VOLUME-8, ISSUE-1, JANUARY-2019 • PRINT ISSN No 2277 - 8160

Original Research Paper | Engineering

AN INVESTIGATION OF MACHINE LEARNING TECHNIQUES FOR CHRONIC KIDNEY DISEASE DETECTION

Santosh Reddy Chada* Fossil Group Inc., USA *Corresponding Author

Srinath Vallakirthy Texas Health Resources, USA

ABSTRACT Nowadays, there has been a recent interest in applying machine learning (ML) techniques to detect different chronic diseases. Chronic kidney disease (CKD) is one such emerging global health problem, and its early detection can be beneficial to take precautions and preventive measures. Hence, there is a requirement of a more sophisticated and specialised system for detecting the CKD in early stages. Recently, ML techniques remain as a focus of researchers for detecting chronic diseases by keeping their potential benefits like adaptability, flexibility and learning by example into consideration.

This work aims to present an investigation of various ML techniques for providing effective detection of CKD. It compares these techniques for their performance on a benchmark dataset of CKD in terms most common performance metrics. The results indicate that MLP is the most accurate ML technique in comparison for detecting the CKD up to 99.75%. Whereas, IB1 is the fastest technique that takes minimum time for building the model from CKD dataset. The comparative analysis of these techniques helps to identify the best performing ML technique. The recognized technique can be considered as a candidate for developing an effective CKD detection system. The proposed work also helps the readers and fellow researchers to better understand the framework for applying ML techniques to detect the disease like CKD.

KEYWORDS :

1. Introduction

Machine learning (ML) is a multi-disciplinary field of artificial intelligence (AI), statistics, probability, information theory, psychology, philosophy, and neurobiology (Rania 2016). ML solves the real world problems by building a model of training on the given data and use that model to predict the unknown data (Bansal et al. 2018). The ML techniques have been recently employed in a wide range of application areas like classification, image processing, natural language processing and network security.

The most critical application of machine learning is pattern matching. Pattern matching attempts to search the relationship of multiple features, finally, resulting in improved efficiency of systems. These methods provide computer-based information systems to explore data patterns. They help to produce information for the hidden association and research knowledge that reveals essential facts that cannot be approached by conventional computer-based systems. Machine learning can be of three types, viz. Supervised learning, Unsupervised learning or Reinforced learning as depicted in Fig.1. In supervised learning, the labels for each class are provided for the classifier at the training stage. In unsupervised learning, the class labels are not known, but the classifier groups the instances having similar features into clusters. Here, each group represents one class of data instances. In reinforcement learning, the classifier makes a classification of each case and is given a score after each classification, to reflect how well it classified the instance. The classifier then adjusts its future actions accordingly (Kumar and Kumar 2010).

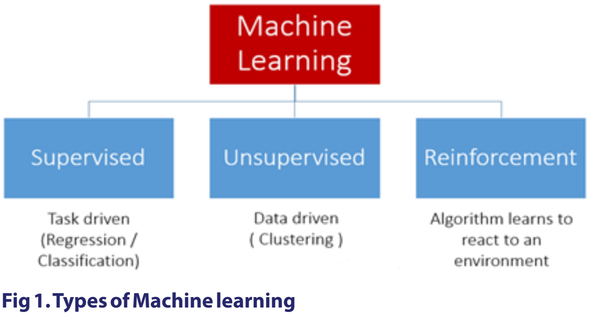

Fig 1. Types of Machine learning

Chronic kidney disease (CKD) has become a worldwide public health problem with increasing incidence and prevalence, high costs, and poor outcomes (Eknoyan et al. 2004). There is even substantially higher prevalence of the earlier stages of CKD, with adverse consequences, including loss of kidney function, cardiovascular disease (CVD), and premature death. Strategies to improve results will require a global effort directed at the earlier stages of CKD (Levey et al. 2005). As per the World Health Organization, it is reported that millions of people around the globe are suffering from CKD and its number is increasing every year(Lakshmi and Nagesh, 2014). Therefore, an early diagnosing technique is the need of the hour for taking precaution or controls well in advance.

Recently, with the advancement of technology, ML techniques have been successfully employed in the medical field for extracting useful information. These techniques can help to obtain the hidden patterns of massive databases and correlation of several variable data.

