[1]Priya Karkare

[2]Vaibhav Narawade

[3]Smita Bharne

The Artistry of Stochastic Gradient Boosting and Gradient Boosting Classifiers in Chronic Renal Disease Classification

Abstract: - Chronic kidney failure is a medical disorder that impairs the kidneys' overall capacity to filter dangerous substances from your blood and maintain overall health. Anemia, weakened bones, poor diet, trauma, and decreased blood pressure are some of the factors that contribute to chronic kidney disease. Anemia, weakened bones, poor diet, trauma, and decreased blood pressure are some of the factors that contribute to chronic kidney disease. This study proposes using a several unsupervised algorithms, evaluate their performance, and determine the optimal combinations with higher accuracy. In addition to scaling features and applying the method to balance the data, this list of processes includes correctly imputed missing data points. The present study has employed six supervised algorithms, including DT, RF, SGB, XgBoost, Ada Boost and Gradient Boosting Classifiers has integrated them with techniques for selecting features based on the Pearson Correlation Coefficient. Classifying clinical data of chronic kidney diseases(CKD) and Non-CKD with an overall accuracy of 99.2% has been achieved by integrating feature reduction approaches with Stochastic Gradient Boosting and Gradient Boosting Classifier. These models were put to the test using a collection of data on chronic kidney illness from the University of California at Irvine, this had four hundred entries of data with twenty-six attributes. The outcomes of various models are examined. The model built with Stochastic Gradient Boosting and Gradient Boosting Classifier method performed the best in terms of correctness using 24 attributes for the small dataset, based on the comparison. The final phase of this study will examine how effectively the machine learning system predicts chronic kidney failure with respect to precision, recall, accuracy, & F1-Score**.**

Keywords: Machine Literacy, Random Forests, Chronic Kidney Failure, DT, RF, SGB, XgBoost, AdaBoost Classifier, Gradient Boosting Classifier.

I. INTRODUCTION

The Chronic kidney damage, or CKD, has been recognized as a worldwide public health risk. Between 1990 and 2010, CKD moved up the global list of diseases that cause death from 27th to 18th place, based toward the 2010 Research on Global Burden of Diseases[1], World Health Organization Study. CKD was responsible for over 38 million out of 58 million total deaths observed in 2005.Women are more likely to suffer from CKD than men, and people who are 65 years of age or older also more probably to suffer it than those who are in the 45 to 64 and 18 to 44 age categories. Current medical data show that an unexpected 10% of people worldwide suffer with CKD. Inconsistent growth of cells that spreads to other parts of the body is the cause of kidney disease. A progressive loss of kidney functioning is a sign of long-term kidney disease, this goes by the name of chronic renal failure as well.Extra fluid and waste products from the blood are removed by your kidneys and excreted in urine. A variety of signs or symptoms may arise when chronic health conditions initially manifest. You may not become aware of the issue unless the problem worsens. If one or both kidneys fail to operate properly, patients may experience symptoms like back pain, abdominal pain, a fever, nosebleeds, itching, and vomiting[2]. Two of the leading causes that may damage the kidneys over time are uncontrolled blood pressure and diabetes. Early detection plays a key role in lowering the mortality rate among those with chronic kidney disease. Renal failure is often the result of late detection of this disease, that requires dialysis or kidney transplantation. The rising number of people with chronic kidney disease (CKD) and a lack of educated experts have resulted in high medical care costs. Medical practitioners use two basic techniques to accurately obtain patient data to diagnose kidney disease. First, the patient undergoes blood and urine tests to check for kidney disease (CKD). A sample of blood can evaluate the state of the kidneys, frequently referred to as glomerular filtration rate (GFR). A kidney's GFR of 60 indicates normal function. Values in the range of 15 to 60 indicate issues with kidney function. Finally, renal failure is indicated if the GFR value is equal to or less than 15 . The second method, a urinalysis test, searches for albumin, which may leak into the urine if the kidney system is damaged. The estimated glomerular filtration rate (eGFR), a urine test, a blood pressure reading, and chronic kidney disease examinations are the only tests available for determining the level of CKD.

[1] *Priya Karkare : Department of Computer Engineering, Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Nerul, Navi Mumbai, India, priya9930688736@gmail.com

[2] Vaibhav Narawade Department of Computer Engineering, Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Nerul, Navi Mumbai, India, vaibhav.narawade@rait.ac.in

[3] Smita Bharne Department of Computer Engineering, Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Nerul, Navi Mumbai, India, smita.bharne@rait.ac.in

Five stages were identified by the Public Order Foundation (NKF) in order problems. It helps doctors to provide excellent care, as each stage requires different examinations and treatments. Doctors determine the severity of health issues by using the glomerular filtration rate (GFR), a calculation based on the individual's gender, age, and creatinine levels in the serum position (linked by blood testing). When the GFR is more than 90 mL/min, the patient enters Phase 1. GFR in Stage 2 mild CKD ranges from 60 to 89 mL/min. Moderate CKD is defined as GFR between 45 and 59 mL/min in Phase 3A. GFR in intermediate CKD, Stage 3B is between 30 and 44 mL/min. GFR in Stage 4 severe CKD ranges from between 15 and 29 mL/min. Stage 5: CKD GFR below 15 mL/min is terminated[3]. Chronic kidney disease (CKD) is associated with risk factors such as diabetes, high blood pressure, heart disease, obesity, a family history of the condition, and advanced age. A kidney's GFR of 60 indicates normal function. Values in the range of 15 to 60 indicate issues with kidney function. Finally, renal failure is indicated if the GFR value is equal to or below 15.

