IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 09, Issue 1 (January. 2019), ||V (III) || PP 40-47

An Effective early stage Disease Prediction Model Using Enhanced Principle Component Analysis and MDRP Algorithm

Basheer. P
FDP Substitute lecturer, Department of Computer Science, KAHM Unity Women's College, Manjeri, Malappuram, Kerala, India[1]
Corresponding Author: Basheer. P

Abstract: Data mining on medical data has great potential to improve the treatment quality of hospitals and increase the survival rate of patients. Data mining techniques are used for a variety of applications. In healthcare industry, data mining plays an important role in early stage of the diseases prediction .It is very helpful to the patients and hospital if the possibility of disease is identified in early stages based on data collected from patients through various methods (either conducting interview or data from various medical tests) By applying different data mining algorithms. So doctors can easily anticipate the future diseases of patients, can give advice or prescribe medicine. As a result patients get cure in early stage of the disease. The main drawbacks of the previous studies are that need accurate and more number of features. The study proposed a Data mining model has been developed using improvised Principle Component Analysis to improve the early stage Disease prediction. The system implements a new IPCA and MDRP algorithm for effective prediction of disease. This also creates a new advanced latent factor model to reconstruct the missing data for fast and accurate disease classification. From the experimental results, early stage Disease Prediction of our proposed algorithm reaches 97.8% accuracy which is faster than the existing system.

Keywords: Data Mining, Machine Learning, Healthcare, Disease prediction, Medical Data Analytics.

Date of Submission: 07-01-2019 Date of acceptance: 22-01-2019

I. INTRODUCTION

Data mining on medical data has great potential to improve the treatment quality of hospitals and increase the survival rate of patients. Data mining lies at the interface of statistics, database technology, pattern recognition, machine learning, data visualization, and expert systems. A database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. Databases contain aggregations of data records or files, and a database manager provides users the capabilities of controlling read and write access, specifying report generation, and analyzing use. Databases and database managers are prevalent in large mainframe systems, but are also present in smaller systems and on personal computers. Databases usually include a query facility, and the database community has a tendency to view data mining methods as more complicated types of database queries. The healthcare data is outspread among to involving several parts of medical systems, healthcare parts, and government hospitals with the advantage of a data mining and a greater awareness is paid to the Disease Prediction. The number of investigations has been managed to selecting the attribute of a disease prediction from a great volume of a data. The greatest amount of the existing work is based on a structured data. The unstructured solitary data can use a convolution neural network. Convolution neural network are make a nerve cell, each nerve cell collect some information entered system and execute operations and the full network indicates a single differentiable result functions. The precise of a disease prediction can be decreased because there is an additional difference in a various related disease because of the weather conditions prevailing in an area in general and living habits of climate the peoples in their particular regions. To reduce this difficulty to combines both the structured and unstructured data. To correctly predict the disease control the problem of a losing and insufficient data. The Big data technology can be used latent characteristic model. In the prior effort only structured data can be used but for the perfect results. In this system can use the unstructured data. In this technique can choose characteristic spontaneously using CNN algorithm. In this proposed system can base on *CNN-MDRP algorithm* to apply the data types. Disease Predictions for structured data use to conventional machine learning algorithm i.e., Naïve Bayesian.

A data mining is an instance of revolving system it is under way in health care. It begins to a very great extent increased to providing the information. Extending the last decade, pharmaceutical companies have been collecting years of investigations and implementation of data into medical databases, while providers have convert their patient data. In similar, modern technical advances have made it easier to receive and inspect

information from lot of sources. In this system, a solution, which considers patient–hospital mutual preference, is provided to guide the appointment scheduling process by means of schedule defragmentation along with the disease diagnosis option. And the healthcare community can use machine learning algorithm for more perfect results.

II. PROBLEM DEFINITION

In this paper[1] Yi Mao, Yixin Chen, Gregory Hackmann has define the topic for "Medical Data Mining for Early Deterioration Warning in General Hospital Wards ". In this paper author explains Data mining on medical data has great potential to improve the treatment quality of hospitals and increase the survival rate of patients. Early prediction techniques have become an apparent need in much clinical area. Clinical study has found early detection and intervention to be essential for preventing clinical deterioration in patients at general hospital units. An early warning system (EWS) designed to identify the signs of clinical deterioration and provide early warning for serious clinical events.

