Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 11(112)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10101788

Machine learning algorithms for predicting of chronic kidney disease and its significance in healthcare

B. Yamini[1], T. Saraswathi[2], P. Radhakrishnan[3], M. Nalini[4*], M. Shanmuganathan[5], and Siva Subramanian.R[6]

Associate Professor, Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur-603203[1]

Assistant Professor, Department of Information Technology, Easwari Engineering College, Ramapuram, Chennai-600089[2]

Assistant Professor, Department of Computer Science & Artificial Intelligence, SR University, Ananthsagar, Hasanparthy, Warangal Urban, Telangana-506371[3]

Associate Professor, Department of Computer Science and Engineering, S.A. Engineering College, Poonamalle-600077[4]

Associate Professor, Department of Computer Science and Engineering, Panimalar Engineering College, Chennai - 600123[5]

Associate Professor, Department of Computer Science and Engineering, R.M.K College of Engineering and Technology, Puduvoyal-601206[6]

Received: 04-June-2023; Revised: 05-March-2024; Accepted: 09-March-2024

Abstract

Chronic kidney disease (CKD) is a progressive disorder that worsens over time, leading to a variety of major health problems including hypertension, anaemia, nerve damage, fractured bones, and cardiovascular diseases. This rate is growing globally by the development of ageing and diabetes and hypertension. Precise prediction of CKD may aid healthcare practitioners in developing tailored treatment plans that address the distinct underlying causes of the illness, reduce the likelihood of complications, and improve patient outcomes. Predicting the course of CKD is crucial because early detection and accurate diagnosis of CKD may improve patient outcomes, prevent complications, and save medical costs. Machine learning (ML) based techniques are taken into consideration for CKD prediction. Several ML algorithms are used, including decision tree (DT), neural networks (NN), random forest (RF), XGBoost (XGB), Gaussian Naive Bayes (GNB), and CatBoost (CB) Classifier. Using six distinct ML algorithms, the kidney disease dataset is used to carry out the empirical method. Different validity ratings are used to evaluate and contrast the generated findings. The results show RF, XGB and CB get higher accuracy of 95 % which is better suited in predicting CKD than DT, NN, and GNB. These algorithms illustrate the highest accuracy and efficiency in the diagnosis of the patients at the beginning CKD stage. An accurate prediction of CKD enables practitioners in the healthcare sector to develop individualized treatment plan, avert the occurrence of complications and achieve a better patient outcome. Early recognition and intervention based on predictive models can prevent unproductive testing, treatments, and hospital beds while also responding quickly to disease management. The predictive models of CKD can identify high-risk patients and facilitate early interventions, ultimately enhancing public health outcomes.

Keywords

Machine learning, Chronic kidney disease, Random Forest, XGBoost, CatBoost, Healthcare.

1.Introduction

A long-term illness known as chronic kidney disease (CKD) develops when the kidneys are injured & can't to adequately filter blood.

When kidneys are not working properly, the body's waste and extra fluid may build up and harm other organs since the kidneys are in charge of eliminating waste and excess fluid. Diabetes, high blood pressure, autoimmune illnesses, infections, blockages of the urinary system, and hereditary conditions are just a few of the numerous potential causes of CKD [1]. Over time, kidney disease may also be brought

*Author for correspondence

on by certain drugs and poisons. CKD normally proceeds in phases and develops slowly over months or years. Patients may not have any symptoms in the early stages, but as the illness develops, they may start to feel weak, dizzy, and sick, lose their appetite, swell, and have trouble sleeping. Patients who are in advanced stages can need kidney transplants or dialysis. CKD is a significant ailment that has to be managed and monitored constantly. Treatment options may include blood pressure and blood sugar management drugs, dietary modifications to lessen the strain on the kidneys, and, in certain circumstances, dialysis or kidney transplantation. For the management of CKD and the prevention of associated consequences, early identification and treatment are essential [2].

Numerous factors make the study of CKD relevant. Prevalence: CKD affects millions of individuals globally and is a prevalent, serious public health issue. The ageing of the population, increased rates of diabetes and high blood pressure, and the incidence of CKD are all contributing factors. Understanding CKD's prevalence and effects on public health requires research. Studying CKD enables the development of screening and preventive measures, risk factor identification, and early detection and slowed progression of CKD. Management and therapy: To avoid problems and enhance results, CKD needs continual management and therapy. Studying CKD advances the management and treatment of people with CKD, identifies useful therapies, and enhances patient care. Economic burden: CKD is linked to high medical expenses, such as those related to dialysis and kidney transplantation. Researching CKD identifies cost-effective healthcare solutions and estimates the disease's economic impact. In conclusion, it is crucial to investigate CKD in order to comprehend its prevalence, recognise risk factors, create screening and preventive plans, and enhance management and treatment, and lower healthcare costs [3].

It has become more commonplace to detect and diagnose CKD using machine learning (ML) algorithms. These algorithms are able to analyse big datasets and spot intricate patterns and correlations that conventional statistical approaches could miss. ML models are capable of learning from historical data & forecasting future events by using a variety of methods and methodologies. ML models in CKD may be utilised for a variety of tasks, including determining the optimal course of therapy for specific patients and predicting the likelihood of getting the illness [4]. To find trends and characteristics linked to the onset and progression of CKD, for instance, ML models may be trained on substantial datasets of patient information, including demographic data, medical history, laboratory findings, and imaging data. The risk of CKD in new patients may then be predicted using these models based on their unique traits and medical history. By examining patient data over time and finding characteristics linked to disease development, ML models may also be used to forecast how CKD will proceed. High-risk patients who may need more rigorous monitoring and care can be identified using this information. Finally, depending on a patient's unique traits and medical history, ML models may be utilised to determine the best therapy alternatives for that patient. ML models may assist inform treatment choices for specific patients by identifying the therapies that are most successful for certain patient types by analysing massive datasets of patient outcomes. Overall, the use of ML models in CKD has the potential to enhance patient diagnosis, prognosis, and treatment results while assisting medical personnel in making better choices about patient care [5].

The objective of this study is to apply ML methods, such as decision tree (DT), neural networks (NN), random forest (RF), XGBoost(XGB), Gaussian Naive Bayes(GNB), and CatBoost(CB) classifier, to predict CKD. The research uses several validity metrics to compare the effectiveness of these algorithms. The goal of this work is to aid clinical professionals in making well-informed choices regarding patient care while also improving patient diagnosis, prognosis, and treatment results. The contributions of this study lie in providing insights into the achievement of various ML model for CKD prediction and highlighting the significance of accurate CKD prediction in clinical practice. The findings indicate the potential for early intervention, personalized treatment strategies, and cost-effective management of CKD, ultimately leading to enhanced patient outcomes and a reduction in healthcare expenditures.

The imperative nature of addressing these gaps in the literature is the driving force for our study. Early detection & accurate prediction of CKD may greatly improve patient outcomes by allowing prompt interventions and individualised treatment plans. A potential approach for CKD prediction is the use of ML algorithms, which can analyse intricate patterns and correlations in large datasets that conventional statistical approaches might overlook.

The major contributions are as under:

1. Providing a detailed analysis of key elements in order to provide a deeper view of CKD.
2. Assessing, contrasting, and providing information on the effectiveness of various ML algorithms for the prediction of CKD.
3. Emphasising the significance of precise CKD prediction for early intervention and better patient outcomes in clinical practise.

This paper structures as follows: Section 2 discusses related works on CKD research, highlighting problems and gaps. Section 3 details the procedure, including experimental design, ML techniques, and dataset. Section 4 presents and analyzes results from ML methods. Section 5 debates these results, comparing algorithmic performances. Section 6 summarizes conclusions, implications, and future research directions, aiming to enhance ML's role in CKD prediction and management.

2.Literature survey

In this section related work was discussed with the major advantages and challenges of the approaches. Raju et al. [6] implies the importance of CKD in this work and uses several ML techniques to diagnose CKD based on the research. Several ML techniques, including support vector machine (SVM), RF, XGB, logistic regression (LR), NN, and naïve bayes (NB), are used in this study. Different validity scores are used to compare and project the outcomes of the experimental process carried out utilising various ML. The analysis suggests that when compared to others, XGB and RF provide superior forecast accuracy. Sinha et al. [7] alludes to a significant issue in the healthcare sector that makes people's lives unpleasant. The author mentions CKD research and how important it is for early diagnosis. In this instance, the author does CKD analysis using several ML. SVM and k-nearest neighbors (K-NN) are used in ML. Different validity ratings are used to compare and project the experimental approach using SVM and K-NN findings. The comparison suggests that K-NN has a higher prediction accuracy than SVM. Almansour et al. [8] aims to use the ML approaches to assist and diagnosis CKD. Based upon the study the author applies ML to perform CKD analysis. In ML, artificial neural networks (ANN) and SVM models are applied. The experimental procedure is carried out using CKD dataset with ANN and SVM models and results obtained are compared and projected. The experimental result implies ANN woks better with accuracy of 99.75 compare to SVM.