II. LITERATURE REVIEW

These days, there is more demand for the development of technologies as well as methods for tracking and predicting chronic kidney sickness. This section will cover recent research on small dataset processing approaches and CKD risk prediction using methods of machine learning.

Salekin, A. et al[15], 24 predictive parameters were employed in the studies. The inaccuracy of the root mean square of the F1-measure, which is 0.1084 creates a detection accuracy of 0.993 for this author. In the paper, Chen et.al [14] which concentrated on the UCI ML data set, three multivariate models were discussed that have overall accuracy rates of over 93% and can distinguish between patients with and without CKD. Compared to SIMCA, KNN and SVM perform better in this study Comparing the SVM model to the other two models is more robust for the composite data set since it can withstand noise disturbances the best. The Support Vector Machine classification technique was employed in the study [13] to identify Chronic KD by author Polat et al. The results of this research indicated, with regard to other chosen techniques, the SVM classifier employing the filtered subset evaluator with the Best First search engine feature selection for the diagnosis of Chronic Kidney Disease approach showed a higher accuracy rate (98.5%). In order to classify gene expression data, the hybrid model that combines RF and SVM was the focus of the Rustam et al[12]. SVM (known as RF-SVM) can efficiently forecast gene expression information using extremely highest dimensions, whereas RF can be more interpretable, more accurate, and more generalizable. Furthermore, it is demonstrated by the simulation outcomes using information from the Gene Expression Omnibus (GEO) repository showing that the recommended RF-SVM algorithm outperforms RFE-SVM on CKD data. Vasquez- Morales et al. [10] provided a neural network model that was 95% accurate. to calculate the rate at which chronic renal failure develops using a 400-case dataset. The objective of their research Ogunleye[9] is to grow a real-time application in order to identify CKD early utilizing machine learning approaches (Naive Bayes and KNN algorithms). A model was created by Deepika et al. in 2020 to evaluate the probability of chronic renal illness using an antiquated CKD dataset [8]. There were twenty-four attributes and one target variable in the dataset. The model was built utilizing the supervised ML technique of Naïve Bayes and KNN, with Naïve Bayes achieving 91% accuracy and KNN achieving an astounding 97%. Furthermore, Senan et al[7] proposed the evaluation of a dataset including 24 features that was gathered from 400 patients was the main focus of this work. In this work, four classification methods were used: support vector machines (SVM), k-nearest neighbors (KNN), random forests, and decision trees. All of the classification algorithms performed admirably. The random forest strategy achieved an accuracy, precision, recall, and F1-score of 92.01% for all criteria, outperforming all other pertinent strategies. A ML-based approach for the prediction of chronic renal illness was presented in [6] by author P. Chittora et al. The authors used ANN and LSVM to predict CKD and seven machine literacy classifiers. applied data collecting on CKD again from the UCI repository. To extract pertinent properties, three selection techniques—filter, wrapper, and embedding method—were used. They also used LSVM to achieve the highest accuracy of 98.46%. Ifraz et al.'s approach[5] for predicting the status of chronic kidney disease (CKD) from clinical data includes feature extraction, data aggregation, preprocessing, and a mechanism for handling missing values. In this work, several physiological variables were utilized to train three different models for trustworthy forecasting using machine learning (ML) approaches such LR, DT classification, and KNN. With an accuracy of almost 97% in this investigation, the LR approach was found to be the most precise in this capacity. Ariful Islam et al. examined twelve different machine learning-based classifiers in a supervised learning environment[4]. In factIn a supervised learning environment, twelve different machine learning-based classifiers were assessed. Of these, the XgBoost classifier produced the best performance metrics: 0.983 for accuracy, 0.98 for precision, 0.98 for recall, and 0.98

for F1-score. A variety of supervised learning approaches are employed model training procedure to produce a strong ML model. SVM and RF achieved the less test accuracy and false-negative rates of all applied learning algorithms, with respective values of 98.67% and 99.33% by author Swain D et al[3]. However, when the validation method was 10-fold cross-validation, SVM outperformed RF. Additionally, Iftikhar H et al[2] utilized a variety of ML models to analyze chronic kidney disease, including LR, probit regression, RF regression, DT , KNN regression, and SVM using four kernel functions. The dataset is an assortment of documents from the district of Buner, Khyber Pakhtunkhwa, Pakistan, which were employed in a case-control investigation concerning patients suffering from chronic kidney disease. The author computed several performance metrics, such as Sensitivity, accuracy, Brier score, Youdent, specificity, F1 score, in order to compare the models' classification and accuracy. The random forest is competitive, but the SVM with the Laplace kernel function beats every other model, as the findings verify. The DNN model developed by the Rahul Sawhney et al[1] outperforms common ML models like SVM and naive Bayes classifiers, achieving 100% accuracy in the diagnosis of chronic kidney disease. The multi-layer perceptron classifier, which is based on the deep neural network made available by the PyTorch library, is thoroughly explained in this paper. An improved substitute for adaptation methods in the classification of chronic renal disease may be neural models. Because they can compute large data heaps fetched from datasets, manage non-linearity in the data, and use the layers of neurons contained in the structure to adapt and learn about the important information on their own.

