Fuzzy Systems and Soft Computing
ISSN : 1819-4362

A SURVEY ON THE ANALYSIS OF DIABETIC DISEASE DATA USING CLUSTERING TECHNIQUES IN DATA MINING

K. Emayavaramban Guest Lecturer, Loganatha Narayanaswamy Government College, Ponneri, India. Mail: emayammsc2002@gmail.com
T. Velmurugan Associate Professor, PG and Research Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Chennai, India;Mail: velmurugan_dgvc@yahoo.co.in

Abstract:
In recent days, disease analysis using computational methods play a very vital role. Medical data analysis is crucial nowadays in the field of diagnostics. A number of approaches utilised for the disease analysis. Finding the disease which is available in the chosen data set is very easily done by using some of the computer algorithms. This survey based analysis is carried out in order to find the diabetic diseases which are used clustering algorithms in data mining particularly. Many researchers done their research work and published research articles based on their approach and concluded research findings on the obtained results. This work has a survey on the analysis of clustering algorithms like k-Means and Fuzzy C-Means and concluded that which algorithm yields best results via some others research results. Finally, suggested that the best algorithm among the chosen two algorithms for the diabetic data analysis.

Keywords:
Clustering Algorithms, Diabetic Data, Survey Based Analysis, Performance of Algorithms.

1. Introduction

Diabetes mellitus is a prevalent and chronic medical condition affecting millions of people worldwide, necessitating advanced methods for its early detection and management. With the advent of data mining techniques, significant strides have been made in the field of medical diagnostics, enabling the analysis of vast amounts of medical data to uncover hidden patterns and insights. The Core aspect of Data Mining called the Clustering algorithms, play a crucial role in segmenting and analysing complex datasets, providing a deeper understanding of disease characteristics and trends. This survey focuses on the application of clustering algorithm, k-means and fuzzy c-means to analyse diabetic disease data. By examining and comparing various research studies, this paper aims to evaluate the performance of these algorithms in identifying and classifying diabetic data. The goal is to determine which algorithm yields the best results, thereby offering valuable insights for future research and practical applications in the diagnosis and management of diabetes.

The significance of accurately analysing diabetic data cannot be overstated, as it directly impacts the ability to predict disease progression, tailor treatments, and improve patient outcomes. Clustering techniques, by grouping similar data points, help in identifying patterns that can lead to early diagnosis and effective intervention strategies. These algorithms have been widely utilised in medical data analysis, each with its own set of strengths and limitations.

Advanced clustering methods allow for the identification of complex patterns within diabetic datasets, providing a more nuanced analysis where data points can belong to multiple clusters with varying degrees of membership. This survey will delve into the comparative analysis of these algorithms, exploring their applicability and effectiveness in diabetic disease data analysis, and will highlight the scenarios in which one method may be preferred over another in clustering techniques.

2. Applications of Data Mining for Diabetic Disease

A Proposal of a high precision diagnostic analysis using k-means clustering technique was made in a research work carried out by Amira Hassan Abed et al. in [1]. In order to prepare date to implement a Clustering model, all the noisy, uncertain and inconsistent data were detected and removed from data set through the pre-processing. The k-means technique was applied on community health diabetes related indicators data set to cluster diabetic patients from the healthy ones with high accuracy and reliability results. In another research work by Hosam El-Sofany et al which was titled as “A Proposed

Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App" proposed a new process [2]. This study effectively predicted diabetes-related features of the dataset by employing a semi supervised model combined with strong gradient boosting.

Provided, a technique of the SMOTE was employed by some researchers in order to meet with the problems of imbalanced classes. Another research work with the Title of "Detection and Prediction of Diabetes Using Data Mining: comprehensive review" done by Khan, Farrukh Aslam, Khan Zeb, Mabrook Al-Rakhami, Abdelouahid Derhab, and Syed Ahmad Chan Bukhari. [3], has a twofold prospects. The Data mining based diagnosis and prediction solutions in the field of glycemic control for diabetes, are properly investigated first. The Second is that, with the help of an investigation, a comprehensive classification and comparison of the techniques that have been frequently used for diagnosis and prediction of diabetes based on important key metrics could be furnished. Provided, the challenges and future research directions in this area which come under the consideration of developing optimised solutions for diabetes detection and prediction, are Highlighted.

A Research work titled as "**PSO-FCM based data mining model to predict diabetic disease"** serves its purpose [4] The work states that, due to the presence of Blood Sugar levels that are higher than the normal, the disease of Diabetics is typically composed. In another perspective, the production of insulin may be regarded insufficient. The rising rate of the diabetes affected patients has been observed to have grown on a rapid scale throughout the world.Another research work known with a title of "Data mining classifier for predicting diabetics**"** created by Singh, Manpreet, Pankaj Bhambri, Inderjit Singh, Amit Jain, and Er Kirandeep Kaur [5] confirms the fact that there are various hardware and software methods used to predict and classify the data. There are many classification algorithms which can be employed to find and affirm whether a person is positively diabetic or not by various physical and biochemical characteristics. Every methodology has their own experimenting process by different techniques and also varies in accuracy. This work specialises in developing a method for the classification of the diabetic data.

A Research work by Saxena, Aditya, Megha Jain, and Prashant Shrivastava titled, "Data Mining Techniques Based Diabetes Prediction." in the "Indian Journal of Artificial Intelligence and Neural Networking" (2021) conveys a similar concept **[6].** This survey paper shows how the multiple diabetes disorders during the earlier stages can be detected with more numbers of approaches and data mining strategies helping one to avoid becoming a chronic patient because of diabetes consequently sparing more lives by the effective predictions. By the use of data mining tools and processes, diabetes is avoided and treatment rates are reduced. Another research titled as, "Likelihood prediction of diabetes at early stage using data mining techniques" created by Islam, M. M., Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra **[7]** conveys that which has become one of the fast growing chronic life threatening diseases. Due to the presence of a relatively long asymptomatic phase, early detection of diabetes is always desired for a clinically meaningful outcome. As the Diabetes has a long term asymptomatic phase levels, 50% of most of the people are unaware of their diabetic nature.

