solvent. With methanol solvent, the optical absorption peak is also formed at (530-600) nm wavelengths indicating that the dye can also work or absorb the green and orange spectrum. While the absorbance peak is on the 663 nm wavelength spectrum with the highest absorbance compared to the other. The results of this absorbance test can also be used to indicate the presence of anthocyanin. Based on absorbance data shows that those are absorbance peaks at visible wavelength range of light, it can be indicated that the teak leaf extract specimen contains anthocyanin. This anthocyanin is very sensitive to visible light so the dye solution is a good candidate for dye materials in the DSSC of solar cell.

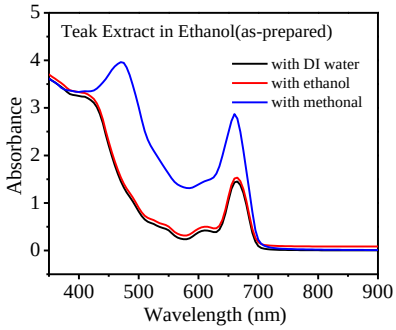


Figure 4. Optical absorption spectra of natural dye solution extracted from tamarind leaves in ethanol (as-prepared)

Figure 5 (a and b) shows the stability studied of natural dye (teak) solution with distilled water and ethanol as prepared and up to 6 weeks. Upon ageing up to two weeks, the optical maximum absorption peak was observed at (λ_{max}= 420 nm, 537 nm, 600 nm and 665 nm) and no color changed. With ethanol solvent, the higher absorption peak intensity was observed than the distilled water solvent. Upon ageing up to three weeks, the absorption peak maximum remained unchanged and peak intensity slightly decreased. This ageing study suggests that the stability of natural dyes (teak) solution (with distilled water and ethanol) is within 2 weeks.

Figure 5. (c) shows the stability studied of natural dye (teak) solution with methanol as prepared and up to 6 weeks. Upon ageing up to five weeks, the optical absorption peak (λ_{max}= 536 nm) intensity slightly increased but decreased up to six weeks. The other absorption peak (λ_{max}= 605 nm, 665 nm), no changed the absorption peak intensity was observed. Upon ageing up to six weeks, the absorption peak maximum remained unchanged and peak intensity significantly decreased. This ageing study suggests that the stability of natural dye (teak) solution is within 5 weeks.

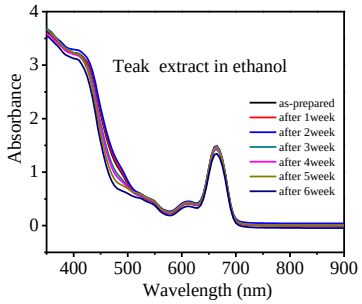


Figure 5. (a) Optical absorption spectra of natural dye solution extracted from Teaks leaves in distilled water (as-prepared and up to 6 weeks)

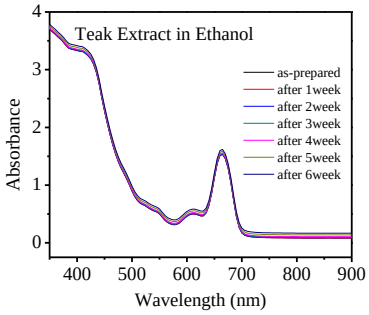


Figure 5. (b) Optical absorption spectra of natural dye solution extracted from Teaks leaves in ethanol (as-prepared and up to 6 weeks)

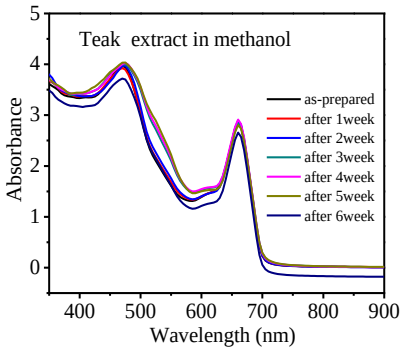


Figure 5. (c) Optical absorption spectra of natural dye solution extracted from teak leaves in methanol (as-prepared and up to 6 weeks)

3.2 Effect of Solvents on AFM images of Natural Dyes from Teak Leaves

The topographies of various surfaces have been studied in many fields due to the significant influence that surfaces have on the practical performance of a given sample. In viewing the opportunity to enhance photovoltaic application, three samples having different solvent (distilled water, ethanol and methanol) were extracted by solution process. Figure 6 (a) shows the. AFM image of Teak leaves films with distilled water. The surface roughness was observed at 0.20 nm. Figure 6 (b-c) shows the. AFM image of Teak leaves films with ethanol and methanol solvents. The surface roughness was observed at 0.40 nm and 1.25 nm. According to the AFM study indicated the entire films exhibited single phase structure for all films. Atomic of roughness increased from 0.20 nm up to 1.25 nm as the solvent changes, the films surface is porous.

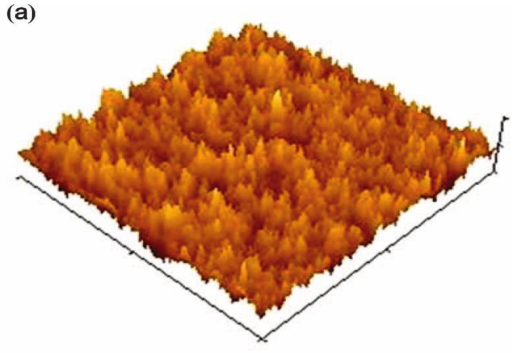


Figure 6. AFM image of Teak leaves films with thickness of (a) distilled water

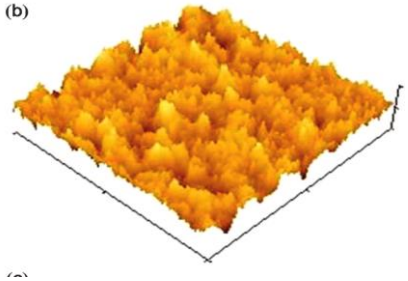


Figure 6. AFM image of Teak leaves films with thickness of (b) ethanol

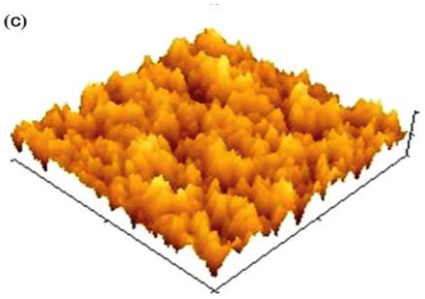


Figure 6. AFM images of Teak leaves films with thickness of (c) methanol

3.3 Effect of solvents on the Wettability of TiO_2 Surface

Wettability involves the interaction between liquid and solid in contact. The wetting behavior of thin film is characterized by the value of contact angle. The contact angle is an important parameter in surface science and its measurement provides a simple and reliable technique for the interpretation of surface energies. Depending on the contact angle, the surface of TiO_2 is classified as shown in table 1.

Table. 1 Classification of TiO_2 substrate on surface energy

$\theta > 150°$	Super-hydrophobic
$65° < \theta < 150°$	hydrophobic
$0° < \theta < 65°$	hydrophilic
$\theta \approx 0$	Super-hydrophilic

Where θ refers to the contact angle. In order to investigate the interfacial properties of the TiO_2 following the different dye concentrations, the wettability of the TiO_2 layers was characterized by contact angle measurements. The surface energy of TiO_2

was also considered from the contact angle measurement.

Figure 7. shows the effect of solvent on contact angle measurement of natural dyes from Teak leaves solutions on TiO_2 coated FTO film with distilled water, ethanol, and methanol. The contact angles $\theta = 32^o$ for bare TiO_2. Effect of solvent (distilled water), the contact angles $(\theta = 25^0), \theta = 20^0$ for ethanol solvent and $\theta = 17^0$ for methanol solvents. The contact angle of liquid drop (fresh Teak solution) on TiO_2 surface is considerably influenced with different solvents, the contact angle decreases and the surface becomes more hydrophilic.

Surface energies of the TiO_2 layers calculated using experimental liquid surface tension and contact angle data and the procedure below.

$$\gamma_l \cos\theta = \gamma_s - \gamma_{sl} \quad (1)$$

To determine solid surface tension, γ_s we invoke an equation of state for interfacial tension.

$$\gamma_s = \frac{\gamma_l (1+\cos\theta)^2}{4} \quad (2)$$

Here, γ_s , γ_l , γ_{sl} are the interfacial surface energies of the solid, liquid and solid- liquid interfaces, respectively. For γ_l, a surface energy of 71.99 mN/m was used for the distilled water, 22.1 mN/m was used for ethanol and 22.7 were used for methanol.

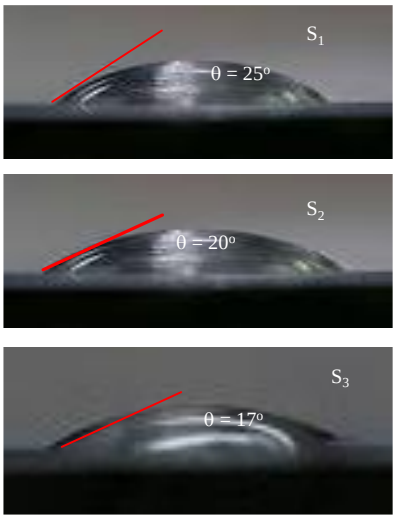


Figure 7. Contact angles of a drop of natural dye from teak leaves solutions on TiO_2/FTO thin film. S_1 (distilled water), S_2 ethanol) and S_3 (methanol)

The surface energies of γ_s, γ_{sl} and contact angle of dye solutions (fresh Teak leaves) on TiO_2 modified and different solvents are summarized in Table 2.

As seen in Table 2, the higher surface energies with methanol solvents of natural dyes from Teak leaves were observed while decreasing the solid-liquid interfacial tension [3]. It may be attributed to the fact that the force of attraction between the liquid molecules and the atoms in the solid becomes more than the force of attraction between the liquid molecules.

Table 2, the higher surface energies with different solvents (distilled water, ethanol and methanol)

Dyes (solvent)	DI water	Ethanol	Methanol
Contact angles	25	20	17
Surface energies (γ_s)	65.4	20.9	21.72
Surface energies (γ_{sl})	0.16	0.02	0.01γ

4. CONCLUSIONS

Natural dyes find use in the colouration of textiles, foods, drugs, and cosmetics. Natural coloring materials can also be extracted using organic solvents such as distilled water, ethanol and methanol. In the first part, the effect of solvent on the optical properties of natural dyes from Teak leaves was studied. The optical maximum absorption peak (λ_{max}) is observed at (437-468) nm and 664 nm in methanol and (435-476) nm and 679 nm in distilled water and ethanol. The maximum absorption peak is relatively higher with methanol solvent for Teak leaves extract. It may be due to its better solubility of natural dye in methanol solvent. In the second study, AFM study indicated the entire films exhibited single phase structure for all films. Atomic of roughness increased from 0.20 nm to 1.2 nm as the solvent changes, the films surface is porous [4]. In another study, the surface

energy of dye solution on TiO_2 film was also calculated from the contact angle measurement. Upon changing solvents, the higher surface energies were observed while decreasing the solid-liquid interfacial tension. It may be attributed to the fact that the force of attraction between the liquid molecules and the atoms in the solid becomes more than the force of attraction between the liquid molecules.

ACKNOWLEDGMENTS

We would like to thank Dr Tin Tun Aung, Acting Rector, Pakokku University. We are also thankful to Dr Win Win Ei (Pro-rector) and Dr Khin Khin Htoot, (Pro-rector), Pakokku University. We would like to express my gratitude to Dr Htun Win (Professor and Head), Dr Win Mar (Professor) and Dr Win Myint Kyu (Professor), Department of Physics, Pakokku University, for their kind permission to give us the opportunity to do research. We would like to express my appreciation to the department of physics, Pakokku University for their use of research facilities.

REFERENCES

[1] J. Liu *et al*., Colloids Surf. A, **302** (2007) 276.
[2] H.S. Shin *et al*., J. Colloid Interf. **274** (2004) 89**.**
[3] J.I. Hussain *et al*., Adv. Mat. Lett. **2** (2011) 188.
[4] S.L. Hsu *et al*., Int. C. Nan. Biosens, **2** (2011) 55.

1. First Author:

Daw Win Win Nwe

Lecturer

Department of Physics

Pakokku University.

2. Second Author:

Daw Aye Myint Sein

Lecturer

Department of Physics

Pakokku University

3. Third Author:

Dr. Thaung Hlaing Win

Lecturer

Department of Physics

Pakokku University.

EFFECT OF REACTION TEMPERATURE ON SIZE AND OPTICAL ABSORPTION PROPERTIES OF COLLOIDAL SILVER NANORODS

San Yu Yu Mon[1], Thiri Win[2], and Myint Myint Kyi[3]

[1,] Department of Engineering Physics, Technological University (Pakokku)
[2,3,] Department of Engineering Physics, Technological University (Magway)
[1]phyu.t.hnin@gmail.com,[2]thiriwin1984@gmail.com,[3]maymoe17416@gmail.com

ABSTRACT

Nanomaterials of noble metals, especially the silver nanorods and silver nanoparticles, have been widely used in different fields of nanoscience such as physics, chemistry, engineering and biochemical fields. In the present work, the formation of colloidal silver nanorods (Ag NRs) have been successfully synthesized by polyol reduction method. In this study, materials were used silver nitrate ($AgNO_3$) as a precursor and polyvinylpyrrolidone (PVP) as a capping agent and stabilizer without adding other chloride ions. The focus of this study was mainly placed on the effect of reaction temperature (150ºC and 170ºC) on the optical absorption of colloidal silver nanorods (Ag NRs). The size and optical absorption of Ag NRs were characterized by UV-vis spectrophotometry and scanning electron microscopy (SEM).

KEYWORDS

Nanoparticles, Reaction temperature, Reducing agent, UV-Vis, SEM.

1. INTRODUCTION

In nanotechnology, nanorods are morphology of nanoscale objects. Each of their dimension ranges from 1-100 nm [1]. One-dimensional metal nanostructures, such as nanorods, nanowires and nanotubes, have been a focused issue of intensive research because of their unique electrical, optical, magnetic, catalytic and other properties being distinctly different from their bulk counterparts [2]. These fascinating properties of nanomaterials strongly depend on size, shape of nanoparticles, their interactions with stabilizers and surrounding media and also on the manner of their preparation. Among the known nanoparticles, silver (Ag) is widely studied because of its characteristic optical and catalytic properties. Preparation of metal nanoparticles such as gold or silver of various shape is very attractive for applications in plasmonics and sensorics. Plasmonic photovoltaics use plasmonic effects inherent to metallic nanostructures to enhance the power conversion efficiency of photovoltaic devices.

The various methods have been actively explored for the synthesis of 1D nanomaterials, including template method, seed-mediated growth method, wet chemical method, chemical reduction method and polyol reduction method. For instance, silver nanorods has been synthesized by polyol reduction method.The polyol is a method to produced metal colloids by heating technique using salt and a solvent like ethylene glycol (EG) and distilled water (DI). The capping agents are generally used to stable the particle and play a role as shape control through selectively binding the facets of nanocrystal. In this experimental technique, the temperature is the most important parameter in controlling nanorod growth due to the kinetic control of the reaction [5].

The main objective of the present work is to investigate the effect of reaction temperature and reaction time on optical absorption and size of Ag NRs. The UV-vis spectrophotometry shows the optical absorption spectra of the Ag NRs. In the case of silver nanorod, the plasmon resonance splits into two modes: one is longitudinal mode and another one is transverse mode. The resonance frequencies as well as the width of the

plasmon absorption band depend on the nanoparticles size. The scanning electron microscopy (SEM) images show the morphologies and sizes of Ag NRs and highly depends on the aspect ratio (i.e. length divided by width) of the nanorods [6].

2. EXPERIMENT

2.1 Synthesis of Colloidal Silver Nanorods (Ag NRs)

Colloidal silver nanorods (Ag NRs) have been developed by polyol reduction method. Firstly,1.25 mM of PVP will dissolved in 20 ml of distilled water, the solution was mixed by stirring for about 15min and the mixture was heated to150ºC.Next, 20 ml distilled water containing 0.1 M ($AgNO_3$) solution was injected into this solution all at one while stirring vigorously. The mixing solution was continuous stirring and heating for about 2 hr and the color of the reaction solution changed to brown-color due to formation of silver nanorods. The reaction temperature was changed from 150ºC to 170ºC to investigate the effect of a reaction temperature on optical absorption spectra of colloidal Ag NRs. In addition, the effect of reaction time (10min and 15min) on optical absorption of silver nanorods was also studied. Preparation steps for colloidal silver nanorods and experimental setup for the synthesis of colloidal Ag NRs are shown in Figure 1 and 2.

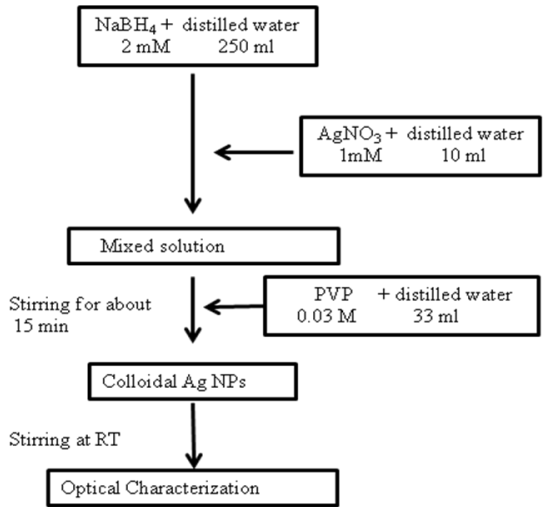


Figure 1. Preparation steps for colloidal silver nanorods

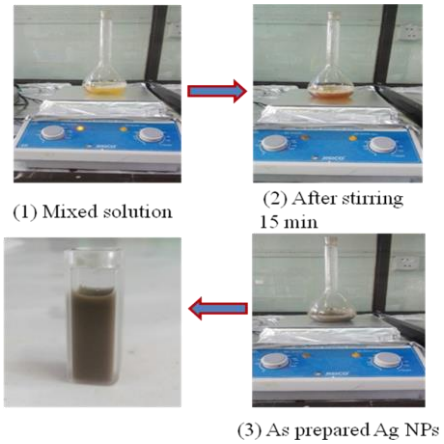


Figure 2. Experimental setup for the synthesis of colloidal Ag NRs

2.2 Characterization Methods

The characterization methods include UV-vis spectrophotometer and scanning electron microscopy (SEM).

2.2.1 UV-vis Spectrophotometry

UV-vis spectrophotometry is a technique to measure absorbance of the sample in the visible region. The UV-vis absorption spectra were measured at room temperature using the GENESYS 10S UV-vis spectrophotometry (Materials Research Lab, Department of Physics, University of Mandalay) Figure 3. The scanning ranges of Ag NRs were recorded from 300 nm to 1000 nm. A glass substrate was used as baseline scanning before the measurement. Baseline measurement was made with distilled water as reference solution. The solution was diluted 10 times with distilled water (DI) before testing. The purpose of this measurement is to investigate the effect of temperature on the optical absorption of colloidal Ag NRs.

2.2.2 Scanning Electron Microscopy (SEM)

Scanning Electron Microscopy (SEM) is a type of electron microscope, which is generally used to directly observe the morphology and size of nanostructures. The SEM images of silver nanorods were measured using the Agilent 8500 SEM as shown in Figure 4. For investigations by SEM was operated at an operating voltage of 15.0kV and a current of 2.2 A. To perform the SEM measurements, we used dip-coating method. Firstly, a thin film was washed in DI water to remove the excess surfactant and dispersed in PIA solution. After that, this thin

film was dipping into the Ag NRs solution and dried in air. The morphological features of metal NRs assembled on Si substrate by SEM. The shape of nanostructures could be observed clearly under SEM.

Figure 3. UV-vis Spectrophotometry (GENESYS 10S)

Figure 4. Scanning Electron Microscopy (Agilent 8500 SEM)

3. RESULTS AND DISCUSSION

This chapter mainly discusses the effect of reaction temperature and reaction time on the optical absorption spectra and shape and size of colloidal silver nanorods (Ag NRs).

3.1 Effect of Reaction Temperature on Absorption Spectra of Ag NRs

In the present study, silver nanorods (Ag NRs) were synthesized by polyol reduction method. It is very interesting to investigate the optical properties of Ag NRs since it strongly absorbs light in the visible region due to surface plasmon resonance effect. Basically, silver nanorods give rise to two absorption bands. One of them is caused by transverse oscillations of the electrons along the long axis (transverse absorption band), which is typically located in the visible wavelength region, and one due to the longitudinal oscillations of the electrons along the short axis (longitudinal absorption band). The red-shift implies the formation of nanorods, which agrees with the color change into opaque brown solution.

The optical absorption and size of the resulting silver NRs was strongly dependent upon the reaction temperature. Figure 5. (a-b) shows the optical absorption spectra of colloidal Ag NRs with different reaction temperature 150ºC and 170ºC.At the reaction temperature (150ºC), TSPR peak maximum (λ_{max}) at 430 nm and the other LSPR peak maximum (λ_{max}) at 550 nm. Upon increasing reaction temperature (170ºC), TSPR peak maximum (λ_{max}) is reached at 400 nm and LSPR peak maximum (λ_{max}) at 720 nm. TSPR peak corresponds to the diameter of the Ag NRs and LSPR peak corresponds to the length of the Ag NRs. By increasing reaction temperature, the longitudinal surface plasmon resonance of Ag NRs was red-shifted, thus suggesting the larger aspect ratio (longer nanorods) with the reaction temperature (170ºC).

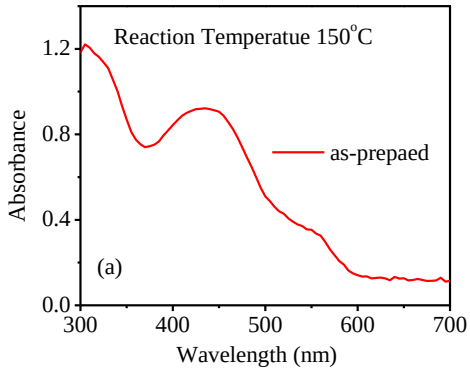
To support our speculation, the scanning electron microscopy (SEM) was employed to estimate the size of Ag NRs on silicon substrates. The Ag NRs were transferred onto the silicon substrates and their sizes were estimated using scanning electron microscopy (SEM). Figure 6. (a-b) shows the SEM images of Ag NRs with different reaction temperature 150ºC and 170ºC. SEM investigation shows that both reaction temperature, the average diameter of NRs around 200 nm and longer nanorods. It indicated that Ag NRs with larger aspect ratios were formed. Achieving the longer nanorods with reaction temperature (170ºC) results in a larger aspect ratio of NR which agrees with the previous spectroscopic results (see Figure 5 (a-b)).

3.2 Effect of Heating Time on Absorption of Colloidal Ag NRs

We have also studied the effect of heating time on the absorption spectra of colloidal Ag NRs for reaction temperature (150ºC and 170ºC). Figure 7 shows the optical absorption spectra of colloidal Ag NRs with different heating time (5 min up to 15 min) for 150º C. Upon the increasing reaction time (5 min up to 10 min), (TSPR and LSPR) peak maximum (λ_{max}) remained unchanged and peak intensity increased. It can be attributed to higher surface charge density of colloidal Ag NRs. Upon increasing reaction time (up to 15 min), peak

intensity (TSPR and LSPR) decreased. It can be attributed to lower surface charge density of colloidal Ag NRs.

Figure 8 shows the optical absorption spectra of colloidal Ag NRs with different heating time (5 min up to 15 min) for reaction temperature 170ºC. When the heating time was increased about at up to 15 min, (TSPR and LSPR) peak intensity increased and slightly blue-shifted. It can be attributed to lower aspect ratio and shorter nanorods (NRs).



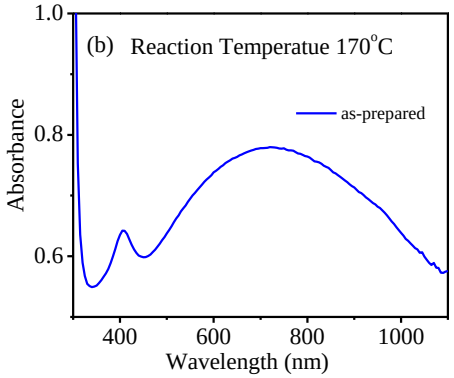
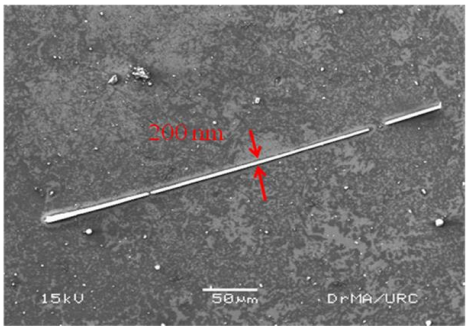


Figure 5. Optical absorption spectra of colloidal Ag NRs with different reaction temperature (a) 150ºC, (b) 170ºC.



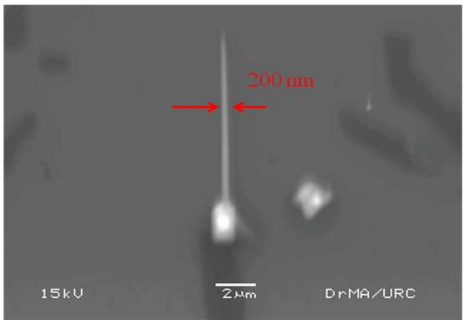


Figure 6. SEM images of Ag NRs with reaction temperature (a) 150ºC and (b) 170ºC.

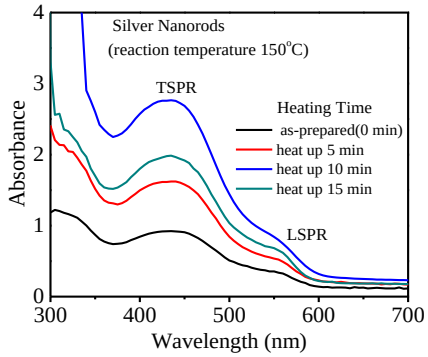


Figure 7 Absorption of colloidal Ag NRs with different heating time (5 min up to 15 min) for reaction temperature 150ºC

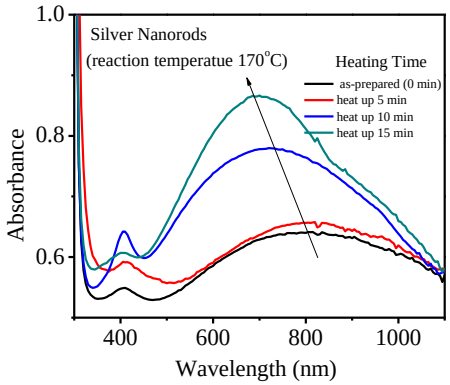


Figure 8 Absorption of colloidal Ag NRs with different heating time (5 min up to 15 min) for reaction temperature170ºC

3.3 Effect of Reaction Time on Optical Absorption of Silver Nanorods

In another studied, Figure 9 shows the effect of reaction time on the absorption spectra of colloidal silver nanorods (NRs) for about 10 min and 15min). At the reaction time (10min), TSPR peak maximum (λ_{max}) at 325 nm and the other LSPR peak maximum (λ_{max}) at 440 nm.

Upon increasing reaction time (15 min), TSPR peak maximum (λ_{max}) is reached at 330 nm and LSPR peak maximum (λ_{max}) at 450 nm. TSPR peak corresponds to the diameter of the Ag NRs and LSPR peak corresponds to the length of the Ag NRs. By increasing reaction time, the longitudinal surface plasmon resonance of Ag NRs was red-shifted, thus suggesting the longer nanorods with the reaction time (15 min).

Figure 10 (a-b) shows the SEM images of Ag NRs with different reaction time (10 min and 15 min). SEM investigation shows that both reaction time, the average diameter of NRs around 350 nm for 10 min and 280nm for reaction time 15 min. Upon increasing heating time, the longer nanorods and smaller diameter ANRs was observed. It indicated that Ag NRs with larger aspect ratios were formed. Figure 5 (a,b) shows achieving the longer nanorods with reaction time 15 min which agrees with the previous spectroscopic.

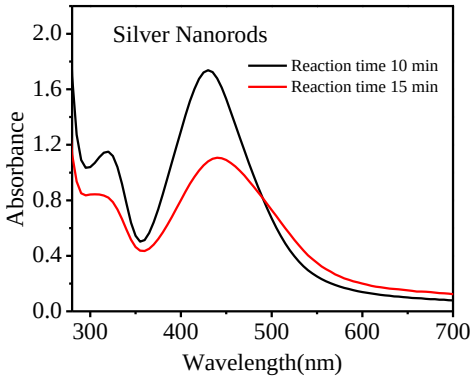
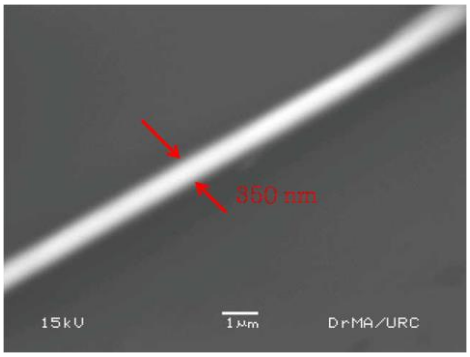


Figure 9. The effect of reaction time (10 min 15min) on the optical absorption of silver nanorods



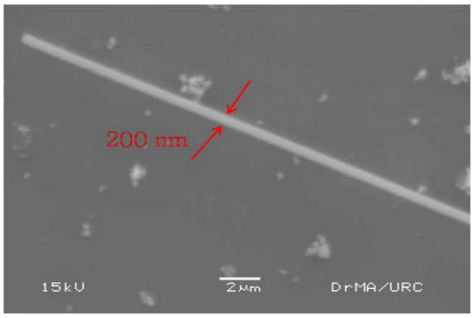


Figure 10 SEM images of Ag NRs with reaction time (a) 10 min and (b) 15 min

Figure 11 shows the effect of reaction time on the absorption spectra of colloidal silver nanorods (NRs) for about 10 min (as-prepared, 1hr, 2 hr and 3 hr). Upon increasing (1hr up to 2hr), the color and absorption peak maximum remained unchanged but the absorption peak intensity significantly increased. It suggests the higher reducing power of silver ions upon increasing reaction time. And then, upon increasing up to 3hr, the color and absorption peak maximum remained unchanged but the absorption peak intensity significantly decreased. It can be attributed to larger diameter of colloidal silver nanorods.

Figure 12 shows the study of colloidal Ag NRs with reaction time 15 min (as-prepared, 1hr, 2hr and 3 hr). Upon increasing reaction time (1hr up to 3 hr), the color and absorption peak maximum remained unchanged but the absorption peak intensity significantly decreased. It can be attributed to larger diameter of colloidal Ag NRs. According to the above results, upon increasing the reaction time over about 1hr, the optical absorption of colloidal silver nanorods was poor.

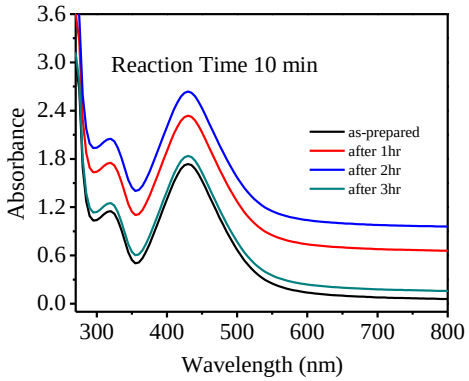


Figure 11. The reaction time (10 min) on the optical absorption of silver nanorods (as-prepared, 1 hr, 2 hr and 3 hr)

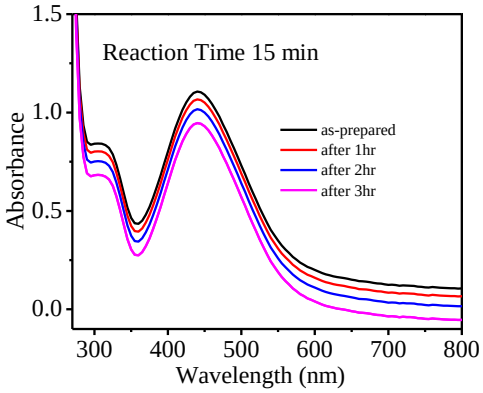


Figure 12. The reaction time (15 min) on the optical absorption of silver nanorods (as-prepared, 1 hr, 2 hr and 3 hr)

4. CONCLUSIONS

Colloidal silver nanorods (Ag NRs) were successfully synthesized using polyol reduction method with different reaction temperatures (150°C and 170°C).It is found that the temperature during the synthesis process influences significantly to the formation AgNRs. In this work, the effect of reaction temperature and reaction time on optical absorption and shape and size of silver nanorods was mainly investigated using UV-vis spectrophotometry and scanning electron microscopy (SEM). By increasing reaction temperature, the longitudinal surface plasmon resonance of Ag NRs was red-shifted suggesting larger aspect ratio or longer nanorods. SEM investigation shows that average diameter of NRs around 200 nm and longer nanorods with reaction temperature (170ºC). In addition, the effect of reaction time at (10 min and 15min) was also studied. According to the SEM images result, the smaller nanorods (200 nm) were observed at reaction time 15 min. Achieving the longer nanorods result in a which agrees with the previous spectroscopic results. An additional effect of reaction time on the λ_{max} (SPR) and absorption peak intensity was also studied. It was found that the λ_{max} (TSPR) was remained unchanged and (LSPR) of Ag NRs was blue-shift suggesting lower aspect ratios or shorter nanorods.

