

**E-COMMERCE CUSTOMER SEGMENTATION USING K-MEANS**

**Submitted by**

**Hemanth Kumar D (231801057)**

**Hitesh Kumar S (231801059)**

**AI23331 - FUNDAMENTALS OF MACHINE LEARNING**

**Department of Artificial Intelligence and Data Science**

**Rajalakshmi Engineering College, Thandalam**

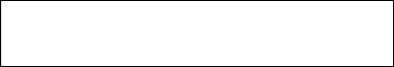
**Nov 2024**

****

**BONAFIDE CERTIFICATE**

**NAME………………………………………………………………………….**

**ACADEMIC YEAR……………SEMESTER………….BRANCH ………**



**UNIVERSITY REGISTER No.**

Certified that this is the Bonafide record of work done by the above students in the Mini Project titled " **CUSTOMER SEGMENTATION USING K-MEANS**" in the subject **AI23331 – FUNDAMENTALS OF MACHINE LEARNING** during the year **2024 - 2025.**

**Signature of Faculty – in – Charge**

**Submitted for the Practical Examination held on -----------------------**

**Internal Examiner External Examiner**

## LIST OF CONTENTS

## Abstract 4

## Abbreviations 7

## Chapter 1: Introduction 6

#### 1.1 Customer Segementation System 6

## 1.2 Problem Statement 7

## Chapter 2: Literature Survey 8

## 2.1 Customer Segmentation Techniques 8

## 2.2 Application of K-means in E-commerce 8

## Chapter 3: Methodology 9

## 3.1 Proposed Workflow 10

## 3.2 Working of K-Means 10

## 3.3 Dataset Detail 11

## 3.4 System Configuration 11

## Chapter 4: Results and Implementation 14

## 4.1 Clusters 17

## 4.2 Results 19

## Chapter 5: Conclusion 20

## References

## ABSTRACT

## This project focuses on customer segmentation in an e-commerce setting using the K-means clustering algorithm. The goal is to group customers based on their purchasing behavior, such as frequency of purchases, spending patterns, and product preferences. By leveraging K-means, we aim to identify distinct customer segments that exhibit similar characteristics, allowing for more targeted marketing strategies and personalized recommendations. The algorithm divides the customer base into clusters, with each cluster representing a unique group with specific needs and behaviors. This segmentation helps the business optimize resource allocation, improve customer retention, and enhance the overall customer experience. The project involves data preprocessing, feature selection, and determining the optimal number of clusters using methods like the Elbow method and Silhouette score. The results provide actionable insights into customer preferences, enabling businesses to tailor their marketing efforts and improve decision-making.

**ABBREVIATION**

|  |  |  |  |
| --- | --- | --- | --- |
| Sr No. | Abbreviation | Meaning |  |
| 1 | CNN | Convolutional neural network |
| 2 | MRI | Magnetic resonance imaging |
| 3 | FLAIR | Fluid attenuated in version recovery weighted | MRI |
| 4 | TR | Time repetition |  |
| 5 | TE | Pulse sequence parameter |  |
| 6 | VGG 16 | Visual Geometry Group |  |
| 7 | FC | Fully connected layer |  |
| 8 | ReLU | Rectified linear unit |  |
| 9 | LRN | Local response normalization |  |
| 10 | SVM | Support vector machine |  |
| 11 | KNN | K nearest neighbor |  |

# CHAPTER 1

## INTRODUCTION

#### CUSTOMER SEGEMENTATION SYSTEM

In today’s rapidly evolving e-commerce industry, businesses are increasingly relying on data-driven approaches to enhance customer experience and optimize marketing strategies. One of the most effective ways to achieve this is through **customer segmentation**, which groups customers with similar behaviors, preferences, and characteristics into distinct segments. By understanding these segments, businesses can better tailor their products, services, and communications to meet the specific needs of each customer group, ultimately leading to improved customer satisfaction and higher revenue.

**Customer segmentation** has become a critical strategy in the e-commerce sector as companies seek to stay competitive in a market characterized by intense competition and high customer expectations. The goal of segmentation is to identify groups of customers that share common attributes and behaviors, such as purchasing patterns, browsing history, demographics, and engagement with marketing campaigns. This enables businesses to target their marketing efforts more efficiently and effectively, ensuring that they are addressing the right customers with the right offers.

