Journal of Pathology Informatics 14 (2023) 100189

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Chronic kidney disease prediction based on machine learning algorithms

Md. Ariful Islam [a,*], Md. Ziaul Hasan Majumder [b], Md. Alomgeer Hussein [c]

[a] Department of Robotics and Mechatronics Engineering, University of Dhaka, Dhaka 1000, Bangladesh
[b] Institute of Electronics, Bangladesh Atomic Energy Commission, Dhaka 1207, Bangladesh
[c] Department of Electrical and Electronic Engineering, University of Dhaka, Dhaka 1000, Bangladesh

ARTICLE INFO

Keywords:
Chronic kidney disease
Machine learning
XgBoost classifier
Classification model

ABSTRACT

Chronic kidney disease (CKD) is a dangerous ailment that can last a person's entire life and is caused by either kidney malignancy or decreased kidney functioning. It is feasible to halt or slow the progression of this chronic disease to an end-stage wherein dialysis or surgical intervention is the only method to preserve a patient's life. Earlier detection and appropriate therapy can increase the likelihood of this happening. Throughout this research, the potential of several different machine learning approaches for providing an early diagnosis of CKD has been investigated. There has been a significant amount of research conducted on this topic. Nevertheless, we are bolstering our approach by making use of predictive modeling. Therefore, in our approach, we investigate the link that exists between data factors as well as the characteristics of the target class. We are capable of constructing a collection of prediction models with the help of machine learning and predictive analytics, thanks to the better measures of attributes that can be introduced using predictive modeling. This study starts with 25 variables in addition to the class property, but by the end, it has narrowed the list down to 30% of those parameters as the best subset to identify CKD. Twelve different machine learning-based classifiers have been tested in a supervised learning environment. Within the confines of a supervised learning environment, a total of 12 different machine learning-based classifiers have indeed been examined, with the greatest performance indicators being an accuracy of 0.983, a precision of 0.98, a recall of 0.98, and an F1-score of 0.98 for the XgBoost classifier. The way the research was done leads to the conclusion that recent improvements in machine learning, along with the help of predictive modeling, make for an interesting way to find new solutions that can then be used to test the accuracy of prediction in the field of kidney disease and beyond.

Introduction

Chronic kidney disease, or CKD, is a condition in which the kidneys are so damaged that they can't filter blood as well as they should. The kidneys' main job is to get rid of waste and extra water from the blood.[8] This is how urine is made. CKD means that waste has built up in the body. This condition is called chronic because the damage happens slowly over a long period of time. It is a disease that affects people all over the world.[7] Because of CKD, you might experience various difficulties with your health. Diabetes, high blood pressure, and heart disease are only 3 of the many conditions that can lead to CKD. In addition to these serious health problems, age and gender also play a role in who gets a CKD.[26] If one or both of your kidneys aren't working right, you may have a number of symptoms, such as back pain, stomach pain, diarrhea, fever, nosebleeds, rash, and vomiting. The 2 most common illnesses that might cause long-term damage to the kidneys are diabetes and high blood pressure.[28] Therefore, the prevention of CKD can be thought of as the control of these 2 diseases. Because chronic kidney disease (CKD) does not often present any symptoms until it has progressed to a more advanced state, many people who have it do not realize they have it until it is too late.

Stages of CKD

Early stages of CKD

CKD in its early stages typically does not present any symptoms. This is due to the fact that the human body can typically adjust to a large decrease in the function of the kidneys. It is common for kidney disease to not be diagnosed until this stage unless a routine test for another issue, such as a test of the blood or urine, discovers a potential problem. If it is discovered at an early stage, treatment with medication and ongoing monitoring with routine tests may help prevent it from progressing to a more advanced state.

CKD in its advanced stages

If kidney disease isn't caught early or keeps getting worse even after treatment, there may be a number of signs.

* Corresponding author.
E-mail address: arif.rme@du.ac.bd (M.A. Islam).

http://dx.doi.org/10.1016/j.jpi.2023.100189
Received 10 September 2022; Received in revised form 5 January 2023; Accepted 5 January 2023
Available online 12 January 2023

Kidney failure is the last stage of CKD. It is also called end-stage renal disease or established renal failure. It is possible that dialysis or a kidney transplant will be needed at some point.

When to see a physician

If you have signs or symptoms of renal illness, make an appointment with your doctor. Renal disease could be prevented from progressing to kidney failure if detected early. During office visits, your doctor may use urine and blood tests to check your blood pressure and kidney function if you have a health condition that makes you more likely to get renal disease. Ask your physician if these tests are required for you.

Tests for CKD

Chronic kidney disease is when a disease or condition makes it hard for the kidneys to work, causing the damage to the kidneys to get worse over time. This can occur when the kidneys are affected by another disease or condition.

Studies[6,8] show that the number of people with CKD who are admitted to hospitals is going up by 6.23 percent every year, even though the global death rate has stayed the same.[8] There are just a few diagnostic tests available to check the status of CKD, including: (i) estimated glomerular filtration rate (eGFR) (ii) a urine test; (iii) a blood pressure reading; (iv) tests for CKD.

eGFR

The eGFR value provides information on how well your kidneys cleanse the blood. If your eGFR number is higher than 90, it means that your kidneys are working well. If the value of your eGFR is less than 60, this indicates that you have CKD.[8]

Urine test

In order to evaluate kidney function, the physician also requests a urine sample. Urine is produced by the kidneys. If your urine contains blood and protein,[24] it is an indication that one or both of your kidneys are not functioning normally.

Blood pressure

The doctor takes your blood pressure because the range of your blood pressure reveals how well your heart is pumping blood. If the patient's eGFR value falls below 15, this means they have reached the end stage of kidney disease. There are just 2two treatments that are now available for renal failure: (i) dialysis and (ii) kidney transplantation. The patient's life expectancy after dialysis is contingent on a number of characteristics, including age, gender, the frequency and length of dialysis treatments, the patient's level of physical mobility, and their mental state.[10] Kidney transplantation is the only option left for the doctor to consider if dialysis cannot be performed successfully. Nevertheless, the price is exorbitantly high.[15]

Other tests

When determining the extent of the damage to your kidneys, it is not uncommon for additional tests to be performed. These may include an ultrasound scan, a magnetic resonance imaging scan, or a computed tomography scan. Their purpose is to look at the kidneys and see if there are any blockages. A needle is used to take a small piece of kidney tissue, and the cells are looked at under a microscope to look for signs of kidney disease. This is done in order to diagnose kidney conditions.

The field of medicine is an extremely important area for the application of intellectually sophisticated systems.[21] Then, data mining could be a big part of finding hidden information in the huge amount of patient medical and treatment data. This is information that doctors often get from their patients to learn more about their symptoms and make more accurate treatment plans.

Literature review

Data mining is the process of using specialized software to find hidden information in a large set of data. Data mining techniques are linked to each other and used in a wide range of places and situations. With data mining technologies, we can make predictions, sort the data, filter it, and put it into groups. The goal of the algorithm is to process a training set that has a collection of attributes and targets, and the objective describes how this should be done. If the dataset is very big, data mining is a good way to find patterns in it. If the dataset is very small, however, we can still reach the same goal with the help of machine learning. Data analysis and pattern recognition are 2 further capabilities of machine learning.[1,16] Because there is such a wide diversity of health datasets, machine learning algorithms are the most appropriate method for enhancing the accuracy of diagnosis prediction.[17,19] The prevalence of machine learning algorithms in the healthcare industry is growing as a direct result of the rapid growth of electronic healthcare datasets.[9]

Using information mining techniques, a variety of types of studies have been carried out in order to extract useful information from datasets pertaining to chronic kidney disease.[22] This was done in order to cut down on the amount of time spent conducting the analysis, and in addition to that, it would increase the precision of the forecast with the assistance of the information mining categorization technique.[5] Data mining is also applied in the treatment and diagnosis of a number of diseases and conditions. Using techniques for information accumulation, different kinds of work have been done to get useful information out of the dataset on chronic kidney disease.

