EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Machine Learning based Exploratory Data Analysis (EDA) and Diagnosis of Chronic Kidney Disease (CKD)

Vaishali Mehta[1], Neera Batra[2], Poonam[3], Sonali Goyal[4], Amandeep Kaur[5], Khasim Vali Dudekula[6*] and Ganta Jacob Victor[7]

[1,2,4,5]Department of Computer Science and Engineering, Maharishi Markandeshwar Deemed to be University, Mullana, Ambala, Haryana, India
[3]LNTE, Panipat, Haryana, India
[6]School of Computer Science and Engineering, VIT-AP University, Amaravati, AP, India
[7]Department of Computer Science and Engineering, Koneru Lakhmaiah Education Foundation, Green fields, Vaddeswaram, Guntur - 522302, Andhra Pradesh, India

Abstract

INTRODUCTION: This research paper presents an exploratory data analysis (EDA) approach to diagnose Chronic Kidney Disease (CKD) using machine learning algorithms.
OBJECTIVES: This paper focuses on early and accurate detection of CKD using a comprehensive dataset of clinical and laboratory parameters to minimize the risk of patients' health complications with timely intervention through appropriate medications.
METHODS: Machine Learning based prediction models including Naive Bayes, KNN, Logistic regression, decision tree, ensemble modelling, Random Forest and Ada Boost.
RESULTS: The results indicate that the Naive Bayes algorithm achieved highest accuracy and sensitivity in detecting CKD.
CONCLUSION: For reduced features and for binary class classification, Naive Bayes classifier gives best performance in terms of accuracy and computational cost. Other algorithms are good for multi-class classification but for binary class, they are little expensive than Naive Bayes.

Keywords: Chronic Kidney Disease, Machine Learning, Classification, Feature selection, Regression

Received on 12 December 2023, accepted on 17 March 2024, published on 22 March 2024

doi: 10.4108/eetpht.10.5512

1. Introduction

In the peritoneal cavity in the back of the human body are two kidneys. Salt, water, and other ions are all balanced by the kidneys' basic function. Additionally, it tracks substances like uric acid, blood urea, magnesium, potassium, calcium, and phosphorus. Kidneys secrete erythropoietin, vitamin D, and renin, three hormones that promote RBC synthesis and maturation, regulate calcium and phosphorus, control blood pressure, and get rid of all metabolic waste products. Chronic renal disease is primarily brought on by high blood pressure and sugar levels. A prevalent issue with kidney function called chronic kidney disease (CKD) results in deteriorating kidney function and kidney failure. In many circumstances, an early diagnostic approach is crucial and critically *significant to determining kidney functionality. Fatigue, swelling around the eyes, ankles, and legs Muscle cramps at night, difficulties falling asleep, and nightly urination are the primary symptoms CKD.

The current diagnostic method relies on the analysis of urine with the aid of serum creatinine levels. This is accomplished using a variety of medical techniques, including ultrasonography and screening. Patients who have hypertension, a history of cardiovascular disease, a current illness, or who have had kidney disease in a family member are all screened during the screening process. This method involves measuring the GFR (Glomerular filtration rate) in a first-morning urine sample as well as estimating from the serum creatinine level.

* Corresponding author. Email: khasim.vali@gmail.com

GFR is a useful metric for assessing renal function and determining the kidney's capacity to filter blood. The eGFR value is expressed in mL/min/1.73 m2, or millilitres per minute per square metre. According to eGFR, the renal function categorises CKD into five stages, as indicated in Table 1 (Baidya et al., 2022).

Table 1. Stages of Chronic Kidney Disease

Stage of CKD	Description	GFR (mL/min/1.73 m^2)
Stage 1	Normal	>=90
Stage 2	Mild CKD	60-89
Stage 3	Moderate CKD	30-59
Stage 4	Severe CKD	15-29
Stage 5	End stage CKD	<15

In order to increase prediction accuracy, this paper focuses on ML approaches by minimizing the features and choosing the best features. The paper is divided into 5 sections. Section 1 describes introduction of CKD. A brief of literature surveyed to carried out this work is presented in Section 2. Section 3 explains the proposed approach. Sections 4 depicts the simulation results and section 5 concludes the research.