Pasadana et al. [9] carries out a significant investigation into CKD analysis and suggests the use of ML to CKD analysis. The author uses ML methods including logical model tree (LMT), NB Tree, RF, J48, random tree (RT), J48graft, reduced error pruning (REP) tree, connectionist temporal classification (CTC) and Simplecart based on the research. The CKD dataset and several ML models are used to conduct the experimental approach. According to the experimental findings, RF predicts more accurately than LMT, NB Tree, RF, J48, RT, J48graft, REPTree, CTC and Simplecart.

Saha et al. [10] performs a deep study on importance of CKD and its earlier prediction to save lives. Based upon the research the author applies ML models to perform CKD analysis. In ML models like RF, NB, multilayer perceptron (MLP), LR and NN optimized by Adam are applied. The experimental procedure is carried out using CKD dataset obtained from National Kidney foundation Bangladesh with different ML models. The experimental result implies NN optimized by Adam performs superior compare to RF, NB, MLP, and LR.

The prevalence of CKD is rising worldwide, and its development must be stopped by early identification. This study analyses patient data to identify CKD infections using eight ML methods. The research analyses their performance using various scores. Gradient boost is only 98% accurate, whereas K-NN and extra tree classifier both reach 99% accuracy [11].

A worldwide health emergency known as CKD is brought on by poor dietary choices and insufficient water intake. Using diagnostic medical data from the University of California Irvine (UCI) repository, this work intends to create a predictive model for CKD prognosis. Data pre-processing, missing value management, aggregation, extraction, and classification are all included in the procedure. With 98.75% accuracy, 100% sensitivity, 96.55% specificity, and 99.03% f1score, a performance tuning layered strategy is suggested. This method might be utilised to create an automated system for assessing the severity of CKD [12].

A disorder known as CKD damages the kidneys and has an impact on general health. End-stage renal illness and patient mortality might result from inadequate diagnosis and treatment. The goal of this study is to create effective ML methods for predicting the incidence of CKD. Rotation Forest

(RotF) demonstrated the best accuracy at 99.2% and was employed in class balance, features ranking, and performance metrics assessments[13].

Glomerular filtration rate (GFR) and kidney damage indicators are used to diagnose CKD, a disorder brought on by high blood pressure and diabetes. This study offers a deep learning (DL) model for early identification and prediction of CKD, a condition for which researchers struggle to make an early diagnosis. The model beats existing classifiers, obtaining 100% accuracy, and employs recursive feature elimination to choose important features [14].

A common condition with substantial risks, such as cardiovascular and end-stage renal disease, is CKD. Accurate illness diagnosis and treatment are made possible by ML algorithms [15]. This study suggests combining feature selection strategies with classification algorithms built on big data frameworks like Apache Spark. Six different ML classification techniques that were utilised. The findings shown that relief-F's chosen features outperformed complete features and chi-squared features to attain the greatest performance at 100% accuracy for the SVM, DT, and Gradient-Boosted Trees (GBT) classifiers.

The prevalent and incurable illness known as CKD damages the kidneys and impairs their capacity to cleanse blood. The only choices are dialysis or kidney transplantation, both of which are expensive and time-consuming. ML methods are being utilised to diagnose CKD in order to solve this problem. The SVM, K-NN, RF, and DT, ML algorithms are the four that are used in this work. These algorithms are trained on a dataset from the UCI repository, and their predictive power for CKD is increased by using data preparation methods [16].

The difficulty of manual diagnosis and growing physician workload is addressed by the research, which provides a ML approach for predicting CKD. Four ML methods are used by the system: SVM, RF, DT, and NB. Kidney illness is predicted using NB using probability, and reports are categorised by DT. When the accuracy scores of each approach are compared, RF performs the best, with an accuracy score of 98.75% [17].

A major health problem brought on by gradual kidney dysfunction is CKD. Severe harm may be avoided with an early diagnosis. In order to determine if a person has CKD or may develop it in the future, this study suggests using a ML model. Utilising an AdaBoost classifier, DT, RF, gradient boosting (GB), NB, LR, and other methods, the model makes use of a dataset from the UCI repository. GB yields an accuracy of 98.22% for the model [18].

The objective of the proposed study is to create and verify a prediction model for assessing CKD, a potentially fatal illness that often affects adults and the elderly. DT Classifier will be used for data input, while NB Classifier will be used for analysis. As a predictor, a clustering technique will be used, and different clusters for every category will be investigated. Principal components analysis (PCA) will be used to train and compare two NN of varying sizes [19].

CKD is a potentially fatal illness that affects a large number of people worldwide. Because it doesn't cause any symptoms, an early diagnosis is essential. Large amounts of data from electronic health records have made ML-based methods for precisely diagnosing this illness possible. This study used the SVM and LR techniques to create an effective clinical diagnostic system. The UCI repository provided the data set. The system employs ML approaches such as hyperparameter tuning to discover minimum relevant information and focuses on accurate and economical prediction of CKD. With an accuracy of 99.17%, the suggested SVM model showed the potential benefit of precise diagnosis in reducing human loss [20].

With the prevalence of kidney transplants and dialysis, CKD, is becoming a bigger problem. Early CKD prediction is critical, and ML may help with this. This study suggests a methodology to predict the state of CKD from clinical data by means of pre-processing, managing missing values, and collaborative filtering. Several methods are used by the model to guarantee accuracy and dependability. The objective is to overcome any overfitting problems caused by inadequate regularisation, noise, or an excess of features in the data in order to develop a reliable model for predicting CKD [21].

The progressive decrease of kidney function associated with CKD impairs the kidneys' capacity to remove waste products from the blood. Dialysis, kidney transplantation, or organ failure may result if it becomes worse. Complications from CKD might include low blood pressure, high blood pressure, weakening of the bones, poor health, and harm to the nervous system. Prompt diagnosis is essential to

avoid deterioration. This research compares several classification algorithms and ensemble stacking techniques with the goal of developing a model to predict CKD using medical variables and ML algorithms [22].

Because CKD has a high risk of sickness and death, it is a serious health problem. Slow-moving, chronic infections are difficult to diagnose, so early diagnosis is essential. The use of ML techniques is essential for diagnosing diseases in their early stages. The goal of this study is to use meta classifiers to evaluate the risk phase of CKD. The method entails pre-processing conventional data using normal scalar and label encoding, then applying basic classifiers like RF and K-NN. Kidney infection risk stage may be determined by using additional tree classifiers as meta classifiers [23].

With a maximum survival period of just eighteen days, CKD is an increasingly common condition that requires early identification by ML algorithms. Using collaborative filtering, attribute selection, data management, and clinical data collecting, this proposal suggests a workflow for forecasting the condition of CKD. The approach guarantees little bias and excellent accuracy for characteristics by using seven machine models. The study places a strong emphasis on domain expertise and real-time data collecting in ML for CKD status prediction. The accuracy of the model is 98.65% [24].

Using ML methods, it is possible to forecast the major health problem of CKD. To evaluate data and choose the optimal model, researchers used a variety of techniques, including LR, DT, RF, and SVM classifiers. With all characteristics, the RF model had the maximum accuracy of 99.1%; with the top five features, the accuracy was 93.5% and 85%, respectively. Additionally, performance analysis and method selection were done using the confusion matrix's real negative values [25].

Chronic illnesses, which last more than three months, are the leading cause of death worldwide owing to their sluggish development and ability to respond to medical therapy. These illnesses are classified as cancer, heart disease, diabetes, asthma, Parkinson's disease, arthritis, rheumatoid arthritis, stroke, and renal disease. Identifying and identifying these illnesses in their early stages is critical, but clinicians struggle to detect and treat them properly [26]. The goal of this study is to develop a highly interpretable model for predicting and detecting CKD using big data analytics, resulting in increased accuracy, efficacy, and reduced computational and temporal complexity.

Early illness prediction and detection, especially in the early stages of CKD, may be greatly aided by artificial intelligence (AI) and ML. This survey compares the accuracy and applicability of several algorithms from earlier research on ML for early-stage CKD prediction. The objective is to identify research gaps and provide a summary of current findings. The accuracy of several ML methods for early CKD prediction is compared, and examples of their use in apps for mobile-based health professionals are also looked at [27].

This work presents an innovative method for using two different datasets from UCI and Kaggle libraries to predict CKD. In order to resolve the imbalance in classes, the dataset is undersampled and trained using a feedforward neural network (FNN) model. The model has a low loss of 0.0259. It obtains an accuracy of 98.88%. In order to show how the model is used in practice and highlight the value of mixed datasets and DL methods for CKD prediction, the paper also includes a case study [28].