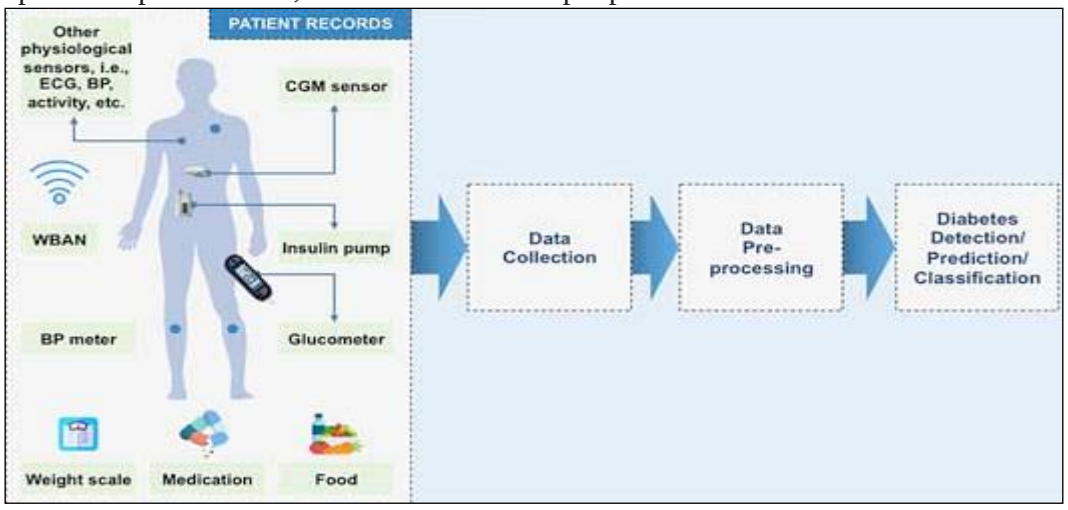

Figure 1: A generic flow of diabetes detection and prediction techniques.

In a Research work by Abdulqadir, Hindreen Rashid, Adnan Mohsin Abdulazeez, and Dilovan Assad Zebari titled, "Data mining classification techniques for diabetes prediction" [8], they affirm and estimate diabetes by its key features and also categorises the relations between conflicting elements.

A significant feature range was provided by the recursive random forest removal function. Random Forest Classifier investigated the diabetes estimate. Greater assistance for medical professionals for exercising care for their patients, is possible with RF, which offers greater precisions of about 75,7813, on par with the Support Vector Machine (SVM). Another Research work titled as "Early prediction of diabetic using data mining", made by Jaber, Fayzeh Abdulkareem, and Joy Winston James [9] states the report by The World Health Organization (WHO), in 2018, that 422 million people throughout the globe are living with diabetes, making it one of the most widespread chronic life-threatening conditions. Early diagnosis is often favoured for clinically relevant findings due to the comparatively longer asymptomatic period associated with diabetes. It is estimated that around 50% of people with diabetes go undiagnosed because of the length of time it takes for symptoms to appear.

A next research paper titled as "Early Prediction of Diabetes Disease Based on Data Mining Techniques", by Elsadek, Salma N., Lama S. Alshehri, Rawan A. Alqhatani, Zainah A. Algarni, Linda O. Elbadry, and Eyman A. Alyahyan [10] The ultimate goal of this particular study was to state the importance of Two Algorithms of Machine Learning: Random Forest and Multi-Layer Perceptron (MLP) using the WEKA environment, can estimate the accuracy in detection of Diabetes in the early stage. The dataset that came from the Machine Learning Repository of UCI, was applied to conduct the Experiments. 16 attributes and 520 instances collected by means of questionnaires from patients at Sylhet-based Sylhet Diabetes Hospital (Bangladesh) made the dataset for this experiment. The findings showed that, better results were yielded by the RF method in comparison with the MLP while enabling early prediction of diabetes with a high accuracy of 97.88%. The research work carried out by Ibrahim, Shahira, and Siti Shaliza Mohd Khairi [11] titled as "Predictive Data Mining Approaches for Diabetes Mellitus Type II Disease", was intended to identify the influential factors that may contribute to DM Type II by comparing the performance of different data mining approaches.

A retrospective cross sectional study was conducted from April 2017 to November 2018, where 684 patients from a public clinic participated. Logistic Regression, Decision Tree, Naïve Bayes and Artificial Neural Network (ANN) are the Four predictive Models that were involved in this Study. To evaluate the performance of the models, the Error measures (Average Squared Error and Misclassification Rate) with ROC Index were used. The Results show that the performance of Logistic Regression-Stepwise outperformed to other predictive models with classification accurateness of 73% and able to predict positive outcome (Y=1) correctly by 90%. The significant inputs that affect DM Type II prediction (Y=1) are Hypertension and Glycated Hemoglobin (HbA1c) given the Root Mean Squared Error (RMSE) of model is 0.424.

A Research paper titled as "Tapping into knowledge: Ontological data mining approach for detecting cardiovascular disease risk causes among diabetes patients." Presented by Alkattan, Hussein, S. K. Towfek, and M. Y. Shams. [12], suggested a novel method of Ontological Data Mining (ODM) for identifying the origins of CVD risk in diabetic patients. By incorporating domain knowledge and semantic linkages into the data mining process, the study aimed to improvise the readability and precision of prediction models. 70,000 records of the patients which were derived from a through a thorough clinical history and physical examination, had made up the dataset which was examined by 11 attributes. Abdelhamid, Abdelaziz A., Marwa M. Eid, Mostafa Abotaleb, and S. K. Towfek, created a study titled: "Identification of cardiovascular disease risk factors among diabetes patients using ontological data mining techniques." [13], in which they showed that they analysed a dataset of diabetes patients and investigated the risk variables that contribute to CVD by using an ontological data mining approach and Light GBM. The increasing risk factors among diabetic patients were made clear by the investigation of the association between diabetes and CVD. The demographics, health behaviours, and physiological indicators which happen to be the common influencers for the inceptions of heart disease, were looked up in this study.

