International Journal of Industrial Engineering
2022;6(2):34-40.
ISSN: 2456-8449

A Novel Data Mining Based Approach for Healthcare Applications

G. Sudha[1*], M. Birunda[1], J. Gnanasoundharam[2], J. Alphas Jeba Singh[1]

[1]*Department of Biomedical Engineering, Muthayammal Engineering College,*
Namakkal, Tamil Nadu, India.
[2]*Department of Electronics and Instrumentation Engineering, St. Joseph's College of Engineering,*
Chennai, Tamil Nadu, India.

[*]*Corresponding author: kavisudhamariyaa@gmail.com*

Abstract. Data mining has been bolstered by a massive increase in the usage of data analysis in all sectors. The greatest significant influence on a person's quality of life today is access to quality healthcare. A patient's health might be adversely affected by sudden changes in short-term patterns. It is extremely difficult to interpret the large amounts of data created by health institutions to make significant decisions about patient health. Personal healthcare is out of reach due to the stress of the job. Clustering techniques like K-means and D-stream are used to determine whether a person is fit or unfit based on their historical and real-time data. Both clustering techniques are used on the biological history database of the patient in question. We tested both algorithms using real-world biological data to ensure they were accurate. The D-stream method, a density-based clustering technique, addresses the K-means algorithm's shortcomings. Finally, we can determine the efficacy and efficiency of both algorithms by computing their performance metrics.

Keywords: K-means clustering, ensemble classification, healthcare data, feature extraction, data mining.

INTRODUCTION

With the goal of extracting useful information from massive data sets, data mining is a critically important and highly stimulating area of study. As a result, Data Mining is growing increasingly popular in healthcare as a means of uncovering previously undiscovered and important information in healthcare data. It is possible to detect health insurance fraud, give patients with affordable medical solutions, discover illness causes, and discover new treatment options through the application of Data Mining techniques in the healthcare business [1].

Another benefit is to researchers in the field of health care who may use it to produce more effective healthcare policies as well as drug recommendation systems and personal health profiles for individuals themselves. It is extremely difficult to interpret the large amounts of data created by health institutions to make significant decisions about patient health. In this database, you'll find information about medical facilities, patients, insurance claims, and more [2]. As a result, a robust tool for assessing and extrapolating essential information from this complicated data is required. Patient management duties are improved through the examination of health data. Data Mining technology can help healthcare organisations group individuals with similar illnesses or health conditions so that they can give them with more effective therapy. Patients' duration of stay, medical diagnosis, and successful information system management may all be predicted using this tool [3].

A variety of cutting-edge medical technologies are being employed to improve patient care while also reducing overall costs. For example, data mining techniques may also be used to study the numerous elements that contribute to illnesses, such as the type of food people eat and the working environment they have, as well as the availability of clean drinking water and health care. As a stream of data, it is possible to think of it as a continuous and ever-changing flow of information [4]. There are several traditional OLAP and mining approaches that are infeasible for real - time stream applications since they need several data scans.

Received: 15.04.2022 Revised: 12.05.2022 Accepted: 28.05.2022

Because data streams may be generated in so many different domains, it's critical to adapt mining methods to work with them. Data stream mining offers a wide range of uses and is a popular research topic. Continuous data mining is the process of finding patterns and models in seemingly endless streams of data to extract knowledge structures hidden within them. Clusters of medical heath data may be formed using these data stream mining techniques. K-means and density-based clustering are two of the key clustering techniques described in this work [5].