Segmentation can be achieved through various methods, and **machine learning algorithms**, particularly **unsupervised learning algorithms**, have gained significant popularity for this task. Among these, the **K-means clustering algorithm** is widely used for customer segmentation due to its simplicity, efficiency, and ability to handle large datasets. K-means allows businesses to identify natural groupings within customer data, making it easier to understand customer behavior and tailor business strategies accordingly.

* 1. **Problem Statement**

The challenge for e-commerce businesses is not only to collect vast amounts of customer data but also to **interpret and analyze** this data to derive actionable insights. With the increasing volume and complexity of customer data, traditional methods of segmentation are no longer sufficient. Companies need advanced techniques that can handle large datasets, uncover hidden patterns, and provide a more granular understanding of customer behavior.

K-means clustering offers a solution to this problem by dividing customers into **homogeneous clusters** based on their attributes, such as purchase history, frequency of interaction, and spending behavior. However, choosing the **right number of clusters** and ensuring that the segmentation is meaningful presents its own challenges. Improper segmentation can lead to inaccurate targeting, wasted marketing efforts, and missed business opportunities.

This research aims to address the challenge of customer segmentation in e-commerce by applying the K-means clustering algorithm to segment customers based on their purchasing behavior.

The project seeks to:

* Explore the effectiveness of K-means clustering for identifying customer segments in the e-commerce domain.
* Determine the optimal number of clusters for meaningful segmentation using methods such as the Elbow method and Silhouette score.
* Analyze the distinct characteristics of each customer segment and provide recommendations for targeted marketing strategies.

By leveraging K-means clustering, this study aims to enhance the understanding of customer behavior and provide e-commerce businesses with actionable insights to improve their marketing strategies, customer retention, and overall business performance.

# CHAPTER 2

## 2.LITERATURE SURVEY

Customer segmentation in the e-commerce industry has become a crucial area of research, as it enables businesses to target specific customer groups more effectively. Over the years, several approaches and algorithms have been explored to enhance the accuracy and efficiency of segmentation. Among these methods, K-means clustering stands out as one of the most widely used unsupervised learning techniques for segmenting customers based on their behavior and attributes. In this chapter, we review key studies and methodologies related to customer segmentation, with a focus on K-means clustering and its application in the e-commerce domain.

**2.1 Customer Segmentation Techniques**

Early customer segmentation strategies were largely based on manual or rule-based approaches, which relied heavily on demographic data and simplistic grouping methods. These methods, while useful, often failed to account for the complexities of customer behavior, especially in dynamic e-commerce environments. As a result, there has been a shift toward data-driven approaches, leveraging machine learning algorithms to uncover hidden patterns in large datasets

.

One of the most popular techniques used in customer segmentation is K-means clustering. K-means is an unsupervised learning algorithm that partitions data into a specified number of clusters by minimizing the variance within each cluster. The algorithm assigns each data point to the nearest cluster center, and iteratively refines the clusters until the centroids stabilize. Studies such as those by Jain (2010) and Xu and Wunsch (2005) have shown that K-means clustering is particularly effective in handling large-scale datasets and can uncover meaningful patterns in customer behavior, such as purchasing frequency, spending habits, and preferences.

**2.2 Application of K-means in E-commerce**

In recent years, there has been significant research focusing on the application of K-means clustering in e-commerce customer segmentation. A study by Bose and Chen (2009) demonstrated the use of K-means to segment customers based on their transaction history, identifying distinct groups such as high-value customers, frequent buyers, and occasional shoppers. The segmentation results were used to design targeted marketing campaigns, which led to increased conversion rates and customer satisfaction.

Similarly, Aggarwal et al. (2015) applied K-means clustering to customer data in the online retail industry, where they used features like browsing patterns, product categories, and purchase frequency. The segmentation enabled the business to offer personalized recommendations, leading to better customer retention and higher sales. The effectiveness of K-means was attributed to its ability to handle both numerical and categorical data, making it suitable for e-commerce environments with diverse customer behavior.