Saeed, Saad Ebrahim, Burair Hassan Al Telaq, and Ahmed M. Zeki [14], made a Research paper, titled as "A Comparative Study on Classifying Diabetes Disease using Data Mining Models." The Paper tells about the field of Data Mining (DM) which helps in enhancing the healthcare system by identifying the main causes of diseases as well as predicting their incidence in the future. But there was an outstanding and immediate problem that was perceived. It happened to be the DM comparison for determining the most accurate algorithm that can serve diabetes patient's data. In the Paper they have used, the Cross-Industry Standard Process for Data Mining (CRISP-DM) as a methodology.

The best performance in terms of accuracy was Decision Tree J48 with an accuracy of 96% higher than both KNN 93.3% when K value equals 3, and SVM which obtained 92.1%. In the KNN model, the study found increasing parameters of K slightly decrease the accuracy, however, this model achieved the highest measurement in terms of ROC in this study.

3. Researches on Diabetic Data Using Clustering Techniques

A research work carried out by Anand, Tanvi, Rekha Pal, and Sanjay Kumar Dub et al [15], shows the utility of the clustering technique, among the various data mining techniques, because clustering is comparatively a good approach to handle with the complex task. In order to identify the best clustering technique, some experiments were conducted. That Clustering technique can easily identify the various impacting factors of DR in a much easily accessible approach. The results after the experiment showed that the performance of K-means is comparatively much effective than other clustering techniques. This analysis would be a boon to identify best algorithm for disease detection and provide preventive measures in advance for a concerned Medical practitioner. "Identifying diabetic patient profile through machine learning-based clustering analysis" is a Research paper created by Gomes, João, João Lopes, Tiago Guimarães, and Manuel Filipe Santos [16] The mission of defining the typology of diabetic patients by building Machine Learning (ML) models from registered clinical information, medication, complementary diagnostic tools, therapeutic and monitoring data, and registered medication data happens to be the primary objective of this Project which is still carried out in collaboration with the Unidade Local de Saúde do Alto Minho (ULSAM).

A research paper which is titled as "Cluster analysis of type II Diabetes Mellitus Patients with the Fuzzy C-means method" made by Lomo, Simeftiany Indrilemta, and Endang Darmawan [17] is a retrospective study of 447 medical record data of type II diabetes mellitus patients aged 18 years old or above, who were hospitalised in the PKU Muhammadiyah Gamping Hospital from 2015-2019. Clustering is using the PCA-Fuzzy C-Means method based on patient's survival status, demographic characteristics, therapy, and blood glucose (BG) levels. Another research titled as "An Efficient and Scalable Technique for Clustering Comorbidity Patterns of Diabetic Patients from Clinical Dataset" was done by Bramesh, S. M., and Anil Kumar KM. [18]. In order to learn the relationships between diabetes patient's clinical profiles and for the classification and categorization which are as an essential pre-processing stage for analysis tasks, the Clustering diabetic patients with comorbidity patterns are necessary to learn.

However, the traditional clustering methods are more difficult to apply as the data are of heterogeneity, which in turn necessitates the development of novel clustering algorithms. The authors recommend here, an effective and scalable clustering technique suitable for datasets made up of attributes which are atomic and set-valued. Another Research paper by the title, "A diabetic prediction system based on mean shift clustering" that is made by Teki, Satyanarayana Murthy, Kuncham Venkata Sriharsha, and Mohan Krishna Varma Nandimandalam [19] states that an abnormal rise in glucose levels may lead to diabetes which lead to a diagnosis of around 30 million people with this disease in our country. In this perspective Indian Council of The Medical Research which is funded by the Registry of People with diabetes in India have taken an initiative and come up with numerous solutions but neither of them materialised or took momentum.

A research work carried out with title, "Detection of type 2 diabetes using clustering methods–balanced and imbalanced pima Indian extended dataset", by Nivetha, S., B. Valarmathi, K. Santhi, and T. Chellatamilan [20], proposed a method that is Dimensionality reduction and clustering technique, which happens to give the highest accuracy for the larger dataset for both balanced and imbalanced datasets. Large and small datasets have been taken for clustering using K-means approach, the Farthest first method, Density based technique, Filtered clustering method and X-means approach. K-means, density based and X-means gives the highest accuracy of 75.64%. For the larger balanced dataset when compared with the smaller balanced dataset. The Paper titled as "Clustering Algorithms in Healthcare" created by Negi, Neerja, and Geetika Chawla, state that there is a dire requirement for a mechanism that supports health researchers in producing effective healthcare policies, building drug recommendation systems, and developing health profiles of personage.

A research work carried out with the title as "Using medical data and clustering techniques for a smart healthcare system", done by Yang, Wen-Chieh, Jung-Pin Lai, Yu-Hui Liu, Ying-Lei Lin, Hung-Pin

Hou, and Ping-Feng Pai [22] showcased the designing of a smart healthcare system based on clustering techniques and medical data (SHCM) in analysing the potential risks and trends in patients under a specific duration. In order to explore results generated by the proposed SHCM system, the Evidence-based medicine was employed. Another research paper titled as "Clusterisation of Diabetes Health Indicators with K-Means Cluster Algorithm." presented by Perwira, Yuda, Wira Apriani, Ahmad Zein, and Sinta Lia Alfaris [23] wanted to show through this study that, enabling the cluster to cluster quickly s to whether a person has diabetes, or pre diabetes or is free from diabetes so that it helps to anticipate diabetes as early as possible. The data used in this study is the result of a survey from the US Behavioural Risk Factor Surveillance System (BRFSS) in 2015 which contains a net data collection of 253,680 survey responses to the CDC's 2015 BRFSS.