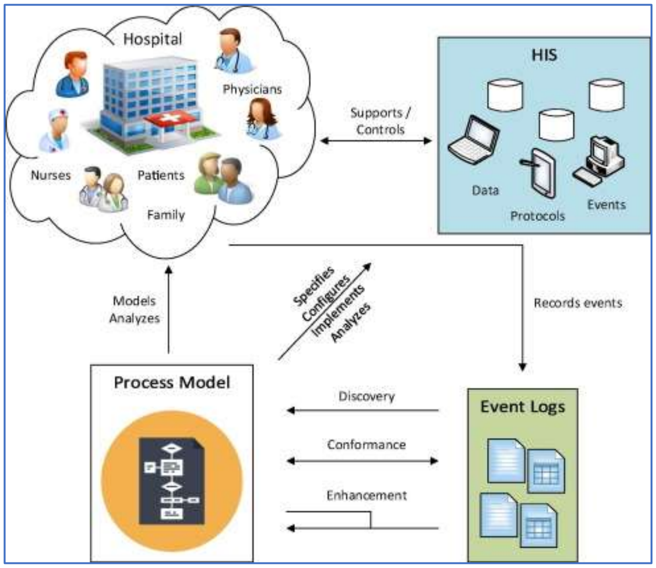

FIGURE 1. Healthcare Applications using Data Mining

Because it cannot handle outliers or find clusters of any forms, the K-means clustering method fails miserably. k and a user-specified time frame are also required. D Stream, a methodology for classifying stream data that used a density-based technique, was developed to overcome these challenges. Online and offline components are used in the method to translate each input record into a grid and to calculate the grid density, which is then used to group the grids depending on the density. Density decaying techniques are employed by the algorithm to record real-time data changes.

Our system generates and adjusts clusters in real time by taking advantage of the complex interactions between decay factor, data density, and cluster structure. An efficient method for removing irregular grids projected to by outliers has also been created. High-speed clustering is possible without sacrificing the quality of the resulting clusters [6]. The results of the experiments reveal that our algorithm is superior in terms of quality and efficiency, and it can find clusters of any shape, as well as properly recognising the changing behaviours of real-time data.

EXISTING WORK

The new power comes from data. Data mining has been bolstered by a massive increase in the usage of data analysis in all sectors. Data mining may be used to identify patterns and variations in patients, identifying trends in Twitter, car wrecks, image recognition, neutralise, trend analysis in genomics farming, identity verification, and many other areas of interest. All Several university and institution-based health care initiatives are under underway with the goal of better serving the needs of the elderly [7].

Faster retrieval of necessary information from medical databases can be achieved using data clustering. It is a real-time smart watch for monitoring, visualising, and analysing physiological information, Health Gear. A blood oximeter was used to assess the user's blood oxygen concentration and pulse as they slept with this system. Two separate algorithms are described in the system, and the entire system's performance is shown in sleep research with twenty subjects [8].

Knowledge Discovery Using Guided Clustering — A Classic Example of Liver Disease Clustering data mining was used to uncover similarities in the dataset of patients with liver disorders [9]. K-means method and SOM parameters are used to identify patterns in the dataset. Clustering approaches can yield relevant results if a domain expert specifies the input restrictions to the algorithm, according to study [10]. For the Intelligent Mobile Health Monitoring System (IMHMS), a patient's biological and environmental data is gathered by deployed sensors and used to deliver medical feedback via mobile devices [11].

Using the Wearable Wireless Body/Personal Area Network, the system collects patient data, mines the data, intelligently forecasts the patient's health state, and offers feedback to the patient via their mobile devices [12]. Patients will be able to view their medical records from any location at any time thanks to their mobile devices. However, the deployment of a data mining architecture for a decision support system has not been completed. For medical applications, a flexible framework was proposed to do proper analysis of clinical information to evaluate people's health state in real time [13].

Data mining techniques are used to create patient- or disease-specific models. Real time categorization of physiological data and continuous assessment of a person's health state are achieved using models. An immediate appraisal and stream analysis of physiological data are both possible in the proposed system [14]. It isn't possible to study the dynamic behaviour of physiological signals in this model since the framework doesn't function with ECG signals.

