International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(1), pp.114-121
DOI: http://dx.doi.org/10.21172/1.81.015
e-ISSN:2278-621X

A NOVEL DETECTION FOR KIDNEY DISEASE USING IMPROVED SUPPORT VECTOR MACHINE

Jyoti Saini[1], R.C Gangwar[2] and Mohit Marwaha[3]

Abstract- Data mining is a process to analyze the number of data sets and extracts the meaning of data. Data mining provides methods and techniques for transformation of the data into useful information for decision making. These techniques can make process fast and take less time to predict the Kidney disease with more accuracy. The healthcare sector assembles enormous quantity of healthcare data which cannot be mined to uncover hidden information for effectual decision making. It becomes more influential in case of kidney disease that is considered as the predominant reason behind death all over the world. In medical field, Data Mining provides various techniques and has been widely used in clinical decision support systems that are useful for predicting and diagnosis of various diseases. These data mining techniques can be used in kidney diseases takes less time and make the process much faster for the prediction system to predict diseases with good accuracy to improve their health. Here one or more (KNN, Naive Bayes, SVM, ISVM) algorithms of data mining will be used for the prediction of kidney disease.
By applying these data mining techniques to kidney disease data which requires to be processed, produces effective results and achieve reliable performance which will help in decision making in healthcare industry. It will help the medical practitioners to diagnose the disease in less time and predict probable complications well in advance.

Keywords - Data Mining, Classification technique, Kidney disease, Dataset, Healthcare, Support vector machine, nural network, Discriminant analysis.'

I. INTRODUCTION

Data mining is an analytical process to investigate definite data from huge volume of data. It is a process that finds previously unknown patterns and trends in databases. This information is further used to construct predictive models. Large amount of data which is generated for the prediction of kidney disease is analyzed traditionally and is too complicated and voluminous to be processed. Data mining is effectively related with data science that involves classification and manipulation of data by applying mathematical and statistical concepts. Data mining is a significant phase in discovery of knowledge and comprises application of discovery and analytical techniques on data to create particular models across data. Data exists everywhere. It can be used to expect the future. Generally the statistical approach is applied. Data mining is an addition of established data examination and statistical approaches in that it includes analytical method drawn from a collection of disciplines. Due to the extensive accessibility of vast,

[1] *Post-Graduate Student, Computer Sc. & Engg, IKG Punjab Technical University, Kapurthala(pb) India*
[2] *Associate Professor, Department of , Computer Sc,Beant college of Engg & Tech, Gurdaspur(Pb) India*
[3] *Associate Professor, Department of , Computer Sc, Beant college of Engg & Tech, Gurdaspur(Pb) India*

complex, information-rich data sets, the capability to extract valuable knowledge concealed in these data and to be active on that knowledge has become progressively more significant in today's competitive world. Consequently data mining is an investigation of huge observational data sets to discover unsuspected associations and to review the data in novel customs that are mutually understandable and valuable to data owner.

1.1 *Data mining techniques*

A. Association **-** It is the greatest known and well researched technique for data mining. Association is also known as relation technique for the reason that patterns which are revealed from the dataset are based on the connection between the items. An association rule divides into two parts, an antecedent (if) and a consequent (then). An antecedent is an item that found in the data. A consequent is an item establish in grouping with the antecedent [1].

B. Classification **-** It is a data mining method that is used to categorize each item in a data set into one of predefined set of group or classes. It is a typical data mining method that is based on machine learning. As through the mainly data mining solutions, classification appears with a degree of certainty. It might be possibility of the item belonging to the class or it might be some other measure of how intimately the item looks like other examples from that class.

C. Clustering **-** A data mining system that make cluster of items having similar characteristics is recognized as clustering. A cluster of data items can be treated as single group. Whereas doing cluster analysis, we primary separating the data set into groups based on resemblance of data and after that assign labels to the group.