ACKNOWLEDGEMENTS

The authors are thankful to Dr. Myint Myint Khaing, Rector, Technological University, Pakokku, for her permission and encouragement to conduct this research work. Special thanks go to members of Materials Research Lab, Department of Physics, University of Mandalay, for their help in experimental work.

REFERENCES

[1] B. Dr. Orhan *et al*., Syn. Appi. Nano, **2** (2012) 978

[2] J.K. Suganuma *et al*., Mat. Chem. Phys, **114** (2009) 333

[3] P.A. Majkova *et al*., Ins. Slo. Sci, **2** (2014) 13

[4] R. Erdal *et al.*, Sil. Nano. Char. App, **102** (2018) 376

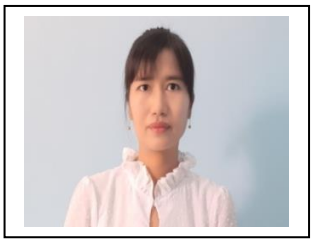
[5] J.K. Harsojo *et al*., Adv. Mat. Res, **1123** (2015) 256

[6] D. Muthuvinothini *et al*, E. J Nanotech, **2** (2015) 1–3

[7] A.N. Taufiq-Yap *et al.*, Nanorods. Synthesis.Dev. App, **12** (2016) 1136

[8] Z.Y. Zhang *et al*., Science, **276** (1997) 377

[9] W. Whatmore, Nanotech. Percep, **1**(2005) 67



1. First Author:
Daw San Yu Yu Mon
Lecturer
Department of Engineering Physics
Technological University (Pakokku).

2. Second Author:
Daw Thiri Win
Lecturer
Department of Engineering Physics
Technological University (Magway).

3. Third Author:
Daw Myint Myint Kyi
Lecturer
Department of Engineering Physics
Technological University (Magway).

INVESTIGATION OF SIZE AND OPTICAL PROPERTIES OF COLLOIDAL SILVER NANOSRTRUCTURES BY SEED-MEDIATED GROWTH METHODS

Thiri Win[1], Myint Myint Kyi[2], and San Yu Yu Mon[3]

[1,2,] Department of Engineering Physics, Technological University (Magway)
[3,] Department of Engineering Physics, Technological University (Pakokku)
[1]thiriwin1984@gmail.com,[2]mayome17416,[3]phyu.t.hnin@gmail.com

ABSTRACT

Colloidal silver nanoparticles (AgNPs) and nanorods (Ag NRs) were prepared by chemical reducing methods and seed-mediated growth method. Synthesis of colloidal silver nanoparticles which involves two steps: preparation of Ag seed solution followed by growth solution. The focus of this work was mainly placed on the effect of reducing agent (Ascorbic acid and Sodium borohydride) on optical absorption of colloidal silver nanorods. In the case of colloidal Ag NRs where two extinction peaks are pronounced corresponding to the transverse and longitudinal surface plasmon resonance (TSPR and LSPR), increasing reducing agents (RA) concentration enabled a red-shift of LSPR suggesting a larger aspect ratio of NRs. The characterization methods used were the UV-vis spectrophotometry, X-ray diffraction (XRD) and Scanning Electron Microscopy (SEM). According to the SEM micrograph, the size of silver nanorods was observed at ~ 250 nm with (AA) and ~ 600 nm with $NaBH_4$.

KEYWORDS

Silver nanostructures, absorption, optical, reducing agent.

1. INTRODUCTION

Nanostructures have been extremely attractive because of their unique properties (e.g., optical, electronic, magnetic and chemical properties) and potential applications in nanoscience and nanotechnology [1]. Important scientific and technological advances based on understanding and control of properties and process at the scale of atoms and molecules (nanometre scale) are taking place in laboratories around the world. Metal nanostructures show unusual and interesting optical properties due to their ability to sustain surface plasmons, which are oscillating electrons that exist at the interface between a real metal and a dielectric [2]. Extreme light concentration and local field enhancement provided by metal nanostructures have lead to tremendous applications such as sensing, optical waveguides, nano-antennas, Raman spectroscopy, near-field optics, and non-linear optics. Fundamental science research and device technologies involving surface plasmons have been among the most active domains in physics, engineering, chemistry, and biology.

The ability to concentrate light to the nano-scale by metal nanostructures comes from their resonant and non-resonant behavior. Firstly, when the nanostructure's dimension is significantly smaller than the excitation wavelength, the resonant effect, localized surface plasmon resonance, occurs at specific optical frequencies and leads to a strong charge displacement and associated field concentration. These resonances are very sensitive to the type of the metal, the structure's shape and the property of the background dielectric. Secondly, non-resonant broadband field enhancement can be achieved due to surface plasmon polarities' propagation along the metal structures such as the feed-gap, lightning rod, and metal cones or wedges, for which enhancing results are strongly dependent on the structures' dimensions. The position and magnitude of the surface plasmon absorption are intensively dependent on the

size of the particles, the refractive index of the solvent and the degree of aggregation. The main objective of the present work is to investigate the effect of reducing agent, sodium borohydride ($NaBH_4$) on the optical extinction and size of silver nanoparticles (Ag NPs) and silver nanorods (Ag NRs).

2. EXPERIMENT

This chapter deals with [the synthesis and characterization of colloidal silver nanoparticles (Ag NPs), and silver nanorods (Ag NRs)] of the present work.

2.1 Synthesis of Colloidal Seed Silver Seed Solution

In the first part of the present work, colloidal silver nanoparticle (Ag NPs) solutions have been developed by a chemical reduction method. In the synthesis, $AgNO_3$ and polyvinylpyrrolidone (PVP) were used, as a metal salt precursor and a stabilizing agent, respectively. Sodium borohydride ($NaBH_4$) was used as a reducing agent. 18.5 mg of $NaBH_4$ was added to 250 ml of distilled water to create a 2 mM solution while 17 mg of $AgNO_3$ was added to 20 ml of distilled water to create a 1mM solution. The sodium borohydride was chilled to roughly 0° C in the ice bath and stirred for about 15 minutes. And then 2 ml of $AgNO_3$ solution was added to the $NaBH_4$ solution using a dropper, at about one drop per second. Upon the complete addition of the $AgNO_3$, stirring was stopped. The solution appeared light yellow, indicating the presence of nanoparticles. Figure 1 shows preparation steps of silver seed solutions.

2.2 Preparation of Silver Nanorods

A 20 ml aqueous solution containing 0.1 M cetyltrimethylammonium bromide (CTAB) was prepared. This solution was heated to 80^o C while stirring to dissolve the CTAB. Next, 4 ml of 0.05 M $AgNO_3$ and 4 ml of 0.1 M freshly prepared ascorbic acid (AA) were added.

And 100 µl of silver seed solution (as-obtained) was added. The molar ratios (AA to $AgNO_3$) were changed to 0.1:1, 0.2:1, 0.4:1 and 0.8:1. Within 2 min, the solution color changed into greenish gray. A 1.67 mM of $NaBH_4$ was prepared in ice-cold water. A total of 50 µL of freshly prepared ice-cold $NaBH_4$ was added to the growth solution. The growth of nanorods was observable within a few minutes after the addition of $NaBH_4$. Flow charts for the synthesis of colloidal Ag NRs are shown in Figure 2.

Figure 1. Preparation steps of silver seed solutions

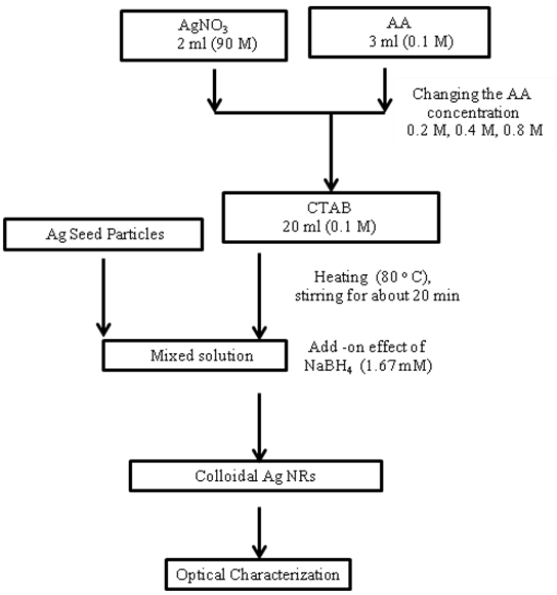


Figure 2. Flowchart of preparation steps for colloidal silver nanorods

3. CHARACTERIZATION

The characterization methods include UV-vis spectrophotometry, FESEM microscopy and X-Ray Diffraction (XRD).

3.1.UV-Vis Spectrophotometry

UV-Vis Spectrophotometry is a technique to measure either absorbance or transmittance of the sample in ultraviolet and visible region. The optical properties of colloidal nanoparticles and nanorods were measured by UV-Vis spectrophotometer: GENESIS 10S

UV-Vis (Materials Research Lab, Department of Physics, and University of Mandalay) Figure 3. The absorption and transmission spectra of NPs and NRs were recorded in the wavelengths ranging from 300-900 nm. A glass substrate was used as baseline scanning before the measurement.

3.2. X-ray Diffraction

The structural characterization of AgNPs was performed by X-ray diffraction. X-ray diffraction (XRD) is a technique to measure the crystal structure of powder or film and the crystallite size of the sample. Figure 4 shows the X-ray diffractometer (BRUKER D8 ADVANCE). Cu K_α radiation of wavelength 1.54056 Å was used.

The observed peaks are assigned the correct Miller indices (h, k and l). Operating conditions are 40 kV and 40 mA. The recorded XRD patterns were analyzed to determine peak position, width and intensity.

3.3. Field Emission Scanning Electron Microscopy

AgNPs solution was transferred onto the silicon substrates by a transfer dip-coating method. Ag NPs solution will run over the wafer relatively evenly, producing a thin film of Ag NPs. The morphological features of metal NPs assembled on Si substrate by field-emission scanning electron microscopy: Agilent 8500 FE-SEM (Defence Services Science and Technology Research Centre (DSSTRC), Pyin Oo Lwin) as shown in Figure 5.

It is a compact field emission scanning electron microscope (SEM) for imaging a variety of nanoscale objects and materials. It is controlled by a desktop computer that is connected to the system via a USB cable. The 8500 FE-SEM software runs on this computer and provides point-and-click operation, configuration and status monitoring of the 8500 FE-SEM.

The images were obtained at an operating voltage of 1000 V and a current of 2.2 A. The size of Ag NPs and Ag NRs thin film were estimated from FESEM image by using "image j" analyzing software.

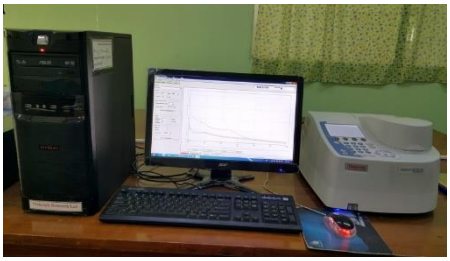


Figure 3. UV-Vis spectrophotometer:

Figure 4. RIGAKU Multiflex X-Ray Diffractometer

Figure 5. Field Emission Scanning Electron Microscopy

4. RESULTS AND DISCUSSION

4.1. Absorption of Seed Solution by XRD Method

Figure 6 shows the absorption spectra of colloidal Ag NPs. The maximum absorption peak was observed at 401 attributing to the s-p transition of electrons. The surface morphological fractures of AgNPs were estimated using X-ray diffraction (XRD). Figure 8 shows the XRD pattern of AgNPs in distilled water solvent. Two diffraction peaks appeared at 2θ angles of 43.33° and 50.40°

corresponding to (111) and (200) planes of face centred cubic (fcc) structure [3].

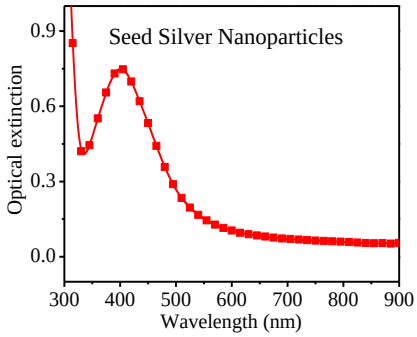


Figure 6. Optical absorption spectra of colloidal silver nanoparticles

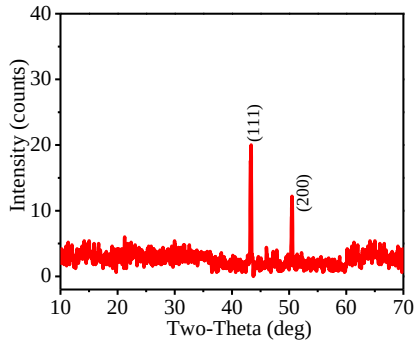


Figure 7. XRD measurements of colloidal silver seed nanoparticles

4.2. Effect of Reducing Agent (AA) Concentration

In addition to colloidal silver nanorods were also developed by seed-mediated growth method. In the case of silver nanorods, electron oscillation can occur in one of two directions (along the short and long axes) depending on the polarization of the incident light. The excitation of the surface plasmon oscillation along the long axis induces a much stronger absorption band in the longer wavelength region, referred to as the longitudinal surface plasmon resonance (LSPR) band. The morphology and aspect ratio of the silver nanorods strongly depend on the reducing agent concentration [4].

As in the previous nanostructures (Ag NPs), we studied the effect of reducing agent (AA) concentration on extinction properties of colloidal Ag NRs. Figure 8 shows the extinction spectra of colloidal Ag NRs with different molar ratios of ascorbic acid to silver nitrate (0.1:1, 0.2:1, 0.4:1 and 0.8:1). Upon increasing reducing agent concentration, the color remained unchanged. For all the molar ratios of ascorbic acid to silver nitrate being studied, two peaks appear for the silver plasmon band. One appeared at 415 nm (for ratio 0.1:1), 420 nm (for ratio 0.2:1), 420 nm (for ratio 0.4:1) and 420 nm (for ratio 0.8:1). These peaks are conventional plasmon band for spherical silver nanoparticles called transverse surface plasmon resonance (TSPR) band. TSPR peak maximum (λ_{max}) is almost invariant upon increasing reducing agent concentrations. A separate plasmon band appeared in the longer wavelength region due to the longitudinal plasmon band of rod-shaped particle and its position depends on the aspect ratio of the rod. The larger aspect ratio, the more red shifted the longitudinal plasmon band. The longitudinal plasmon resonance (LSPR) band of silver nanorods synthesized at different molar ratios appeared at 570 nm (for ratio 0.1:1), 565 nm (for ratio 0.2:1), 549 nm (for ratio 0.4:1) and 533 nm (for ratio 0.8:1) wavelengths, respectively. LSPR band maximum (λ_{max}) is blue-shifted upon increasing reducing agent concentrations which can be ascribed to the higher reducing power of silver ions at higher reducing agent concentration. The colloidal silver nanorods (molar ratios 0.1:1) were transferred onto the silicon substrates and their sizes were estimated using scanning electron microscopy (SEM). Figure 9 shows SEM images of colloidal Ag NPs. The average NRs size is 260 nm for the molar ratio of 0.1:1.

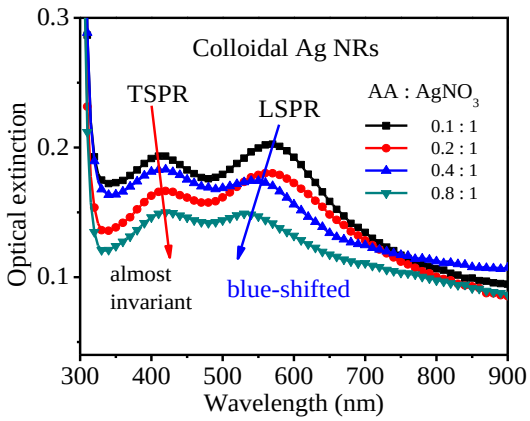
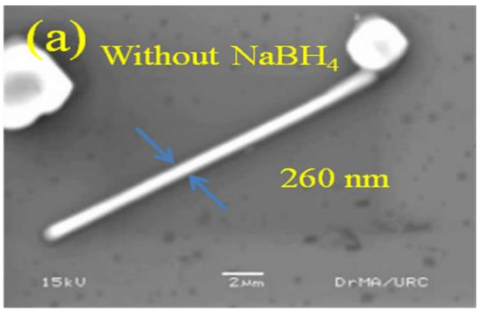


Figure 8. Optical absorption spectra of Ag

NRs with different molar ratios of AA to AgNO3 (0.1:1, 0.2:1, 0.4:1 and 0.8:1)



(0.1:1)

Figure 9. SEM images of colloidal Ag NRs with molar ratios of ascorbic acid to silver nitrate (0.1:1)

4.3. Add-on Effect of Strong Reducing Agent ($NaBH_4$)

The formation of Ag NRs by seed-mediated growth method involves at least two steps: preparation of seed particles solution and growth of nanorods from seed particles solution. In this section, we have discussed the effect of strong reducing agent ($NaBH_4$) on optical extinction of Ag NRs grown from as-obtained seed solution. Figure 10. (a-c) shows the extinction spectra of colloidal Ag NRs [different molar ratios of AA to $AgNO_3$ (0.1:1), (0.2:1), (0.4:1)] grown from seed solutions with and without $NaBH_4$. Upon introducing $NaBH_4$, both LSPR and TSPR peak positions were almost invariant. However, LSPR and TSPR peak intensity significantly increased which can be attributed to higher surface charge density of colloidal Ag NRs induced by $NaBH_4$, thus expecting the smaller diameter of Ag NRs. In order to determine the size of Ag NRs, scanning electron spectroscopy (SEM) images were acquired.

Figure11. (a-b) shows the SEM images of colloidal Ag NRs [AA to $AgNO_3$ ratios of (0.1:1)] grown from aged seed solutions with and without $NaBH_4$. With $NaBH_4$, average diameter of NRs decreased from ~ 700 nm to ~ 590 nm for (0.1:1). The smaller diameter of Ag NRs with $NaBH_4$ (observed from SEM) agrees well with our speculation that increased extinction band (LSPR) of NRs with $NaBH_4$ would produce smaller diameter of NRs.

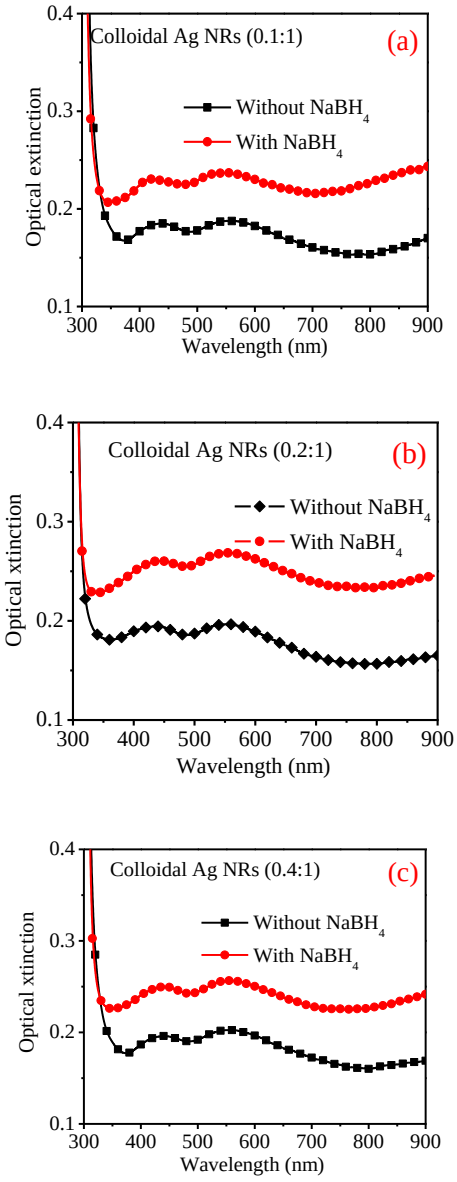


Figure 10. (a-c) Optical absorption of Ag NRs (molar ratios 0.4:1) with and without strong reducing agent ($NaBH_4$)

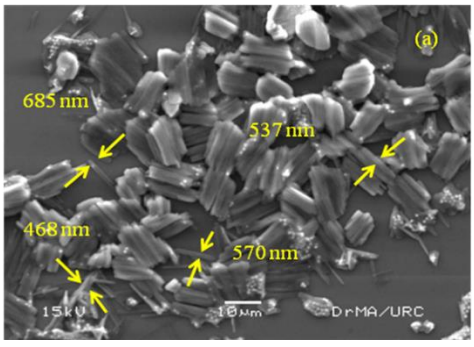


Figure 11. (a) SEM images of Ag NRs with $NaBH_4$ at molar ratios of ascorbic acid to silver nitrate (0.1:1)

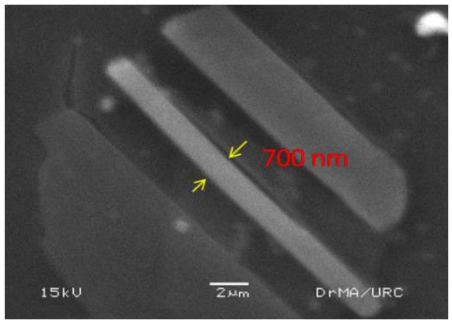


Figure 11. (b) SEM images of colloidal Ag NRs without $NaBH_4$ at molar ratios of ascorbic acid to silver nitrate (0.1:1)

5. CONCLUSIONS

Metal nanostructures have attracted considerable attention due to both fundamental interest and potential applications in nanoscience and nanotechnology. Colloidal silver (Ag NPs) and silver nanorods (Ag NRs) were developed by chemical reduction method and seed mediated growth method. The size and shape of nanostructures can be easily controlled by varying type and concentration of reducing agents (RA) and stabilizer. In this work, the effect of reducing agent concentration on size and optical absorption of colloidal Ag NPs and Ag NRs was mainly investigated using UV-vis spectrophotometry, XRD and SEM. It can be concluded that in the case of colloidal Ag NPs, the absorption band maximum (λ_{max}) was observed at 401 nm and surface morphology of silver nanoparticles which was confirmed by XRD micrographs.

Moreover, as an add-on effect, introducing the reducing agent (AA and $NaBH_4$) resulted in a blue-shift of λ_{max} and increase of EBI. In the case of colloidal Ag NRs where two absorption peaks are pronounced corresponding to the transverse and longitudinal surface plasmon resonance (TSPR and LSPR), increasing reducing agent (AA) concentrations enabled a blue-shift of LSPR suggesting a lower aspect ratio of NRs. An additional effect of strong RA ($NaBH_4$) on the λ_{max} (LSPR) and EBI was also studied. It was found that the λ_{max} (LSPR) was blue-shifted and EBI increased, and the diameter of NR was observed at around 260 nm with reducing agent (AA). With $NaBH_4$, average diameter of NRs decreased from ~ 700 nm to ~ 590 nm for ratios (0.1:1). The smaller diameter of Ag NRs with $NaBH_4$ (observed from SEM) agrees well with our speculation that increased extinction band (LSPR) of NRs with $NaBH_4$ would produce smaller diameter of NRs. This work demonstrates that, the RA ($NaBH_4$) plays a key role in tuning the size and shape of nanostructures, thus improving the optical extinction properties.

ACKNOWLEDGEMENTS

Special thanks go to members of Materials Research Lab, Department of Physics, University of Mandalay, for their help in experimental work.

REFERENCES

[1] J. Liu *et al*., Colloids Surf. A, **302** (2007) 276.

[2] H.S. Shin *et al*., J. Colloid Interf. **274** (2004) 89**.**

[3] J.I. Hussain *et al*., Adv. Mat. Lett. **2** (2011) 188.

[4] S.L. Hsu *et al*., Int. C. Nan. Biosens, **2** (2011) 55.

1. First Author:
Daw Thiri Win
Lecturer
Department of Physics
Technological University (Magway).
2. Second Author:
Daw Myint Myint Kyi
Lecturer
Department of Physics
Technological University (Magway).
3. Third Author:
Daw San Yu Yu Mon
Lecturer
Department of Physics
Technological University (Pakokku).

EFFECT OF REDUCING AGENTS CONCENTRATION ON SURFACE PLASMON RESONANCE PROPERTIES OF COLLOIDAL SILVER NANOPARTICLES

Myint Myint Kyi[1], Thida Nwe[2], and Thiri Win[3]

[1,2,3] Department of Engineering Physics, Technological University (Magway)
[1]thaung.naing001@gmail.com,[2]thidanwe@gmail.com,[3]thiriwin1984@gmail.com

ABSTRACT

The synthesis of nanomaterials is currently one of the most active in nanoscience branches. Silver nanoparticles (AgNPs) are one of the most important nanoparticles used in catalysis and photonics. In the present work, colloidal silver nanoparticles (AgNPs) were prepared by chemical reduction of metal salt silver nitrate ($AgNO_3$) with the aid of reducing agent tri sodium citrate ($Na_3C_6H_5O_7$). The famous of this study was mainly placed on the effect of reducing agent (sodium citrate and ascorbic acid) concentrations (1%, 2%, 3%) on optical absorption of colloidal silver nanoparticles. The optical and surface features of silver nanoparticles were characterized by UV-vis spectrophotometry and scanning electron microscopy (SEM). The maximum absorption peak of colloidal silver nanoparticles was observed at ~ 420 nm and the stability of colloidal silver nanoparticles is ~ 15 days.

KEYWORDS

Silver nanoparticles, Reducing Agent, Concentration, UV-Vis, SEM.

1. INTRODUCTION

Nanostructures have been extremely attractive because of their unique properties (e.g., optical, electronic, magnetic and chemical properties) and potential applications in nanoscience and nanotechnology [1]. Nanoparticles are usually the particles of size 1nm to 100nm. Physicochemical properties of nanopaSilver nanoparticles (AgNPs) are increasingly used in various fields, including medical, food, health care, consumer, and industrial purposes, due to their unique physical and chemical properties. These include optical, electrical, and thermal, high electrical conductivity, and biological properties such as mechanical, chemical, magnetic, optical or electrical properties are different compared to bulk materials for the volume to surface ratio. Nanotechnology is now a day's spreading rapidly and researchers are focusing more on nanoscience due to increasing usage and applications of nanomaterials based on their various properties [2].

In particular, silver nanoparticles (AgNPs) have become an intensive area of scientific research because silver nanoparticles are now being widely used in different cases such as industrial usage, optical sensors, catalysts and organic solar cells. There are two approaches of synthesis of nanoparticles: the bottom-up and top-down approaches [3].

Top-down approach is a separation method by which an external force is applied to a solid (bulk) that leads to its fragmentation into smaller particles. Ball milling and attrition are example of top-down approach. Bottom-up approach is to form the nanoparticles from atoms of gas or liquid, based on atomic transformation or molecular condensations [4]. In this study, we will follow the bottom-up approach to prepare silver nanoparticles.

There are many methods of synthesis silver nanoparticles including physical method, chemical reduction method, biological method and liquid-liquid method [5] and so on. Among these methods, we used chemical reduction method.

This method is simple, cost effective and also it gives the desire shape of silver nanoparticles. We can control the size and size distribution of colloidal nanoparticles by optimizing the

experimental parameters, such as the molar ratio of capping agent to the precursor salt, the molar ratio of reducing agent to the precursor salt and so on [6].

UV-vis spectrophotometry is used to measure absorption spectra of the prepared colloidal silver nanoparticles. The position and magnitude of surface plasmon absorptions are intensively dependent on the size of metal nanoparticles and the degree of aggregation. The absorption wavelength of silver nanoparticles is about 420 nm.

The absorption peak maximum (λ_{max}) can be used to get information about the size of the nanoparticles. Scanning Electron Microscopy (SEM) is used to measure average size and size distribution of nanoparticles. In this study, we demonstrated the effect of reducing agent concentrations (sodium citrate) and surface plasmon resonance properties of colloidal silver nanoparticles.

2. EXPERIMENT

2.1 Synthesis of Colloidal Silver Nanorods (Ag NRs)

The silver nanoparticle colloids were prepared by chemical reduction method. Silver nitrate ($AgNO_3$) was as a metal salt precursor, sodium citrate ($Na_3C_6H_5O_7$) was used as reducing agent and as well as stabilizing agent.

In this experiment, we used various sodium citrate concentrations (1%, 2% and 3%). All the solutions were prepared in distilled water. In typical experiment, 0.02 g of $AgNO_3$ was dissolved in 100 ml of distilled water to get 1mM solution. This solution was heated with continuous stirrer for about 30 min at 150°C.

After the solution reaches boiling temperature we added 10 ml of 39 mM (sodium citrate solution) one drop per second with constant stirring for 10 min. After that the solution changed into yellow color. At that time, heating was stopped and the solution was cooled at room temperature with continuous stirring. The change in yellow color represents the production of silver nanoparticles. The flowchart and experimental setup for the synthesis of colloidal silver nanoparticles of the present work are shown in Figure 1 and Figure 2.

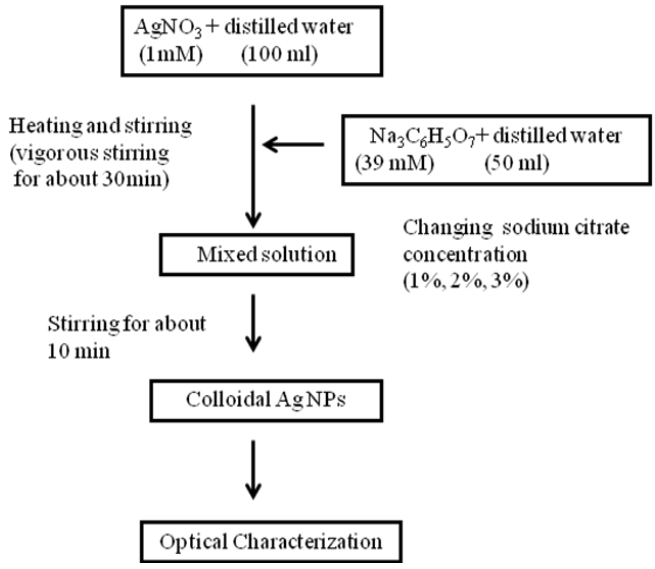


Figure 1. Preparation steps for colloidal silver nanorods

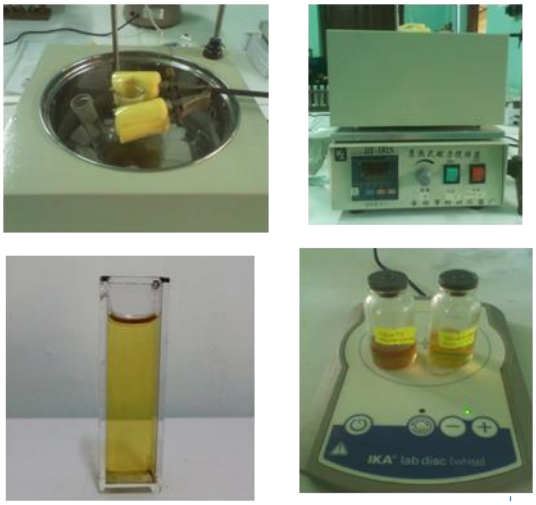


Figure 2. Experimental setup for the synthesis of colloidal Ag NRs

2.2 Characterization Methods

The characterization methods include UV-vis spectrophotometer and scanning electron microscopy (SEM).

2.2.1 UV-vis Spectrophotometry

UV-vis spectrophotometry is a technique to measure absorbance of the sample in the visible region. The UV-vis absorption spectra were measured at room temperature using the GENESYS 10S UV-vis spectrophotometry (Materials Research Lab, Department of Physics, University of Mandalay) Figure 3. The scanning ranges of Ag NRs were recorded from 300 nm to 1000 nm. A glass substrate was used as baseline scanning before the

measurement. Baseline measurement was made with distilled water as reference solution. The solution was diluted 10 times with distilled water (DI) before testing. The purpose of this measurement is to investigate the effect of temperature on the optical absorption of colloidal Ag NPs.

2.2.2 Scanning Electron Microscopy (SEM)

Scanning Electron Microscopy (SEM) is a type of electron microscope, which is generally used to directly observe the morphology and size of nanostructures. The SEM images of silver nanorods were measured using the Agilent 8500 SEM as shown in Figure 4. For investigations by SEM was operated at an operating voltage of 15.0kV and a current of 2.2 A. To perform the SEM measurements, we used dip-coating method. Firstly, a thin film was washed in DI water to remove the excess surfactant and dispersed in PIA solution. After that, this thin film was dipping into the Ag NRs solution and dried in air. The morphological features of metal NRs assembled on Si substrate by SEM. The shape of nanostructures could be observed clearly under SEM.

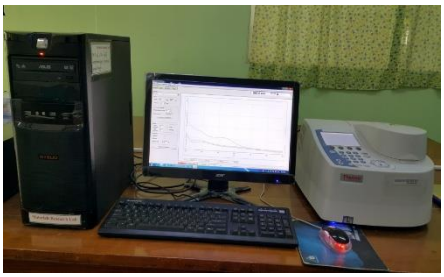


Figure 3. UV-vis Spectrophotometry

Figure 4. Scanning Electron Microscopy (Agilent 8500 SEM)

3. RESULTS AND DISCUSSION

In this chapter, we mainly discuss the effect of reducing agent concentration on optical absorption, size and stability of silver nanoparticles (AgNPs) of the present work.