Heart attack forecast data was utilised as a framework to determine the effectiveness of clustering algorithms. Prediction accuracy and cluster assignment visualisation highlight the relationship between error and attribute in the result. The classifier algorithm's performance is also shown. Clusters based on density are the most accurate predictors, according to the results of the comparison [15-16]. Different image processing applications for healthcare are discussed in [17-18].

PROPOSED SYSTEM

We've developed a method for efficiently grouping datasets retrieved from medical databases. Clustering the patient records into distinct groups based on the qualities of the test results may assist physicians identify the patient's ailment in an efficient manner, and the assessment processes are listed below. Oxygen saturation, ABPsys, ABPdias, Tachycardia, genetics, overweight, and cigarette smoking are all part of the dataset. Patients' health state can be predicted with the use of risk factors such as these.

SpO2, ABPsys, ABPdias, and HR may all be gleaned from the MIMIC database, while the person's actions have an impact on the other characteristics. The values of these properties are all discrete. You may expect the data to be in pre-processed form. Data features will be retrieved from the available data and then clustering algorithms will be used to the features extracted during the model building phase. The following properties are extracted for each signal of interest x between those X observed vital signs. It is the difference between x(t) and the linear interpolation (i.e., mean value throughout the time span) that is measured by the offset feature.

Comparing today's values against those of prior years is what this study focuses on. Signal change may be measured using a slope function. A patient's health might be adversely affected by sudden changes in short-term patterns. Current signal measurement drifts from a given normalcy range is measured by the distance feature When the measurement falls inside the usual range, it is zero. As the signal's distance from normalcy range increases, so does its danger level. If a patient has an immediate measurement that is outside of the normal range, it may not be essential, but their/her persistence in these settings increases the danger.

Risk functions and overall risk elements will be computed from the aforesaid risk components. Clustering algorithms will make use of these variables as an input for the construction of clusters. K-means and D-stream algorithms are used in the system's planned flow. The attributes listed above will be used to compare two clustering techniques. Classification is the process of grouping data samples together according to a predetermined set of criteria. Each data point is predicted to belong to a certain class using the classification method. A patient's illness pattern, for example, might be used to classify them as "high risk" or "low risk" patients using a data classification technique. Class categories are known in advance; hence this is a form of supervised learning. Classification may be done in two ways: binary or multilayer. As opposed to binary categorization, the multiclass method considers more than two alternative classifications such as "high," "medium," and "low" risk patients. There are two sets of data: one for training and the other for testing.

The classifier was developed using a training dataset. It is possible to verify the accuracy of the classifier by utilising a test dataset. Data mining techniques such as classification are commonly employed in healthcare organisations. For the analysis of microarray data, Hu et al. utilised a variety of classification methods, including decision trees, SVM, and an ensemble approach. The root node refers to the node at the very top of the tree, where all the labels are connected.

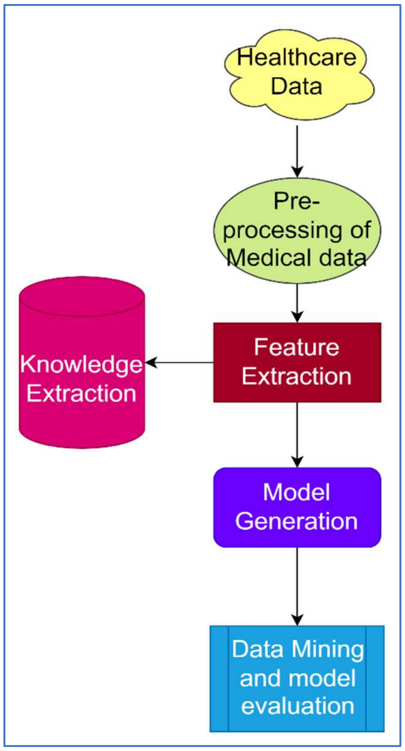

FIGURE 2. Proposed System Architecture

For instance, a financial institution's decision tree determines whether a loan should be granted. It is not necessary to have any prior understanding of the issue area to arrive at a choice. Tree-like graphs are used to build Decision Trees, a type of classifier. Decision Trees are most commonly used in production research study to calculate the probability of a certain outcome.