After Analyzing above literature study, there are lots of gaps were identified because of that gaps result accuracy were affected like, majority of the papers in the literature examined the not numerical feature outliers present within the data preparation stage. The majority of the literature used incorrect data to train its models, resulting in an unequal model. A large portion of this literature failed to take consideration the use of feature selection to find the appropriate and efficient number of features. As an outcome in this circumstance, the models received a set of irrelevant features. This boosts the expense of the check-up for this kidney sickness. Most of the accuracy in this paper has been collected from an unbalanced dataset, where most of the authors have not used any methods to balance the data, due to which the result accuracy is overestimated. But this will be considered a useless result because the data set here is unbalanced or not accurate. In future, we use more features, classifiers and data preprocessing steps for unbalanced data, we can get more accuracy which will also be considered trustworthy.

Table 1. Review of the appropriate research for all investigations.

Author	Year	Approach	Dataset	Accuracy
Ariful Islam et al[3*4]	2023	XgBoost	UCI Repository	XgBoost with 98%
Ifraz et.al.[5]	2021	DT, KNN	UCI Repository	DT with 96.25%
P. Chittora et.al[6]	2021	ANN, LSVM	UCI Repository	98.46% with LSVM
Senan et.al.[7]	2021	RF	UCI Repository	RF with 96%
Deepika[8]	2020	KNN and Naïve Bayes supervised	UCI Repository	KNN with 97%
Ogunleye[9]	2020	XGBoost (Extreme Gradient Boosting)	UCI Repository	Accuracy 98.7%
Vasquez-Morales[10]	2019	neural network model	4000 instances	NN with 95 percentage
Rady & Anwar[11]	2019	PNN, SVM, MLP	UCI Repository has 361 records	PNN with 96.7%

Rustam[12]	2019	RF-SVM	dataset with 48 samples 36 trained & 12 testing records	RF-SVM with 83.4%
Polat[13]	2017	SVM	UCI Repository	98 percentage
Chen[14]	2016	KNN, SVM	UCI Repository	SVM with 97 percentage

III. RESEARCH METHODOLOGY

In this situation, we provide an overview of our suggested approach for the identification of persistent kidney illness.

A. Summary of our Proposed Model

All of the data used in this work was gathered from the raw data and the UCI was pre-processed using an approach called pre-processing of data to make it suitable for use with ML classification algorithms[10]. Fig 1 depicts a proposed generalized diagram of a system. The dataset is then subjected to more analysis in order to address any missing data. Using the label encoding approach, Numerical data was generated from the category data[14]. After pre-processing, two groups are created from the data set. One is the training set, which contains greater than 75% of the data needed to predict the values of the attributes. The other is the test set, which has a 25% testing portion allocated to it. Next, training data is used to test the algorithms GB, KNN, XGB, ADB, RF, DT, and GNB. We then used the test dataset to apply training methods and analyze each algorithm's recall, f1-score, accuracy, and precision. To choose most accurate, dependable algorithm for diagnosing Chronic Kidney Disease, we also evaluated performance and prediction accuracy of the six selected algorithms.