Polat et al.[21] offered directions that combined a total of 6 classifiers and 3 outfit measures. k-nearest neighbors (kNN), naive Bayes (NB), support nector machine (SVM), preference tables, random forest (RF), and J48 were some of the classifiers that were used. The authors of Polat et al.[21] looked into a number of possible treatments for chronic renal disease by using the k-means algorithm and Apriori. A test that uses SVM, DT, NB, and KNN computations was developed in order to diagnose chronic kidney disease (CKD). Ani et al.[4] modified many characterizations of computations, including DT, NB, Linear discriminant analysis (LDA) classifiers, randomized subspace, and kNN, as well as back propagation network (BPN). The decision tree (DT) and NB characterization approaches were employed in order to anticipate CKD and reduce the mortality rate that was caused by CKD[3,4]: Have you come up with a way to figure out how bad chronic kidney disease is in its early stages? They made use of several different neural network algorithms. Wickramasinghe et al.[27] described a research focus entirely by trying to bring relevant data from such a patient's medical record and then implementing a framework computation to these documents, which now provides CKD patients a rational eating regimen strategy. This focus was produced by attempting to extract information from a patient's medical notes. A strategy for information retrieval was suggested by Arora and Sharma.[5] This method incorporates identification capabilities, such as a planned release for implementation in Weka's equipment. In essence, Eroğlu and Palabaş[12] was able to remember informational indexes approaches as well as the methods by which it is possible to anticipate recurrent kidney infection. Therefore, there is little doubt that data mining is a more effective tool for predicting long-term kidney diseases. Charleonnan et al.[7] investigated 4 different machine learning approaches, namely kNN, SVM, logistic regression (LR), and DT classifiers. The dataset of patients with CKD is used to make these predictive models. Then, the performances of these models are compared to find out which classifier is the best at predicting which patients will get CKD. Using machine learning, Qin et al.[22] diagnosed CKD. The UCI machine learning repository's CKD dataset has many missing values. kNN imputation was used to fill in missing values, which finds full samples with identical measurements for each incomplete sample. Real-life medical scenarios often have missing values because patients miss measurements. After completing the dataset, 6 machine learning algorithms (LR, RF, SVM, kNN, NB, and feed forward neural network (FFNN)) were used to create models. With 99.75% accuracy, RF performed best. By examining the established models' misjudgments, they created a

model that combines LR and RF utilizing perceptrons, which achieved 99.83% accuracy after 10 simulations. They hypothesized that this technology may be used to diagnose complex diseases using clinical data. Almasoud and Ward[3] wants to test machine learning algorithms' capacity to forecast CKD with the fewest features. Analysis of variance (ANOVA), Pearson's correlation, and Cramer's V tests were used to remove redundant features. 10-fold cross-validation was used to train and test LR, SVM, RF, and gradient boosting. The gradient boosting classifier has a 99.1% accuracy rate. Hemoglobin is more important for RF and gradient boosting in identifying CKD. Their results are high compared to earlier studies, although they've reached fewer characteristics. Three simple tests can indicate CKD for $26.65.

However, in order to make clinical decisions regarding testing, treatment, and referral, a significant amount of accurate information regarding the risk of nephropathy progression is urgently required. As a result, the focus of this section was on the most recent developments in the study of CKD. Throughout this study, we tried to evaluate the possibility of a variety of machine learning algorithms, each of which could potentially provide an early diagnosis of CKD. There has been a substantial amount of research carried out on this subject; nonetheless, we are strengthening our strategy by making use of predictive modeling. As a result, in our methodology, we study the link that exists between the data variables and the characteristics of the target class. Because predictive modeling allows for a more accurate measurement of attributes to be introduced, we are able to use machine learning and predictive analytics to compile a set of prediction models. This is made possible by the improved ability of predictive modeling to introduce new attributes or identify the most important features responsible for CKD. The research on the detection of CKD is based on 1 dataset which is available in the UCI machine learning repository. The dataset includes 24 input features used by the maximum research mentioned above. No work has been found for detecting CKD based on the least number of predictors. In addition to improving the accuracy using this dataset, we tried to reduce the number of input features utilizing principle component analysis (PCA) and tried to develop a machine learning-based model showing the highest accuracy. A total of 12 distinct machine learning-based classifiers have been investigated, and the outcomes have been listed to compare with the previous studies.

Experimental data

CKD dataset

This approach makes use of a dataset from the UCI Machine Learning Repository[11] referred to as CKD. A total of 24 features and 1 target variable are included in the CKD Dataset. It can be broken down into 2 categories, yes or no.

The dataset has 25 attributes, 11 of which are numerical and 14 of which are nominal. For the purposes of training machine learning algorithms to make predictions, the entire dataset of 400 instances is utilized. Out of a total of 400 cases, 250 are classified as having CKD, and the remaining 150 are classified as having non-CKD. The attributes that can be found in the dataset are depicted in Fig. 1.

CKD dataset with PCA

Principal component analysis, also known as PCA, is a method that reduces the number of dimensions used to describe data while preserving as much of the original data's information as feasible.[20] When working with datasets that contain a significant number of features, PCA is a very helpful tool. The dataset that is available for the detection of CKD contains 24 input features, but it is necessary to know the contribution of each input feature to produce the outcome. Table 1 describes the overall scenario of the CKD dataset after PCA, in addition to the CKD original dataset. In the following section, the whole process will be implemented.

Methodology

The outputs of each classifier were tested with a variety of different assessment parameters, and a 10-fold cross-validation was used to check for over-fitting in the findings. In addition, the method of nested cross-validation has been of assistance in refining the model's underlying parameters. The evaluations will be carried out by utilizing the Jupyter Notebook web tool in conjunction with the Python 3.3 programming language. Several Sciket-learning libraries, which is a free platform for machine learning systems built on the Python programming language, were utilized. This analysis takes into consideration the following assessment measures: sensitivity, specificity, area under the curve (AUC), and accuracy as measured by the F1-measurement. Depending on the values of its parameters, every model generates uniquely diverse outputs.

In order to analyze the CKD dataset, a number of experiments were carried out using different types of machine learning algorithms, including SVM, KNN, LGBM, and hybrids. In this study, Fig. 2 demonstrates the overarching structure of the CKD diagnostic process. During the preprocessing stage, the mean technique was used to compute the missing numerical values, while the mode approach was used to compute the missing nominal values. Both methods are referred to together as the mode method. The recursive feature elimination (RFE) and principal component analysis (PCA) algorithm was utilized in order to choose the features of relevance that are connected with the characteristics of importance for CKD diagnosis. These carefully chosen characteristics were provided to illness classifiers so that diagnoses could be made. In this work, the diagnosis of CKD was performed using a variety of classifiers, including SVM, KNN, LGBM, Xg, CatBoost, Ada, hybrid, and others. The machine learning model that has been suggested will achieve good classification performance with a limited number of features and will achieve optimal performance measures by using PCA.