The condition known as CKD affects millions of individuals worldwide [29]. The severity of the illness may be lessened by using ML techniques for early identification. Based on accuracy rate, this research seeks to identify the optimal ML technique for CKD analysis. Experiments using widely used classifiers showed that the accuracy rate of the J48 algorithm is greater than that of other traditional methods. Reducing the severity of CKD and predicting it early are the objectives of ML. With the greatest rates of morbidity and death, kidney disease ranks as the third most important cause worldwide [30]. It is difficult to diagnose and requires intensive medical care, such as hemodialysis. It is infectious and affects around 10% of the population. Identification of CKD may be attributed to hypertension, heart disease, diabetes, potassium, and salt levels. Severe medical treatments such as hemodialysis are necessary for treatment.

The literature review demonstrates a trend of increasing use of ML techniques in both the prediction and diagnosis of CKD due to the highly complicated nature of the disease and the severe consequences it entails. Such studies show that ML algorithms can be effectively used in the identification of individuals at risk of developing

CKD and to aid in the early intervention to prevent further complications and maximise patient outcomes by leveraging clinical data. Although there are encouraging results in the studies concerning the implementation of the ML models, the CKD analysis implementation still encounters challenges and limitations. Model interpretability issues, the class imbalance problem, overfitting, and data quality are among such problems. Additionally, it is crucial to evaluate the performance of ML algorithms across various healthcare settings and patient demographics to ensure their effectiveness and adaptability.

Overall, the literature review focuses on the necessity of continuous researches that are based on ML techniques for the prediction and diagnosis of CKD, the purpose being the improvement of healthcare outcomes and, consequently, a relief from the burden of this disease on both individuals and healthcare systems around the world.

3.Methodology

The present study concerns the methodology employed to investigate CKD prediction. Worldwide, this disorder affects millions of individuals and, if untreated, may have catastrophic side effects. The kidneys are essential organs that produce urine by filtering waste and extra fluid from the circulation. When the kidneys are ill or injured, they are unable to carry out these tasks properly [31]. As a consequence, the body accumulates waste materials and extra fluids, which may cause a variety of symptoms and difficulties. The likelihood of having CKD may be increased by a number of risk factors [32]. The underlying cause of CKD may not always be identified. The necessity for analysis is thought to be crucial for understanding CKD at an early stage. On a dataset that contains patient clinical and laboratory data, ML algorithms may predict CKD. ML is used to analyse the existence of CKD more effectively. In this study, many ML methods are taken into consideration to conduct CKD prediction. The overall methodology is given in *Figure 1*. Details description about each step are given below.

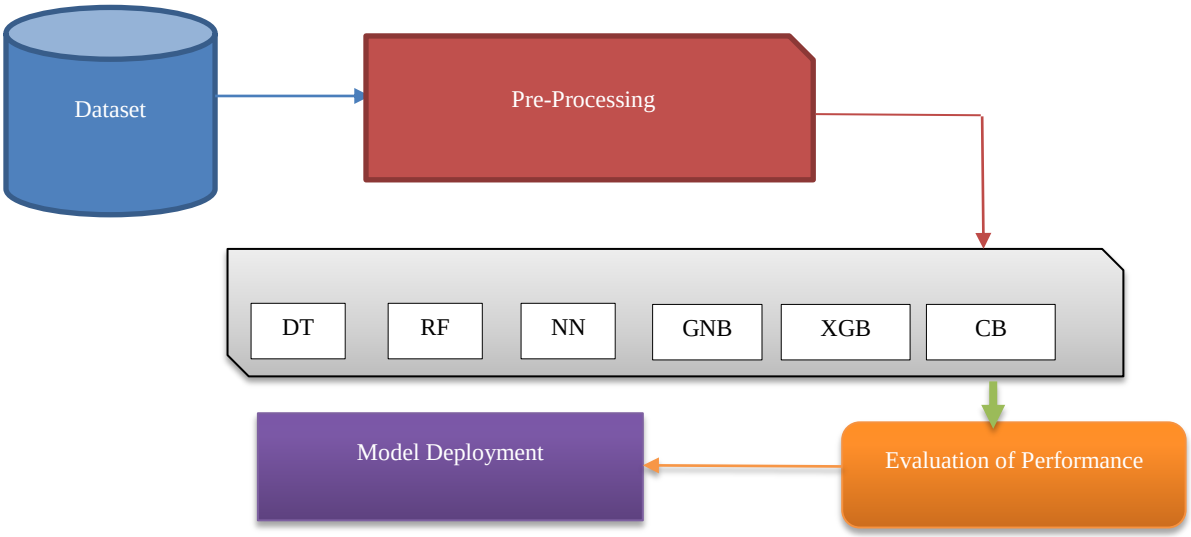

Figure 1 Overview of methodology

3.1Data collection

To create precise and efficient ML models for CKD prediction, high-quality data must be gathered from a range of sources. It's very important that data is collected in a responsible way, with the right level of informed permission, and that patient privacy is always maintained. Public CKD datasets [33] from UCI are taken into account in this study. This dataset consists of 400 instances with 25 features. Out of 25 features one is class label. There 11 numeric and 14 nominal features in the dataset. The purpose of the dataset is to predict whether the person has presence of CKD disease or not. The attributes in the dataset are: BP - blood pressure SG - specific gravity AL - albumin SU - sugar RBC - red blood cells PC - pus cell PCC - pus cell clumps BA - bacteria BGR - blood glucose random BU - blood urea SC - serum creatinine SOD - sodium pot - potassium hemo - hemoglobin PCV - packed cell volume WC - white blood cell count RC - red blood cell count HTN - hypertension DM - diabetes mellitus CAD - coronary artery disease APPET - appetite pe - pedal edema ANE - anemia class – class.

3.2Data pre-processing

The creation of precise and efficient ML models for CKD prediction requires the cleaning and preprocessing of data. They make sure the data is accurate, relevant, and in an analytically sound manner. This enhances the models' generalizability and accuracy, which are essential for accurate CKD prediction. Data transformation, splitting, cleansing, and imbalanced data operations are carried out. Here mean & median are calculated for the non-missing values and used in computing of the missing values.

3.3ML models:

ML, a subfield of AI, uses statistical models and techniques to enable computers to learn from data and make predictions or judgments without explicit programming. The basic idea behind ML is to give a computing system instruction to find patterns in datasets and use those patterns to make predictions or take actions. The procedure entails feeding the computer a tremendous amount of data and allowing it to learn from that data. The computer then makes predictions about fresh data that it has never seen before using the patterns it has learned. The right ML algorithm must be used for the prediction of CKD so that it can effectively analyse the input data and provide precise predictions [34]. The choice of a model is essential for getting precise and trustworthy results. For categorization problems, some well-liked ML algorithms include DT, RF, SVM, NB, and NN. The model used for the prediction of CKD relies on the unique features of the dataset and the required performance criteria. The DT, NN, RF, XGB, GNB, and CB models are taken into consideration in this study.

The chosen models include a variety of approaches, including ensemble techniques, NN, and DT. Due to its variety, the dataset and its underlying patterns may be thoroughly explored. By using the combined knowledge of many models, ensemble techniques such as RF and XGB improve prediction accuracy and resilience. GNB is included to provide a probabilistic perspective, and its simplicity balances the complexity of NN and ensemble approaches. In datasets with a combination of numerical and categorical factors, CB specialized handling of categorical features is essential to assuring a more accurate portrayal of the CKD prediction issue. In summary, taking into account the features of the dataset and the intricacy of the underlying interactions in the setting of CKD, the selected ML models provide a comprehensive approach to CKD prediction.

a) DT: DT is a form of ML technique that is used for regression analysis & classification. It is a visual depiction of every decision-making option depending on certain parameters. The tree is made up of nodes, which stand in for critical decisions, and edges, which represent the results or potential directions. Each node in a DT represents a characteristic or attribute of the data being examined, and the branches extending from the node indicate potential values for that characteristic [35]. Up until a predetermined halting requirement is satisfied, the data is divided recursively depending on the values of the features to create the tree. The ultimate choice or result is represented by the tree's leaves. When attempting to predict a categorical output feature based on a collection of input characteristics, DT are often utilized. A DT for instance, may be used to forecast whether a buyer would buy a product based on their demographic data and previous buying behavior. Because they are simple to understand and can handle both category & numerical data, DT are widely used. They are also effective in terms of computing and work well with big datasets. Over fitting, which happens when a DT is too complicated and matches the training data too closely, might harm a DT ability to generalize to new data. Overfitting may be avoided by methods like pruning, which entails cutting limbs from the tree [36]. Hyperparameter applied are: Criterion: entropy(Entropy is a measure of impurity in a set of examples), max_depth: 5(Limiting the maximum depth of the tree to 5 helps control the complexity of the model), min_samples_split: 3(This parameter sets the minimum number of samples required to split an internal node) , min_samples_leaf: 2(This parameter determines the minimum number of samples required to be at a leaf node), max_features: sqrt(This parameter controls the number of features considered when looking for the best split at each node), class_weight: balanced(Class_weight adjusts the weights of classes in the input data, particularly useful for imbalanced datasets).