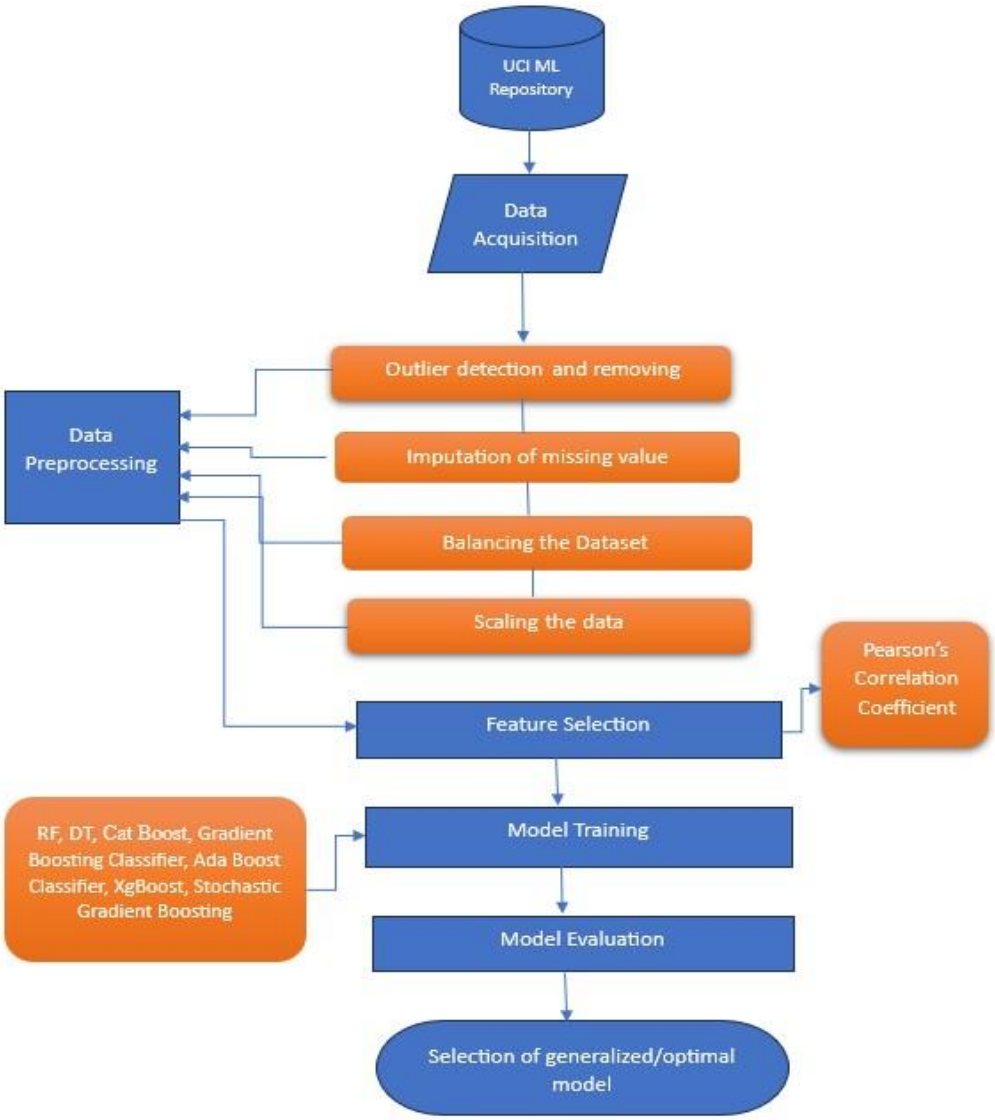

Fig 1. Workflow of the proposed method

B. Performance Measurement Indicator

Accuracy, recall, f1-score, precision, and other performance evaluation metrics have been employed to evaluate the efficacy and accuracy of our model's six ML strategies. Positive categorization was determined to have occurred upon a person's diagnosis of CKD. Negative classification, on the other hand, happens when the presence of CKD is not found. The above indicators' performance outcomes rely on TP, TN, FP, and FN.

Table 2. An explanation of the various evaluations.

TP	True positive	Model recognized CKD accurately.
TN	True negative	In this no CKD was accurately recognized by the model.
FP	False positive	The model misclassified a case as having chronic kidney disease (CKD), misidentifying non-CKD individuals as having the disease.
FN	False negative	A patient with CKD was mistakenly classified by the model as a non-CKD case.

C. Algorithm Phase

Chronic Kidney Failure is the input.
Output of the Dataset: Prediction Model with High Accuracy
Step 1: Enter data
Step 2: Prepare the data, detect and remove outliers.
Transform value numbers into category ones in step 2.1.
Step 2.2: To replace absent numerical values, Use the mean.
Step 2.3: Mode Replacement for Categorical values that are missing.
Step 3: Feature Selection for feature reduction.
Step 3.1: Create ML Classifier Model.
Create a model step 3.2.
Step 4: Using confusion matrix, evaluate the mathematical accuracy of the models generated.
Step 5: Select one of the best CK Model for prediction.

IV. IMPLEMENTATION

Data are regarded as the primary component of the study. The 400 instances and 25 variables that make up the proposed dataset for this work were obtained from the Apollo Hospital in Karaikudi, Tamil Nadu, using the UCI ML Repository [16]. Employing a dataset of 400 occurrences, we tested our method using 25 attributes—11 numerical features, 14 nominal characteristics, and ONE target class. They fall into two categories: non-CKD (150) and CKD (250). Using label encoding, we converted all of the dataset's categorical data into numerical values. For example, the terms "not-CKD" and "CKD" stand for 0 and 1.

Outliers: Extreme values that deviate from the characteristic central tendency are called outliers.
The primary cause of incorrect outliers, sometimes referred to as data noise, is data input problems.

The box plot below Fig 2 shows the largest outlier found in cad and contains 34 outliers, 22 outliers in ba, 16 outliers in su, 9 outliers in sc, 5 outliers in sod, 3 outliers in bp, 2 outliers in pot, rc, and pcv, and 1 outlier in al. In this case, the outliers were eliminated using the Z-Score method. There are now 324 rows in the dataset. This dataset will be used for further processing. As explained in the section on missing data, the extreme data points in this study that are outside of the range considered to be appropriate for medical care were considered missing data and have now been rectified. To find outliers in the CKD dataset, box plots have been employed.