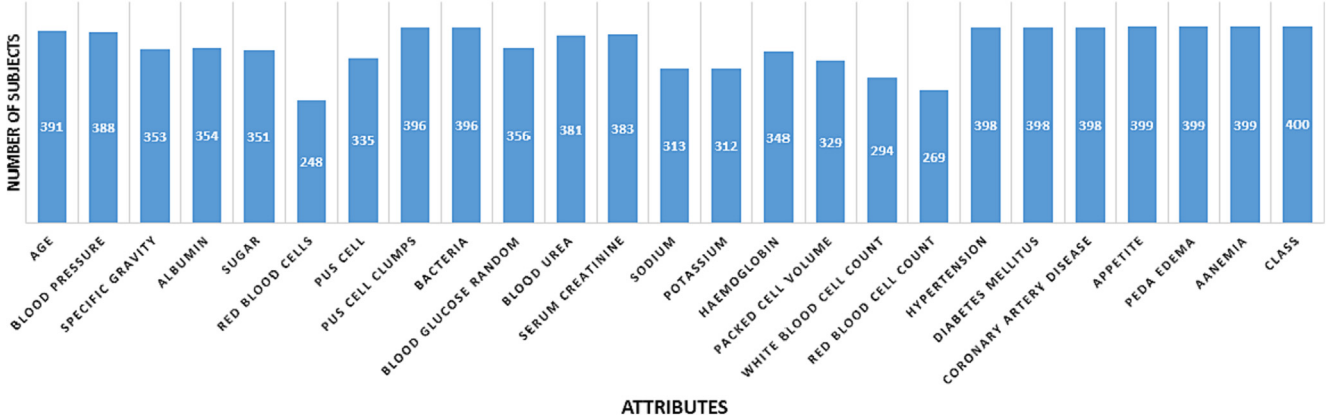

Fig. 1. CKD dataset used in previous paper.

Table 1
In-depth descriptions of each feature in the main CKD dataset.

CKD dataset	CKD dataset with PCA	Attributes meaning	Category	Scale	Missing
age	-	Age	Numerical	Years	9
bp	-	Blood pressure	Numerical	mm/Hg	12
sg	sg	Specific gravity	Nominal	1.005 to 1.025	47
al	al	Albumin	Nominal	0 to 5	46
su	su	Sugar	Nominal	0 to 5	49
rbc	-	Red blood cells	Nominal	Abnormal, Normal	152
pc	-	Pus cell	Nominal	Abnormal, Normal	65
pcc	-	Pus cell clumps	Nominal	Not present, Present	4
ba	-	Bacteria	Nominal	Not present, Present	4
bgr	bgr	Blood glucose random	Numerical	mgs/dl	44
bu	-	Blood urea	Numerical	mgs/dl	19
sc	sc	Serum creatinine	Numerical	mgs/dl	17
sod	-	Sodium	Numerical	mEq/L	87
pot	pot	Potassium	Numerical	mEq/L	88
hemo	-	Hemoglobin	Numerical	gms	52
pcv	pcv	Packed cell volume	Numerical	P cv	71
wc	wc	White blood cell count	Numerical	cells/cumm	106
rc	rc	Red blood cell count	Nominal	millions/cmm	131
htn	-	Hypertension	Nominal	No, Yes	2
dm	dm	Diabetes mellitus	Nominal	No, Yes	2
cad	-	Coronary artery disease	Nominal	No, Yes	2
appet	-	Appetite	Nominal	Poor, Good	1
pe	-	Peda edema	Nominal	No, Yes	1
ane	-	Anemia	Nominal	No, Yes	1
Classification	Classification	Class	Nominal	Not CKD, CKD	0

Data preprocessing

It is essential that the quality of the data be high in order for data mining methods to deliver efficient performance reasonable cost. The CKD dataset as a whole needs to have the variables that are missing from the database filled. When continuous characteristics are present, the approaches can be synchronized in order to construct discrete haracteristics in certain circumstances. Each instance of them has values that are noisy and has some that are missing. The original data is preprocessed to improve the behavior of medical data. The act of modifying the raw data and making it appropriate for use in a machine learning model is referred to as data preprocessing. The data preprocessing steps are shown in Fig. 3. It is important to draw attention to the fact that the data contains a significant number of unaccounted-for numbers. The specifics of each variable are shown in Table 1, which may be found here. The data of the 3 mentioned datasets should be cleaned up because they contain not-a-numbers (NaNs), and the numerical features should be converted to floats. Simply put, we were given the directive to get rid of all rows that had NaNs, and there was no threshold for this action; this means that any row that contained even a single NaN was to be removed.

The act of modifying the raw data and making it appropriate for use in a machine learning model is referred to as "data preprocessing." The data preprocessing steps are shown in Fig. 3. It is important to draw attention to the fact that the data contains a significant number of unaccounted-for numbers.

The data of the 3 mentioned datasets should be cleaned up because they contain not-a-numbers (NaNs), and numerical features should be converted to floats. Simply put, we were told to remove all rows with NaNs, and there as no threshold for this action; this meant that any row with even a single NaN was to be removed.

In order to make the processing in a computer easier, each category variable, also known as a nominal variable, was given a code. When it came to the values of rbc and pc, normal and abnormal were represented by the numbers 1 and 0, respectively, in the coding system. Both being present and not being present were assigned the numbers 1 and 0 for he pcc and ba values, respectively. The answers "yes" and "no" were assigned the values "1" and "0" for the variables htn, dm, cad, pe, and ane, respectively. For the appetizing value, the values 1 and 0 were assigned, respectively, to the categories of good and poor. Although the initial data description describes three variables—sg, al, and su—as being of a categorical type, the values of these 3 variables are still based on numeric information; hence, these variables were handled as though they were numeric variables. Every single one of the category variables was turned into a factor. An independent number ranging from 1 to 400 was assigned to every sample that was taken. The dataset is lacking a significant number of values, and there are only 158 instances that are complete. There are missing values in a considerable number. Before reaching a diagnosis, it's possible that the patients will be missing some measurements for a variety of reasons in general. As a consequence of this, missing values will be included in the data if the diagnostic categories of the samples are unknown and a technique that is analogous to imputation is necessary. Following the encoding of the categorical variables, the missing values in the initial CKD dataset were processed and filled up first. After the categorical variables had been encoded, this step was carried out. When there are missing values for numerical variables, those gaps are filled up using the median value of the relevant variable for the samples that have all of the data. On the other hand, the missing values for the category variables are filled in by selecting the category from the complete samples that appears the most frequently in the variable that corresponds to it. This is done so as to ensure that all of the samples are represented accurately.

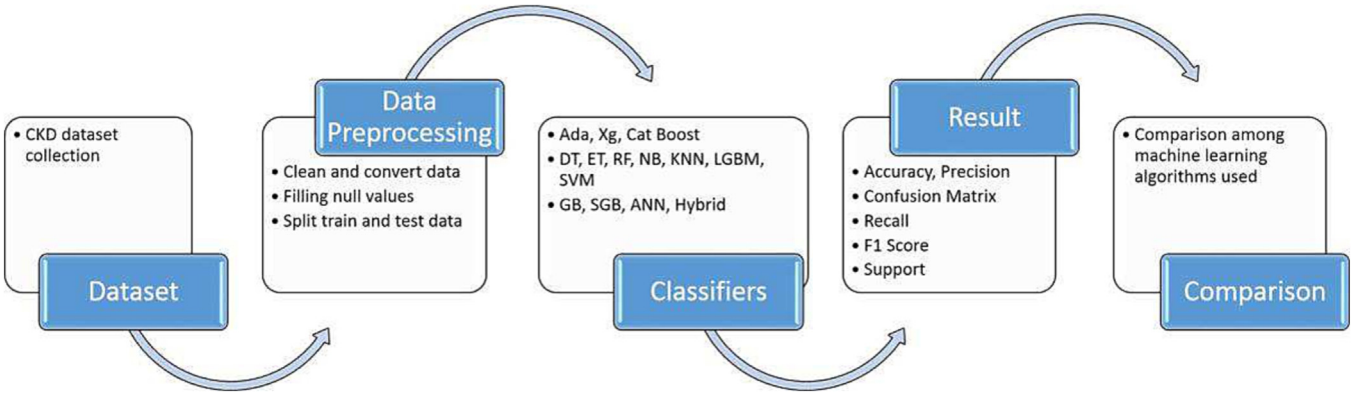

Fig. 2. Proposed model utilizing several machine learning classification algorithms.