A next research paper titled as Abdulqadir, Hindreen Rashid, Adnan Mohsin Abdulazeez, and Dilovan Assad Zebari. "Data mining classification techniques for diabetes prediction" created by Abdulqadir, Hindreen Rashid, Adnan Mohsin Abdulazeez, and Dilovan Assad Zebari..[24], where the study aims to evaluate the efficiency of the methods used that are based on classification. The most widely employed techniques and the strategies with the best precision have also been highlighted by the researchers. Many analyses include multiple Machine Learning algorithms for various disease assessments and predictions to improve overall issues. The detection and prediction of diseases is an aspect of classification and prediction. Yet another research paper tiled as **"**An Improved Robust Fuzzy Local Information K-Means Clustering Algorithm for Diabetic Retinopathy Detection" done by Naz, Huma, Tanzila Saba, Faten S. Alamri, Ahmed S. Almasoud, and Amjad Rehman. **[25]** This particular study presents the Robust Fuzzy Local Information K-Means Clustering algorithm, an advanced iteration of the classical K-means clustering approach, integrating localised information parameters tailored to individual clusters. Comparative analysis is conducted between the performance of Robust Fuzzy Local Information K-Means Clustering and Modified Fuzzy C Means clustering, which incorporates a median adjustment parameter to augment Fuzzy C Means for diabetic retinopathy detection.

A research work carried out by Singh, Swapnil, and Vidhi Vazirani. [26], which goes under the titled of "Classification vs clustering: Ways for diabetes detection." The authors claim to have used established machine learning algorithms, namely Logistic Regression, K Nearest Neighbours, Support Vector Machine, Gaussian Naïve Bayes, Decision Tree, Random Forest, Multi-Layer Perceptron, and XGBoost. Apart from supervised learning algorithms, the unsupervised learning algorithms for classification via clustering was implemented. Provided they have also used K-means and agglomerative clustering. Preechasuk, Lukana, Naichanok Khaedon, Varisara Lapinee, Watip Tangjittipokin, Weerachai Srivanichakorn, Apiradee Sriwijitkamol, Nattachet Plengvidhya, Supawadee Likitmaskul, and Nuntakorn Thongtang. Have presented a Resaerch paper with the title, "Cluster analysis of Thai patients with newly diagnosed type 2 diabetes mellitus to predict disease progression and treatment outcomes: a prospective cohort study." [27]. In this study they have attempted to aim at categorising Thai T2D into subgroups using variables that are commonly available based on routine clinical parameters to predict disease progression and treatment outcomes and it happens to be a cohort study.

Data-driven cluster analysis was performed using a Python program in patients with newly diagnosed T2D (n=721) of the Siriraj Diabetes Registry using five variables (age, body mass index (BMI), glycated hemoglobin (HbA_{1c}), triglyceride (TG), high-density lipoprotein cholesterol (HDL-C)). Disease progression and risk of diabetic complications among clusters were compared using the X^2 and Kruskal-Wallis test. Cox regression and the Kaplan-Meier curve were used to compare the time to diabetic complications and the time to insulin initiation. "An Extended Approach to Predict Retinopathy in Diabetic Patients Using the Genetic Algorithm and Fuzzy C-Means." Is yet another Research Paper made by Ghoushchi, Saeid Jafarzadeh, Ramin Ranjbarzadeh, Amir Hussein Dadkhah, Yaghoub Pourasad, and Malika Bendechache [28] in which they have presented that the growth region algorithm for the aim of diagnosing diabetes, considering the angiography images of the patients' eyes. Provided, this study integrated two methods, including fuzzy C-means (FCM) and genetic algorithm (GA) to predict the retinopathy in diabetic patients from angiography images. The developed algorithm was applied to a total of 224 images of patients' retinopathy eyes. As clearly confirmed by the obtained results, the GA-FCM method outperformed the hand method regarding the selection of initial points.

4. Survey on the k-Means and Fuzzy C-Means Algorithms

The research work carried out by Jamel, Ahmed, and Bahriye Akay in the title, "A survey and systematic categorization of parallel k-means and fuzzy-c-means algorithms." done [29] This research study presents a survey of parallel K-means and Fuzzy-c-means clustering algorithms based on their implementations in parallel environments such as Hadoop, MapReduce, Graphical Processing Units, and multi-core systems. The Research paper which carries the title, "Selection of optimal number of clusters and centroids for k-means and fuzzy c-means clustering: A review" created by Pugazhenthi, A., and Lakshmi Sutha Kumar. [30], states that most commonly used image segmentation algorithms such as K-means and Fuzzy c-means clustering face the specific important problem in selecting the optimal number of clusters and the corresponding cluster centroids. Multiple works were presented on the limitations of the claimed clustering algorithms to improve the efficient isolation of clusters. This paper enumerates the works done by different researchers in selecting the initial number of clusters and the centroids using K-means and Fuzzy c-means clustering. A research paper titled as "Clustering of cities based on their smart performances: a comparative approach of fuzzy c-means, k-means, and k-medoids." done by Kenger, Ömer N., Zülal Kenger, Eren Özceylan, and Beata Mrugalska [31] states as its main goal of this study is to categorise cities using a scientific manner (clustering algorithms) based on SCI data and to present how the chosen approaches work for dealing with the associated problems. The primary innovation of the present study is the use of clustering techniques in reports where the indexes are used. It is known by results that using the three clustering techniques suggested here for grouping the cities on the basis of their smart indicators was much effective than others.

"Comparison of Fuzzy C-Means and K-Means Clustering Performance: An Application on Household Budget Survey Data", happens to be one of the good Research papers by Cinaroglu, Songul [32]. The comparison of FCM and k-means (KM) classification performance for the grouping of households in terms of sociodemographic and out-of-pocket (OOP) health expenditure variables is the aim of the study. Health expenditure variables have heavily skewed distributions and that the shape of the variable distribution has a measurable effect on classifiers. Incorporating Bayesian data generation procedures into the variable transformation process will increase the ability to deal with skewness and improve model performance. A another research paper titled as "Comparison of K-Means Algorithms and Fuzzy C-Means Algorithms for Clustering Customers Dataset." done by Tahyudin, Imam, Galih Firmansyah, Arul Gading Ivansyah, Windiya Ma'arifah, and Lisna Lestari. [33]. This study aims to compare the value of the validity of the K-Means and Fuzzy C-Means algorithms to determine customer clusters to identify cross selling and build a business service strategy.

