Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 3
Open Access

Advancing Chronic Kidney Disease Prediction through Machine Learning and Deep Learning with Feature Analysis

Shiddarth Dey Tusar[1,2], S. M. Ahad Ali Chowdhury[1,2], Md. Jalal Uddin Chowdhury[1,2,*], Rana M Pir[1], H M Nur A Alam[3], Muhammad Rezaur Rahman[4], Md Nurul Absar Siddiky[5], Muhammad Enayetur Rahman[6]

[1]Department of CSE Leading University, Sylhet 3112, Bangladesh

[2]DeepNet Research and Development Lab, Sylhet 3100, Bangladesh

[3]Boston Children's Hospital, 300 Longeood Avenue, Boston, MA 02115

[4]Department of Research and Innovation, Agile Crafts, Khilgaon, Dhaka, 1219, Bangladesh

[5]Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, Charlotte, NC, 28223, USA

[6]Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, 23529, USA

tusardey77@gmail.com, smahadalichowdhury@gmail.com, jalal_cse@lus.ac.bd, ranapir_cse@lus.ac.bd, Hafiz.Alam@childrens.harvard.edu, refayetbd@gmail.com, msiddiky@uncc.edu, mrahm011@odu.edu

*Corresponding Author: Md. Jalal Uddin Chowdhury

Article Info

Article type:
Research

Article History:
Received: 2024-07-16
Accepted: 2024-10-12
Published: 2024-11-25

Keywords:
Kidney Disease, Machine Learning, Feature Analysis, LASSO Algorithm, Random Forest, Deep Neural Network, and Feature Analysis.

A B S T R A C T

Introduction: Chronic Kidney Disease (CKD) is a significant health issue, ranking as the fourth leading cause of mortality worldwide. The traditional diagnosis and treatment process, reliant on medical experts, is time-consuming. Therefore, there is an urgent need for more efficient diagnostic methods to improve patient outcomes and reduce mortality rates. In this study, we employ Machine Learning (ML) and Deep Learning (DL) techniques to predict CKD based on important features. Feature analysis was performed using a correlation matrix and the LASSO algorithm to identify the most relevant features for model training. We evaluated several ML and DL classifiers, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbor (KNN), Decision Tree (DT), XG-Boost (XGB), Ada- Boost (AB), and Deep Neural Network. Our results demonstrated that the RF, LR, and SVM classifiers achieved the highest performance, with accuracy rates of 96.66%, 96.66% and 96% on average, respectively. This study identifies the most effective classifiers for CKD prediction and proposes a public model to facilitate early diagnosis, potentially reducing CKD-related mortality.

INTRODUCTION

One of the most important organs of the mortal body, the kidneys, diligently performing the vital task of filtering blood and expelling waste substances from the body [1]. These bean- shaped structured organs houses intricate networks of tubules and nephrons, introducing a complex filtration process essential for maintaining equilibrium. The kidneys always maintain the balance of electrolytic situations and blood pressure, just like they take care of the role of producing necessary hormones demanded for making red blood cells and metabolism of calcium [2]. This complex interplay underlines the essentiality of kidneys in supporting fleshly functions and underscores the urgency with which optimal functioning of the kidneys should be conserved. CKD is a notorious health challenge characterized by time-related corrosion of renal function [3]. Fastidious assessments of the Glomerular Filtration Rate (GFR) are critical indices of renal function that decide and manage CKD hinge on [4]. Advancing to further stages of CKD requires treatment interventions that are adapted, in terms of lifestyle modifications to pharmacological therapies, to help effectively relieve the related complications. In the face of the rapidly increasing frequency of CKD glob- ally, timely detection and visionary management policies are required to reduce the growing burden of this debilitating condition [5].

Due to the growing demand for heightened personal skills, croakers and experimenters have highly leaned on ML techniques for better comprehension and early detection of CKD [6]. Naturalistic, massive data sets from these studies aim to unlock subtle patterns within these massive data sets using the likes of NB, DTs, and SVM advanced algorithms in predicting the course of falling ill with CKD and its progression [7]. Through the power of ML, clinicians hope to usher in a new period of individualized therapeutic regimens in which early discovery and regular treatment become the bedrock of CKD operation, therefore offering renewed stopgap to the individualities scuffling with this insidious sickness. Numerous experimenters have formerly worked on determining CKD using ML [8].

Many researchers have already worked on predicting CKD using ML. Revathy Ramesh et al. [9] proposed a model using ML prognosticating CKD. It discusses the significance of early discovery of CKD and presents a vaticination frame that utilizes data preprocessing, data metamorphosis, and colorful classifiers to prognosticate CKD at an early stage. The algorithms used in this trial are DT, RF, and SVM. Surya Krishnamurthy et al. [10] proposed this paper exploring the development and evaluation of artificial intelligence (AI) models for prognosticating the threat of CKD in cases over a 6 or 12- month period. The study focuses on exercising minimum variables similar as sex, age, comorbidities, and specifics uprooted from electronic health records (EHR). The study employs colorful ML algorithms, including LR, DT, Light- GBM, and deep neural network similar as Convolutional Neural Networks (CNN) and Bidirectional Long Short- Term Memory (BLSTM). These algorithms are applied to different datasets, including added up, temporal-daily, and temporal-yearly data. Dibaba Adeba Debal et al. [11] proposed a model using ML prognosticating CKD. This paper focuses on the operation of ML ways, specifically RF, SVM, and DT algorithms, for the vaticination of CKD. The authors employ these algorithms on a dataset containing 19 features with numerical and nominal values, along with the class markers indicating the presence or absence of CKD. Two point selection styles, Recursive point Elimination with Cross-Validation (RFECV) and Univariate Feature Selection (UFS), are used to identify the most prophetic features for erecting the models.

