IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

A Machine Learning Methodology for Diagnosing Chronic Kidney Disease

Saraswathi. P[1], Vidya Shree. CH[2], P. Geethika[3], Shree Latha[4], M Priyanka[5]
Assistant Professor, Department of Computer Science[1]
Students, Department of Computer Science[2,3,4,5]
Rao Bahadur Y Mahabaleswarappa Engineering College, Bellary, Karnataka, India

Abstract: *Chronic kidney disease (CKD) is a global health problem with high morbidity and mortality rate, and it induces other diseases. Since there are no obvious symptoms during the early stages of CKD, patients often fail to notice the disease. Early detection of CKD enables patients to receive timely treatment to ameliorate the progression of this disease. Machine learning models can effectively aid clinicians achieve this goal due to their fast and accurate recognition performance. In this study, we propose a machine learning methodology for diagnosing CKD. The CKD data set was obtained from the University of California Irvine (UCI) machine learning repository, which has a large number of missing values. KNN imputation was used to fill in the missing values, which selects several complete samples with the most similar measurements to process the missing data for each incomplete sample. Missing values are usually seen in real-life medical situations because patients may miss some measurements for various reasons. After effectively filling out the incomplete data set, six machine learning algorithms (logistic regression, random forest, support vector machine, k-nearest neighbor, naive Bayes classifier and feed forward neural network) were used to establish models. Among these machine learning models, random forest achieved the best performance with 99.75% diagnosis accuracy. By analyzing the misjudgments generated by the established models, we proposed an integrated model that combines logistic regression and random forest by using perceptron, which could achieve an average accuracy of 99.83% after ten times of simulation. Hence, we speculated that this methodology could be applicable to more complicated clinical data for disease diagnosis.*

Keywords: Logistic Regression, Random forest, Support Vector Machine, k-nearest neighbour, Naïve Bayes classifier, Neural network

I. INTRODUCTION

CHRONIC kidney disease (CKD) is a global public health problem affecting approximately 10% of the world's population. The percentage of prevalence of CKD in China is 10.8% and the range of prevalence is 10%-15% in the United States. According to another study, this percentage has reached 14.7% in the Mexican adult general population. This disease is characterized by a slow deterioration in renal function, which eventually causes a complete loss of renal function. CKD does not show obvious symptoms in its early stages. Therefore, the disease may not be detected until the kidney loses about 25% of its function. In addition, CKD has high morbidity and mortality, with a global impact on the human body. It can induce the occurrence of cardiovascular disease.CKD is a progressive and irreversible pathologic syndrome. Hence, the prediction and diagnosis of CKD in its early stages is quite essential, it may be able to enable patients to receive timely treatment to ameliorate the progression of the disease. Machine learning refers to a computer program, which calculates and deduces the information related to the task and obtains the characteristics of the corresponding pattern. This technology can achieve accurate and economical diagnoses of diseases; hence, it might be a promising method for diagnosing CKD. It has become a new kind of medical tool with the development of information technology and has a broad application prospect because of the rapid development of electronic health record. In the medical field, machine learning has already been used to detect human body status, analyze the relevant factors of the disease and diagnose various diseases. For example, the models built by machine learning algorithms were used to diagnose heart disease diabetes and retinopathy, acute kidney injury cancer and other diseases. In these models,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