Moreover, research by Deng and Wu (2017) explored the combination of K-means clustering with other data analysis techniques, such as dimensionality reduction (e.g., PCA), to improve clustering performance. By reducing the number of features while retaining essential information, the researchers were able to achieve more accurate and interpretable customer segments. This hybrid approach helped overcome some of the challenges associated with high-dimensional e-commerce data.

**2.3 Challenges and Future Directions**

Despite the success of K-means clustering in customer segmentation, there are challenges to consider. One of the major limitations is the selection of the optimal number of clusters, which can significantly impact the quality of the segmentation. Methods like the Elbow method and Silhouette score have been proposed to help determine the appropriate number of clusters, but these methods are not always foolproof, especially when data is noisy or highly variable.

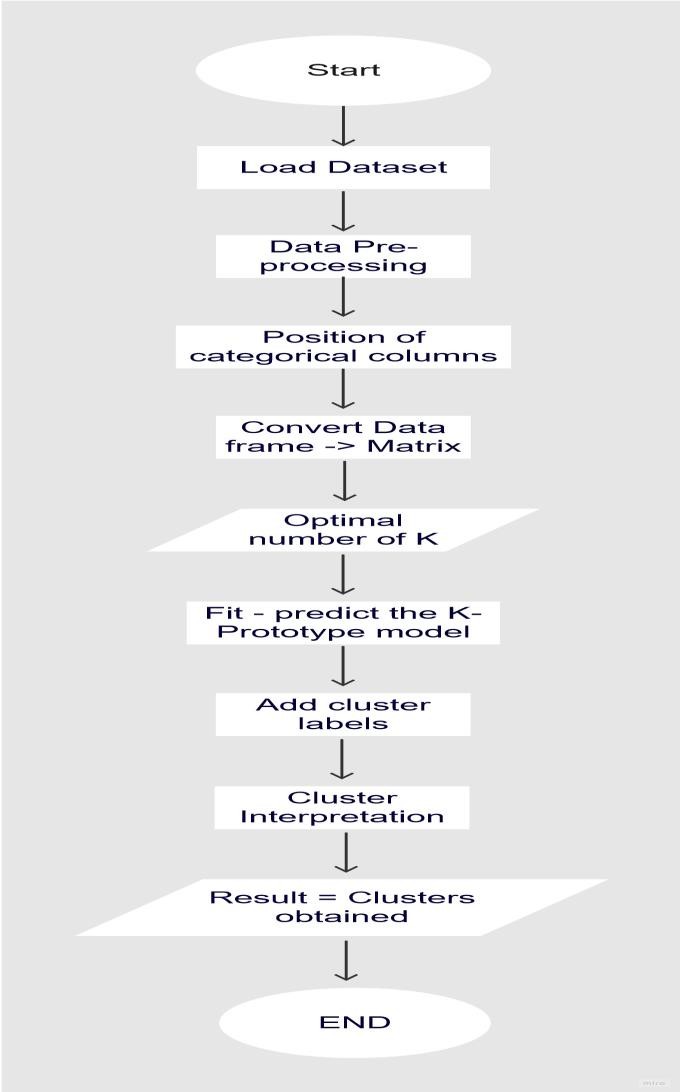
Additionally, K-means clustering assumes that clusters are spherical and of equal size, which may not always hold true in real-world e-commerce datasets. To address this, variations of K-means (e.g., K-medoids or DBSCAN) have been introduced to handle more complex cluster shapes and outliers. Further research is needed to explore how K-means can be improved or combined with other algorithms for more robust customersegmentation.

# CHAPTER 3

# 3.METHODOLOGY

**3.1PROPOSED WORKFLOW**

Firstly we take the dataset, in our case, it is an E-Commerce dataset, load them using pandas. Next, we search for any missing or duplicate values in any of the rows, as the K-Prototype Algorithm cannot handle any missing or duplicate values. Next, we seek the categorical column positions from the dataset, and we convert the data frame to a matrix. Next, we find the optimal number of ‘K’ using the elbow method and then we build the k-prototype model and use the fit\_predict for it. Next, we add the cluster labels and then do interpretation to get the different clusters in which we can identify the customer spending traits.

.

### 3.2 Working of CNN model

In this e-commerce customer segmentation project, the **K-means clustering algorithm** is employed to group customers into segments based on their purchasing behaviors. The working of the K-means algorithm can be broken down into a series of steps that help identify distinct groups of customers with similar characteristics, which can then be used for targeted marketing and personalized customer engagement.