In this paper, we propose a novel approach for predicting CKD using a streamlined dataset. Previous works have considered all features of the dataset, potentially including unnecessary ones. We applied the Correlation Matrix and Lasso Algorithm to the dataset to identify the dominant features with a deeper impact on CKD prediction. By excluding unnecessary features and focusing on dominant ones during preprocessing, training, and testing, we aim to reduce the redundant load on classifiers, saving time and increasing efficiency. This may help reduce the redundant cargo on classifiers and end up saving time and increase in proficiency.

LITERATURE REVIEW

Researchers have shown a lot of interest in the idea of creating surveillance and fore- casting methods and processes for diseases that significantly affect human health. This section provides an in-depth analysis of the interconnected research conducted by several researchers in the same field.

The authors of this paper [9] researched on the prediction of CKD using ML models. The paper explains how early detection of CKD is vital and also indicate how the predictive framework can be used for the early prediction of the prevalence rate of CKD. The experiment was performed using the DT, RF, and Class 3 SVM model. The classifiers were trained using input parameters obtained from rewards where CKD patients were involved and then developed models that can be used to predict CKD. The results show that the second model which employed the RF classifier gave the best accuracy of 99.16% compared to the other models that were used. The SVM model had an accuracy of 98.33% while the DT model had an accuracy of 94.16%, while the. A small amount of test data of only 120 instances are given. A larger amount of training data might have resulted in more accurate results.

This research [10], introduces AI-driven prediction models, predicting potential CKD over a 6- and 12-month period in patients, using minimal EHR variables such as sex, age, comorbidities, and medications. Classifiers include various LR, DT, RF, LightGBM, deep learning with CNN and bidirectional LSTMs. This work evaluates those classifiers over three datasets: aggregated, temporal quarterly, and temporal monthly. The highest AUROCs still were 0.957, 6 months and 0.954, 12 months, for the best algorithm. Remaining higher-risk predictors include older age, diabetes mellitus and gout, and use of some specific medications. These deep neural networks associated with the classical models in these types of predictions, indeed present in the prediction by integrating temporal information. Most laboratory results were unavailable in the present study. The most indispensable part of the study, though, is considered on the global generalization of the whole population, the amount and level of noise level of the dataset. Clinical practice from these findings requires the use of further clinical trials and, more so, features.

This paper [12] aims to understand how the applied ML algorithms are put to use in predicting CKD stages. The classification performance of patient data was compared between J48 and RF in this paper. A decision algorithm is applied for this study. The decision algorithm is applied for this study where the algorithm attained good performance in terms of precision, recall, and specificity on the CKD Stage 3A, although not very good. J48 yielded accuracy with the proportion of 94% in stage 3B, parallel sensitivity, and good specificity compared to RF. For stage 4, it was 95% with parallel specificity, while for stage 5, RF performs better with J48 in terms of accuracy and specificity. In this concern, J48 resulted in an accuracy of 85.5% under 0.03 s, while RF resulted in an accuracy of 78.25% at 0.28 s. However, the performance is a function of the size and characteristic of the dataset and cannot be transferred to another dataset or an algorithm. This needs further research and validation.

This paper [13] explores the perceptions of ML algorithm performance on CKD. Testing involves a dataset with 24 features from 400 cases diagnosed with CKD. Algorithms applied in this study: SVM, KNN, DT, and RF. In these algorithms, the data misclassification is done based on the dataset to classify and predict CKD. Here the results show that the RF algorithm is outstanding in its performance, as it has achieved ratings of 100 in terms of accuracy, perfection, recall, and F1-score criteria. While the DT model showed an equal performance to recommend its estimable one with an accuracy of 99.17%. Even though these promising issues are bound to a few certain limitations, the study is mainly based on the specific dataset taken from UCI ML Repository, causing the result bias. The assumption that the listed features are fully representative can affect generalization power. Similarly, in the work, there are no discussions regarding ethical issues and external validation of the models proposed.

In paper [14], the authors demonstrated their study about predicting CKD using ML techniques. Precisely, the authors applied the DT algorithm and the NB algorithm to the prediction classes with the aid of clinical data. The result of the work shows that the DT algorithm tends to have a better rate of accuracy at 99.25% compared to the NB one of 98.75\%. For this DT algorithm, sensitivity and specificity were obtained to be 99.33% and 99.20%, respectively,

which tops those obtained through NB. A larger number of attributes have been considered here, like age, bp, sg, al, su, bc, pc, pcc, etc. Keeping only serious ones could result in faster processing.