algorithms based on regression, tree, probability, decision surface and neural network were often effective. In the field of CKD diagnosis, Hodnel and et al. utilized image registration to detect renal morphologic changes .Vasquez-Morales et al. established a classifier based on neural network using large-scale CKD data, and the accuracy of the model on their test data was 95% .In addition, most of the previous studies utilized the CKD data set that was obtained from the UCI machine learning repository. Chen et al. used k-nearest neighbor (KNN), support vector machine (SVM) and soft independent modelling of class analogy to diagnose CKD, KNN and SVM achieved the highest accuracy of 99.7%.In addition, they used fuzzy rule-building expert system, fuzzy optimal associative memory and partial least squares discriminant analysis to diagnose CKD, and the range of accuracy in those models was 95.5%-99.6%. Their studies have achieved good results in the diagnosis of CKD. In the above models, the mean imputation is used to fill in the missing values and it depends on the diagnostic categories of the samples. As a result, their method could not be used when the diagnostic results of the samples are unknown. In reality, patients might miss some measurements for various reasons before diagnosing. In addition, for missing values in categorical variables, data obtained using mean imputation might have a large deviation from the actual values. For example, for variables with only two categories, we set the categories to 0 and 1, but the mean of the variables might be between 0 and 1. Polat et al. developed an SVM based on feature selection technology, the proposed models reduced the computational cost through feature selection, and the range of accuracy in those models was from 97.75%-98.5% J. Aljaaf et al. used novel multiple imputation to fill in the missing values, and then MLP neural network (MLP) achieved an accuracy of 98.1% .Subas et al. used MLP, SVM, KNN, C4.5 decision tree and random forest (RF) to diagnose CKD, and the RF achieved an accuracy of 100%. In the models established by Boukenze et al., MLP achieved the highest accuracy of 99.75%. The studies of focus mainly on the establishment of models and achieve an ideal result. However, a complete process of filling in the missing values is not described in detail, and no feature selection technology is used to select predictors as well. Almansour et al. used SVM and neural network to diagnose CKD, and the accuracy of the models was 97.75% and 99.75%, respectively. In the models established by Gunarathne et al., zero was used to fill out the missing values and decision forest achieved the best performance with the accuracy was 99.1%.

1.1 Problem Definition

Chronic kidney disease is detected during the screening of people who are known to be at threat by kidney problems, such as those with high blood pressure or diabetes and those with a blood relative to Chronic Kidney Disease(CKD) patients. So, early prediction is necessary for combating the disease and providing good treatment.

Having CKD increases the chances of having heart disease and stroke. Managing high blood pressure, blood sugar, and cholesterol levels—all factors that increase the risk for heart disease and stroke—is very important for people with CKD.

1.2Objectives

The main goal of this project is to determine whether or not a patient is at risk of developing a chronic disease, and to do so, we used classification techniques such as logistic regression, random forest, support vector machine, k-nearest neighbor, Naive Bayes classifier, and feed forward neural network.

II. LITERATURE SURVEY

[1] Z. Chen et al., “Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers,” Chemometr. Intell. Lab., vol. 153, pp. 140-145, Apr. 2016.

The feasibility of two in-house fuzzy classifiers, fuzzy rule-building expert system (FuRES) and fuzzy optimal associative memory (FOAM), for diagnosis of patients with chronic kidney disease (CKD) was investigated. A linear classifier, partial least squares discriminant analysis (PLS-DA), was used for comparison. The CKD data used in this work were taken from the UCI Machine Learning Repository. Composite datasets were created by adding different levels of proportional noise to evaluate the robustness of the two fuzzy approaches. Firstly, 11 levels of proportional noises were added to each numeric attribute of the training and prediction sets one after another, and then these simulated training and prediction sets were combined in pairs. Thus, a grid with 121 groups of simulated data was generated, and classification rates for these 121 pairs were compared. Secondly, the performances of two fuzzy

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

classifiers using the simulated datasets, in which 11 levels of noise were randomly distributed to each numeric attribute, were compared and the average prediction rates of FuRES and FOAM were 98.1 ± 0.5% and 97.2 ± 1.2%, respectively, with 200 bootstrap Latin partitions. The PLS-DA can give 94.3 ± 0.8% with the identical evaluation. Confluent datasets comprised of the original and modified datasets were also used to evaluate FuRES, FOAM, and PLS-DA classification models. The average prediction rates of FuRES and FOAM obtained from 200 bootstrapped evaluations were 99.2 ± 0.3% and 99.0 ± 0.3%. PLS-DA yields slightly worse accuracy with 95.9 ± 0.6%. The results demonstrate that both FuRES and FOAM perform well on the identification of CKD patients, while FuRES is more robust than FOAM.
These two fuzzy classifiers are useful tools for the diagnosis of CKD patients with satisfactory robustness, and can also be used for other kinds of patients
Summary: In the present work, two fuzzy classifiers, FuRES and FOAM, are applied for the classification of chronic kidney disease (CKD) patients. Based on the CKD data cited from the UCI Machine Learning website, their feasibility and robustness were investigated