In this paper [2] Asha Unnikrishnan and Senthil Kumar B has presented, Biomedical information extraction is always a difficult process due to its huge set of results and unknown keywords. Some medical terminologies are not aware by the users; often it's very difficult to retrieve such content from the website like PubMed. Several work used concept hierarchies for easy search navigation, however the biomedical query retrieval processing time always higher than the normal data retrieval. Ranking, summarizing and categorization have been proposed together to improve the searching efficiency. Results categorization and result summarization based on the concept hierarchy for biomedical databases is the focus of this work. A natural way to organize biomedical citations is according to their keywords and tags, additionally the conceptual similarity can also be performed in the proposed system. In this paper, a new BioSearch engine is proposed with effective data mining algorithms with less energy for query processing. The proposed system contains Predictive data caching technique for fast data retrieval; this has been performed with the help of facet order and concept hierarchy methods. The proposed system also provides the auto query incremental algorithm to ease the search. Finally the retrieved data's are ranked and summarized using RII (Ranked Inverted Index) algorithm. This helps to summarize and simplify the result to the user view.

This paper [3] Neesha Jothia, Nur'Aini Abdul Rashidb has presentedThe knowledge discovery in database (KDD) is alarmed with development of methods and techniques for making use of data. One of the most important step of the KDD is the data mining. Data mining is the process of pattern discovery and extraction where huge amount of data is involved. Both the data mining and healthcare industry have emerged some of reliable early detection systems and other various healthcare related systems from the clinical and diagnosis data. In regard to this emerge, we have reviewed the various paper involved in this field in terms of method, algorithms and results. This review paper has consolidated the papers reviewed inline to the disciplines, model, tasks and methods. Results and evaluation methods are discussed for selected papers and a summary of the finding is presented to conclude the paper.

In the paper [4] Senthil Kumar B and Bavitha Varma E has presented text categorization is an important and well-studied area of pattern recognition, with a variety of modern applications. Effective spam email filtering systems, automated document organization and management, and improved information retrieval systems all benefit from techniques within this field. The problem of feature selection, or choosing the most relevant features out of what can be an incredibly large set of data, is particularly important for accurate text categorization. The proposed systems (i) use well known pre-processing method porter and Lancaster for train the dataset. (ii) A number of feature selection metrics have been explored in text categorization, among which information gain (IG), chi-square (CHI), Mutual information (MI), Ng-Goh-Low (NGL), Galavotti-Sebastiani-Simi (GSS), Relevancy Score (RS), Multi-Sets of Features (MSF) Document frequency (DF) and odds ratios (OR) are considered most effective. Pruning techniques are also proposed using ignore the feature based on TF and DF to further reduce the set of possible features (typically words) within a document prior to applying a method of feature selection. (iii) Finally classify the selected feature based on two algorithm KNN and Navie bayes. Two benchmark collections were chosen as the testbeds: Reuters-21578 and small portion of Reuters Corpus Version 1 (RCV1). There are two classifiers and both data collections, and that a further increase in performance is obtained by combining uncorrelated and high-performing feature selection methods.

The Paper [5] J.Sasitha Burvin, Dr.K.Dhanalakshmi presents General health examination is an integral part of healthcare in many countries. Identifying the patients at risk is important for early warning and preventive intervention. Data mining is a well- known technique used by health organizations for classification of diseases. Dengue is a fast emerging pandemic-prone viral disease in many parts of the world. It should also be noted that dengue can be co-morbidity with other disorders. It can also be detected in the patients with chronic disease. Hereby using Fever based dataset to be load and finding the patients affected by dengue fever prediction by machine learning. Association rule mining algorithm useful for analyzing, predicting patient's behaviors. In the proposed work, a graph-based decision tree is proposed to predict the dengue fever earlier and

reduce the mortality rate and classify different activities of patients in more accurate manner. The patients affected with dengue fever are divided into those who are affected with dengue fever, or dengue fever with chronic disease warning signs. Based on the complication of the patient disease, prioritize the patients so that they will get effective treatment in timely and accurate manner.

