Comparative analysis of machine learning techniques based on chronic kidney disease dataset

Vineeta Gulati¹, Neeraj Raheja²

¹Research Scholar, CSE Department, Maharishi Arkaishwar (Deemed to be University), Mullana 133207, India
²Associate professor, CSE Department, Maharishi Markandeshwar (Deemed to be University), Mullana 133207, India

vineetaarora04@gmail.com neeraj_raheja2003@mmumullana.org

Abstract. Kidney failure is one of the chronic diseases that becoming a very common health issue in the world. It is a state in which kidney is damaged and cannot clean blood as well as they should. Excess fluid and waste may cause more health issue in our body and takes long period to diagnose. Moreover, with no early symptoms, the disease is detected only at a later critical end stage. As this Chronic Disease is becoming threat in today’s world, research is being done on a large scale to predict the presence of this chronic disease using machine learning. Machine learning is playing such a tremendous role in healthcare system like identify diseases and diagnoses, drug discovery and manufacturing, smart health record with the use of various techniques like Support Vector Machine, Decision Tree, Naïve Bayes, Random Forest etc. This paper comparatively analyzes the accuracy of pre-existing techniques for prediction of chronic kidney disease based on data from various research papers. Furthermore, this study also considers the different attributes from either already existing database or from real life database by using multiple techniques of machine learning. It is concluded that working with real life datasets with all possible attributes taken into consideration yields the accurate prediction for the presence of chronic kidney disease using machine learning.

1. Introduction

Kidney is very essential part of our body or we can say that we cannot survive without kidney because it filters waste and excess fluid from our body. But now a days kidney failure becoming a very common health issue in the world. It is a chronic disease means long lasting or takes long period to diagnose [1]. Chronic Kidney Disease is a state in which kidney is damaged and cannot clean blood as well as they should. Excess fluid and waste may cause more health issue in our body and this problem threatens all over the world [2]. As kidney disease progresses and irreversible destruction for our body, it may finally lead to kidney failure, which requires dialysis or a kidney transplant for survival In India 1 in 10 people estimated to be suffering from chronic kidney disease. Early analysis can lead to prevention of kidney failure. The best way to evaluate kidney function or guess kidney disease stages are to observe the Glomerular Filtration Rate (GFR) regularly [3]. GFR is calculated using age, gender, race and blood creatinine level.

Five stages of chronic kidney disease on basis of GFR:
Kidney disease is usually asymptomatic early and can go undiagnosed until it develops. It is often referred as silent disease [4].

- **Stage 1**: Normal kidney function stage is when calculated GFR is greater than 90 ml(GFR>90ml).
- **Stage 2**: If approximate GFR is in between 60-80 ml then minor loss of kidney function started.
- **Stage 3**: If calculated GFR is in between 30-59 ml then minor to harmful stage is to be considered.
- **Stage 4**: If approximate GFR is in between 15-29 ml then we can predict that the major loss of kidney function started.
- **Stage 5**: If calculated GFR less than 15 ml(GFR<15 ml) then this is the last stage of kidney failure which needs dialysis or transplant of kidney for survival.

**Dialysis**: When 90% of your kidney function fail then you require a artificial process to clean your blood from spare fluid, we have two process for this implementation. When an artificial kidney is used to fritter away fluid from your blood then this is called hemodialysis[5]. But when through pistula a tube pipe is fit in your arm or belly to purify your blood then this called peritoneal. Both the process is performed under the doctor supervision.

**End-Stage Renal Disease (ESRD)**: When transplantation or dialysis is only option to treat complete and stable kidney failure.

**Glomerular Filtration Rate (GFR)**: The rate at which the kidneys filter out waste and surplus fluid; calculated in millilitres per minute.

**Proteinuria**: When urine contains higher levels of an unusual protein called albumin. Though, Kidney disease is asymptomatic in early stage but if we are able to diagnose this disease in early stage through accurate medical action then we can prevent chronic kidney[6]. If a professional system guess the kidney disease by examine the patient’s symptoms like blood, urine test etc[7]. Then it gives doctors a proper time to heal or cure the patient disease on time. In current scenarios machine learning algorithms are playing important and accurate role in healthcare sector.

2. **Machine learning**

Machine Learning is outstanding within Artificial Intelligence area, which focuses more on to develop system which performs intelligently and also it has covered the largest real life impact for business.
The main function of ML is to enable machine to work in self-learning mode without being programmed explicitly. With the help of its algorithm implementation, it makes programs able to learn, grow, and change by themselves when exposed to new data. If the measurable performance of some task that is assigned by computer programs is improved as it gains more and more experience, then it is called Machine Learning, or we can say that when a machine takes decisions and makes predictions on data, then it is said to be the machine learning concept.