In the present paper [15], early prediction of CKD by ML is discussed. CKD often progresses without the individual feeling any symptoms until the later stages. So, the models take the data from the CKD patients and make use of DT and SVM for building the predictive model. DT: Makes a report for the prediction of CKD. SVM: Classifies and predicts relationships between coordinates in multi-dimensional spaces. For DT: Total Instances: 400. Number of Correctly Predicted: 367. Incorrect Predictions: 33. Accuracy (F1 Measure): 0.8959. Precision: 0.8503. Recall: 0.9467. The results were: Total Instances: 400; Correctly Predicted: 387; Incorrect Predictions: 13; Accuracy (F1 Measure): 0.9579; Precision: 0.9308; Recall: 0.9866. More and better data are needed to highlight ways of making this model more precise. Dataset limitations include its size, as well as missing values in.

This paper [11] prescribes the application of ML techniques for the prediction of CKD. The techniques use algorithms based on RF, SVM, and DT. Models trained are implemented against a data set prepared with 19 features and class labels, finally indicating the presence of CKD. Other feature selection techniques include Recursive Feature Elimination with Cross-Validation and Univariate Feature Selection. In the binary classification performance with CKD versus non-CKD, SVM was slightly the best with 99.8%, and RF came almost equally at 99.7% after the RFECV procedure. In the multi-class classification with the five stages of CKD levels, RF with RFECV actually gained the best accuracy of 79%. The characteristics that are limiting to this work are that the dataset is quite small, highly prone to overfitting, and the scope of the study was not set on unsupervised algorithms, including deep learning. Future efforts will be focused on realizing the mobile-based systems for real-time monitoring and applications of the evaluation with deep or unsupervised models.

The paper [16] discusses an offering entitled "Empirical Evaluation of ML Techniques for CKD Prophecy," which reflects an empirical comparison of ML techniques for predicting CKD. In the current paper, the authors try to evaluate and compare several ML algorithms' performance in classifying instances with or without CKD and achieve a base from the UCI ML repository. The paper evaluates a few of the ML algorithms with others, comprising NB (NB), LR, Multi-layer Perceptron (MLP), J48, SVM (SVM), NB Tree, and one hybrid algorithm, called CHIRP (Composite Hypercube on Iterated Random Projection). The experimental results show that CHIRP performs better performance-wise, compared to all others used on it. Several evaluations of performance are conducted based on error criteria, which include: MAE, RAE, RMSE, RRSE, Precision, Recall, F-Measure, and Accuracy. All these criteria consistently respect the order of the CHIRP algorithm in maximizing the accuracy and minimizing the error rate compared to other approaches.

The purpose of a literature review is to present an outline of the latest research and descriptions that are relevant to what they are doing or the topic of inquiry so that the material may be organized on a piece of paper. To obtain the best results, feature analysis is required for important variables that have a greater impact on the kidney. After reviewing multiple methodologies from various researchers, it emerged that many studies did not employ more feature analysis techniques employing distinct classifiers. In addition, after considering the findings of several studies, it emerged that ML methods were not applied extensively. After using different preprocessing procedures, a balanced set of data is generated, increasing the accuracy of the prediction model.

METHODS

The methodology gives a glimpse on how to use the power of ML techniques for pre- dicting CKD with precision and reliability. It comprised several critical steps, including data collection, data preprocessing, feature analysis, data splitting, classification using diverse ML algorithms, and performance evaluation. Workflow of this entire paper is shown at Figure 1.

A. Data Collection

The dataset used in this research was collected from the Kaggle repository [17], offering accurate records from 400 patients, each characterized by 25 distinct features pertinent to CKD diagnosis. Table I provides a feature description.

B. Data Preprocessing

Data preprocessing is an essential procedure before model development and training in ML. After

obtaining the relevant dataset, the next step is preprocessing it for developing models and training. The collected dataset went through extensive preprocessing procedures to ensure its suitability for accurate analysis [18]. The features 'red blood cells', 'pus cell', 'pus cell clumps', 'bacteria', 'hypertension', 'diabetes mellitus', 'coronary artery disease',

TABLE I. Input Features

No	Features	Data types	Illustration
1	age	Numerical	Represents the age of the patient
2	blood_pressure	Numerical	Indicates the blood pressure measurement of the patient
3	specific_gravity	Numerical	Refers to the density of urine relative to water
4	albumin	Numerical	A protein found in the blood that can leak into the urine when kidney function is impaired
5	sugar	Numerical	Represents the presence of sugar in the urine
6	red_blood_cells	Nominal	Indicates the presence or absence of red blood cells in the urine
7	pus_cell	Nominal	Represents the presence or absence of pus cells in the urine
8	pus_cell_clumps	Nominal	Indicates the presence or absence of clumps of pus cells in the urine
9	bacteria	Nominal	Represents the presence or absence of bacteria in the urine
10	blood_glucose_random	Numerical	Represents the random blood glucose level of the patient
11	blood_urea	Numerical	Measures the amount of urea nitrogen in the blood
12	serum_creatinine	Numerical	Measures the level of creatinine in the blood
13	sodium	Numerical	Measures the concentration of sodium in the blood
14	potassium	Numerical	Measures the concentration of potassium in the blood
15	haemoglobin	Numerical	Measures the amount of hemoglobin in the blood
16	packed_cell_volume	Numerical	Also known as hematocrit, measures the volume percentage of red blood cells in the blood
17	white_blood_cell_count	Numerical	Measures the number of white blood cells in the blood
18	red_blood_cell_count	Numerical	Measures the number of red blood cells in the blood
19	hypertension	Nominal	Indicates the presence or absence of hypertension (high blood pressure)
20	diabetes_mellitus	Nominal	Indicates the presence or absence of diabetes mellitus
21	coronary_artery_disease	Nominal	Indicates the presence or absence of coronary artery disease
22	appetite	Nominal	Represents the patient's appetite status
23	peda_edema	Nominal	Indicates the presence or absence of pedal edema (swelling of the feet)
24	aanemia	Nominal	Indicates the presence or absence of anemia
25	class	Nominal	Represents the target variable indicating the presence or absence of chronic kidney disease

'appetite', 'pedal edema', 'anemia', and 'class' are nominal and were converted to 1 and 0. There were many missing or null values in the dataset, which were addressed using two methods: random sampling for features with a high proportion of null values and mean/mode sampling for features with a lower proportion of null values.