In this paper [6] Parvez Ahmad and Saqib Qamar Data mining is gaining popularity in disparate research fields due to its boundless applications and approaches to mine the data in an appropriate manner. Owing to the changes, the current world acquiring, it is one of the optimal approach for approximating the nearby future consequences. Along with advanced researches in healthcare monstrous of data are available, but the main difficulty is how to cultivate the existing information into a useful practices. To unfold this hurdle the concept of data mining is the best suited. Data mining have a great potential to enable healthcare systems to use data more efficiently and effectively. Hence, it improves care and reduces costs. This paper reviews various Data Mining techniques such as classification, clustering, association, regression in health domain. It also highlights applications, challenges and future work of Data Mining in healthcare.

In this paper [7] Shakuntala Jatav have gives the contents based on" An Algorithm for Predictive Data Mining Approach in Medical Diagnosis". It contains large and composite data is required extremely interesting pattern of diseases & different types of machine learning techniques are used to makes perfect decisions. For the medical research the advanced data mining techniques are used to discover knowledge in database. This paper has to be used to predict the disease for Kidney, Diabetes, and Liver disease using lot of input functionalities. The data mining categorization techniques, such as Support Vector Machine (SVM) and Random Forest (RF) are examined on Diabetes, Kidney and Liver disease datasets. The presentation of these techniques is interconnected, based on precision, accuracy, recall, f measure as well as time.

In this paper [8] Suresh P has given the concept for "Study and Analysis of Prediction Model for Heart Disease: An Optimization Approach using Genetic Algorithm". In the medical field the heart disease diagnosis is the biggest task. The grouping large clinical and pathological data heart disease diagnosis is very difficult. In heart disease prediction the complication, increased amount and clinical professionals about the good and perfect. Machine learning is used to provide the efficient support for predicting heart disease with perfect case of training and testing. The important of this work is to study diverse prediction models for the heart disease and choosing important heart disease feature using genetic algorithm.

In This paper [9] Shirsath concept of big data is provides the beneficial merits like, accurate medical data analysis before the disease prediction of perfect data can be securely stored and used. And the accuracy of a medical data analysis can reduced the incomplete medical data; the genetic disease can break for using the medical data prediction. In the purpose of disease prediction can collect the hospital data in the particular region. The misplaced data can be used inactive factor model to aim the incomplete data. In medical data prediction have three types such as: a) Hospital data, b) Structured data, c) Unstructured Data. a) Hospital data: The hospital data is a highest volume of datasets of a patient can be given by a hospital and the data can be saved in the data centre to protect the patient privacy and security of saved data, can be create a security access mechanism. b) Structured data: The structured data is nothing but the scientific experiments data, patient's basic information like patient's age, life habits, gender, weight , height etc. c) Unstructured Data :Unstructured Data is a information of patients medical history, patients weakness, and doctors interrogation and diagnosis**.**

In this paper [10] Sreekanth Rallapalli, Suryakanthi T has provide the "Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm". In this concept explains Electronic Health Care records can be stored in an enterprise database or cloud databases. In this data can be efficiently processed. The predictive analysis helps the doctors, physicians to identify the patient's admission details or other information. The main challenging task is identifying the strong indicators for an accurate disease prediction.

III. PROPOSED SYSTEM

The term discusses details and brief explanation about the proposed methodology and the steps involved in that proposed system. The optimized IPCA and MDRP (*Multimodal Disease Risk Prediction*) algorithm has been expanded with the new optimal classification algorithms, which can handle large category dataset more rapidly, accurately and effectively, and keeps like good scalability at the same time. This term discuss about the algorithms and methodologies.

Contributions of the proposed System:

The followings are the contributions of the proposed system.

- The system implements a new improvised and MDRP algorithm for effective prediction of disease.
- This also creates a new advanced latent factor model to reconstruct the missing data for fast and accurate disease classification. The system developed with the intension of high accuracy and less training overhead.