3.1. Effect of Reducing Agent Concentration

We have studied the effect of reducing agent ($Na_3C_6H_5O_7$) concentrations on optical absorption properties of AgNP colloids. Figure 5 shows the absorption spectra of colloidal AgNPs with different reducing agent concentrations (1%, 2% and 3%). Upon increasing reducing agent concentration, the color and surface plasmon resonance peak maximum (λ_{max}) remained unchanged. The absorption spectra of silver nanoparticles (AgNPs) solution peaks at λ_{max}= 420 nm is attributing to the s-p transition of electrons. The absorption peak intensity decreases with increasing reducing agent concentration. It means the formation of larger nanoparticles with increased concentrations of $Na_3C_6H_5O_7$. According to above results, the absorption wavelength of reducing agent (sodium citrate) concentration (1%) is better than the other concentrations (2% and 3%).

To support our speculation that the incorporation of reducing agent can reduce the nanoparticles size, the scanning electron microscopy (SEM) was employed to estimate the size of Ag NPs on silicon substrates. The colloidal Ag NPs were transferred onto the silicon substrates and their sizes were estimated using scanning electron microscopy (SEM).

Figure 6 (a-c) shows SEM images of Ag NPs with reducing agent concentrations (1%, 2% and 3%). The average NPs size is 50 nm for reducing agent concentration 1%, 80 nm for reducing agent concentration 2% and 90 nm reducing agent concentration 3%. It was found that Ag NPs size increased with increasing sodium citrate concentrations which agreed well with our previous speculation (increased NP size attributed to decrease absorption peak intensity).

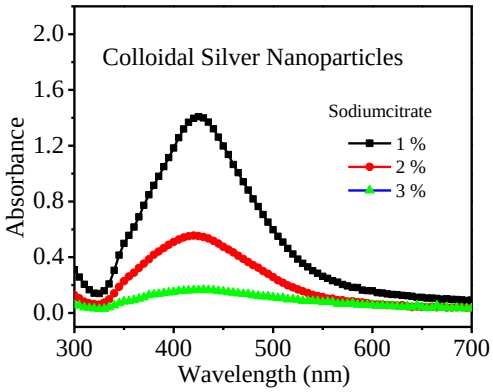


Figure 5. Absorption spectra of colloidal AgNPs with different reducing agent concentrations ($Na_3C_6H_5O_7$)

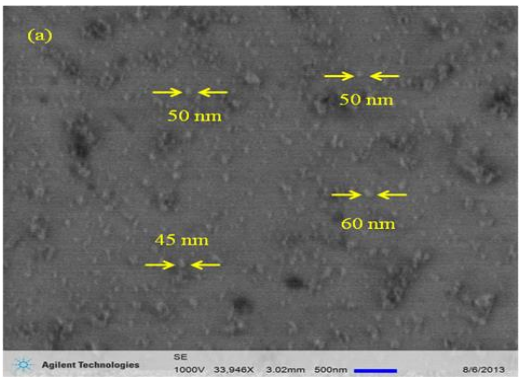


Figure 6. (a) SEM images of Ag NPs with reducing agent concentrations (1%)

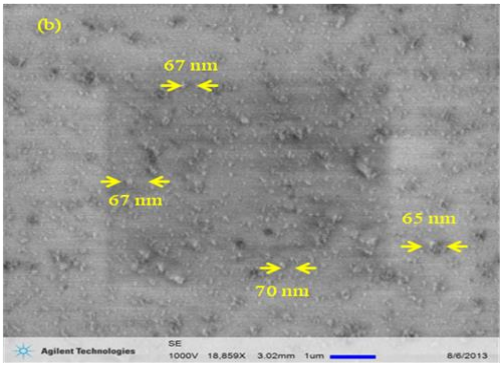


Figure 6. (b) SEM images of Ag NPs with reducing agent concentrations (2%)

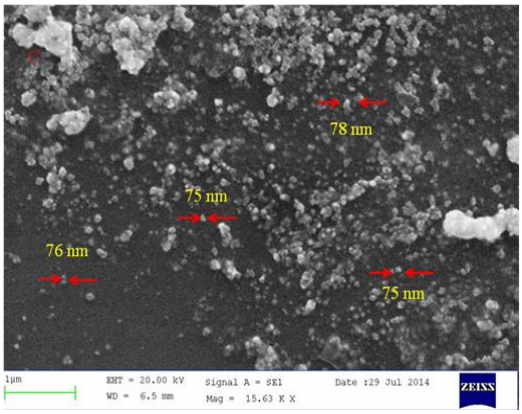


Figure 6. (c) SEM images of Ag NPs with reducing agent concentrations (3%)

3.2. Effect of Reducing Agent Concentration (AA)

We have studied the effect of reducing agent ascorbic acid (AA) concentration (varying AA to $AgNO_3$ molar ratio) on extinction properties of Ag NPs colloids. Figure 7 shows the extinction spectra of colloidal Ag NPs with different molar ratios of AA to $AgNO_3$ (1%,2%,3%). Upon the increasing AA concentration, the color remained unchanged. The λ_{max} is 401 nm, 410 nm, and 422 nm for Ag NPs colloids with different molar ratio. The λ_{max} is red-shifted when reducing agent concentration is increased. In addition, the extinction peak intensity decreases with increasing reducing agent concentration. It suggests the formation of larger NPs with increased concentrations of AA.

The colloidal Ag NPs were transferred onto the silicon substrates and their sizes were estimated using scanning electron microscopy (SEM). Figure 8 (a-c) shows SEM images of colloidal Ag NPs. The average NPs size is 56 nm for the molar ratio of 1%, 74 nm for the molar ratio of 2% and 88 nm for molar ratio 3%. It was found that Ag NPs size increased

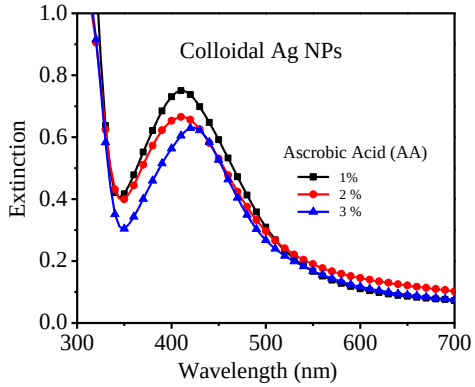


Figure 7. Absorption spectra of colloidal AgNPs with different reducing agent (AA) concentrations

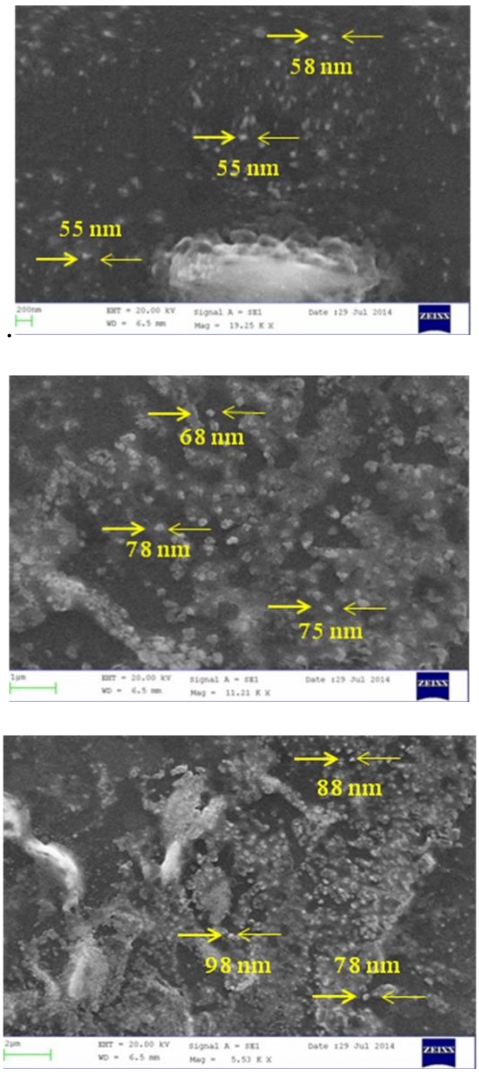


Figure 8. (a-c) SEM images of Ag NPs with reducing agent concentrations (1%,2% and 3%)

3.3. Stability of Colloidal Silver Nanoparticles

In order to examine the stability of AgNP colloids, we investigated the optical properties of colloidal AgNPs upon aging for 5-20 days. Figure 9 shows the absorption spectra of as-prepared and aged AgNPs colloid (reducing agent concentration 1%). Upon ageing for 5 days up to 15 days, the color and peak position of AgNPs remained unchanged. The optical absorption peak intensity increased. Increased peak intensity can be attributed to the higher concentration of surface electrons in aged AgNP colloids. After ageing up to 20 days, the optical absorption peak intensity decreased. This ageing study indicated that AgNP colloids were stable only up to 15 days for reducing agent concentration 1%.

The stability of AgNPs was studied upon ageing the colloids for 5 days up to 15 days. Figure 10 shows absorption spectra of as-prepared and aged AgNPs colloid (reducing agent concentration 2%). After ageing 5 days up to 10 days, the optical absorption peak position remained unchanged and peak intensity increased but decreased up to 15 days. There was no color changed. We obtained maximum absorption peak intensity after 10 days. This ageing studied, AgNP colloids were stable within 10 days for reducing agent concentration (2%). The stability of AgNPs was examined upon ageing the colloids for 5-15 days.

Figure 11 shows absorption spectra of as-prepared and aged AgNP colloids (reducing agent concentration 3%). Upon ageing 5 days up to 10 days, the absorption peak intensity increased suggesting that the higher surface charge density of colloidal AgNPs. After ageing up to 15 days, peak position remained unchanged and peak intensity decreased. It means the formation of larger nanoparticles with aged of AgNP colloids. This ageing study indicated that AgNP colloids with lower reducing agent concentration (1%) yield a better stability as compared to the stability of other colloids with higher reducing agent concentrations (2% and 3%).

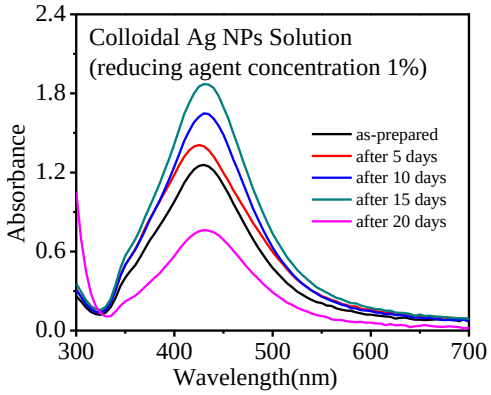


Figure 9 Absorption spectra of as-prepared and aged AgNPs colloid (reducing agent concentration 1%)

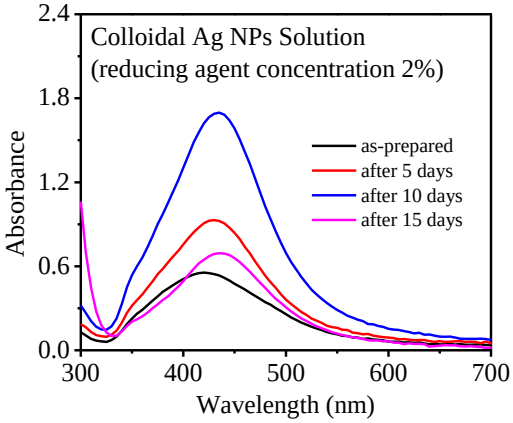


Figure 10 Absorption spectra of as-prepared and aged AgNPs colloid (reducing agent concentration 2%)

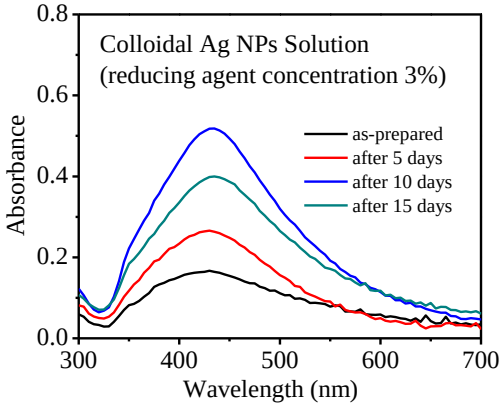


Figure 11 Absorption spectra of as-prepared and aged Ag NPs colloid (reducing agent concentration 3%)

4. CONCLUSIONS

AgNPs are mainly used for antimicrobial and anticancer therapy, and also applied in the promotion of wound repair and bone healing, or as the vaccine adjuvant, anti-diabetic agent and biosensors. Silver nanoparticles (AgNPs) were synthesized by chemical reduction method. Three different samples were prepared using different reducing agent (ascorbic acids and sodium citrate) concentrations (1%, 2% and 3%). In this present work, the effect of reducing agent concentration on size and optical absorption of colloidal AgNPs was mainly examined using UV-vis spectrophotometry and scanning electron microscopy (SEM). It can be concluded that in the case of colloidal AgNPs, upon increasing reducing agent concentration, the absorption peak maximum (λ_{max}) remained unchanged and absorption peak intensity decreased suggesting the formation of larger NP size which was confirmed by SEM micrographs. Upon increasing reducing agent concentration, the Ag NPs size increased from ~ 50 nm to ~ 80 nm for sodium citrate and ~ 56 nm up to 88 nm for AA). The stability study indicated that the NPs could be stable up to 15 days.

ACKNOWLEDGEMENTS

Special thanks go to members of Materials Research Lab, Department of Physics, University of Mandalay, for their help in experimental work.

REFERENCES

[1] J. Liu *et al*., Colloids Surf. A, **302** (2007) 276.

[2] D.R. Duran N *et al*., Silver Nanopar, **13** (2012) 12.

[3] M. K. Fahadi *et al*., J. Eng. Env. Sic, **34** (2011) 281.

[4] F.M. Al-Khateeb *et al*., Adv. Appl Phys and Mater Sci, **134** (2018) 217.

[5] G.M. Azizan *et al*., Int. J. Auto and Mechan Engi, **10** (2014) 1920.

[6] T. Markvart *et al*., J. Appl. Phys. **29**, (2003) 300.

[7] D. Muthuvinothini A *et al*., J. Nanotechnol, **2** (2015) 3.

[8] T.X. Chen Li *et al*., J. Open Surf Sci, **3** (2011) 76.

[9] A.C. Chan *et al*., J. Phys. Chem. B **109 (40)** (2005) 40.

[10] K. Sung *et al*., Langmuir, **14** (1998) 602.

1. First Author:

Daw Myint Myint Kyi

Lecturer

Department of Physics

Technological University (Magway).

2. Second Author:

Daw Thida Nwe

Lecturer

Department of Physics

Technological University (Magway).

3. Third Author:

Daw Thiri Win

Lecturer

Department of Physics

Technological University (Magway).

EXTRACTION OF RULES FROM TRAINED NEURAL NETWORK FOR TEACHING EVALUATION

Soe Moe Aye[1], Aye Mya Sandar[2], and Thuzar Aung[3]

[1,2,3] Information Technology Supporting and Maintenance Department, University of Computer Studies (Mandalay)
[1]soemoeaye2479@gmail.com,[2]ayemyasandarpaing@gmail.com

ABSTRACT

Data mining is utilized education for analysis, forecasting and judging the enforcement of student, teacher and other that have a role in the educational system. Teaching evaluation attempts to methodically analysis teachers' teaching process in accordance with certain teaching principles and criteria, and evaluate its value, positive and negative consequences, so as to make better the standard of teaching. It is not only an essential part of the teaching process, but also the basis of all powerful and wealthy teaching. So the idea of this paper is to give the evaluation results using Neural Networks (NNs) and extract the rules using M-of-N algorithm from trained neural network and calculate the accuracy. NNs attain high accuracy is classification, prediction and many other applications but decision process of NN is hard to be understand by users. Taking into consideration, the architectural features and flexible and self-adaptive actions of NNs, a teaching quality assessment system for teachers was established applying the Backpropagation algorithm. Taking out rules from trained neural network is one of the results for users understand the large amount of connectionist NN structure. By use of feed forward network and backpropagation algorithm, this system can determine for evaluation assessment from students to teachers.

KEYWORDS

M-of-N, Neural Network, Backpropagation, feed forward network

1. INTRODUCTION

Educational quality has taken a lot of regard as the current higher education teaching remodel continues to deepen and expand. The core to elaborating education quality is to advance teaching quality, and teacher evaluation is a principal tool for doing so. As a consequence, educational executive needs the evolution and refinement of a system for examining teaching quality. Conventional approaches to evaluating teaching quality, on the other hand, are unsettled due to their restrictions. As a consequence, a scientific and impartial model for evaluating the teaching quality of college teachers must be established. Teacher assessment is aimed to secure teacher quality and to uphold professional learning with the target of developing forthcoming performance [1].

In this area, we prefer to initiate appropriate apply of data mining in education, called Educational Data Mining (EDM). EDM is a multidisciplinary research area generated as the application of data mining in the educational field. EDM is the proceeding of raw data transformation from large educational databases to effective and purposeful information which can be utilized for a special understanding of students and their learning situations, enhancing teaching reinforcement in addition for decision making in educational systems.

Recognizing the growing importance of neural networks in the EDM community, this system purposes to assess the teaching evaluation with deep learning networks that have long been contemplated by a critical subject on which teaching achievements should be intense, to hold up a speedy and powerful training. Neural networks can enhance education through personalized learning,

recommendation systems, and automated assessment.

A feedforward Artificial Neural Network (ANN) is a type of neural network in which the information flows in only one direction, from the input layer, through one or more hidden layers, to the output layer, left out any feedback loops. It is called "feedforward" now that the data only moves forward through the network and never loops back on itself. Backpropagation Artificial Neural Network (ANN) is a supervised learning algorithm used to train a feedforward neural network. It is the exalted popular and long-established algorithm for training ANNs [7]. As any neural network needs to be trained for the achievement of the task, Backpropagation is an algorithm that is applied for the training of the NN.

Backpropagation Neural network is the most famous machine learning method. As an intention technology, it can be used as an assuming of online teaching assessment. The teaching assessment process mostly requires powerful decision-making and exploration. One of the alternatives to bring out unbiased and non-prejudicial information from student data for driving smart education is the treatment of ANNs. Hence, ANN is good enough for developmental smart education, intelligent tutoring, and charity academic advising.

The planned system can able to forecast the performance of teachers' applying predication model and also classify factors that influence the performance of teachers'. In this study use Backpropagation and an effective algorithm for extracting M-of-N rules from trained neural networks.

A comprehensive explanation of the backpropagation procedure can be found in Segment 2. In Segment 3, Rule Extraction and M-of-N algorithm explained. Segment 4 reports Accuracy. At the closing segment the paper is completed.

2. BIOLOGICAL BASIS OF ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANN) are a technique related to analyses of the brain and nervous system as illustrated in Figure 1. These networks take as a biological neural network but they put to utilize a reduced set of conceptions from biological neural systems. Unambiguously, ANN replicas make believe the electrical movement of the brain and nervous system. Processing elements (a neurode or perceptron) are banded together to another processing elements. In general, the neurodes are prepared in a layer or vector, with the output of one layer serving as the input to the next layer and perhaps other layers.

A neurode may be linked to all or a subset of the neurodes in the successive layer, with these relations put on the synaptic joints of the brain. Weighted data signals getting in a neurode assume the electrical excitation of a nerve cell and subsequently the transference of information through the network or brain. The input values to a treating element, in, are multiplied by a connecting weight, wn,m, that counterfeit the strengthening of neural byways in the brain. It is through the adaptation of the relation strengths or weights that learning is reflected in Artificial Neural Networks. Period the learning aspect, the network find out by harmonizing the weights so as to be talented to evaluate the accurate class label of the input sampling [2].

Each and every of the weight-adjusted input values to a treating element are moreover accumulated applying a vector to scalar function such as summation (i.e., $y = \Sigma wijxi$), averaging, input maximum, or mode value to generate a single input value to the neurode. Formerly the input value is assessed, the processing element then utilizes a transfer function to bring its output.

The transfer function transfigures the neurode's input value. For the most part, this transmutation contains the employ of a sigmoid, hyperbolic-tangent, or other nonlinear function. The procedure is reiterated between layers of treating elements as far as a final output value, on, or vector of values is generated by the NN.

In theory, to counterfeit the asynchronous action of the human nervous system, the processing elements of the ANN should also be stimulated with the weighted input signal in a nonsynchronous manner. Almost software

and hardware accomplishments of ANNs, anyhow, carry though a more discretized approach that warrants that each processing element is stimulated once for each demonstration of a vector of input values. [6]

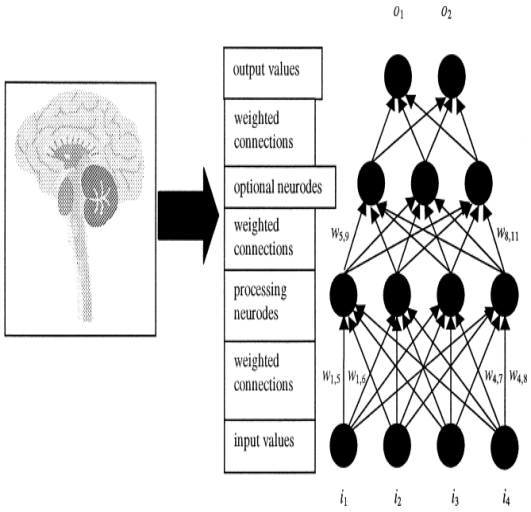


Figure 1. Sample Artificial Neural Network Architecture

2.1. Feedforward Neural Networks

The feed forward model is a fundamental variety of artificial neural network whereas the relations between units do not structure a cycle. Feedforward neural networks were the original kind of ANN imagined and are easier than their counterpart, recurrent NNs.

They are named feedforward for the reason that information only proceeds forward in the network (no loops), initial through the input nodes, then throughout the hidden nodes (if present), and at last across the output nodes. Feedforward neural networks are essentially applied for supervised learning in states where the data to be accomplished is not sequential or time-dependent. [8]

It was pointed out in the initiation such feedforward NNs have the property that information (i.e. computation) streams forward across the network, i.e. there are no loops in the computation graph (directed acyclic graph, or DAG).

Other means of saying this is that the layers are linked in this manner that no layer's output

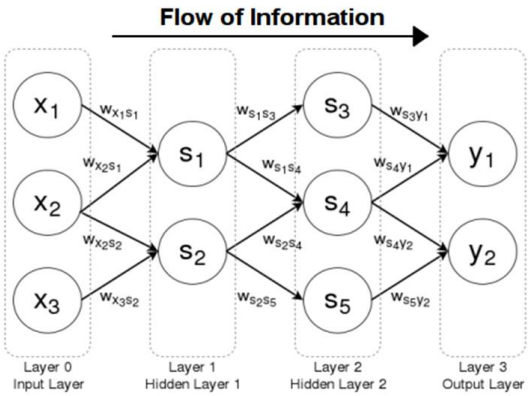


Figure 2. A Feedforward Neural Network with Information Flowing Left to Right

relies on itself. In the previous figure, this is fair since each layer (other than the input layer) is only related to the layer straightforwardly to its left. Feedforward NNs, by having DAGs for their enumeration graphs, have a highly uncomplicated learning algorithm matched to recurrent NNs, which have cycles in their dependency graphs.

In a general sense, if one is stated a graph illustrating a feedforward network, it can ever be combined into layers in such a way that each layer insists only on layers to its left.

This can be executed by achieving a topological type on the nodes and arrangement nodes with the same depth into the similar layer, where layer li contains each node in the graph with depth i in the topological sort. At that times, systematizing layers 0,…,l0,…,lk from left to right, each layer's nodes will only rely on nodes from layers to its left.

2.2. Backpropagation

Backpropagation is one of the major considerations of a neural network. The essential points of view for Backpropagation are the iterative, recursive and advanced method through which it enumerates the updated weight to enhance the network until it is not able to perform the function for which it is being trained.

Derivatives of the activation function to be known at network design time is expected to Backpropagation.

Recollect the following Backpropagation neural network instance diagram to be aware

of: In what manner Backpropagation Algorithm acts:

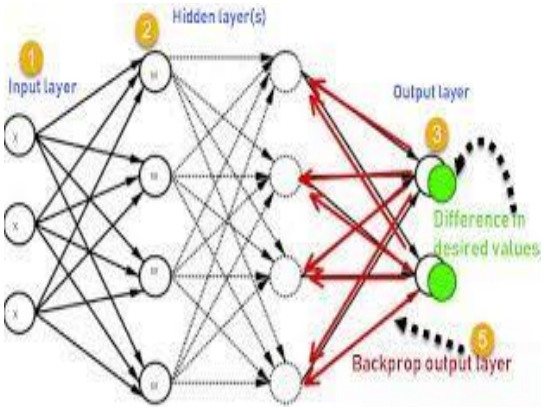


Figure 3. Backpropagation Neural Network Architecture

i. Inputs X, reach throughout the assembled path

ii. Input is formed utilizing actual weights W. The weights are regularly blindly preferred.

iii. Compute the output for each neuron from the input layer, to the hidden layers, to the output layer.

iv. Evaluate the error in the outputs

ErrorB= Actual Output – Desired Output

v. Traverse return from the output layer to the hidden layer to adapt the weights in order, the error is reduced.

Hold reiterating the operation awaiting the wanted output is attained. [9]

2.3. Backpropagation Algorithm

At the start, every edge weights are designated at random. For every inputs in the training data set, the ANN is triggered and its output is recognized. These outputs are relative to what we preliminarily note and be hopeful to output, and the error "propagates" return to the previous level. The error will be noted and the weight will be "adjusted" consequently. This procedure is replicated as far as the output error is following the settled standard.

Set up all weights and biases in network;
While terminating condition is not satisfied {
for each training tuple X in samples {
for each input layer unit j {
O j = I j;
for each hidden or output layer unit j {
I j = Σ i wij O i + Θ j ;
O j = 1/ (1 + e –t j); }
//Backpropagate the errors:
for each unit j in the output layer
Err j = O j (1- O j)(T j –Oj);
Err j = O j (1- O j) Σk Errkwjk ;
for each weight wij in nework {
Δwij = (l)ErrjOi;
wij = wij + Δwij }
for each bias Θ j in network {
Δ Θ j= (l)Errj ;
Θ j= Θ j +Δ Θ j ;}
}}

3. EXTRACTION OF RULES

NNs bring about great accuracy in classification, prediction and many other applications. ANNs have the competency to learn from examples and exhibit some ability for generalization beyond the training data. The knowledge learned by neural networks is make hard to understand by users because it is concealed in a large number of connections. Extracting rules from trained NNs is one of the results for familiarity with the networks.

3.1. M-of-N algorithm

M-of-N algorithm is examines weights of the neural network to discover M-of-N rules. M-of-N rules especially attractive to report the action of a unit when there are several weights with uniform values associated to unit. It goes in search of rules with a Boolean demonstration. The expression is pleased when M of N sets are satisfied. The rule has the following manner:

If {M-of-N antecedents are true} then…

This rule form is a generalization of conjunction of conjunctive and disjunctive rules. For M=1, the rules can be recorded as a number of disjunctions, For M=N, the rules can be explicit as a conjunction of the conditions. There are our stages in the M-of-N algorithm.

Step1. Clustering

Weights are clustered with similarly weighted links. The consequent set of clusters S is specified by

S := {C1, C2, ……..,Cm}

Ck := {wv1, wv2, ……..,wvl}

Wherein Ck means a single cluster with the weights wv1, wv2, ……..,wvl.

Step2. Rule Generation

The rule generation strategy aims to confirm and dispose of the clusters that are not likely to have any outcome on the computation of the resultant. Clusters that cannot transform whether the network input be more than the bias are removed. The rule generation algorithm separates the set of cluster S into disjunctive sets Sp and SN by:

Sp := { Ck / Ck€ S ^ Wck 0}

SN := { Ck / Ck€ S ^ Wck 0}

Wherein Sp the set of all positively weighted clusters, SN is the set of all negatively weighted clusters and Wck is the summed weight of a cluster Ck.

At the last moment ∞ be the incoming activation a set of clusters can dispatch. The minimal and maximal incoming activation is then stated by:

∞min := min { wj, 0}

∞max := max { wj, 0}

When w1, w2, ….,wn are incoming activation link weights of a neuron. With each hidden and output neurons.

i. If ∞min < θ, excluded each cluster and block.

ii. Elicitate sets of positively weight cluster P ϵ (Sp), with the constraint wp > θ, wherein wp us the summed weight of P or 0 if P = θ.

iii. With each set of clusters P raised:

   a) Figure out an incoming activation ∞ = ∞min + wp.

   b) If an incoming activation is larger than P and go on with the later clusters.

   c) Elicitate minimal sets of negatively weighted clusters N, with the necessity wN < ∞ - θ, wherein wN is the summed weight of N.

   d) With all set of clusters N construct, construct a rule for P and N.

Step3. Rule Pruning

Rule Pruning is required, since the formed rules are under definite conditions redundant. This implies that too precise rules are originated where formerly more common rules reside.

Step4. Rule Reformulation

In the last stage, rules are clarified every time possible to remove weights and thresholds. M-of-N simplified a rule by examining it to decide the workable blends of antecedents that exceed the rules threshold. This examine may outcome in excess of one rule. If the removal of weight and biases demands revising a single rule with upwards of five rules, then the rule is left in its initial states.

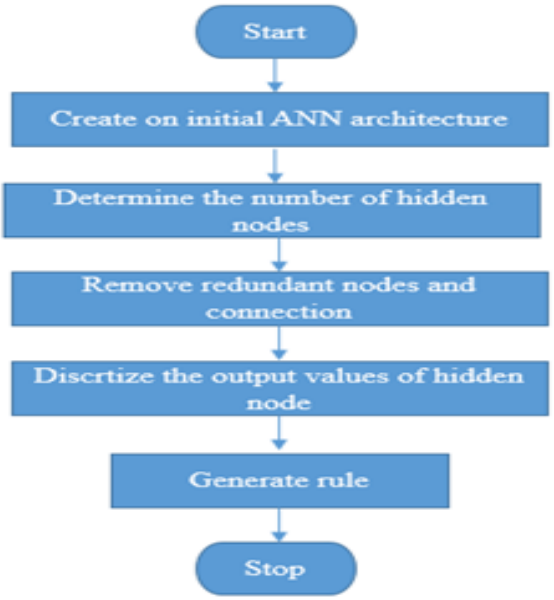


Figure 4. Flow Chart of the Rule Extraction Algorithm

4. ACCURACY MEASUREMENT

Classification and prediction can be matched and judge conforming to the succeeding standard [3]:

Accuracy: The accuracy of a classifier appertain to capability of a considering classifier to rightly forecast the class label of new data.

Speed: This contains the computational cost.

Robustness: This is the capability of classifier or forecaster to construct precise prediction stated noisy or data with missing values.

Scalability: That specify to capability to design the classifier or predictor given great deal of data.

The correctness and accuracy of the machine learning model can be find out using the Mean Squared error (MSE). The MSE of a process for guesstimating an recognized aggregate scales the average of the squares of errors to audit the accrracy that is, the dissimilarity between the actual value and the estimated value of the model. The MSE keep track of the quantity of an estimator – it is consistently non-negative, and values near to zero are better. The mean squared erroris defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Pi - Ai)2$$

where Pi = Predicted output for data point i;

Ai =Definite output for the data point i;

n = Total data points.

Calculation of the mean of the twelve means from "samples of 100"		
Estimated Values	Find the error	Square the errors
98	97-100 = -3	$(-3)^2 = 9$
99	99-100 = -1	$(-1)^2 = 1$
98	98-100 = -2	$(-2)^2 = 4$
106	104-100 = 4	$(4)^2 = 16$
97	97-100 = -3	$(-3)^2 = 9$
95	94-100 = -4	$(-4)^2 = 16$
99	99-100 = -1	$(-1)^2 = 1$
101	101-100 = 1	$(1)^2 = 1$
97	97-100 = -3	$(-3)^2 = + 9$
96	95-100 = -5	$(-5)^2 = 25$
99	94-100 = -6	$(-6)^2 = 36$
99	99-100 = -1	$(-1)^2 = 1$

Figure 5. Illustrates the Sample Calculation the Mean for These Twelve "mean of 100" Samples:

Add all of the squared errors up: 9 +1 +4 +16 +9+16+1+1+9+25+36+1 = 128.
Find the mean squared error: 128 / 12 = 10.66.

The MSE value increases as the value of the number of record increase (Number of record 100 - MSE 10.66 , Number of record 200 - MSE 12.11 and Number of record 300 - MSE 13.69). So, it is obvious from the above implementation that as the number of record count increase then the error rate will be high.

This system represents an adequate system model for assessment and forecasting of teachers' performance in higher institutions and trainings of learning applying data mining machinery. To carry out successfully the purposes of this effort, it contains a two-layer Artificial Neural Network (ANN) with adjustable hidden-to-output weights (backpropagation of errors).ANN method provides a robust and approximation the target function for many classification, prediction and clustering problem. NNs achieve high accuracy in predictions for this system.

Evaluating the teaching quality of teachers enables the educational management department to gain a systematic, comprehensive, and accurate understanding of the teaching standards and quality across various courses. This evaluation process also helps identify effective teaching practices and areas that require improvement, facilitating timely teaching reforms and providing a solid foundation for decision-making [4-5].