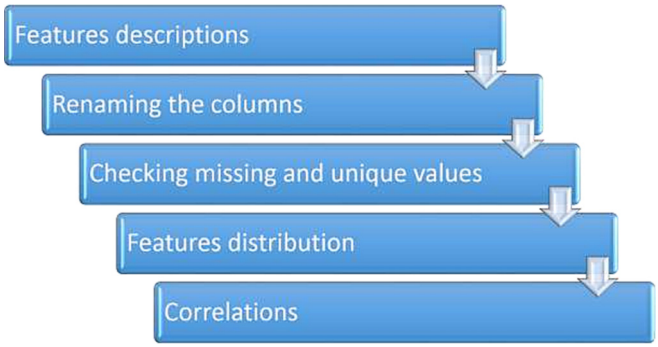

Fig. 3. CKD dataset used in this paper.

When it comes to physiological measurements, it makes sense for people whose physical conditions are comparable to have comparable physiological measurements. To fill in the missing numbers, the method based on a kNN is used since physiological measurements are expected to be comparable for individuals who are under similar physical situations. For instance, healthy people should have physiological measurements that are consistent within a given range. When comparing diseased individuals, the physiological measurements of a diseased individual should be comparable to those of a person with a similar degree of the same condition. In particular, there should not be significant discrepancies in the data collected from physiological measurements between individuals whose circumstances are comparable. This strategy, which has been utilized in the field of hyperuricemia, should also be adapted to the diagnostic data of other disorders.

The research was conducted using the CKD dataset.[11] There are 400 rows and 14 columns in this dataset. The output column "class" has a value of either "yes" or "no." The responses "yes" and "no" were assigned the values "1" and "0," respectively. The value "1" shows that the patient is a CKD patient, while the value "0" indicates that the patient is not a CKD patient. Fig. 4 shows the categorical columns view of the dataset without PCA. Fig. 5 shows a categorical view of the dataset with PCA where diabetes-mellitus is present over 250 instances who are not CKD patients and 140 instances who are CKD patients. In the target class distribution (Fig. 6), almost 250 rows are CKD patients and 150 are non-CKD patients.

A number of the features have categories that are not balanced. Stratified folds are going to be required for the cross-validation process. Let's also check to see if there is any imbalance between the classes.

A sample percentage of patients with chronic renal disease: 62.5%. The percentage of samples that do not have chronic renal disease is 37.5%. It is clear that there is not a significant imbalance between the classes.

In the heat map (Fig. 7), the absolute values of the correlations between the class label and features show that blood pressure, specific gravity, albumin, sugar, blood urea, serum creatinine, blood glucose random, and sodium all have positive links. whereas hemoglobin, potassium, white blood cell count, and red blood cell count have negative links. The heatmap of data shown in Fig. 7 can be observed to see the correlations between the features.

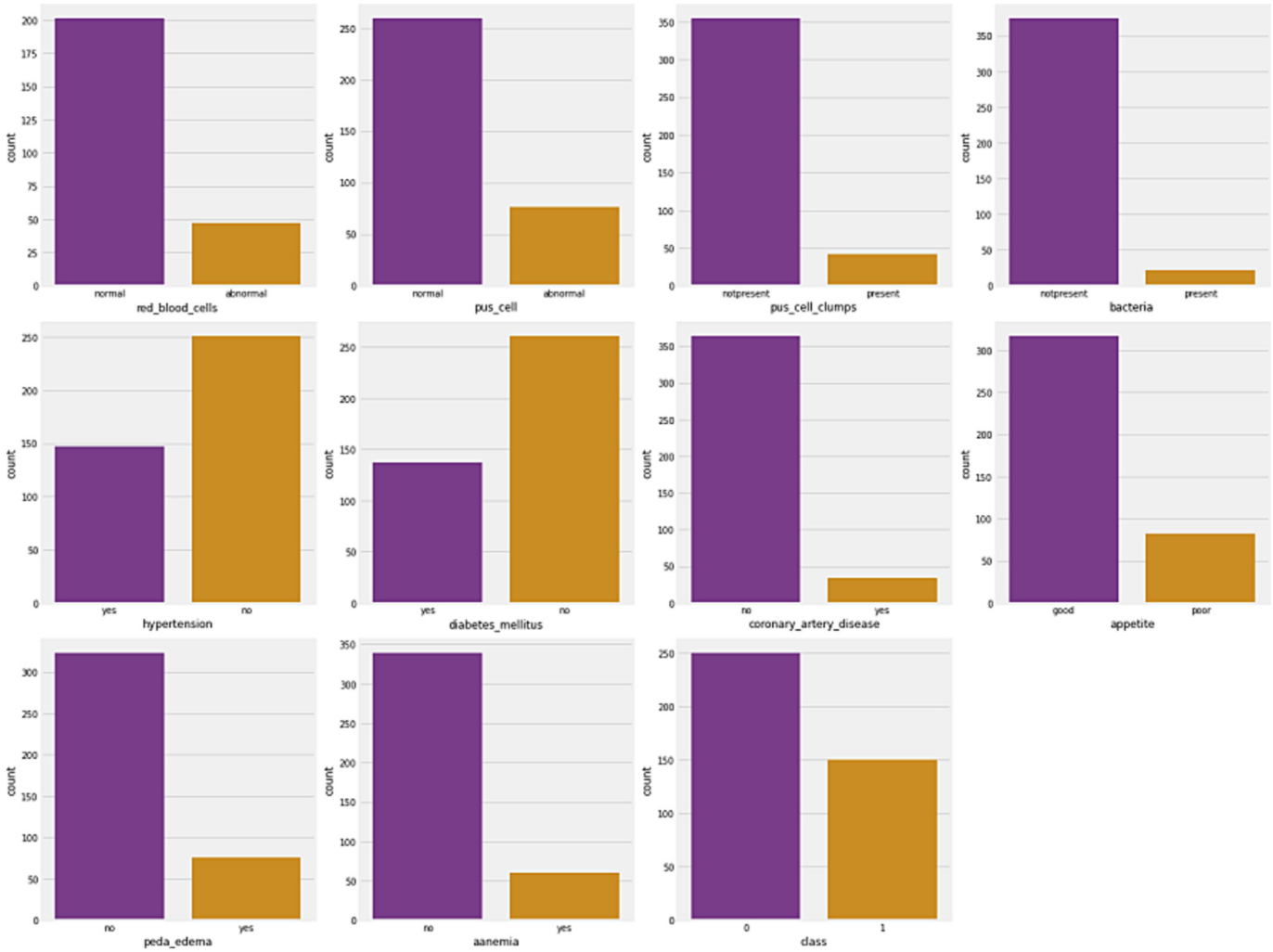

Fig. 4. Categorical columns view of the dataset.