In this work, we compared the performance of various supervised ML techniques, namely, BayesNet, Naive Bayes, Multi-Layer Perceptron (MLP) neural network, SMO, IB1, J48, AdaBoosted Decision tree J48, and bagged decision tree J48 based on defined performance metrics using benchmarked CKD dataset. Evaluation of the classifiers on a variety of parameters is very significant because different classifiers are designed by keeping in mind to optimize different criteria. For example, SVM are intended to minimize the structural risk and hence maximize the accuracy, whereas neural network is designed to reduce empirical risk and accordingly optimize root mean squared error (RMSE). It is common that one classifier may show optimal performance on one set of metrics and suboptimal on another set of metrics. Major contributions of this empirical comparison of various ML techniques to identify best performing technique using CKD dataset in terms of defined performance metrics.

Rest of paper is organized as follows. Section 2 describes the analysis of related work in comparing the ML techniques for detecting chronic diseases. Section 3 provides the experimental setup and methodology for conducting experiments. This includes preparation of benchmark dataset and definition of performance metrics. The results of best classifiers are compared and analyzed in terms of defined metrics in section 4. Finally, the concluding remarks and future research guidelines are highlighted in section 5.

2. Literature review

Several researchers utilized ML techniques for detection of CKD.

Given below is the comparative work done by different researchers.

Vijayarani and Dhayanand (2015) suggested predicting the kidney disease using six attributes of renal affected disease. Glomerular filtration Rate is a measured feature for prediction of kidney disease. The authors analyzed the performance of Naive Bayes and Support Vector Machine techniques for predicting kidney disease. The reported results proved the better performance of SVM in comparison to Naïve Bayes.

Ramya and Radha (2016) employed and compared a set of ML techniques for detecting kidney disease using test data of patient medical report. The authors used 1000 records with 15 attributes in their experimental work. Their results prove that RBF (Radial Basis Function) has a better detection accuracy of chronic kidney disease. Jena and Kamila (2015) analyzed CKD dataset by ML techniques, namely Naive Bayes, Support Vector Machine, Multilayer Perceptron, J48, Conjunctive Rule, and Decision table. The authors used 25 different attributes for detecting CKD. Their research shows that for CKD prediction, Multilayer Perceptron gives better results in comparison to the other techniques.

Sinha and Sinha (2015) designed a decision support system for predicting CKD using Support Vector machine and k-Nearest Neighbor technique. Their results indicate that k-NN is capable of producing high detection accuracy than SVM technique.

Al-Hyari et al. (2014) employed SVM and Logistic Regression (LR) techniques for detecting CKD. Their results indicate that SVM can give more detection accuracy than other methods.

Kumar (2016) empirically compared the performance of six ML techniques, namely, Random Forest Classifiers, Sequential Minimal Optimization (SMO), Naive Bayes, Radial Basis Function (RBF), Multilayer Perceptron (MLP), and Simple Logistic for detecting CKD from a given dataset. The authors indicated that Random Forest technique is the best performing techniques in comparison to the other methods for detecting CKD.

Bansal et al. (2018) performed a comparative analysis of four machine learning algorithms, i.e., J48, Naïve Bayes, Random Forest and Multilayer Perceptron for detecting dementia. They proved that J48 is performing best among all the techniques for the detection of Dementia after reducing the features using the CFSSubsetEval method.

The comparative analysis done in the aforementioned text has the following drawbacks: Most of the researchers used different CKD dataset for comparison, and hence different preprocessing has been done by all the researchers in their work. Thus their work cannot be critically evaluated on the same platform. To solve this purpose, a set of experiments is performed using a common CKD dataset to identify the best performing techniques for detecting CKD. The accurate and early detection of CKD can help in taking precautions and controls to prevent the severe problems of CKD.

3. Experimental methodology

This section describes the evaluation dataset, its preprocessing strategy, the formation of training and testing dataset and experimental setup.

3.1. The proposed methodology

We performed experiments on Intel R Core 2 Duo CPU E 4500 @ 2.20 GHz and 2GB RAM. We conducted a 2-class classification of the dataset using well known open source publicly available machine learning tool called WEKA (Witten et al. 2011) to classify CKD dataset. We conducted a set of experiments using default parameters of WEKA implemented classifiers in the knowledge flow environment. The stages of the investigations and their interaction are described as follows and depicted in Fig. 2.

1. **Preprocessing model:** This model performs a conversion of characteristic features to numeric features and normalization of features is performed for Training and Test CKD dataset similar to the method described by Kumar and Kumar, (2011).
2. **CMD detection module:** This module involves two phases, namely the training phase and testing phase. 1) Training Phase: Here, the classifier is learnt using training dataset. The output of this phase is a trained to a model which optimized using 10 cross-validations. 2) Testing Phase: Here, trained model is given input of Test dataset to predict the class label.
3. **Performance computation module:** After the testing phase, the performance analysis module computes the defined performance metrics.