2. Literature Review

The amount of medical data collected daily has significantly increased as a result of the technological revolution and the rise in health issues. As a result, finding the best method for predicting or diagnosing underlying disease has turned into a top priority for medical researchers. We have made an attempt to highlight important research on the use of machine learning classification strategies for the diagnosis of chronic kidney disease (CKD) in recent literature. Researchers have conducted studies in a variety of fields, including designing new machine learning models for data already available, applying existing algorithms onto a real dataset, as well as using various tools and evaluating the respective performance associated with each tool—that is, determining these tools/techniques which needs little computational work but offer high accuracy.

In first paper (Almasoud & Ward, 2013), the authors present two distinct frameworks: one for MLs, which categorizes many aspects of kidney disease diagnosis whereas the other is the framework of medically specific domains relevant to MLKDD. This article conducts a thorough literature analysis on the use of machine learning to diagnose kidney problems. Statistical metrics like accuracy, sensitivity, specify, etc. are utilized to assess the performance of diagnostic (prediction) procedures. This paper defines the effectiveness and precision of a used ML method for KDD which is found to be highly dependent on a number of factors, including the data being used for input, the process of preparing the data, the training dataset, and particularly the techniques used in feature selection. It is concluded that using ML in conjunction with comparison of utilizing either MLs or the GFR tests increases the accuracy of the procedure.

In (Baidya et al., 2022), the authors define how to develop an extremely precise CKD diagnosis method developed by utilizing an adequate feature set. This paper analyses various feature optimization methodologies combined with a max voting ensemble model by using five existing classifiers. This paper explains the best performance of the feature optimizer LDA with our ML model is a total in-accuracy of 0.5%, while the maximum inaccuracy of the ML model without any feature optimization is 1.62%.

In (Chittora et al., 2021), authors present a classifier (called ANFIS) that uses the neuro-fuzzy concept to determine the presence of chronic kidney disease and for academic research, they use the results of blood tests on a number of patients. Based on experimental findings, it is suggested that neuro-fuzzy rule-based classifier outperforms the existing classifiers. In comparison to the other classifiers, ANFIS has provided accuracy that is 3% to 4% higher. In (Elhoseny et al., 2019), authors examined frequently gathered emergency department (ED) data and created models of prediction with the ability to recognize ED patients at significant risk for severe kidney damage early on. They created classifiers based on ma-chine learning to predict the beginning of acute kidney injury stages 1 and 2 within the next 24 to 72 hours using creatinine-based worldwide agreement criteria. By using Monte Carlo cross validation, predictive performance was assessed outside of the sample.

In (Laaksonen & Oja, 1996), authors first apply class balancing then features are ranked and analyzed. Finally, numerous ML models are trained and assessed using a variety of performance indicators. The outcomes showed that the Rotation Forest (RotF) was the most efficient model, outperforming the comparison models with an Area Under the Curve (AUC) of 100%, Precision, Recall, F-Measure, and Accuracy equal to 99.2%. Authors in (Manonmani & Balakrishnan, 2020), explored a study uses KNN, Naive Bayes, and Decision Tree, using bag-ging and random subspace approaches to improve the models' ability to classify data accurately.

Authors in (Nikhila, 2021) proposed a model with the greatest performance, eXtreme Gradient Boosting (XGBoost), had an AUC of 89%.

3. Proposed Methodology

Prediction and analysis of CKD-suffering patients is the primary focus of this work. Moreover, the emphasis has also been made on minimizing the computational cost and maximizing the accuracy. As discussed in section 1 of the paper, numerous optimized data analysis strategies such as data preprocessing, feature selection and exploratory data

analysis have also been used in the machine learning algorithms to achieve the goal. The entire process of proposed work is illustrated in the flowchart in Fig. 1.