Standard learning algorithms trade greater for majority class samples and regularly accomplish poorly on minority (positive) class samples. This unfairness outcomes in performance model, which is generates building a predictive model challenging. In spite of that an ANN leveraging clustering and connecting can obtain balanced data that can accurately judge for teaching evaluation.

The crucial to obtain a great paradigm with exact predictions is to discover "optimal values of W — weights" that downsizes the prediction error. This is attained by "Backpropagation algorithm" and this constructs ANN a learning algorithm because by learning from the errors, the model is upgraded.

5. EXPERIMENTAL DESIGN AND RESULT

The system is proposed to give that the evaluation results may contain data relating to individual teachers after learning and teaching the end of each courses. The proposed system can provide the Institute , Department,

Training Course and other teaching centres must be shared the course evaluation results with students by publishing course evaluation summaries.

Student evaluation is one of the established components of the Institutes' quality assurance system. Evaluation results must be used to improve teaching performance and also students must be assured of the anonymity of results. Viewing students' assessment results as data rather than evaluations can also help to write about teaching evaluations that are almost universal in recommending the use of multiple sources of data. In this system, Admin, canter director for training (Rector for Institute), students and teachers can login, only when he/she enters a correct username and password, the system will perform what he/she want to do.
A web-based evaluation system intends to use for evaluation assessment from students to teachers. System allows users to provide evaluation, which is used to improve training environment Feedback questions and answers are stored in database according to their related question categories so that system has the ability to reuse questions and ability to add new instructor-supplied questions. All students evaluation data are encrypted and stored in database as so that no one can know except the one (centre director) who has the decryption key. Centre Director can view all of grading results for each subject. He/she can also check their suggestion reports that are sent by students.

Administrator can develop evaluation forms by quickly browse through the question banks and by adding new questions. Admin can manage the date of the evaluation form from response time to close time. The assessment questions and quantity may differ contingent on the nature of the course/subject. The question charging process is done by admin taking question bank of some of the classes are categorized by subjects that taught by professional teacher. Admin have the authorization to delete and edit the records of question according to the changed course as shown in Figure 6.

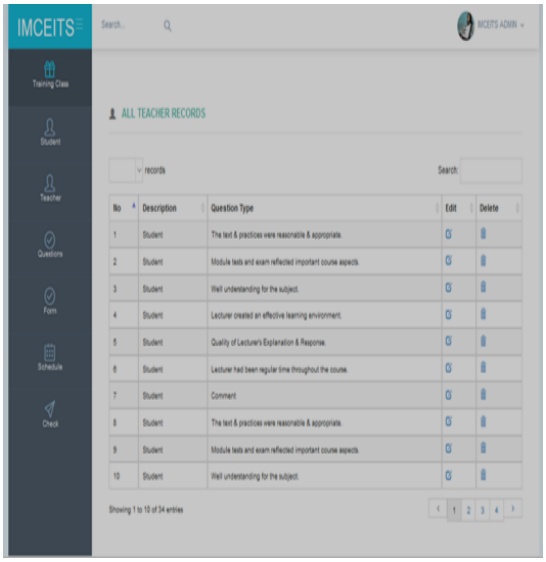


Figure 6. Questions Records

Admin can choose the right questions in dataset of the Center for Information and Communication Technology Training (CICTT), Yangon depend on the nature of subject whether it related to practical or theory as in Figure 7.

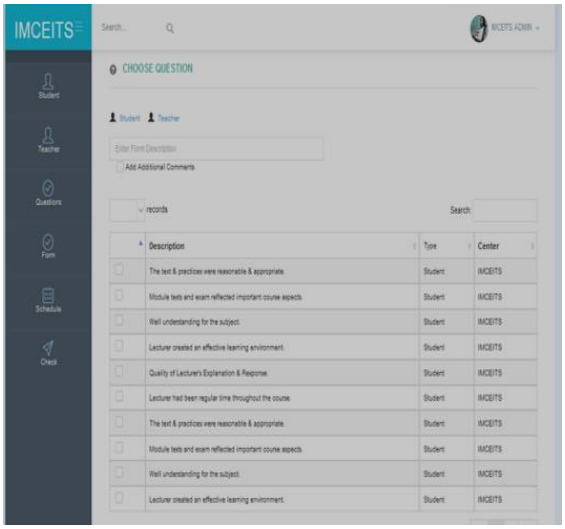


Figure 7. Choose Questions

Figure 8 shows the evaluation form for student that give feedback to teachers from teaching centre by course schedule. First, he/she must choose the name of teachers that he/she wants to give feedback and then complete form.

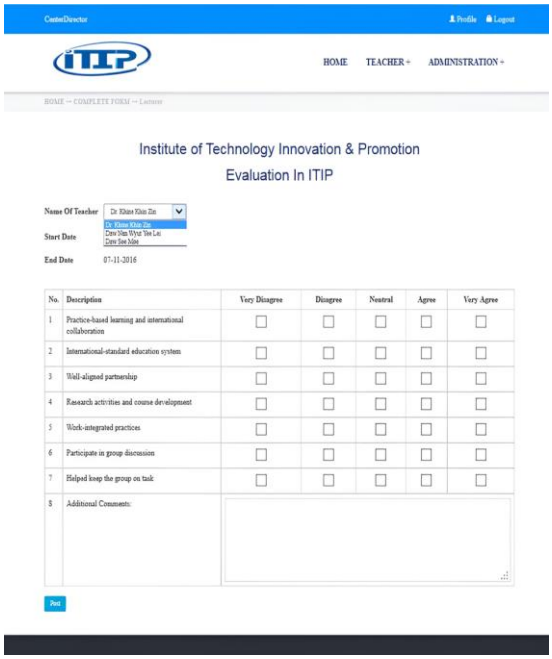


Figure 8. Evaluation Form

Figure 9 shows the personal result by each teacher' evaluation form. The teacher can view the average percentage of his/her own assessment result according to course and their course' schedule.

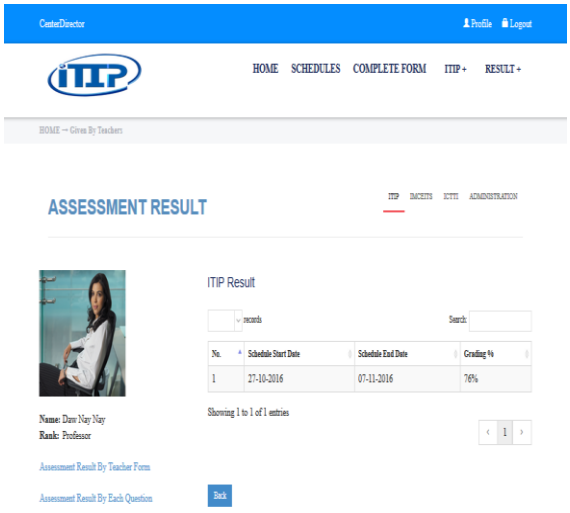


Figure 9. Personal Result for Each Schedule

Figure 10 shows the assessment result by each question. The teacher can select form (according to the courses) and view his/her evaluation result by each question.

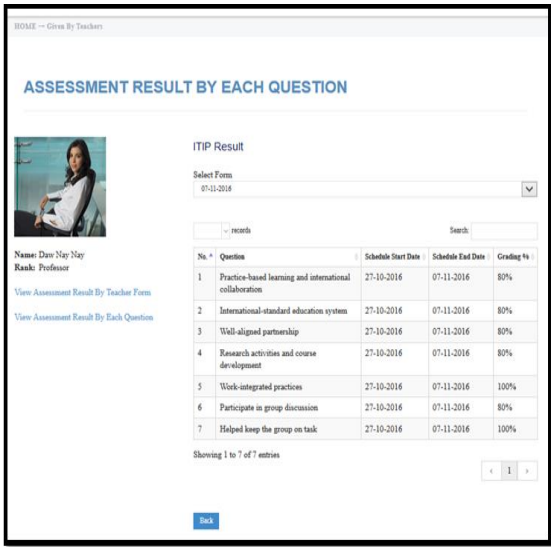


Figure 10. Personal Result for Each Question

In this study, prediction of Teachers' performance assessment of the subject is presented. The format of a questionnaire given to students in the form of questions has maximun 20 questions in which each question represents the study variables. The study variables are Presentation of Content, Theory/Content Knowledge, Course Content, Encouraging of Participation/Discussion and so on. Students need to tick the desired category 1 (Very Disagree), 2 (Disagree), 3(Normal), 4 (Agree) and 5(Very Agree).

In this system, the technique uses 20 input neurons in input layer of neural network, 10 neurons for 1 hidden layer and 1 neuron for result to forecast the teacher performance of subjects. The nearly weight values 0.2, 0.4, 0.6, 0.8 and 1 are represented by category, respectively of every learning variable to determine the prediction of teachers' performance of the subject. To determine the level of teachers' performance of the course is based on professional teaching activities in questionnaire. Output from the last layer of neural network, there are five possibilities based on defined weighted output, namely Grade A (Between 100% and 80%), Grade B (Between 80% and 60%), Grade C (Between 60% and 40%), Grade D (Between 40% and 20%)and Grade E (under 20%). Category for "Grade A" is determined by the minimum error rate of the target that is 1 and nearly 1 and the others are "Grade B" is target rate nearly 0.8 , "Grade C" is nearly 0.6, "Grade D" is nearly 0.4 and nearly 0.2 is "Grade E" .

The procedure of designing a neural network model is a logical process. This system followed the step designing methodology presented in Figure 11.

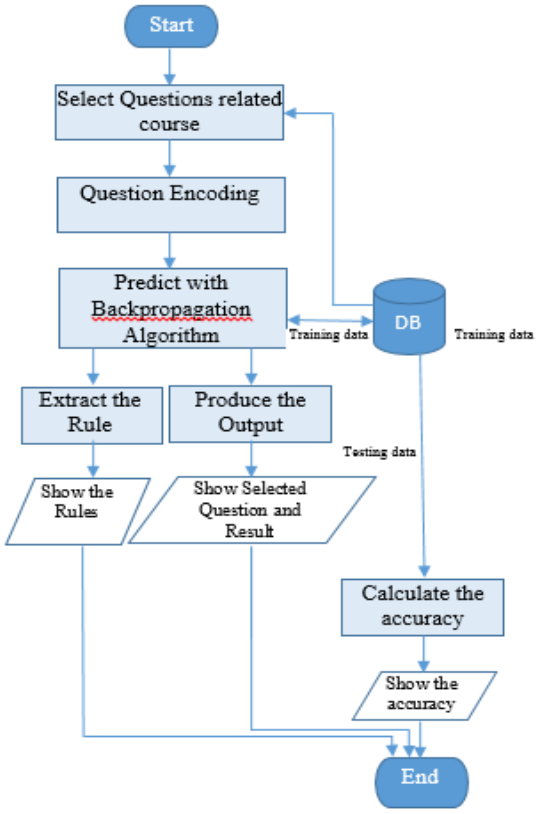


Figure 11. Personal Result for Each Question

6. CONCLUSIONS

By employing the BP algorithm, we can construct a three-layer neural network that accurately identifies the teaching quality and effectiveness of teachers by adjusting the weight system. Using Backpropagation NN for teaching quality assessment can enhance objectivity and accuracy. Compared to traditional subjective evaluation methods, BP neural network can obtain a comprehensive assessment of a teacher's abilities through extensive data training. This objective evaluation method reduces subjective bias and personal preferences, resulting in more impartial and reliable assessment outcomes.

The results of teaching quality evaluation using BP neural network can provide targeted feedback and improvement suggestions for teachers. By analysing the output of the NN, we can understand a teacher's performance in different aspects and offer corresponding advice and training. This helps teachers continuously improve their teaching methods and skills, enhancing teaching quality and promoting students' learning outcomes and satisfaction. In conclusion, Backpropagation NN has vast potential in evaluating the teaching quality of training centres. We presume more uncomplicated set of rules to be produced if more connections are retired. At long last, we reform each node dividing conditions in the Backpropagation in the context of the network inputs that are not detached. A node breaking situation can be reformed as M-of-N rules.

7. ACKNOWLEDGEMENTS

We would principally want to thank Prof. Dr. San San Tint, Pro Rector, Editor in Chief of JITRI committee and each of Technology Program Committee has offered our comprehensive personal and professional advice and taught our a good deal about both scientific paper and life for the most part. We would also want to show our gratitude to the teachers from our department for exchange their pearls of wisdom with us, and we thank 3 “anonymous” reviewers for their so-called insights.

REFERENCES

[1] Danielson, C. (2010). Evaluations that help teachers learn. Educational Leadership, 68(4), 35–39.

[2] Fayyad, U., “Mining Databases: Towards Algorithms for Knowledge Discovery”, IEEE Bulletin of the Technical Committee on Data Engineering, Vol.21 No1, pp.41-48, 1998.

[3] Jiawei Han, Jian Pei, Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, University of Illinois at Urbana-Champaign

[4] Ling, W. (2022) Research on Classroom Teaching Quality Evaluation Method Based on Machine Vision Analysis. Frontiers in Educational Research.5 (3):79-84. doi:10.25236/ FER. 2022. 050314.

[5] Shi, Y. (2022) Application of Artificial Neural Network in College-Level Music Teaching Quality Evaluation. Wireless Communications and Mobile Computing, 2022: 7370015. doi:10.1155/2022/7370015.

[6] Steven Walczak, Narciso Cerpa, in Encyclopedia of Physical Science and Technology (Third Edition), 2003.

[7] https://www.linkedin.com/pulse/feedforward-vs-backpropagation-ann-saffronedge1

[8]https://brilliant.org/wiki/feedforward-neural-networks.

[9]https://www.guru99.com/backpropogation-neural-network.html

Authors

Biography

First A. Author Soe Moe Aye received M.C.Sc degree in University of Computer Studies (Yangon) in 2010. Now she is a lecturer of Information Technology Supporting and Maintenance, University of Computer Studies (Mandalay).

Second B. Author Aye Mya Sandar received M.C.Sc degree in University of Computer Studies (Mandalay) in 2005. Now she is a Associate Professor of Information Technology Supporting and Maintenance, University of Computer Studies (Mandalay).

Third C. Author Thu Zar Aung received M.C.SC degree from University of Computer Studies (Meikhtila) in 2010.She works as a lecturer in Information technology supporting and maintenance department at the University of Computer Studies(Mandalay).

.

AUTOMATIC RECOGNITION OF RICE PLANT DISEASE USING RESIDUAL FEATURES AND SUPPORT VECTOR MACHINE

Aye Mya Sandar[1], Soe Moe Aye[2], and Thuzar Aung[3]

[1,2,3] Information Technology Supporting and Maintenance Department, University of Computer Studies (Mandalay)
[1]ayemyasandarpaing@gmail.com, [2]soemoeaye2479@gmail.com,
[3]tza.thuzaraung@gmail.com

ABSTRACT

Automatic detection and recognition of different diseases from the rice leaf images is important part for agricultural research area. The paper studies and compared handcrafted features, statistical features and deep features extraction methods by using Support Vector Machine (SVM) classifier. Deep learning learns automatically all high-level features in one pass rather than having to engineer them ourselves. Transfer learning approach is used to extract the features by applying residual network (ResNet). The performance of the system has been analyzed on rice leaf disease dataset with total of 120 images collected from UCI data repository. Classification result has been evaluated for each disease by using 10-folds cross validation method. The overall accuracy of disease recognition achieved 90.83% by using features of ResNet-18 Convolutional Neural Network (CNN) and SVM classifier.

.

KEYWORDS

Disease detection, Rice leaf, Deep features, Transfer learning, ResNet, SVM.

1. INTRODUCTION

Agriculture is one of the important sectors and the backbone of economy in Myanmar. Rice is essential food for citizens in our country. Farmers in rural areas face the problem of loss many yields because they have poor knowledge how to prevent or control the pests and disease on the plants timely. Farmers need to detect diseases parts of the field in early stage.

Farmers detect and recognize the symptoms of diseases manually. It is a time and effort costly than automatic recognition. Computerized system should be created to detect and recognize diseases. There are many rice plants' diseases. The paper has classified three types of plant diseases such as Leaf smut, Brown spot and Bacterial leaf blight which are acquired from UCI machine learning repository.

The basic steps of rice disease recognition system are image acquisition, preprocessing, feature extraction and classification. Previous studies [2, 3, 4, 5, 6, 8, 12] have performed preprocessing and feature extraction based on image processing techniques such as image enhancement, region segmentation, colour processing, filtering and so on. For classification of disease, machine learning techniques are used such as Naïve Bayes (NB), Decision Tree (DT), Neural Network (NN), SVM and others. Ensemble classifiers can also be used to classify diseases using fusion of number of classifiers. Deep NN (DCNN) is an end-to-end learning technique for extracting features and classifying the types of diseases.

In the rest of the paper, section 2 contains the related works, section 3 describes the proposed rice leaf disease recognition including data acquisition, preprocessing, feature extraction and classification, section

4 discusses the experimentation of the system and the last section concludes the study.

2. RELATED WORK

In paper [1], the authors have surveyed the different image processing and machine learning methods for detection and recognition of rice plants diseases. The paper explored the preprocessing methods, region segmentation, feature extraction methods and machine learning techniques for classification.

A. A. Joshi et al. [2] have studied the recognition system for controlling and monitoring rice diseases. Segmentation of infected part were processed by using YCbCr colour space in the preprocessing step. The segmented images are resized to 200 by 200 pixels and then were extracted colour features from the RGB components and nine shape features. The classification of diseases was implemented by using Minimum Distance Classifier and k-Nearest Neighbour classifier. They only used colour and shape features.

B.S. Ghyar et al. [3] explored the system for detection of rice leaf diseases using texture and colour features. Firstly, the input RGB image is converted to L*a*b colour space and filtered with median filtering. The disease region is segmented with k-mean clustering methods. After segmentation, area features, statistical features from grey level co-occurrence matrix (GLCM) and colour features. The extracted features are selected to improve the performance by using Genetic algorithm. The classifier are SVM and Artificial Neural Network (ANN).

In paper [4], three types of features are used in the study such as colour, shape and texture features. They employed the system not only the background removal but also segmentation methods to detect infected leaf regions. They also segmentation methods to detect infected leaf regions. They also used feature normalization methods such as min-max normalization and SVM classifiers. The paper [5] also employed the region segmentation on HSV colour space with threshold calculated by OTSU's algorithm. Colorimetric model, ellipse fitting are used to extract features and SVM as classifier.

N. Kitpo et al. [6] have designed rice disease detection system using IoT architecture and Drone device. GPS sensor is used to detect early disease portion and define the position in real time. The proposed system have IoT architecture, including sensor devices, gateway, M2M services and application. The infected rice leaves images are collected by using drone patrolling the rice fields. Before the segmentation, resize the image to 300 by 400 pixels and histogram equalization is used to adjust the image contrast. To separate the healthy and disease region, convert to HSV colour space and used a threshold of Hue value for green colour. The system has used colour features, shape features and SVM classifier.

S. Ramesh et al [7] studies a classification system for rice plants whether the leaf is healthy or infection by using mean, standard deviation, GLCM features and artificial neural network classifier. The leaf image is resized to 256 by 256 pixels and k-mean clustering is used to segment the infected regions. T. Islam et al. [8] have presented different colour classed by calculating the RGB percentage comparing with predefined pixel values for each colour. The percentage values are applied as features to recognize diseases by using Gaussian Naïve Bayes classifier. In [9], the authors have compared the two classifier SVM and CNN. For segmentation of disease portion, k-mean clustering is used in this paper.

Md. J. Hasan et al. [10] have used transfer learning for detecting and classifying rice disease using the CNN features extracted from inception-v3 network and SVM classifier. The paper also used data augmentation technique called random flipping. The authors have experimented nine different rice diseases. In paper [11], the images are classified with K-NN classifier, J48 classifier, Naïve Bayes classifier a Logistic Regression. M. E. Pothen et. al. [12] proposed a system to recognize rice diseases by using image processing techniques. Firstly, the leaf

image is segmented with Otsu's method and significant features are extracted by using Local Binary Pattern (LBP) and histogram of oriented gradient (HOG) features. They have compared two features by using SVM classifier with linear, polynomial and radial basis function kernels.

The reference [7] paper has classified the rice leaves as healthy or disease leaves. The research [10] described the system to recognize nine different disease types such as brown spot, false smut, bacterial leaf blight, leaf scald, leaf smut, red stripe, sheath blight, tungro virus and blast. The most commonly used disease types and number of images analyzed in the previous studies are listed in table 1.

Table 1. Disease Types of previous studies

ID	Types of Disease	No. of Images
[2]	Rice bacterial blight, Rice blast, Rice brown spot, Rice sheath rot	115
[3]	Rice blast, Brown spot, Healthy	60
[4]	Bacterial leaf blight, brown spot, leaf smut	120
[5]	Acute type, Chronic type, white type	90
[8]	Rice brown spot, Rice bacterial blight, Rice blast	60
[11]	Leaf smut, Bacterial blight, Brown spot	480
[12]	Bacterial leaf blight, Brown spot, Leaf smut	120

3. PROPOSED SYSTEM

The proposed system has four steps: image acquisition, preprocessing, feature extraction and classification. The architecture of the proposed system has shown in figure 1.

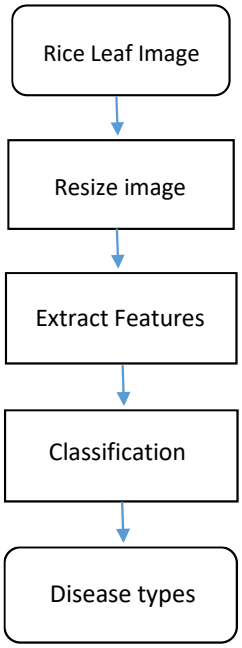


Figure 1. Architecture of the rice disease recognition system

3.1. Data acquisition

The infected rice leaves are collected from UCI data repository [13]. The images are captured manually and the diseases names in dataset have provided by farmers. The number of leaves images are 40 infected leaves in each class. The characteristics of three types of rice leaf diseases are as follows.

Bacterial leaf blight: The leaves are elongated lesions on the tips and margins of leaf and turns white to yellow and then greyish with black dots [4, 6, 11]. Sample image of bacterial leaf blight disease as show in figure 2.

Figure 2. Bacterial Leaf Blight

Brown spot: The infected leaf has dark brown to purple colour and round to oval shaped lesions on leaves [4, 6, 11]. Sample image of brown spot leaf as shown in figure 3.

Figure 3. Brown Spot

Leaf smut: The infected leaf has small spots linear lesions and turns grey and dry on leaf tips [4, 11]. Sample image of leaf smut has shown in figure 4.

Figure 4. Leaf Smut

3.2. Image Preprocessing

Input image is pre-processed to get the better result for the next processes. The proposed system needs to resize the input image to fix the input size of network into 224 by 224 by 3-pixel dimension.

3.3.Feature Extraction

The feature of infected leaves are extracted by using pre-trained CNN called residual network (ResNet-50). Deep learning need not be featuring engineering and learns high dimensional features automatically. Transfer learning learns the meaningful features from the previous trained network.
The residual network contains a number of residual blocks. A residual block contains convolutional layer followed by batch normalization layer and ReLU activation and again convolutional layer and batch normalization layer. The feature extraction process using residual network has shown in figure 5.

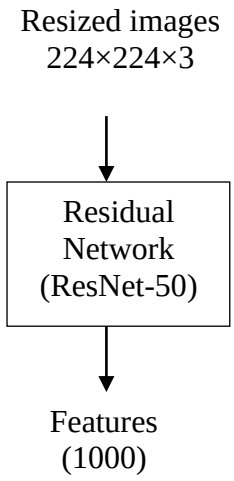


Figure 5. Feature extraction using ResNet-50

ResNet-50 have learned to extract the generic features and transfer these on our labelled data. The number of extracted features are 1000 from the fully connected layer of the network. These features are feed to SVM classifier to classify the disease types. The features extracted from first residual block have visualized as in figure 6, figure 7 and figure 8.

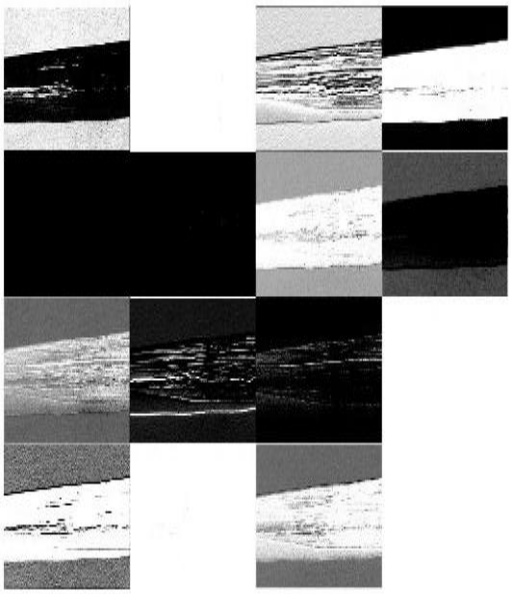


Figure 6. Features visualization for bacterial leaf blight in first residual block

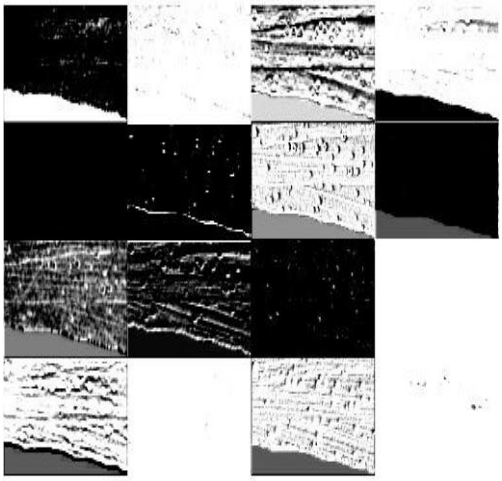


Figure 7. Features visualization for brown spot in first residual block

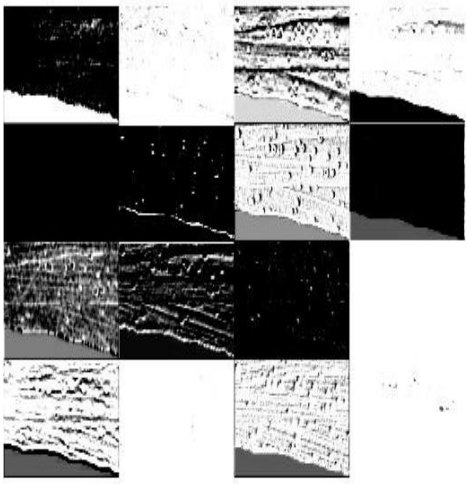


Figure 8. Features visualization for leaf smut in first residual block

3.4. Classification

The rice diseases recognition system used SVM classifier to classify the different rice plant leaf diseases. SVM performs the classification process on multidimensional hyper-plain for different class labels.

4. EXPERIMENTAL RESULTS

The number of correct classified images are described in table 2 for each class, leftmost column is actual class labels and row is predicted classes labels. The evaluation results of classification 97.5% for bacterial leaf blight, 92.5% for brown spot and 83.5% for leaf smut in accuracy by using 10 folds cross validation. Bacterial leaf blight is the most correctly classify rice leaf disease in the system. Leaf smut can misclassify as one of two other classes. To compare the proposed system, we analyzed the other features such as local binary pattern and a number of statistical features.

Table 2. Classification Results

	Bacterial Leaf blight	**Brown spot**	**Leaf smut**
Bacterial Leaf blight	39	0	1
Brown spot	0	37	3
Leaf smut	4	4	33

Local binary pattern (LBP) is the mostly used as feature descriptors in image classification tasks. Other features used in rice disease detection system are colour features such as mean and standard deviation of RGB colour (6 features), features from gray level co-occurrence matrix (4 features) such as contrast, correlation, energy, homogeneity, and other significant features such as mean, standard deviation, entropy, root mean square (RMS), variance, smoothness, kurtosis and skewness and inverse different movement (IDM). Total number of statistics features is nineteen. SVM also used as a classifier to analyze the features. The comparison result has shown in table 3.

Table 3. Classification Results Using Svm Classifier

Features	**Accuracy**
LBP	60%
19 Statistical features	72.5%
ResNet-18 features	89.17%
ResNet-50 features	90.83%

5. CONCLUSIONS

Automatic rice leaf diseases detection and recognition system is important for agriculture sector. Farmers need to detect early the disease portion in the field and to prevent failure of rice plants. The proposed system does not need prior knowledge about disease and no need to design features. Transfer learning is proper used for system on limited data to collect. ResNet-50 network is proper network structure for image classification system. The proposed system used features extracted from ResNet-50 CNN and SVM multiclass classifier. This work will be extended by combining features statistical, handcrafted, and deep features to improve the recognition.

6. ACKNOWLEDGEMENTS

We would especially like to thank Prof. Dr. San San Tint, Pro Rector, Editor in Chief of JITRI committee and each of Technology Program Committee has provided our extensive personal and professional guidance and taught a great deal about both scientific paper and life in general. We would also like to show our gratitude to the teachers from our department for sharing their pearls of wisdom with us during the course of this research, and we thank 3 “anonymous” reviewers for their so-called insights.

REFERENCES

[1] J. P. Shah, H. B. Prajapati, and V. K. Dabhi, “A survey on detection and classification of rice plant diseases,” in 2016 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC), Bangalore, India, Mar. 2016, pp. 1–8, doi: 10.1109/ICCTAC.2016.7567333.

[2] A. A. Joshi and B. D. Jadhav, “Monitoring and controlling rice diseases using Image processing techniques,” in 2016 International Conference on Computing, Analytics and Security Trends (CAST), Pune, India, Dec. 2016, pp. 471–476, doi: 10.1109/CAST.2016.7915015.

[3] B. S. Ghyar and G. K. Birajdar, “Computer vision based approach to detect rice leaf diseases using texture and color descriptors,” in 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, Nov. 2017, pp. 1074–1078, doi: 10.1109/ICICI.2017.8365305.

[4] H. B. Prajapati, J. P. Shah, and V. K. Dabhi, “Detection and classification of rice plant diseases,” Intell. Decis. Technol., vol. 11, no. 3, pp. 357–373, Aug. 2017, doi: 10.3233/IDT-170301.

[5] Jun. Zhang, Lifei. Yan, and Jinyi. Hou, “Recognition of rice leaf diseases based on salient characteristics,” in 2018 13th World Congress on Intelligent Control and Automation (WCICA), Changsha, China, Jul. 2018, pp. 801–806, doi: 10.1109/WCICA.2018.8630414.

[6] N. Kitpo and M. Inoue, “Early Rice Disease Detection and Position Mapping System using Drone and IoT Architecture,” in 2018 12th South East Asian Technical University Consortium (SEATUC), Yogyakarta, Indonesia, Mar. 2018, pp. 1–5, doi: 10.1109/SEATUC.2018.8788863.

[7] S. Ramesh and D. Vydeki, “Rice Blast Disease Detection and Classification Using Machine Learning Algorithm,” in 2018 2nd International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), Ghaziabad, India, Sep. 2018, pp. 255–259, doi: 10.1109/ICMETE.2018.00063.

[8] T. Islam, M. Sah, S. Baral, and R. Roy Choudhury, “A Faster Technique on Rice Disease Detectionusing Image Processing of Affected Area in Agro-Field,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, Apr. 2018, pp. 62–66, doi: 10.1109/ICICCT.2018.8473322.

[9] G. Verma, C. Taluja, and A. K. Saxena, “Vision Based Detection and Classification of Disease on Rice Crops Using Convolutional Neural Network,” in 2019 International Conference on Cutting-edge Technologies in Engineering (ICon-CuTE), Uttar Pradesh, India, Nov. 2019, pp. 1–4,

doi: 10.1109/ICon-CuTE47290.2019.8991476.

[10] Md. J. Hasan, S. Mahbub, Md. S. Alom, and Md. Abu Nasim, "Rice Disease Identification and Classification by Integrating Support Vector Machine With Deep Convolutional Neural Network," in 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, May 2019, pp. 1–6, doi: 10.1109/ICASERT.2019.8934568.

[11] K. Ahmed, T. R. Shahidi, S. Md. Irfanul Alam, and S. Momen, "Rice Leaf Disease Detection Using Machine Learning Techniques," in 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, Dec. 2019, pp. 1–5, doi: 10.1109/STI47673.2019.9068096.

[12] M. E. Pothen and Dr. M. L. Pai, "Detection of Rice Leaf Diseases Using Image Processing," in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, Mar. 2020, pp. 424–430, doi: 10.1109/ICCMC48092.2020.ICCMC-00080.

[13] https://archive.ics.uci.edu/ml/datasets/Rice+Leaf+Diseases

Authors

Biography

First A. Author Aye Mya Sandar received M.C.Sc degree in University of Computer Studies (Mandalay) in 2005. Now she is a Associate Professor of Information Technology Supporting and Maintenance, University of Computer Studies (Mandalay).

Second B. Author Soe Moe Aye received M.C.Sc degree in University of Computer Studies (Yangon) in 2010. Now she is a lecturer of Information Technology Supporting and Maintenance, University of Computer Studies (Mandalay).

Third C. Author Thu Zar Aung received M.C.SC degree from University of Computer Studies (Meikhtila) in 2010.She works as a lecturer in Information technology supporting and maintenance department at the University of Computer Studies(Mandalay).