1. **Initialization**

The K-means algorithm requires the user to specify the number of clusters, KKK, before the clustering process begins. In this project, the goal is to segment customers into distinct groups based on features like purchase frequency, total spend, and other relevant customer behaviors. Initially, the number of clusters is either decided based on business goals or selected using methods such as the **Elbow method** or **Silhouette score**.

For example, in this project, we might start with a guess of K=3K = 3K=3 clusters, but this will be adjusted later based on the analysis of the clustering performance.

1. **Cluster Assignment (First Step)**

Once KKK is chosen, the algorithm begins by randomly selecting KKK points as the initial **centroids** (central points) for each cluster. These centroids represent the "center" of each cluster.

Next, the algorithm assigns each customer (data point) to the nearest centroid. The proximity between the customer and the centroid is typically calculated using **Euclidean distance**:

Distance=(x1−x2)2+(y1−y2)2\text{Distance} = \sqrt{(x\_1 - x\_2)^2 + (y\_1 - y\_2)^2}Distance=(x1​−x2​)2+(y1​−y2​)2​

Where x1,y1x\_1, y\_1x1​,y1​ are the features of a customer and x2,y2x\_2, y\_2x2​,y2​ are the coordinates of the centroid. The customer is assigned to the cluster whose centroid is closest.

1. **Recalculating Centroids (Second Step)**

After all the customers have been assigned to a cluster, the next step is to **recalculate** the centroids of the clusters. The new centroid is computed by taking the mean (average) of all customers assigned to that cluster across all features. This gives us the updated "center" of the cluster.

1. **Iteration and Convergence**

Steps 2 and 3 are repeated iteratively:

1. **Assign customers to the new centroids**.
2. **Recalculate the centroids** based on the updated customer assignments.

This process continues until the centroids no longer change significantly, or the maximum number of iterations is reached. When the centroids stabilize, meaning no customer changes clusters, the algorithm has **converged**, and the final clusters are determined.

1. **Evaluating the Clusters**

Once the clusters are formed, the next step is to analyze the customer segments:

* **Cluster profiling**: Examine the features that define each cluster. For example, one cluster may consist of customers who frequently make large purchases, while another may include customers who make smaller, infrequent purchases.
* **Cluster visualization**: Use visual techniques like **scatter plots** or **pairwise plots** to understand how customers are grouped and to visualize the separations between clusters.

1. **Determining the Optimal Number of Clusters**

In practice, it’s important to select the right number of clusters for meaningful segmentation. The **Elbow method** and **Silhouette score** are two techniques used to evaluate different values of KKK and determine the optimal number of clusters:

* **Elbow Method**: Plot the sum of squared distances (inertia) from each point to its assigned centroid for various values of KKK. The optimal KKK is typically chosen where the inertia curve forms an "elbow," meaning adding more clusters no longer significantly reduces the inertia.
* **Silhouette Score**: Measures how similar each data point is to its own cluster versus other clusters. A higher silhouette score indicates well-separated clusters, and the optimal KKK is chosen where the silhouette score is maximized.

1. **Business Insights and Targeted Marketing**

Once the segmentation is complete, the results can be used to create **targeted marketing strategies** for each customer segment. For example:

* High-spending customers can be offered loyalty programs or exclusive deals.
* Price-sensitive customers may be targeted with discounts or special offers.
* Infrequent buyers can receive personalized recommendations to encourage repeat purchases.

Each segment is treated according to its characteristics, leading to more personalized customer engagement, increased retention, and higher sales.

#### 3.3: DATASET DETAIL

**Kaggle**: A popular platform for datasets and data science competitions.

* For example, you can use the **"E-commerce Customer Segmentation"** dataset on Kaggle.

## 3.4 System Configuration

# 3.3.1 Software Requirements

Visual Studio Code, Python 3.9, PIP

Modules: Keras, Tensorflow, CV2, NumPy, Flask

# Visual Studio Code:

Visual Studio Code is a streamlined code editor with support for development operations like debugging, task running, and version control. It aims to provide just the tools a developer needs for a quick code-build- debug cycle and leaves more complex workflows to fuller featured IDEs, such as Visual Studio IDE.