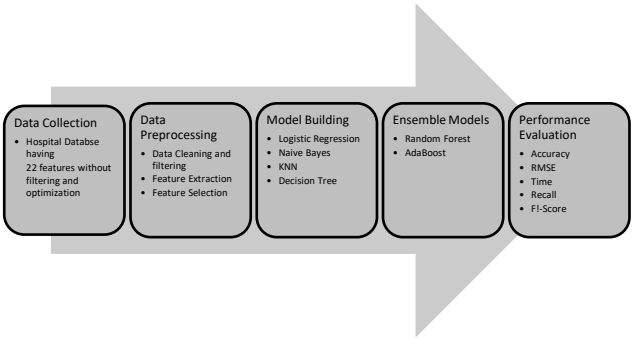


Figure 1. Process flow of Proposed Model

3.1. Data Collection

The dataset used in this study is collected from private medical clinic. The dataset of 500 patients was collected consisting of 23 attributes where 22 are independent variables and 1 is the target variable Class. Discrete and continuous both types of data used in present work which is depicted in Table 2.

Table 2. Feature Description of the Dataset

Feature Abbreviation	Description	Type	Permissible Range
Age	Age	Numeric	<40
BP	Blood Pressure	Numeric	<120/80
SUG	Sugar	Numeric	99-140
AL	Albumin	Numeric	3.20-4.60
BU	Blood Urea	Numeric	17-43 mg/dL
SOD	Sodium	Numeric	136-146
CRET	Creatinine	Numeric	0.51-0.95
POT	Potassium	Numeric	3.50-5.10
HB	Hemoglobin	Numeric	12-15
RBC	Red Blood Cells	Nominal	Normal
PC	Pus Cells	Numeric	0-5
BA	Bacteria	Nominal	Not Present
BGR	Blood Glucose Random	Numeric	135–214 mg/DL
WC	White Blood Cells Count	Numeric	4.5-11.0×10^9/L
PCV	Packed Cell Volume	Numeric	35.11 + 5.14
RC	Red Blood Cell Count	Numeric	3.80-4.80
HTN	Hypertension	Nominal	NO
DM	Diabetes Mellitus	Numeric	<100mg/DL
ANE	Anemia	Nominal	NO
SP	Specific Gravity	Numeric	1.005 - 1.025
CAD	Coronary Artery Disease	Nominal	NO
APPET	Appetite	Nominal	Good
Class	Decision Class	Nominal	Y/N

3.2. Data Preprocessing & Filtering

As shown in Table 2, the dataset consists of continuous as well as discrete values. However, the dataset contains the missing and duplicate values which are major reason of performance degradation. Missing values are replaced by mean for continuous variables and by median for discrete variables. For accurate analysis, the text data is replaced by numbers (e.g., NA by 0, OK by 1, Low by 0 etc.). Moreover, the dataset also contains some unimportant features which also affect the accuracy as such features do not add any value to the prediction. The proposed approach handles such attributes by using feature extraction and feature selection (Mangla et al., 2020) as discussed in next section.

Feature Extraction

Feature extraction or feature elimination is a technique which reduces the model dimensions keeping only the most important features for further processing. This process not only reduces the number of features in a dataset, it can also reduce the redundant amount of information in data analysis while minimizing the information loss (S. Sharma et al., 2016). Hence, it helps to reduce the computational efforts and improves the model efficiency. In our approach, we have used the Principal Component Analysis (PCA) as a feature extraction tool. PCA works on finding out the correlation between features so that least correlated features can be used for the analysis process. This process helps to reduce high variance hereby reducing the overfitting (Goyal et al., 2023). Fig. 4 shows the heatmap of correlation between features.

Feature Selection

For more accurate results, our model performs feature selection following feature extraction. The feature selection selects from the set of reduced features the most promising and reasonable features and ignores the irrelevant ones. One of the most efficient feature selection techniques is Lasso regularization(Radovic et al., 2017). Lasso regularization technique works on absolute sum of coefficients. The cost function in Lasso regularization is defined as:

Cost function = Loss + λ + Σ ||w|| Here,
Loss = sum of squared errors
λ = penalty
w = slope of the curve

This method adds a penalty to different model parameters to avoid over-fitting. The coefficients of grossy attributes thus gradually reduce to zero, hence removing such features from the model and retaining only the most important features. Table 3 represents the selection features for model building.