b) NN: Inspired by the structure and operation of the human brain, NN are a form of ML algorithm. It is made up of a collection of linked nodes, or neurons, which cooperate to learn from and anticipate data input. The neuron, which takes in inputs, processes them, and then generates an output, is the fundamental unit of a NN [37]. A NN generally consists of many linked layers of neurons, with each layer processing information from the one before. The input layer is the top layer of the network, while the output layer is the bottom layer. Hidden layers are

the layers that lie between. Back propagation, a technique used to train NN, involves modifying the weights of the connections between neurons in order to reduce the error between the expected output and the actual output. The network is exposed to a collection of labelled samples during training, and by changing its weights, it learns to spot patterns in the data. After being trained, the network may be utilized to generate predictions based on fresh, unexplored data. They are able to work with both numerical and categorical data and can learn intricate non-linear connections between input and output data. The training of NN, however, may be time-consuming and costly computationally. They might also be challenging to analyse and comprehend the methodology behind [38]. Hyperparameter applied are: Number of layers: 3(This parameter defines the depth of the neural network), number of neurons per layer: (128, 64, 32) (These parameters determine the width of each layer in the neural network), activation function: Rectified Linear Unit (ReLU) for hidden layers, sigmoid for the output layer, learning rate: 0.001(The learning rate controls the step size of weight updates during training), batch size: 32(A batch size of 32 balances computational efficiency and model stability during training), optimizer: Adam(Adam is an adaptive optimization algorithm commonly used for training NN), regularization: dropout (rate = 0.5)(A dropout rate of 0.5 implies that each neuron has a 50% chance of being dropped out during each training iteration, encouraging robustness and generalization).

c) RF: A particular kind of ML technique called RF is used in both classification & regression analysis. It is an ensemble approach that aggregates the predictions from many DT and combines them to get a conclusion [39]. A RF is made up of many DT, each of which has been trained using a random subset of the training data and features. By doing so, over fitting is decreased and prediction accuracy is raised. The final conclusion is based on the majority vote of the forecasts from all the DT, which are created by the RF algorithm during training using various subsets of the data. Because of its high accuracy, ability to handle both category and numerical data, and resistance to over fitting, RF is a well-liked method. It can also handle big datasets and is computationally efficient. Applications for RF include forecasting consumer behavior, identifying fraud, and diagnosing illnesses. RF can tell you how important each feature in the dataset is, which one of its key benefits. The dimensionality of the data may be decreased and feature selection can be done using this information. By detecting data points that are outliers in relation to the DT, RF may also be utilized for unsupervised learning tasks like anomaly identification. All things considered, RF is a strong and adaptable algorithm that may be used to a variety of ML problems. It is especially helpful for datasets with lots of characteristics and erratic data [40]. Hyperparameter applied are: n_estimators=100(This parameter determines the number of DTs to be included in the RF), criterion=gini(Gini impurity is a metric commonly used in DTs and RFs), max_depth=None(Setting max_depth to None allows the DTs in the RF to grow without any restrictions on depth), min_samples_split=2(This parameter determines the minimum number of samples required to split an internal node in each DTs), and min_samples_leaf=1(min_samples_leaf specifies the minimum number of samples required to be at a leaf node).

d) XGB: Based on the GB framework, XBG is a ML technique. It is a strong and well-liked method that is often used to regression, classification, and ranking issues. In order to minimize a certain objective function, XGB builds a group of DT. The method creates DT repeatedly during training, trying to fix the mistakes of the previous tree with each new tree [41]. DT are given regularization using XGB, which helps to minimize over fitting and boost model accuracy. The capacity of XGB to deal with missing data is one of its important characteristics. It is also capable of learning intricate non-linear correlations between the input and output variables and can handle both numerical & categorical data. XGB is capable of processing big datasets and is 43 computationally efficient. XGB is superior to boosting algorithms like AdaBoost and GB in a number of ways. It offers several hyper parameters that may be tweaked to enhance the efficiency of the model and employs a distributed computing technique to scale to extremely big datasets. The relevance of each feature in the dataset may also be determined via XGB, which is useful for feature selection and dimensionality reduction. Applications for XGB include forecasting customer turnover, detecting illnesses, and spotting fraudulent transactions. In conclusion, XGB is a strong & adaptable algorithm that can be used to a variety of ML problems. It is a desirable option for many applications because to its handling of missing data, regularization approaches, and scalability [42]. Hyperparameter applied are: learning_rate=0.1(The learning rate controls the step size of updates during the boosting process), n_estimators=100(This

parameter determines the number of boosting rounds), subsample=1.0(Subsample specifies the fraction of samples to be used for training each tree), Scolsample_bytree=1.0(Colsample_bytree specifies the fraction of features to be randomly sampled for training each tree)
e) GNB: The GNB. ML method is based on the Bayes theorem and the assumption of feature independence. Data are categorized into many groups using this probabilistic classification process. The data is normally distributed since GNB requires that each feature's probability distribution is Gaussian. The procedure determines the likelihood of each feature given the class before determining the probability of each class given the characteristics using Bayes' theorem. The data point is then given the class with the greatest probability. The ease of use and effectiveness of GNB are two of its key benefits. It can handle both numerical and categorical data and can be trained fast on huge datasets. Additionally resilient to noise and irrelevant characteristics, GNB is effective with tiny datasets. Numerous applications, including as text categorization, spam filtering, and medical diagnostics, employ GNB. It is very helpful for cases when there are more characteristics compared to training examples. The assumption of feature independence made by GNB, meanwhile, is one of its primary flaws and may not hold true in many real-world issues. On certain datasets, GNB may not perform as well as other, trickier algorithms like RF or XGB. Overall, GNB is a straightforward and effective approach that may be used for classification problems, especially when there are many more features than training samples. Before using it to address a particular issue, however, one should carefully analyse its presumptions and constraints [43]. GNB does not have hyperparameters to tune like other ML algorithms,

f) CB: A technique for ML called CB is especially good at handling classification and regression issues. It is an acronym for Categorical Boosting and was created by the Russian search engine business Yandex. CB is well-suited for datasets that include a mix of categorical and numerical variables because of its reputation for handling categorical characteristics in a stable way [44].

These are some of the main traits and qualities of the CB algorithm:
GB architecture: CB is a member of the family of GB algorithms, which combines the predictions of a group of weak prediction models typically DT built repeatedly. It repeatedly adds new models that fix the mistakes caused by the prior models in order to minimise a differentiable loss function.

Dealing with categorical characteristics: CB uses an approach to handle categorical characteristics known as ordered boosting. Categorical variables are transformed into numerical values using an algorithmic approach, which enhances the accuracy of the model's forecasts. When working with datasets that have a lot of categorical attributes, this is quite helpful.

Missing value support built-in: CB includes functionality for managing missing values in the data. Without any user-initiated pre-processing, it can manage missing values in both category and numerical characteristics.

CB is designed to take use of the processing capabilities of current graphics processing units (GPU) in an effective manner. Faster training and prediction times are made possible by this, particularly when working with huge datasets.

Regularisation and the avoidance of over fitting: Several methods are provided by CB to avoid over fitting, including L2 regularisation, bagging, and learning rate decay. These methods assist the model become more generic and prevent it from being too dependent on the training set of data.

Simple to use application programming interface (API): A user-friendly API offered by CB makes it easier to train and assess models. It provides interoperability with other ML frameworks like scikit-learn and interacts nicely with well-known programming languages like Python and R.

Due to its efficient handling of categorical characteristics and competitive performance on a variety of ML problems, CB has grown in prominence in both academic and industry contexts. It has been used to a variety of tasks, such as fraud detection, recommendation engines, picture categorization, and other things. Hyperparameter applied are: random_seed=42(Setting a random seed ensures reproducibility of results across different runs of the algorithm), Depth of the trees(Controlling the depth of the DTs helps prevent overfitting), Number of boosting rounds(iterations=100)(This parameter determines the number of iterations or rounds of boosting), l2_leaf_reg=3(A regularization parameter of 3 indicates the strength of regularization applied),

and learning_rate=0.1(The learning rate controls the step size of updates during training).

3.4Model training and performance:

Using measures like F1-Score , recall, precision, and accuracy, train the model on the training set and assess its performance on the testing set and the Equation 1 to 4 given below [45, 46].

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

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

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{F1} - \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

3.5Model deployment:

Use the trained model to forecast the likelihood that new patients will develop CKD in a clinical environment. It is essential to remember that the amount and quality of patient data used for learning and testing the prediction model affect its performance. Therefore, for precise and trustworthy CKD prediction, thorough data collection, cleaning, and preprocessing are crucial.

4.Results

This section describes in detail about the experimental results carried out in the methodology. The dataset is split into 70% for training and 30% for testing. The experimental procedure is carried out using Python programming language together with data science libraries like TensorFlow, Scikit-Learn, Pandas, and NumPy to build data processing and run our algorithms.