C. Feature Analysis

Feature analysis was the most crucial part of this research that enabled us to identify the most potent predictors for CKD. We have applied both correlation matrix analysis and the Lasso algorithm in this study.

The correlation matrix can enable one to see the inter-relationship between the different features; on the other hand, the Lasso algorithm is able to identify only the valid features for accurate prediction of CKD [19]. Specifically, the correlation matrix revealed 8 features of prime importance; Figure 2 indicates the result with it, complemented by a personal 14 features that the Lasso algorithm said are indispensable. Figure 3 demonstrates the result.

D. Data Splitting

To facilitate robust model testing and training, the dataset was purposely partitioned into two distinct subsets: 80% designated for model training and the remaining 20% reserved for rigorous testing and validation [20].

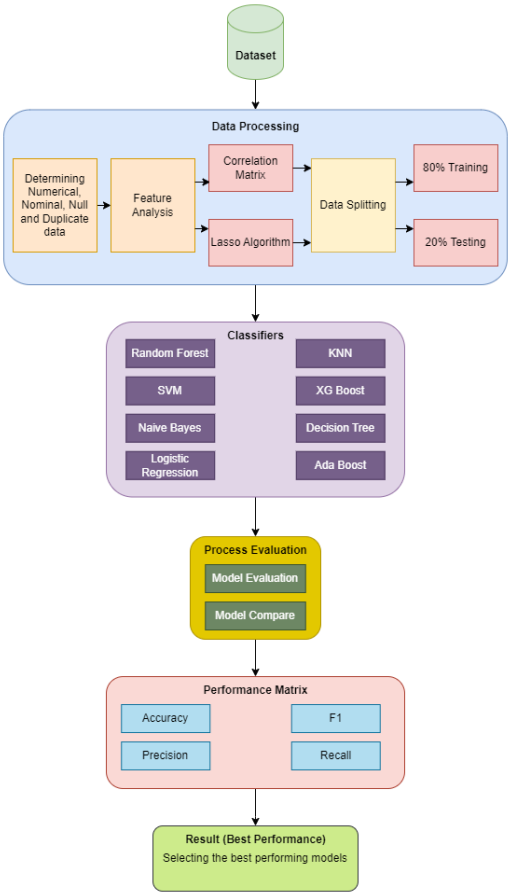

Fig 1. Overall workflow diagram

E. Classifiers

An array of state-of-the-art ML classification algorithms was enlisted for the task at hand. These included Decision Tree (DT), Random Forest (RF), AdaBoost (AB), XG-Boost (XGB), k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB) and Logistic Regression (LR).

I. Decision Tree

It is a supervised learning algorithm that is applied for both classification problems and regression problems. This model is one of the most supportive techniques for problems relevant to classification problems. A DT can be described as a tree structure or a tree containing many separators based on numerous features describing an outcome. There are two kinds of nodes a. Decision Nodes: It is a branched node. b. Leaf Node: This node is a decision and does not have any further classification. Decisions are made based on the features of the dataset, which in graphical terms, make the tree present all possible solutions to a problem under specified conditions in the form of a tree starting from the root and branching out [21].

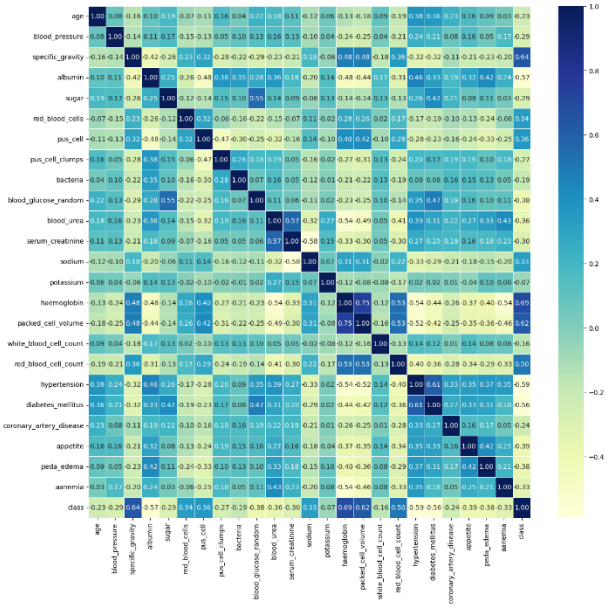

Fig. 2. Feature Importance from correlation matrix

II. Random Forest

A common supervised ML algorithm used to solve classification and regression problems in ML. It is based on the idea of ensemble learning: let's find a way to group many classifiers together to solve a complex problem and improve model performance. "RF is a classifier built by a number of DTs on various subsets of the given dataset and taking the average with optimizing the predictive accuracy of that dataset," as its name explains. In RF, the final prediction does not depend on any one tree; rather, it is the result of the majority of votes received by any of the pre- diction

trees in the ensemble. The more the number of trees in the forest, the lesser the variance and hence greater in accuracy; it saves from overfitting problems [22].