In our dataset 11 features are eliminated during this process. Resulting dataset is shown in Table 3.

Table 3. Set of features after Feature Optimization

Feature Abbreviation	Description	Type	Permissible Range
Age	Age	Numeric	<40
BP	Blood Pressure	Numeric	<120/80
SUG	Sugar	Numeric	99-140
AL	Albumin	Numeric	3.20-4.60
BU	Blood Urea	Numeric	17-43 mg/dL
SOD	Sodium	Numeric	136-146
CRET	Creatinine	Numeric	0.51-0.95
POT	Potassium	Numeric	3.50-5.10
HB	Hemoglobin	Numeric	12-15
PCV	Packed Cell Volume	Numeric	35.11 + 5.14
RBC	Red Blood Cells	Nominal	Normal

Further in the Decision Tree modelling process, the features are filtered in advance by IG(information gain) and Ginni index (GI) parameters (Ren et al., 2016).

3.4. Model Building

Different machine learning classifiers and regression models have been implemented and analyzed for the problem under consideration as explained below (Qezelbash-Chamak et al., 2022).

Naive Bayes

Naive Bayes classifier is a supervised learning technique which is used in text classification of high dimensional dataset. It predicts or classifies the target variable on the basis of probability of individual features using following method:

$$Y = armax\ P(y) \prod_{i=1}^{N} P(x_i|Y),$$

where P(y) is the posterior probability and $P(x_i|Y)$ is the conditional probability.

KNN

The KNN classifiers classify the closest data points in the training set (Laaksonen & Oja, 1996). Our model uses Euclidean Distance based method to find the closest data points. Different values of K are applied to the training algorithm for CKD dataset as shown in Table 5. The results of KNN algorithm are highly sensitive to the value of K. A large value of K may reduce the accuracy, whereas smaller value may add more noise to the output.

Logistic regression

Logistic Regression is one of the most efficient supervised learning algorithms used in binary classification. Moreover, it is used in non-linear classification where data points are not linearly separable. It uses a sigmoid function and predicts the probability of output class accordingly. The value of sigmoid function varies gradually and therefore, the change in the output is also gradual until stopping criteria is met (Qezelbash-Chamak et al., 2022).

Decision Tree

Decision Tree is a top-down tree-based classification technique which calculates entropy and information gain to choose the best classifier. DT is an efficient technique for multi-class classification. We implemented the DT model on our dataset for prediction of classes based on the values - Normal, Mild, Moderate, Severe and End stage.

Ensemble Modeling

Ensemble modeling is a machine learning technique which utilizes the power of multiple ML models and combines them to achieve maximum accuracy (Nikhila, 2021; N. Sharma et al., 2021). Two commonly used ensemble modeling techniques are - Random Forest and AdaBoost which are used in present research.

Random Forest

Random Forest is an ensemble learning technique which shows the best accuracy among other classifiers. It uses bagging approach which considers random samples of data with replacement. These random samples are supplied to different decision trees of randomly selected feature subsets. The algorithm finally uses voting method for final prediction. For optimal selection of feature subset RF uses information gain and ginni index.

AdaBoost

AdaBoost is another ensemble learning technique which works on the weights of input parameters. While RF uses bagging to draw random samples of the dataset, AdaBoost uses boosting technique to assign weights to individual

parameters and then they are selected according to assigned weights. In RF, multiple classifiers (usually decision trees) run in parallel, but in boosting different classifiers run in sequential manner.

SVM

SVM is a widely used supervised learning algorithm used for classification and regression problems. SVM is known to be a "model-free" method in a way that it doesn't depend on the distribution of problem parameters. Each data point is represented as a n-dimensional vector. The SVM model separates them into n-1-dimensional separating hyperplane to two discriminant classes. It aims to maximize the distance between the hyperplane and data points on each side. Research demonstrates that SVM has similar or improved accuracy for disease classification in comparison to Logistic Regression.

Linear Regression

Linear regression is a supervised machine learning model which is used to predict the linear relationship between the independent variables and the single target variable. In other words, the target variable can be predicted based on the linear combination of input/independent variable.

