Medical Data Mining Techniques for Health Care Systems

R. Naveen Kumar ¹, M. Anand Kumar ²
Research Scholar¹, Associate Professor²
Department of Computer Science¹, Department of Information Technology²
Karpagam University, Coimbatore, Tamil Nadu, India

Abstract:
Due to the sequence in the information technology, the prevalence of the healthcare organizations conserves their data electronically. Enormous progress in medical data leads to be scarce in the mining of well-informed in series from the mass data. There is a necessity for accomplished analysis tools to resolve covered relatives and desire in data. Data mining can represent new biomedical and healthcare details for clinical preference. The relationship comes together from the data of patients composed in database aid in the resultant progression. This inspection investigates the worth of a variety of Data Mining techniques such as classification, clustering, association, regression in health domain. This paper work presents a brief preamble of the techniques that are currently used in medical field followed by virtues and demerit of the existing techniques. This survey also highlights applications, challenges and future issues of Data Mining in healthcare. Finally, this work provides the recommendations for appropriate selection of available Data Mining techniques.

Keywords: Association, Classification, Clustering, Data Mining, Decision making and Healthcare.

I. INTRODUCTION
Medical data mining has been an enormous latent process for exploring veiled patterns in data sets of medical sphere. In healthcare, despite the fact that data mining is not broadly used, its reputation now increasingly accepted in the medical datasets for its earlier discovery progression. Data mining can perk up Decision-making by discovering patterns and trends in large amounts of multifaceted data. There are two chief goals of data mining-prediction and portrayal. Prediction involves some variables or fields in the data set to envisage mysterious or future values of other variables of curiosity. On the other hand narrative focuses on verdict patterns recitation of the data that can be interpreted by humans. The data generated by the health organizations is very immense and multifaceted due to which it is complicated to investigate the data, in order to make significant announcement concerning patient's health. This data contains details regarding hospitals, patients, medical assert, cure cost etc. So, there is a must to create a commanding tool for scrutinizing and extracting significant information from this intricate data. The analysis of health data improves the healthcare by enhancing the concert of patient organization tasks. The results created by Data Mining technologies improve the progression of predicting the comparable patients and clustering them under a challenging group based on illness or fitness issues, so that healthcare involvement offers them effectual treatments. It can also be constructive for forecasting the span of life of patients in hospital, for medical investigation and constructing plan for effectual information system management. Current technologies are used in medical field to augment the medical services in cost valuable manner. Data Mining techniques are also used to investigate the assortment of factors that are accountable for diseases such as food type, diverse working situation, learning level, livelihood conditions, accessibility of pure water, health care services, artistic, ecological and farming factors[1].

The rest of this paper is organized as follows: Section 2 - reviews the related work on data mining techniques on Parkinson’s disease; Section 3 - presents the finding from the survey of Heart disease Diagnosis-A and Brest Cancer Diagnosis-B using mining techniques; Section 4 - discusses the survey on Diabetes Disease Diagnosis using data mining technique; Section 5 - presents the summary of the problem; Section 6 - presents the statement of the problem; Finally, Section 7 - concludes the work presented in this paper.

II. DATA MINING FOR DISEASE DIAGNOSIS
With the fast increase in population, there is a substantial quantity of augmentation in the health linked diseases. Numerous diseases are strongly associated with a symptom which makes it complicated for the doctors to forecast the precise diseases on one go. This is where data mining appears into backing; it helps in forecasting the disease which is almost perfect. Even-though the forecasting is not extremely accurate, it gives the doctor a concise idea what the disease might be. Thus, Data mining is not a substitution to doctors as an alternative, it is a accolade to them to envisage the diseases in advance stages. [2][3]

A. PARKINSON’S DISEASE:
In the Parkinson’s illness is a neurodegenerative ailment, the malady impinges on by brain cells (neurons) in human brain. The neurons make a chief chemical called dopamine. The dopamine send signal to the fraction of brain that pedals travel. The petite signals can aid those parts of the brain work enhanced. The decrease of dopamine in the brain makes the person immobile. The four types of symptoms of
Parkinson’s bug are: tremor, rigidity, Bradykinesia and postural volatility.

When quiver occurs, it pulse by hands, arms, legs or jaws. The sign of firmness makes limbs and trunk rigid. Bradykinesia is a sign which leads to sluggish travels. Postural volatility causes gloominess and poignant changes. The basic indicator affects 75-90% of citizens with Parkinson’s bug. The work in [4] obtained through tell-monitoring dataset, by UCI Irvine machine erudition repository. The dataset comprises two object curriculums (i.e) Motor-UPDRS and Total-UPDRS. This vocation only concentrates on attribute relevancy and not on idleness, and the time intricacy is highly compared.