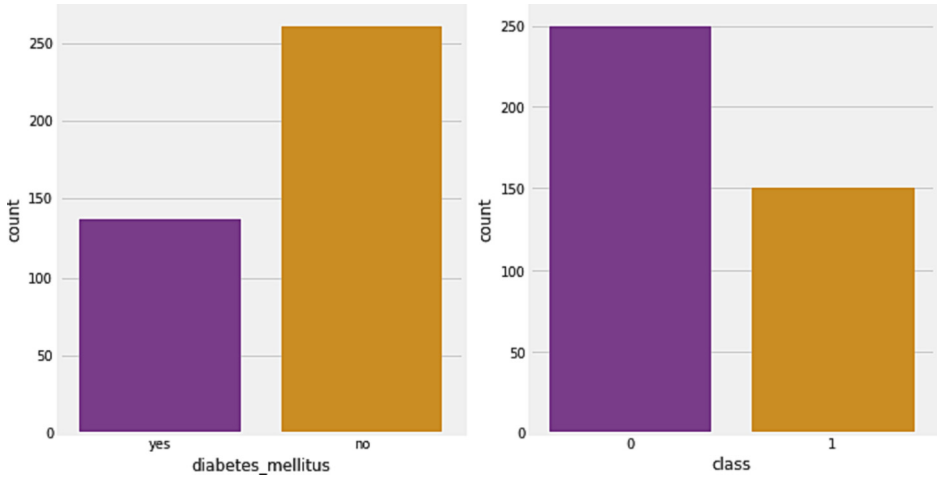

Fig. 5. Categorical columns view of the dataset with PCA.

The positive and negative correlations are shown in Fig. 8.

The dataset that is used is then split into 2 groups: the first group is used for testing the samples, and the second group is used for training the samples. The proportion of data obtained by testing is 30%, while the proportion obtained through training is 70%.

CKD dataset with PCA

Eliminating variables that are neither helpful for prediction nor connected to response variables can be accomplished by extracting feature vectors or predictors. Because of this, the building of the model would not be affected by variables that are not linked to the problem at hand, which would lead to the models making accurate predictions. Fig. 9 displays the results of the procedure for extracting important variables from the data.

The following machine learning models were developed for the goal of making a diagnosis of chronic kidney disease (CKD). Each model was produced by applying the matching subset of features or predictors to the whole CKD datasets. Generally, in sickness diagnosis, diagnostic samples are distributed in a multidimensional space. This section is where the predictors that are used during the data classification process may be found (CKD or non-CKD). The many categories that the samples of data fall under cause them to congregate in a variety of distinct parts of the space. As a result, there is a line that separates the 2 groups, and the distances that exist between the samples that are contained inside the same category

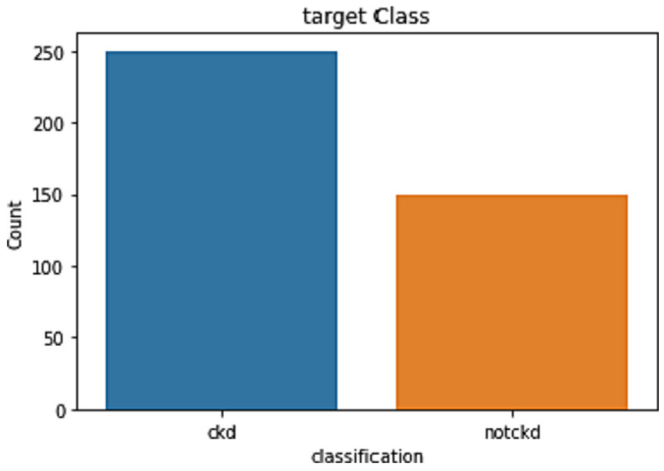

Fig. 6. Target class distribution.

are reduced. The aforementioned approaches to disease diagnosis are the ones that we use since they are the most effective in terms of classification.

Classifiers

In a supervised learning system, classification algorithms play a vital role in both training and testing. Classifier algorithms absorb knowledge from the training dataset and implement it in the testing dataset to generate the desired output.

AdaBoost

AdaBoost is an excellent machine learning approach for creating highly accurate prediction rules by combining and boosting relatively weak and inaccurate rules. It has a compact mathematical basis and increases the efficiency of multiclass classifier problems in practical applications. AdaBoost takes an iterative approach in order to improve the performance of weak classifiers by allowing them to study and improve from their own mistakes. The ability to reduce noise is improved when the AdaBoost is put into the stopping condition.[20]

Decision tree

When it comes to solving categorization issues, one of the most effective and widely used strategies for supervised machine learning is known as the decision tree. A decision tree is a type of tree structure that is similar to a flowchart. In a decision tree, each internal node represents a test that is performed on a feature, each branch represents the outcome of the test, and each leaf node contains a class label.[24] The decision tree starts with the root node of the tree, compares the value of various variables and then moves to the next branch until it reaches the end leaf node. In classification issues, the decision tree enquires, and based on the answers, it splits the data into subsequent sub branches. It makes use of many techniques to examine the population divide and parameters that allow for the most homogeneous sets[29].

XGBoost

XGBoost is commonly known to offer smart solutions to structured data problems through the implementation of the gradient boosted trees technique. Each regression tree in a gradient boosting regression setup acts as the weak learner, and it does so by assigning a continuous score to each input data point in the form of a leaf. XGBoost reduces a formalized objective function by merging a convex loss function based on the difference between the observed and target outputs with a weighting parameter for model computational complexity. Adding new trees that forecast the residuals or errors of earlier trees, which are then integrated with earlier trees to produce the final prediction, is how the training process is carried out iteratively.[13]

Fig. 7. The heatmap of data with correlation profile.

CatBoost

CatBoost is a decision tree gradient boosting technique that takes extremely little time to predict. It is used for a variety of functions, including weather forecasting, self-driving automobiles, personal assistants, and recommendation tems. In the majority of settings, CatBoost by default uses one-hot encoding for categorical features with a limited number of distinct values. Symmetric trees, also known as balanced trees, are used in CatBoost to describe trees where the splitting condition holds true for all nodes at the same level of the tree. These trees use the scaling feature exclusively for decision tree classification, not for catBoost.[13]

K-nearest neighbor (KNN)

The kNN is a simple supervised learning approach widely applied to resolve classification and regression issues. In kNN, the decision is made by calculating the Euclidian distances between a query and each example in the data, hoosing the value of the example (k) that is closest to the query, and either choosing the most common label for classification or the average of the labels for regression. However, this method is prohibited in dynamic web mining due to its lazy learning and dependency on the good value of k. The value of k is automatically chosen to increase the accuracy of the kNN algorithm.The kNN is a simple supervised learning approach widely applied to resolve classification and regression issues.[25] In kNN, the decision is made by calculating the Euclidian distances between a query and each example in the data, choosing the value of the example (k) that is closest to the query, and either choosing the most common label for classification or the average of the labels for regression. However, this method is prohibited in dynamic web mining due to its lazy learning and dependency on the good value of k. The value of k is automatically chosen to increase the accuracy of the kNN algorithm (Ayodele, 2010).

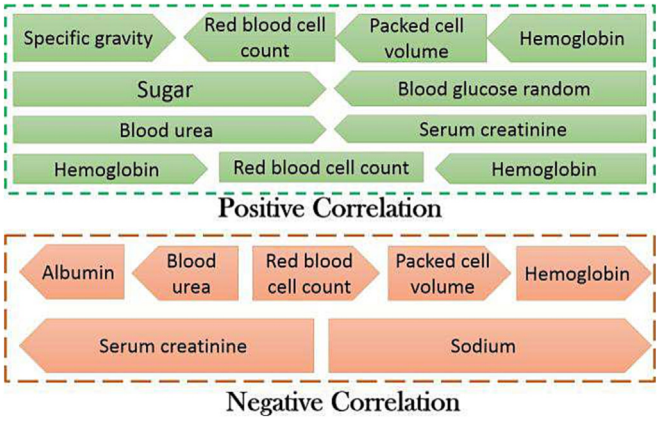

Fig. 8. Positive and negative correlations between the features.