The method of validity testing is the Silhouette Coefficient method which combining the Cohesion and Separation methods. The results presents that the Silhouette Coefficient validity method can determine the validity value of the K-Means and Fuzzy C-Means algorithms. The clustering results of the Fuzzy C-Means algorithm on the mall customers dataset are more accurate than the K-Means algorithm. The research work carried out "A Performance Study of Probabilistic Possibilistic Fuzzy C-Means Clustering Algorithm." done by Vijaya, J., and Hussian Syed [34]. In this paper, the proposed PPFCM clustering technique is quantitatively evaluated based on several metrics and the accuracy of the clustering outcome as well as the execution output are investigated. A comparative study is made with the proposed PPFCM clustering with the traditional clustering methods, and the results are plotted. In this work, six benchmark datasets based on different application is used for evaluating the performance of PPFCM clustering method. A next research paper titled as An improved ant-based algorithm based on heaps merging and fuzzy c-means for clustering cancer gene expression data." done by Bulut, Hasan, Aytuğ Onan, and Serdar Korukoğlu [35]. The algorithm utilizes the feature selection algorithm to overcome the high-dimensionality problem encountered in bioinformatics domain. Based on extensive empirical analysis on microarray data, clustering quality of the ant-based clustering algorithm is enhanced with the use of fuzzy c-means algorithm and heaps merging heuristic.

The performance of the proposed clustering scheme is compared with k-means, PAM algorithm, CLARA, self-organizing map, hierarchical clustering, divisive analysis clustering, self-organizing tree algorithm, hybrid hierarchical clustering, consensus clustering, Ant Class algorithm and fuzzy c-means clustering algorithms. The experimental results indicate that the proposed clustering scheme yields

better performance in clustering cancer gene expression data. A paper titled as k-means and fuzzy c-means fusion for object clustering." made by Heni, Ashraf, Imen Jdey, and Hela Ltifi [36], where this phase uses either deductive or inductive learning. Inductive learning algorithms derive a set of classification rules (or standards) from a set of already classified examples. The goal of these algorithms is to produce classification rules to predict the assignment class of a new case.

Among the available knowledge we can mention the choice of the initial cluster centers which is a very important factor in the final definition of clusters. For this purpose, we proposed evolutionary algorithm EK-means based on the combination of k-means and fuzzy c-means by touching the initialization phase of the centroids. The another paper titled as "Brain tumor segmentation using K-Means clustering and fuzzy C-Means algorithms." Done by Gayathri, K., and D. Vasanthi. in [37]. The tumor is of different types and they have different characteristics and different treatment. Normally the anatomy of the brain can be viewed by the MRI scan or CT scan. MRI scanned image is used for the entire process. The research paper titled as "Binary Classification of Rainfall Level by K-means and Fuzzy C-means Clustering." by Mittal, Amit, Sandeep Kumar Sunori, Ashish Saxena, Mehul Manu, Manoj Chandra Lohani, Somesh Sharma, and Pradeep Juneja. [38], states that, using some efficient machine learning techniques can effectively solve the classification and regression problem pertaining to rainfall. This thought has been a motivation in developing this research article. The secondary data with two attributes viz. humidity and maximum temperature has been used to train binary classification models to cluster it into 2 classes viz. 'Heavy rainfall' & 'Light rainfall'.

A research work titled as "An Improved Robust Fuzzy Local Information K-Means Clustering Algorithm for Diabetic Retinopathy Detection" made by Naz, Huma, Tanzila Saba, Faten S. Alamri, Ahmed S. Almasoud, and Amjad Rehman [39], had presented that the Robust Fuzzy Local Information K-Means Clustering algorithm, an advanced iteration of the classical K-means clustering approach, integrating localised information parameters tailored to individual clusters. A Comparative analysis was conducted between the performance of Robust Fuzzy Local Information K-Means Clustering and Modified Fuzzy C Means clustering, which incorporates a median adjustment parameter to augment Fuzzy C Means for diabetic retinopathy detection.

Ren, Jie, and Ping Zhu, came up with a Research work under the title, "Grouping fuzzy granular structures based on k-means and fuzzy c-means clustering algorithms in information granulation" [40] They have stated in the study that fuzzy information granulation, instead of discussing single fuzzy granules, it is common to consider a fuzzy granular structure arising from a set of fuzzy information granules. Different fuzzy granular structures could be generated from different approaches and perspectives in the same universe by dividing the object into a number of meaningful fuzzy information granules. However, a specific task usually requires only a selection of representative fuzzy granular structures. Therefore, the main aim of this paper is to group fuzzy granular structures efficiently and accurately. Until the known part, the distances between two fuzzy granular structures and illustrate the relevant properties, was introduced subsequently.

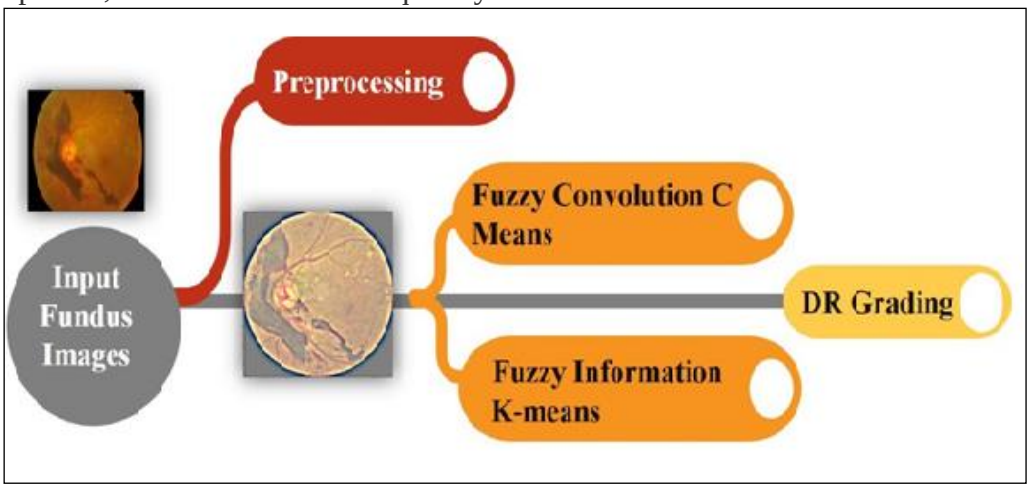

Figure 2: Proposed Model for Diabetic Retinopathy Severity Classification.