Peyman Mohammadi et al [5] obtained their revise on medical verdict by separating 11 data mining algorithms into five groups, which are practical to a dataset of patient’s irrefutable variables data with Parkinson’s bug (PD), to study the ailment string. The dataset includes 22 properties of 42 citizens, that all of our algorithms are functional to this dataset. The verdict table with 0.9985 association coefficients has the finest accurateness and Decision Stump with 0.7919 correlation coefficients has the least accuracy. The work of Hariganesh S et al [6] discusses the Parkinson’s disease of remote tracking used by eleven techniques with 4406 training dataset and 1469 data of test set. The elevated accuracy of correlation coefficient in dataset is 99.85% and 99.67 % produced by M5Rules algorithm.

The research work in [6] developed the voice dimensions of disease chiefly spotlights the speech signals [6]. The Parkinson dataset is a series of biomedical voice measurements from 31 people, 23 trait features in Parkinson’s disease. The error rate of confusion matrix of 2*2 matrix is the output. The major objective is to get the lowest amount of error rate with the minimum characteristic of Parkinson’s dataset. The random tree gives 100% accuracy with zero error rates. Shianghan et al [7] offered three models to investigate the Parkinson’s disease for error probability premeditated by logistic regression analysis, decision tree analysis and NN analysis. Error probability of 5.15% is produced by logistic regression and neural network exhibits 23.73%. At last, the neural net analysis holds the highest error probability and is concluded as the best analysis in Parkinson’s disease.

The work [8] was presented in the speech of vocal sound examination for the Parkinson’s disease patients to be evaluated by the health control (HC) people. The speech was assessed for four features like NHR, SPLD, RFPC and F0 SD. The disease pretentious person speaks through microphone. The voice test for vocal task will be performed by speech subsystem measures calculated by correct assessment rate. The categorization accurateness is almost 85%.

The work [9] was urbanized to be evaluated for the performance of Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The ANN and SVM are designed by metrics like accurate, truly positive, false positive, positive predictive value and negative predictive value. The authors in [10] utilized ANN which has two type of MLP. They are: back- propagation learning and Radial Basis Function classifier, to produce the highest accuracy of data set. The MLP network has accuracy output with 93.22% and RBF has accuracy output with 86.44%. In [11] the authors proposed a speech recognized model by Mel-frequency cepstral coefficients (MFCC) and Vector Quantization (VQ). The MFCC is developed through speech analysis using signals and frequent domain. Code book is calculated using VQ. 20 phonations are used for normal speech and patient with Parkinson’s disease. Finally, VQ consequence with the codebook in normal speech of rate is 90% and patient with PD is 95%.

In [12] it presents the boosting committee, a machine which has been developed. Boosting is a kind of filtering technique. It is used for the neural networks with back propagation. They mainly work on voting for the proposal. They are the three types of experts in the preparation of set. The bulk appointments amplify the recital tempo in excessive data set. The highest positive rate in NN is 74%. The obtainable work chiefly concentrates on the discovering of regulations from the obtainable set of patterns but fall short to grip the data preprocessing specifically.

III. A - SURVEY ON HEART DISEASE DIAGNOSIS USING MINING TECHNIQUES

This piece presents the literature survey associated to heart bug dataset forecast by means of data mining techniques for judgment & achieved dissimilar probabilities for diverse methods as discussed below.

- A gifted Heart bug prophecy System (IHDDS) urbanized by via data mining techniques. Immature Bayes, Neural Network, and conclusion Trees was projected in the vocation. The effort has its own might to get suitable fallout. To erect this system veiled patterns and affiliation were used. It is web-based, user friendly & stretchy.

- To widen the multi-parametric trait with linear and nonlinear (HRV - Heart Rate Variability) a narrative practice was projected by authors [15]. To attain this, they have used classifiers like Bayesian Classifiers, Classification based on Multiple Association Rules, C4.5 and Support Vector Machine.

- The guess of Heart bug, Blood heaviness and Sugar with the support of neural networks was projected by the novelist of [14]. The dataset contains records with 13 attributes in each trace. The supervised networks i.e. Neural Network with back broadcast algorithm is used for preparation and difficult data.

- The crisis of identifying unnatural organization regulations for heart bug guess was deliberate by Carlos


http://ijesc.org/
Three constraints were introduced to diminish the number of patterns. They are as follows: 1. the attributes have to emerge on only one side of the regulation. 2. Split the attributes into groups. i.e. tedious groups. 3. In a decree, there must be limited number of attributes. The consequence of this is two groups of rules, the occurrence or deficiency of heart bug.