4.1Experimental results of CKD prediction using various ML models:

Figure 2 displays the experimental findings from the prediction of CKD using several ML models. The effectiveness of any model is assessed using the accuracy measure. The findings are explained as follows:

CB model predicted CKD with an accuracy of 0.95. The DT model's accuracy was 0.8167. In terms of accuracy, this model did rather worse than others. The NN model's accuracy was 0.8667, which means that it accurately diagnosed CKD 0.8667 of the time in the studies. Compared to the DT model, this model performed better but not as well as some other models.

The RF model successfully identified CKD with a 0.95 accuracy score. This model performed similarly to the XGB and CB Classifier models. The XGB model also attained an accuracy of 0.95, which indicates that it accurately identified CKD in the tests with 0.95 accuracy. This model fared as well as the RF and CB Classifier models. The GNB model successfully identified CKD 0.9333 of the time, with an accuracy of 0.9333. In comparison to the CB classifier, RF, and XGB models, this model performed well but somewhat worse.

Overall, the CB, RF, and XGB models outperformed the competition, predicting CKD with the accuracy 0.9500. The accuracy of the NN and DT models was lower, at 0.8667 and 0.8167, respectively. With an accuracy of 0.9333, the GNB model likewise did well.

Figure 3 illustrates the experimental results from many ML models that were used to the recall measure prediction of CKD. Recall is the proportion of actual positive cases that the model correctly identified; it is also known as sensitivity or true positive rate.

With a recall of 1.00, the CB accurately identified each case of CKD in the experiments. This demonstrates that the model had no false negatives and correctly detected every instance of the sickness. With a recall of 0.9828, the DT model correctly identified 98.28% of the positive cases of chronic renal illness in the experiments. This model performed well at identifying the great majority of positive scenarios, even in the face of a few false negatives. With a recall of 0.9429, the NN model identified 0.9429 of the experimentally verified positive cases of CKD. This model had more false negatives than the previous two. With a recall score of 0.9737, the RF model successfully identified 0.9737 of the positive cases of CKD in the investigations. This model exhibited fewer false negatives than the NN model, although performing somewhat worse than the DT model. With a recall of 1.00, the XGB model was able to accurately identify each case of CKD that was positive in the research. Similar to the CB Classifier, this model accurately detected every instance of the disease with no false negatives. GNB model had a recall of 0.9861 and correctly identified 0.9861 of the positive cases of CKD. This model does not have a lot of false negatives.

Overall, the CB Classifier and XGB models, both of which had 1.00 recall, were able to identify all positive cases of CKD. Similarly, the GNB model performed well, with a recall of 0.9861. Of the models in the table, the NN model performed the worst in terms of recall, while the DT and RF models performed rather well despite having considerably lower recall rates.
Figure 4 displays the results of many ML models' experiments that were used to predict CKD using the accuracy measure. The proportion of cases that were really positive among those that were predicted to be positive is known as precision. The following is an explanation of the results:

The CB Classifier accurately predicted 0.9231 of positive outcomes with a precision of 0.9231. This implies that the model did not provide a large number of false positive predictions. With an accuracy of 0.7308, the DT model was able to accurately predict 0.7308 of the favorable outcomes. This model produced a higher number of false positive predictions than the CB Classifier. With a precision of 0.8462, the NN model produced 0.8462 of the positive predictions that were really accurate. While this model performed better than the DT model, it had a slightly higher false positive prediction rate than the CB Classifier. With a precision of 0.9487, the RF model accurately predicted 94.87% of the positive observations it produced.

This model showed a reduced rate of false positive predictions and performed better than the NN model. Predicting 0.9231% of the favourable events accurately, the XGB model also achieved an accuracy of 0.9231. As expected from the CB Classifier, this model produced very few incorrectly positive predictions. With a precision of 0.9103, the GNB model accurately predicted 0.9103 of its positive observations. This model performed somewhat worse than the XGB and CB Classifier models in terms of false positive predictions.
All things considered, the CB Classifier, XGB, and RF models demonstrated excellent accuracy performance and high true positive prediction rates. The DT and NN models produced more false positive predictions because of their lower precision. The GNB model performed well even if it wasn't as good as the best models.

Figure 5 displays the experimental results from many ML models that were used to predict CKD using the F1-Score metric. CB Classifier: With an F1-Score of 0.9600, the CB Classifier achieved an excellent overall performance grade. This model balanced memory and accuracy to give accurate forecasts of CKD. The DT model fared somewhat worse overall than other models, as shown by its F1-Score of 0.8382. This model has a lower F1-Score due to its weaker accuracy and recall. A F1-Score of 0.8919 for the NN model was achieved, indicating a moderate overall performance. It was still not as good as the top models, although performing better than the DT model.

The RF model's F1-Score of 0.9610 indicates that it performs well overall. Like the XGB and CB Classifier models, this model successfully struck a compromise between recall and accuracy. XGB: With an F1-Score of 0.9600, the XGB model's performance was on par with the RF and CB Classifier models. This system, which carefully balanced memory and accuracy, generated accurate forecasts of chronic renal disease.

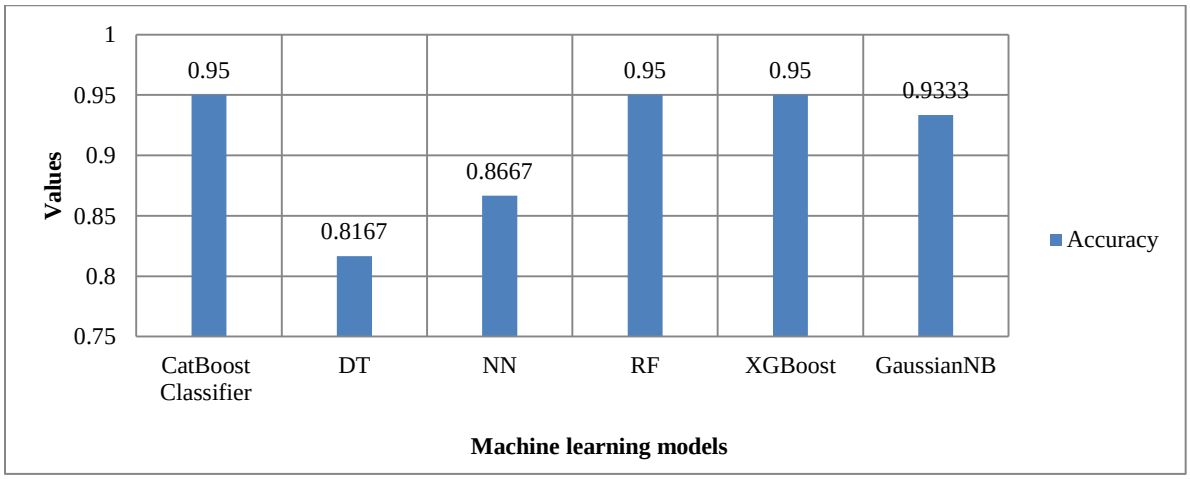

Figure 2 Experimental results of CKD Prediction using ML using accuracy metrics

The F1-Score of the GNB model is 0.9467, indicating a generally good performance. This model was not as excellent as the top models, despite achieving a reasonable balance between recall and accuracy. The RF, XGB, and CB Classifier models performed the best overall, with high F1-Score s of 0.9600. Through careful consideration of both recall and accuracy, these models generated accurate predictions. The GNB model also did well, with an F1-Score of 0.9467. As shown by their lower F1-Score, the DT and NN models underperformed relative to the best models overall.