Random forest

Random forest is a supervised learning classifier that associates a series of decision tree algorithms with various subsets of the provided datasets. It is capable enough to be used for large-scale problems and simple enough to be customized for various ad hoc learning tasks. To improve the prediction accuracy of the given dataset, it takes the average value from each tree and predicts the final outputs. A probabilistic machine learning technique called NB, which is based on the Bayesian theorem, has been successfully employed for a wide range of tasks, but it excels in solving natural language processing (NLP) issues. it used a simple mathematical formula for calculating conditional probabilities. However, its classification efficiency gradually falls if features are not independent and when the attributes are not independent, and it cannot handle continuous non-parametric characteristics.[20]

Gradient boosting

Gradient boosting (GB) classifiers are a type of machine learning method that brings together numerous weak learning models to develop a powerful predictive model. GB frequently makes use of decision trees. It

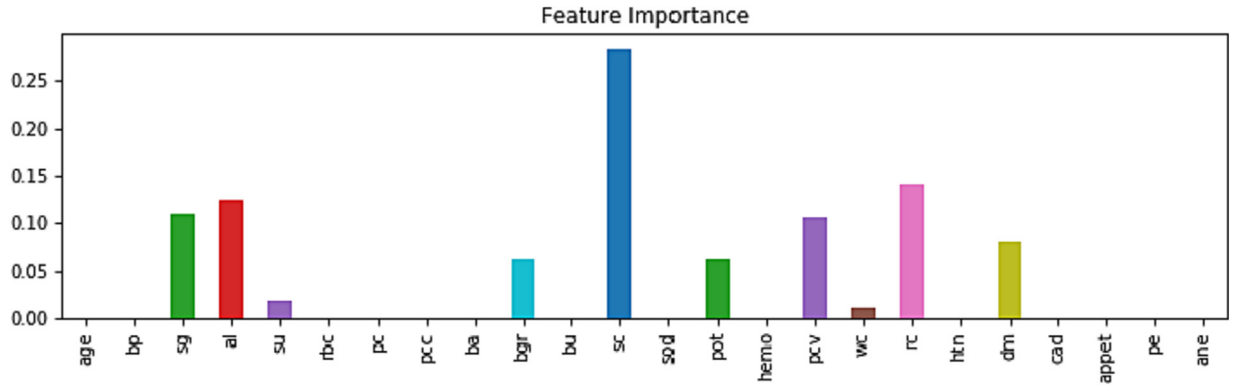

Fig. 9. Important features identification from the dataset.

is predicated on the hypothesis that when merged with earlier models, the best next model will minimize the overall prediction error. Setting the desired results for this subsequent model in order to reduce mistakes is the important concept. The gradient of the error with regard to the prediction is used to determine the goal outcomes for each case. Each model moves closer to making the best predictions feasible for each training example while minimizing prediction error.[13]

Stochastic gradient boosting

Stochastic gradient boosting, the training dataset is subsampled, and then each learner is trained using random samples produced by the subsampling, which provides the foundation of neural networks. By combining findings that have low correlation, this lowers the correlation between the outcomes from different students, giving us a better overall result. The gradient descent is approximated probabilistic. Because the algorithm only calculates the gradient for 1 randomly chosen observation at each step rather than the gradient for the entire dataset, the result is an approximation (Ayodele, 2010).

Light gradient boosting machine (LGBM)

LGBM is a dispersed, strong gradient boosting framework for sorting, classification, as well as data science application development that is based on the decision tree method and it requires less RAM to run while handling massive data sizes. When compared to other algorithms, Light GBM grows trees vertically, or leaf-wise, as opposed to other algorithms, which grow trees horizontally. It will be decided to grow the leaf with the highest delta loss. A leaf-wise method can reduce loss more than a level-wise strategy when expanding the same leaf.[13]

Extra tree

The Extra Trees method functions by combining the forecasts from various decision tree algorithms. An extra-trees regression employs averaging to increase predictive accuracy and reduce over-fitting. It does this by implementing a meta predictor that fits a number of randomized decision trees on different subsamples of the dataset. The Extra Trees approach is faster and shortens the process overall in terms of computational cost and execution time, but it arbitrarily selects the separation point and does not choose the best one.[13]

Support vector machine

Support vector machine (SVM) is a promising classical learning method for classification and regression problems and also solves various linear, non-linear, and practical difficulties. The statistical learning theory is the foundation of SVM and it projects targeted data using a kernel function to categorize in a high-dimensional feature space so that data points can be classified even though they are linearly non-separable (Ayodele, 2010).

Artificial neural network (ANN)

An artificial neural network (ANN) is a computational system that consists of a huge number of simple processors (neurons) that receive inputs and produce outputs by weight adjustment according to the defined activation functions. A feed-forward back-propagation learning technique is employed to train artificial neural networks by updating the weight coefficients efficiently in every epoch and minimizing errors. ANN systems intend to use a few "administrative" rules that are supposed to be applied to humans.[20]

Hybrid machine learning (HML)

HML usually integrates already-existing techniques or adds strategies from other domains to improve upon each other. HML is incredibly effective in estimating data and overcoming the limitations of current machine learning techniques. because no one machine learning technique is suitable for all problems, and while some techniques excel at handling noisy data, they might not be able to handle input spaces with many layers. Others may not be the best for handling sparse data even though they could scale rather well on high-layered input space. These conditions are a good reason to employ HML to improve the competing methods and to make up for their shortcomings.[13]

Performance metrics

In this particular investigation, the value of CKD was made positive, whereas the value of non-CKD was made negative. The performance of the machine learning models was evaluated with the help of the confusion matrix, which was utilized to display the specific results. Fig. 10 is an example of the format for the confusion matrix.[23]

A true positive (TP) result for the CKD samples indicates that the disease was accurately identified. In the case of the CKD samples, a false-negative (FN) result implies that an inaccurate diagnosis was made. In the case of non-KD samples, a false-positive (FP) result means that the samples were wrongly diagnosed. The term "true negative," abbreviated as "TN," denotes that the non-CKD samples were identified accurately. The performance of the model was evaluated using a number of different metrics, including

		Predicted CKD	
		ckd	not ckd
Ground Truth	ckd	TP	FP
	not ckd	FN	TN

Fig. 10. Confusion matrix.

accuracy, sensitivity, specificity, precision, recall, and F1 score.[23] These are determined by applying the formulae that are shown in Eqs. 1, 2, 3, and 4.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$Recall = Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{FP + TN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

Result and discussion

The results of each classifier have been reviewed with a variety of criteria for evaluation, and the 10-fold cross-validation method has been utilized to validate the results against overfitting.[23]

In addition, the layered cross-validation method has been utilized with the goal of fine-tuning the models' respective parameter settings. The Google Colab web application, written in Python, is used to run the experiments, which are ried out using that programming language. This project has utilized Sciket-learn,[23] which is open- source software for the machine learning library in Python, in several different ways. Accuracy, F1-score, precision, and recall are the evaluation metrics that were taken into consideration for this study.

The various values that are assigned to each model's parameters result in the production of distinct sets of outputs as shown in Table 2. The best performance for XgBoost was achieved with an accuracy of 0.9833 for the original CKD dataset. The accuracy has improved for the same classifier to 0.9916 with the dataset after implementing PCA. The adaBoost, random dorest, gradient boosting, LGBM, and Extra Tree show the same accuracy (0.9833) as the original CKD dataset. For this dataset, the highest accuracy is 0.9833, which is further improved to 0.9916 with the dataset after implementing PCA for the XgBoost classifier (Table 3). The performance of KNN and MLP is quite poor. The ANN classifiers show 60% accuracy for both datasets due to the lack of abundant data. The table compiles the results of the experiments conducted on each model, including their testing and training accuracy, F1-measure, precision, recall, and confusion matrix.

