Data Mining Techniques for the Prediction of Kidney Diseases and Treatment: A Review

Ms. Astha Ameta[1], Ms. Kalpana Jain[2]

[1] College of Technology and Engineering,
Udaipur, India
Ameta.astha@gmail.com
[2] Assistant Professor, Department of Computer Science and Engineering,
College of Technology and Engineering,
Udaipur, India
Kalpana_jain2@rediffmail.com

Abstract: *The healthcare sector has vast amount of medical data which is not properly analyzed and mined to discover useful information and interesting patterns. Applying data mining techniques on such domain can help medical practitioners to predict even the crucial diseases with ease. Data mining techniques consist of detection of Anomaly, learning the Association rules, Classification, Clustering, Regression, Time series analysis and Summarization. Kidney related problems and their treatments are non-trivial. Timely and accurate diagnosis in such cases can save many lives. In this paper, data mining techniques for prediction of kidney disease and treatment has been analyzed and reviewed.*

Keywords: Data mining, Naïve Bayes, SVM, ANN, KNN, Decision trees, Backpropagation, MLP, K-means Clustering.

1. Introduction

Data Mining is an inter-disciplinary field of computer science that focuses on discovering properties or patterns of data sets. Data mining is the extraction process of useful knowledge and valuable information from huge data sets. The extracted information can be used in defining rules to kick start the decision making process. Data mining is also called knowledge mining as it extracts knowledgeable information from large heap of data. Knowledge discovery in database follows six basic steps in its mining process:

a. Cleaning: This process removes outliers or noise from data.
b. Integration: It integrates various data sources.
c. Selection: Selection process selects the relevant data from data source.
d. Transformation: This step transforms the data into considerable form.
e. Data mining: Necessary step where techniques are applied to mine data pattern.
f. Pattern Evaluation: Evaluates the interesting patterns showing information.
g. Knowledge Presentation: Visualization and knowledge representation techniques represent the extracted information to the user.

Data mining techniques involve classification, clustering, Anomaly detection, time series analysis, Association rule learning, regression and summarization.

Classification is a tool to analyze the data which is used to mine the models that can define about important data classes. Such analysis helps us with a good understanding of the vast amount of data. Complex amount of data is generated in the field of healthcare which remains unused. By finding effective techniques which properly utilize data in extracting meaningful information from ocean of data can make the diagnosis and prediction process more accurate.

Kidney disease is a gradual but huge problem in medical science. Patients suffering from such diseases need special care and attention. The treatment in such cases is also non trivial. So, their timely and accurate prediction may prove beneficial to patients suffering from such diseases. The risk factor for kidney disease involves high blood pressure, diabetes, family history, heart related diseases. This review paper focuses on data mining techniques used for prediction of kidney diseases as well as their treatment like dialysis.

II. REVIEW OF PREDICTION OF KIDNEY DISEASE:

In this section, we review the classification techniques and the prediction of kidney diseases using these techniques.

Naïve Bayes classifiers are defined as the probabilistic classifiers which applies Bayes' theorem with naïve independence assumptions among the features. [1] employed naïve Bayes and SVM as the classification techniques for predicting kidney diseases. The main focus of research work is to find the best classification algorithm on the basis of classification accuracy and execution time performance factors. The algorithm with higher accuracy and minimum execution time is chosen as the best algorithm. In this classification, different accuracy rate is shown by each classifier. ANN has the maximum classification accuracy and it is considered as a better algorithm for classification. The performance of SVM is observed better than the other algorithm with respect to minimum execution time.

Artificial Neural Network (ANN) is a computational model. Its basis is the functions and structure of biological neural net. These are statistical and non-linear techniques which models the complex relationship between inputs and outputs. [2] used SVM and ANN for diagnosing acute renal failure. The results

showed that an artificial neural network outperformed a support vector machine. For prediction of death of patient, the artificial neural network gained classification accuracy of 0.8268 having a sensitivity of 0.8325 and a specificity of 0.8435. For prediction need for dialysis for patients, the artificial neural network measured classification accuracy of 0.8229 having a sensitivity of 0.8310 and a specificity of 0.8433. The obtained results suggest that artificial neural network provides a good model for kidney disease prognosis. [9] used ANN for diagnosing kidney stone disease. The goal of the research work is to diagnose disease of kidney stone with the use of three different neural network algorithms, each having a different architecture and variety of characteristics. [9] compared the performance of all three neural networks with respect to its accuracy, time needed to create model, and size of training data set. They made use of Learning vector quantization (LVQ), Radial basis function (RBF) networks and two layers feed forward perceptron trained with back propagation training algorithm for diagnosis of kidney stone disease. The multilayer perceptron with two hidden layers and back propagation algorithm is the best model for diagnosis of kidney stone disease. In [11], stone disease is diagnosed with the use of DA(discriminant analysis) and ANN in genetic polymorphisms. Here, the association of information of genetic polymorphisms with stone disease is identified, the analytic and predictive tool is discriminant analysis and artificial neural networks. The result shows that with genetic variables, DA classified 64% of participants. When all related factors were considered together, DA classified 74% whereas ANN classified 89% of participants. Hence, ANN gives better prediction results than the DA, if all the related factors, genetic as well as environmental, are simultaneously considered.