Classification, Regression, and Clustering are types of problems mainly solved through Machine Learning [9]. Which technique we are going to use depends mainly upon the type and category of data we provide to our model. The available techniques are Supervised, Unsupervised, Semi-Supervised, and Reinforcement to implement machine learning algorithms.

2.1 Supervised Learning
It is the simplest and easiest type of learning method. When your dataset acts as a teacher or you can say guide your model for training, then this type of learning is called supervised learning [10]. In this approach, your model gets trained automatically and starts making predictions and taking decisions.

2.2 Unsupervised Learning
When your model tries to find patterns from the given dataset or you can say that it learns through its observations and finds structure in the data, then this type of learning is called unsupervised learning [11][12].

2.3 Semi-supervised Learning
From the name, it is clear that it lies between supervised and unsupervised learning. Practically, the cost of labeling is very high in some situations, so we need that much skilled person who can work with only a few labels because the majority of cases are absent.

2.4 Reinforcement Learning
This type of learning is completely different from supervised and unsupervised learning. Here in this learning, we perform a feedback loop between an agent and an environment [13]. To make the relation or to connect the agent or environment, we give a set of actions. Video games are the best example of Reinforcement. Reinforcement goes through the following steps:
- Agent observed the input state.
- Agent performs an action instructed by the decision-making function.
- Action performed and now the agent receives reinforcement from the environment.
- This state-action information is stored for further use.

3. Machine learning in healthcare
Technology world’s new inventions giving opportunity for people to develop a system which helps the doctors to detect the chronic disease in early stage and cure the disease. Machine learning is playing a tremendous role in healthcare systems [5][14]. There are many applications of machine learning in healthcare sectors like identifying diseases and diagnoses, drug discovery and manufacturing, medical imaging diagnoses, smart health records, better radiotherapy, etc. [8]. There are many chronic diseases in this world like cancer, lungs problem, heart disease, kidney failure, etc. With no early symptoms and people coming to know only at the end stage, so here machine learning actually providing so many solutions to early detect your problem and cure them.

3.1 Requirement of Decision Support in Healthcare
Many people die every year because of not being able to detect disease early in the health care system. Health information technology construction suggested a few strategies, like association, significant consumer selection of clinician society and IT adoption [15].

[8] The main function of ML is to enable machine to work in self-learning mode without being programmed explicitly. With the help of its algorithm implementation, it makes programs able to learn, grow, and change by themselves when exposed to new data. If the measurable performance of some task assigned by computer programs is improved as it gains more and more experience, then it is called Machine Learning, or we can say that when a machine takes decisions and makes predictions on data, then it is said to be the machine learning concept.

Classification, Regression, and Clustering are types of problems mainly solved through Machine Learning [9]. Which technique we are going to use depends mainly upon the type and category of data we provide to our model. The available techniques are Supervised, Unsupervised, Semi-Supervised, and Reinforcement to implement machine learning algorithms.

2.1 Supervised Learning
It is the simplest and easiest type of learning method. When your dataset acts as a teacher or you can say guide your model for training, then this type of learning is called supervised learning [10]. In this approach, your model gets trained automatically and starts making predictions and taking decisions.

2.2 Unsupervised Learning
When your model tries to find patterns from the given dataset or you can say that it learns through its observations and finds structure in the data, then this type of learning is called unsupervised learning [11][12].

2.3 Semi-supervised Learning
From the name, it is clear that it lies between supervised and unsupervised learning. Practically, the cost of labeling is very high in some situations, so we need that much skilled person who can work with only a few labels because the majority of cases are absent.

2.4 Reinforcement Learning
This type of learning is completely different from supervised and unsupervised learning. Here in this learning, we perform a feedback loop between an agent and an environment [13]. To make the relation or to connect the agent or environment, we give a set of actions. Video games are the best example of Reinforcement. Reinforcement goes through the following steps:
- Agent observed the input state.
- Agent performs an action instructed by the decision-making function.
- Action performed and now the agent receives reinforcement from the environment.
- This state-action information is stored for further use.

3. Machine learning in healthcare
Technology world’s new inventions giving opportunity for people to develop a system which helps the doctors to detect the chronic disease in early stage and cure the disease. Machine learning is playing a tremendous role in healthcare systems [5][14]. There are many applications of machine learning in healthcare sectors like identifying diseases and diagnoses, drug discovery and manufacturing, medical imaging diagnoses, smart health records, better radiotherapy, etc. [8]. There are many chronic diseases in this world like cancer, lungs problem, heart disease, kidney failure, etc. With no early symptoms and people coming to know only at the end stage, so here machine learning actually providing so many solutions to early detect your problem and cure them.