According to the evaluation results, all of the models have excellent performance in terms of detecting CKD with accuracy of greater than 97% by utilizing hemoglobin, specific gravity, and albumin, sugar, blood glucose random, serum creatinine, potassium, packed cell volume, white and red blood cell count, and diabetes mellitus characteristics, shown in Table 1. By concentrating on precision and recall, one can see that AdaBoost, XgBoost, LGBM, and Extra Tree have greater than 98%, with the exception of kNN and ANN, which indicates that most of the models were correct identifying the participants who did not have the disease or were healthy. Through the utilization of this model, an accurate prediction of chronic kidney disease can be made. It is an innovation that has the potential to assist the medical community in furthering their understanding of biomedical science.

However, certain restrictions placed on the dataset that was used are essential to this investigation. First, it is anticipated that the sample size, which is just 400 instances, will be low, which may have an effect on the dependability of the investigations. Second, the problem identification dataset is an additional dataset that contains the same attributes as the datasets to be evaluated for their performance.

Table 4 and Fig. 11 shows the comparison between previous studies and the proposed study. Chittora et al.[9] used kNN, SVM, and ANN classifiers to predict the CKD with highest accuracy of 96.5% for SVM. Almasoud and Ward[3] found 98.5% accuracy with RF classifier. Islam et al.[18] obtained slightly greater accuracy with RF classifier. SVM outperformed than other classifiers conducted by Gudeti et al.[14] Aljaaf et al.[2] got as usual accuracy shown in Fig. 11. In a nutshell, the previous studies were conducted based on kNN, RF, NB, GB, SVM, and ANN algorithms to predict the CKD based on the CKD dataset that is publicly available. We found highest accuracy with GB algorithm which was 0.990, where we found an accuracy of 0.983 with XgBoost and hybrid algorithms for CKD dataset. But the XgBoost algorithm shows the highest accuracy, which is 0.992. Our proposed model shows the highest accuracy till now.

Table 2
Performance of various algorithms for CKD original dataset.

Classifiers	Testing accuracy	Training accuracy	Confusion matrix	Precision	Recall	F1-score
Ada boost	0.975	1.0	72 0 3 45	0.98	0.97	0.97
Decision tree	0.96	0.97	71 1 4 44	0.96	0.96	0.96
XgBoost	0.983	1.0	72 0 2 46	0.98	0.98	0.98
CatBoost	0.966	1.0	71 1 3 45	0.97	0.97	0.97
KNN	0.65	0.796	47 25 17 31	0.66	0.65	0.65
Random forest	0.975	0.996	72 0 3 45	0.98	0.97	0.97
Naïve Bayes	0.94	0.94	66 6 1 47	0.95	0.94	0.94
Gradient boosting	0.975	1.0	72 0 3 45	0.98	0.97	0.97
Stochastic gradient boosting	0.975	1.0	72 0 3 45	0.98	0.97	0.97
LGBM	0.983	1.0	72 0 2 46	0.98	0.98	0.98
Extra tree	0.975	1.0	72 0 3 45	0.98	0.97	0.97
SVM	0.93	0.97	65 7 1 47	0.94	0.93	0.93
ANN	0.60	0.64	72 0 48 0	0.36	0.60	0.45
Hybrid	0.9833	0.975	71 1 1 47	0.98	0.98	0.98

Table 3
Performance of various algorithms for CKD dataset with PCA.

Classifiers	Testing accuracy	Training accuracy	Confusion matrix	Precision	Recall	F1-score
Ada boost	0.983	1.0	72 0 2 46	0.98	0.98	0.98
Decision tree	0.975	0.98	72 0 3 45	0.98	0.97	0.97
XgBoost	0.9916	1.0	72 0 1 47	0.99	0.99	0.99
CatBoost	0.975	0.985	72 0 3 45	0.98	0.97	0.97
KNN	0.59	0.76	50 22 27 21	0.58	0.59	0.59
Random forest	0.975	0.99	72 0 3 45	0.98	0.97	0.97
Naïve Bayes	0.8833	0.9	62 10 4 44	0.89	0.88	0.88
Gradient boosting	0.975	1.0	72 0 3 45	0.98	0.97	0.97
Stochastic gradient boosting	0.975	1.0	72 0 3 45	0.98	0.97	0.97
LGBM	0.983	1.0	72 0 2 46	0.98	0.98	0.98
Extra tree	0.9833	1.0	72 0 2 46	0.98	0.98	0.98
SVM	0.9666	0.946	70 2 2 46	0.97	0.97	0.97
ANN	0.6	0.6357	72 0 48 0	0.36	0.60	0.45
Hybrid	0.958	0.978	71 1 4 44	0.96	0.96	0.96

Table 4
Comparison between previous studies and the proposed study.

Classifiers	Previous studies					This study	
	PANKAJ CHITTORA	Marwa Almasoud	Md. Ashiqul Islam	Bhavya Gudeti	Ahmed J. Aljaaf	CKD dataset	CKD dataset with PCA
Ada boost						0.975	0.983
Decision tree						0.960	0.975
XgBoost						0.983	0.992
CatBoost						0.966	0.975
KNN				0.788		0.650	0.590
Random forest	0.910	0.985	0.989			0.975	0.975
Naïve Bayes			0.940			0.940	0.883
Gradient boosting		0.990				0.975	0.975
Stochastic gradient boosting						0.975	0.975
LGBM						0.983	0.983
Extra tree						0.975	0.983
SVM	0.965	0.970		0.993	0.950	0.930	0.967
ANN	0.960				0.981	0.600	0.600
Hybrid						0.983	0.958

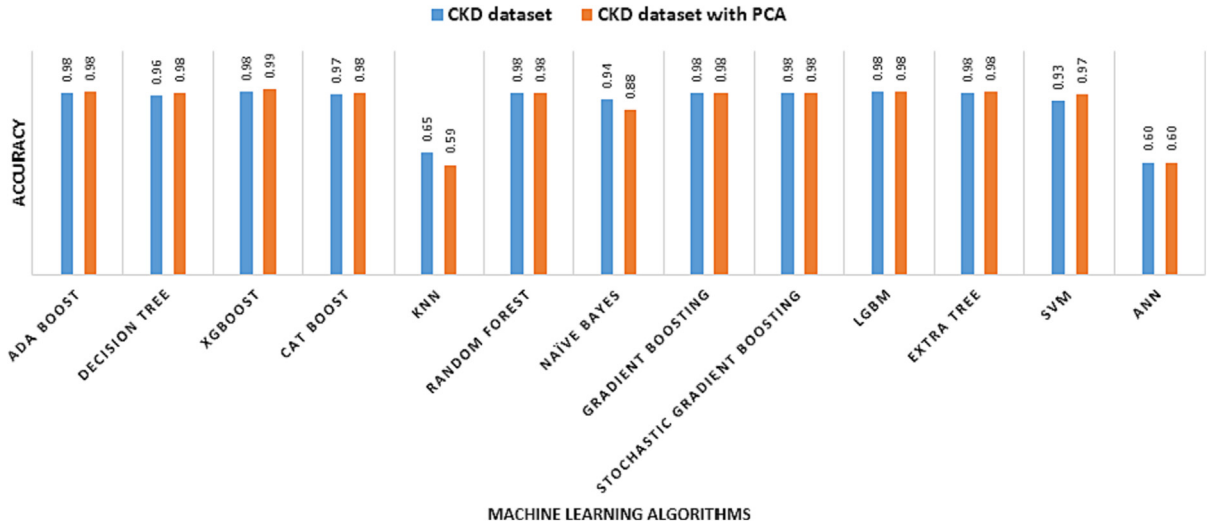

Fig. 11. Performance of the proposed model.