- So the system initially collects and make score for every label, this partially makes an ensemble approach to improve the detection speed.
- EPCA for feature selection and dimensionality reduction

A Proposed model is generated or selected to predict the best possibility of an outcome.

IV. METHODOLOGIES

Latent factor model

Data preprocessing is a data mining real world data is often incomplete, inconsistent, or lacking in certain behaviors or trends, and is likely to contain many errors in that data. Data preprocessing is one of the proven methods of resolving such issues, problems and so on. Data preprocessing prepares raw data for further processing. Proposed used latent factor model to reconstruct the missing data from the medical records collected from a hospital in different Medical unit.

IPCA (improvised Principle Component Analysis)

Perhaps the most widely used algorithm for manifold learning is EPCA. The proposed system utilizes a model for disease classification and prediction. It is a combination of Principal component analysis and the non linear trick. EPCA begins by computing the covariance matrix of the $m \times n$ matrix $\mathbf{X}$WPCA steps:

The Step By Step Approach For EPCA

1. Taking the entire dataset ignoring the initial class labels
2. Find initial and starting component
3. Compute the *d*-dimensional mean vector values continuously Compute the covariance matrix of the original or standardized d-dimensional dataset X (here: d=3); alternatively, compute the correlation matrix values effectively.
4. Compute the eigenvectors and eigen values of the covariance matrix (or correlation matrix).
5. Sort the values in descending /ascending order.
6. Choose the k vectors that correspond to the largest values where the number of dimensions of the new feature subspace ($k \leq d$).
7. Construct the projection matrix W from the k selected eigenvectors.
8. Transform the original dataset X to obtain the k dimensional feature subspace Y ($Y=WT \cdot X$).

The important process of disease Classification and prediction is the analysis of patterns and grouping those into different subset.

Predictive model

A predictive model analysis in data mining is a procedure by which a model is generated or selected to predict the best likelihood of an outcome. In certain scenarios, the model is selected on the basis of detection theory to guess the probability of an outcome given with a group of input data. For example, given patients data, ranging from disease attributes values and classes which decide how likely a data classified. A predictive analysis in data mining, such as classification, starts from a given classification of the data items. From that it derives a situation based on the properties of the data objects that permit to predict the association to a specific class. For example, the prediction could be based on a partitioning of the attribute values along with each measurement. Predictive data mining comprises of combining the predicted classifications from different models, or from similar type of models for the purpose of different learning data. At the same time, predictive data mining is also used to tackle the intrinsic volatility of outcome when applying composite models to compare small data sets. Suppose, if the task of data mining is to construct a model for classification of predictive types of data, and the data set that is involved in mining is

relatively small and then the data and predictive data mining is the most common type of data mining procedure.

The proposed system performs the prediction model based on the medical dataset. The proposed system successfully analyses the prediction based on the given training dataset. The system also predicts the score for the chance based on the prediction. The proposed system implements a semi supervised classifier which does not depends on the training dataset completely. The system performs the statistical properties to estimate the score of every attribute. The system finally provides the prediction accuracy over the given dataset.

Data category	Item	Description
	Patient Demographics	The Details such as Patient's gender, age, height, weight, etc.
Structured data	**Living habits**	The patient smokes, has a genetic history, etc. So, Whether these details are exposed.
	Diseases	Patient's disease, such as cerebral infarction, etc
Unstructured text data	**Patient's readme ill health**	Patient's readme sickness and medical history

Table 4.1: Item Taxonomy In Hospital Data

MDRP (Multimodal Disease Risk Prediction)