Chauhan, Rishabh, Rajan Yadav, and Rajesh Kumar Gupta made the Research work, "Comparison of Fuzzy-C-Means Method with K-Means Algorithm and Self Organizing Maps Using Image Quantization" [41] They claim to develop some quick and accurate variations of the hard and fuzzy c-

means algorithms coupled with SOM and a variety of starting strategies, and then compare the resulting quantizers on a wide range of images. In the end it became evident that the images produced by fuzzy c-means algorithm were higher than the K-means and SOM in terms of PSNR value. Due to their widespread use in this sector, along with Hierarchical clustering and Density Based-Spatial Clustering of Applications with Noise (DBSCAN), we have picked FCM, K-means, and SOM for comparison.

Table 1: Comparison of Results

Paper Ref.No.	Researcher	Methods Used	Result & Accuracy
4	Raja, J. Beschi, and S. Chenthur Pandian	PSO-FCM	PSO-FCM Provides the highest accuracy
5	Singh, Manpreet, Pankaj Bhambri, Inderjit Singh, Amit Jain, and Er Kirandeep Kaur	KNN	100% Accuracy
6	Saxena, Aditya, Megha Jain, and Prashant Shrivastava	Artificial Neural network(ANN)	Provides the highest accuracy
7	Islam, M. M., Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra	Naïve Bayes Algorithm, Logistic Regression Algorithm, and Random Forest Algorithm	Random forest has been found having best accuracy on this dataset
8	Abdulqadir, Hindreen Rashid, Adnan Mohsin Abdulazeez, and Dilovan Assad Zebari	Random Forest Classifier and Support Vector Machine	RF offers 75% greater precisions than SVM
9	Elsadek, Salma N., Lama S. Alshehri, Rawan A. Alqhatani, Zainah A. Algarni, Linda O. Elbadry, and Eyman A. Alyahyan	Naïve-Bayes Algorithm, the Logistic Regression Algorithm, and the Random Forest Algorithm	Random Forest Provided the best accuracy for this dataset when evaluated using both ten-fold Cross-validation and the percentage split method.
10	Elsadek, Salma N., Lama S. Alshehri, Rawan A. Alqhatani, Zainah A. Algarni, Linda O. Elbadry, and Eyman A. Alyahyan	Random Forest and Multi-Layer Perception(MLP)	Random Forest Produces 97.88%
11	Ibrahim, Shahira, and Siti Shaliza Mohd Khairi	Logistic Regression, Decision Tree, Naïve Bayes, and Artificial Neural Network (ANN)	Logistic Regression-Stepwise outperformed to other predictive models with classification accurateness of 73% and able to predict positive outcome correctly by 90%
14	Saeed, Saad Ebrahim, Burair Hassan Al Telaq, and Ahmed M. Zeki	KNN, Decision Tree J48	The best performance in terms of accuracy was Decision Tree J48 with an accuracy of 96% higher than both KNN93.3%
15	Anand, Tanvi, Rekha Pal, and Sanjay Kumar Dubey	K-Means clustering technique	K-Means is better than other clustering techniques
16	Gomes, João, João Lopes, Tiago Guimarães, and Manuel Filipe Santos	Machine Learning (ML) model from registered clinical information,	Classification models also point to the development of Deep Learning algorithms

		medication, Complementary diagnostic tools, and registered medication data	
17	Lomo, Simeftiany Indrilemta, and Endang Darmawan	PCA-Fuzzy C-Means Method	54.1% were male 90.6% were >= 45 years old; 66.4% has comorbidities
18	Bramesh, S. M., and Anil Kumar KM	Novel clustering Algorithm	Proposes an efficient and scalable technique to cluster the data with atomic and set-valued attributes
19	Teki, Satyanarayana Murthy, Kuncham Venkata Sriharsha, and Mohan Krishna Varma Nandimandalam	K-Means and Naïve Bayes	Clustering and classification gives promising results with an utmost accuracy rate of 99.42%
20	Nivetha, S., B. Valarmathi, K. Santhi, and T. Chellatamilan	Clustering using K-Means approach, Farthest first method, Density based technique, Filtered clustering method and X-Means approach	K-Means, density based and X-Means give the highest accuracy of 75.64%
23	Perwira, Yuda, Wira Apriani, Ahmad Zein, and Sinta Lia Alfaris	K-Means Clustering method	K-Means Clustering method where this method has been quite successful and is widely used by many researchers to cluster and predict
27	Ghoushchi, Saeid Jafarzadeh, Ramin Ranjbarzadeh, Amir Hussein Dadkhah, Yaghoub Pourasad, and Malika Bendechache	Fuzzy C-Means (FCM) and genetic algorithm	The result of the analysis demonstrated that the proposed method was efficient and effective
29	Jamel, Ahmed, and Bahriye Akay	K-Means and Fuzzy C-Means clustering algorithms	K-Means and Fuzzy C-Means algorithms are integrated with metaheuristic and other traditional algorithms produces best result
39	Naz, Huma, Tanzila Saba, Faten S. Alamri, Ahmed S. Almasoud, and Amjad Rehman	K-Means Clustering and Modified Fuzzy C- Means clustering	94.4% accuracy rate and an average execution time of 17.11 Seconds

5. Conclusion

After a comprehensive survey of existing research and an in-depth analysis of clustering algorithms applied to diabetic data, it is evident that both k-Means and Fuzzy C-Means have their merits in medical data analysis. However, Fuzzy C-Means generally provides better results in terms of accuracy and flexibility, especially when dealing with overlapping data points common in diabetic datasets. This algorithm's ability to handle the inherent uncertainty in medical data makes it a more suitable choice for diabetic disease analysis. Future research could further explore the integration of Fuzzy C-Means with other techniques to enhance diagnostic precision.