DETERMINATION OF RADIONUCLIDES CONCENTRATION IN SOIL SAMPLES FROM HALIN HOT SPRING BY USING HPGe GAMMA DETECTOR

Zin Mo Naing[1], Yin Nu[2], and Aye Aye Soe[3]

[1,2,3] Department of Natural Science, University of Computer Studies (Mandalay)
[1] zinmonaing.zmn@gmail.com

ABSTRACT

The objective of this research work is to investigate the natural radioactive nuclides in soil samples from Halin hot springs by Gamma Spectroscopic method. The samples were collected from the six different places near the hot spring in Halin, Wetlet Township, Sagaing Region. The collected samples were measured by using high resolution HPGe gamma detector and analyzed by the Gamma Vision 32 software. From the analyzed spectrum, ^{228}Ac, ^{212}Pb, ^{208}Tl, ^{214}Pb, ^{214}Bi, ^{212}Bi and ^{40}K radioisotopes were investigated. The activities of these nuclides in soil samples were calculated by using analyzed data. Then, the efficiency of HPGe detector were calculated by using analysed by full energy peak of experimental data.

KEYWORDS

Radionuclides concentration, Hot spring, Soil samples, HPGe detector,

1. INTRODUCTION

Radioactive decay occurs when an unstable (radioactive) isotope transforms to a more stable isotope. It is emitting a subatomic particle such as an alpha or beta particle. Gamma radiation is not a mode of radioactive decay (such as alpha and beta decay). Uranium, radium, and thorium occur in three natural decay series, headed by uranium-238, thorium-232, and uranium-235, respectively. In nature, the radionuclides in these three series are approximately in a state of secular equilibrium. The activities of all radionuclides within each series are nearly equal. [2]

Some of radioactive nuclides were found in Hot springs. **Hot springs** occur when water seeps into the earth, which is heated by magma (molten rock beneath the surface of the earth). The earth's pressure causes the water to rise above the surface forming a **hot spring**. The water from the **hot spring** is heated by geo thermal heat. The temperature of rocks within the earth increases with depth. It is geothermal gradient. If water percolates deeply into the crust, the water gets heated when it comes into contact with the hot rocks .In volcanic areas, water may come into contact with very hot heated by magma. [5,6]

2. MATERIAL AND METHOD

2.1. Study Area

The soil samples were collected from Halin hot spring, Halin Village, Wetlet Township, Sagaing Region. Halin, one of the significant Phu site, which lies about 2 miles south-east of Shwebo. It is situated at the west bank of the Irrawaddy River. All sampling area is covered within 150 square ft. The location map as shown in Figure 1.

Figure 1. Map of Sampling Area

2.2. Sample Collection

The six soil samples (S-11, S-12, S-21, S-22, S-31 and S-32) were collected from different places near the hot spring. These samples were cleaned by removing unwanted materials by meshes. Then, cleaned samples were weighted 500g each. These weighted samples were placed in plastic containers of 250 cm^3 volume and measured with HPGe detector.

S-11andS-12 S-21andS-22 S-31andS-32
Figure 2. Soil Samples Collected Area

Figure 3. Six Soil Samples from Different Places near Hot Spring

2.3 Experimental Procedure

The experimental work was performed at the University Research Center, University of Mandalay. The spectral analysis of the radionuclides of these samples were carried out using γ-ray spectrometer equipped with high purity germanium (HPGe) detector. The detector was placed inside a thick lead shield to reduce the background radiation. The detector output signal was taken to the PC equipped with a MCA (Multi Channel Analyzer).The software utilized for data acquisition is Gamma Vision-32 software, including peak search and nuclide identification modules. The system was calibrated for energy and efficiency on a regular basis. The experimental arrangement for HPGe detection system were shown in Figure 4. Each sample was counted for a period of 2 hours. Measurements with an empty plastic container under identical condition, were also carried out to determine the background counts. The latter was subtracted from the measured spectra of each sample to obtain the net radionuclide concentration.

2.3.1 Measurement of Energy Calibration and Efficiency Calibration

The energy of radioactive elements in soil samples were unknown. The standard radioactive sources of known energies were used to calibrate the spectrometer. The gamma rays spectra of seven standard sources ^{241}Am (59.95 keV), ^{133}Ba (81 keV, 302 keV and 356.02 keV), ^{57}Co (122 keV and 136 keV), ^{137}Cs (661.66 keV),^{54}Mn (834.85 keV), ^{22}Na (511 keV and 1274 keV) and ^{60}Co (1173.24 keV and 1332 keV) were used for 1800 seconds. The energy calibration curve is as shown in Figure 5. Establishing a direct relationship between photo peak energy and channel number could do the energy calibration process. The amplifier gain was adjusted until peaks were obtained at the desired channel numbers. In doing so, the energy calibration curve was obtained.

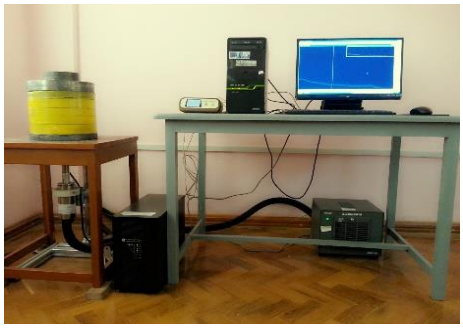


Figure 4. Experimental Arrangement for HPGe Detection System

Table 1. Energy Calibration Data for HPGe Detector

Sr. No	Channel	Energy (keV)
1	427.93	81.01
2	1460.2	276.29
3	1599.95	302.71
4	1881.08	355.86
5	2028.58	383.7
6	3497.95	661.62
7	6202.9	1173.23
8	7044.88	1332.51

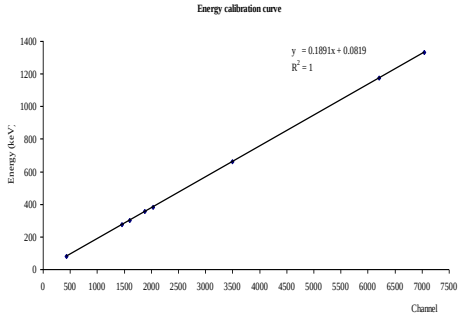


Figure 5. Energy Calibration Curve of the HPGe Detector

2.4 Data Analysis

2.4.1 Calculation of the Activity of Radionuclides

The activity of radionuclides the six soil samples from hot spring were calculated by following equation

$$A = \frac{N_A}{m\epsilon P_\gamma T}$$

Where, N_A = net area count for sample
m = mass of sample
P_γ = gamma ray intensity
T = counting time
ε= efficiency of the interest gamma energy. [1]

2.4.2 Calculation of the efficiency of Gamma-ray Detector from Experimental Data

An efficiency calibration of HPGe detector was performed with the use of the some standard sources of known energy and intensity. The detector efficiency can be calculated by using the energy produced from the sources and the net peak area under the full energy peak. The prepared plastic containers (samples) were placed on the detector window.

The full energy peak efficiency ε for γ-rays energy E can be measured as follow

$$\varepsilon = \frac{N_A}{A_t t P_\gamma}$$

Where, N_A = net areas of the full energy peak
A_t = the present activity of the standard sources
t = counting time
P_γ = emission probability for standard source [3]

2.5. RESULTS AND DISCUSSIONS

From the experimental results, the background spectrum is shown in Figure (6). The energy spectra for six soil samples are shown in Figure 7 to Figure12. By using

the analysed data, the calculated activity for each sample were show in Table 3 to Table 6 and graphically show in Figure14. The efficiency calibration data of the HPGe detector was tabulated was shown in Table 2. The efficiency curve of the HPGe detector was shown in Figure 13.

According to experimental results, we observed that the activity of ^{40}K is highest and that of ^{214}Bi is lowest in all samples. The activity of the ^{208}Tl is the highest in S12 and that is lowest in S-11. Activity of the ^{214}Pb is the highest in S-21 and that is lowest in S-22. Activity of the ^{212}Pb is the highest in S-11 and that is lowest in S-21. Activity of the ^{228}Ac is the highest in S-22 and that is lowest in S-21. Activity of the ^{214}Bi is the highest in S-22 and that is lowest in S-12. Activity of the ^{212}Bi is the highest in S-11 and that is absent in S-12, S-21, S-22, S-31 and S-32.

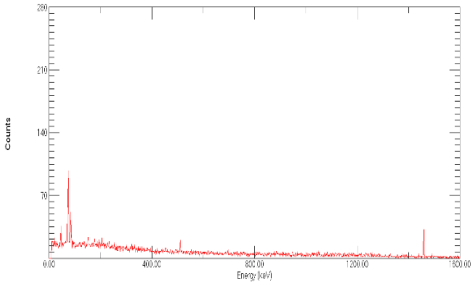


Figure 6. Background Spectrum

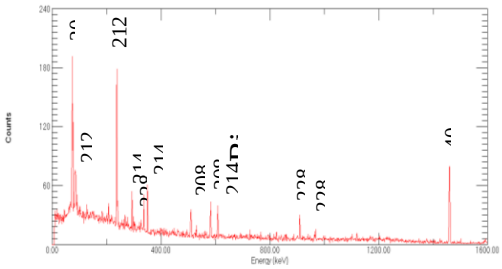


Figure 7. Soil Sample S-11 Spectrum

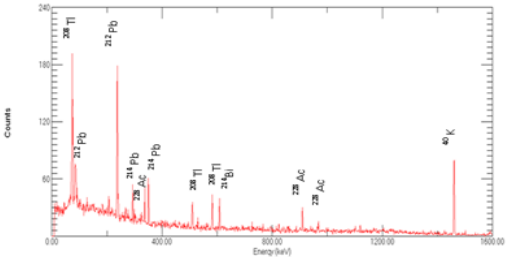


Figure 8. Soil Sample S-12 Spectrum

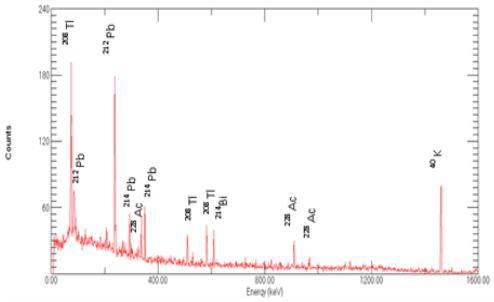


Figure 9. Soil Sample S-21 Spectrum

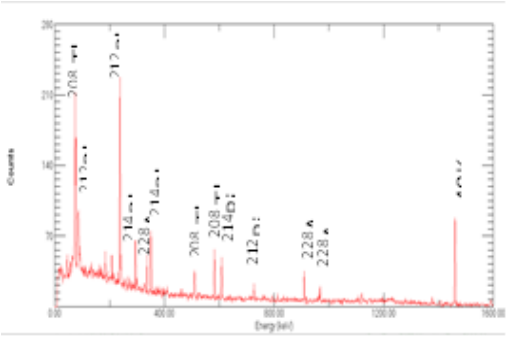


Figure 10. Soil Sample S-22 Spectrum

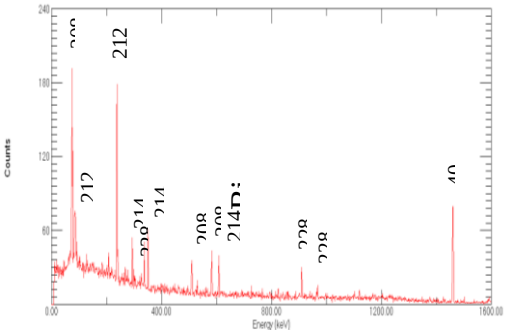


Figure 11. Soil Sample S-31 Spectrum

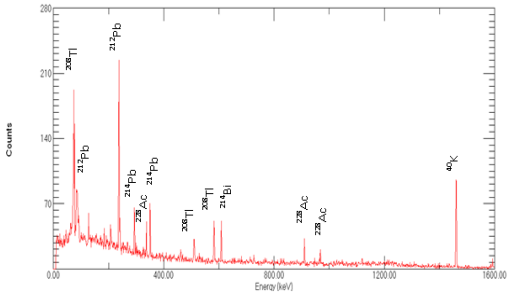


Figure 12. Soil Sample S-31 Spectrum

Table 2. Efficiency calibration data of HPGe detector

Sr. No	Gamma Energy (keV)	Efficiency
1	59.54	0.006094
2	88.04	0.013850
3	122.07	0.017806
4	165.85	0.015632
5	391.71	0.007142
6	661.62	0.004164
7	898.02	0.003049
8	1173.23	0.002390
9	1332.51	0.002098
10	1836.01	0.001585

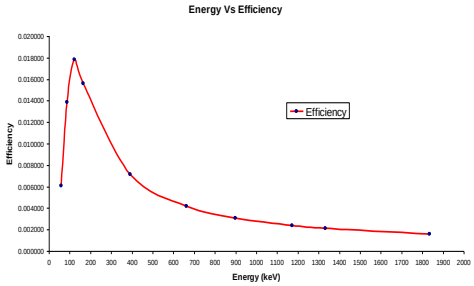


Figure 13. Efficiency Calibration curve of the HPGe detector

Table 3. Activity of the Various Elements of Six Soil Samples

Samples	Activity (Bq kg^{-1})				
	^{228}Ac	^{214}Pb	^{208}TI	^{206}TI	^{214}Bi
S-11	20.6	12.9	12.1	6.93	11.02
S-12	19.8	15.2	12.1	6	10.74
S-21	19.8	14	15.5	6.86	18.05
S-22	26.5	11.9	33.9	16.6	33.06
S-31	24.9	12.6	12.3	5.78	14.74
S-32	28.3	23.4	22.3	15.8	24.8

Table 4. Activity of the Various Elements of Six Soil Samples

Samples	Activity (Bq kg^{-1})		
	^{228}Ac	^{228}Ac	^{40}K
S-11	15.3	4.89	893
S-12	63.3	19	881.2
S-21	35.6	35.5	842.2
S-22	170	61.2	829.8
S-31	156	28.8	790.1
S-32	154	17.8	932

Table 5. Activity of the Various Elements of Six Soil Samples

Samples	Activity (Bq kg^{-1})				
	^{214}Pb	^{212}Pb	^{214}Pb	^{212}Pb	^{212}Bi
S-11	-	26	17.76	296.4	50.13
S-12	-	27.5	12.26	-	-
S-21	-	25.5	8.25	-	-
S-22	-	45.8	14.7	-	-
S-31	97.7	22.8	3.003	-	-
S-32	30.1	24.7	12.51	-	-

Table 6. Activity of the Various Elements of Six Soil Samples

Samples	Activity (Bq kg^{-1})				
	^{208}TI	^{214}Pb	^{208}TI	^{212}Pb	^{228}Ac
S-11	115	66.7	114.7	49	84.46
S-12	482	56.5	180	35.4	82.84
S-21	400	41.2	31.99	3.35	7.15
S-22	346	49.7	183.3	42.9	43.46
S-31	434	39.7	122.6	18.3	-
S-32	343	67.5	136	36.1	-

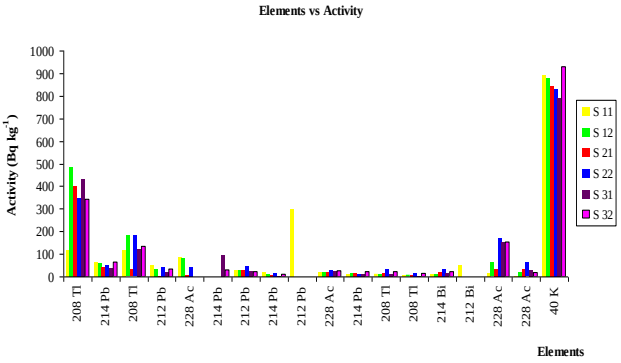


Figure 14. Elemental Vs Activity Graph for All Samples

3. CONCLUSIONS

Hot spring water dissolved large amounts of minerals. The mineral water in hot springs can reduce stress by relaxing tense muscles and sleep well at night. When spending time in a mineral bath, the skin absorbs minerals and nutrients that smooth and heal. In this study, the results show that no toxic chemical substances were found by gamma detection method. It can be assumed that these radionuclides originated from the hot spring, not from the surrounding area. These radionuclides were the products of U and Th natural radioactive series. Therefore, this hot springs should use bathing, relaxation and medical therapy for human health.

ACKNOWLEDGEMENTS

The authors would like to thank Dr San San Tint, Pro Rector, University of Computer Studies (Mandalay), for her encouragement. We are also grateful to Professor Dr Nwe Nwe Htoon, Head of Department of Natural Science, University of Computer Studies (Mandalay) for her suggestions.

REFERENCES

[1] S. A. Saqan, M. K. Kullab, A. M. Ismail, (2000). "Radionuclides in hot mineral spring water in Jordan, *Journal of Environmental Radioactivity*'' *52-99-107.*

[2] Report of the United Nations Scientific Committee on the effects of atomic radiation report to the General Assembly, United Nations, New York, USA, 2000.

[3] Copper, J. R, Randle, K, Sokhi, R. S (2003), "Radioactive Releases in the Environment impact and Assessment,'' England, John Wiley & Sons, Inc, Publication.

[4] www.britannica.com.

[5]https://hive.rochesterregional.org/2021/07/health-benefits-of-mineral-baths

Authors

Biography

First A. Daw Zin Mo Naing graduated the M.Sc (2016) and MRes (2016) degrees from Shwebo University and Mandalay University. She is now serving as an Assistant Lecturer at the Department of Natural Science in University of Computer Studies (Mandalay).

Second B. Daw Yin Nu graduated M.Sc(Physics) degree in 2005 at the University of Magway. She first served as a Tutor in 2005 at (TU) Magway. And then She was promoted to Assistant Lecturer in 2013 at the (TU) Magway. She was transferred and promoted as a Lecturer to the UCS(Maubin) in 2017. She is currently serving as an Associate Professor in the Department of Natural Science at the University of Computer Studies (Mandalay). She has received the publication paper from (TUMGJRI) 2023, Vol-1, Issue-3. So, she would like to keep on doing research in this field.

Third C. Daw Aye Aye Soe graduated M.Sc (Physics) degree in 2005 from Mandalay University. She is now serving as an Associate Professor at the Department of Natural Science in University of Computer Studies (Mandalay).

PERFORMANCE COMPARISON BETWEEN DISCRETIZATION BASED CLASSIFIERS

Myint Thidar[1], Thuzar Hnin[2], and Khin Mu Aye[3]

[1, 2, 3]Faculty of Information Science, University of Computer Studies (Mandalay)
[1]myintthidar1989@gmail.com, [2]thuzarhninn99@gmail.com,
[3]khinmuaye70@gmail.com

ABSTRACT

Today the majority of data mining and machine learning methods employ classification based on discrete data and clustering. In practice, most data have continuous properties. Discretization tackles this issue by identifying numerical intervals that are more concise to describe and specify. One of the critical data pre-processing phases in knowledge extraction is the discretization of continuous attributes. A good discretization method not only reduces the demand for system memory and improves the efficiency of data mining and machine learning algorithms, but it also makes the knowledge retrieved from the discretized dataset more compact, understandable, and usable. In this research, two untrained approaches are proposed: Equality of Frequency (EF) and Width Interval (EW)), as well as using discretion based on CART as one supervised way. The classification accuracy results show that, when compared to other algorithms, one discretization method outperforms the others.

KEYWORDS

Supervised, Unsupervised, Discretization, Classifiers.

1. INTRODUCTION

Various types of classification algorithms in data mining and machine learning can only be used for data that is discrete. However, most data have a continuous feature. The datasets often include a combination of data that is discrete, nominal, and continuous. Intervals between values in a continuous series define discrete numbers. An attribute's number of continuous numbers may be infinite, whereas the quantity of discrete numbers can be limited and might have a final value. The procedure for transforming from continuous to discrete values is referred to as discretization [3].

Discretization is typically used for sorting and rearranging continuous attribute variables into classified characteristics [3]. Furthermore, discretization improves the accuracy and speed of learning. The discretization methods are classified into three types: static versus dynamic, global versus local, and supervised versus unsupervised. As a result, discretization is critical in the area of machine learning [7]. Machine learning algorithms come in two basic types, supervised and unsupervised. The unsupervised type takes the dataset and determines how best to *cluster* the inputs into likely groups of common data. Supervised algorithms, on the other hand, know the classification of the input data, and learn how to best separate the dataset into the known classes.[4]

2. DISCRETIZATION METHODS

The discretization of data is a general-purpose initial processing approach it decreases the number of different results for an individual variable that is continuous by partitioning its domain of values into a limited number of discontinuous intervals and subsequently assigning specific descriptive labels to these intervals. As a result, rather than the delicate individual values at this higher level of knowledge representation, data are processed or reported, resulting in a simplified representation of data in the context of the

data explorations and data mining procedure [5]. Typically, Typically, discretization procedures include continuous value sorting, determining cut points, and lastly implementing the procedure for conversion [2].

The purpose of discretization is to discover a set of cut points that will divide the range into a limited number of intervals. There are primarily two discretization jobs. The initial objective is to determine the quantity of discrete intervals. Only a few discretization algorithms do this; frequently, the user must select the number of intervals or supply a heuristic rule. The second challenge is to determine the breadth, or borders, of the intervals given a continuous feature's value range [5]. The discretization approaches' hierarchical framework is shown in Figure (1) [5].

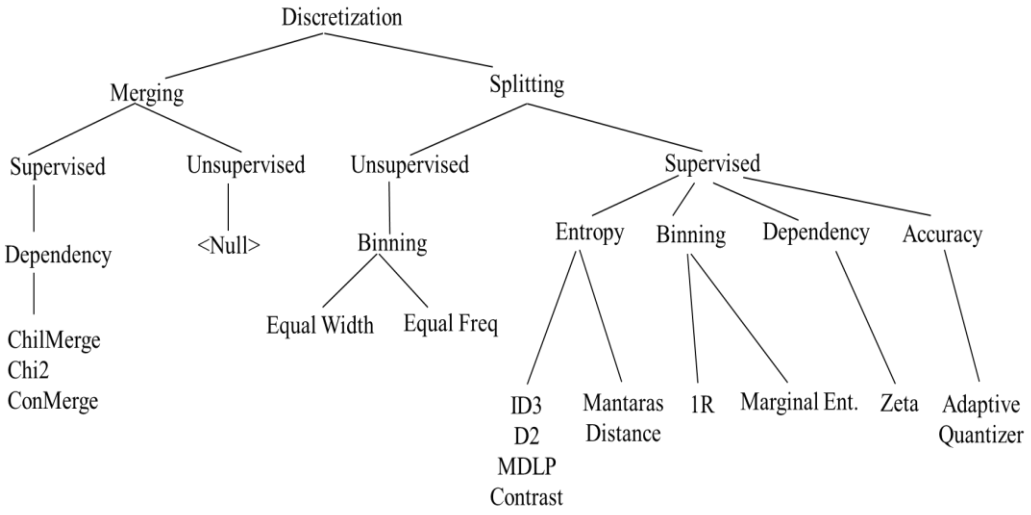


Figure 1. A Hierarchical Framework for Discretization Method

2.1. Methods for unsupervised discretization

Unsupervised discretization procedures include simple discretization (equal-width and equal-frequency interval binning) and more complex ones that use grouping analysis, such as k-means the discretization process. The width or frequency set by the user divides continuous ranges into subranges [5].

2.1.1 Discretization of Equal Width Intervals

The most basic discretization approach is equal-width interval discretization, which separates a feature's observed value range into k equal-sized bins, where k is a user-specified number. Sorting A = a_0, a_1, a_2, a_{n-1}, a_n, a_{min} = a_0, and a_{max} = a_n are all continuous features, to find the minimum and maximum observed values. A is an array of continuous values. Using the following formula, divide the variable's observed value range into equally sized groups to compute the interval [3].

$$\text{Interval} = \frac{(a_{max} - a_{min})}{K} \quad (1)$$

$$\text{Boundaries} = a_{min} + (i * \text{interval}) \quad (2)$$

i = 1...k-1 can be used to form the borders. Figure 2 depicts the discretization processes for an Equal Width Interval. Table 1 below provides an example of these steps. Where I stands for the instance, Ag represents the age of the people, C represents the value of the classes Female and Male, A represents the sorted value of Age, and D represents the discretized value of Age.

Input: data consists of an array with values for continuous attributes, A = {a_0, a_1, a_2, … a_{n-1}, a_n} and the number of components is denoted by k, where k > 0;

Output: discrete-valued data.

Step 1: A is the sort value in ascending order.

Step 2: Use equation 1 to calculate the interval.

Step 3: Bin the data using the formula for defining boundaries and get the A split point.

Step 4: The array's attribute value must be placed within the same limits.

Figure 2. Equal Width (EW) Interval Discretization Algorithm

Table 1. Example for Discrete Intervals of Equal Width

I	1	2	3	4	5	6	7	8	9	10
Ag	20	15	50	30	70	45	60	75	35	65
C	M	M	F	M	M	F	F	F	M	F
A	15	20	30	35	45	50	60	65	70	75
Here a_{min} = 15, a_{max} = 75 and let k=3 So, Interval = (75-15)/3 = 20 then Boundary = 15+20 = 35 0 = [15,35); 1 = [35, 55); 2 = [55, 75]										
D	0	0	1	0	2	1	2	2	1	2

2.1.2 Discretization of the Equal Frequency Interval

The algorithm of equal frequency identifies the discretized attribute's values between the lowest and maximum, in ascending order, all values are sorted, and then separates the continuous data were sorted into k intervals with around n/k data occurrences in each interval with adjacent values. Many occurrences of a continuous value with similar frequency may result in the occurrences being assigned to various bins. By splitting intervals within the domain with the same data point distribution, this approach attempts to address the restrictions of equal-width interval discretization. Because occurrences of data with similar values must be arranged in the same space, generating exactly k equal frequency intervals is not always achievable. Proportional k-interval discretization is another name for this technology [5].

Steps of equal frequency discretization are presented in Figure 3. Table 2 below provides an example of these steps. Where I is the instance, Ag is the people's age, C is the Female and Male class value, A sorted Age value, and D is a discretized Age value.

Input: A = {a_0, a_1, a_2, …. a_{n-1}, a_n}, and k is the total number of pieces. If k is greater than zero.

Output: Discrete-valued data.

Step 1: Sort value A in ascending order.

Step 2: Divide the number of elements in the array by the number of sections to calculate the data instances in each interval.

Step 3: An array attribute's value must be placed within the same constraints.

Figure 3. Equal Frequency (EF) Discretization Algorithm

Table 2. Example for Discrete Intervals of Equal Frequency

I	1	2	3	4	5	6	7	8	9	10
Ag	20	15	50	30	70	45	60	75	35	65
C	M	M	F	M	M	F	F	F	M	F
A	15	20	30	35	45	50	60	65	70	75
Here a_{min} = 15, a_{max} = 75 and set k to 3 As a result, Interval = 10/3 = 3.3 and up (interval). Then interval = 3; each binning had about 3 elements, and the rise is added to the last part 0 = [15,35); 1 = [35, 60); 2 = [60, 75]										
D	0	0	1	0	2	1	2	2	1	2

2.2.Supervised Discretization Method

Supervised discretization approach transforms numerical data into categorical data and selects discretization nodes based on class information [3]. CART is used on the 1R discretization method and the various methods of discretizing are to evaluate the training data so that the time can be reduced a great deal without considerably impacting the resulting classifier's accuracy.

2.2.1 1R discretization method

1R is a binning-based supervised discretization algorithm. 1R separates the range of continuous values into a series of discontinuous intervals and adjusts the boundaries based on the continuous values' class labels after sorting the continuous values. Each interval should have a minimum number of instances (6 by default), except the last interval, which includes the remainder of instances that have not yet been grouped in any interval. The adjustment at the boundary prevents one from terminating an interval if the next occurrence has the identical class label as the interval's majority group label up to that point [1].

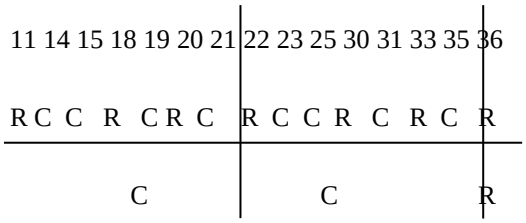


Figure 4. Cut-Points Searched by 1R

2.2.2 CART-Based Discretization

Machine learning algorithms are programs that a computer uses to determine the set of rules by which to classify a given dataset.

Input: Data is a continuous-valued array.

A = {a0, a1, a2, … an-1, an}

Output: The values in data are discrete.

Step 1: Sort A in ascending order.

Step 2: In the case of class data, utilize 1R discretization.

Step 3: To determine the best-split point for A, compute the Gini index for each split point.

Step 4: Select the split point that has the lowest Gini index.

Step 5: When the Gini index falls below a particular threshold, the partition on each split point is terminated recursively.

Figure 5. CART Algorithm for Discretization

Classification algorithms take an input dataset and classify it according to a predetermined rule or set of rules. This paper focuses on Decision Trees, which are a particular type of classifier, and the Classification and Regression Tree (CART) algorithm that grows decision trees and the speed at which a classifier can be built. CART is a widely used binary decision tree technique, with the Gini index serving as the evaluation function for splitting. For each attribute, the Gini index measures employ a binary split. As the separating attribute, the attribute with the lowest Gini index is picked [4].

Table 3. Example for CART Discretization

I	1	2	3	4	5	6	7	8	9	10
Ag	20	15	50	30	70	45	60	75	35	65
C	M	M	F	M	M	F	F	F	M	F
A	15	20	30	35	45	50	60	65	70	75
Use 1R discretization method Ag <= 50 = S1 Male (4 out of 6) Ag > 50 = S2 Female (3 out of 4) =7/10 = 70% Determine the Gini index at each split point $Gini(D)=1-\sum_{j=1}^{n} p_j^2$ = 1-0.5 = 0.5 Gini $_A$ (D) = $\frac{\lvert D1\rvert}{\lvert D\rvert}$ Gini (D_1) + $\frac{\lvert D2\rvert}{\lvert D\rvert}$ Gini (D_2) Gini (S1) = 0.2667 Gini (S2) = 0.15 0 = [15, 60) 1 = [60, 75]										
D	0	0	0	0	1	0	1	1	0	1

3. CLASSIFICATION

The categorization concept essentially refers to the distribution of data on a dataset around distinct classifications [3]. Many branches are working on Bayesian classification, neural networks, decision trees, and evolutionary algorithms, among other things. Among these, the decision tree has grown in popularity for a variety of reasons: (a) humans can interpret it more easily than neural networks or a Bayesian-based approach; (b) it is more efficient for huge amounts of training data than neural networks, which would take a long time on thousands of repetitions. (c) Unlike other technologies, a decision tree algorithm does

not require domain knowledge or prior information; and (d) it has high classification accuracy [1].

3.1 Algorithm of Decision Trees

A decision tree is a type of tree structure that looks like a flowchart and is generated by a recursive divide-and-conquer process that partitions the data. Each internal node in a decision tree represents an attribute test, every connected branch represents a test outcome, and each leaf node is connected with the desired class (or class). Classifying an unidentified instance starts using the primary node and proceeds through until inside nodes it reaches a node of a leaf [1].

Data classification takes two steps to use the decision tree technique procedure that includes both learning and categorization. During the learning phase, the classification algorithm analyzes previously existing education data to construct a model. The model is represented by a decision tree or categorization rules. Test data, on the other hand, is used during the classification phase to assess the precision of the categorization rules of a decision tree. The rules are used to classify fresh data if the precision is within a reasonable range. To create the tree, it must be decided which domains will be used and in what order. The Gini index, like ID3, creates decision trees from a set of training data utilizing the concept of information entropy. The Gini index makes use of the fact that any data attribute can be utilized to make a choice that divides the data into smaller groupings [1].

As a heuristic, the Gini index is utilized in classic Gini index classification to determine the characteristic that will best divide the training tuples into different classes. Gini index If a data collection D comprises n classes' examples. The fundamental Gini equation Gini (D) is described as follows:

$$\text{Gini}(D) = 1 - \sum_{j=1}^{n} p_j^2 \qquad (3)$$

Class j's relative frequency is denoted in D. The Gini index genial (D) is determined as follows when a data set D is partitioned into two subsets D1 and D2:

$$\text{Gini}_A(D) = \frac{|D1|}{|D|}\text{Gini}(D_1) + \frac{|D2|}{|D|}\text{Gini}(D_2) \qquad (4)$$

For each attribute, all feasible binary splits are explored. As the splitting subset for a variable with discrete values, the subset that offers the lowest Gini index for that attribute is chosen. The decision tree techniques ID3, C 4.5, and C5.0 modules are employed in this research. [3]

4. PROPOSED SYSTEM

In this work, unsupervised (EW and EF) and supervised (CART) discretization approaches were used to transform the continuous value of a characteristic Table 4 displays the UCI dataset as discrete values. Algorithms for machine learning classification (ID3, C4.5, and C5.0) are used, we used discrete datasets of converted values to do categorization. The quantity of components utilized in the discretization of EW and EF procedures is (k= 3 to 22; k+= 2), correspondingly, the means (3, 5, 7... 21) are utilized, and classification performance is determined at each k number, in Table 5, we have arithmetic mean of all part numbers based on each classification algorithm's great performance, in this situation k = 11, we calculated classification accuracy based on the value of k. As previously stated, the CART discretization method calculates the number of components using the 1R discretization method, each data set property is subdivided into a distinct number of components.