Proposed this model effectively use the text data to predict whereas the patient is at high risk or not. The output value is C, which indicates whether the patient is amongst the high-risk population.
C0 indicates the patient is at high-risk.
C1 indicates the patient is at low-risk.
The Structured data (S-data): Uses the patient's structured data then to predict whether the patient is at high-risk.
Text data (T-data): Uses the patient's unstructured text data then to predict whereas the patient is at high-risk
Step 1: deftrain_nn_SGD (nn_structure, X, y, iter_num=3000,):
Step 2: W, b= setup_and_init_weights (nn_structure)
Cnt = 0
m = len(y)
Step 3:avg_cost_func = []
Iterations'.format (iter_num))
Step 4: while cnt<iter_num:
If cnt%50 == 0:
For i in range (len(y)):
Delta = {}
Step 5: for l in range (len (nn_structure), 0, -1):
If l == len (nn_structure):
Delta[l] =calculate_out_layer_delta(y[i,:], h[l], z[l])
avg_ += np.linalg.norm ((y[i,:]-h[l]))
Step 6: else:
If l > 1:
Delta[l] = calculate_hidden_delta (delta[l+1], W[l], z[l])
tri_b[l] = delta [l+1]
# complete the average calculation
Avg = 1.0/m * avg
avg_ _func.append (avg_value)
C += 1
Return
The output value is C, which indicates whether the patient is high-risk or Not.

V. RESULT AND ANALYSIS

Experimental Results

This section describes the implementation process. Implementation is the realization of an application, or execution of plan, idea, model, design of a research. This section explains the software, datasets and modules which are used to develop the research. Then experimental term is performed on an Intel I3 Processor with a RAM capacity 4GB. The algorithms are implemented in Dot net and are run under Windows platform.

Data Set:

The data used in this study contains real time hospital data, and the data store in database. The dataset is general composed of structured and unstructured text data. The structured data is which includes the laboratory data and the patient's basic information such as the patient's age, gender and life habits, and then etc. Whereas, the unstructured text data includes the patient's narration of his/her illness, the doctor's interrogation records and diagnosis, etc. Patient database is collected from Disease Dataset (DD) available on the UCI Repository. The attributes considered are age: age, sex, height, weight (resting blood pressure), chol (cholesterol

in mg/dl), FBS (fasting blood sugar > 120 mg/dl), smokes, blood and disease info. There are a total of 500 patient records in the databaset.

Metrics	**Dataset**	**Existing**	**Proposed System**
	DS1(50)	95	99
Detection Accuracy (%)	**DS2(100)**	93	98.8
	DS3(120)	93	98.5
	DS4(150)	90	98

Table 5.1 Performance Evaluation

Data Preprocessing

This phase includes extraction of data from disease Dataset in a uniform format. The step involves transforming the data, which involves removal of missing fields, normalization of data, and removal of anomalies, which refers not important data. Out of the 500 available records, 25 tuples have missing attributes. These have been excluded from the data set. For proposed system, data points were automatically centered at their mean and scaled to have unit standard deviation. No changes need be made to the data sets for EPCA.

Results And Analysis

The experiments are designed so that the different parts of the work could be evaluated. These include the evaluation of the features of the above dataset, the feature selection and also the feature creation methods. To this aim, first the features which were selected by the feature selection method named as EPCA and their importance are discussed. Second, all the four possible combinations of the feature selection and creation methods are theoretically analyzed over the dataset. Finally the performance of this proposed work Scheme was compared with the existing algorithms based on the following parameters.

- ***Accuracy*** *– Determines the correctness*
- ***Precision*** *–Repeated process same result*
- ***Time taken*** *– Determines the processing time involved.*

TP, TN, FN and FP these terms are described by Sensitivity, specificity and accuracy.

A. Accuracy:

The accuracy is measured by following formula this measured in terms of percentage.

Accuracy = (TN + TP)/ (TN+TP+FN+FP) (Number of correct evaluations)/Number of all evaluations)

B. Precision:

A class is a number of true positives (i.e. the number of instances correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class) that is called Precision.

The equation is:

$$\text{Precision} = TP / (TP + FP)$$

C. ***Time taken***

This Determines the processing time involved for completed entire data set prediction process.

This experiment has been done through the hospital Dataset. The dataset is preprocessed by latent factor model and features are selected effectively and finally the prediction process is made by MDRP.