III. AdaBoost

It is an ensemble classifier developed by Yoav Freund and Robert Schapire in 1995, winning the Gödel Prize in 2003, for use in conjunction with other learning algorithms in the construction of this statistical classification meta-algorithm. AB Acquisition is a general technique that can be applied to a wide range of learning problems to boost performance. The boosting that arises is the linear combination of the weak hypotheses; this combines each boosting step into a weighted sum reflection of the final boosted classifier. Originally, AB was designed for binary classification; it generalizes to multiclass problems. It adapts to focusing on instances that are misclassified by earlier classifiers and, therefore, is considered not very prone to overfitting [23]. Since weak learners can be only slightly better than random guessing, the final model will converge to a strong learner.

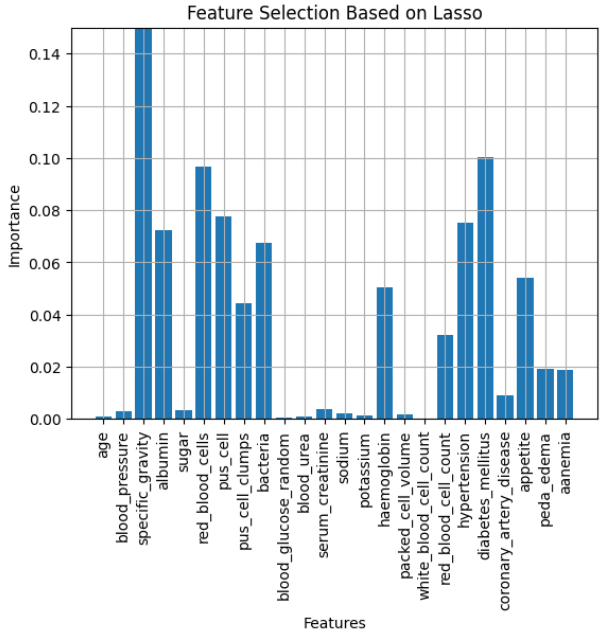

Fig. 3. Feature importance from LASSO

IV. XGBoost

Stands for Extreme Gradient Boosting. It is a very effective and scalable ML algorithm for a supervised task regarding classification, regression, and ranking. As such, it belongs to the family of gradient boosting algorithms that combine forecasts from many weak learners into a strong learner, most commonly trees. XGB optimizes model performance through an iterative building of a set of DTs. In other words, the later trees in a sequence focus on error correction based on previous steps. It employs optimization by gradient descent to reduce a predefined loss function and also effectively handles big data. That is why some of the major features of XGB include regularization techniques for preventing overfitting yet improving generalization performance. It consists of L1 (Lasso) and L2 (Ridge) regularization terms in the objective function that penalize complex models, directing the solution towards simpler models [24].

V. KNN

k-Nearest Neighbour is one of the simplest ML algorithms based on the supervised learning technique. Basically, the K-NN algorithm assumes that the new case or data is put into a category most similar to the available cases, with the new case put into it. The K-NN algorithm is a lazy algorithm since it stores all the available data but classifies a new data point based on the similarity. Therefore, when new data comes in, it is straightforward to use the K-NN algorithm to classify it into the appropriate category. The K-NN algorithm is most often used for classification. KNN is a non-parametric algorithm, insinuating that the algorithm does not presume anything about the data distribution [25]. During the training phase, the KNN algorithm only stores the dataset and, when new data is received, classifies that data based on its similarity to the data stored in the dataset.

VI. SVM

SVM is one of the most widely used supervised learning algorithms for classification, as well as regression, problems. But it is mainly used in ML for Classification 10 problems. The most important thing that an SVM algorithm tries to do is to drive the best-fitted optimal line or hyperplane dividing n-dimensional spaces into classes so that, in future, we can easily put a new data point into the correct category. This best decision boundary is referred to as a hyperplane and the support vectors that help in making this hyperplane are basically the extreme points/vectors [26].

VII. Naïve Bayes

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used

for solving classification problems. It is primarily used for text classification that includes a high-dimensional training dataset. NB Classifier is one of the most elementary and effective algorithms for Classification. It helps build ML models fast enough to make swift predictions. It is a probabilistic Classifier, meaning that it predicts based on the probability of an object [27].

VIII. Logistic Regression

It is a statistics method of analyzing a dataset used for predicting the value of a dependent variable based on one or more independent variables. The most common LR model is a binary outcome, which is something that takes on two possible values, such as ”yes” and ”no,” or ”true” and ”false.” Multinomial LR is helpful for modeling a situation in which more than two possible discrete outcomes exist. Classification problems cover that important type of analysis for which LR may be used. This is where an attempt is made to determine whether a new sample fits best into a particular category. Since aspects of cyber security involve a classification problem, LR will be useful as an analytic technique [28].