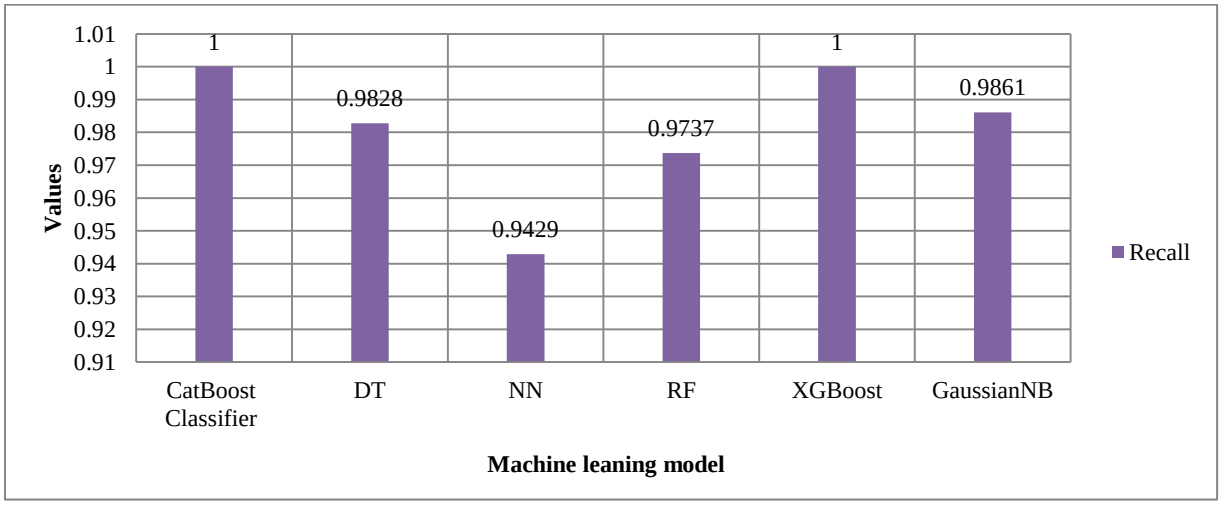

Figure 3 Experimental results of CKD Prediction using ML using recall metrics

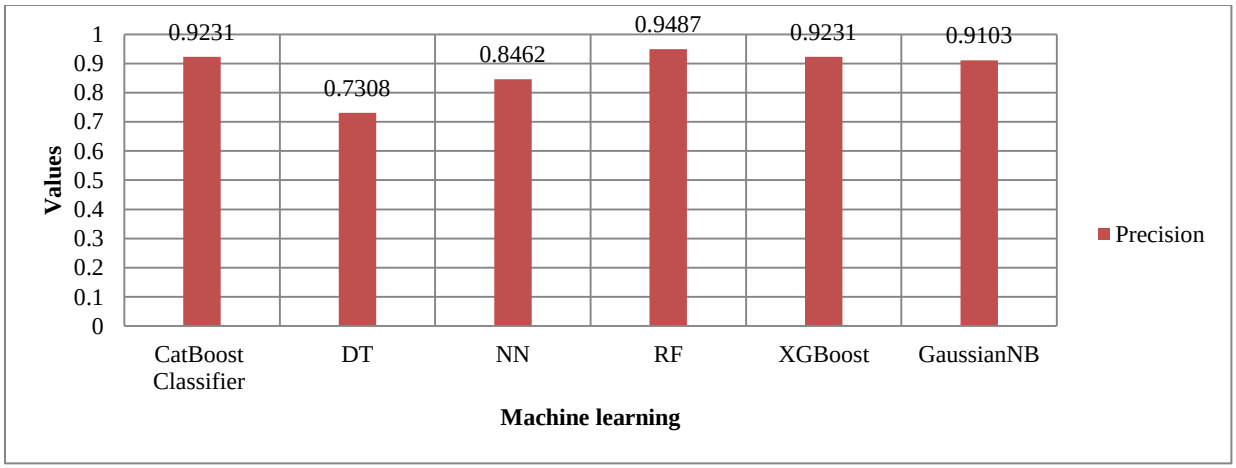

Figure 4 Experimental results of CKD Prediction using ML using precision metrics

5.Discussion

Figure 6 displays experimental findings for predicting CKD using several ML models, with assessment metrics for each model including accuracy, recall, precision, and F1-Score.

Across all criteria, the CB Classifier consistently outperformed the competition. With a high accuracy score of 0.9500, it successfully predicted CKD 0.95 of the time. Additionally, it had a recall score of 1.00, accurately identifying every instance of the condition. The accuracy was 0.9231, meaning that 0.9231 of the model's positive predictions were correct. The F1-Score of 0.9600 indicates a performance that strikes a balance between recall and accuracy. When compared to the CB Classifier, the DT model had a lower accuracy of 0.8167, suggesting a comparatively greater number of wrong predictions. However, it managed to properly identify 0.9828 of positive instances with a high recall of 0.9828. Since the accuracy was 0.7308 of the optimistic predictions really came true. The F1-Score of 0.8382 indicates a modest overall performance, with a trade-off between recall and accuracy.

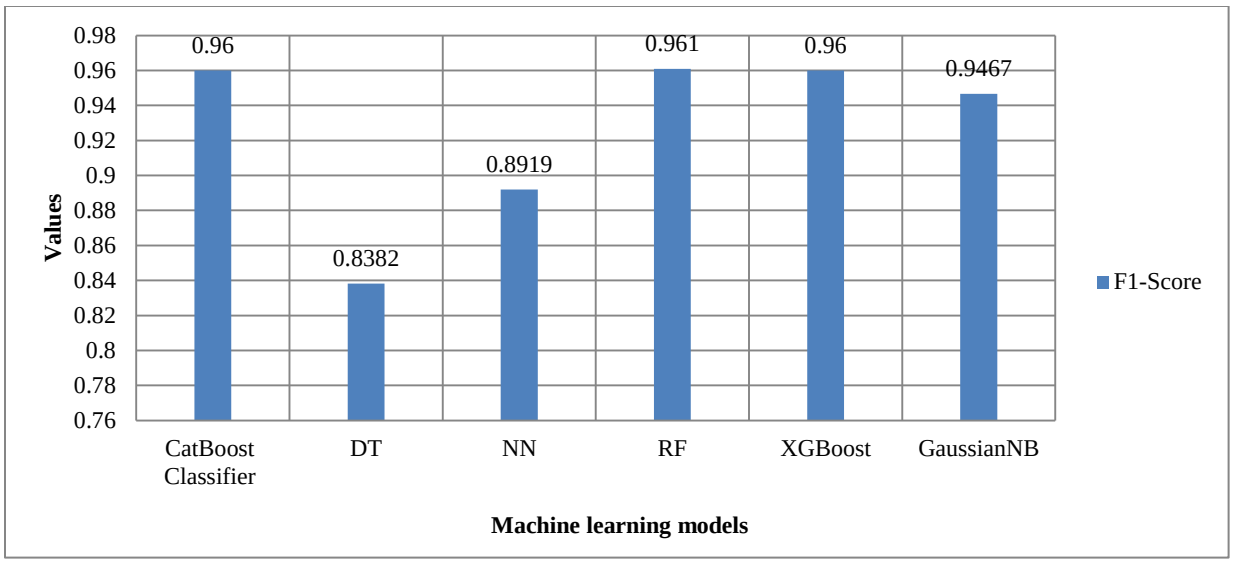

Figure 5 Experimental results of CKD prediction using ML using F1-Score metrics

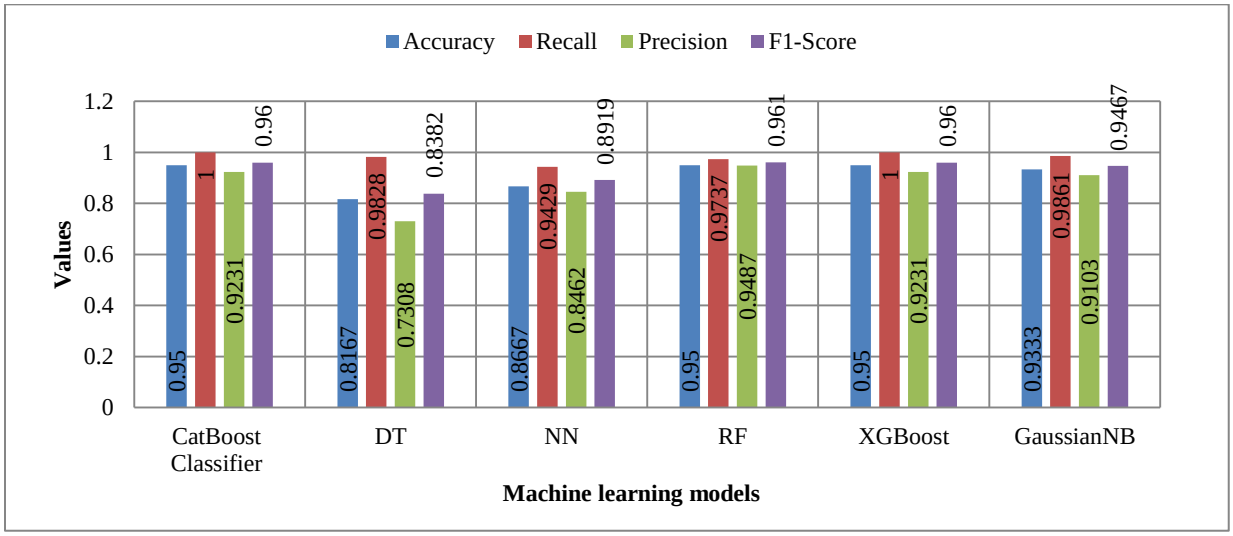

Figure 6 Experimental results of CKD Prediction using ML using different metrics

The accuracy of the NN model, 0.8667, was higher than that of the DT model but lower than that of the CB classifier. Recall of 0.9429 implies that of positive instances were accurately recognized. The F1-Score of 0.8919 suggests a performance that is well balanced in terms of recall and accuracy. The RF model's accuracy score of 0.9500, which is comparable to that of the CB Classifier and XGB, shows that it is efficient in predicting CKD. It properly identified 0.9737of positive cases. Since the accuracy was 0.9487 of the positive predictions really materialized as positives. The CB Classifier-like performance is shown by the F1-Score of 0.9610. The XGB model achieved an accuracy of 0.9500. Additionally, it exhibited 100% recall, properly recognizing every affirmative instance. Since the accuracy was 0.9231 of the positive predictions really materialized as positives. The F1-Score achieved by CB Classifier is 0.9600. The GNB model's accuracy, was 0.9333, indicated prominent in predicting CKD with a recall of 0.9861. Overall, the CB Classifier and XGB models exhibited strong performance across various metrics, including accuracy, recall, precision, and F1-Score, demonstrating their effectiveness in predictive analysis. The accuracy and precision of the DT and NN models were significantly lower, while recall was excellent. While the GNB model didn't quite match the top models in terms of accuracy, it still did well across the criteria.