# Python:

**PIP:**

Python is an interpreted, high-level, general purpose programming language created by Guido Van Rossum and first released in 1991, Python's design philosophy emphasizes code Readability with its notable use of significant Whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

It is the package management system used to install and manage software packages written in Python.

# NumPy:

NumPy is a general-purpose array-processing package. It provides a highperformance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

* A powerful N-dimensional array object
* Sophisticated (broadcasting) functions
* Tools for integrating C/C++ and Fortran code
* Useful linear algebra, Fourier transform, and random number capabilities

# Tensor Flow:

Tensor flow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google.

# Keras:

Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, R, Theano, or Plaid ML. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible. Keras contains numerous implementations of commonly used neural- network building blocks such as layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier to simplify the coding necessary for writing deep neural network code.

# OpenCV:

OpenCV (Open source computer vision) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by willow garage then Itseez (which was later acquired by Intel). The library is cross platform and free for use under the open source BSD license. OpenCV supports some models from deep learning frameworks like TensorFlow, Torch, PyTorch (after converting to

an ONNX model) and Caffe according to a defined list of supported layers. It promotes Open Vision Capsules. which is a portable format, compatible with all other formats.

# 3.3.2Hardware Configuration

* + - 1. Processor: Intel core i5 or above.
      2. 64-bit, quad-core, 2.5 GHz minimum per core
      3. Ram: 4 GB or more
      4. Hard disk: 10 GB of available space or more.
      5. Display: Dual XGA (1024 x 768) or higher resolution monitors
      6. Operating system: Windo

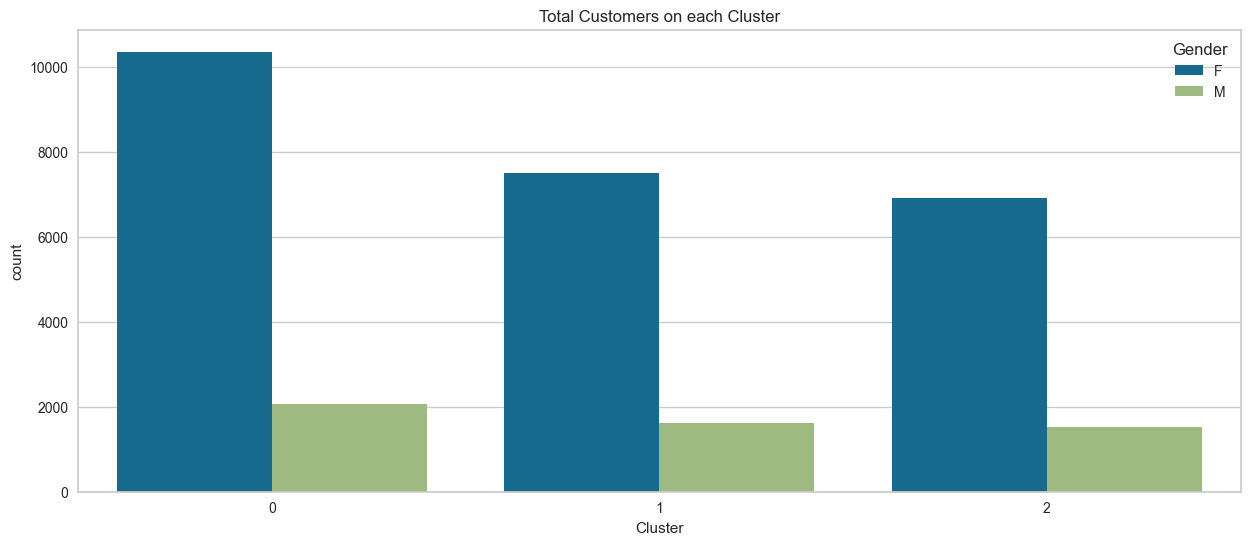
# CHAPTER 4

# 4.RESULTS AND IMPLEMENTATION

#### 4.1 CLUSTERS

#### 

#### 

.