CKD will be especially useful in this setting for forecasting outcomes by identifying or listing persons who are at risk. When treating patients, it will be very helpful to list those at risk and assign them scores to forecast the outcomes of their treatment. However, a sizeable portion of the population at high risk of chronic kidney disease (CKD) can still be recognized or identified within the community by using CKD risk factor predicting without having to be admitted to a hospital for treatment. This is possible through the use of CKD risk factor predicting. This method has the advantage of requiring much less time for the prediction process, which enables medical professionals to begin treatment for patients with CKD at the earliest possible stage and to further classify a wider population of patients in a shorter amount of time.

Furthermore, this methodology may be relevant to the clinical data of various disorders that are being diagnosed during the course of actual medical practice. On the other hand, throughout the process of constructing the model, due to constraints of the conditions, the data samples that are available are very limited, including just 400 samples in total. Having said that, this investigation does have some restrictions that come from the dataset that was employed. To begin, the size of the dataset, which is just 400 occurrences, is deemed to be rather low, which may have an effect on the dependability of the results. Second, it is difficult to locate another dataset that has the same features as this one in order to make a meaningful comparison between the 2 sets of data.

In the future, a significant amount of data that is both more sophisticated and representative will be gathered for the purpose of training the model in order to increase its generalization performance while also enabling it to determine the severity of the condition. We have faith that as the quantity and quality of the data continues to improve, this model will evolve into a more refined and accurate representation of reality.

Conclusion

This study presented a number of different machine learning algorithms with the intention of making a CKD diagnosis at an earlier stage. The models that are constructed using CKD patients are then trained and validated using the input parameters that were discussed earlier. Studies have been done on the associations between different factors so that the number of features can be cut down and redundant information eliminated. When applying a filter feature selection approach to the remaining attributes, it was discovered that hemoglobin, albumin, and specific gravity had the biggest impact when it comes to predicting CKD. This was the case after the method was used. This work presented a number of different machine learning algorithms with the intention of making a CKD diagnosis at an earlier stage. The original CKD dataset has been preprocessed first to validate the machine learning-based detection models. After that, the PCA has been performed to identify the most dominant features, thereby detecting CKD. The models that are constructed using CKD patients are then trained and validated using the input parameters that were discussed earlier. The accuracy of such algorithms was the primary criterion that was utilized in evaluating their overall performance.

Declaration of Competing Interest

All authors have participated in

(a) conception and design, or analysis and interpretation of the data;

(b) drafting the article or revising it critically for important intellectual content; and

(c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

Author's name	Affiliation
Md. Ariful Islam	Department of Robotics and Mechatronics Engineering, University of Dhaka, Dhaka, Bangladesh
Md. Ziaul Hasan Majumder	Institute of Electronics, Bangladesh Atomic Energy Commission
Md Alomgeer Hussein	Department of Electrical and Electronic Engineering, University of Dhaka, Dhaka, Bangladesh

References

1. Akhter T, Islam MA, Islam S. Artificial neural network based covid-19 suspected area identification. J Eng Adv 2020;1:188–194.
2. Aljaaf AJ, Al-Jumeily D, Haglan HM, et al. Early prediction of chronic kidney disease using machine learning supported by predictive analytics. 2018 IEEE Congress on Evolutionary Computation (CEC). IEEE; 2018. p. 1–9.
3. Almasoud M, Ward TE. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int J Soft Comput Appl 2019;10.
4. Ani R, Sasi G, Sankar UR, Deepa O. Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE; 2016. p. 1287–1292.Arora, M., Sharma, E.A., 2016. Chronic kidney disease detection by analyzing medical datasets in weka. International Journal of Computer of machine learning algorithms. New advances in machine learning 3, 19–48.
5. Arora M, Sharma EA. Chronic kidney disease detection by analyzing medical datasets in weka. Int J Comput Mach Learn Algor New Adv Mach Learn 2016;3:19–48.
6. Banik S, Ghosh A. Prevalence of chronic kidney disease in Bangladesh: a systematic review and meta-analysis. Int Urol Nephrol 2021;53:713–718.
7. Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N. Predictive analytics for chronic kidney disease using machine learning techniques. 2016 Management and Innovation Technology International Conference (MITicon). IEEE; 2016.pp. MIT–80.
8. Chen Z, Zhang X, Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. Int Urol Nephrol 2016;48:2069–2075.
9. Chittora P, Chaurasia S, Chakrabarti P, et al. Prediction of chronic kidney disease-a machine learning perspective. IEEE Access 2021;9:17312–17334.
10. Cueto-Manzano AM, Cortés-Sanabria L, Martínez-Ramírez HR, Rojas-Campos E, Gómez-Navarro B, Castillero-Manzano M. Prevalence of chronic kidney disease in an adult population. Arch Med Res 2014;45:507–513.
11. Dua D, Graff C. UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml 2017.
12. Eroğlu K, Palabaş T. The impact on the classification performance of the combined use of different classification methods and different ensemble algorithms in chronic kidney disease detection. 2016 National Conference on Electrical, Electronics and Biomedical Engineering (ELECO). IEEE; 2016. p. 512–516.
13. Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intel Learn Syst Appl 2017;9(01):1.
14. Gudeti B, Mishra S, Malik S, Fernandez TF, Tyagi AK, Kumari S. A novel approach to predict chronic kidney disease using machine learning algorithms. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE; 2020. p. 1630–1635.
15. Heung M, Chawla LS. Predicting progression to chronic kidney disease after recovery from acute kidney injury. Curr Opin Nephrol Hypertens 2012;21:628–634.
16. Islam M, Shampa M, Alim T, et al. Convolutional neural network based marine cetaceans detection around the swatch of no ground in the bay of bengal. Int J Comput Digit Syst 2021;12:877–893.
17. Islam, M.A., Akhter, T., Begum, A., Hasan, M.R., Rafi, F.S. Brain tumor detection from MRI images using image processing.
18. Islam MA, Akter S, Hossen MS, Keya SA, Tisha SA, Hossain S. Risk factor prediction of chronic kidney disease based on machine learning algorithms. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE; 2020. p. 952–957.
19. Islam MA, Hasan MR, Begum A. Improvement of the handover performance and channel allocation scheme using fuzzy logic, artificial neural network and neuro-fuzzy system to reduce call drop in cellular network. J Eng Adv 2020;1:130–138.
20. Mahesh B. Machine learning algorithms-a review. Int J Sci Res (IJSR)[Internet] 2020;9:381–386.
21. Polat H, Danaei Mehr H, Cetin A. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. J Med Syst 2017;41:1-11.

22. Qin J, Chen L, Liu Y, Liu C, Feng C, Chen B. A machine learning methodology for diagnosing chronic kidney disease. IEEE Access 2019;8:20991–21002.
23. Rácz A, Bajusz D, Héberger K. Multi-level comparison of machine learning classifiers and their performance metrics. Molecules 2019;24:2811.
24. Ravizza S, Huschto T, Adamov A, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nat Med 2019;25:57–59.
25. Saringat Z, Mustapha A, Saedudin RR, Samsudin NA. Comparative analysis of classification algorithms for chronic kidney disease diagnosis. Bull Elect Eng Inform 2019;8:1496–1501.
26. Subasi A, Alickovic E, Kevric J. Diagnosis of chronic kidney disease by using random forest. CMBEBIH 2017. Springer; 2017. p. 589–594.
27. Wickramasinghe M, Perera D, Kahandawaarachchi K. Dietary prediction for patients with chronic kidney disease (ckd) by considering blood potassium level using machine learning algorithms. 2017 IEEE Life Sciences Conference (LSC). IEEE; 2017. p. 300–303.
28. Zhang L, Wang F, Wang L, et al. Prevalence of chronic kidney disease in china: a cross-sectional survey. The Lancet 2012;379:815–822.
29. Ayodele Olugbenga E, Alebiosu C Olutayo. Burden of chronic kidney disease: an international perspective. Adv Chronic Kidney Dis 2010;17(3):215–224.Elsevier.