RESULTS AND DISCUSSION

Clusters are formed on a medical database using the above-described procedures. A open dataset from open. swiss has been used for the data analysis. In the dataset, there are 107 cases and 14 attributes. Age, gender, Hypertension, Lipid, Difficulty Breathing, etc. are some of the characteristics. Those systems will be evaluated based on the accuracy of their predictions. It is possible to attain three different levels of categorization accuracy: 71-80%, 81-90%, and 90-100%. Category A includes just Nave Bayes, whereas category B includes Decision tree and Naïve Bayes approach. The performance from Different classifiers in terms of MAE, RMSE, RAE and RRSE values of the constructed classifiers in this study are summarised below Table 1.

TABLE 1. Classifier Performance Comparison

Classifier	Kappa Statistics	MAE	RMSE	RAE	RRSE
Ada Boost M1	0	0.1482	0.2425	211.21 %	130.95 %
Decision Table	0.0594	0.0705	0.1736	100.47 %	93.78 %
J-Rip	0.3306	0.0558	0.1671	79.55 %	90.24 %
J48	0	0.0685	0.1851	97.66 %	99.98 %
Lazy IBK	0.9715	0.0049	0.0406	7.01 %	21.95 %
Lazy K-Star	0.9715	0.0042	0.0412	6.02 %	22.24 %

The graphic indicates the PRC Area, Roc Curve, MCC, F-Measure, Recollect, Accuracy, Prediction Accuracy, TP Rate parameter of all the classifications in the table. A further conclusion to be drawn from this study is the fact that no one data mining approach provides consistent findings for all forms of healthcare data. Data mining algorithms perform better or worse depending on the nature of information we use in our experiments. If you're looking to get the most out of your data mining efforts, you may combine techniques like clustering and classification or correlation with clustering and classification, for example, to get better results.

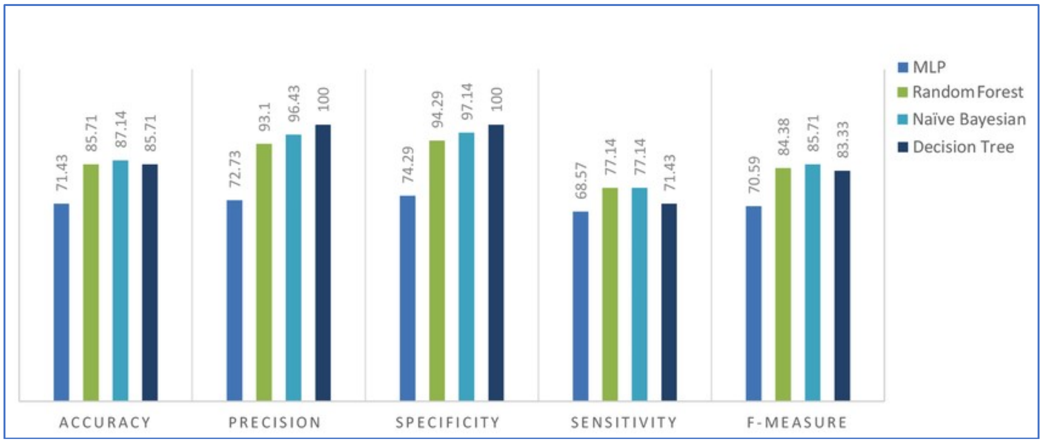

FIGURE 3. Performance Analysis of Algorithms

In Figure 3, we have shown the performance analysis of algorithms. While this isn't a complete list, we've found that clustering or classification using decision tree, random forest and naïve Bayesian approaches have achieved better results than single standard approaches. So, if you want greater outcomes, hybridization is a viable choice. This study examines the use of data mining in medical institutions, as well as the many methodologies and obstacles that Data Mining in medicine faces. A wide range of people, from doctors and insurers to consumers and healthcare organisations, can profit from the use of data mining. Using data mining information, doctors can readily identify the most effective remedy, patients can get cost-efficient treatments, the healthcare sector can better manage its customers, and healthcare insurance can uncover any occurrences of medical claim fraud. Analysis and descriptive abilities make Data Mining a popular tool in the healthcare industry. The error rate analysis has been shown in Figure 4.