Support Vector Machine (SVM) is a supervised learning classification model which makes analysis of data for classification and regression analysis. [1] used SVM as one of the techniques and it's results were found better than the other technique used. [10] employed SVM which includes use of various machine learning techniques applied on irregular as well as unbalanced diabetes dataset, such as support vector machine classification and feature selection methods. Other use is Visualization of the risk factors to give physicians intuitive information about clinical pattern of each patient. Linear SVM classifiers combined with either embedded or wrapper feature selection methods give the best results. This methodology can predict the onset of diabetic nephropathy few months before the actual diagnosis from an unbalanced and irregular dataset with high prediction performance, which cannot be achieved when using statistical methods for example, logistic regression and *t*-test.

K-NN, abbreviated as K nearest neighbor is a classification algorithm which is a subset of supervised learning. [3] performed metabolic investigation in patients with nephrolithiasis. For knowledge extraction in the form of decision rules, various data mining techniques such as Clustering and Classification were employed. The results find out the morbidity risk and disease recurrence risk. In the research, it was found that support vector machine performed as the better classifier as compared to K-nearest neighbor. A new decision support system is introduced to predict chronic kidney disease (CKD) in [14]. Techniques used are K-NN and SVM and performance comparison parameters are accuracy,

precision and execution time taken. K-NN classifier performed better than SVM.

Decision tree uses a tree-like model or graph of decisions and their possible results. [6] compared the features and performance of data mining techniques such as Decision Tree, Rule Based Classifiers with Logistic Regression as a standard statistical data mining method for predicting outcome of kidney transplants over a 5-year horizon. It was found that classification trees and rule based classifiers are more fast and easy to interpret rather than logistic regression and classification. The experimental results also speak that rule based classifiers and decision tree classifiers are efficient approaches for pattern extraction from kidney transplantation dataset. In [5], approaches like data preprocessing, data transformations, and data mining are used to exploit information about the interaction between many of these parameters (measured) and survival of patient. Two different data mining algorithms are used for extracting knowledge in the decision rules form which were being used by an algorithm for decision-making, that predicts survival of patients which are new or unseen. Vital parameters detected by data mining are interpreted to know how significant they are medically. The methods introduced in [5] have been tested and applied with the use of data collected at four dialysis sites.

Back propagation is a well-known method of training artificial neural networks. It is a learning technique that adjusts weight in neural networks by propagating weight changes backward from the sink to the source node. ". In [8], two neural network techniques, (a) Back Propagation Algorithm, (b) Radial Basis Function and one non-linear classifier Support Vector Machine are compared with respect to their efficiency and accuracy. The main focus of this thesis is to identify the best tool for kidney stone identification so as to reduce the diagnosis time and improve the efficiency and accuracy. The back propagation algorithm (BPA) significantly improves the conventional classification technique for use in medical field.

MLP, abbreviated as Multilayer Perceptron is a feedforward ANN model used to map input data sets to appropriate outputs. It contains multiple layers of nodes in a directed graph where every layer is fully connected to its next layer. An MLP is a network of perceptron. [4] predicted kidney disease with the use of multiple machine learning algorithms such as Support Vector Machine, Multilayer Perceptron, Decision Tree (C4.5), Bayesian Network and K-Nearest Neighbor. Here, these algorithms are compared and the most efficient algorithm on the basis of multiple criteria is defined. Observation for experimental results says that MLP and C4.5 have the best rates. However, when compared to the ROC curve, C4.5 has proved itself as the most efficient algorithm.

III. REVIEW OF PREDICTION OF TREATMENT FOR KIDNEY DISEASE:

Prediction about the treatment of kidney disease, like dialysis, is as important as prediction of the disease itself. It tells about the survivability of patient via dialysis. [12] identified whether a patient requires hemodialysis or not. It makes use of entropy function to identify hemodialysis' key features, which are used as dimensions in cluster analysis. The requirement of dialysis for patients is identified by these key features. In this work,

association rules of each cluster are found and in this way, hidden rules (if any) are identified. Thus, [12] does the following:

1. It predicts patients with high probability to perform hemodialysis by finding key features.
2. To categorize the patients, it applies k-means clustering with key features.
3. Association rules are identified from each cluster using data mining technique.
4. Whether a patient requires hemodialysis is determined by mined rules.