[2] L. Zhang et al., "Prevalence of chronic kidney disease in china: a cross-sectional survey," Lancet, vol. 379, pp. 815-822, Aug. 2012.
We did a cross-sectional survey of a nationally representative sample of Chinese adults. Chronic kidney disease was defi Ned as eGFR less than 60 mL/min per 1·73 m² or the presence of albuminuria. Participants completed a lifestyle and medical history questionnaire and had their blood pressure measured, and blood and urine samples taken. Serum creatinine was measured and used to estimate glomerular fi ltration rate. Urinary albumin and creatinine were tested to assess albuminuria. The crude and adjusted prevalence of indicators of kidney damage were calculated and factors associated with the presence of chronic kidney disease analysed by logistic regression. Findings 50 550 people were invited to participate, of whom 47 204 agreed.. In rural areas, economic development was independently associated with the presence of albuminuria. The prevalence of chronic kidney disease was high in north (16·9% [15·1–18·7]) and southwest (18·3% [16·4 –20·4]) regions compared with other regions. Other factors independently associated with kidney damage were age, sex, hypertension, diabetes, history of cardiovascular disease, hyperuricaemia, area of residence, and economic status. Interpretation Chronic kidney disease has become an important public health problem in China. Special attention should be paid to residents in economically improving rural areas and specify c geographical regions in China. Funding The Ministry of Science and Technology (China); the Science and Technology Commission of Shanghai; the National Natural Science Foundation of China; the Department of Health, Jiangsu Province; the Sichuan Science and Technology Department; the Ministry of Education (China); the International Society of Nephrology Research Committee; and the China Health and Medical Development Foundation.
Summary: The prevalence of chronic kidney disease is high in developing countries. However, no national survey of chronic kidney disease has been done incorporating both estimated glomerular fi ltration rate (eGFR) and albuminuria in a developing country with the economic diversity of China. We aimed to measure the prevalence of chronic kidney disease in China with such a survey

[3] A. Singh et al., "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," J. Biomed. Inform, vol.53, pp. 220-228, Feb. 2015.
Predictive models built using temporal data in electronic health records (EHRs) can potentially play a major role in improving management of chronic diseases. However, these data present a multitude of technical challenges, including irregular sampling of data and varying length of available patient history. In this paper, we describe and evaluate three different approaches that use machine learning to build predictive models using temporal EHR data of a patient. The first approach is a commonly used non- temporal approach that aggregates values of the predictors in the patient's medical history. The other two approaches exploit the temporal dynamics of the data. The two temporal approaches vary in how they model temporal information and handle missing data. Using data from the EHR of Mount Sinai Medical Center, we learned and evaluated the models in the context of predicting loss of estimated glomerular filtration rate (eGFR), the most common assessment of kidney function. Our results show that incorporating temporal information in patient's medical history can lead to better prediction of loss of kidney function. They also demonstrate that exactly how this information is incorporated is important. In particular, our results demonstrate that the relative

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

importance of different predictors varies over time, and that using multi-task learning to account for this is an appropriate way to robustly capture the temporal dynamics in EHR data. Using a case study, we also demonstrate how the multi-task learning based model can yield predictive models with better performance for identifying patients at high risk of short-term loss of kidney function.

Summary: In this study we presented three different methods to leverage longitudinal data: one that does not use temporal information and two methods that capture temporal information. These methods address some of the challenges faced in using EHR data, rather than data from controlled studies, in building models. These challenges include irregularly sampled data and varying lengths of patient history.

[4] H. Polat, H.D. Mehr, A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," J. Med. Syst., vol. 41, no. 4, Apr. 2017.

As Chronic Kidney Disease progresses slowly, early detection and effective treatment are the only cure to reduce the mortality rate. Machine learning techniques are gaining significance in medical diagnosis because of their classification ability with high accuracy rates. The accuracy of classification algorithms depend on the use of correct feature selection algorithms to reduce the dimension of datasets. In this study, Support Vector Machine classification algorithm was used to diagnose Chronic Kidney Disease. To diagnose the Chronic Kidney Disease, two essential types of feature selection methods namely, wrapper and filter approaches were chosen to reduce the dimension of Chronic Kidney Disease dataset. In wrapper approach, classifier subset evaluator with greedy stepwise search engine and wrapper subset evaluator with the Best First search engine were used. In filter approach, correlation feature selection subset evaluator with greedy stepwise search engine and filtered subset evaluator with the Best First search engine were used. The results showed that the Support Vector Machine classifier by using filtered subset evaluator with the Best First search engine feature selection method has higher accuracy rate (98.5%) in the diagnosis of Chronic Kidney Disease compared to other selected methods.