With respect to an accuracy measure, recall, precision, and F1-Score, CB Classifier, RF, and XGB outperformed the other models consistently. Efficiently, both approaches identified CKD patients with the CB Classifier model coming out the best and featuring an outstanding combination of accuracy and recall. While both tree and neuron model performances were worsened by lower accuracy, both of them still showed satisfactory recall so that they could be used where recording all positive cases was crucial. In a general sense, GNB performed fine, offering a performance that did a trade-off between recall and accuracy.

Overall, the results point to the potential use of ML models, particularly CB classifier, RF, XGB, & GNB, for the diagnosis and prognosis of CKD. For flexibility in various application settings, these models provide adjustable trade-offs between accuracy, recall, precision, and F1-Score.

5.1Study limitations

Representativeness and the quality of the data utilized have a great effect on the output the models achieve. The use of quality data is highly encouraged.

The study outcomes may be of general application to all or some patient groups but not, for example, to different healthcare settings. How models perform in practice maybe be affected by diversities in the population in a context of patients' demographics and disease qualities, and even variations in health practices.

While multiple models have great predictive capability, the inherent complexity of these models may limit their readability, thus preventing medical professionals from identifying the root causes of predictions. As the study's robustness and generalizability get assessed, external validation could be done by making use of other independent datasets.

A complete list of abbreviations is summarised in *Appendix I.*

6.Conclusion and future work

Different ML algorithms were evaluated considering the CKD dataset. CB, RF, and XGB models exhibit the most potential for accurately predicting CKD. The models demonstrated a notable level of accuracy, recall, precision, and F1-Score, signifying their ability to accurately detect positive instances of CKD while maintaining a low rate of false positives. The significance of predicting kidney disease arises from the fact that CKD is an escalating public health issue that has a widespread impact on millions of individuals globally. The precise anticipation of CKD enables healthcare professionals to devise individualized therapeutic strategies that address the particular root causes of the ailment, diminish the likelihood of adverse outcomes, and enhance patient results. In addition, timely prediction has the potential to facilitate prompt referral of patients to specialists and, in certain instances, obviate the necessity for dialysis or kidney transplantation. The result indicate that ML-oriented methodologies possess the capability to precisely anticipate CKD. This can lead to a noteworthy enhancement in patient results and a reduction in healthcare expenditures. Nevertheless, additional investigation and verification are imperative to ascertain the dependability and applicability of these models' outcomes prior to their implementation in clinical settings. Future work can be extended by using different ML, DL and hybrid models and applying feature engineering to choose relevant variable subset.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability

The dataset utilized in this paper is publicly available at https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease.

Author's contribution statement

B. Yamini, T. Saraswathi: Writing, software setup and analysis, **P. Radhakrishnan, M. Nalini:** Writing, revisison and Interpretations of the results. **M. Shanmuganathan, Siva Subramanian.R**: Writing, reviewing and analysis of experiment.

References

[1] Wang W, Chakraborty G, Chakraborty B. Predicting the risk of chronic kidney disease (CKD) using machine learning algorithm. Applied Sciences. 2020; 11(1):1-17.

[2] Wickramasinghe MP, Perera DM, Kahandawaarachchi KA. Dietary prediction for patients with chronic kidney disease (CKD) by considering blood potassium level using machine learning algorithms. In life sciences conference 2017 (pp. 300-3). IEEE.

[3] Gunarathne WH, Perera KD, Kahandawaarachchi KA. Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney

disease (CKD). In 17th international conference on bioinformatics and bioengineering 2017 (pp. 291-6). IEEE.
[4] Elhoseny M, Shankar K, Uthayakumar J. Intelligent diagnostic prediction and classification system for chronic kidney disease. Scientific reports. 2019; 9(1):1-14.
[5] Rajendran T, Rajathi SA, Balakrishnan C, Aswini J, Prakash RB, Subramanian RS. Risk prediction modeling for breast cancer using supervised machine learning approaches. In 2nd international conference on automation, computing and renewable systems 2023 (pp. 702-8). IEEE.
[6] Raju NG, Lakshmi KP, Praharshitha KG, Likhitha C. Prediction of chronic kidney disease (CKD) using data science. In international conference on intelligent computing and control systems 2019 (pp. 642-7). IEEE.
[7] Sinha P, Sinha P. Comparative study of chronic kidney disease prediction using KNN and SVM. International Journal of Engineering Research and Technology. 2015; 4(12):608-12.
[8] Almansour NA, Syed HF, Khayat NR, Altheeb RK, Juri RE, Alhiyafi J, et al. Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study. Computers in biology and medicine. 2019; 109:101-11.
[9] Pasadana IA, Hartama D, Zarlis M, Sianipar AS, Munandar A, Baeha S, et al. Chronic kidney disease prediction by using different decision tree techniques. In journal of physics: conference series 2019 (pp. 1-6). IOP Publishing.
[10] Saha A, Saha A, Mittra T. Performance measurements of machine learning approaches for prediction and diagnosis of chronic kidney disease (CKD). In proceedings of the 7th international conference on computer and communications management 2019 (pp. 200-4). ACM.
[11] Baidya D, Umaima U, Islam MN, Shamrat FJ, Pramanik A, Rahman MS. A deep prediction of chronic kidney disease by employing machine learning method. In 6th international conference on trends in electronics and informatics 2022 (pp. 1305-10). IEEE.
[12] Srivastava S, Yadav RK, Narayan V, Mall PK. An ensemble learning approach for chronic kidney disease classification. Journal of Pharmaceutical Negative Results. 2022:2401-9.
[13] Dritsas E, Trigka M. Machine learning techniques for chronic kidney disease risk prediction. Big Data and Cognitive Computing. 2022; 6(3):1-15.
[14] Singh V, Asari VK, Rajasekaran R. A deep neural network for early detection and prediction of chronic kidney disease. Diagnostics. 2022; 12(1):1-22.
[15] Abdel-fattah MA, Othman NA, Goher N. Predicting chronic kidney disease using hybrid machine learning based on apache spark. Computational Intelligence and Neuroscience. 2022; 2022:1-12.
[16] Kashyap CP, Reddy GS, Balamurugan M. Prediction of chronic disease in kidneys using machine learning classifiers. In 1st international conference on computational science and technology 2022 (pp. 562-67). IEEE.
[17] Yogish HK. Prediction of chronic kidney disease using machine learning technique. In fourth international conference on cognitive computing and information processing 2022 (pp. 1-6). IEEE.
[18] Bhowmick R, VM AX. Machine learning models for analysis and prediction of chronic kidney disease. In 9th international conference on advanced computing and communication systems 2023 (pp. 937-42). IEEE.
[19] Hassan R, Sharan B, Kumari N, Rafiq T, Thakur G, Bhargav R. Prediction of chronic diseases using machine learning classifiers. In 10th international conference on computing for sustainable global development 2023 (pp. 885-9). IEEE.
[20] Swain D, Patel H, Patel K, Sakariya V, Chaudhari N. An intelligent clinical support system for the early diagnosis of the chronic kidney disease. In 2nd international symposium on sustainable energy, signal processing and cyber security 2022 (pp. 1-5). IEEE.
[21] Anil D, Naimudden S, Reddy AS, Lavanya A. Prediction of chronic kidney disease using various machine learning algorithms. In international conference on innovative data communication technologies and application 2023 (pp. 156-61). IEEE.
[22] Vimala C, Subramani C, Bhagat V, Sravani MV. Analysis of different algorithms to predict chronic kidney disease. In international interdisciplinary humanitarian conference for sustainability 2022 (pp. 1051-4). IEEE.
[23] Navyasree V, Surarchitha Y, Reddy AM, Sree BD, Anuhya A, Jabeen H. Predicting the risk factor of kidney disease using meta classifiers. In 2nd Mysore sub section international conference 2022 (pp. 1-6). IEEE.
[24] Deepa R, Priscilla R, Pandi A, Renukadevi B. An early prediction model for chronic kidney disease using machine learning. In international conference on networking and communications 2023 (pp. 1-10). IEEE.
[25] Pilli VS, Pamidi K, Poovammal E. Chronic kidney disease prediction. In international conference for advancement in technology 2023 (pp. 1-4). IEEE.
[26] Raj RS, Kusuma M. A comprehensive analysis of chronic health diseases using big data. In international conference on evolutionary algorithms and soft computing techniques 2023 (pp. 1-5). IEEE.
[27] Saumya L, Manimozhi I. A survey on machine learning algorithms for the detection of chronic kidney disease. In 2nd international conference on automation, computing and renewable systems 2023 (pp. 1832-7). IEEE.
[28] NJ S, Venkatesh K. CKD prediction using mixed dataset concept and deep learning: a combined approach for accurate diagnosis. In international conference on research methodologies in knowledge management, artificial intelligence and telecommunication engineering 2023 (pp. 1-6). IEEE.