D. Neural network **-** Neural network is a set of input/output divisions and every connection has a mass present on it. Throughout the learning period, network learns by regulating the masses so as to be capable to predict the exact labels of class of the input tuples.

E. Decision tree **-** It is the mainly used data mining method and its representation is simply understandable. The origin of the decision tree is a easy question or state that has numerous answers. Each answer directs to a set of questions or conditions that assists to verify the data so that we can take a concluding decision based on it [1].

F. Regression **-** it is a data mining (machine learning) method used to fit an equation to a set of data. The easiest form of regression, linear regression, apply the formula of a straight line (y = mx + b) and verifies the suitable values for m and b to calculate the value of y based upon a specified value of x. Highly developed techniques, for example multiple regression, permit the use of more than one input variable and permit for the appropriate of more difficult models, such as a quadratic equation [1].

1.2 Improved SVM

Modified SVM (Support Vector Machine) aims to adapt two or more classifiers of any kind to new datasets [41]. Problem is how to select best classifier for adaptation. Solution to this problem is to select classifier with best parameters after estimating performance of each classifier on sparsely labeled dataset. General problem of binary classification task is considered on original dataset D^o, which made up of majority of unlabeled instances D_u^o and limited number

of labeled instances D_l^o,therefore, the original dataset is:

$D^o = D_l^o \cup D_u^o$

There are one or more subordinate datasets D_1^s……,D_M^s which is different from the original dataset. The subordinate classifier f_k is used to train each of the subordinate datasetsD_k^s. We have,

$D_l^o = \{(x_i, y_i)\}_{i=1}^N$

Where, x_i is the i_{th} data vector and $y_i \in \{-1, +1\}$ is its binary label. Data vector x always include a constant 1 as its first element, such that,$x_i \in R^{d+1}$, where d is the number of features. There exist multiple subordinate datasets as D_1^s, D_M^s with$D_k^s = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$, where $x_i^k \in R^{d+1}$ and $y_i^k \in \{-1, +1\}$. The subordinate dataset description is different from the original dataset. The subordinate classifier $f_k^s(x)$ has been trained from each subordinate dataset D_k^s , which gives us the result of prediction of data label through the sign of its decision function, i.e. $\hat{y} = f^s(x)$.

The traditional SVM trains the $f(x)$ from the labeled dataset D_l^o . Modified SVM is used to adapt a combination of multiple existing classifiers $f_1^s(x) \ldots\ldots, f_M^s(x)$ to the new classifier. The traditional SVM trains the $f(x)$ from the labeled datasetD_l^o. The decision boundary is determined by the kernel function$\langle x, x' \rangle = \langle \Phi(x), \Phi(x') \rangle$, where $\Phi(x)$ is a feature vector. The kernel function is the inner product of two projected feature vectors. Delta function is used in Modified SVM in the form of $\Delta f(x) = w^T \Phi(x)$ on the basis of $f^s(x)$:

$$f(x) = f^s(x) + \Delta f(x) = f^s(x) + w^T \Phi(x) \quad (4.1)$$

Where, w are the parameters predicted from the labeled dataD_l^o. As defined earlier, the objective is to make a group of subordinate classifiers and adapt this group to new classifier$f(x)$. By using Eq.(3), the adapted classifier's form is:

$$f(x) = \sum_{k=1}^M t_k f_k^s(x) + \Delta f(x) = \sum_{k=1}^M t_k f_k^s(x) + w^T \Phi(x) \quad (4.2)$$

Where, $t_k \in (0,1)$ is the weight of each subordinate classifier $f_k^s(x)$, which sums to one as $\sum_{k=1}^M t_k = 1$.