This system starts with loading the numerical dataset, followed by applying the discretization methods (EF, EW, 1R). After preprocessing, the user selects the desired classification method (ID3, C4.5, C5.0, and CART) and applies it to the preprocessed data. The result is a decision tree with rules generated from the classification method. Then, this system evaluates the performance of decision tree by using the dataset and then this system calculates the

accuracy. Finally, this system compares the accuracy and then determines which result performs better. The best result leads to optimal performance. System flow diagram about classification methods is shown in Figure 6.

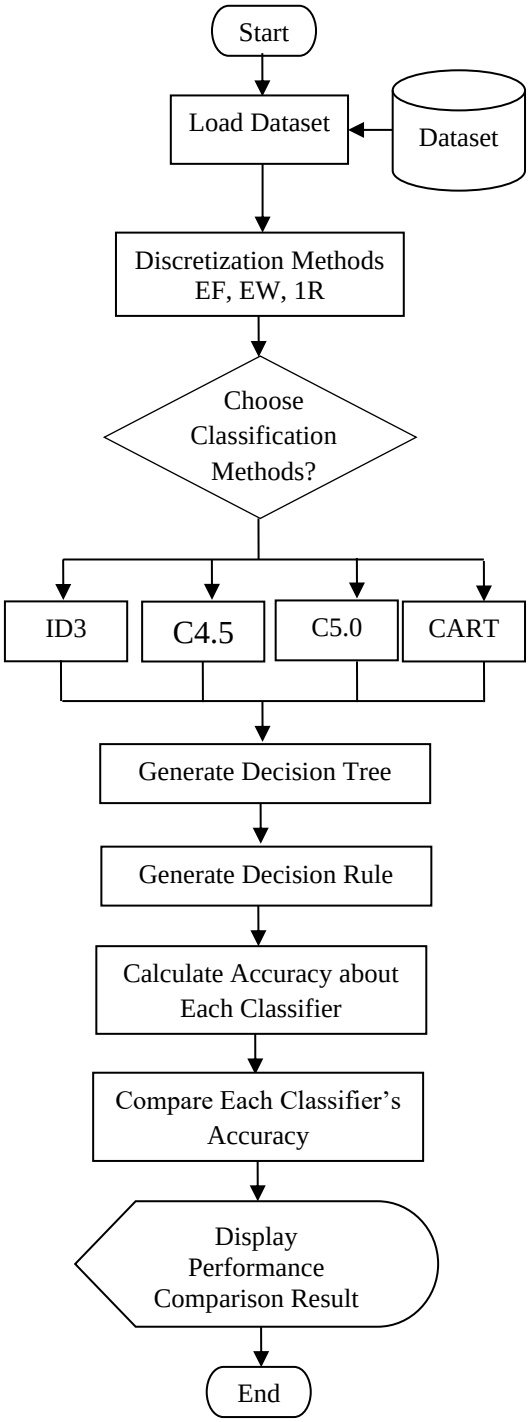


Figure 6. System Flow Diagram about Classification Methods

4.1 Decision Tree

A binary decision tree using the discretization method for the vehicle data set is shown in Figure 7.

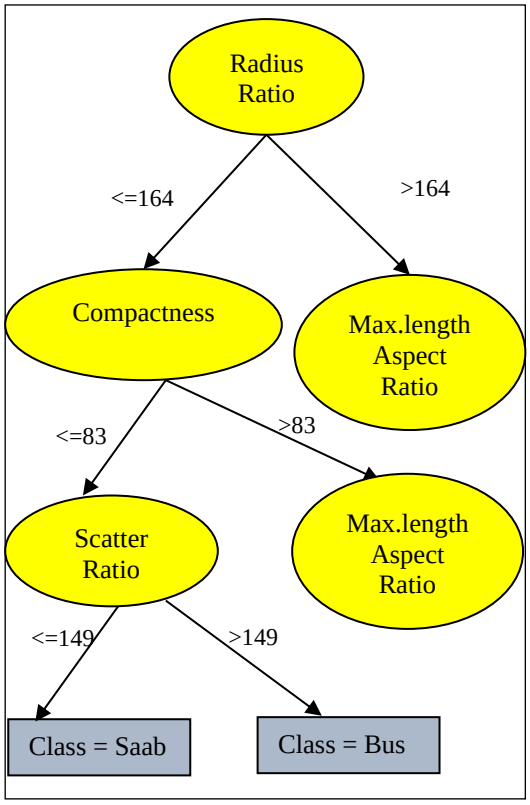


Figure 7. Binary Decision Tree using Discretization Method for Vehicle Data Set

4.2 Decision Rule

Binary decision rules using the discretization method for the vehicle data set are as follows:

- Rule1: IF Radius Ratio < = 164 AND Compactness < = 83 AND Scatter ratio < = 149 THEN Class = saab
- Rule 2: ELSE IF Radius Ratio <= 164 AND Compactness <= 83 AND Scatter ratio > 149 THEN Class = bus.
- Rule 3: ELSE IF Radius Ratio < =164 AND Compactness > 83 AND Maxlengthaspectratio <= 8 AND Elongatedness < = 48 AND Scatter ration < = 142 AND Circularity <= 39 THEN Class = opel

- Rule 4: ELSE IF Radius Ratio < =164 AND Compactness > 83 AND Maxlengthaspectratio <= 8 AND Elongatedness <= 48 AND Scatter ration <= 142 AND Circularity > 39 THEN Class = saab
- Rule 5: IF Radius Ratio < = 164 AND Compactness > 83 AND Maxlengthaspectratio <= 8 AND Elongatedness < = 48 AND Scatter ration > 142 THEN Class = bus.

According to experimental results, there are 4 classes, 18 features, and 56 instances of data in the vehicle data set. Minimum Reduction Accuracy (MRA) 2 is equal to 16 rules and 100 percentage of accuracy in the classification method of CART.

4.3 Experimental Result

The holdout accuracy method is a common technique used in machine learning for evaluating the performance of a classification model. It involves splitting the dataset into two subsets: a training set and a testing set. The hold-out scheme randomly partitions a data set into two mutually exclusive subsets, of which one is the training data set, and the other is the test data set. The training data is used for inducing a classifier and the test data is then used for accuracy estimation. The holdout accuracy estimate is usually taken *n* times for different partitions. The accuracy of the classifier is then computed as the number of correct predictions and total number of predictions. [7].

Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

All data sets are obtained from UCI Machine learning Repository (UCI-MLR) (Dua & Graff,2019). *Https://Archive.Ics.Uci.Edu/Ml/Datasets*

Where NA denotes the number of attributes, NI is the number of instances, and NC is the number of classes.

Table 4. Dataset Names and Properties

Dataset Name	Number of Instances (NI)	Number of Attributes (NA)	Number of classes (NC)
Australian	690	14	3
Breast	699	9	2
Glass	214	10	6
Heart	270	13	2
Vehicle	846	18	4
Iris	150	4	3
Wine	178	13	3
Pima	768	8	2
Ionos	351	34	3
Thyroid	215	5	2

The examination of the findings shown in Table 5 from 10 datasets demonstrates that the EF approach has a good classification precision for only Australian and Ionos datasets and in EW approach has a high accuracy in the classification of Wine and Throyid datasets while in CART method provides 100 percentage categorization precision for the vehicle datasets with eighteen features and four classes. In most cases, supervised CART gives greater precision than unsupervised EF and EW.

Table 5. Discretization Method of Classification Accuracy

Datasets Name	**ID3**		**C4.5**		**C5.0**		**CART**
	EF	EW	EF	EW	EF	EW	
Australian	81.24	80.11	81.45	79.56	93.34	92.77	93.89
Breast	97.42	97.42	93.82	93.82	95.42	95.42	95.99

Glass	68.69	68.69	55.61	55.61	74.81	73.81	75.82
Heart	82.22	82.22	58.15	58.15	79.63	79.63	81.85
Vehicle	85.23	84.23	80.11	82.34	95.11	97.22	100
Iris	95.33	95.33	89.33	89.33	96.00	96.00	90.24
Wine	89.00	90.75	83.11	85.11	85.24	86.34	88.22
Pima	75.91	75.91	51.69	51.69	75.13	75.13	79.89
Ionos	69.23	67.76	71.45	65.23	74.24	72.22	75,67
Thyroid	70.50	71.23	69.55	70.11	72.56	76.25	78.77

5. CONCLUSIONS

Discretization of continuous features is useful in data pre-processing before applying a variety of machine learning and data mining techniques to real-valued data sets. Discretization algorithms have established advantages and disadvantages in comparison to one another. Even though the equal-width discretization approach is the most appealing among the others, there have been reports of significant data loss after the discretization procedure fails because the size of the gap is incorrectly measured between the intervals of equal length. Other techniques, such as the EW and EF, need the user to enter the settings, whereas CART uses a 1R discretization method that decides on the parameters automatically. It is conceivable to state that in this scenario if the parameters entered by the user are not right, the results produced may be affected. In the data categorization process, the data discretization procedure is critical because it is the CART-based discretization method that is affected by the type of data in the discretization and supervised, static, and global manner for data discretion is mostly correct, and enhances the data categorization probability is relatively high. As a consequence of this research in algorithmic classification methods (ID3, C4.5, C5.0, CART), one discretization approach outperforms the others. Thus, we show that there exists a new successful technique for discretization.

ACKNOWLEDGMENTS

I would want to thank everyone who contributed directly or indirectly to the success of this work for their encouragement, excellent ideas, valuable direction, and assistance in writing this paper.

REFERENCES

[1] Gorunescu, F. (2011) "Data Mining: Concepts, models, and techniques", *Springer Science & Business Media*, vol. 12.

[2] Yan, J., Zhang, Z., Xie, L., & Zhu, Z. (2019) "A unified framework for decision tree on continuous attributes", *IEEE* Access, 7, 11924-11933.

[3] Toulabinejad, E., Mirsafaei, M., & Basiri, A. (2023) "Supervised discretization of continuous-valued attributes for classification using the RACER algorithm", *Expert Systems with Applications*, 121203.

[4] Barbieri, M., Dayton, K., Hebert, W., & Robinson, K. (2020) 'Experiments with Decision Tree Classifiers–Discretization of Numerical Attributes".

[5] Thaiphan, R., & Phetkaew, T. (2018) "Comparative analysis of discretization algorithms on a decision tree", *IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pp. 63-67.

[6] Alasadi, S. A., & Bhaya, W. S. (2017) "Review of data preprocessing techniques in data mining", *Journal of Engineering and Applied Sciences*, vol. 12, pp. 4102-4107.

[7] Yacob, Y. M., Sakim, H. A. M., & Isa, N. A. M. (2012) "A Discretization Algorithm Based on Extended Gini Criterion", *International Journal of Machine Learning and Computing*, vol. 2, no. 3.

Authors
Biography

First A. Myint Thidar received M.C.Sc from the University of Computer Studies, Mandalay (UCSM) in 2017. She is now working as an Assistant Lecturer at the University of Computer Studies, (Mandalay) and she is interested in Machine learning, Database Management Systems, and Software Engineering.

Second B. Thuzar Hnin

Third C. Khin Mu Aye

COVID-19 DIAGNOSIS SYSTEM BY USING IMPROVED NAIVE BAYESIAN (INB) CLASSIFIER

Yin Min Tun[1], Badounmar[2], and Kyawt Yin Khaing[3]

[1,2,3] Faculty of Computer Science, University of Computer Studies (Mandalay)
[1]yinmintun.ymtkyaw@gmail.com

ABSTRACT

Today, COVID-19 disease is a worldwide health problem. It is a communicable disease and the diagnosis reports are received between 12-24 hours of illness. Conducting the test on a large population is not practical due to the spread of COIVD-19 worldwide, remote and high altitude places, and the exponential growth of the virus. In this situation, an automated medical diagnosis system enhances the medical care and it can also reduce costs. So, this system is proposed as the medical diagnosis system by classifying the COVID-19 disease. For classification, this system presents the improved Naive Bayesian (INB) classifier. This classifier classifies the class label of unknown instants and it produces more accurate result than the original Naive Bayesian classifier. So, this system helps the user to quickly know the COVID-19 disease status and to reduce the check-up costs.

KEYWORDS

COVID-19, Classification, Improved Naive Bayesian.

1. INTRODUCTION

COVID-19 is the infectious disease caused by the coronavirus that was first discovered in Wuhan City, China and then spread throughout the world. The first case was discovered in Wuhan, China, in December 2019. On December 31, 2019, the coronavirus disease 2019 was initially reported to the World Health Organization (WHO). On January 30, 2020, the COVID-19 outbreak was declared a global health emergency. Covid-19 is a very serious disease because it mainly attacks the lungs in the human body and can cause death for the sufferers, especially such as congenital diseases or a weak immune system. COVID-19 has rapidly affected our day-to-day life, businesses, public health, food systems, education systems and employment. With the emergence and spread of COVID-19, different methodologies are used to analyze these COVID-19 cases.

So, this system proposes the medical diagnosis system that can check the patient who suffers the COVID-19 disease or not. For checking, this system uses the classification method. There are many classification methods that are Bayesian classifier, decision tree, k-nearest neighbor (KNN) classifier, support vector machine and so on. Among them, the proposed system uses the Bayesian classifier for COVID-19 disease classification. Bayesian classifier is applicable if the classes originate from some statistical distribution. Bayesian classifiers are based on the Bayes formula from probability theory. Bayesian classifiers have performed well on wide range of applications-including medical diagnostics, text characterization, and information retrieval and spam filtering. This classifier offers greater classification accuracy.

Among many Bayesian classifiers, this system proposes the improved Naive Bayesian (INB) classifier. The proposed classifier is enhanced the original Naive Bayesian (NB) classifier. The original NB classifier faces a little problem about probability estimation for the user inputted data samples. In this situation, the INB classifier can solve this problem. So, the

proposed system gives the correct and accurate COVID-19 disease result for the patient.

2. Related Work

In 2021, Kuyo, Mwalili and Okango [1] presented machine learning approaches for classifying the distribution of Covid-19 sentiments. The goal of this study was to categorize and forecast the regional distribution of both positive and negative feelings toward the COVID-19 epidemic using machine learning techniques. Machine learning has been used in this study to categorize the spatial distribution of Covid-19 Twitter sentiments as either positive or negative.

In 2021, Bhatia and Malhotra [2] presented the Naive Bayes classifier for predicting the coronavirus. Creating a system to forecast the coronavirus illness was the main goal of this paper. As a result, the current study presented several data mining methods for predicting COVID-19 positive instances.

In 2022, Yoshikawa [3] predicted the naive bayes classifier infection in a close contact of COVID-19. Then, this system discussed a comparative test for predictability of the predictive model and healthcare workers in Japan. Using posterior probabilities, a machine learning model was developed that can assess the probability of infection in close contact with COVID-19 patients during the incubation period based on the contact status reported from the index case. A verification test was performed on 169 new close contacts to confirm genuine predictability, and the machine learning model was compared with four experienced healthcare workers for the predictability.

3. Machine Learning

Machine learning might be used to describe the process of improving performance in the future by drawing on prior knowledge, in this case, historical data. This field focuses only on autonomous learning techniques. Learning is the process of autonomously modifying or improving an algorithm based on prior "experiences" without the need for outside human assistance [4].

To educate machines how to handle data more effectively, machine learning is used. The user may occasionally be unable to decipher patterns in the data or draw conclusions from it after seeing it. In that instance, the amount of available datasets is driving up demand for machine learning. Machine learning is used by many industries, including the military and medical, to retrieve pertinent data. Learning from the data is the aim of machine learning. The topic of teaching robots to learn on their own has been extensively researched. Numerous programmers and mathematicians use a variety of techniques to try to solve this challenge [5].

Machine learning comes in different varieties. These include reinforcement, multi-task, ensemble, neural network, supervised, unsupervised, semi-supervised, and instance-based learning. Support vector machines, Naive Bayes, and decision trees are examples of supervised learning [5]. Among them, the proposed system uses the Naive Bayesian classifier.

3.1. Naive Bayesian Classifier

Probabilistic classifier based on applying Bayes theorem is called a naive Bayesian classifier. A Naive Bayesian model is particularly helpful in the field of medical science for diagnosing heart patients since it is simple to construct and does not require complex iterative parameter estimates. The Naive Bayesian classifier, which is popular because it frequently outperforms more complex classification techniques, performs surprisingly well for its simplicity [6]. When compared to other classifiers, this one has the lowest error rate [7].

The posterior probability, P(c|x), can be computed using the Bayes theorem in the forms P(c), P(x), and P(x|c). The naive bayes classifier makes the assumption that a predictor's (x) value has an independent impact on a given class (c) regardless of the values of other predictors. Class conditional independence is the term for it [6].

3.2. Improved Naive Bayesian (INB) Classifier

Due to the probability calculation problem from the original Naive Bayesian classifier, this system proposes the improved Naive Bayesian (INB) classifier. The proposed classifier firstly checks the user inputted symptom value that are included in the training data or not. If the user inputted symptom (feature) value doesn't include in the training data, the original Naive Bayesian classifier can't produce the class result because of the zero probability result.

To solve this problem, the proposed improved Naive Bayesian classifier firstly checks the feature (attribute) value of unknown sample. INB classifier considers the feature value at the multiplication of feature probability calculation from step 4 of original Naive Bayesian classifier. So, the INB classier can produce the more precise result than the original NB classifier.

The processing steps of INB classifier is as follows:

1. Each data sample is represented by n-dimensional feature vector, X = (x1, x2… xn) depicting n-measurements made on the sample from n- attributes, respectively, A1, A2,….., An.
2. Suppose that there are m classes, C1, C2, …, Cm. Given an unknown data sample X, Naïve Bayesian classifier assigns an unknown sample X to the class Ci if and only if
   $P(C_i \backslash X) > P(C_j \backslash X)$ for $1 \leq j \leq m, j \neq I$ (1)
   The class Ci for which $P(C_i \backslash X)$ that is maximized, called maximum posteriori hypothesis. By Bayes theorem,
   $P(C_i \backslash X) = P(X \backslash C_i)\, P(C_i) / P(X)$ (2)
3. As P(X) is constant for all classes, only $P(X \backslash C_i)\, P(C_i)$ need to be maximized.
4. Given data sets with many attributes, the Naive assumption of class conditional independence is made. Thus,
   $P(X \backslash C_i) = \prod^{n} P(x_k \backslash C_i)$ (3)
   The probability $P(x1 \backslash C_i)$, $P(x2 \backslash C_i)$,… $P(x_n \backslash C_i)$ can be estimated from the data samples.
5. In order to classify an unknown sample X, $P(X \backslash C_i)\, P(C_i)$ is evaluated for each class $C_{i,.}$ Sample X is then assigned tothe class C_i if and only if
   $P(X \backslash C_i)\, P(C_i) > P(X \backslash C_j)\, P(C_j)$ for $1 \leq j \leq m, j \neq i$ (4)
   In other words, it is assigned to the class C_i for which $P(X \backslash C_i)\, P(C_i)$ is the maximum [8].

4. PROPOSED SYSTEM DESIGN

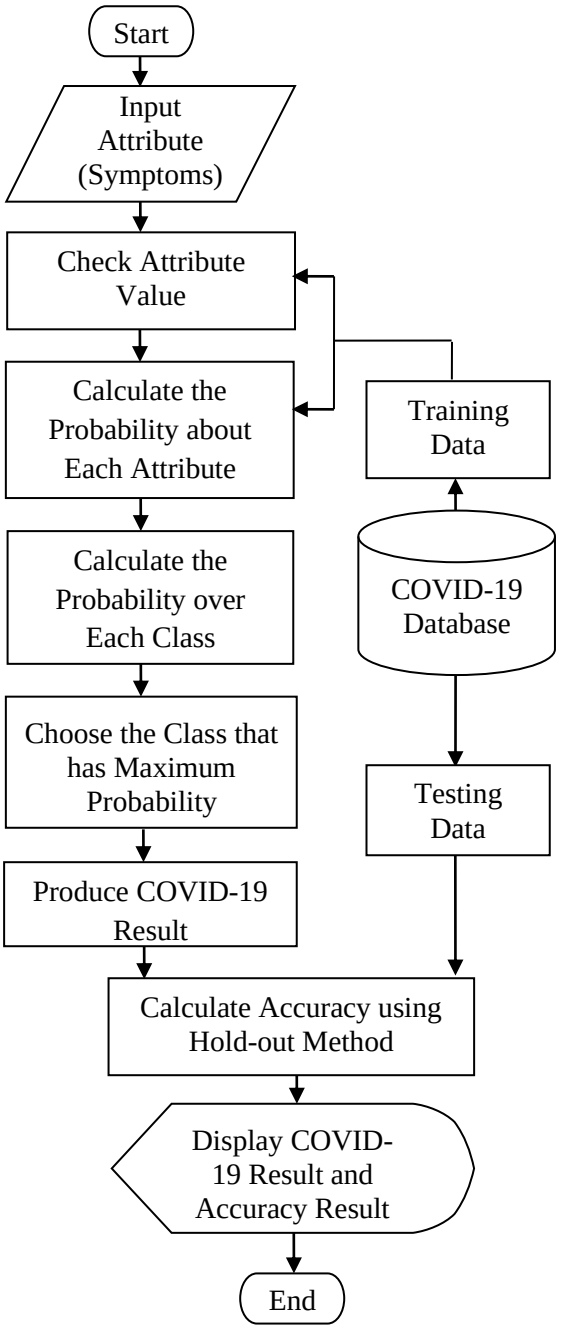


Figure 1. System Flow Diagram

System flow diagram is shown in Figure 1. In this system, the user must firstly input the attributes (symptoms) about the unknown sample that has no class label. Based on the unknown sample, this system classifies the COVID-19 disease by using improved Naive Bayesian (INB) classifier.

According to the INB, this system checks the user inputted attribute value (symptom value) that is included in the training COVID-19 data or not. If the attribute value is included in the training data, this system calculates the probability about each attribute. Then, this system continues to calculate the probability about unknown sample over each class. After calculating each probability, this system chooses the COVID-19 disease class that has maximum probability. Finally, this system calculates the accuracy by using the COVID-19 testing data. For patient, this system produces the COVID-19 disease (presence or absence) result.

4.1. COVID-19 Attribute Information

In this system, eleven attributes (symptoms) of COVID-19 disease and its class are used for classification. These COVID-19 data are obtained from the https://www.kaggle.com/datasets/meirnizri/covid19-dataset". The size of COVID dataset is "58.45 MB". COVID-19 disease attribute and its information are shown in Table 1.

Table 1. COVID-19 Attribute and Its Information

ID	Attribute (Symptom) and Its Information	
	Attribute	Information
1	Patient Type (A1)	Type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
2	Pneumonia (A2)	Whether the patient already have air sacs inflammation or not.
3	Pregnancy (A3)	Whether the patient is pregnant or not.
4	Diabetes (A4)	Whether the patient has diabetes or not.
5	Copd (A5)	Indicates whether the patient has Chronic obstructive pulmonary disease or not.
6	Asthma (A6)	Whether the patient has asthma or not.
7	Inmsupr (A7)	Whether the patient is immunosuppressed or not.
8	Hypertension (A8)	Whether the patient has hypertension or not.
9	Cardiovascular (A9)	Whether the patient has heart or blood vessels related disease.
10	Renal chronic (A10)	Whether the patient has chronic renal disease or not.
11	Other disease (A11)	Whether the patient has other disease or not.
12	Class	**Yes** mean that the patient was diagnosed with covid in different degrees. **No** means that the patient is not a carrier of covid or that the test is inconclusive.

4.2. Step by Step Classification Process

As a sample, the COVID-19 disease is tested by using ten patient records. To know the COVID-19 disease result, the user (patient) first inputs the unknown case (sample) that includes sex = male, age = 40, patient type = hospitalization, pheumonia = yes, pregnancy = no, diabetes = yes, copd = no, asthma = no, inmsupr = no, hypertension = yes, cardiovascular = no, renal chronic = yes, other disease = no. By using the training ten records from Table 2, this system checks the COVID-19 disease result. Sample training data is shown in Table 2.

Table 2. Sample COVID-19 Disease Data

ID	A1	A2	A3	A4	…	Class
1	1	Yes	No	No	…	Yes
2	2	Yes	No	Yes	…	No
3	2	No	No	Yes	…	Yes
4	1	No	No	No	…	No
5	1	No	No	Yes	…	Yes
6	2	Yes	No	No	…	Yes
7	1	No	No	No	…	Yes
8	1	Yes	No	Yes	…	Yes
9	2	No	No	Yes	…	Yes
10	1	Yes	No	No	…	No

For checking, this system uses the improved Naive Bayesian (INB) classifier. The probability results about each symptom(attribute) are shown in Table 3.

Table 3. Probability Results for Each Attribute

Attribute (Symptom) Name	Probability Result	
	Class: Yes	Class: No
patient type: hospitalization	0.428	0.333
Pheumonia: yes	0.428	0.666
Pregnancy: no	1	1
Diabetes: yes	0.571	0.333
Copd: no	1	1
Asthma: no	1	1
Inmsupr: no	0.857	1
hypertension : yes	0.571	0.666
Cardiovascular: no	1	1
renal chronic: yes	0.142	0.666
other disease : no	1	1

This system calculates the probability of each attribute that are faced from the patient. For probability calculation, this system calculates each attribute probability based on each class of the disease (Yes and No).

After calculating and producing each attribute probabilities, this system obtains the "Yes" probability that is "0.00504" and the "No" probability that is "0.00981". These "Yes" and "No" probability result values are obtained from each of user inputted disease attributes probabilities.

So, this system produces the "**COVID-19 is No (Absence)**" result to the patient according to the inputted unknown case. Because of the "No" probability is greater than the "Yes" probability.

4.3. Experimental Results of the System

This system uses the holdout method to measure the performance of the system. Hold out method splits up the dataset into a 'train' and 'test' set. The training set is what the model is trained on, and the test set is used to see how well that model performs on unseen data. Hold-out method uses the 80% of data for training and the remaining 20% of the data for testing. The precision method is as follows:

$$\text{Precision } (i, j) = n_{ij} / n_i \quad (5)$$

where, the i and j represents all the classified data and the correct classified data. The n_j and ni represents the number of i and j.

To show the performance of the system, this system is tested by using 1000 patients record about COVID-19 disease. These records are obtained from the www.kaggle.com. Moreover, this system compares the performance between the original Naive Bayesian (NB) classifier and the improved Naive Bayesian (INB) classifier. Table 4 shows the experimental result of the system. Figure 2 shows the performance comparison results between original Naive Bayesian (NB) classifier, improved Naive Bayesian (INB) classifier of the system.

Table 4. Experimental Results of the System

Number of Testing COVID-19 Disease Record	Precision (%)	
	Original Naive Bayesian Classifier	Improved Naive Bayesian Classifier
200	87%	90%
400	84%	88%
600	90%	95%
800	91%	95%
1000	89%	92%

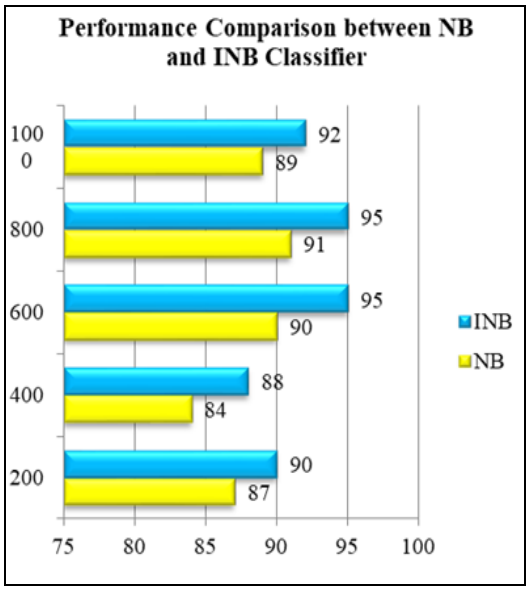


Figure 2. Performance Comparison Result of the System

This system is tested by using 200, 400, 600, 800 and 1000 COVID-19 disease records. According to the testing results, the improved Naive Bayesian classifier is especially noticeable at higher volumes of records (600, 800, and 1000), where the precision consistently increases compared to the original classifier. This indicates that the improved Naive Bayesian classifier is effective in enhancing its accuracy in predicting positive cases of COVID-19.

5. CONCLUSIONS

Classification system for COVID-19 disease is implemented by using improved Naive Bayesian classification technique. The system can classify the disease result quickly and correctly. It helps patients for testing their COVID-19 disease status. Then it also supports the junior doctors when they classify the disease problem. Besides, it can save time and money because patients don't need to go and consult the specialists. By comparing the original and improved Naive Bayesian classifier, this system points out the INB is more accurate than the NB classifier.

REFERENCES

[1] Kuyo, Mwalili and Okango (2021) "Machine Learning Approaches for Classifying the Distribution of Covid-19 Sentiments", *Scientific Research Publishing*, Vol. 11, pp 620-632.

[2] Bhatia and Malhotra (2021) "Naive Bayes Classifier for Predicting the Novel Coronavirus", *IEEE, Intelligent Communication Technologies and Virtual Mobile Networks*, pp. 880-883.

[3] Yoshikawa (2022) "Can Naive Bayes Classifier Predict Infection in a Close Contact of COVID-19? A Comparative Test for Predictability of the Predictive Model and Healthcare Workers in Japan", *Elsevier, Journal of Infection and Chemotherapy*, Vol. 28, pp 774-779.

[4] Das and Behera (2017) "A Survey on Machine Learning: Concept, Algorithms and Applications", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 5, No. 2, pp.1031-1039.

[5] Ayon (2016) "Machine Learning Algorithms: A Review", *International Journal of Computer Science and Information Technologies*, Vol. 7, No. 3, pp.1174-1179.

[6] Vembandasamy, Sasipriya and Deepa (2015) "Heart Diseases Detection using Naive Bayes Algorithm", *International Journal of Innovative Science, Engineering & Technology*, Vol. 2, No. 9, pp. 441-444.

MYANMAR TRADITIONAL FOOD CLASSIFICATION USING DEEP CONVOLUTIONAL NEURAL NETWORK

Nyo Mar Min[1], and May Mon Khaing[2]

[1] Faculty of Computer Systems and Technologies, University of Computer Studies (Taunggyi)
[2] Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City)
[1]nyomarmin@ucstgi.edu.mm

ABSTRACT

Food image categorization is a relatively new area of study, as it is gaining importance in the health and medical professions. Future applications such as calorie prediction and diet monitoring systems, will surely benefit from automated food recognition capabilities. This paper proposed Myanmar Traditional Food Classification using Deep Convolutional Neural Network. Deep learning with AlexNet framework is applied in this system to train and fine-tune on natural images of Myanmar food. The classification results efficiently support the calories estimation, contributing to a smarter and healthier lifestyle. A new dataset called Myanmar Traditional Food (MTF-dataset) has been created, containing 4500 images across 15 classes. The proposed system achieves a high accuracy of about 98%. A comparison of three optimization techniques reveals that the SGD with Momentum optimizer yields better results for the proposed system.

KEYWORD

Deep Learning, AlexNet dataset, Myanmar Traditional Food (MTF-dataset), Convolutional Neural Network (CNN), SGD

1. INTRODUCTION

Automatic food recognition is a developing research subject, not merely from the perspective of social networks. Indeed, because to its growing advantages from a medical standpoint, researchers are concentrating on this field. To combat obesity, tools for automatically identifying foods will make it simpler to calculate calories, determine food quality, and develop diet monitoring systems. A healthy diet gives people the nutrition their bodies need, serving as the cornerstone of human survival. People have gradually started to pay attention to the nutritional balance in their everyday diets as living standards have continued to rise. For food images classification and recognition, computer vision technology has been used popular and offers a new, quick, and affordable way to assess the nutritional content of food. As a result, computer vision research has steadily focused on the technology for classifying food images [1].

Since food monitoring is an important role in health-related area, it is becoming more essential in our daily lives. Today, people are dependent on smart technologies, among them deep learning is very popular to used that gives automated detection and classification and these techniques perform the better accuracy than other techniques. Deep learning has many advantages over traditional machine learning techniques, involving automatic features extraction, processing large and complex data, increase performance, solving structured and unstructured data and predict modelling etc., This paper proposed Myanmar Traditional Food Classification Using Deep Convolutional Neural Network (CNN). CNN is different unlike the traditional artificial neural networks; convolutional neural networks have the ability of assessing the process directly from image pixels [2]. A 2D convolution layer has been applied which creates a convolution kernel.

The system in this paper used own created dataset which include Myanmar Traditional Food images (4500 images). To classify the food, first food images from the dataset are entered into the system and then pre-processing is done such as resize the images. The system used 80 % of the images as the training data and the rest 20% are used for the testing data. The features extractions and classification are done using CNN(AlexNet) and then output classification result [3]. The comparison results of the optimization techniques are presented and suggest the better optimization techniques to get the high accuracy results.

2. RELATED WORK

The authors suggested a technique for classifying Asian food images in which they combined the Convolutional Block Attention Module (CBAM) with the Mobile NetV2, VGG16, and ResNet50 [4]. To achieve smoother discrimination, the authors of this paper employed a mixed data enhancement algorithm (Mix-up). The results of applying the mixed data enhancement algorithm (Mix-up) and introducing the attention mechanism (CBAM) were shown respectively through experimental comparison. The researchers classified food images using the UECFOOD100 dataset, which was produced by Yoshiyuki Kawano of the University of Electro-Communications in Japan. The experimental data set contains 14361 images that show 100 different varieties of Asian food such as rice, fried rice, curry rice, tempura rice, udon noodles, etc.