References

[1]. Abed, Amira Hassan, and Mona Nasr, "Diabetes disease detection through data mining techniques", *International Journal of Advanced Networking and Applications,* Vol. 11, Issue 1, 2019, pp. 4142-4149.
[2]. El-Sofany, Hosam, Samir A. El-Seoud, Omar H. Karam, Yasser M. Abd El-Latif, and Islam ATF Taj-Eddin, "A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App", *International Journal of Intelligent Systems,* 2024, Issue 1, pp.1-13.https://doi.org/10.1155/2024/6688934
[3]. Khan, Farrukh Aslam, Khan Zeb, Mabrook Al-Rakhami, Abdelouahid Derhab, and Syed Ahmad Chan Bukhari. "Detection and prediction of diabetes using data mining: a comprehensive review", *IEEE Issue* 9, 2021 pp. 43711-43735.
[4]. Raja, J. Beschi, and S. Chenthur Pandian. "PSO-FCM based data mining model to predict diabetic disease", *Computer Methods and Programs in Biomedicine* Issue 196, 2020 pp. 105659. https://doi.org/10.1016/j.cmpb.2020.105659
[5]. Singh, Manpreet, Pankaj Bhambri, Inderjit Singh, Amit Jain, and Er Kirandeep Kaur. "Data mining classifier for predicting diabetics.", *Annals of the Romanian Society for Cell Biology,* 2021, pp. 6702-6712.
[6]. Saxena, Aditya, Megha Jain, and Prashant Shrivastava. "Data Mining Techniques Based Diabetes Prediction", *Indian Journal of Artificial Intelligence and Neural Networking*, 2021.
[7]. Islam, M. M., Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. "Likelihood prediction of diabetes at early stage using data mining techniques", In *Computer vision and machine intelligence in medical image analysis*, pp. 113-125. Springer, Singapore, 2020.
[8]. Abdulqadir, Hindreen Rashid, Adnan Mohsin Abdulazeez, and Dilovan Assad Zebari. "Data mining classification techniques for diabetes prediction", *Qubahan Academic Journal* 1, Issue 2, 2021, pp. 125-133.
[9]. Jaber, Fayzeh Abdulkareem, and Joy Winston James. "Early prediction of diabetic using data mining", *SN Computer Science* 4, Issued. 2, 2023, pp. 169.
[10].Elsadek, Salma N., Lama S. Alshehri, Rawan A. Alqhatani, Zainah A. Algarni, Linda O. Elbadry, and Eyman A. Alyahyan. "Early Prediction of Diabetes Disease Based on Data Mining Techniques", In *Computational Intelligence in Data Science: 4th IFIP TC 12 International Conference, ICCIDS 2021, Chennai, India, March 18–20, 2021, Revised Selected Papers 4*, pp. 40-51. Springer International Publishing, 2021.
[11].Ibrahim, Shahira, and Siti Shaliza Mohd Khairi. "Predictive Data Mining Approaches for Diabetes Mellitus Type II Disease", *International Journal of Global Optimization and Its Application* 1, Issue 2, 2022, pp. 126-134.
[12].Alkattan, Hussein, S. K. Towfek, and M. Y. Shams. "Tapping into knowledge: ontological data mining approach for detecting cardiovascular disease risk causes among diabetes patients", *J. Artif. Intell. Metaheuristics* 4, Issued 1, 2023, pp. 08-15.
[13].Abdelhamid, Abdelaziz A., Marwa M. Eid, Mostafa Abotaleb, and S. K. Towfek. "Identification of cardiovascular disease risk factors among diabetes patients using ontological data mining techniques", *Journal of Artificial Intelligence and Metaheuristics* 4, Issue 2, 2023, pp. 45-53.
[14].Saeed, Saad Ebrahim, Burair Hassan Al Telaq, and Ahmed M. Zeki. "A Comparative Study on Classifying Diabetes Disease using Data Mining Models", In *2021 International Conference on Data Analytics for Business and Industry ICDABI*, pp. 438-442. IEEE, 2021.
[15].Anand, Tanvi, Rekha Pal, and Sanjay Kumar Dubey. "Cluster analysis for diabetic retinopathy prediction using data mining techniques", *International Journal of Business Information Systems* 31, Issue 3, 2019, pp.372-390.
[16].Gomes, João, João Lopes, Tiago Guimarães, and Manuel Filipe Santos. "Identifying diabetic patient profile through machine learning-based clustering analysis", *Procedia Computer Science* Issue 220, 2023, pp. 862-867.
[17].Lomo, Simeftiany Indrilemta, and Endang Darmawan. "Cluster analysis of type II Diabetes Mellitus Patients with the Fuzzy C-means method", *Annals of Mathematical Modeling* 3, Issue 1, 2023, pp. 24-31.