II. Related Work

N. SRIRAAM et al [4] presented data mining approach for parametric evaluation to improve the treatment of kidney dialysis patient. Results on the basis of dialysis parameter combination suggested that its classification accuracy by Association mining is found between the ranges of 50-97.7%. Therefore, such approach can be providing benefits to the clinician and they can easily select the level of dialysis that is basically required for specific patient. K.R. Lakshmi et al [5] have been reported that performance comparison of Artificial Neural Networks, Logical Regression and Decision Tree are used for Kidney dialysis survivability. Different accuracy measures (classification accuracy, specificity and sensitivity) were selected for the estimation of data mining techniques. They achieved results by using 10 fold cross-validations and confusion matrix for each technique. They found ANN shows better results using Kidney dialysis of patient records.Morteza Khavanin Zadeh et al [8] This research described the prediction of early risk of AVF failure in patients by using supervised classification. Authors used different approaches to predict probability of complication in new haemodialysis patients. But mainly for those patient whom have been suggested by nephrologists to AVF operation. J. Van Eyck et al [9] After elective cardiac surgery explored data mining techniques for predicting acute kidney injury with Gaussian process & machine learning techniques (classification task & regression task). Xudong Song et al [10] introduced data mining decision tree classification method, and proposed a new variable precision rough set decision tree classificaAbeer. Y. Al- Hyari et al [11] proposed in their research by using Artificial Neural Network (NN), Decision Tree (DT) and Naïve Bayes

(NB) to predict chronic kidney disease. The proposed NN algorithm as well as the other data mining algorithms demonstrated high potential in successful kidney diseases.

III. Proposed Work

A. ***Initialize the dataset*** **-** The dataset is mined, uploaded and transformed into the required matrix form with the help of data mining tool Matlab.

B. ***Apply k nearest neighbor algorithm and evaluate accuracy*** **-** K nearest neighbor algorithm is applied to the dataset and accuracy is calculated.This algorithm assigns the object to the class which is most common in its neighbors.

C. ***Apply linear discriminant algorithm and evaluate accuracy*** **-** Linear discriminant algorithm is applied to the dataset and accuracy is calculated separately. This algorithm finds linear combination of features that separates two classes of objects.

D. ***Apply naive bayes algorithm and evaluate accuracy*** **-** This algorithm is applied to the dataset again separately and then the accuracy of this algorithm is calculated. Naive Bayes considers each of the features to contribute independently to the probability that the person has heart disease.

E. ***Apply support vector machine algorithm and evaluate accuracy*** **-** Support vector machine algorithm is applied to the dataset and accuracy is calculated. This algorithm finds the best hyper plane that separates the data points of one class from the data points of another class.

F. ***Apply improved support vector machine algorithm and evaluate accuracy*** **-** Support vector machine algorithm is improved by introducing specific kernel features and then the accuracy is calculated.

G. ***Comparative analysis of the algorithms in terms of accuracy, precision and recall*** **-** Comparative analysis of the entire algorithms is done and the result of performance is calculated in terms of accuracy, precision and recall.