The authors of [5] suggested a novel approach to classifying Thai food dishes by employing CNN to classify and localize each dish's ingredients. MobileNet is used to identify foods, while the YOU ONLY LOOK ONCE (YOLO) network is utilized to classify ingredients and localizer. Finally, the networks are moved to the Raspberry Pi 3 platform in order to resemble a mobile phone's limited resources and processing power.

By identifying suitable regions using a variety of methods and classifying them using a range of attributes, a two-step strategy to recognize multiple-food images was proposed in [6]. To identify multiple suitable regions, the authors combined the outputs of multiple region detectors, such as the JSEG region segmentation in the initial phase, a circle detector, and Felzenszwalb's deformable part model (DPM). The authors used a feature-fusion-based food recognition method in the second step to bound the candidate regions' bounding boxes using a variety of visual features, such as Gabor texture features, histogram of oriented gradient (HoG), and bag-of features of SIFT and CSIFT with spatial pyramid (SP-BoF). A new food image dataset comprising 100 categories with bounding boxes on each food item was created by the authors. The dataset included 9060 images total, with approximately 100 images in each category. It included both of multiple and single food-item images.

In this work, Deep Convolutional Neural Network with parameter fine tuning (transfer learning) is employed for the classification of Myanmar food, aiming to achieve higher performance results. Due to the lack of a massive labelled food dataset in deep learning architecture, Myanmar foods are used to create the own Dataset.

3. SYSTEM DESIGN

The system comprises two parts, as shown in Figure 1: the training and testing phases. In the training phase, the data from the training dataset is first loaded, followed by the pre-processing step, such as resizing the image. After that, the standardization process is completed. Feature extraction is performed using CNN, and the model is trained.

In the testing phase, the testing image is first input, and then it undergoes the pre-processing step, such as resizing the image. The image can be combined with other food images such as MontHinGah, EiKyarKway, and Samusar, as shown in Figure 2.

And then, select the region of interest, as shown in Figure 3. Feature extraction is performed using CNN, followed by the classification process, which involves matching the testing image with the images from the trained model and outputting the result by labelling the image. Then, the accuracy is the primary metric used to evaluate the performance of the trained DCNN model in food classification, indicating its effectiveness in correctly identifying food

items. It is calculated using the following formula [7]:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

TP = True Positive, TN = True Negative,
FP = False Positive, FN = False Negative

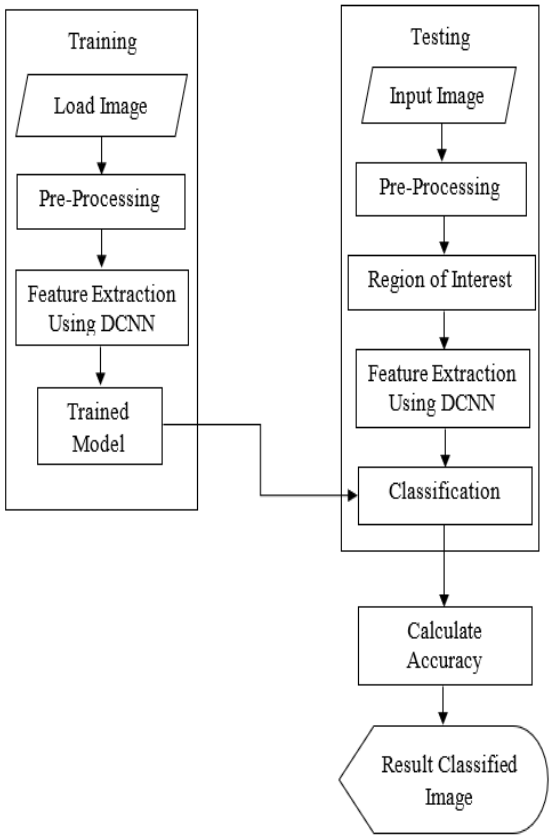


Figure 1. System Flow of the proposed system

Figure 2. Different kind of food together in an image

Figure 3. Select the Region of Interest in an image

4. DATASET

The system in this paper used fifteen classes of Myanmar food. The classes (attributes) of the dataset are as shown in Table 1. The images are collected from the smartphone camera, and some images are obtained from search engines, namely Google. All in all, a total of 4500 images are collected and saved in the dataset. The images are in various settings, with varying backgrounds and dishes. Therefore, the images are resized to 512 x 512 pixels to suit the CNN model to be used. The system used 80 % of the images as the training data, while the remaining 20% were used for testing data. Samples of the images in the dataset can be seen in Figure 4.

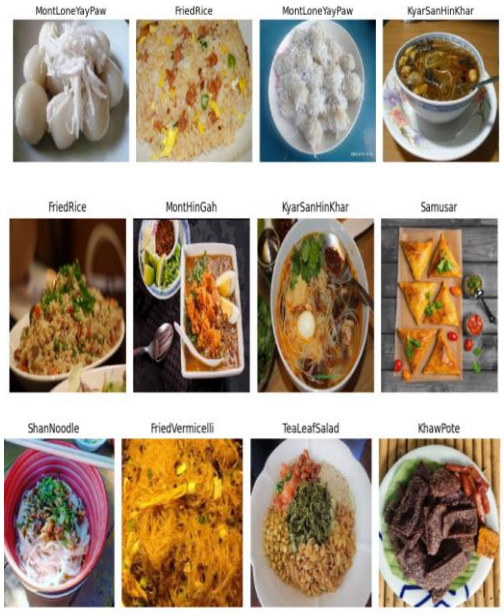


Figure 4. Some images in the training dataset

Table 1. Classes (Attributes) Name in the Dataset

Sr. No	Class Name	No. of Images in each class
1	EiKyarKway	300
2	FriedNoodle	300
3	FriedRice	300
4	FriedVermicelli	300
5	KhawPote	300
6	KyarSanHinKhar	300
7	MontHinGah	300
8	MontLoneYayPaw	300
9	RakhineMontTi	300
10	RicePancake	300
11	Samusar	300
12	ShanNoodle	300
13	TeaLeafSalad	300
14	TofuNway	300
15	TofuSalad	300

5. PRE-PROCESSING

The pre-processing step reduced the size of multiple resolution images to produce an MTF image with a specific format (512 x 512). This improvement reduced processing time while maintaining image quality. And then, the resized images are manually labelled.

6. SYSTEM IMPLEMENTATION

The primary core processes of the proposed system are as follows: data collection, pre-processing, food region detection (ROI), feature extraction and classification. Figure 5. shows the proposed framework for food image classification.

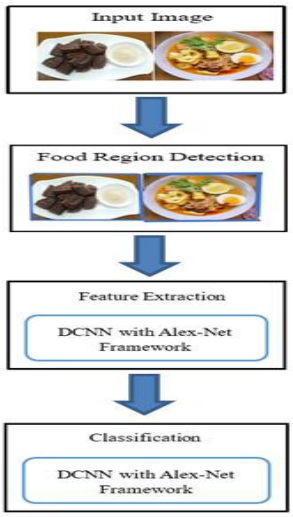


Figure 5. Block Diagram of the proposed system

6.1 Food Region Detection (ROI)

This step is pivotal for robust identification and accurate localization of food items, laying the foundation for subsequent tasks such as nutritional analysis, portion estimation and overall food understanding within the context of image-based applications. In this system, the ROI is manually performed to obtain candidate regions.

6.2 Feature Extraction

After food region detection, the input image is classified using deep learning with the AlexNet framework. The two phases of the AlexNet Deep Convolutional Neural Network are feature extraction and classification, each with a number of layers. The feature extraction phase includes five layers, while the classification phase is composed of three layers. To identify various features of the input image, it is first fed into the feature detection layer of the deep convolutional neural network. Multiple layers of convolution, activation, max pooling, and normalization are used in different sequences by the AlexNet feature extraction layer. In Convolutional Layer 1, the filter size is 11 x 11, the stride is 4, there are 96 filters, and the activation function used is Rectified Linear Unit (ReLU). Additionally, this layer includes other operations such as Local Response Normalization and Max Pooling with a filter size of 3x3 and a stride of 2. In Convolutional Layer 2, it features 5 x 5 filters, 256 in total, and employs ReLU activation along with

similar operations. Convolutional Layer 3 utilizes a 3x3 filter size with 384 filters and employs ReLU as the activation function, while Convolutional Layer 4 mirrors the specifications of Layer 3. Furthermore, Convolutional Layer 5 incorporates 3 x 3 filters, 256 filters, ReLU activation, and Max Pooling (3 x 3, Stride 2), unfolding the architecture in a step-by-step fashion for feature extraction [3]. Figure 6. shows the architecture of AlexNet framework [8].

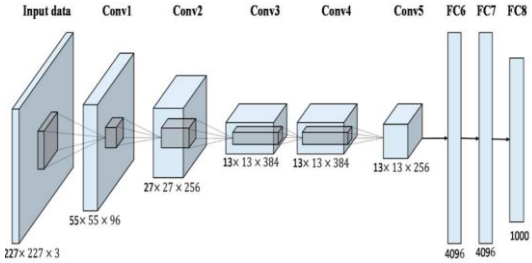


Figure 6. Architecture of AlexNet framework

6.3 Classification

In the classification stage, the feature extracted images are classified in the fully connected layers of this stage. The extracted features are fed into a multilayer perceptron (MLP) neural network with dimensions 4096 × 4096 × N during the classification phase. Afterward, the DCNN with the AlexNet framework is employed to classify food images and output the results by labelling them [3]. According to the conducted experiments, the achieved accuracy stands at 98%. The detailed experiment will be elaborated in the following section.

7. EXPERIMENTAL RESULT

This section presents a comparison of classification accuracy based on varying epochs, as detailed in Table 2. The accuracy outcomes are assessed across three distinct optimizers, namely Adam, SGD, and SGD with momentum. These optimizers are trained over different epochs (10 epochs, 30 epochs, and 50 epochs). The system demonstrates improved accuracy with an increased number of epochs. Specifically, the achieved high accuracy rates are 96% for Adam, 97% for SGD, and 98% for SGD with momentum, respectively. Particularly, among the three optimizers, SGD with momentum consistently yields superior accuracy results.

Table 2. Accuracy by different Epoch

Optimizer	Epoch (10) Accuracy	Epoch (30) Accuracy	Epoch (50) Accuracy
Adam	22%	44%	96%
SGD	25%	50%	97%
SGD with momentum	26%	52%	98%

The training accuracy vs. epochs and training/validation loss vs. epochs are show in Figure 7. Additionally, the more epochs the system uses, the less data is lost.

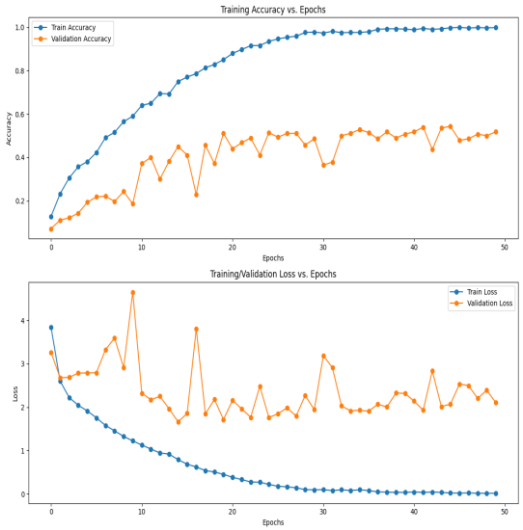


Figure 7. The accuracy and the data lost when classification using different epochs.

The classification results when matching the testing image with the training images in the dataset are shown in Figures 8, 9, and 10. In Figure 8 and Figure 9, the classification output is correct, but in Figure 10, the classification result is incorrect because the pixel value of jelly mushroom (Kywet Na Ywet Mo) is nearly similar to the color of sticky rice (Nga Chake Khaw Pote).

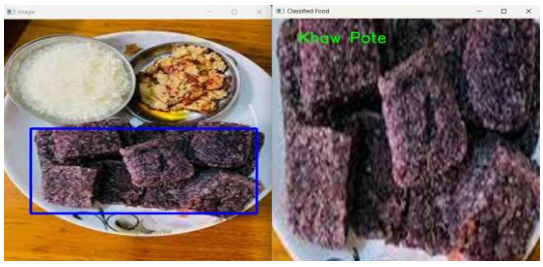


Figure 8. The input image and the classification result of the selected region of the image

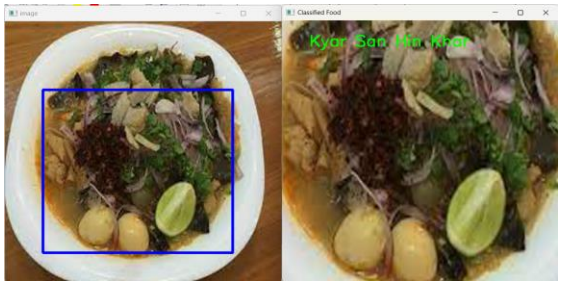


Figure 9. The input image and the classification result of the selected region of the image

Figure 10. The input image and the classification result of the selected region of the image

8. CONCLUSIONS

The system in this paper classifies Myanmar food images using deep learning algorithms (DCNN). The images are collected using a smart phone camera and some images are obtained from the internet. The images are resized during the pre-processing steps. Feature extraction and classification are performed using DCNN (AlexNet). The accuracy, calculated with different epochs, varies across different optimizers. Among them, the SGD with momentum optimizer achieves better accuracy results.

ACKNOWLEDGEMENTS

I would like to thanks to anyone who support me in any corners to do this research.

REFERENCES

[1] L.Zhou, C. Zhang and F. Liu, et al., “Application of Deep Learning in Food: A Review,” Comprehensive Reviews in Food Science and Food Safety, 2019, vol. 18, pp. 1793-1811.

[2] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” International Conference on Engineering and Technology (ICET), 2017, pp. 1–6.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton., “ImageNet classification with deep convolutional neural networks,” 2012, pp. 1097–1105.

[4] X.Bing, H. Xiaopei, Q. Zhijian, “Asian Food Image Classification Based on Deep Learning,” IEEE Journal of Computer and Communications, 2021.

[5] K. Sukvichai, P.Maolanon, K.Sawanyawat, and W.Muknumporn, “Food Categories Classification and Ingredients Estimation Using CNNs on Raspberry Pi 3”, International Conference of Information and Communication Technology for Embedded Systems ,2019.

[6] Y. Matsuda, H. Hoashi and K. Yanai, "Recognition of Multiple-Food Images by Detecting Candidate Regions," 2012 IEEE International Conference on Multimedia and Expo, Melbourne, VIC, Australia, 2012, pp. 25-30, doi: 10.1109/ICME.2012.157.

[7] https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

[8] X.Han, Y. Zhong, L. Cao, and L. Zhang,"Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification”,2017,https://doi.org/10.3390/rs9080848.

Authors Biography

Nyo Mar Min obtained a B.C. Tech (Hons) degree from the University of Computer Studies (Taunggyi) in 2007 and an M.C. Tech degree from the University of Computer Studies Mandalay (UCSM) in 2010, respectively. She continues her studies for a Ph.D. in IT at the University of Technology (Yatanarpon Cyber City). Now, she is a Lecturer at the Faculty of Computer Systems and Technologies at the University of Computer Studies (Taunggyi), Myanmar.

Dr. May Mon Khaing got her Bachelor of Engineering from Technological University (Monywa) in 2006 and Master of Engineering from West Yangon Technological University, Yangon in 2010. She continues her studies for Doctor of

Philosophy (IT) in University of Technology (Yatanarpon Cyber City), Myanmar and she got it finally in 2014. Now she has been working as a Professor in Faculty of ICT in UTYCC. Her specialization is Data Communication and Networking, Artificial Intelligence and Computer Architecture. She is doing research in this fields and she published 5 papers till now.

MACHINE LEARNING BASED HAND CAPTURE RECOGNITION

Thin Zar Aung[1], Nan Saw Kalayar[2], and Moe Phyu Phyu Khaing[3]

[1,2] Faculty of Information Science, University of Computer Studies (Taunggyi)
[1]*thinzaraung@ucstgi.edu.mm*,[2] *nansawkalayar@ucstgi.edu.mm*,
[3]*moephyuphyukhaing@ucstgi.edu.mm*

ABSTRACT

Sign language is the most significant way of communication for those who are deaf or hard of hearing. Building a bridge between the two people who wish to communicate thus becomes essential. Many algorithms have been developed recently to assist those who are not familiar with sign language, but there aren't many that work well. In a time where human-computer interaction is always changing, it is more important than ever to create methods for input commands that are both efficient and intuitive. At the forefront of these advancements is hand gesture recognition, which provides a flexible and natural method of user-to-device communication. The idea and specifications for a hand gesture recognition system that is intended to identify and decipher hand gestures that represent the numbers 0 through 9 are presented in this paper. In order to provide a more accurate performance analysis, the research employs a Histogram of Gradients for feature extraction and Support Vector Machines and K-Nearest Neighbours for the classification stage.

KEYWORDS

Histogram of Gradients, K-Nearest Neighbours Algorithm, Support Vector Machine, Hand Capture, Machine Learning

1. INTRODUCTION

Individuals who are deaf or hard of hearing can communicate using sign language. In order to successfully transmit the speaker's words and thoughts, that method of communication combines the simultaneous use of face expressions, hand orientation and movement, finger spells, body language, head motions, and eye gazes. The subject of hand gesture recognition [1] is a fascinating one with numerous applications in various domains, including computer games, robotics, human-computer interaction, automatic sign-language interpretation, and so forth. There are a few methods for recognising hand gestures in the literature that solely rely on the analysis of pictures and videos [1], [2]. However, complex movements and inter-occlusions that define hand gestures are often too intricate to be captured by a bi-dimensional representation.

The creation of natural and intuitive interfaces has become essential in the rapidly changing field of human-computer interaction. A revolutionary step in this regard is the Hand Gesture Recognition System, which provides a fresh method of user-device communication. This system seeks to reinvent how people interact with technology by deciphering and comprehending the complex language of hand gestures, making it more approachable, interesting, and responsive. The goal of the Hand Gesture Recognition System is to close the communication gap between responsiveness of machines and human expression. Its main goal is to give consumers a wide range of hand gestures that they can use to effortlessly interact with electronic gadgets. Whether used for virtual reality, gaming, or hands-free device control, this technology offers a creative and intuitive way to interact.

The system's scope includes a wide range of applications, from gaming and entertainment to everyday practical use. The Hand Gesture Recognition System aims to precisely understand hand movements by utilizing computer vision and machine learning technologies. This enables users to interact with digital information more naturally and

immersive, as well as to transmit commands and manipulate virtual objects. The Hand Gesture Recognition System integrates complex software algorithms with hardware elements like cameras or sensors. Hand movements are recorded by the hardware and converted into digital data, which is then processed by the software to recognize particular gestures. Subsequently, the system converts these identified motions into significant instructions, enabling a smooth and engaging user interface.

The following part is how this paper is structured: The suggested gesture recognition system's relevant past works are discussed in Section II. The suggested system design and the methods for extracting and identifying the hand region are described in Section III. The accuracy analysis and experimental findings are presented in Section IV. Section V presents the study work's conclusion in its final form.

2. RELATED WORKS

The results of the literature review offer insight into the many approaches that can be used to accomplish hand gesture recognition. It also aids in comprehending the benefits and drawbacks of the different approaches. The camera module and the detection module are the two primary sections of the literature review. The many cameras and markers that are available for use are identified by the camera module. The pre-processing of images and feature extraction is handled by the detection module.

Data gloves, hand belts, and cameras are the most often utilized techniques for gathering input from the user that has been witnessed. Data gloves are used in the gesture recognition methods [1] and [2] to extract input. To read hand movements, a hand belt equipped with an accelerometer, gyroscope, and Bluetooth was employed [3]. In [4], the authors employed a Bumblebee2 stereo camera along with an inventive Senz3D camera to record colour and depth data. By [5], a monocular camera was employed. Simple web cameras have been used by cost-effective models to implement their systems [6]. A Kinect depth RGB camera was utilized in [7] to record color streams for the approaches. In addition to standard photos, depth cameras also record additional depth information for every pixel (depth images) at a frame rate [8]. Using the color space, the majority of technologies enable the robust extraction of a hand region. The background issue is not entirely resolved by them. This background issue was fixed in [9] by employing a monochromatic glove—a pattern of augmented reality markers in black and white. Even while built-in cameras don't provide depth information, they use less computational power. Therefore, the webcam that comes with the laptop is employed in this suggested model instead of the need for any extra cameras or hand markers like gloves.

Pre-processing of the image has been done using a variety of techniques and algorithms, such as edge identification, noise reduction, and smoothing, which are followed by various segmentation approaches for boundary extraction, or separating the foreground from the background. To get rid of noise, the authors of [10] employed a morphology algorithm that carried out image dilation and erosion. Following binarization, the contours were smoothed using a Gaussian filter [11]. By using the SAD (Sum of Absolute Differences) technique to compare the left and right pictures, a depth map was created in order to carry out segmentation. The contours in [12] were found using the Theo Pavildis Algorithm, which only visits the boundary pixels. The computational costs are reduced by using this approach. In [6], the hand palm's largest contour was selected, and it was then made simpler using polygonal approximation. The process of classifying involves grouping distinct items according to how similar they are to one another. The method recognizes 25 hand postures using a classifier based on Euclidean distance [13]. In [14], the Support Vector Machine (SVM) classifier was employed. We depart from other conventional techniques by recognizing gestures without the use of gloves or other hand markers. Our model made the system more affordable by using the laptop's built-in webcam instead of buying any additional cameras.

3. PROPOSED SYSTEM

In this section, the detail step by step processes of the proposed system are presented. As the pre-processing step, resizing the image, colour

correction and segmentation on the image are performed. The next of the proposed system is feature extraction, in which HOG is the main part of this step. In the classification of the system, KNN and SVM are used to get the better performance. Figure 1 shows the system design of the proposed system.

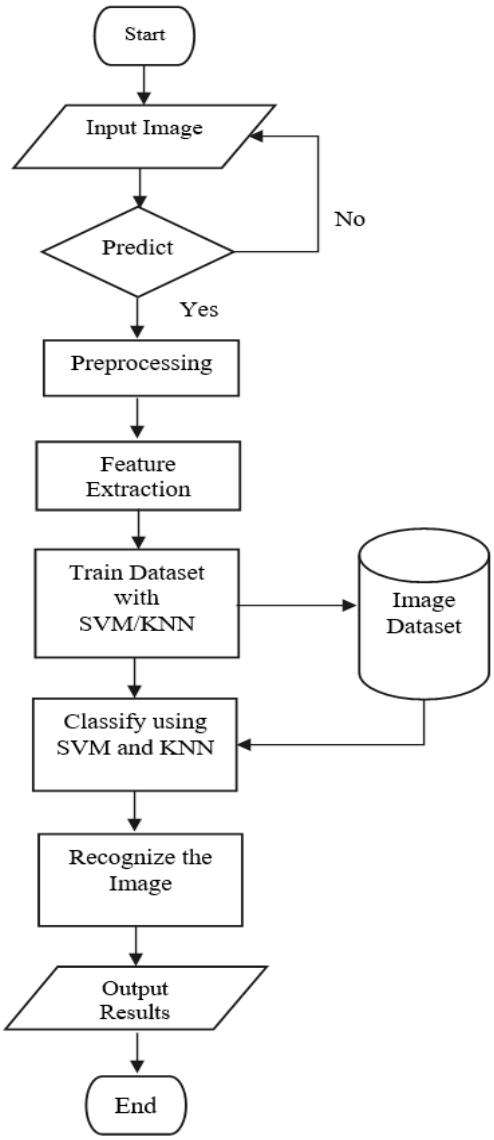


Figure 1. Proposed system design

3.1. Histogram of Gradients

When a single feature description is computed for an image or image-patch, it is referred to as a global feature representation, or histogram of oriented gradients, or HoG. A vector with numerous histograms is the description. Every histogram tracks the frequency of gradient directions within a certain local area of the image. In computer vision and image processing, the Histogram of Oriented Gradients, or HOG, is a feature extraction method that is frequently employed. It has numerous uses in picture recognition and object detection.

In many computers vision tasks, the objective is to represent complicated visual input in a more compact and comprehensible form; in these situations, feature extraction plays a critical role. This is accomplished using HOG by concentrating on the gradient orientation distribution inside an image. HOG's main goal is to capture the orientations and local intensity gradients, which are crucial for describing the forms and structures of objects. Pre-processing is done on the input image to increase its resistance to noise and changes in lighting. Normalizing pixel intensities, contrast normalization, and grayscale image conversion are common pre-processing techniques.

The gradient orientations and magnitudes of picture pixels are computed via HOG. This stage facilitates the identification of texture borders and edges. A histogram of gradient orientations is calculated for every cell. The distribution of gradient orientations within the cell is represented by the histogram, which is based on the orientations being quantized into bins. Every block has normalization applied to it to make the algorithm more resilient to variations in contrast and lighting. The final HOG description for the image is created by concatenating the normalized histograms from each block. The orientations and geographical distribution of gradients are captured by this descriptor. The HOG picture process of this suggested system is displayed in Figure 2.

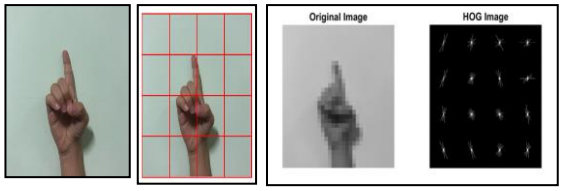


Figure 2. HOG of Proposed System

3.2. K-Nearest Neighbours Algorithm

Using the K-Nearest Neighbours (KNN) technique, machine learning tasks involving regression and classification can be effectively

and intuitively handled. KNN uses its K nearest neighbours in the training dataset to predict the label or value of a new data point by leveraging the similarity notion. This article will introduce the k-Nearest Neighbours technique, also known as the supervised learning algorithm (KNN), emphasizing its user-friendliness. KNN utilizes the Euclidean Distance technology. Every instance in the dataset can be represented as a point in N-dimensional space, according to the underlying theory. Additionally, KNN employs a parameter K to indicate the number of instances that must pass before classifying a new instance, with the majority class being selected as the winner.

Because of its simplicity and ease of use, the (K-NN) method is a popular and adaptable machine learning technique. No presumptions on the distribution of the underlying data are necessary. It is a versatile option for a range of dataset types in classification and regression applications because it can handle both numerical and categorical data. This non-parametric technique bases its predictions on how similar the data points in a particular dataset are to one another. By comparison, K-NN is less susceptible to outliers than other algorithms.

The K-NN algorithm locates the K closest neighbors, using a distance metric like Euclidean distance, to a given data point. The majority vote or the average of the K neighbors are then used to establish the class or value of the data item. Using this method enables the algorithm to anticipate outcomes based on the local structure of the data and adjust to various patterns.

3.2.1. Fine KNN

Fine KNN likely refers to a K-Nearest Neighbors (KNN) algorithm that has been carefully optimized or fine-tuned for performance on a specific task or dataset. Fine KNN optimization process might involve tuning hyper parameters such as the number of neighbors (K), the distance metric, or other algorithm-specific parameters to improve the algorithm's performance.

When constructing a Fine KNN model, practitioners typically use techniques such as grid search, random search, or Bayesian optimization to explore the hyperparameter space and find the combination of parameters that yields the best performance on a given dataset. Additionally, cross-validation is often employed to assess the model's generalization performance and prevent overfitting during the tuning process.

3.2.2. Medium KNN

Medium KNN could refer to the choice of the number of neighbors (K), the distance metric, or other parameters that influence the behavior of the algorithm. For example, in the context of choosing the number of neighbors (K), Medium KNN might mean using a moderate value for K, neither too small nor too large, to balance the bias-variance trade-off and achieve reasonable performance.

However, without more context or specification, it's challenging to precisely determine what "Medium K-Nearest Neighbors" refers to. It's possible that it could be a specific configuration or variant of the KNN algorithm used in a particular context or by a specific practitioner.

3.2.3. Cubic KNN

Cubic KNN doesn't correspond to a standard machine learning algorithm, but it could be a term used to describe a variant or adaptation of the traditional K-Nearest Neighbors (KNN) algorithm. KNN is a non-parametric, instance-based learning algorithm used for classification and regression tasks. It works by finding the K nearest neighbors of a given data point in the feature space and making predictions based on the labels or values of those neighbors.

It's possible that "Cubic K-Nearest Neighbors" refers to a variant of KNN where a cubic transformation is applied to the input features before computing distances between data points. This transformation could involve raising each feature to the power of 3, effectively creating a cubic feature space. However, this interpretation would significantly change the nature of the algorithm and is not a standard approach. Another interpretation is that Cubic KNN involves using a distance metric that

emphasizes cubic relationships between data points. This could mean modifying the Euclidean distance metric used in traditional KNN to incorporate cubic terms, potentially giving more weight to features that exhibit cubic relationships.

3.3. Support Vector Machine

Strong machine learning algorithms like Support Vector Machine (SVM) are utilized for tasks including regression, outlier identification, and linear or nonlinear classification. Text classification, picture classification, handwriting recognition, spam detection, face detection, gene expression analysis, and anomaly detection are just a few of the many applications for SVMs. Because SVMs can handle high-dimensional data and nonlinear relationships, they are versatile and effective in a wide range of applications.

A supervised machine learning approach called Support Vector Machine (SVM) is used for regression as well as classification. The SVM algorithm's primary goal is to locate the best hyperplane in an N-dimensional space that may be used to divide data points into various feature space classes. The hyperplane attempts to maintain the largest possible buffer between the nearest points of various classes. The number of features determines the hyperplane's dimension. The hyperplane is essentially a line if there are just two input features. The hyperplane transforms into a 2-D plane if there are three input features. If there are more than three features, it gets hard to imagine. The SVM classifiers—Linear, Quadratic, Cubic, and Gaussian—transform the data using a kernel trick technique, and then, using the results of these transformations, determine the best border between the possible outputs.

3.3.1. Quadratic SVM

A Quadratic Support Vector Machine (QSVM) is a variant of the Support Vector Machine (SVM) algorithm used for classification tasks. SVMs are powerful supervised learning models used for both classification and regression tasks. They work by finding the optimal hyperplane that best separates classes in a high-dimensional space.

The standard SVM algorithm seeks to find a linear hyperplane that maximizes the margin between different classes of data points. However, in some cases, the data may not be linearly separable, meaning that a linear boundary cannot effectively separate the classes.

To address this limitation, the QSVM extends the standard SVM algorithm to allow for non-linear decision boundaries by using a quadratic kernel function. This kernel function maps the input features into a higher-dimensional space, where the data may become separable by a hyperplane. In other words, the QSVM implicitly performs feature space transformation to project the data into a higher-dimensional space where linear separation might be feasible.

3.3.2. Cubic SVM

A Cubic Support Vector Machine (Cubic SVM) is an extension of the traditional Support Vector Machine (SVM) algorithm that incorporates cubic kernel functions. Similar to quadratic SVMs, which use quadratic kernel functions, cubic SVMs aim to handle non-linearly separable data by mapping it into a higher-dimensional space where linear separation may be feasible.

In general, the kernel function in SVM defines the similarity between pairs of data points in the input space. By employing a cubic kernel function, the SVM can effectively model more complex decision boundaries compared to linear SVMs.

Cubic SVMs are beneficial for handling data with complex non-linear relationships between features. However, like other kernel-based SVMs, they may suffer from increased computational complexity and are susceptible to overfitting, particularly when dealing with high-dimensional data or datasets with a large number of samples. Regularization techniques and careful selection of hyper parameters are essential for obtaining well-performing models.

3.3.3. Fine Gaussian SVM

A "Fine Gaussian Support Vector Machine" likely refers to a Support Vector Machine (SVM) that uses a Gaussian (also known as

Radial Basis Function, RBF) kernel and has been finely tuned or optimized for performance on a specific task or dataset.

The Gaussian kernel is a popular choice in SVMs for capturing complex, nonlinear relationships between data points. It maps the input data into a higher-dimensional space using a Gaussian function, allowing the SVM to find nonlinear decision boundaries. "Fine" in this context likely means that the SVM has been carefully optimized or fine-tuned. This optimization process may involve tuning hyper parameters such as the regularization parameter (C) and the kernel bandwidth parameter (σ) to achieve the best possible performance on a given dataset. Fine-tuning ensures that the SVM generalizes well to unseen data and maximizes its predictive accuracy.

When constructing a Fine Gaussian SVM, practitioners typically employ techniques such as grid search, random search, or Bayesian optimization to explore the hyper parameter space and find the combination of parameters that yields the highest performance. Additionally, cross-validation is often used to assess the model's performance and prevent overfitting during the tuning process.

4. Experimental Result

In this section, the experimental results of the proposed system are described. The "Kernel Trick" is a method used in Support Vector Machines (SVMs) to convert data (that is not linearly separable) into a higher-dimensional feature space where it may be linearly separated. This technique enables the SVM to identify a hyperplane that separates the data with the maximum margin, even when the data is not linearly separable in its original space. The kernel functions are used to compute the inner product between pairs of points in the transformed feature space without explicitly computing the transformation itself. This makes it computationally efficient to deal with high dimensional feature spaces. For the experimental result evaluation, 1500 images are collected as the dataset. 75% of collected data are used as training dataset and 25% are used as testing dataset. The average accuracy of the proposed system on the system's dataset are shown in the table1.