3.2. Benchmark dataset

In this set of experiments, we evaluated different ML techniques using Chronic Kidney Disease (CKD) Dataset from the UCI Machine Learning Repository (UCIML). This database contains 400 instances and 24 integer attributes, two class, namely, chronic kidney disease (ckd), not chronic kidney disease (notckd). There are 400 labelled records to be used as a training and test data set. Each record consists of 24 features and 01 class type as depicted in Table 1. The total number of instances of ckd class and non-ckd class are 250 and 150 respectively.

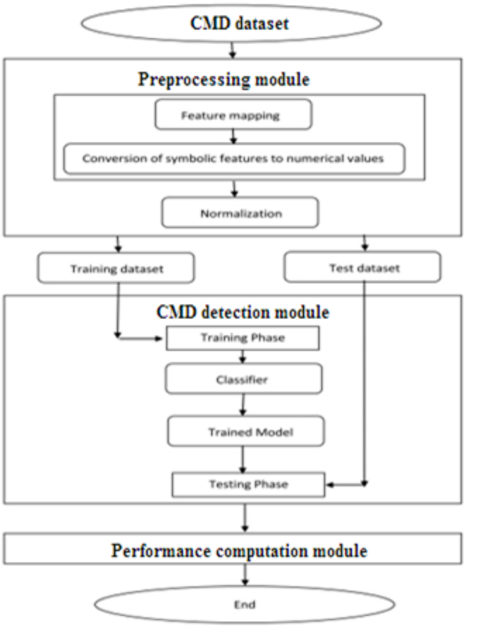

Fig.2 The proposed methodology

Table 1. The features of CKD dataset

Sr no.	Representation	Feature type	Description
1.	Age	Numerical	Age
2.	Bp	Numerical	Blood pressure
3.	Sg	Nominal	Specific gravity
4.	Al	Nominal	Albumin
5.	Su	Nominal	Sugar
6.	Rbc	Nominal	Red blood cells
7.	Pc	Nominal	Pus cell
8.	Pcc	Nominal	Pus cell clumps
9.	Ba	Nominal	Bacteria
10.	Bgr	Numerical	Blood glucose random
11.	Bu	Numerical	Blood urea
12.	Sc	Numerical	Serum creatinin
13.	Sod	Numerical	Sodium
14.	Pot	Numerical	Potassium
15.	Hemo	Numerical	Haemoglobin
16.	Pcv	Numerical	Packed cell volume
17.	Wc	Numerical	White blood cell count
18.	Rc	Numerical	Red blood cell count
19.	Htn	Nominal	Hypertension

20.	Dm	Nominal	Diabetes mellitus
21.	Cad	Nominal	Coronary artery disease
22.	Appet	Nominal	Appetite
23.	Pe	Nominal	Pedal oedema
24.	Ane	Nominal	Anaemia
25.	Classe	Nominal	Class

3.3. Performance analysis metrics

Performance metrics evaluate the performance of machine learning based CKD detection quantitatively. Several metrics have been proposed for the purpose. Most of the metrics can be computed from the confusion matrix which gives the classification possibilities of events. The possibilities are as described below:

- True positive (TP): Number of samples detected as CKD when it is actually CKD.
- True negative (TN): Number of samples detected as non-CKD when it is actually non-CKD.
- False positive (FP): Number of samples detected as CKD when it is actually non-CKD.
- False negative (FN): Number of samples detected as non-CKD when it is actually CKD.

From the confusion matrix, we computed correctly classified instances, incorrectly classified instances, detection accuracy, false positive rate, kappa statistics, MAE, and RMSE for comparative analysis. In addition, ROC is potted to a graphical view of a comparison of various machine learning techniques for detecting CKD.

3.4. Experimental setup

In this work, various ML techniques, namely, BayesNet, Naive Bayes, Multi-Layer Perceptron (MLP) neural network, SMO, IB1, J48, AdaBoosted Decision tree J48, and bagged decision tree J48 are employed to detect the CKD from the given dataset. The detail of these techniques can be explored in (Kumar et al.,2010; Witten et al., 2011). Experiments were conducted using WEKA - Waikato Environment for Knowledge Analysis (WEKA) software tool (Witten et al., 2011). The results were computed on Intel R Core 2 Duo CPU E 4500 @ 2.20 GHz and 2GB RAM.