3.4. Exploratory Data Analysis

Exploratory data analysis helps to have better understanding of the data variables and to gain more insights of the problem. EDA is a powerful tool to reach some conclusion and remove some biases. Therefore, the next logical step is Exploratory Data Analysis.

A Univariate Analysis is performed in this paper. For better understanding we have divided the data set into numeric and categorical variables. Attempt has been made to explore the distribution of our variables using the kde plot and box as shown in Fig. 2. Kde plot helps to visualize the distribution of variables, shape of distribution and the concentration of data. On the other hand, a Boxplot helps to illustrate the median, the Quantiles and Outliers.

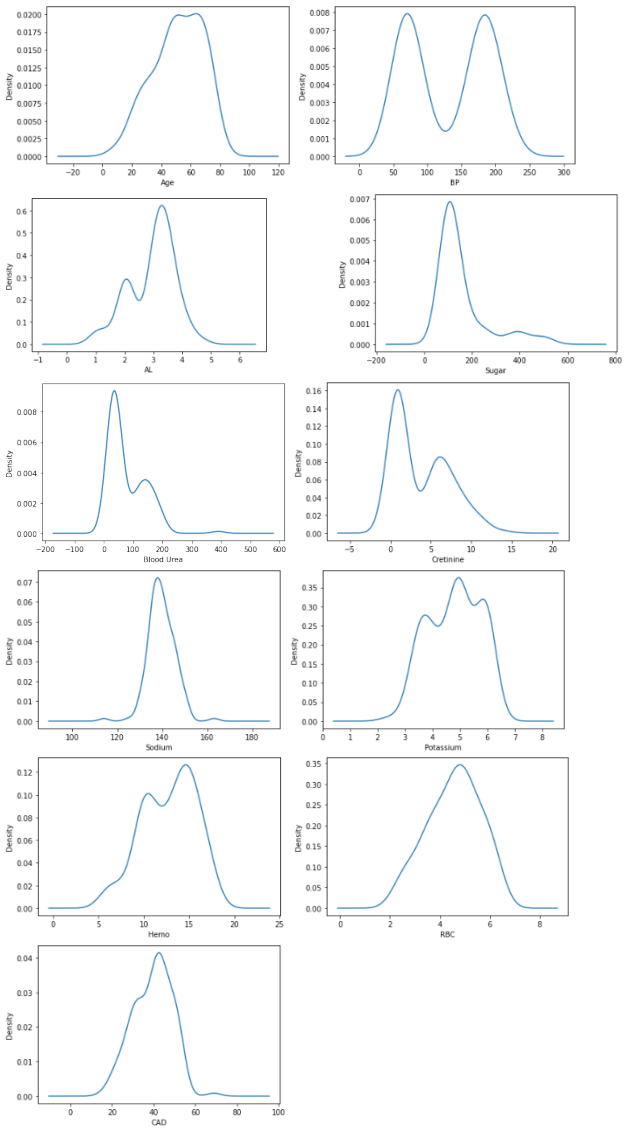


Figure 2. Kde plots of all 11 optimized variables.

Observations

(i) AGE - It is slightly right skewed with high concentration of data between 40-60. There are few outliers on the lower spectrum of age.
(ii) BP - as observed earlier also they are discrete variables. Tend to concentrate values on the lower end.
(iii) AL, CRET - Discrete variables. Little left skewed with a distribution close to normal.
(iv) SUGAR - Right skewed, Outliers. It has outliers but we should not see them as outliers as they are actually true values. Due to a skewed distribution, it is showing presence of outliers.
(v) SOD, RBC - Both are little left skewed with a distribution close to normal.
(vi) POT, HEMO - Few outliers.

Haemoglobin and potassium have a slight right skewness but overall, we can see a normal distribution.

Mostly all the variables have outliers ranging from few to many.

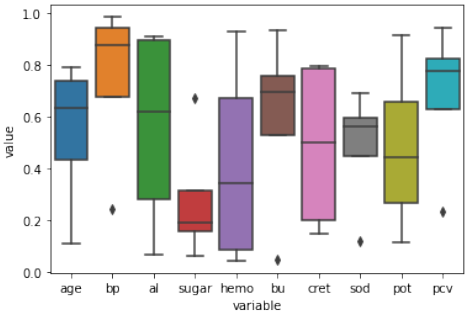


Figure 3. Boxplot of Feature Set.