Table 1. Accuracy Results of Proposed System

Classifiers	Average
Quadratic SVM	99.7%
Cubic SVM	99.4%
Fine SVM	98.8%
Fine Gaussian SVM	87.4%
Cubic KNN	83.5%
Medium KNN	82.3%

According to the analysis, Quadratic SVM achieve the best accuracy result of proposed system. In this system, simplifies the process of loading images, performing gesture recognition, and assessing accuracy. And then incorporating HOG (Histogram of Oriented Gradients) feature extraction enhances the system's accuracy and robustness by effectively. SVM (Support Vector Machine) classifiers to achieve high accuracy in recognizing the gestures of the proposed system.

5. Conclusions

A notable advancement in computer vision and pattern recognition is the creation of the hand gesture detection system for Myanmar number identification, which makes use of KNN classifiers, support vector machines, and HOG feature extraction. This suggested solution demonstrates how well HOG features capture crucial visual data from hand motions, allowing precise identification of the best model to select for particular deployment scenarios while maintaining accuracy and computational efficiency. This technology advances the field of gesture recognition by providing a solid platform for future research and application across a range of industries.

Acknowledgements

I would like to thank a lot to all teachers and my friends who give a valuable suggestion to me.

Authors

Biography

Thin Zar Aung received the B.C.Sc (Hons) degree from University of Computer Studies, Taunggyi in 2003, and with a master's degree in computer science from University of Computer Studies Mandalay (UCSM) in 2008. She currently works as a lecturer at the Faculty of Information Science, University of Computer Studies (Taunggyi), Myanmar. Machine Learning, Computer vision, and operation research are her main areas of interest.

Nan Saw Kalayar received the Ph. D (IT) from University of Computer Studies (Yangon), Myanmar. Now she is working as the Professor, Faculty of Information Science, University of Computer Studies (Taunggyi). Her research interest fields are Data Science, Machine Learning, Deep Learning, Image Processing.

Moe Phyu Phyu Khaing received the B.C.Sc in Computer Science from University of Computer Studies (Monywa) in 2018 and M.C.Sc from University of Computer Studies, Mandalay (UCSM) in 2022. She is a tutor at University of Computer Studies.

REFERENCES

[1] G. Luzhnica and et al., "A Sliding Window Approach to Natural Hand Gesture Recognition using a Custom Data Glove", IEEE, New York, Mar 19, 2016, pp.81-90.

[2] J. H. Kim and et al., "3-D Hand Motion Tracking and Gesture Recognition Using a Data Glove", IEEE, New York, July 5, 2009, pp.1013-1018.

[3] C. Hung and et al. "Home outlet and LED array lamp controlled by a smartphone with a hand gesture recognition", ICCE, IEEE, New York, Jan 7, 2016, pp.5-6.

[4] Q. Wang and et al., "A real time hand gesture recognition approach based on motion features of feature points", CSE, 17th IEEE, New York, Dec 19, 2014, pp.1096-1102.

[5] J. Dulayatrakul and et al., "Robust implementation of hand gesture recognition for remote human-machine interaction", 7th ICITEE, Oct 29, 2015, pp. 247-252.

[6] T. H. Tai and et al, "Embedded virtual mouse system by using hand gesture recognition", ICCE-TW, IEEE, Taiwan, Jun 6, 2015, pp. 352-353.

[7] Y. Chen and et al., "A real-time dynamic hand gesture recognition system using kinect sensor", ROBIO, IEEE, New York, Dec 6, 2015, pp. 2026-2030.

[8] W. L. Chen and et al., "Depth-based hand gesture recognition using hand movements and defects", ISNE, Taiwan, May 4, 2015, p. 1-4.

[9] H. Ishiyama and et al., "A robust real-time hand gesture recognition method by using a fabric glove with design of structured markers", Greenville, New York, IEEE, Mar 19, 2016, pp. 187-188.

[10] R. Suriya and et al., "A survey on hand gesture recognition for simple mouse control", ICICES, India, IEEE, Feb 27, 2014, pp. 1-5.

[11] K. Chanda and et al., "A new hand gesture recognition scheme for similarity measurement in a vision based barehanded approach", 3rd ICIIP, New York, IEEE, 2015, pp. 17-22.

[12] D. H. Lee and et al., "A Hand gesture recognition system based on difference image entropy", 6th IEEE, Seoul, Nov 30, pp. 410-413.

[13] G. Luzhnica and et al., "A sliding window approach to natural hand gesture recognition using a custom data glove", Greenville, New York, IEEE, Mar 19, 2016, pp. 81-90.

[14] Y. Chen and et al., "Rapid recognition of dynamic hand gestures using leap motion", New York, IEEE, Aug 8, 2015, pp. 1419-1424.

DESIGN AND IMPLEMENTATION OF SHOPPING SYSTEM USING MVC SOFTWARE ARCHITECTURE

Myintzu Phyo Aung[1], Zin Mar Oo[2], and Theint Theint Htwe[3]

[1,2,3]Faculty of Information Science, University of Computer Studies (Mandalay)
[1]myitzu.mm@gmail.com, [2]zmo.zinmaroo@gmail.com,
[3]theinttheint1784@gmail.com

ABSTRACT

Development of web applications by using traditional methods can make large limitations. Such limitations may include a lot of time consuming and can make number of unexpected errors. Because in these traditional methods, all of the codes are written in a single file. Most of the languages have their inbuilt tags which are used to process the request and fetch data according to the request. As a result, a novel approach of technology, MVC framework, is built and used to deal with such issues. This paper presents a design and implementation for web application in MVC framework. As a case study, web-based E-commerce application is presented. There will be an admin panel and a user panel. As an admin, who can maintain the product availability, adding new products and viewing sale reports. The system provides user friendly, safe and time saving e-commerce application.

KEYWORDS

MVC Architecture, Web Application, Servlet, JSP

1. INTRODUCTION

Nowadays, with the prevalence of cloud computing, the demand for modular and scalable Web application development technologies is urgent. Dynamic contents and ubiquitous user interactions make Web applications increasingly complicated. A majority of current web applications leverage a Model-View-Controller (MVC) architectural style. Since the MVC triad does not provide feature-based modularization, Web applications in pure MVC style are experiencing scalability and maintainability issues [1]. **Model View Controller** (MVC) design pattern specifies that an application consists of a data model, presentation information, and control information. The pattern requires that each of these parts must be separated into different objects. MVC is not only an architectural pattern, but also a complete application [2]. MVC design patterns are well-known patterns because which can be used for interactive software system architectures. The main objective of MVC framework is to separate the main components such as data manipulation (model), display/interface (view) and the process (controller) so that it is neater, structure and easily developed. [3]

2. RELATED WORK

MVC architecture: detail insight to the modern web application development is presented in [4]. In their work, key insights of layers, advantages, disadvantages and uses of MVC architecture were presented. Related technologies can be used in conjunction with MVC based web applications were also presented.

Design and implementation of web application based on MVC Laravel architecture was presented in [5]. The authors presented that web application development using traditional methods has many issues. Therefore, MVC architecture, a novel approach, is used by many

software developments companies to solve such issues and more efficient implementation can be resulted because of standardized and scalable development.

3. MVC ARCHITECTURE

A software architecture is a description of subsystems and components of a software system and the relationships between them.

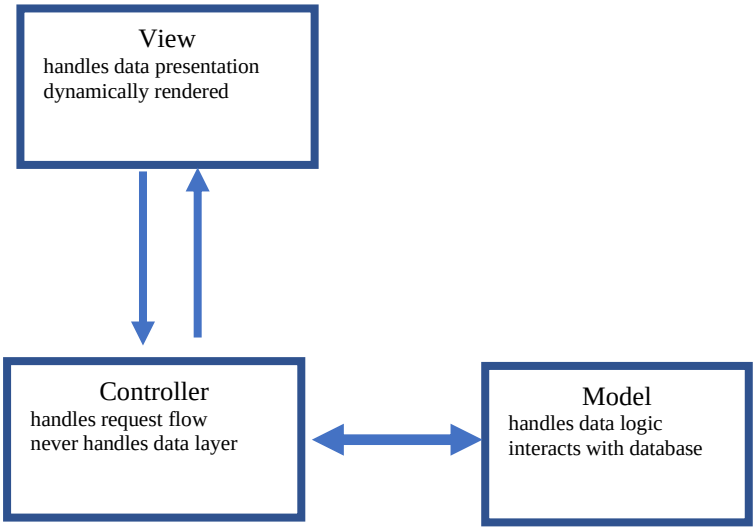


Figure 1. Model View Controller (MVC) Architecture, Functionalities of Each Part

To achieve standard software systems, an architectural pattern can be used as a means of representing, sharing and reusing knowledge of designs. The Model-View-Controller pattern nowadays has widely been used by companies and developers [6]. The MVC architecture is shown in **Figure 1.**

There are many features of MVC framework. First of all, the architecture can provide a clear separation of business logic, User Interface logic and input logic. Because of full control over HTML and URLs and also power URL mapping component, it is easy to design web application architecture that have comprehensible and searchable URLs. It can also support Test Driven Development (TDD) [2].

The Model-View-Controller (MVC), an architectural pattern, separates an application into three main logical components: The Model, the View, and the Controller. For handling specific development aspects of an application, all of these components must be built. Because of its features and advantages of MVC, MVC become the most frequently used industry-standard web development frameworks to create scalable and extensible projects.

3.1. Model

Model is an important layer of MVC application. It manages the information in the form of data which is used to represent the output with the help of views. It depicts the database records. It manages the application data. It basically contains the application data, logic definition, function specification, business rule involvement. A model possibly is a single object or it is some composition of objects. This layer can manage the data and also allow the database communication i.e. insert, retrieve and update data in the database.

3.2. View

Views are used to prepare the interface of the application for the interaction of users and application [7]. A view has the responsibility to display all data of the

model. It only shows the required attributes and hides the unnecessary attributes.

3.3. Controller

The responsibility of controller is to interpret the inputs from the user, and then give commands to the model and/or the view to change appropriately [8]. But it is needed to ensure that the separation of concerns by keeping the view and model independent of each other [9].

3.4. Advantages of MVC

MVC separates the Data Layer and the Expression Layer, and divides the Application objects into three classes: Model class, View class and Control class. For the Business and Data Logic, Model class must be implemented. View class must be implemented for the Display Logic, and Control class is used to implement the Control Processing. MVC holds the three logic types independency to build the Business and Data Logic oriented business and design the control and display logic-oriented application. As the result, the changing of business processing does not modify the business and data logic. Furthermore, the changing of business principles and algorithms only modifies Model class and not needed to change other parts. So MVC separates the data accessing.

4. PROPOSED SYSTEM

4.1. System Flow Diagram

Figure 2 presents the System Flow Diagram of the Proposed E commerce system. There are two roles of users, Staff and Admin.

Admin user can add or update brand, category, product and can view reports. The Staff can search, view and sale products.

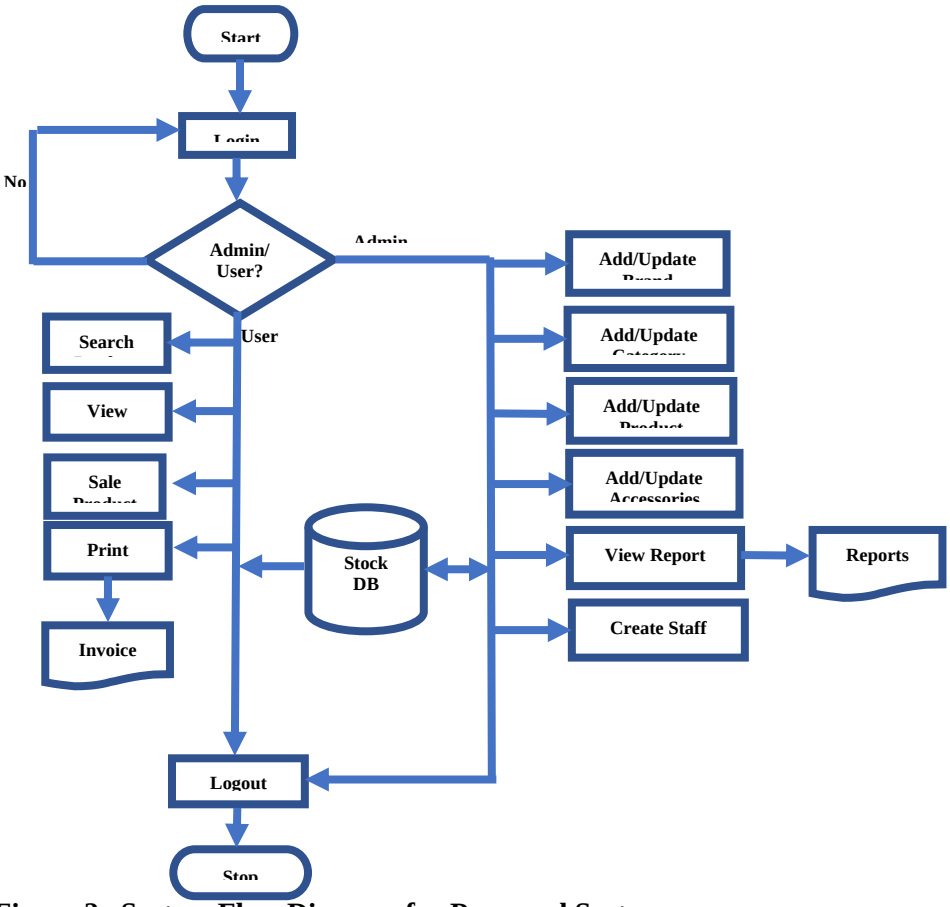


Figure 2. System Flow Diagram for Proposed System

Before the development of the application is started, the type of architecture to be used and data model design should be decided. One approach to do is to examine the functional components of the application and the way through which they will communicate each other. The Model-View-Controller or MVC software architecture pattern, which is currently widely used in web design architecture will be used for the proposed system [10].

4.2. Designing Data Model

The conceptual model of storage can be defined by entities and their relationships. The entity relationship (ER) model is a structural design of the database. It is a framework which is created with specialized symbols for the purpose of defining the relationship between the database entities. ER diagram is created based on three components: entities, attributes, and relationships.

The ER model of the proposed system includes: The Admin entity keeps the administrator information with the administrator's name and password; Product entity holds the information about the product; InstockRecord entity stores the quantity of each product by the date it was added; and SalesRecord entity records every transaction made at the point of sale. This entity is linked to relationships like Product, connecting customers with their bought products, and InventoryRecord, monitoring the availability and movement of products within the inventory. Entity Relationship Model for proposed system is depicted in Figure 3.

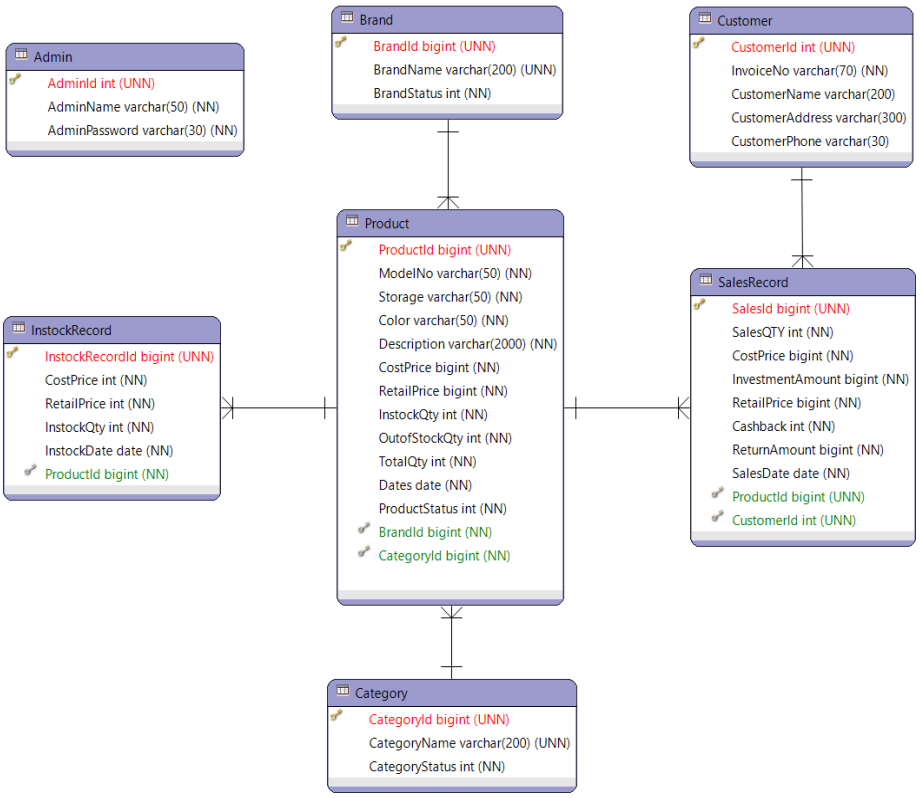


Figure 3. SQL Schema

4.3. CRUD Matrix

CRUD stands for Create, Refer, Update and Delete which are four basic functions of persistent storage and used for bottom-up data modelling. It can be used to analyse when and what events happens to each entity and its attributes. A CRUD Matrix is a useful tool for identifying the Data Items which are involved in the Data Access for each application. SQL consists of only 4 statements, sometimes referred to as CRUD:

Create-INSERT is to store new data; Read-SELECT is for retrieve data; Update and Delete are to change or modify data and delete respectively. CRUD Matrix for proposed system is described in Table 1.

In this system, there are two roles of user: User and Administrator. As an Administrator, who can Create, Refer, Update and Delete permissions are granted on Admin, Product, Brand and Category Tables and only Refer permission is granted on InstockRecord, SalesRecord and Customer Tables. As a user who is granted only Refer operation on Product, InstockRecord, SalesRecord tables and nothing is granted on Admin, Brand, Category and Customer tables.

Table 1. CRUD Matrix for Proposed System

Entity	User	Administrator
Admin	-	C,R,U,D
Product	R	C,R,U,D
Brand	-	C,R,U,D
Category	-	C,R,U,D
InstockRecord	R	R
SalesRecord	R	R
Customer	-	R

4.4. Development of Application

In this paper, the proposed system is developed using MVC architecture in Java technologies, composed of three tiers Figure4.

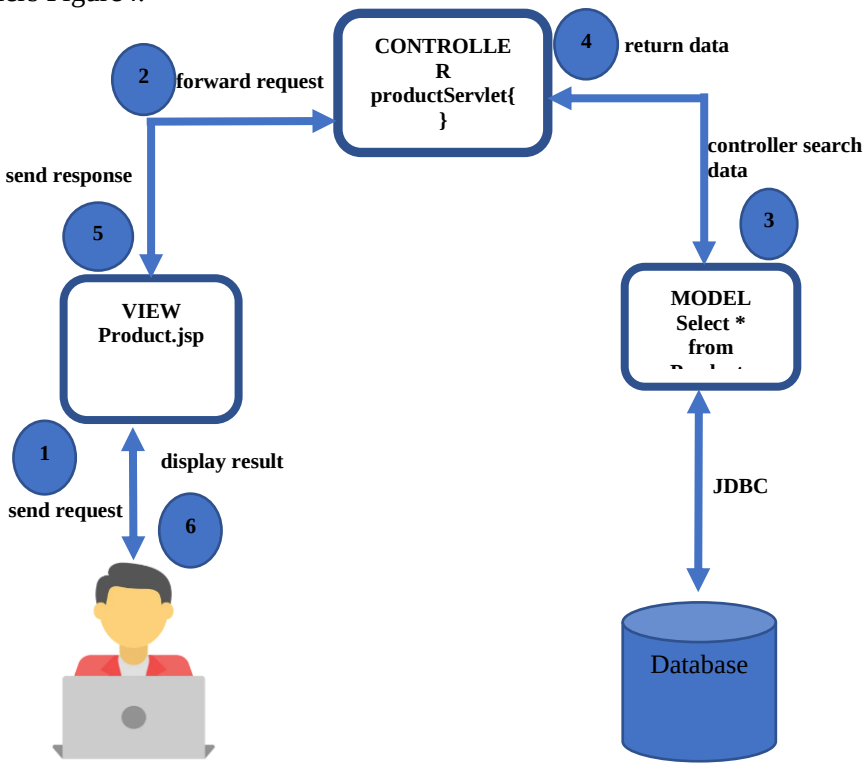


Figure 4. MVC Components

The first tier is input and output on the client machine, such as a web browser. Business logic and view control belong to the second tier. Servlets, Java Server Pages (JSPs), and JavaBean are the components of this tier. The database system belongs to the third tier. A Servlet is a Java plug-in for a web server. If the user gives the **request** from a browser to a **Servlet,** then the Servlet (Controller) will get the request parameters from the browser and pass them to the **JavaBean (Model)** class. JavaBean goes into a database and performs any database operations. After that, the data will be read. The JavaBean class will give the result to the Servlet in object format. The Servlet goes to View pages in JSP, which are

defined in presentation logic. JSP (View) gives the response to the browser page [11].

4.4.1 Implementation Technologies

Model

For Model part of the system MySQL database, JDBC and JavaBeans are used for data storage, Database connectivity and business logic of data.

MySQL

MySQL is a very popular open-source relational database management system (RDBMS) which is ideal for both small and large applications because it is very fast, reliable, scalable, and easy to use [12].

JDBC

To access data from Java Programming Language, Java Database Connectivity (JDBC) API provides that feature. Using the JDBC API, data sources can be accessed virtually from relational databases to spreadsheets and flat files. To use the JDBC API with a particular database management system, JDBC technology-based driver works to mediate between JDBC technology and the database [13]. For the implementation of this system, the data model is described in ER model and CRUD Matrix, Figure 3 and Table 1 respectively.

Java Beans

Java Beans or EJBs is used for Model part of the system which applies business rules and stores the data for application. Many java Bean classes are needed to implement the Model part of the system. Some of these classes are LoginBean.java, Brand.java, Category.java, SaleReportBean.java etc.

Table 2. Example SaleReportBean.java

```
while(rs.next()){
    SaleReportBean dailyReportBean = new SaleReportBean();

    dailyReportBean.setInvestmentamount(rs.getInt("InvestmentAmount"));
    dailyReportBean.setReturnamount(rs.getInt("ReturnAmount"));
    dailyReportBean.setProfit(rs.getInt("Profit"));

    SimpleDateFormat ft = new SimpleDateFormat ("dd/MM/yyyy");
    String strDate=ft.format(rs.getDate("SalesDate"));

    dailyReportBean.setStringdate(strDate);

    dailyReportList.add(dailyReportBean);

}
```

Java Server Pages

The View represents the user interface. In J2EE, the Java Server Pages (JSP) is the view of the application which contains only presentation constructs: it receives all the information needs from the controller. JSP is a scripting language that allows HTML programmer to mix static HTML with dynamically generated HTML [14].

Table 3. Example DailyReport.jsp

```
<table class="table responsive table-bordered afont" id="jdTable">
 <thead class="table-heading">
        <tr>
            <th width="8%">No </th>
            <th >Investment Amount </th>
            <th > Return Amount</th>
            <th>Profit</th>
            <th>Sale Date </th>
            <th>Action </th>

          </tr>
   </thead>
   <tbody>
```

Servlet

The Java Servlet (**Controller**) specification defines an application programming interface for communication between the web/application server and the application program. The HttpServlet class in Java implements the servlet API specification: servlet classed used to implement specific functions are defined as subclasses of this class [15].

In this system many servlet classes (eg., LoginServlet, BrandServlet, CatgoryServlet, ProductServlet, DailySaleReportServlet) act as a Controller. First it reads parameter form the request. And then that parameter is forwarded to Model. Once it retrieves necessary data from Model, it forwards response to JSP, view of the application.

Table 4. Example DailySaleReport.servlet

```
public class ViewDailySaleReportServlet extends HttpServlet {

    @Override
    protected void doGet(HttpServletRequest req, HttpServletResponse resp)
            throws ServletException, IOException {
        // TODO Auto-generated method stub
        req.setCharacterEncoding("utf-8");
        resp.setCharacterEncoding("utf-8");

    DailySaleReportDao viewDao= new DailySaleReportDao();
        List<SaleReportBean> dailyRlist = viewDao.ViewAllDailySaleReport();
        req.setAttribute("dailyRlist", dailyRlist);

        RequestDispatcher view = req.getRequestDispatcher("/Admin/ViewDailySaleReport.jsp");
        view.forward(req, resp);
    }


}
```

5. RESULT AND DISCUSSION

MVC design pattern is very helpful for developers for developing web applications. For developing large scale and Dynamic Web Applications, MVC is the powerful and leading programming paradigm. MVC architecture reduces the complexity of the application because the application is divided into three major components: model, view and controller. Moreover, MVC architecture provides higher level of abstraction to their web applications and elements of applications can be tested simply. Table 5 describes the testing results for functionalities of the system. Figure 5, 6, and 7 illustrates SaleReport, SaleReportDetail and Instock pages of the system.

Table 5. Testing Result of the Functions of the System

No	Function Test	Test Form	Expected Result	Result
1.	Login	Open Login Form Enter Admin/ Staff name and password Admin/ Staff name and password verification	Login Form appeared Admin/ Staff name and password entered save on Form If verified, the Home page of the web application is displayed according to access rights. If not verified, the Login page will appear again.	Done Done Done
2.	AddProduct/ AddCategory/ AddBrand	Open a Add Form for adding product or category or brand Enter the required data related to the product or category or brand that you want to add and then perform add process	AddProduct or AddCategory or AddBrand Form appeared If succeed, the insert successful dialog box is displayed.	Done Done
3.	ViewProductList	Click the product list tab from Home page	The product list is displayed with the product name and its total quantities.	Done
4.	ViewInstockRecordList	Click the instock record list tab from Home page of application.	Instock Record List is appeared product name and quantity according to inserted date of each product.	Done
5.	SearchProduct	On the product list form, enter the product name that you want to search	The product list is displayed according to the searched product name.	Done
6.	UpdateProduct/ UpdateCategory/ UpdateBrand	Open a Update Form for modifying the product or category or brand Enter the required data related to the product or category or brand that you want to modify and then perform update process	UpdateProduct or UpdateCategory or UpdateBrand Form appeared If succeed, the update successful dialog box is displayed.	Done Done
7.	ViewReport	Click the view report tab on Home Page of web application.	The daily, weekly, monthly, and yearly reports are displayed according to the user's preferences.	Done

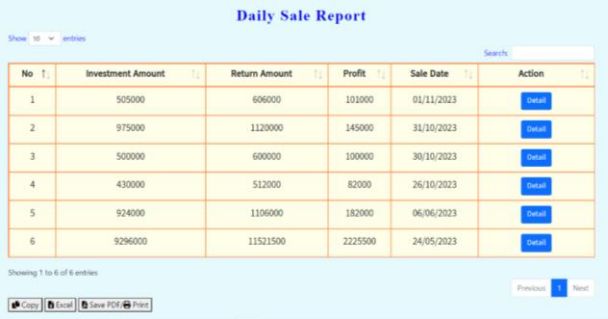


Figure 5. Daily Sale Report

Figure 6. Daily Sale Report Detail

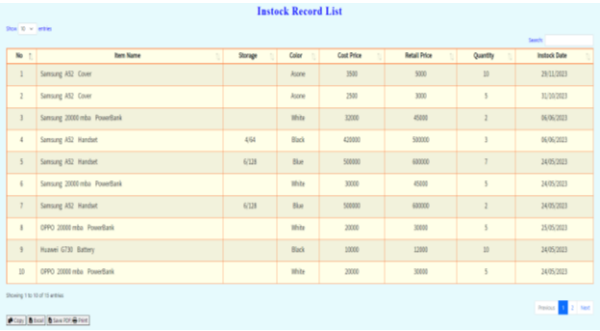


Figure 7. Viewing in Stock Items

6. CONCLUSIONS

A web application that is scalable, transparent and portable can he implemented using MVC architecture. Since MVC architecture divides the logic of application into three layers, different developers can work simultaneously on these three layers of the same application. Generally, the expense for setting up the MVC architecture is usually high. But large project built in MVC architecture will be processed faster than non-MVC project.

REFERENCES

[1] M. Ma, J. Yang, P. Wang, W. Liu and J. Zhang (2019) *Light-Weight and Scalable Hierarchical-MVC Architecture for Cloud Web Applications, 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/* pp. 40-45.

[2]https://www.geeksforgeeks.org/mvc-design-pattern/

[3] Andri Sunardi and Suharjito (2019) *MVC Architecture: A Comparative Study Between Laravel Framework and Slim Framework in Freelancer Project Monitoring System Web Based*, 4th International Conference on Computer Science and Computational Intelligence.

[4] Abdual Majeed and Ibtisam Rauf (2018) MVC Architecture: *A Detailed Insight to the Modern Web Applications Development*, Journal of Solar and Photoneergy Systems.

[5] Thanh Son Huynth et. al, *Design and Implementation of Web Application Based on MVC Laravel Architecture* (2022) European Journal of Electrical Engineering and Computer Science, Volume 6, Issue 4.

[6] Ebenezer et.al, *A Review on Software Architectural Patterns* (2021) Global Scientific Journals, Volume 8, Issue 8.

[7]https://www.prepbytes.com/blog/java/mvc-architecture-in-java/

[8]https://coderwall.com/p/l-a79g/the-principles-of-the-mvc-design-pattern

[9]https://www.prepbytes.com/blog/java/mvc-architecture-in-java/

[10] Nur Syufiza et. al (2021) *Information System Development with MVC Approach: A Case Study on Team Communication Strategy*, Turkish Journal of Computer and Mathematics Education, Volume 12.

[11] ICTTI *Advanced Java Lecture Notes* (2020) Information and Communication Technology Training Institute.

[12]https://www.w3schools.com/MySQL/default.asp

[13]https://docs.oracle.com/javase/8/docs/technotes/guides/jdbc/

[14]https://docs.oracle.com/javase/8/docs/technotes/guides/jdbc/

[15] Abraham Silberschatz et. al, *Database System Concepts* (2022) Mc Graw Hill Education.

Authors

Biography

Myintzu Phyo Aung (Ph.D) got her M.C.Sc and Ph.D (IT) degrees in 2008 and 2014 respectively from University of Computer Studies, Mandalay. She has over 18 years of experience in teaching on Information Science and Computer Science subjects. She served as a teaching faculty at University of Computer Studies, Mandalay, University of Computer Studies (Myitkyina) and Myanmar Institute of Information Technology. Currently she served as an Associate Professor at University of Computer Studies (Mandalay). She had published many International Conference Papers and Journals and gave presentations on her research work. The followings are her selected publications at International Conference and Journals.

1. *Construction of Finite State Machine for Myanmar Noun Phrase* (2013), MJIIT-JUC Joint International Symposium, Japan.
2. *Identification and Translation of Myanmar Noun Phrase to English* (2014), International Journal of Computational Linguistics and Natural Language Processing, India.
3. *New Phrase Chunking System for Myanmar Natural Language Processing* (2014), Applied Mechanics and Materials, Malaysia.
4. *Extraction of Myanmar Noun Phrase using Maximum Matching Approach (2015)*, International Workshop on Plasma Application and Hybrid Functionally Materials, USA.
5. *Constructing Myanmar Phrase Structure Grammar for Myanmar Noun Phrase Extraction* (2016), International Conference on Science and Engineering, Myanmar.
6. *Stochastic Context Free Grammar for Statistical Parsing of Myanmar Natural Language Processing (2018)*, International Conference on Computer Applications, Myanmar.
7. *Proposed Framework for Stochastic Parsing of Myanmar Language (2018)*, Springer Journal, Japan.
8. *Phrase Based Translation Model for Myanmar Language* (2019), Proceedings of Joint International Conference on Science, Technology and Innovation, Myanmar.

Zin Mar Oo got her B.C.Sc(Hons.) and M.C.Sc degrees in University of Computer Studies (Pinlon) and University of Computer Studies (Meiktila). She worked as a teaching faculty at University of Computer Studies (Meiktila), University of Computer Studies (Monywa), and University of Computer Studies (Myitkyina). Currently, she works as a Lecturer at University of Computer Studies (Mandalay). Her selected publications are followings.

1. *Implementation of Software Agent for Agricultural Fulfilment System*, Proceedings of the second conference on Applied Information and Communication Technology
2. *Travel Package Recommendation System*, Journal of Innovative Research, Volume 1
3. *An Approach of Advice for Peasant Aid System*, Myanmar Universities' Research Conference, 2019
4. *Iris recognition Attendance System*, Scientific Journal of Innovative Research, Volume 1.

Theint Theint Htwe got her B.C.Sc (Hons.) and M.C.Sc degrees from University of Computer Studies (Pakokku) and University of Computer Studies, Yangon, respectively. She worked as teaching faculty at University of Computer Studies, Yangon, University of Computer Studies (Pakokku), University of Computer Studies (Mandalay), and University of Computer Studies (Loikaw). Currently, she works as a Lecturer at University of Computer Studies (Mandalay). Her selected publications are followings.

1. *Decision Support System for House Interior Design*, Proceedings of the Conference on Parallel and Soft Computing, 2009.
2. *Car Recommender System Based on Customer's Reviews Using Model Based Filtering, University Journal of Science, Engineering and Research, 2020.*