The experimental layout using WEKA is as depicted in Fig. 3. It consists of different components of the knowledge flow environment for loading the dataset (Arff Loader), assigning the class feature of dataset (Class Assigner), picking the class values (Class ValuePicker), preparing training and test dataset using 10 cross-validation (Cross Validation Folder Maker), Classifier, classifier performance evaluator and finally text viewer, and graph chart viewer component.

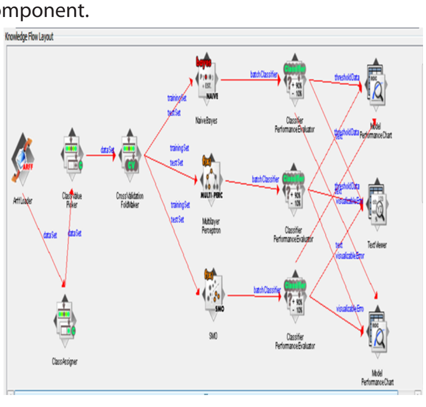

Fig. 3 Experimental layout

4. Experimental results and discussion

Since different ML techniques have been built upon different concepts and optimize a different set of metrics. So, they take different time for developing the respective model and execution time. The ML techniques have been compared on the basis of time taken to build the model and results are shown in Table 2 and Fig. 4-6. It can be observed that IB1 classifier learns quickly in comparison to other techniques. MLP took the maximum time for building the model using CKD dataset.

Table 2. Comparative analysis of time taken to build the model

Model	Time taken (Seconds)
BayesNet	0.13
Naive Bayes	0.03
MLP	17.75
SMO	0.12
IB1	0.02
J48	0.08
AdaBoosted J48	0.09
Bagged J48	0.18

We compared different ML models for their performance on the basis of different metrics, namely, correctly classified instances (CCI), Incorrectly classified instances (ICCI), Detection accuracy (DA), and False Positive Rate (FPR), Kappa Statistics, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as shown in Table 3 and Fig. 4-6. It can be observed from Table 3 that MLP reported the maximum classified instances with a minimum value for RMSE. However, it took maximum time for building a model of the given dataset as shown in Table 1.

Table 3: Comparative results

Model	CCI	ICCI	Detection accuracy (%)	FPR (%)	Kappa	MAE	RMSE
BayesNet	395	5	98.8	0.8	0.974	0.013	0.104
Naive Bayes	380	20	95	3	0.896	0.048	0.205
MLP	399	1	99.8	0.2	0.995	0.009	0.062
SMO	391	9	97.8	1.3	0.953	0.023	0.150
IB1	383	17	95.8	1.3	0.911	0.043	0.206
J48	396	4	99	1.4	0.979	0.023	0.081
AdaBoosted J48	398	2	99	0.6	0.989	0.019	0.083
Bagged J48	397	3	98.8	0.8	0.984	0.035	0.090