3.1 Requirement of Decision Support in Healthcare
Many people die every year because of not being able to detect disease early in the health care system. Health information technology construction suggested a few strategies, like association, significant consumer selection of clinician society and IT adoption [15].
3.2 Decision Support

Health protection system in machine learning depends on the computer’s large computing capacity and on doctor’s logic competence [13]. Doctors and machine both search for pattern but doctors cannot estimate heartbeat of each patient as efficient machine can do. So, machine will do all these tasks and present the outcome for confirmation to doctor. Decision support system in healthcare. Decision support system help in early disease exploration by maintaining the awareness of health issue, background knowledge of individual patient to doctors [12]. It keeps track of insurance policy, refund detail, charge, account receivable, account payable. This system also helps in Identification of patient’s situation and then doctor suggest when and how to use the drug to cure your disease.

4. Literature review

This table present the work done for prediction of chronic kidney disease using machine learning, here I, Study how dataset collection (Open Source or Real life Data) and their pre-processing affects the accuracy level of our study.

| Author name | Study purpose and concept | Repository used | Research findings |
|-------------|---------------------------|-----------------|-------------------|
| Ramya S and Radhan N [2] | Presented different stages of kidney failure on basis of GFR and by using Random Forest, Radial Basis Function and Back propagation neural network classifier to identify which level of chronic disease is present in patient. Radial Basis Function gives maximum accuracy of 85.3%. | Data is collected through different medical reports of patients available in different laboratory in Coimbatore. 1000 records with 15 different attributes like age, gender, creatanine level, uric acid etc. | On basis of sensitivity, specificity, accuracy and kappa Radial Basis Function provide the accuracy of 85.3%. |
| Allassaf R A, Alsulaim K A, Alroomi, N Y, Alsharif N S, Aljubeir M F, Olatunji S O and Alturayeif N S [6] | The focus question is to predict CKD or not by two attribute selection process one is Correlation coefficient and another is recursive attribute elimination. | Data is Collected from King Fahd University hospital(KFUD) , 244 Records earlier with 491 features but after all refinement and cleaning it contains only 57 features. | ANN, SVM and Naïve Bayes provide 98% accuracy. Limitation: Only creatinine and urea correlation give highest accuracy, many attributes are ignored in this study to achieve the maximum accuracy. |
| Tekale S, Shingavi P, Wandhekar S and Chatorikar [7] | Presented a model for identify chronic kidney disease in patient by using pre existing dataset of patient and clean the data to achieve the better result, which is provided to our model for implementation process. | Datasets downloaded from UCI repository, 400 patient records with 25 aspect. Attributes used are red blood cell, Blood Pressure, white blood cell etc. Raw data needs to be clean so that properly use for our model like missing value removed by WEKA function “Replace Missing Value” with NA or by | Accuracy level given by Decision Tree is 91.75% and SVM is 96.7%. Calculation time is less. Limitation: Strength of the data is not higher because of size and missing aspect. |
| Authors                | Methodology                                                                                                                                                                                                 |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Aljaaf A J, Aljumeily D, Haglan H M, Alloghani M, Baker T, Hussain A J and Mustafina J [8] | The purpose of the study is to develop a CKDPS system by using different approaches like finding the correlation between input parameters, such as creatinine and urea, and providing only correlated data to model for better prediction. Machine learning classifiers used in this study are regression tree, support vector machine, logistic regression, and multilayer perceptron neural network. Dataset given by Apollo Hospital available online at UCI Machine learning Repository (400 Record of patient with 24 attributes). This study starts with 24 attributes but after removing all missing values, noise, redundancy left only with 7 parameters strongly correlated to each other, given to the model for implementation. Multilayer perceptron neural network gives 99.5% accuracy with 98% sensitivity and 40 millisecond time consumption. Limitation: Value of most of the attributes is neglected to achieve better accuracy, but some attributes affected by prescribed drugs like blood pressure of the patient may affect our model prediction. |
| Revathy S, Bharathi B, Jeyanthi P and Ramesh M [13] | Examine the model for chronic disease prediction by using decision tree, random forest and support vector machine algorithms of machine learning. Adaboost and Logitboost two boosting techniques of data mining are used to enhance the performance of the model. R programming tool is used for statistical analysis and confusion matrix with two dimensions CKD or NONCKD is generated. For clean data, replace all the numerical values by mean or missing values by mode and then provide the dataset to model for training as well as testing purpose. Random forest provides the accuracy level of 99.16%. Limitation: Confusion matrix means two or more classes for output prediction, in that case, it is very difficult to predict that both the class equally involved or all the accuracy may belong to only one class. |
| Feng B, Zhao Y Y, Wang J, Yu H, Potu S, Wang J and Guo Y [16] | Presented an independent way to predict CKD without GFR measurement, which basically overcome the incorrect and ambiguous issues related to GFR. 703 patient record, with all metabolite feature on basis of blood sample used from randomized hospitals. With all metabolite feature RNN gives 95.6% accuracy, SVM 95.4%, Decision Tree – 95.4%. |
Random forest, Decision tree and support vector machine run parallel with randomized input values by 10 fold cross validation.