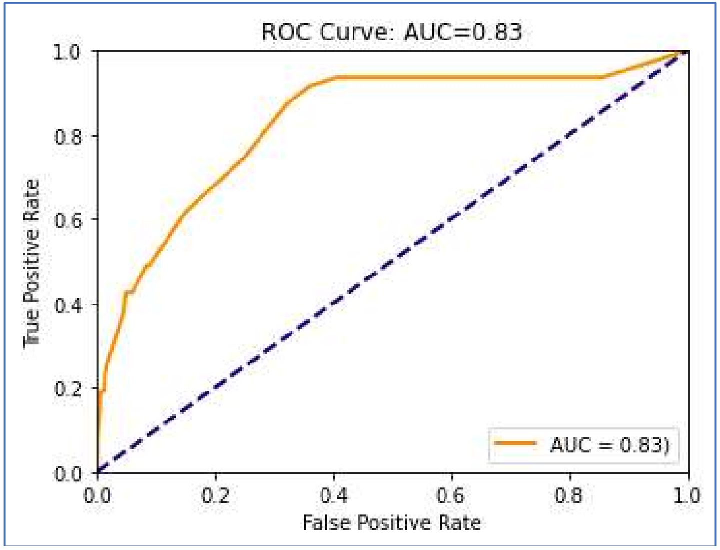

FIGURE 4. Error Rate Analysis

Medical professionals use data mining techniques to make informed decisions about how to improve patient health, reduce the cost of health care, and identify health insurance fraud, among other things. Using Data Mining in the medical area also comes with several problems for healthcare researchers. For example, some Data Mining algorithms demand settings from the user. The user's parameters are considered while using these strategies. According to the settings that are set by the user, the outcomes might be somewhat different the selection and use of parameters might be difficult for certain individuals, especially if they lack the necessary knowledge. A sample medical dataset has been shown in Table 2.

TABLE 2. Sample Medical Dataset

Age	Heart Rate	Blood pressure	Heart Problem
65	78	150/70	Yes
37	83	112/76	No
71	67	108/65	No

Enhanced and safe exchange of health data is necessary for the efficient use of data mining in health organisations. To get around the security concerns, researchers and health care providers must enter into contractual arrangements that place restrictions on their ability to share information. A systematic method to building the data warehouse is also required. There has been a rise in the availability of large datasets (both text and non-textual) on the internet in recent years. For this reason, sophisticated data mining techniques are also critical in order to reveal hidden information from this data.

CONCLUSIONS

Because the number of classes to be produced cannot be predicted with certainty, K-means cannot handle arbitrary cluster formation. Superior in quality and efficiency, the D-stream algorithm finds clusters of arbitrary forms and effectively recognises the changing behaviours of real-time large datasets. As a result, D-stream will be a better fit in the biological field. Biomedical data may be analysed and used to forecast the present health condition of patients using this approach. Patients in ICUs and the elderly can both benefit from the system's planned application. Also, the system may be utilised by professionals to maintain track of patients' health statuses. The adaptive nature of the proposed system is demonstrated by the fact that it can process several physiological signals. The algorithms such as K-means, D-stream can be used to predict a patient's present health state by examining past biological data. Because D-stream allows for clustering, it can better predict a person's health than K-means, which does not. Users with little or no knowledge of the application data can also benefit from D-stream because it does not need the use of K-values. Using historical data, D-stream is better than K-means in creating clusters, thanks to its parameter-free nature.