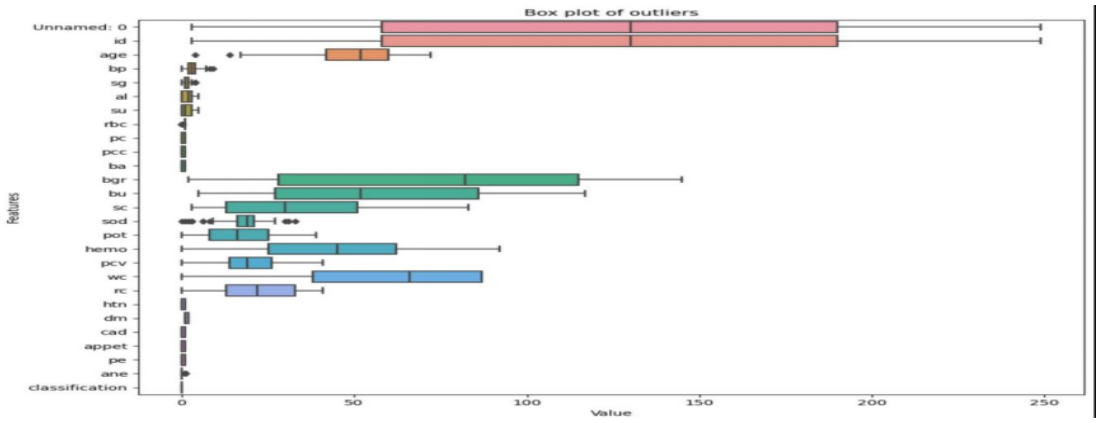

Fig 2. Box plot for outliers

A. Data Preprocessing

Following the completion of our collection of many unfiltered data sets from the UCI Repository, we pre-processed the dataset. We have discovered in this section that the majority of the dataset's columns contain missing values. Because The dataset contains values that are missing, it is difficult to provide an accurate result. After the pre-processing method was finished and no missing values were discovered, one of the most often used approaches was used. The label encoding technique was used to convert the category data into numerical values [14].

Fig 3 illustrates the association between twenty-five distinct features. The strength of connection is indicated by the bar on the right, which goes from darker to brighter [15]. The color is lighter and the link is greater the higher the value. The value range in the bar is -0.8 to 0.6, as can be seen. Here, feature "id" has a substantial correlation with features "rc," "pcv," and "hemo," all of which have values close to 0.6. characteristics "htn" and "ane" show less link with the characteristic "hemo," considering the fact that the color is darker and the value is closer to -0.6 or -0.7. A few more correlated features between the "pe" and "ane" and "classification" and "pot" features have values that range from roughly 0.1 to 0. Once more, the correlation values between 'rbc', 'hemo', 'rc', 'pcv' respectively, 0.34, 0.37, and 0.34. Week association between feature “classification”, "id," "sg," "hemo," "pcv," and "rc" is shown, with values ranging from -0.68 to -0.83. Once more, there is a significant link between feature Cad, Dm and features PCC, WC, and HTN.