The normal range for creatinine in the blood may be 0.84 to 1.21 milligrams per deciliter (Mayo Clinic). Higher than that it indicates kidney malfunction. So, it can be a big indicator for classification problem. We could confirm it in a better way by visualizing the relation between two latter.

Also, the very far off value for Cret certainly point to very abnormal Outliers which needs domain knowledge to be dealt properly. We can see abnormal values of 0 in sodium values which should not be a case.

4. Implementation and Results

This section explains the results obtained after simulation process. The dataset is collected from private hospital of around 500 patients. Approximately 200 values were found to be inconsistent which were either removed or corrected during data preprocessing. The resulting dataset consists of data of 458 patients. After data filtering, next step is feature optimization. During this phase 11 features were selected for the training phase. However, for the comparison the simulation is performed first for entire dataset and then for selected features.

Image in Fig. 4 represents the correlation among selected features which in effect varifies the effectiveness of the feature optimization process.

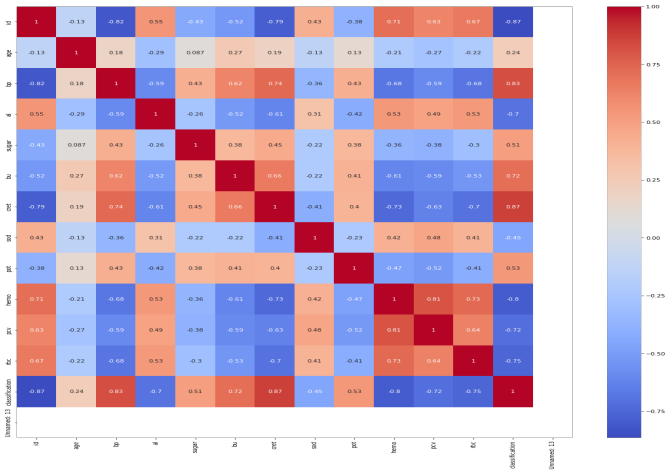


Figure 4. Correlation between model parameters

The parameters used for model performance evaluation are - Accuracy, Sensitivity, Specificity, Time (in ms), RMSE and F1-score.

Table 4 and 5 presents the simulation results of various performance measures. Naive Bayes, Random Forest and Adaboost, have shown the maximum efficiency in terms of accuracy and estimated error. However, the time taken by RF and Adaboost is more as compared to Naive Bayes for binary classification. This might increase for a larger dataset. Hence, for binary classification Naive Bayes gives the best results.

Table 4. Effectiveness of Different Classifiers after Feature Optimization

Classifier	Sensitivity	Recall	F1 Score
Naive Bayes	1.0	0.98	0.99
KNN	1.0	0.96	0.92
Logistic Regression	1.0	1.0	1.0
Decision Tree	1.0	1.0	1.0
Random Forest	1.0	0.98	0.99
AdaBoost	1.0	1.0	1.0
SVM	1.0	0.96	0.98
SVM(Poly)	0.96	1.0	0.90
SVM(rbf)	0.90	0.83	0.89
Linear Reg	0.94	0.80	0.89

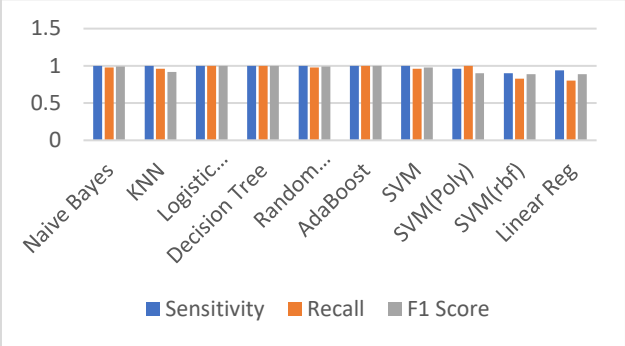
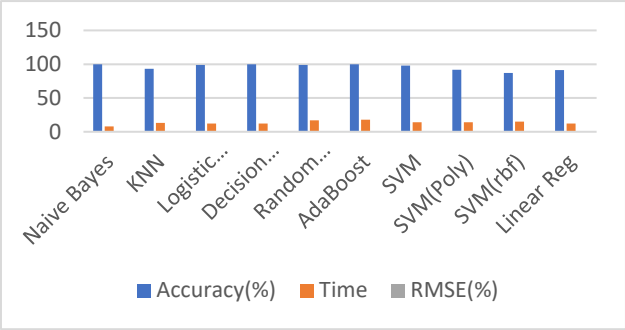