IX. Deep Neural Network

Deep neural networks are more complex artificial neural networks, usually with a much deeper hidden layer structure, and can also include many different layers such as convolutional layer (Conv), maximum-pooling layer (Pooling), dense or fully connected layer, and other specific layers. These extra layers are designed to allow the model to grasp issues much better and to use clever answers for complicated projects. A deep neural network contains more layers (i.e. depth) than ANN and additional layers add complexity to the model, however, it helps to process inputs in a much more concise manner for getting as close as possible to an ideal solution[29].

Neural Network Architecture: The neural network is designed in a sequential model with some numbers of layers[30]. A neural network contains some input layer, respectively, 24 neurons corresponding to the input dimensions. Then two hidden layers with 11 and 8 neurons are defined, followed by dropout layers to avoid overfitting. Then batch normalization layers are added in the subsequent layers to ensure stabilization and acceleration of the training process. There is a single output layer with one neuron based on the sigmoid activation function for the last classification.

Compilation and Training: The Adam optimizer with learning rate 0.001 and binary cross-entropy as loss functions is used. Fit the model on the train dataset (X train, Y train) up to 100 epochs with a batch size of 32. Early stopping will be set with a patience parameter of 10 epochs to counter overfitting.

Evaluation: The final step is the testing of the model using the test data set. Evaluation of a trained model with test dataset. Validation accuracy shows how the model relates to the test dataset, by giving a percentage of the correctly predicted labels. The final testing is performed by deploying the test data set. Calculation of the estimated metrices related functions is done from scikit-learn.

Using the unique strengths and characteristics of each algorithm, this study aimed to comprehensively evaluate their efficacy in predicting CKD, thus informing clinical decision-making processes.

F. Performance Indicator

The key point was performance evaluation, and continuous attention was paid to assessing model effectiveness considering 2 different sets of features, the 8 features taken forward using the correlation matrix and the 14 features that were put through the Lasso algorithm for selection. The 10-fold approach allowed for an in-depth comparison of the performance of the models by means of the subsets of features, thus yielding very important information on the best approach toward CKD prediction.

RESULTS

A. Performance of models Using All Features The dataset

In the beginning we run the training using all 25 features on classifiers such as RF, SVM, NB, LR, KNN, XGB, DT and AB. From here we can seen that the RF and LR has the most higher level of accuracy of 99% followed by SVM, XGB and AB of 97% accuracy, DT and NB having 96% and 95% accuracy and the lowest accuracy is of KNN which is 73%. Training dataset through DNN with all 25 features for 100 epochs with a batch size of 32 we get a Validation Accuracy of 65%. Table II shows the accuracy, precision, recall and f1 score for all classifiers.

TABLE II. Performance for all models

Algorithms	Accuracy	Precision	Recall	F1 Score

Random Forest	99%	0.99	0.99	0.99
SVM	97%	0.97	0.97	0.97
Naive Bayes	95%	0.95	0.95	0.95
Logistic Regression	99%	0.99	0.99	0.99
KNN	73%	0.79	0.72	0.73
XG Boost	97%	0.98	0.97	0.98
Decision Tree	96%	0.97	0.96	0.96
ADA Boost	97%	0.98	0.97	0.98
DNN	65%	0.99%	65%	78%

B. Performance of models Using Features Analysis by Correlation Matrix

We run the training set using 8 dominant features which are more impactful found by the correlation matrix on classifiers such as RF, SVM, NB, LR, KNN, XGB, DT, and AB. From here we can seen that the RF, SVM, LR and XGB has the most higher level of accuracy of 94% followed by DT, AB and NB of 93%, 91% and 90% accuracy, the lowest accuracy is of KNN which is 89%. training with DNN we get a very

TABLE III. Performance analysis using Correlation matrix

Algorithms	Accuracy	Precision	Recall	F1 Score
Random Forest	94%	0.94	0.94	0.94
SVM	94%	0.94	0.94	0.94
Naive Bayes	90%	0.91	0.90	0.90
Logistic Regression	94%	0.94	0.94	0.94
KNN	89%	0.90	0.89	0.89
XG Boost	94%	0.94	0.94	0.94
Decision Tree	93%	0.93	0.93	0.93
ADA Boost	91%	0.91	0.91	0.91
DNN	88.75%	100%	65%	78%

satisfying Validation Accurracy of 88.75%. Table III shows the results, containing the percentage and scores for accuracy, precision, recall, and F1-score for all these 8 classifiers and DNN.

C. Performance of models Using Features Analysis by Lasso Algorithm

Now we run the training using 14 dominant features found by the Lasso algorithm which are more impactful in determin- ing CKD. From here we can seen that the RF, SVM, LR and XGB has the most higher level of accuracy of 97% followed by DT, AB of 96% accuracy, the lowest accuracy is of NB and KNN which is 94%. Training with DNN we get Validation Accurracy of 65%. Table IV is the result table containing the percentages and scores for accuracy, precision, recall, and F1- score for all these 8 classifiers and DNN.

TABLE IV. Performance analysis using LASSO Algorithm

Algorithms	Accuracy	Precision	Recall	F1 Score
Random Forest	97%	0.98	0.97	0.97
SVM	97%	0.97	0.97	0.97
Naive Bayes	94%	0.94	0.94	0.94
Logistic Regression	97%	0.97	0.97	0.97
KNN	94%	0.94	0.94	0.94
XG Boost	97%	0.98	0.97	0.97
Decision Tree	96%	0.96	0.96	0.96
ADA Boost	96%	0.96	0.96	0.96
DNN	65%	100%	65%	78%

D. Performance Anlysis using ROC Curve

In order to demonstrate the performance of each classifier, a Receiver Operating Characteristic curve is drawn. It means that it is a plot of True Positive Rate against False Positive Rate of a classifier. The more the AUC score the better the result we will have.