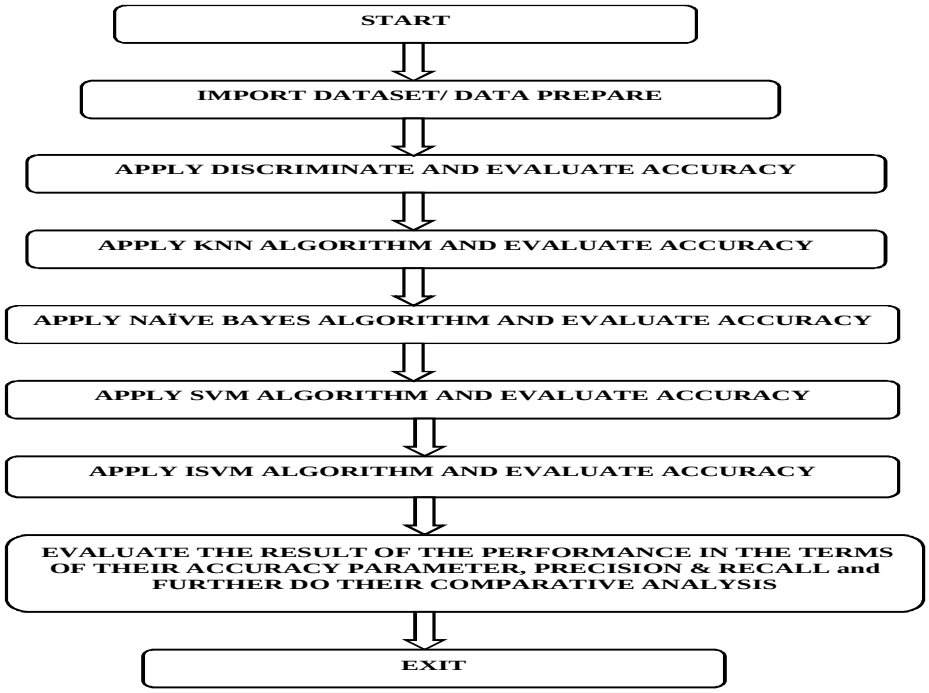

Fig 1 Flow diagram of Proposed System

IV. RESULTS AND DISCUSSION

The objective of this proposed work is to have greater accuracy, as high precision and recall metrics. For the implementation of our proposed algorithm I have used Matlab version 2015, with i7 processor with ram of 8gb having processor speed 2.7ghz, for fast optimization of our algorithm we initialized Matlab pool using the 'local' profile, nevertheless as we can see that the output performance of various classifier in figures the performance of improved SVM (support vector machine) Outperformed the rest of overall classification rate is 99.475% which is the highest accuracy achieved present in the literature.

A. Accuracy

This refers to the ability of the classifiers to correctly measure the intrusions from the training dataset. This is defined as the ratio of correctly classified data to the total classified data.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Where, TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

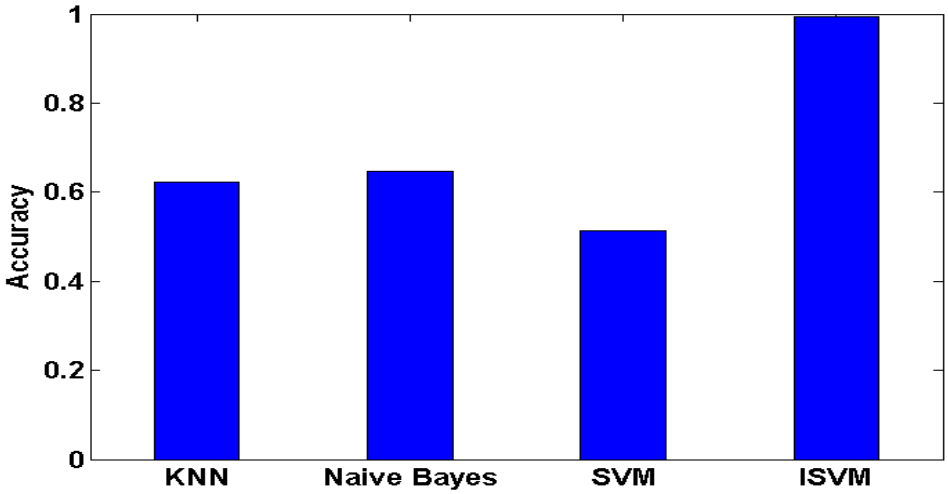

Fig 2 Prediction of Accuracy between various data mining techniques

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data. Above fig 5.10 shows the comparison between the existing and proposed accuracy. In the fig 2 it is clear that proposed method is give 100% accuracy which is very effective as compare with existing.

B. Precision

Precision is the fraction of retrieved instances that are relevant to also called positive predictive value. In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query. Above fig 5.8 shows the comparison between the existing and proposed Precision. In the above fig 3 it is clear that proposed method is give near about 100% Precision which is very effective as compare with existing.