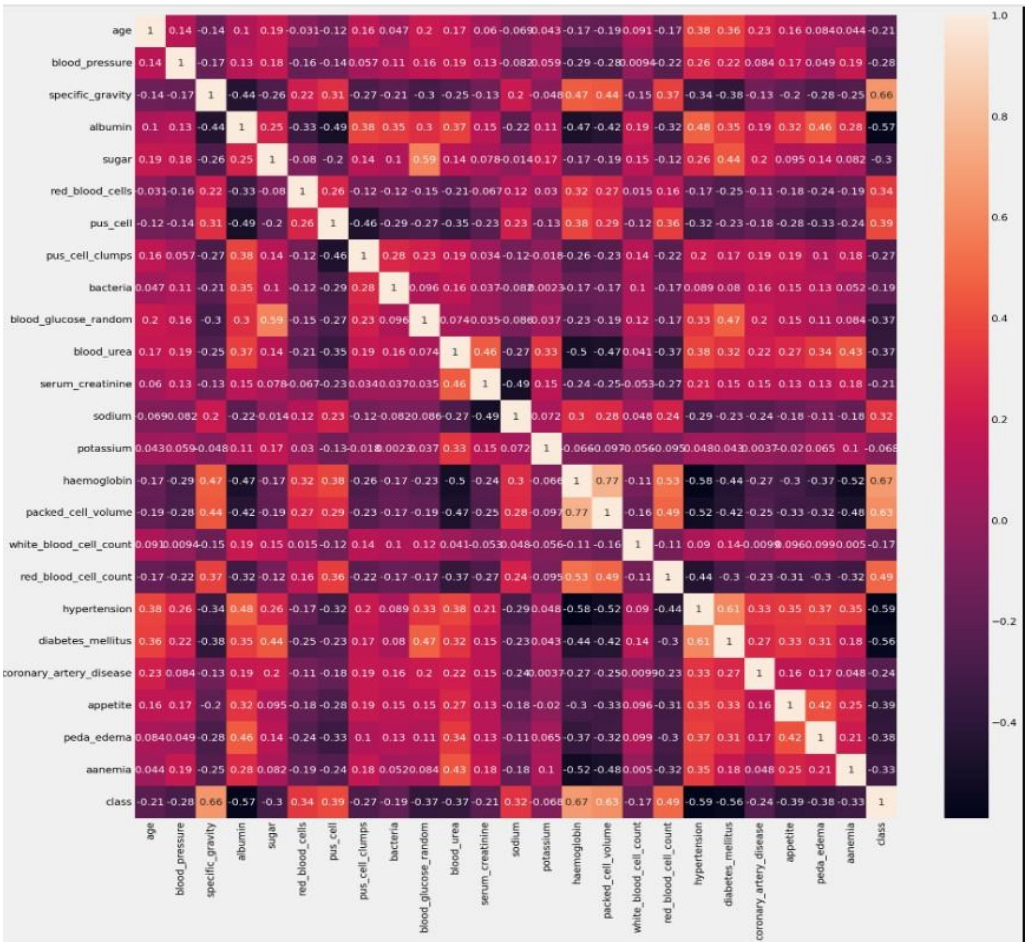

Fig 3. HeatMap

B. Pearson's Correlation Coefficient

To find the degree of linear dependency between two continuous variables, apply the correlation coefficient formula. It generates numbers in the range of -1 to +1. The features can be ordered as shown once the Pearson correlation between each feature and the aim variable (Class) has been applied in Fig 3: According to a Pearson correlation study, potassium has the lowest association with the target variable and hemoglobin the greatest. According to this, hemoglobin is the most important characteristic and potassium is the least important.

C. PCA-based CKD dataset

Variables that are unrelated to response variables or not predictively useful can be excluded by obtaining predictors or feature vector extraction. Consequently, factors unconnected to the matter at hand would not impact the model's development, leading to precise forecasts from the models. demonstrates the results of the procedure that was used to take significant variables out of the data. The purpose of the machine learning models that follow is to diagnose chronic kidney disease (CKD). Every CKD dataset was subjected to the corresponding subset of traits or predictors in order to create each model.

D. Algorithms for Machine Learning Classification

a) Decision Tree: When predicting The root node of a decision tree is where the class of an algorithm for a particular dataset starts. Taking into account the training sample's attribute value, it divides the data into several classes [15]. This software compares the original property values with the record (real dataset) attribute values. After that, it advances to the node after it, pursuing the branch in accordance with the result of the comparison. To get the most information gain, an attribute is split first when using a decision tree method, which maximizes the information gain value.

b) Random Forest: Random forests are useful in many different applications, such as feature selection, recommendation engines, and image classification. It generates decision trees using randomly selected data samples. Next, it compiles predictions from every tree and uses voting to determine which option is the best. The RF method also has the ability to provide a fair indicator of the feature relevance. The ability to determine the relative significance of each feature in a forecast is the most amazing feature.

c) AdaBoost: Ada-Boost classifier is an effective technique that raises the accuracy of the final classifier by combining several under performing classifiers. It can therefore provide a robust classification with a high accuracy rate. AdaBoost is an iterative ensemble technique. Fixing classifier weights, and for every iteration, training a sample of data are the goals of utilizing ADB in order to guarantee accurate predictions of anomalous occurrences.

d) Gradient Boosting: This machine learning method is used to develop a prediction model for tasks like regression, classification, and other ones. When dealing with relatively weak learning techniques or hypotheses, the Gradient Boosting Classifier is used. It is followed by several weeks, which strengthens either the student or the theory. As an ensemble boosting process, gradient boosting (GB) starts with a "regression tree" and uses "weak learners." All things considered; The GB model uses a sequential sampling strategy and adds a technology that reduces the loss function. The loss function determines the difference between predicted and actual numbers. Better accuracy is produced by GB because it lessens bias and volatility. This technique also has an excellent option for least squares regression, which is quite simple to understand.

e) Stochastic Gradient Boosting: The foundation of neural networks, stochastic gradient boosting, entails sub sampling the training set of data, followed by training Every learner with the resulting arbitrary samples. Findings with less association are integrated to improve our end result, which is a decreased Relationship between the information from various learners. For the gradient descent, estimates based on probability are used. The approach just estimates the gradient for each step; it does not compute the gradient throughout the whole dataset (Ayodele, 2010). Rather, it computes the gradient for a single observation selected at random.

f) Extreme Gradient Boosting: XGBoost is an ML algorithm that functions as a group and adheres to a sequence. It improves prediction by combining collective of poor learners. Using an efficient distributed library, this technique aims to provide maximum flexibility and efficiency. This tree-building strategy is also used by Portable XGB, where the optimal node splits are indicated by the Similarity Score and Gain.

V. RESULT AND DISCUSSION

There are 400 instances in the dataset for this study, and there are 25 attributes total—Fourteen nominal features and eleven numerical features. Label encoding was used to transform the category data in this dataset into a numerical value[14]. We used the test data to calculate the classifier's accuracy score and compared them after the testing and training phases were finished. We have since trained and tested. We calculated the f1-score, recall, accuracy, and precision to assess the models' performance. Positive classification only happens if an individual

exhibits signs of renal illness. On the other hand, a person is classified negatively if they do not develop chronic renal disease. Table 3 displays the outcome of the prediction.

Table 3. Performance evaluation.

	DT	RF	SGC	GBC	XGBOOST	ABC
Recall	0.78	1.00	1.00	1.00	1.00	1.00
Precision	1.00	0.98	0.98	0.98	0.98	0.98
F1 Score	0.87	0.99	0.99	0.99	0.99	0.99
Support	58	58	58	58	58	58
Tp	40	39	39	39	39	39
Tn	45	58	58	58	58	58
Fp	13	0	0	0	0	0
Fn	0	1	1	1	1	1

A. Outcome without of feature selection

In this section, results were evaluated using each of the six machine learning classification techniques using 80% training data and 50% testing data. There was a comparison table made for each algorithm. Following a comparison of all classifiers, the techniques employed by XGBoost, RF, SGB, ABC, and GBC were shown to have the highest training accuracy (100). Table 4 presents a comparison of accuracy, recall, and precision.