[18].Bramesh, S. M., and Anil Kumar KM. "An Efficient and Scalable Technique for Clustering Comorbidity Patterns of Diabetic Patients from Clinical Datasets", *International Journal of Modern Education and Computer Science* 11, Issue 6, 2022, pp. 35.
[19].Teki, Satyanarayana Murthy, Kuncham Venkata Sriharsha, and Mohan Krishna Varma Nandimandalam. "A diabetic prediction system based on mean shift clustering", *nutrition* 84, Issue S2, 2021, pp. S177-S181.
[20].Nivetha, S., B. Valarmathi, K. Santhi, and T. Chellatamilan. "Detection of type 2 diabetes using clustering methods–balanced and imbalanced pima indian extended dataset", In *Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI-2019)*, pp. 610-619. Springer International Publishing, 2020.
[21].Negi, Neerja, and Geetika Chawla. "Clustering Algorithms in Healthcare", In *Intelligent healthcare: Applications of ai in ehealth*, pp. 211-224. Cham: Springer International Publishing, 2021.
[22].Yang, Wen-Chieh, Jung-Pin Lai, Yu-Hui Liu, Ying-Lei Lin, Hung-Pin Hou, and Ping-Feng Pai. "Using medical data and clustering techniques for a smart healthcare system", *Electronics* 13, Issue 1, 2023, pp. 140.
[23].Perwira, Yuda, Wira Apriani, Ahmad Zein, and Sinta Lia Alfaris. "CLUSTERIZATION OF DIABETES HEALTH INDICATORS WITH K-MEANS CLUSTER ALGORITHM", *INFOKUM* 10, Issue 03, 2022, pp. 64-70.
[24].Abdulqadir, Hindreen Rashid, Adnan Mohsin Abdulazeez, and Dilovan Assad Zebari. "Data mining classification techniques for diabetes prediction", *Qubahan Academic Journal* 1, Issued. 2, 2021, pp. 125-133.
[25].Naz, Huma, Tanzila Saba, Faten S. Alamri, Ahmed S. Almasoud, and Amjad Rehman. "An Improved Robust Fuzzy Local Information K-Means Clustering Algorithm for Diabetic Retinopathy Detection", *IEEE Access* 2024.
[26]. "Cluster analysis of Thai patients with newly diagnosed type 2 diabetes mellitus to predict disease progression and treatment outcomes: a prospective cohort study", *BMJ Open Diabetes Research and Care* 10, Issue 6, 2022, pp.e003145.
[27].Ghoushchi, Saeid Jafarzadeh, Ramin Ranjbarzadeh, Amir Hussein Dadkhah, Yaghoub Pourasad, and Malika Bendechache. "An Extended Approach to Predict Retinopathy in Diabetic Patients Using the Genetic Algorithm and Fuzzy C-Means", *BioMed Research International* 2021, Issue 1, 2021, pp. 5597222.
[28].Singh, Swapnil, and Vidhi Vazirani. "Classification vs clustering: Ways for diabetes detection", In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pp. 1-8, IEEE, 2022.
[29].Jamel, Ahmed, and Bahriye Akay. "A survey and systematic categorization of parallel k-means and fuzzy-c-means algorithms", *Computer Systems Science and Engineering* 34, Issue 5, 2019, pp. 259-281.
[30].Pugazhenthi, A., and Lakshmi Sutha Kumar. "Selection of optimal number of clusters and centroids for k-means and fuzzy c-means clustering: A review", In *2020 5th International conference on computing, communication and security (ICCCS)*, pp. 1-4. IEEE, 2020.
[31].Kenger, Ömer N., Zülal Kenger, Eren Özceylan, and Beata Mrugalska. "Clustering of cities based on their smart performances: a comparative approach of fuzzy c-means, k-means, and k-medoids", *IEEE Access* 2023.
[32].Cinaroglu, Songul. "Comparison of Fuzzy C-Means and K-Means Clustering Performance: An Application on Household Budget Survey Data", In *Intelligent and Fuzzy Techniques: Smart and Innovative Solutions: Proceedings of the INFUS 2020 Conference, Istanbul, Turkey, July 21-23, 2020*, pp. 54-62. Springer International Publishing, 2021.
[33].Tahyudin, Imam, Galih Firmansyah, Arul Gading Ivansyah, Windiya Ma'arifah, and Lisna Lestari. "Comparison of K-Means Algorithms and Fuzzy C-Means Algorithms for Clustering Customers Dataset", In *2022 1st International Conference on Smart Technology, Applied Informatics, and Engineering (APICS)*, pp. 211-216. IEEE, 2022.
[34].Vijaya, J., and Hussian Syed. "A Performance Study of Probabilistic Possibilistic Fuzzy C-Means Clustering Algorithm." In *Advances in Computing and Data Sciences: 5th International Conference,*

ICACDS 2021, Nashik, India, April 23–24, 2021, Revised Selected Papers, Issue 5, pp. 431-442. Springer International Publishing, 2021.
[35].Bulut, Hasan, Aytuğ Onan, and Serdar Korukoğlu. "An improved ant-based algorithm based on heaps merging and fuzzy c-means for clustering cancer gene expression data." *Sādhanā* 45, Isuue 1, 2020, pp. 160.
[36].Heni, Ashraf, Imen Jdey, and Hela Ltifi. "k-means and fuzzy c-means fusion for object clustering", In *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, Issue 1, pp. 177-182. IEEE, 2022.
[37].Gayathri, K., and D. Vasanthi. "Brain tumor segmentation using K-Means clustering and fuzzy C-Means algorithms", *Int J Sci Res Comput Sci Eng Inf Technol IJSRCSEIT* 2, Isdue 2, 2017.
[38].Mittal, Amit, Sandeep Kumar Sunori, Ashish Saxena, Mehul Manu, Manoj Chandra Lohani, Somesh Sharma, and Pradeep Juneja. "Binary Classification of Rainfall Level by K-means and Fuzzy C-means Clustering", In *2022 3rd International Conference for Emerging Technology (INCET)*, pp. 1-5. IEEE, 2022.
[39].Naz, Huma, Tanzila Saba, Faten S. Alamri, Ahmed S. Almasoud, and Amjad Rehman. "An Improved Robust Fuzzy Local Information K-Means Clustering Algorithm for Diabetic Retinopathy Detection." *IEEE Access* 2024.
[40].Ren, Jie, and Ping Zhu. "Grouping fuzzy granular structures based on k-means and fuzzy c-means clustering algorithms in information granulation." *Iranian Journal of Fuzzy Systems* 20, Issue 5, 2023, pp. 9-31.
[41].Chauhan, Rishabh, Rajan Yadav, and Rajesh Kumar Gupta. "Comparison of Fuzzy-C-Means Method with K-Means Algorithm and Self Organizing Maps Using Image Quantization", In *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pp. 397-402. IEEE, 2023.