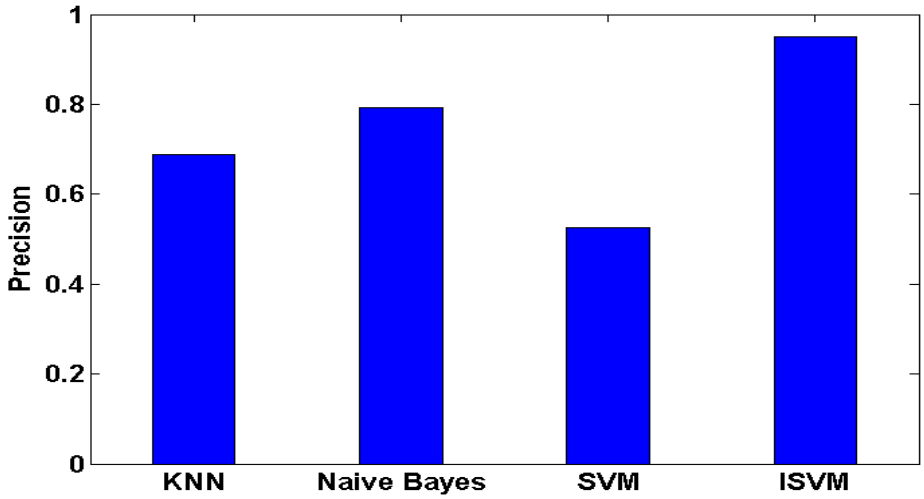

Fig 3 Prediction of Precision between various data mining techniques

C. Recall

Recall is the fraction of relevant instances that are retrieved or also called sensitivity. Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved. Above fig 5.9 shows the comparison between the existing and proposed recall. In the fig 4 it is clear that proposed method is give near about 100% recall which is very effective as compare with existing.

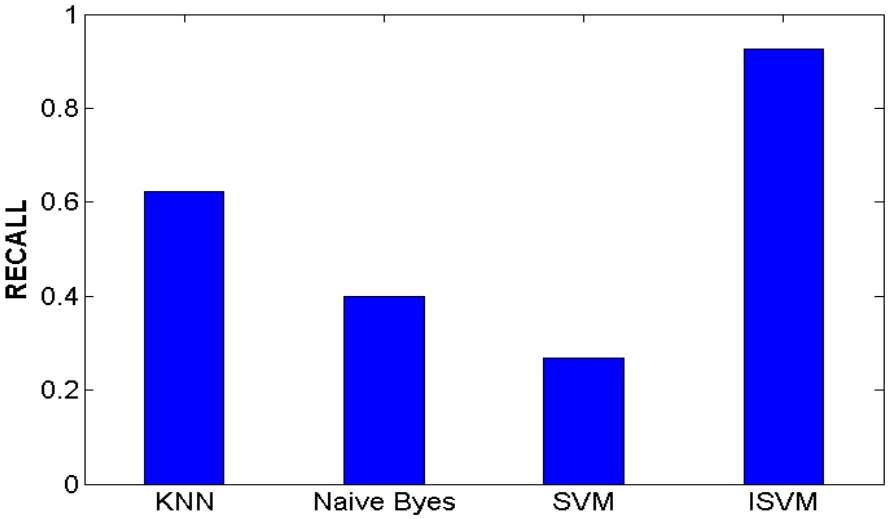

Fig 4 Prediction of Recall between various data mining techniques

Table 1 Prediction accuracy between various data mining techniques

Data mining technique	**Precision**	**Recall**	**Accuracy**
KNN	0.688685	0.446154	62.2237
Naïve Bayes	0.792013	0.4	64.7479
SVM	0.525984	0.267241	51.3202
Proposed	0.949345	0.925747	99.475

V. CONCLUSION

Medical related information is highly voluminous in nature in the healthcare industry. It can be derived or retrieved from various sources which are not entirely applicable in this feature. In this work, kidney disease prediction system was developed using classification algorithms (KNN, Naive Bayes, SVM, ISVM) through Matlab data mining tool to predict effective and better accurate results regarding whether the patient is suffering from kidney disease or not. As the kidney disease patients are increasing world-wide each year and huge amounts of data is available for research, where different data mining techniques are used in the diagnosis of kidney disease. So, different techniques used have shown different accuracies depending upon the number of attributes taken and tool used for implementation. In the result portion of the dissertation, it is clearly shown that our proposed system give accurate results in case of each parameter (accuracy, precision, and recall). Precision value of our proposed system is 0.949384% which is very large as compare to previous system. In the case of recall values our proposed system attains higher value i.e. 0.925747% and same is true in the case of accuracy.