# 4.2.RESULTS

# 

# 5..CONCLUSION

In brain tumor detection we have studied about feature based existing work. In feature based we have study about image processing techniques likes image pre-processing, image segmentation, features extraction, classification. And also study about deep learning techniques CNN and VGG16.In this system we have detect the tumor is present or not if the tumour is present then model return’s yes otherwise it return no. and we have compared CNN with the VGG 16 Model. The result of comparison VGG 16 is more accurate than CNN. However, not every task is said to be perfect in this development field even more improvement may be possible in this application. I have learned so many things and gained a lot of knowledge about development field.

### References

[1]L.Guo,L.Zhao,Y.Wu,Y.Li,G.Xu,andQ.Yan,“Tumordetection in MR images using one- class immune feature weighted SVMs,” IEEE Transactions on Magnetics, vol. 47, no. 10, pp. 3849–3852,2011.

[2]R.Kumari,“SVMclassificationanapproachondetectingabnormalityinbrainMRIimages,”Inter nationalJournalofEngineeringResearchandApplications,vol.3,pp.1686–1690,2013.

1. DICOM Samples Image Sets, [http://www.osirix-viewer.com/.](http://www.osirix-viewer.com/)
2. “Brainweb:SimulatedBrainDatabase” [http://brainweb.bic.mni.mcgill.ca/cgi/brainweb1.](http://brainweb.bic.mni.mcgill.ca/cgi/brainweb1)
3. Obtainable Online: [*www.cancer.ca/~/media/CCE* 10/08/2015](http://www.cancer.ca/~/media/CCE%2010/08/2015).
4. J. C. Buckner, P. D. Brown, B. P. O’Neill, F. B. Meyer , C. J. Wetmore,J. H Uhm, "Central nervous system tumors." In *Mayo Clinic Proceedings*,Vol. 82, No. 10, pp. 1271- 1286, October 2007.
5. Deepa , Singh Akansha. (2016). - Review of Brain Tumor Detection from tomography.

*International Conference on Computing for Sustainable Global Development (INDIACom)*

1. R. A. Novellines, M. D. - *Squire's fundamentals of radiology*; Six Edition; UPR, 2004.
2. preston, D. c. (2006). Magnetic Resonance Imaging (MRI) of the Brain and Spine from Basics. *casemed.case.edu* .
3. Hendrik RE. (2005) Glossary of MR Terms from *American College of Radiology* .

.

[11].A. Demirhan, M. Toru, and I. Guler, “Segmentation of tumor and edema along with healthy tissues of brain using wavelets and neural networks,” IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1451–1458, 2015.

1. Nilesh Bhaskarrao Bahadure, A.K. (2017, March 6). Retrieved from https:/[/www.hindawi.com/journals/ijbi/2017/9749108/](http://www.hindawi.com/journals/ijbi/2017/9749108/).
2. S. Mohsin, S. Sajjad, Z. Malik, and A. H. Abdullah, “Efficient way of skull stripping in MRI to detect brain tumor by applying morphological operations, after detection of false background,” *International Journal of Information and Education Technology*, vol. 2, no. 4,

pp. 335–337, 2012.

1. Gavale, P. M., Aher, P. V., & Wani, D. V. (2017, April 4). Retrieved from <https://www.irjet.net/archives/V4/i4/IRJET-V4I462.pdf>.
2. N. Gordillo, E. Montseny, and P. Sobrevilla, “State of the art survey on MRI brain tumor segmentation,”*Magnetic Resonance Imaging*, vol. 31, no. 8, pp. 1426–1438, 2013.
3. Samantaray, M.(2016, November 3). Retrieved from <http://ieeexplore.ieee.org/document/7727089/> .
4. Nandi, A. (2016, April 11) Retrieved from <http://ieeexplore.ieee.org/document/7449892/>
5. C. C. Benson and V. L. Lajish, “Morphology based enhancement and skull stripping of MRI brain images,” in *Proceedings of the international Conference on Intelligent Computing Applications (ICICA ’14)*, pp. 254–257, Tamilnadu, India, March 2014.
6. S. Z. Oo and A. S. Khaing, “Brain tumor detection and segmentation using watershed segmentation and morphological operation,” *International Journal of Research in Engineering and Technology*, vol. 3, no. 3, pp. 367–374, 2014.
7. R. Roslan, N. Jamil, and R. Mahmud, “Skull stripping magnetic resonance images brain images: region growing versus mathematical morphology,” *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 3, pp. 150– 158, 2011.