Summary: In this study, wrapper and filter methods have been utilized on data set of CKD. Two different evaluators have been used for each method. For filter approach, Cfs Subset Eval with Greedy stepwise search engine and Filter Subset Eval with Best First search engine have been used.

[5] C. Barbieri et al., "A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis," Comput. Biol. Med., vol. 61, pp. 56-61, Jun. 2015.

Chronic Kidney Disease (CKD) anemia is one of the main common comorbidities in patients undergoing End Stage Renal Disease (ESRD). Iron supplement and especially Erythropoiesis Stimulating Agents (ESA) have become the treatment of choice for that anemia. However, it is very complicated to find an adequate treatment for every patient in each particular situation since dosage guidelines are based on average behaviors, and thus, they do not take into account the particular response to those drugs by different patients, although that response may vary enormously from one patient to another and even for the same patient in different stages of the anemia. This work proposes an advance with respect to previous works that have faced this problem using different methodologies (Machine Learning (ML), among others), since the diversity of the CKD population has been explicitly taken into account in order to produce a general and reliable model for the prediction of ESA/Iron therapy response. Furthermore, the ML model makes use of both human physiology and drug pharmacology to produce a model that outperforms previous approaches, yielding Mean Absolute Errors (MAE) of the Hemoglobin (Hb) prediction around or lower than 0.6 g/dl in the three countries analyzed in the study, namely, Spain, Italy and Portugal.

Summary: This paper has presented a reliable ML approach to predict Hb values in patients undergoing secondary anemia to CKD. The work is the result of a long experience of the authors in this problem, with some previous works in which the produced models were not completely satisfactory.

[6] V. Papademetriou et al., "Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study," Am. J. Med., vol. 130, no. 12, Dec. 2017.

Early stages of chronic kidney disease are associated with an increased cardiovascular risk in patients with established type 2 diabetes and macro vascular disease. The role of early stages of chronic kidney disease on macro vascular

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

outcomes in prediabetes and early type two diabetes mellitus is not known. In the Outcome Reduction with an Initial Glargine Intervention (ORIGIN) trial, the introduction of insulin had no effect on cardiovascular outcomes compared to standard therapy. In this post hoc analysis of ORIGIN, we compared cardiovascular outcomes in subjects without to those with mild (Stages 1-2) and/or moderate chronic kidney disease (Stage 3). Methods. Two co-primary composite cardiovascular outcomes were assessed. The first was the composite endpoint of non-fatal MI, non-fatal stroke, or death from cardiovascular causes; and the second was a composite of any of these events plus a revascularization procedure, or hospitalization for heart failure. Several secondary outcomes were pre- specified including micro vascular outcomes, incident diabetes, hypoglycaemia, weight, and cancers. Complete renal function data were available in 12,174 out of 12,537 ORIGIN participants. A total of 8,114 had no chronic kidney disease (67%) while 4,060 had chronic kidney disease stage 1-3 (33%). When compared to non-CKD participants, the risk of developing the composite primary outcome (nonfatal myocardial infarction, nonfatal stroke, or cardiovascular death) in those with mild to moderate chronic kidney disease was 87% higher; hazard ratio (HR): 1.87; 95% confidence interval (CI): 1.71-2.04 ($p<0.001$).

Summary: In high-risk patients with dysglycemia (pre-diabetes and early diabetes), mild and moderate chronic kidney disease significantly increased cardiovascular events.

III. PROPOSED SYSTEM

Proposed several machine learning models in this section, the classifiers were first established by different machine learning algorithms to diagnose the data samples. Among these models, those with better performance were selected as potential components. By analyzing their mis judgments, the component models were determined. An integrated model was then established to achieve higher performance.