REFERENCES

[1]. K. P. Bennett and O. L. Mangasarian, 1992, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization methods and software*, **1(1)**, pp. 23-34.

[2]. O. L. Mangasarian and W. H. Wolberg, 1990, "Cancer diagnosis via linear programming," *University of Wisconsin-Madison Department of Computer Sciences*, pp. 1-8

[3]. W. H. Wolberg and O. Mangasarian, T. F. Coleman, Y. Li, 1990, "Pattern recognition via linear programming: Theory and application to medical diagnosis," *Large-Scale Numerical Optimization, SIAM Publications*, Citeseer, pp. 22-30.

[4]. U. Ojha and S. Goel, 2017, "A study on prediction of breast cancer recurrence using data mining techniques," *7th Int. Conf. on Cloud Computing, Data Science & Eng.-Confluence* 2017 pp. 527-530.

[5]. L. R. Borges, 1989, "Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection," *Group*. **1(369),** pp. 15-19.

[6]. S. Saxena and K. Burse, 2012, "A survey on neural network techniques for classification of breast cancer data," *Int. J. of Eng. and Advanced Techn.*, **2(1)**, pp. 234-237.

[7]. UCI Machine Learning Repository: Breast Cancer Wisconsin Dataset, https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29

[8]. D. Verma and N. Mishra, 2017, "Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques," *Int. Conf. on Intelligent Sustainable Systems (ICISS)* pp. 533-538.

[9]. W. H. Wolberg and O. L. Mangasarian, 1990, "Multi surface method of pattern separation for medical diagnosis applied to breast cytology," *Proc. of the National Academy of Sciences*, **87(23),** pp. 9193-9196.

[10]. Y. Chen and L. Tu, 2007, "Density-based clustering for real-time stream data," *Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining,* pp. 133-142.

[11]. N. Oliver and F. Flores-Mangas, 2006, "HealthGear: a real-time wearable system for monitoring and analyzing physiological signals," *Int. Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)* pp. 1-4

[12]. V. Singh and S. Nagpal, 2010, "A guided clustering technique for knowledge discovery–a case study of liver disorder dataset," *Int. J. of Computing and Business Res.*, **1(1)** pp. 78-85.

[13]. M. M. Uba, R. Jiadong, M. N. Sohail, M. Irshad and K. Yu, 2019, "Data mining process for predicting diabetes mellitus-based model about other chronic diseases: A case study of the northwestern part of Nigeria," *Healthcare Tech. Letters*, **6(4)** pp. 98-102.

[14]. P. P. Jayaraman, A. R. Forkan, A. Morshed, P. D. Haghighi and Y. B. Kang, 2020, "Healthcare 4.0: A review of frontiers in digital health," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **10(2)** pp. 1-23.

[15]. A. Pika, M. T. Wynn, S. Budiono, A. H. Ter Hofstede, W. M. van der Aalst and H. A. Reijers, 2020, "Privacy-preserving process mining in healthcare," *Int. J. of Environmental Res. and Public Health*, **17(5)**, pp.1-28.

[16]. F Joseph, and S Murugan, 2018, "Hybrid windowing adaptive FIR filter technique in underwater communication," *Int. J. of MC Square Scientific Res.* **10(2),** pp. 17-21.

[17]. Ahmed AK. Tahir and S. Anghelus, 2022, "Improving Iris Recognition Accuracy Using Gabor Kernels with Near-Horizontal Orientations," *Int. J. Adv. Sig. Img. Sci*, **8(1)**, pp. 25–39.

[18]. A. Hussaindeen, S. Iqbal and T. D. Ambegoda, 2022, "Multi-Label Prototype Based Interpretable Machine Learning for Melanoma Detection," *Int. J. Adv. Sig. Img. Sci*, **8(1)**, pp. 40–53.