I. ROC Curves using CM's 8 features

RF, SVM, LR and XGB seems to have the highest AUC score of 0.94 followed by DT of score 0.93. NB and AB

has a score of 0.91 and KNN has the lowest score of 0.90. Figure 4 shows the ROC Curves and table V shows the AUC scores.

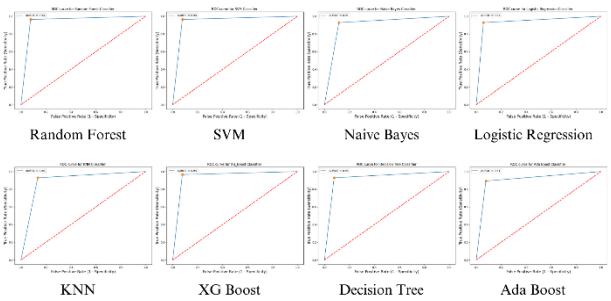

Fig. 4. ROC Curves of classifiers using CM's features

II. ROC Curves using Lasso Algorithm's 14 features

SVM and LR seems to have the highest AUC score of 0.97 followed by RF and XGB of score 0.96. DT and AB has a score of 0.95, NB and KNN has the lowest score of 0.94. Figure 5 shows the ROC Curves table VI shows the AUC scores.

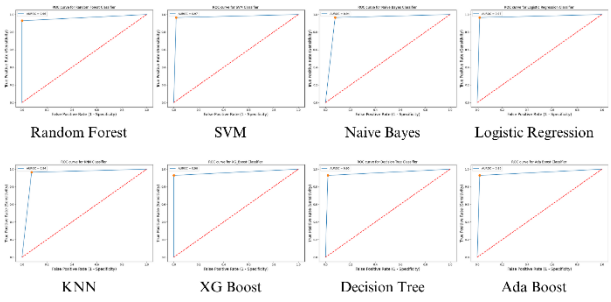

Fig. 5. ROC Curve of classifiers using LASSO's features

E. Discussion

Comparing the performance of ML models using different sets of features. Using 25 Features, RF and LR achieved the highest accuracy of 99%. SVM, NB, XGB, DT, and AB also performed well with accuracies ranging from 95% to 97%. KNN had the lowest accuracy at 73%. Using 8 Features RF, SVM, and LR maintained high accuracies but slightly decreased to 94%. NB, KNN, XGB, DT, and AB also experienced decreases n accuracy, with values ranging from 89% to 94%. Using 14 Features RF, SVM, LR, and XGB maintained their high accuracies at 97%. NB, DT, and AB had slightly lower accuracies compared to the 25-feature scenario but still performed well at 94% to 96%. KNN had the lowest accuracy among the models at 94%.

CONCLUSION

In this paper, our aim was to minimize the quantity of features, thereby achieving a faster and more optimized response by eliminating unnecessary ones. We used a correlation matrix and the Lasso algorithm to determine the dominant feature and trained our models using them. Upon analyzing the results, we found that, on average, RF and LR achieved the highest accuracy, with an average of 97%. The model with the lowest performance to date is KNN, with an average accuracy of only 85%. This method of predicting CKD will assist doctors in diagnosing the disease more efficiently and quickly, without adding unnecessary burden to the model. In the future, we aim to develop a user interface for this method, ensuring its accessibility to a specific user base. We will also train these models using data from local private and public hospitals, thereby enhancing their accuracy for practical use.

REFRENCES

[1] Taylor, J. S. (2017). Stakes and kidneys: why markets in human body parts are morally imperative. Routledge.
[2] Barboza, G.D., Guizzardi, S., Talamoni, N.T.: Molecular aspects of intesti- nal calcium absorption. World Journal of Gastroenterology: WJG 21(23), 7142 (2015)
[3] Stanzick, K.J., Li, Y., Schlosser, P., Gorski, M., Wuttke, M., Thomas, L.F., Rasheed, H., Rowan, B.X., Graham, S.E., Vanderweff, B.R., et al.: Discovery and prioritization of variants and genes for kidney function in¿ 1.2 million individuals. Nature Communications 12(1), 4350 (2021)
[4] Dehlin, J.P., Morrison, K.L., Twohig, M.P.: Acceptance and commitment ther- apy as a treatment for scrupulosity in obsessive compulsive disorder. Behavior modification 37(3), 409–430 (2013)
[5] Fernandez-Prado, R., Esteras, R., Perez-Gomez, M.V., Gracia-Iguacel, C., Gonzalez-Parra, E., Sanz, A.B., Ortiz, A., Sanchez-Ni no, M.D.: Nutrients turned into toxins: microbiota modulation of nutrient properties in chronic kidney disease. Nutrients 9(5), 489 (2017)
[6] Silveira, A.C.d., Sobrinho, A., Silva, L.D.d., Costa, E.d.B., Pinheiro, M.E., Perku- sich, A.: Exploring early prediction of chronic kidney disease using machine learning algorithms for small and imbalanced datasets. Applied Sciences 12(7), 3673 (2022)