Table 5. Performance/Efficiency of Different Classifiers after Feature Optimization

Classifier	Accuracy (%)	Time	RMSE(%)
Naive Bayes	100	8	0
KNN	93	13	0
Logistic Regression	99	12	0.02
Decision Tree	100	12	0
Random Forest	99	17	0.02
AdaBoost	100	18	0
SVM	98	14	0.04
SVM(Poly)	92	14	0.08
SVM(rbf)	87	15	0.24
Linear Reg	91.3	12	0.14



In case of Chronic Kidney Disease, early diagnosis of mortality is very important for those patients who are at a greater risk so that preventive measures can be taken which mught help in reducing the mortality. Research has been made to develop effective ML models for such cases. Present work focuses on developing variety of machine learning models consisting of linear model, KNN, NB, DT, RF, SVM and LR and selected the optimal one with the best performance in terms of computational cost and accuracy. Moreover In the current study, emphasis is given to feature importance and feature extraction which is missing in most of the literature surveyed. Also significant efforts have been made for data preprocessing to achieve better results. Undoubtedly, none of the previous models emphasised feature extraction and optimization as a crucial tool in predicting the mortality risk for CKD patients (Almasoud, M., & Ward, T. E. (2019)). Our investigation reveals that Blood Urea and Creatinine are major factors affecting the mortality in CKD. The elevated blood urea and Creatinine level increase the risk of fatal cases (Elhoseny, M., Shankar, K., & Uthayakumar, J. (2019). Another key feature affecting the mortality prediction is the age of the patient, which is found to be missing in most of the previous experiments. Present research demonstrates and validates considerable performance for CKD early diagnosis while taking into account all the crucial features as depicted in Table 4-5.

In the previous studies (disussed in section 2 above), authors did not focus upon feature optimization and exploratory data analysis for improving overall model performance. Therfore, the accuracy of the models proposed in literature is comparatively low. Our main focus is on feature optimization and EDA which have proven to be very useful tools in improving overall models performance and decreasing computaiotnal cost. Further, in order to validate the findings achieved by the proposed method, authors perform a comparative analysis with finding obtained by researchers in (Farman et. al.). This comparative analysis is illustrated in Table 6 which clearly indicates that the proposed method yields better results in terms of accuracy:

Table 6. Comparative Analysis of Proposed Approach with Existing One

Classifier	Accuracy % [1]	Accuracy %
Random Forest	98	100
Naïve Bayes	93	89.2
Decision Tree	96	97.89
AdaBoost	99	97.84

5. Conclusion and Future Scope

Prediction of chronic kidney disease at early phases is very crucial and plays a great role in the medical treatment. Hence, an attempt has been made in this work to design an accurate machine learning model which predicts the disease in advance with given database of around 500 patients. Out of 22 features present in the database, 12 were extracted using efficient feature optimization techniques. Ten machine learning classifiers were implemented and tested for the available dataset. The simulation was done on features before and after optimization. The results prove that for reduced features and for binary class classification,

Naive Bayes classifier gives best performance in terms of accuracy and computational cost. Naive Bayes, Random Forest and AdaBoost models give 100% accuracy for given dataset. But the computational time for Naive Bayes in lowest among all. Hence, it can be concluded that for binary class classification, Naive Bayes algorithms are best. Other algorithms are good for multi-class classification but for binary class, they are little expensive than Naive Bayes.