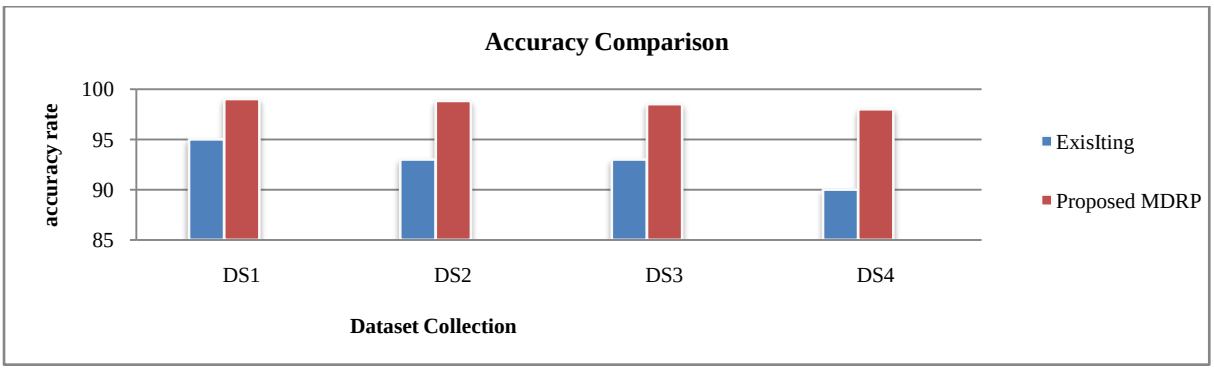

Figure 5.1 Accuracy Comparison

Performance comparison of proposed system with existing approaches based On Disease prediction Result accuracy

From the results shown in the graphs, it can be observed that the proposed MDRP based approaches provides better accuracy and increased true positive rate when it is analyzed with different number of datasets. The system finally performs the analysis to show the accuracy of the proposed system.

Precision Comparison Chart:

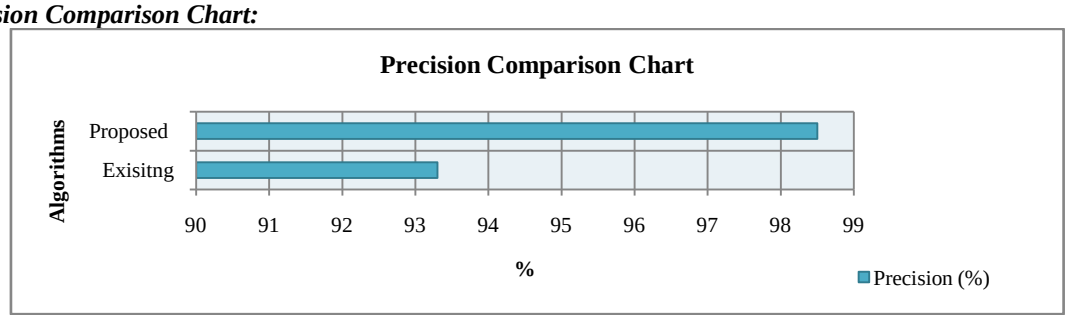

Figure 5.2 Precision comparisons Graph between existing and proposed

Execution Time comparison chart:

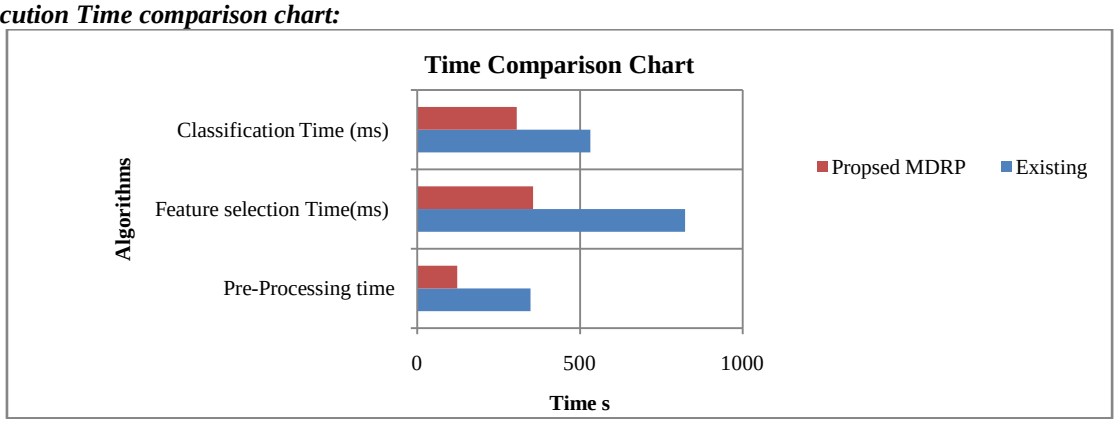

Figure 5.3 Time comparison between existing and proposed

This graph observed that the performance is very promising compared to the existing methods that have been explored in the previous term. The next term deals with the presentation of the conclusion and enhancements.