- The vocation in [17] builds a verdict tree with database of enduring for a medical problem

**B. SURVEY ON BREAST CANCER DIAGNOSIS**

The work [39] had discussed that arithmetic neural networks can be used to execute breast cancer decision effectively. The researcher has compared algebraic neural network with Multi Layer Perceptron on WBCD database. Radial basis function, universal decline Neural Network and Probabilistic Neural Network were used for organization and their on the whole routine were 96.18% for Radial Basis Function, 97% for PNN, 98.8% for GRNN and 95.74% for MLP. This work used statistical neural network structures applied to diagnose breast cancer to improve the detection rate considerably. Xin Yao [40] et al. had attempted to execute neural network for breast cancer diagnosis. Negative correlation training algorithm was used to molder a trouble routinely and resolve them. This article has discussed two approaches such as evolutionary approach and ensemble approach, in which evolutionary approach can be used to devise compact neural network involuntarily.

The ensemble approach was aimed to undertake bulky problems but it was in advancement. The works [38] have done an examination on data mining techniques for gene selection classification. This article is in conformity with the majority used data mining techniques for gene selection and cancer classification. Mainly they have four major rising fields as Neural network, Machine learning, genetic and cluster based techniques which are specified for the future improvement.

The work [41] have presented an article on an intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers with Bayesian classifiers and ANN for prognosis and diagnosis of breast cancer disease. Wisconsin diagnostic breast cancer datasets were used to employ SVM model to offer division between the malignant and benign breast masses. These datasets involve quantity taken according to Fine Needle Aspirates. An inflammation felt during the assessment approximately gives evidences to the size of tumour and its texture. In research work [19] a comparison is performed on neural network, MLP, PNN, SOM and RBF to classify the presence of heart disease using WBC and NHBCD data. According to the work RBF had produced its best only in training set whereas PNN had proved its best performance both in training and testing set. The work demonstrated that statistical neural networks can be efficiently used for breast cancer diagnosis. By pertaining several neural network structures a diagnostic scheme was performed fairly.

The work [20] explored the applicability of decision trees for the exposure of high-risk breast cancer groups. The dataset produced by the Department of Genetics of faculty of Medical Sciences with 164 controls and 94 cases. To statistically legalize the association found, incarnation tests were used. They establish a high-risk breast cancer group unruffled of 13 cases and only 1 control, with a Fisher Exact Test which is used for validation value of 9.7*10^-6 and a p-value of 0.017. These consequences showed that it is probable to discover statistically important relations with breast cancer by obtaining a decision tree and picking the finest folio. The work [21] examined the ability of the classification SVM with Tree Boost and Tree Forest. In analyzing the DDSM dataset, the mining mammographic accumulation features categorize true and false cases. Increasing diagnostic accuracy is shown promising by SVM witnessed by the largest area under the ROC curve comparable to values for tree boost and tree forest.

The work [22] explored that the genetic algorithm model yielded better results for the detection of breast cancer patient classification along with the expression and complexity of the classification rule. WBC dataset is used for analyze and it is performed by artificial neural network, decision tree, logistic regression, and genetic algorithm. The result reveals that the accuracy and positive predictive value of each algorithm were used as the evaluation metric and finally genetic algorithm produce more acceptable accuracy. The work [23] constructed classification rules use the Particle Swarm Optimization Algorithm for breast cancer datasets. This paper handles the feature subset selection as a preprocessing step which infers the rule using fuzzy logic with genetic algorithm and the selected features dataset are used for classification rules developed by PCO and the result shows the rate of accuracy defining the underlying attributes effectively.

The research work [24] performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. When comparing RBF and MLP neural network models, it was found that RBF had good predictive capabilities and also the time taken by RBF was less. The hierarchical granulation based rough set theory was proposed in the work [25]. This method adopted to determine the rules with minimal attributes using lower approximations of rough set theory. The simulation is done on WBC dataset and the result analysis showed that the proposed work produced effective result on classification.

The paper [26] presented a novel approach by performing reduced dimensionality and normalizing of the WBC data with 360 instances. The preprocessed data is then observed by a rough set method for generating classification rules. To optimize the generated rules the rough set reduction technique was applied to find all reducts of the data which contains the minimal subset of attributes and determined
optimizing 30 rules out of 472. The paper [27] applied SVM and ANN on the WBC data. The results of SVM has 97% accuracy while comparing ANN prediction models which hold 89. This work mainly contributes to the decision making for biopsy.

The work [28] presented advancement for early breast cancer diagnosis by integrating ANN and multiwavelet based sub band image decomposition. The MIAS dataset was used for testing the proposed approach was tested using the MIAS mammographic databases and the result shows that the multiwavelet performs better with areas ranging around 0.96 under ROC curve. The proposed approach could aid the radiologists in mammogram psychoanalysis and analytic verdict making.

IV. DIABETES DISEASE DIAGNOSIS USING DATA MINING TECHNIQUE
A good number of researches have been reported in literature on diagnosis of diabetes disease diagnosis. This section deals with discussion of some related work in diabetes disease diagnosis using data mining technique. The work in [29] proposed by the author based on neuropathy diabetics which is a kind of nerve disorder due to diabetic mellitus. This is mainly affected by the long term diabetics. This paper deals with the symptoms and risk factors of neuropathy taken into the consideration for deployment of fuzzy based relation equation. It is linked with the Multilayer perceptron in composition of binary relation using fuzzy inference model.