3.1 Block Diagram

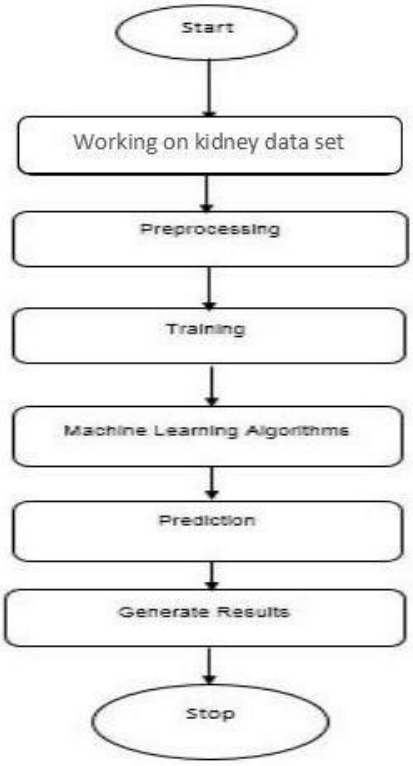

Fig 1. Block diagram of proposed method

Advantages:

- Highest accuracy
- Reduces time complexity

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

IV. METHODOLOGY

4.1 Logistic Regression:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function. There is a dataset given which contains the information of various users obtained from the social networking sites. There is a car making company that has recently launched a new SUV car. So the company wanted to check how many users from the dataset, wants to purchase the car. For this problem, we will build a Machine Learning model using the Logistic regression algorithm. The dataset is shown in the below image. In this problem, we will predict the purchased variable by using age and salary.

4.2 Support Vector Classifier:

Support Vector Machines (SVM) can handle both classification and regression problems. In this method hyperplane needs to be defined which the decision boundary is. When there are a set of objects belonging to different classes then decision plane is needed to separate them. The objects may or may not be linearly separable in which case complex mathematical functions called kernels are needed to separate the objects which are members of different classes. SVM aims at correctly classifying the objects based on examples in the training data set. Following are the advantages of SVM: it can handle both semi structured and structured data, it can handle complex function if the appropriate kernel function can be derived.

As generalization is adopted in SVM so there is less probability of over fitting. It can scale up with high dimensional data. It does not get stuck in local optima. Following are disadvantages of SVM: its performance goes down with large data set due to the increase in the training time. It will be difficult to find appropriate kernel function. SVM does not work well when dataset is noisy. SVM does not provide probability estimates. Understanding the final SVM model is difficult.

Support Vector Machine finds its practical application in cancer diagnosis, fraud detection in credit cards, handwriting recognition, face detection and text classification etc. So, among the three approaches of Logistic Regression, Decision Tree and SVM the first approach to attempt will be the logistic regression approach, next the decision trees (Random Forests) can be tried to see if there is significant improvement. When the number of observations and features are high then SVM can be tried out.

Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

4.3 Random Forest

Random forest is an extension over bagging. It takes one extra step where in addition to taking the random subset of data, it also takes the random selection of features rather than using all features to grow trees. When you have many random trees. It's called Random Forest Suppose there are N observations and M features in training data set. First, a sample from training data set is taken randomly with replacement. A subset of M features are selected randomly and whichever

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

feature gives the best split is used to split the node iteratively The tree is grown to the largest Above steps are repeated and prediction is given based on the aggregation of predictions from n number of trees Handles higher dimensionality data very well. Handles missing values and maintains accuracy for missing data.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

The random forest algorithm is actually a bagging algorithm: also here, we draw random bootstrap samples from your training set. However, in addition to the bootstrap samples, we also draw random subsets of features for training the individual trees; in bagging, we provide each tree with the full set of features. Due to the random feature selection, the trees are more independent of each other compared to regular bagging, which often results in better predictive performance (due to better variance-bias trade-offs), and I'd say that it's also faster than bagging, because each tree learns only from a subset of features.

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Suppose we have 1000 observations in the complete population with 10 variables. Random forest

4.4 K-Nearest Neighbors:

K Nearest Neighbour (KNN) Algorithm is a classification algorithm It uses a database which is having data points grouped into several classes and the algorithm tries to classify the sample data point given to it as a classification problem. KNN does not assume any underlying data distribution and so it is called non-parametric. Advantages of KNN algorithm are the following: it is simple technique that is easily implemented. Building the model is cheap. It is extremely flexible classification scheme and well suited for Multi-modal classes. Records are with multiple class labels. Error rate is at most twice that of Bayes error rate. It can sometimes be the best method. KNN outperformed SVM for protein function prediction using expression profiles. Disadvantages of KNN are the following: classifying unknown records are relatively expensive. It requires distance computation of k-nearest neighbours.