Table 4. Dataset without Feature Selection.

	Recall	**Precision**	**F1 Score**	**Train Accuracy**	**Test Accuracy**
Random Forest Classifier	1.00	1.00	1.00	100	100
Stochastic Gradient Boosting	1.00	1.00	1.00	100	100
XgBoost	1.00	0.98	0.99	100	98.9
Ada Boost Classifier	1.00	1.00	1.00	100	100
Gradient Boosting Classifier	1.00	1.00	1.00	100	100
Decision Tree Classifier	0.98	0.97	0.97	98.6	96.9

Table 5. Result of machine learning model.

	UCI Repository (Dataset1)	**Kaggle Repository (Dataset2)**	**Pakistan hospital (Dataset3)**
Random Forest	1.000000	1.000000	0.991667
Stochastic Gradient Boosting	1.000000	1.000000	1.000000
XgBoost	1.000000	1.000000	1.000000
Ada Boost	0.991667	1.000000	0.991667
Gradient Boosting	0.991667	1.000000	1.000000
Decision Tree	0.908333	0.975000	0.941667

B. Proposed Model's Comparison With Earlier Research

There are just a few research that use supervised methods and algorithms have tackled the problem of early detection of CKD. Still, a few notable studies were published that used unsupervised and semi-supervised learning techniques for the identification of CKD. Relevant research with comparable performance can be found in Table 5.5. The

comparison table makes it clear that CKD has never been detected in a prior study with a detection accuracy higher than 99.1%. By comparison, 99.2% accuracy was the highest achieved by the provided approach with the Gradient Boosting Classifier and Random Forest Classifier and a maximum accuracy of 100% with the Cat Boost algorithm and Stochastic Gradient Boosting. The majority of research did not use feature selection methods and those that did provide an explanation for the exclusion of particular features did not clean the data or use noisy data to build models. The research eliminates the less significant variables and identifies the most important ones for illness prediction.

Table 6. Comparison of the proposed and present work.

AUTHOR	REFERENCES	YEAR	DATASET	ALGORITHM	ACCURACY
Ariful Islam et.al.	[4]	2023	UCI Repository	XgBoost	98%
Ifraz et.al.	[5]	2021	UCI Repository	DT	96.25%
P. Chittora, et.al.	[6]	2021	UCI Repository	LSVM	98.46
Deepika	[8]	2020	UCI Repository	KNN	97%
Vasquez-Morales	[10]	2019	4000 instance	NN	95%
Rady & Anwar	[11]	2019	UCI Repository has 361 records	PNN	96.7%
Rustam	[12]	2019	dataset with 48 samples 36 trained & 12 testing records	RF-SVM	83.4%
Chen	[14]	2016	UCI Repository	SVM	97%
Proposed system			UCI ML Repository	Stochastic Gradient Boosting and Gradient Boosting Classifier	99.02%
			https://www.kaggle.com/abhia1999/chronic-kidney-disease	Stochastic Gradient Boosting and Gradient Boosting Classifier	99.02%
			The record includes more than 172 individuals, of whom 109 cases are expected to have chronic renal failure and the remaining 63 cases are not expected (Pakistan)	Random Forest Classifier	100%

VI. CONCLUSION

A potentially fatal illness called Chronic Kidney Disease (CKD) impact nearly 14% of the world's population. Patients can receive therapy at the lowest potential risk and expense by correctly diagnosing illness in its early stages. This research examines the efficacy of ML algorithms in identifying CKD using minimal tests or features. To do this, we use seven machine learning classifiers: DT, RF, SGB, XgBoost, Ada Boost Classifier, and Gradient Boosting Classifier, using a small data collection of 400 records. The Pearson's coefficient features were used to compute each classifier's results. This paper suggests a novel method that addresses missing values, handles the presence of outliers, and involves data preparation in order to forecast whether CKD status is positive or negative. Testing and training have been done

on the classifiers. Better results were obtained with the gradient boosting approach and SGB, as indicated by the F1-score (99%), accuracy (100%), and precision (98%). This is the best outcome among previous studies with fewer characteristics. We wish to compare the results using a different dataset with similar qualities or use a larger dataset in the future to validate our findings because the data used in this study was rather little. In addition, in order to contribute to the decline in the frequency of CKD, our intention to make use of pertinent datasets in order to predict the probability that An individual with risk characteristics for CKD, such as diabetes, hypertension, and a family history of kidney failure, would experience kidney failure in the future or not.

REFERENCES

[1] Rahul Sawhney, Aabha Malik, Shilpi Sharma, Vipul Narayan, A Comparative Assessment Of Artificial Intelligence Models Used For Early Prediction And Evaluation Of Chronic Kidney Disease, Decision Analytics Journal, Volume 6, 2023,100169, Issn 2772-6622, Https://Doi.Org/10.1016/J.Dajour.2023.100169.