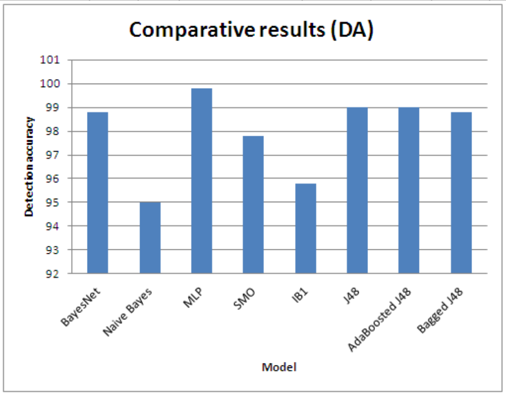

Fig. 4. Comparative results in terms of Detection Accuracy

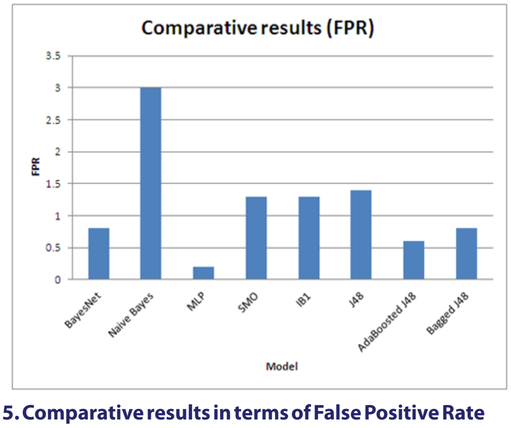

Fig. 5. Comparative results in terms of False Positive Rate

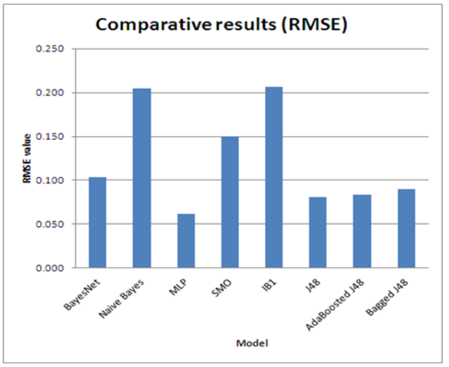

Fig. 6. Comparative results in terms of RMSE

5. Conclusions

In this study, we explored the literature of comparative work to identify promising techniques and research gaps of existing comparative work. To that end, a set of experiments is conducted for investigating the performance of ML techniques for detecting chronic kidney disease. Different ML techniques are compared in terms of a variety of performance metrics, namely, the time taken to build the model, CCI, ICCI, Kappa statistics, MAE, and RMSE. It can be observed that IB1 classifier learns from CKD dataset quickly in comparison to other techniques. But, it exhibits the detection accuracy up to 95.75%. The MLP neural network technique is found to be the most accurate in detecting the chronic kidney disease from CKD dataset up to 99.75% of detection accuracy. However, MLP takes more training time from a given dataset. As a future extension of this work, more ML techniques will be compared based upon the realistic dataset representing chronic diseases.

REFERENCES

1. Al-Hyari, A.Y., Al-Taee, A.M. and Al-Taee, M.A., 2014. Diagnosis and classification of chronic renal failure utilising intelligent data mining classifiers. International Journal of Information Technology and Web Engineering (IJITWE), 9(4), pp.1-12.
2. Bansal, D., Chhikara, R., Khanna, K. and Gupta, P., 2018. Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia. Procedia Computer Science, 132, pp.1497-1502.
3. Eknoyan, G., Lameire, N., Barsoum, R., Eckardt, K.U., Levin, A., Levin, N., Locatelli, F., Macleod, A., Vanholder, R., Walker, R. and Wang, H., 2004. The burden of kidney disease:
4. improving global outcomes. Kidney international, 66(4), pp.1310-1314.
5. Jena, L. and Kamila, N.K., 2015. Distributed data mining classification algorithms for prediction of chronic-kidney-disease. International Journal of Emerging Research in Management &Technology, 4(11), pp.110-118.
6. Kumar, G. and Kumar, K., 2011. AI based supervised classifiers: an analysis for intrusion detection, Proc. Of International Conference on Advances in Computing and Artificial Intelligence, ACM, pp. 170-174.
7. Kumar, G. and Kumar, K., 2012. An information theoretic approach for feature selection, Security and Communication Networks, 5(2), pp. 178-185.
8. Kumar, G., Kumar, K. and Sachdeva, M., 2010., The use of artificial intelligence based techniques for intrusion detection: a review, Artificial Intelligence Review, 34(4), pp. 369-387.
9. Kumar, M., 2016. Prediction of chronic kidney disease using random forest machine learning algorithm. International Journal of Computer Science and Mobile Computing, 5(2), pp.24-33.
10. Lakshmi, K.R., Nagesh, Y. and 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), p.242.
11. Levey, A.S., Eckardt, K.U., Tsukamoto, Y., Levin, A., Coresh, J., Rossert, J., Zeeuw, D.D., Hostetter, T.H., Lameire, N. and Eknoyan, G., 2005. Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO). Kidney international, 67(6), pp.2089-2100.
12. Padmanaban, K.A. and Parthiban, G., 2016. Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease. Indian Journal of Science and Technology, 9(29).
13. Raina, C.K., 2016. A review on machine learning techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 4(3), pp.395-399.
14. Ramya, S. and Radha, N., 2016. Diagnosis of chronic kidney disease using machine learning algorithms. International Journal of Innovative Research in Computer and Communication Engineering, 4(1), pp.812-820.
15. Sinha, P. and Sinha, P., 2015. Comparative study of chronic kidney disease prediction using KNN and SVM. International Journal of Engineering Research and Technology, 4(12), pp.608-12.
16. UCIML, UCI Machine Learning Repository: Kidney failure Data Set [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease#
17. Vijayarani, S. and Dhayanand, S., 2015. Data mining classification algorithms for kidney disease prediction. International Journal on Cybernetics and Informatics (IJCI).
18. Witten, I., Frank, E. and Hall, M., 2011. Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.