VI. CONCLUSION

The study and research proposed a new classification and prediction scheme for hospital medical disease data. The system studied the main two problems in the literature survey, which are prediction accuracy and classification delay. The study overcomes the above two problem by applying the effective enhanced MDRP. The system effectively identifies and prediction the disease, the sub type which is referred as the percentage of class such as normal and disease. The experimental result shows that integrated extended proposed algorithm shows better quality assessment compared to traditional research techniques. From the experimental results, prediction accuracy of our proposed algorithm reaches 98.8% with a convergence speed which is faster than the existing system. As further work, improvements can easily be done since the coding is mainly structured or modular in nature. In the system can changing the existing modules or adding new modules can append improvements. Further enhancements can be made to the application by expanding the existing modules future research may use the model to identify the existing area of research in the field of data mining in other dataset and use of other classification algorithms. As further work, use this model as a functional base to develop an appropriate data mining system for classification performance.

REFERENCES

[1]. Yi Mao, Yixin Chen, Gregory Hackmann, "Medical Data Mining for Early DeteriorationWarning in General HospitalWards". IEEE, 2011.

[2]. Asha Unnikrishnan, Senthil Kumar. B. "Biosearch: A Domain Specific Energy Efficient Query Processing and Search Optimization in Healthcare Search Engine". Journal of Network Communications and Emerging Technologies. Volume 8, Issue 1, January (2018).

[3]. Neesha Jothi, Nur'Aini Abdul Rashid. "Data Mining in Healthcare – A Review". Organizing Committee Of Information Systems International Conference. Volume 72, 306 – 313, 2015.
[4]. Senthil Kumar B, Bhavitha Varma E. "A Survey on Text Categorization". International Journal of Advanced Research in Computer and Communication Engineering. Vol. 5, Issue 8, August 2016.
[5]. J.Sasitha Burvin, Dr.K.Dhanalakshmi. "Pandemic Disease Detection And Prevention System Using Mining With Graph-Based Approach". International Journal of Pure and Applied Mathematics. Volume 118 No. 20 2018.
[6]. Parvez Ahmad, Saqib Qamar. "A Techniques of Data Mining In Healthcare: A Review". International Journal of Computer Applications. Volume 120 – No.15, June 2015.
[7]. Shakuntala Jatav And Vivek Sharma. "An Algorithm For Predictive Data Mining Approach In The Medical Diagnosis". International Journal Of Computer Science & Information Technology (Ijcsit) Vol 10, No 1, February 2018.
[8]. Suresh P And 2m Ananda Raj D. "Study Of Prediction And Analysis Of Prediction Model For Heart Disease: An Optimization Approach Using The Effective Genetic Algorithm". International Journal Of Pure And Applied Mathematics Volume 119 No. 16 2018, 5323-5336.
[9]. Shraddha Subhash Shirsath, Prof. Shubhangi Patil. "Disease Prediction Using The Machine Learning Over Big Data". International Journal Of Innovative Research In Science, Engineering And Technology. Vol. 7, Issue 6, June 2018.
[10]. Sreekanth Rallapalli, Suryakanthi T. "Predicting The Risk Of Diabetes In Big Data Electronic Health Records By Using The Scalable Random Forest Classification Algorithm". International Conference On Advances In Computing And Communication Engineering (Icacce). Doi: 10.1109/Icacce.2016.8073762(2016).

Basheer. P. "An Effective early stage Disease Prediction Model Using Enhanced Principle Component Analysis and MDRP Algorithm." IOSR Journal of Engineering (IOSRJEN), vol. 09, no. 01, 2019, pp. 40-47.