[7] combined temporal abstraction(TA) with data mining techniques for analysis of biochemical data of dialysis patients to develop a decision support system. The mined temporal patterns are helpful to predict hospitalization of hemodialysis patients and accordingly, suggest quick/immediate treatments to avoid hospitalization. The experimental results show that various data mining methods can be combined in an efficient manner, and more adequate patterns can be sought for practical applications. Also, we can add domain knowledge prior to data analysis by addition of the TA method, so as to make mining results more likely comprehended by the clinician.

Performance of three data mining techniques – ANN, Decision tree, logistic regression for prediction of survivability of kidney dialysis are compared in [13]. Information is extracted in classification rule form. ANN performed best in terms of accuracy and performance for kidney dialysis prediction.

IV. CONCLUSION

This paper presents a review on data mining techniques used for the prediction of kidney diseases and their treatment like dialysis. It is evident from the paper that data mining technique, like classification, is highly efficient in prediction of kidney related diseases. Various classification techniques, like support vector machine (SVM), artificial neural networks (ANN), Bayesian network (BN), Naïve Bayes, decision tree, rule based, discriminant analysis, clustering technique like k-means, and temporal abstraction techniques are efficient in prediction of different kidney diseases depending on the data. Various feature selection methods can further improve the classification results by optimizing the process of classification by removing the irrelevant and redundant data.

REFERENCES

[1] Vijayarani, S., & Dhayanand, S., 2015. Data mining classification algorithms for kidney disease prediction. *International Journal on Cybernetics and Informatics (IJCI)*.

[2] Chao, C. M., Yu, C. T., & Cheng, B. W., 2012. DATA MINING MODELS FOR DIAGNOSING ACUTE RENAL FAILURE. *International Journal of Organizational Innovation (Online)*, *4*(4), 155.

[3] Adl, S. T., Givchi, A., Saraee, M., & Eshraghi, A., 2011. Disordered Metabolic Evaluation in Renal Stone Recurrence: A Data Mining Approach. *Journal of Applied Computer Science & Mathematics*, (11).

[4] Boukenze, B., Haqiq, A., & Mousannif, H., 2017. Predicting Chronic Kidney Failure Disease Using Data Mining Techniques. In *Advances in Ubiquitous Networking 2*, *pp. 701-712*. Springer Singapore.

[5] Kusiak, A., Dixon, B., & Shah, S., 2005. Predicting survival time for kidney dialysis patients: a data mining approach. *Computers in biology and medicine*, *35*(4), *pp. 311-327*.

[6] Fouad, M., Hagag, M., & Akl, A. Prediction Of Long Term Living Donor Kidney Graft Outcome: Comparison Between Different Machine Learning Methods.

[7] Yeh, J. Y., Wu, T. H., & Tsao, C. W., 2011. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems*, *50*(2), *pp. 439-448*.

[8] Abhishek, G. S. M. T., & Gupta, D., 2012. Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis. *International Journal of Computer Science and Information Technologies*, *3*(3), *pp. 3900-3904*.

[9] Kumar, K., & Abhishek, B., 2012. Artificial neural networks for diagnosis of kidney stones disease.

[10] Cho, B. H., Yu, H., Kim, K. W., Kim, T. H., Kim, I. Y., & Kim, S. I., 2008. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. *Artificial intelligence in medicine*, *42*(1), pp. 37-53.

[11] Chiang, D., Chiang, H. C., Chen, W. C., & Tsai, F. J., 2003. Prediction of stone disease by discriminant analysis and artificial neural networks in genetic polymorphisms: a new method. *BJU international*, *91*(7), *pp. 661-666*.

[12] Lu, T. C., & Tseng, C. Y. (2012). Hemodialysis key features mining and patients clustering technologies. *Advances in Artificial Neural Systems*, *2012*, 6.

[13] Lakshmi, K. R., Nagesh, Y., & Krishna, M. V. (2014). Performance comparison of three data mining techniques for predicting kidney dialysis survivability. *International Journal of Advances in Engineering & Technology*, *7*(1), 242.

[14] Sinha, P., & Sinha, P. (2015, December). Comparative study of chronic kidney disease prediction using KNN and SVM. In *International Journal of Engineering Research and Technology* (Vol. 4, No. 12, December-2015). IJERT.

[15] Han J. & Kamber M.,2000. data mining Concepts and Techniques, *Morgan Kaufmann Publishers*, *pp. 337-342*.