[29] Jaya LA, Pm JP, Lokesh K, Dev AD, Jagathieshwaran J, Sneha M. Predicting the severity of chronic kidney disease with J48. International conference on computing, communication, and intelligent systems 2023 (pp.556-61). IEEE.
[30] Saraswat T, Pathak S, Sachdeva S, Sahu K, Sawhney R. Kidney disease detection and identification using artificial intelligence. In international conference on cloud computing, data science & engineering 2023 (pp. 537-43). IEEE.
[31] Rysz J, Gluba-brzózka A, Franczyk B, Jabłonowski Z, Ciałkowska-rysz A. Novel biomarkers in the diagnosis of chronic kidney disease and the prediction of its outcome. International Journal of Molecular Sciences. 2017; 18(8):1-17.
[32] Poorani K, Karuppasamy M. Comparative analysis of chronic kidney disease prediction using supervised machine learning techniques. In international conference on information and communication technology for intelligent systems 2023 (pp. 87-95). Singapore: Springer Nature Singapore.
[33] Rubini L. Early stage of chronic kidney disease UCI machine learning repository. Chronic Kidney Disease. 2015.
[34] Gracious LA, Jasmine RM, Pooja E, Anish TP, Johncy G, Subramanian RS. Machine learning and deep learning transforming healthcare: an extensive exploration of applications, algorithms, and prospects. In 4th IEEE global conference for advancement in technology 2023 (pp. 1-6). IEEE.
[35] Yamini B, Kaneti VR, Nalini M, Subramanian S. Machine learning-driven PCOS prediction for early detection and tailored interventions. SSRG International Journal of Electrical and Electronics Engineering. 2023; 10:61-75.
[36] Bansal M, Goyal A, Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal. 2022; 3:100071.
[37] Verma VK, Verma S. Machine learning applications in healthcare sector: an overview. Materials Today: Proceedings. 2022; 57:2144-7.
[38] Humayun M, Sujatha R, Almuayqil SN, Jhanjhi NZ. A transfer learning approach with a convolutional neural network for the classification of lung carcinoma. Healthcare 2022; 10:1-15. MDPI.
[39] Nithya T, Kumar VN, Gayathri S, Deepa S, Varun CM, Subramanian RS. A comprehensive survey of machine learning: advancements, applications, and challenges. In second international conference on augmented intelligence and sustainable systems 2023 (pp. 354-61). IEEE.
[40] Kamala SP, Gayathri S, Pillai NM, Gracious LA, Varun CM, Subramanian RS. Predictive analytics for heart disease detection: a machine learning approach. In 4th international conference on electronics and sustainable communication systems 2023 (pp. 1583-9). IEEE.
[41] Yamini B, Sudha K, Nalini M, Kavitha G, Subramanian RS, Sugumar R. Predictive modelling for lung cancer detection using machine learning techniques. In 8th international conference on communication and electronics systems 2023 (pp. 1220-6). IEEE.
[42] Subramanian RS, Prabha D. Ensemble variable selection for naive bayes to improve customer behaviour analysis. Computer Systems Science & Engineering. 2022; 41(1):339-55.
[43] Kamel H, Abdulah D, Al-tuwaijari JM. Cancer classification using gaussian naive Bayes algorithm. In international engineering conference 2019 (pp. 165-70). IEEE.
[44] Luo M, Wang Y, Xie Y, Zhou L, Qiao J, Qiu S, et al. Combination of feature selection and catboost for prediction: the first application to the estimation of aboveground biomass. Forests. 2021; 12(2):1-21.
[45] Subramanian RS, Sudha K, Ramana KV, Sivakumar S, Nithyanandhan R, Nalini M. Hybrid variable selection approach to analyse high dimensional dataset. In 7th international conference on computing methodologies and communication 2023 (pp. 1489-95). IEEE.
[46] Raju SS, Dhandayudam P. Prediction of customer behaviour analysis using classification algorithms. In AIP conference proceedings 2018. AIP Publishing.

Dr. B. Yamini has completed her Bachelor of Engineering in Computer Science and Engineering from Mailam Engineering College, in the year 2003. She pursued her Master of Technology in Information Technology from Sathyabama University, Chennai in the year 2007. She was awarded the Doctor of Philosophy in Computer Science and Engineering from Sathyabama Institute of Science and Technology, Chennai in the year 2020. She has published papers in various International and National Conferences and Journals. She is currently working at SRM Institute of Science and Technology, College of Engineering and Technology, School of Computing as Assistant Professor in the Department of Networking and Communications. Her areas of interests include Network Security, Cyber Forensics, Image Processing, Information Retrieval system, Machine Learning, Deep Learning and Cloud Computing.
Email: yamini.subagani@gmail.com

T. Saraswathi, M.E (Computer Science) is an Assistant Professor in Easwari Engineering College (Autonomous), Chennai, Tamilnadu, India. She has a total experience of 10 years in teaching and industry. She is currently pursuing her Ph.D in Information and Communication Engineering, Anna University. She has successfully completed many industry related workshop and has a rich

knowledge in the real time projects related to python programming. She has presented her research ideas in national and international conferences. She has published patents, research papers during her service.
Email: saand1986@gmail.com

P. Radhakrishnan is a fascinating Researcher in School of Computing, SRM Institute of Science And Technology. He is currently working as an Assistant Professor in School of CS and AI, SR University – Warangal having 11+ years of teaching experience. He has also attended international conferences and published his research work in reputed journals.
Email: rksiva13@gmail.com

Dr. M. Nalini is an Associate Professor in the Department of Computer Science and Engineering, S.A Engineering College, Chennai, India. She received her B.E. in Computer Science and Engineering from Anna University, Chennai in 2010 and M.Tech. in Computer Science and Engineering from B.S.Abdur Rahman Crescent Institute of Science & Technology, Chennai, India in 2012. She was awarded Ph.D. in Computer Science and Engineering from St. Peter's Institute of Higher Education and Research, Chennai, India in 2018. She has 10 years teaching experience. Her research interests include Data Mining, Machine Learning, Big Data Analytics and Networking. She has published many articles in reputed Journals.
Email: nalini.tptwin@gmail.com

Dr. M. Shanmuganathan, is currently Associate Professor, Department of Computer Science and Engineering, Panimalar Engineering College, Chennai - 600123, TamilNadu, India having more than 21 years of teaching experience. His areas of interests include Software Engineering, Artificial Intelligence, Machine Learning, Deep Learning and MCDM.
Email: shanmail2k@gmail.com

Siva Subramanian.R is an Associate Professor in the Department of Computer Science and Engineering, RMK College of Engg and Tech, Chennai, India. He received her B.E. in Computer Science and Engineering from Anna University, Chennai in 2009 and M.Tech. in Computer Science and Engineering from Bharath University, Chennai, India in 2013. He has 10 years teaching experience. His research interests include Data Mining, Machine Learning, Big Data Analytics and Networking. He has published many articles in reputed Journals.
Email: sivasubramanian12@yahoo.com

Appendix I

S. No.	Abbreviation	Description
1	AI	Artificial Intelligence
2	API	Application Programming Interface
3	ANN	Artificial Neural Networks
4	CKD	Chronic Kidney Disease
5	CB	CatBoost
6	CTC	Connectionist Temporal Classification
7	DT	Decision Tree
8	FNN	Feedforward Neural Network
9	GNB	Gaussian Naive Bayes
10	GFR	Glomerular Filtration Rate
11	GB	Gradient Boosting
12	GPU	Graphics Processing Units
13	GBT	Gradient-Boosted Trees
14	K-NN	K-Nearest Neighbors
15	LR	Logistic Regression
16	LMT	Logical Model Tree
17	NN	Neural Networks
18	NB	Naive Bayes
19	ML	Machine Learning
20	MLP	Multilayer Perceptron
21	PCA	Principal Component Analysis
22	RF	Random Forest
23	REP	Reduced Error Pruning
24	RotF	Rotation Forest
25	ReLU	Rectified Linear Unit
26	SVM	Support Vector Machine
27	UCI	University of California Irvine
28	XGB	XGBoost