The authors of [30] devised an automatic retinal diabetic detection using a multilevel perceptron neural network. To evaluate the optimal global threshold to minimize the pixel classification errors, the network is trained by this algorithm. The performance of the proposed work improves by the detection and adequate index based on neuro fuzzy subsystem. The fuzzy set and linguistic variable are used in [31] to diagnose diabetes. This paper used the maximum and minimum relationship to deal with uncertainty availed in the dataset. Data sets of forty patients were collected to produce this relationship.

In paper [32] presented a binary classification based neural network techniques used diabetes diagnosis problem. The assessment is done based on three benchmark data sets obtained from UCI machine learning repository.

In the work [33] fuzzy membership function is used in connection with fuzzy neural network to detect the diabetes in early stages. This work uses two experimental examinations for medical data. In the paper [34] developed a substitute pruning technique based on the Minimal Description Length principle. It can be viewed as substitutions between theory complexity and data prediction accuracy. This work proposed a greedy search algorithm to prune the fuzzy ARTMAP categories one by one. The results proved that fuzzy ARTMAP pruned with the MDL principle gave improved rectal with less categories shaped compared to original fuzzy ARTMAP and other machine learning systems.

The authors in [35] proposed a fuzzy clustering technique which determined the number of appropriate clusters based on the prototype spirit. Diverse experiments for algorithm evaluation were executed, which showed an improved presentation to the distinctive widely used K-means clustering algorithm and data was taken from the UCI Machine Learning Repository [36]. In the work [37] proposed a new SSVM for classification of diabetes patient problems was considered. It is called Multiple Knot Spline SSVM (MKS-SSVM). To assess the usefulness of their method, they carried out a research on Pima Indian diabetes dataset. First, hypothetical of MKS-SSVM was obtainable. Then, application of MKS-SSVM and assessment with SSVM in diabetes disease diagnosis was agreed. The results of this study demonstrated that the MKS-SSVM was effectual to perceive diabetes disease diagnosis.

V. SUMMARY
The survey reveals that there are lots of existing techniques available for detection of disease diagnosis by performing classification or clustering techniques. But still there is a low concentration on the quality of the dataset, when it is incomplete. There are very few papers which work on data preprocessing, feature selection and reduction.

VI. STATEMENT OF THE PROBLEM
The existing techniques used for disease diagnosis are still suffering from false detection rate, due to the raw nature of dataset. The insufficient and incomplete dataset may lead to false alarms in diagnosis of accurate results. The preprocessing of dataset is less concentrated in the existing work, which leads to major setback in the overall process. The future work of this survey will start with a data preprocessing technique to produce a complete dataset instead of using raw dataset.

VII. CONCLUSION
Over the past few decades, the automated problem-solving tools have been intended to assist the physician with a clear sense of medical data. In healthcare, data mining is becoming increasingly more essential. From the above study it is examined that the support vector machine due to its regularization parameter is often recommended by a lot of the researches to evade over-fitting. With the help of its kernel trick it can box to build an expert inference system. In clustering models k-means based forecasting and variants of it also produce promising results. In evolutionary based models, particle swarm optimization with its variants contributed more innovatively in many disease diagnosis works. It also assumes the real number code, and it is determined honestly by the solution. The selection of data mining approaches depends on the nature of the dataset. If the dataset consists of the labeled features then the classification techniques can be suggested for best prediction. If the dataset is with unlabelled features then the clustering techniques are most excellently suited for pattern
recognition. If the optimization of the results needs to be improvised, bio inspirational based techniques are best suited. Still the diagnosis of disease suffers from high false alarm and detection rate is low. In the future work it is planned to propose a novel approach to reduce the false alarm rate in the situation of incomplete dataset handling.

VIII. REFERENCES:


[3] Rahul Isola, Student Member, IEEE Rebeck Carvalho, Student Member, IEEE, Amiya Kumar Tripathy, Member, IEEE Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and k-NN


[12] Mehmet can “Boosting committee Machines to Detect the Parkinson’s disease by Neural Networks”.


[22] Chang Pin Wei and Liou Ming Der, “Comparision of three Data Mining techniques with Genetic Algorithm in analysis of Breast Cancer data”.[Online].


[38] “Multicategory Classification Using Support Vector Machine for Microarray Gene Expression Cancer Diagnosis” Dr.S.Santhosh Baboo1, Mrs.S.Sasikala 2 38 Vol.10 Issue 15 (Ver.1.0) December 2010


[40] Xin Yao, Yong Liu., “Neural Networks for Breast Cancer Diagnosis” 0-7803-5536 - 1999 IEEE.