The dataset considered for our study is quite small collected from a single medical center. Work can be extended on huge dataset of multiple hospitals. Moreover, multi-class classification can also be another area of work where in spite of CKD and NCKD prediction, stages of CKD can be predicted.

References

[1] Almasoud, M., & Ward, T. E. Almasoud, M., & Ward, T. E. Detection of Chronic Kidney Disease Using Machine Learning Algorithms with Least Number of Predictors. 2014. Int. Jr. of Adv. Comp. Sci. and Applications (IJACSA), 10(8). http://dx.doi.org/10.14569/0100813

[2] Baidya, D., Umaima, U., Islam, M. N., Shamrat, F. M. J. M., Pramanik, A., & Rah-man, M. S. A Deep Prediction of Chronic Kidney Disease by Em-ploying Machine Learning Method. 2022. In: 6th International Conference on Trends in Electronics and Informatics, ICOEI 2022 - Proceedings, 1305–1310. https://doi.org/10.1109/9776876

[3] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Le-onowicz, Z., Jasinski, M., Jasinski, L., Gono, R., Jasinska, E., & Bolshev, V. Prediction of Chronic Kidney Disease - A Machine Learning Per-spective. 2021. IEEE Access, 9, 17312–17334. https://doi.org/10.1109/3053763

[4] Elhoseny, M., Shankar, K., & Uthayakumar, J. Intelligent Diagnostic Pre-diction and Classification System for Chronic Kidney Disease. 2019. Scientific Reports. 9(1), 1–14. https://doi.org/10.1038/s41598-019-46074-2

[5] Goyal, S., Batra, N., & Chhabra, K. Diabetes Disease Diagnosis Using Machine Learning Approach. 2023. Lecture Notes in Networks and Systems, 47(3), 229–237. https://doi.org/10.1007/978-981-19-2821-5_19

[6] Laaksonen, J., & Oja, E.. Classification with learning k-nearest neighbors. 1996. In: IEEE International Conference on Neural Networks - Conference Proceedings, 3, 1480–1483. https://doi.org/10.1109/ICNN.1996.549118

[7] Mangla, M., Akhare, R., Deokar, S., & Mehta, V.. Employing Machine Learning for Multi-perspective Emotional Health Analysis. 2020. Emotion and In-formation Processing, 199–211. https://doi.org/10.1007/978-3-030-48849-9_13

[8] Manonmani, M., & Balakrishnan, S. Feature Selection Using Improved Teaching Learning Based Algorithm on Chronic Kidney Disease Dataset. 2020. Procedia Computer Science, 171, 1660–1669. https://doi.org/10.1016/ 04.178

[9] Nikhila. Chronic kidney disease prediction using machine learning ensemble algorithm. In: Proceedings of IEEE 2021 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2021, 476–480. https://doi.org/10.1109/ 9397144

[10] Qezelbash-Chamak, J., Badamchizadeh, S., Eshghi, K., & Asadi, Y. A survey of machine learning in kidney disease diagnosis. Machine Learning with Applications. 2022. 10, 100418. https://doi.org/10.1016/j.mlwa.2022.100418

[11] Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. Minimum re-dundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics. (2017). 18(1), 1–14. https://doi.org/10.1186/S12859-016-1423-9/FIGURES/6

[12] Ren, Y., Zhang, L., & Suganthan, P. N. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions. 2016. IEEE Computational Intelligence Magazine, 11(1), 41–53. https://doi.org/10.1109/MCI.2015.2471235

[13] Sharma, N., Dev, J., Mangla, M., Wadhwa, V. M., Mohanty, S. N., & Kakkar, D. A Heterogeneous Ensemble Forecasting Model for Disease Prediction. 2021. New Generation Computing, 39(3–4), 701–715. https://doi.org/10.1007/S00354-020-00119-7/TABLES/6

[14] Ali, Farman, Khalid, Hira, Zhaid K., Muhhamad, Mehmood, Gulzar & Shuaib Quershi, Muhammad. Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease. 2023. Computational Intelligence and Neuroscience, Hindawi. https://doi.org/10.1155/2023/9266889.