With the growth in training set size the algorithm gets computationally intensive, Noisy / irrelevant features will result in degradation of accuracy. It is lazy learner; it computes distance over k neighbours. It does not do any generalization on the training data and keeps all of them. It handles large data sets and hence expensive calculation. Higher dimensional data will result in decline in accuracy of regions. KNN can be used in Recommendation system, in medical diagnosis of multiple diseases showing similar symptoms, credit rating using feature similarity, handwriting detection, analysis done by financial institutions before sanctioning loans, video recognition, forecasting votes for different political parties and image recognition.-

A k-nearest neighbor (KNN) based bagging pruning algorithm for ensemble KNN classification is proposed in this paper. Redundant bags are discarded without reducing the performance of the ensemble classifier. Ten VCI binary classification datasets are used to evaluate the performance of the proposed pruning algorithm against single and bagging classifiers. Results show that the proposed bagging pruning improves the classification accuracies on most of the datasets with use less number of base classifiers thereby reducing computational requirements.

4.5 Naïve Bayes:

NAÏVE BAYES This algorithm is simple and is based on conditional probability. In this approach there is a probability table which is the model and through training data it is updated. The "probability table" is based on its feature values where one needs to look up the class probabilities for predicting a new observation. The basic assumption is of conditional independence and that is why it is called "naive". In real world context the assumption that all input features are independent from one another can hardly hold true. Naïve Bayes (NB) have the following advantages: implementation is easy, gives good performance, works with less training data, scales linearly with number of predictors and data points, handles continuous and discrete data, can handle binary and multi-class classification problems, make

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

probabilistic predictions. It handles continuous and discrete data. It is not sensitive to irrelevant features. Naïve Bayes has the following disadvantages Models which are trained and tuned properly often outperform NB models as they are too simple. If there is a need to have one of the features as "continuous variable" (like time) then it is difficult to apply Naive Bayes directly, even though one can make "buckets" for "continuous variables" it's not 100% correct. There is no true online variant for Naive Bayes, so all data need to be kept for retraining the model. It won't scale when the number of classes are too high, like > 100K.

Even for prediction it takes more runtime memory compared to SVM or simple logistic regression. It is computationally intensive especially for models involving many variables. Naïve Bayes can be used in applications such as Recommendation System and forecasting of cancer relapse or progression after Radiotherapy.

4.6 Neural Network

A neural network is a series of algorithms that endeavour's to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. The concept of neural networks, which has its roots in artificial intelligence, is swiftly gaining popularity in the development of trading systems.

Neural networks, in the world of finance, assist in the development of such process as time-series forecasting, algorithmic trading, securities classification, credit risk modelling and constructing proprietary indicators and price derivatives.

A neural network works similarly to the human brain's neural network. A "neuron" in a neural network is a mathematical function that collects and classifies information according to a specific architecture. The network bears a strong resemblance to statistical methods such as curve fitting and regression analysis.

A neural network contains layers of interconnected nodes. Each node is a perceptron and is similar to a multiple linear regression. The perceptron feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear.

In a multi-layered perceptron (MLP), perceptron's are arranged in interconnected layers. The input layer collects input patterns. The output layer has classifications or output signals to which input patterns may map. For instance, the patterns may comprise a list of quantities for technical indicators about a security; potential outputs could be "buy," "hold" or "sell."

Hidden layers fine-tune the input weightings until the neural network's margin of error is minimal. It is hypothesized that hidden layers extrapolate salient features in the input data that have predictive power regarding the outputs. This describes feature extraction, which accomplishes a utility similar to statistical techniques such as principal component analysis.

V. CONCLUSION

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. After unsupervised imputation of missing values in the data set by using KNN imputation, the integrated model could achieve a satisfactory accuracy. Hence, we speculate that applying this methodology to the practical diagnosis of CKD would achieve a desirable effect. In addition, this methodology might be applicable to the clinical data of the other diseases in actual medical diagnosis. However, in the process of establishing the model, due to the limitations of the conditions, the available data samples are relatively small, including only 400 samples. Therefore, the generalization performance of the model might be limited. In addition, due to there are only two categories (ckd and notckd) of data samples in the data set, the model can not diagnose the severity of CKD. In the future, a large number of more complex and representative data will be collected to train the model to improve the generalization performance while enabling it to detect the severity of the disease. We believe that this model will be more and more perfect by the increase of size and quality of the data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