[7] Segal, Z., Kalifa, D., Radinsky, K., Ehrenberg, B., Elad, G., Maor, G., Lewis, M., Tibi, M., Korn, L., Koren, G.: Machine learning algorithm for early detection of end-stage renal disease. BMC nephrology 21, 1–10 (2020)

[8] Bai, Q., Su, C., Tang, W., Li, Y.: Machine learning to predict end stage kidney disease in chronic kidney disease. Scientific reports 12(1), 8377 (2022)

[9] Revathy, S., Bharathi, B., Jeyanthi, P., Ramesh, M.: Chronic kidney disease pre- diction using machine learning models. International Journal of Engineering and Advanced Technology 9(1), 6364–6367 (2019)

[10] Krishnamurthy, S., Ks, K., Dovgan, E., Luˇstrek, M., Gradiˇsek Pletiˇc, B., Srini- vasan, K., Li, Y.-C., Gradiˇsek, A., Syed-Abdul, S.: Machine learning prediction models for chronic kidney disease using national health insurance claim data in taiwan 9(5), 546 (2021)

[11] Debal, D.A., Sitote, T.M.: Chronic kidney disease prediction using machine learning techniques. Journal of Big Data 9(1), 109 (2022)

[12] Senan, E.M., Al-Adhaileh, M.H., Alsaade, F.W., Aldhyani, T.H., Alqarni, A.A., Alsharif, N., Uddin, M.I., Alahmadi, A.H., Jadhav, M.E., Alzahrani, M.Y., et al.: Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering 2021 (2021)

[13] Kaur, C., Kumar, M.S., Anjum, A., Binda, M., Mallu, M.R., Al Ansari, M.S.: Chronic kidney disease prediction using machine learning. Journal of Advances in Information Technology 14(2), 384–391 (2023)

[14] Ilyas, H., Ali, S., Ponum, M., Hasan, O., Mahmood, M.T., Iftikhar, M., Malik, M.H.: Chronic kidney disease diagnosis using decision tree algorithms. BMC nephrology 22(1), 273 (2021)

[15] Tekale, S., Shingavi, P., Wandhekar, S., Chatorikar, A.: Prediction of chronic kid- ney disease using machine learning algorithm. International Journal of Advanced Research in Computer and Communication Engineering 7(10), 92–96 (2018)

[16] Khan, B., Naseem, R., Muhammad, F., Abbas, G., Kim, S.: An empirical evalu- ation of machine learning techniques for chronic kidney disease prophecy. IEEE Access 8, 55012–55022 (2020)

[17] Rubini, S.P. L., Eswaran, P.: Chronic Kidney Disease. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5G020 (2015)

[18] Crone, S.F., Lessmann, S., Stahlbock, R.: The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. European Journal of Operational Research 173(3), 781–800 (2006)

[19] Fang, H., Huang, C., Zhao, H., Deng, M.: Cclasso: correlation inference for compositional data through lasso. Bioinformatics 31(19), 3172–3180 (2015)

[20] Joseph, V.R., Vakayil, A.: Split: An optimal method for data splitting. Techno- metrics 64(2), 166–176 (2022)

[21] Ilyas, H., Ali, S., Ponum, M., Hasan, O., Mahmood, M.T., Iftikhar, M., Malik, M.H.: Chronic kidney disease diagnosis using decision tree algorithms. BMC nephrology 22(1), 273 (2021)

[22] Petkovic, D., Altman, R., Wong, M., Vigil, A.: Improving the explainability of random forest classifier–user centered approach. In: Pacific Symposium on Bio-computing 2018: Proceedings of the Pacific Symposium, pp. 204–215 (2018).

[23] Ying, C., Qi-Guang, M., Jia-Chen, L., Lin, G.: Advance and prospects of adaboost algorithm. Acta Automatica Sinica 39(6), 745–758 (2013)

[24] Asselman, A., Khaldi, M., Aammou, S.: Enhancing the prediction of student per- formance based on the machine learning xgboost algorithm. Interactive Learning Environments 31(6), 3360–3379 (2023)

[25] Zhu, X., Ying, C., Wang, J., Li, J., Lai, X., Wang, G.: Ensemble of ml-knn for classification algorithm recommendation. Knowledge-Based Systems 221, 106933 19 (2021)

[26] Chowdhury, M. J. U., Hussan, A., Hridoy, D. A. I., & Sikder, A. S. (2023, March). Incorporating an Integrated Software System for Stroke Prediction using Machine Learning Algorithms and Artificial Neural Network. In 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0222-0228). IEEE.

[27] Reddy, E.M.K., Gurrala, A., Hasitha, V.B., Kumar, K.V.R.: Introduction to naive bayes and a review on its subtypes with applications. Bayesian reasoning and gaussian processes for machine learning applications, 1–14 (2022)

[28] Nusinovici, S., Tham, Y.C., Yan, M.Y.C., Ting, D.S.W., Li, J., Sabanayagam, C., Wong, T.Y., Cheng, C.-Y.: Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of clinical epidemiology 122, 56–69 (2020)

[29] Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Muller, K.-R.: Explain- ing deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE 109(3), 247–278 (2021)

[30] Chowdhury, M. J. U., & Kibria, S. (2023, December). Performance Analysis for Convolutional Neural Network Architectures using Brain Tumor Datasets: A Proposed System. In 2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP) (pp. 195-199). IEEE.