REFERENCES

[1] Ms. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, "A data mining approach for prediction of heart disease using neural networks", international journal of computer engineering and technology", Vol. 3 , No. 3 , pp. 30-40, (2012).

[2] M.A. Nishara Banu and B. Gomathy," Disease Forecasting System Using Data Mining Methods",Vol. 1, No. 4, pp.130-133, (2014).

[3] Aqueel Ahmed and Shaikh Abdul Hannan,"Data Mining Techniques to Find Out Kidney Diseases", International Journal of Innovative Technology and Exploring Engineering(IJITEE), Vol. 1, No. 4, (2012)

[4] N.Sriraam, V.Natasha and H.Kaur,"data mining approaches for kidney dialysistreatment ", journal of Mechanics in Medicine and Biology, Vol. 06, No. 02,(2006).

[5] K.R. Lakshmi, Y. Nagesh and M. Veera Krishna," Performance comparison of three data mining techniques for predicting kidney disease survivability", International Journal of Advances in Engineering & Technology. Vol. 7, No. 1, pp. 242-254,(2014).

[6] Abeer Y. Al-Hyari," Chronic Kidney Disease Prediction System Using Classifying Data Mining Techniques", library of university of Jordan, 2012.

[7] DSVGK Kaladhar, Krishna Apparao Rayavarapu* and Varahalarao Vadlapudi, ”Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis”, Open Access Scientific Reports, Vol. 1 , No. 12 ,pp. , 2012.

[8] Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri,” Data Mining Performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients”,International journal of hospital research, Vol. 2, No. 1, pp 49-54,(2013).

[9] J. Van Eyck, J. Ramon, F. Guiza, G. Meyfroidt, M. Bruynooghe, G. Van den Berghe, K.U. Leuven,” Data mining techniques for predicting acute kidney injury after elective cardiac surgery”, Springer, (2012).

[10] Xudong Song, Zhanzhi Qiu, Jianwei Mu,” Study on Data Mining Technology and its Application for Renal Failure Hemodialysis Medical Field”, International Journal of Advancements in Computing Technology(IJACT) ,Vol. 4, No. 3, pp. , 2012.

[11] Abeer Y. Al- Hyari,” Chronic Kidney Disease Prediction System Using Classifying Data Mining Techniques”, library of university of Jordan, 2012.

[12] Luck, M. M., A. Yartseva, G. Bertho, E. Thervet, P. Beaune, N. Pallet, and C. Damon. "Metabolic Profiling of 1H NMR Spectra in Chronic Kidney Disease with Local Predictive Modeling." In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 176-181. IEEE, 2015.

[13] Dr. S. vijayarani , Mr. S. Dhayanand et al . “Kidney Disease Prediction Using SVM AND ANN Algorithms”, International Journal of Computing and Business Research (IJCBR), vol. 6 , issue. 2 , pp . 2229-6166 , 2015.

[14] Sunitha Devi1. P, Sowjanya. CH, Sunitha. K.V.N, (2014) “A Review of Supervised Learning Based Classification for Text to Speech System”, International Journal of Application or Innovation in Engineering & Management, Volume 3, Issue 6, June page no 79-86.

[15] Khyati Chaudhary, Bhawna Mallick, “Exploration of Data mining techniques in Fraud Detection: Credit Card”, International Journal of Electronics and Computer Science Engineering, Volume1,Number 3, page no 1765-1771.