A large 16282 metabolite feature dataset with clinical blood samples were used independently and after their prediction then only selected 69 metabolite feature dataset is used and conclude the difference.

Further we apply unsupervised learning to validate that with selected metabolite is relate to critical factor of CKD.

With selected metabolite feature Models give maximum accuracy of 98%.

| Anandajayam P, Aravindkumar S, Arun P, Ajith.A [17] | Real life data of 43400 records collected from hospitals. | RNN algorithm provide 97.62% of accuracy to predict CKD. |
|-----------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| Purposed both a structural and un-structural real life data for early detection of chronic kidney disease. | A Digital format of patient’s record like statistical data information is also used called electronic health record. | Limitation : Dealing with both structural and un-structural combine form may be time consuming so for better result in future use only MRI scan or RAY. |
| A big real life data plays a major role for better output | 80% of dataset use for training and 20% use for testing. | |
| Naïve bays, KNN, SVM and RNN algorithms are implemented on refined data. | | |

5. **Accuracy analysis of various techniques**

The following study shows the accuracy comparison of various techniques, using different attributes from either pre-existing database or real life dataset to predict the presence of chronic kidney disease in an individual.

- Figure 2 shows the accuracy comparison of SVM and DT algorithms with 14 attributes of 400 patient records downloaded from UCI Repository[18]. SVM provides the best result with 96.7% accuracy but strength of the data is not higher because of size and missing aspect.
Figure 2. Accuracy comparison of SVM and DT techniques based on 14 attributes of UCI Repository

- Figure 3 shows the accuracy comparison of LOGR, SVM, MLP and RPART algorithms with 24 attributes of dataset downloaded from UCI Repository[19]. MLP gives the best result with 99.5% accuracy but finds only the correlation between input parameter to get the desired result for example cretanine and urea have strong relationship, Value of most of the attributes is neglected to achieve the better accuracy.

Figure 3. Accuracy comparison of SVM, LOGR, MLP and RPART techniques based on 24 attribute of UCI Repository dataset.

- Figure 4 shows the accuracy comparison of SVM, DT, RF algorithms with 25 attributes of dataset downloaded from UCI Repository. RF gives the best result with 99.16% accuracy[20].
Figure 4. Accuracy comparison of SVM, DT and RF techniques based on 25 attributes of UCI Repository dataset.

- Figure 5 shows the accuracy comparison of DT, NB, KNN and RNN algorithms with 43400 records structural as well as un-structural collected from hospitals[21]. RNN gives the best result with 97.62% accuracy. This may be time consuming so for better result in future use only MRI scan or R-RAY.

Figure 5. Accuracy comparison of NB, KNN, RNN and DT techniques based on structural as well as un-structural records.

- Figure 6 shows the accuracy comparison of SVM, DT and RF with 703 patient records, with all metabolite features on basis of blood sample used from randomized hospitals[16]. RNN gives the best result with 95.6% accuracy.
Figure 6. Accuracy comparison of SVM, DT and RNN with all metabolite features.

Figure 7. Accuracy comparison of algorithms with selected 69 metabolite features[16] are shown below:

6. Conclusion
This paper provides a Comparative Analysis of Machine Learning Techniques based on chronic kidney disease dataset with different attributes from either already existing database or from real life database. Numbers of machine learning algorithms are used like Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), Recurrent Neural Network (RNF) etc. The study includes how dataset affect the accuracy level of an algorithm for instance, with only 14 attributes SVM results in the accuracy of 96.7%, with 24 attributes MLP results in the accuracy of 99.5%. However, the value of most of the attributes is neglected to achieve a higher accuracy[8]. In such a case, some attributes which are affected by factors such as prescribed drugs affecting blood pressure of the patient, we may have a high probability for a wrong prediction. Hence, to work with real life data with all the possible attributes for prediction of chronic kidney disease is more accurate, like as shown in the figure 6, with all metabolite features on basis of blood sample used from randomized hospitals, RNN gives the best result with 95.6% accuracy. Therefore, in the future
working with real life datasets with all possible attributes taken into consideration yields the more accurate prediction for the presence of chronic kidney disease using machine learning.