REFERENCES

[1]. Z. Chen et al., "Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers," Chemometr. Intell. Lab., vol. 153, pp. 140-145, Apr. 2016.
[2]. A. Subasi, E. Alickovic, J. Kevric, "Diagnosis of chronic kidney disease by using random forest," in Proc. Int. Conf. Medical and Biological Engineering, Mar. 2017, pp. 589-594.
[3]. L. Zhang et al., "Prevalence of chronic kidney disease in china: a crosssectional survey," Lancet, vol. 379, pp. 815-822, Aug. 2012.
[4]. A Singh et al., "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," J. Biomed. Inform., vol. 53, pp. 220-228, Feb. 2015.
[5]. M. Cueto-Manzano et al., "Prevalence of chronic kidney disease in an adult population," Arch. Med. Res., vol. 45, no. 6, pp. 507-513, Aug. 2014.
[6]. H. Polat, H.D. Mehr, A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," J. Med. Syst., vol. 41, no. 4, Apr. 2017.
[7]. Barbieri et al., "A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis," Comput. Biol. Med., vol. 61, pp. 56-61, Jun. 2015.
[8]. V. Papademetriou et al., "Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study," Am. J. Med., vol. 130, no. 12, Dec. 2017.
[9]. N. R. Hill et al., "Global prevalence of chronic kidney disease - A systematic review and meta-analysis," Plos One, vol. 11, no. 7, Jul. 2016.
[10]. M. M. Hossain et al., "Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts," IEEE Trans. Ultrason. Ferr., vol. 66, no. 3, pp. 551-562, Mar. 2019.
[11]. M. Alloghani et al., "Applications of machine learning techniques for software engineering learning and early prediction of students' performance," in Proc. Int. Conf. Soft Computing in Data Science, Dec. 2018, pp. 246258.
[12]. Gupta, S. Khare, A. Aggarwal, "A method to predict diagnostic codes for chronic diseases using machine learning techniques," in Proc. Int. Conf. Computing, Communication and Automation, Apr. 2016, pp. 281-287.
[13]. L. Du et al., "A machine learning based approach to identify protected health information in Chinese clinical text," Int.
[14]. J. Med. Inform., vol. 116, pp. 24-32, Aug. 2018.
[15]. R. Abbas et al., "Classification of foetal distress and hypoxia using machine learning approaches," in Proc. Int. Conf. Intelligent Computing, Jul. 2018, pp. 767-776.
[16]. M. Mahyoub, M. Randles, T. Baker and P. Yang, "Comparison analysis of machine learning algorithms to rank alzheimer's disease risk factors by importance," in Proc. 11th Int. Conf. Developments in eSystems Engineering, Sep. 2018.
[17]. Alickovic, A. Subasi, "Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier," J. Med. Syst., vol. 40, no. 4, Apr. 2016.
[18]. Z. Masetic, A. Subasi, "Congestive heart failure detection using random forest classifier," Comput. Meth. Prog. Bio., vol. 130, pp. 56-64, Jul. 2016.
[19]. Q. Zou et al., "Predicting diabetes mellitus with machine learning techniques," Front. Genet., vol. 9, Nov. 2018.
[20]. Z. Gao et al., "Diagnosis of diabetic retinopathy using deep neural networks," IEEE Access, vol. 7, pp. 3360-3370, Dec. 2018.
[21]. R. J. Kate et al., "Prediction and detection models for acute kidney injury in hospitalized older adults," Bmc. Med. Inform. Decis., vol. 16, Mar. 2016.
[22]. N. Park et al., "Predicting acute kidney injury in cancer patients using heterogeneous and irregular data," Plos One, vol. 13, no. 7, Jul. 2018.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

[23]. M. Patricio et al., "Using resistin, glucose, age and BMI to predict the presence of breast cancer," BMC CANCER, vol. 18, Jan. 2018.

[24]. X. Wang et al., "A new effective machine learning framework for sepsis diagnosis," IEEE Access, vol. 6, pp. 48300- 48310, Aug. 2018.

[25]. Y. Chen et al., "Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B," Comput. Biol. Med., vol. 89, pp. 18-23, Oct. 2017.

[26]. Hodneland et al., "In vivo detection of chronic kidney disease using tissue deformation fields from dynamic MR imaging," IEEE Trans. BioMed. Eng., vol. 66, no. 6, pp. 1779-1790, Jun. 2019.