[2] Iftikhar H, Khan M, Khan Z, Khan F, Alshanbari Hm, Ahmad Z. A Comparative Analysis Of Machine Learning Models: A Case Study In Predicting Chronic Kidney Disease. Sustainability. 2023; 15(3):2754. Https://Doi.Org/10.3390/Su15032754.

[3] Swain D, Mehta U, Bhatt A, Patel H, Patel K, Mehta D, Acharya B, Gerogiannis Vc, Kanavos A, Manika S. A Robust Chronic Kidney Disease Classifier Using Machine Learning. Electronics. 2023; 12(1):212. Https://Doi.Org/10.3390/Electronics12010212.

[4] Md. Ariful Islam, Md. Ziaul Hasan Majumder, Md. Alomgeer Hussein, Chronic Kidney Disease Prediction Based On Machine Learning Algorithms, Journal Of Pathology Informatics, Volume 14, 2023, 100189, Issn 2153-3539, Https://Doi.Org/10.1016/J.Jpi.2023.100189.

[5] Gazi Mohammed Ifraz, Muhammad Hasnath Rashid, Tahia Tazin, Sami Bourouis, Mohammad Monirujjaman Khan, "Comparative Analysis For Prediction Of Kidney Disease Using Intelligent Machine Learning Methods", Computational And Mathematical Methods In Medicine, Vol. 2021, Article Id 6141470, 10 Pages, 2021. Https://Doi.Org/10.1155/2021/6141470.

[6] Chittora P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., Jasinski, M.F., Jasiński, Ł., Goňo, R., Jasińska, E., & Bolshev, V.E. (2021). Prediction Of Chronic Kidney Disease - A Machine Learning Perspective. Ieee Access, 9, 17312-17334.

[7] Senan Em, Al-Adhaileh Mh, Alsaade Fw, Aldhyani Thh, Alqarni Aa, Alsharif N, Uddin Mi, Alahmadi Ah, Jadhav Me, Alzahrani My. Diagnosis Of Chronic Kidney Disease Using Effective Classification Algorithms And Recursive Feature Elimination Techniques. J Healthc Eng. 2021 Jun 9;2021:1004767. Doi: 10.1155/2021/1004767. Pmid: 34211680; Pmcid: Pmc8208843.

[8] Deepika B, Rao Vkr, Rampure Dn, Prajwal P, Gowda Dg, Et Al (2020) Early Prediction Of Chronic Kidney Disease By Using Machine Learning Techniques. Am J Comput Sci Eng Surv Vol. 8 No. 2:7.

[9] Ogunleye A, Wang Qg. Xgboost Model For Chronic Kidney Disease Diagnosis. Ieee/Acm Trans Comput Biol Bioinform. 2020 Nov-Dec;17(6):2131-2140. Doi: 10.1109/Tcbb.2019.2911071. Epub 2020 Dec 8. Pmid: 30998478.

[10] G. R Vasquez-Morales, S. M. Martinez- Monterrubio, P. Moreno-Ger, And J. A. Recio-Garcia, ‘‘Explainable Prediction Of Chronic Renal Disease In The Colombian Population Using Neural Networks And Case-Based Reasoning,’’ Ieee Access, Vol. 7, Pp. 152900–152910, 2019.

[11] Rady, El Houssainy & Anwar, Ayman. (2019). Prediction Of Kidney Disease Stages Using Data Mining Algorithms. Informatics In Medicine Unlocked. 15. 100178. 10.1016/J.Imu.2019.100178.

[12] Rustam, Zuherman & Sudarsono, Ely & Sarwinda, Devvi. (2019). Random-Forest (Rf) And Support Vector Machine (Svm) Implementation For Analysis Of Gene Expression Data In Chronic Kidney Disease (Ckd). Iop Conference Series: Materials Science And Engineering. 546. 052066. 10.1088/1757-899x/546/5/052066.

[13] 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 Apr;41(4):55. Doi: 10.1007/S10916-017-0703-X. Epub 2017 Feb 27. Pmid: 28243816.

[14] 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 Dec;48(12):2069-2075. Doi: 10.1007/S11255-016-1346-4. Epub 2016 Jun 22. Pmid: 27334750.

[15] Salekin, A., & Stankovic, J. (2016). Detection Of Chronic Kidney Disease And Selecting Important Predictive Attributes. In W-T. Fu, K. Zheng, L. Hodges, G. Stiglic, & A. Blandford (Eds.), Proceedings - 2016 Ieee International Conference On Healthcare Informatics, Ichi 2016 (Pp. 262-270). Article 7776352 (Proceedings - 2016 Ieee International Conference On Healthcare Informatics, Ichi 2016). Institute Of Electrical And Electronics Engineers Inc.. Https://Doi.Org/10.1109/Ichi.2016.36.

[16] Ckd Prediction Dataset. Available Online:Https://Www.Kaggle.Com/Datasets/Abhia1999/Chronic-Kidney-Disease (Accessed On 27 June 2022).

[17] Kaggle,“Chronic Kidney Disease Dataset,” https://www.kaggle .com/abhia1999/chronic-kidney-disease.