7. References
[1] 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), 608-12
[2] 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), 812-820
[3] Subasi A, Alickovic E and Kevric J 2017 Diagnosis of chronic kidney disease by using random forest In CMBEBIH, pp. 589-594 Springer, Singapore
[4] Shailaja K, Seetharamulu B and Jabbar M A 2018 March Machine Learning in Healthcare: A Review In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). pp. 910-914 IEEE
[5] Ganiger S and Rajashekararaih K M M 2018 December Chronic Diseases Diagnosis using Machine Learning International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET). pp. 1-6 IEEE
[6] Allassaf R A , Alsulaim K A, Alroomi, N Y, Alsharif N S, Aljubeir M F, Olatunji S O and Alturayeif N S 2018 November Preemptive Diagnosis of Chronic Kidney Disease Using Machine Learning Techniques International Conference on Innovations in Information Technology (IIT). pp. 99-104 IEEE
[7] Tekale S, Shingavi P, Wandhekhar S and Chatorikar A 2018 Prediction of chronic kidney disease using machine learning algorithm Int. J. Adv. Res. Comput. Commun. Eng. 7(10), 92-96
[8] Aljaaf A J, Al-Jumeily D, Haqlan H M, Alloghi M, Baker T, Hussain A J and Mustafina J 2018 July Early prediction of chronic kidney disease using machine learning supported by predictive analytics IEEE Congress on Evolutionary Computation (CEC) . pp. 1-9
[9] Mahrishi M., Morwal S. (2020) “Index Point Detection and Semantic Indexing of Videos—A Comparative Review”. In: Pant M., Kumar Sharma T., Arya R., Sahana B., Zolfagharinia H. (eds) Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing, vol 1154. Springer, Singapore.
[10] Amirgaliev Y, Shamiliulu S, and Serekk A 2018 October Analysis of chronic kidney disease dataset by applying machine learning methods 12th International Conference on Application of Information and Communication Technologies (AITC). pp. 1-4 IEEE
[11] Almasoud M and Ward T E 2019 Detection of chronic kidney disease using machine learning algorithms with least number of predictors International Journal of Soft Computing and Its Applications 10(8)
[12] Ray S 2019 February A quick review of machine learning algorithms International conference on machine learning, big data, cloud and parallel computing (COMITCon). (pp. 35-39) IEEE
[13] Revathy S, Bharathi B, Jeyanthi P and Ramesh M 2019 October Chronic kidney disease prediction using machine learning models International Journal of Engineering and Advanced Technology (IJEAT) Volume-9 Issue-1 ISSN 2249-8958
[14] Rabby, A. S. A., Mamata, R., Laboni, M. A., & Abujar, S. (2019, July). Machine Learning Applied to Kidney Disease Prediction: Comparison Study. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCCNT) (pp. 1-7). IEEE.
[15] Dhillon A and Singh A 2019 Machine learning in healthcare data analysis: a survey Journal of Biology and Today's World 8(6), 1-10
[16] Feng B , Zhao Y Y, Wang J, Yu H, Potu S, Wang J and Guo Y 2019 May The application of machine learning algorithms to diagnose CKD stages and identify critical metabolites
features International Work-Conference on Bioinformatics and Biomedical Engineering (pp. 72-83) Springer Cham

[17] De S and Chakraborty B 2019 September Development of chronic kidney disease prediction system (CKDPS) using machine learning technique In International Conference on Intelligent Data Communication Technologies and Internet of Things (pp. 155-164) Springer Cham

[18] Veerasekaran K and Sudhakar P 2019 September Machine learning based predictive models and its performance evaluation on healthcare systems International Conference on Intelligent Data Communication Technologies and Internet of Things (pp. 677-684) Springer Cham

[19] Almasoud M and Ward T E 2019 Detection of chronic kidney disease using machine learning algorithms with least number of predictors International Journal of Soft Computing and Its Applications. 10(8)

[20] Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, and Ye Z 2019 Comparison and development of machine learning tools in the prediction of chronic kidney disease Progression Journal of translational medicine. 17(1), 119

[21] Anandajayam P, Aravindkumar S, Arun P, Ajith A 2019 Prediction of Chronic Disease by Machine Learning Proceeding of International Conference on Systems Computation Automation and Networking.