Undergraduate Conference on Intelligent Computing and Systems (UCICS 2025)
26-27 February, 2025; Varendra University, Rajshahi, Bangladesh

Comparative Analysis of Machine Learning Algorithms for Chronic Kidney Disease Prediction

M. Morsedur Rahman
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi 6204, Bangladesh
morsed.ruet.ece18@gmail.com

Sagor Chandro Bakchy
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi 6204, Bangladesh
sagorchandro.10@gmail.com

Nafia Islam Shishir
Computer Science & Engineering
Varendra University,
Rajshahi 6204
Bangladesh
nafia@vu.edu.bd

Md. Sajidur Rahman
Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Rajshahi 6204, Bangladesh
mdsajidurrahman375@gmail.com

***Abstract*— In recent days, chronic kidney disease (CKD) has been recognized as one of the most significant health problems globally. The defining feature of CKD is a progressive deterioration in renal function over time. Since kidney damage develops slowly over a long period of time, early detection and appropriate treatment may be able to save the lives of many. Machine learning classifier algorithms have emerged as a reliable tool to identify the disease at its early stages, providing a means to intervene and manage it sooner than other methods. In this paper, the performance of 10 models is evaluated on the dataset of CKD collected from the UCI ML repository for the classification of CKD. The training data in this study was augmented by applying the SMOTE technique and Gaussian noise. In case of missing values handling, for numerical and categorical variables KNN imputation and mode imputation for features were utilized respectively. Combining the filter, wrapper and embedded feature selection strategies led to the identification of the most important 13 features. Extra Tree Classifier, XGBoost, Gradient Boosting and Random Forest performed better than other algorithms with an accuracy of 99.17%. When compared to the other nine methods, Extra Tree Classifier performed extremely well in case of precision, recall and F1 score. For this proposed approach, the error rate and training time were all comparatively low at 0.0083, and 0.0787 seconds respectively. This paper illustrates the performance comparison of ten different machine learning (ML) algorithms and the importance of feature selection for predicting CKD.**

Keywords*— *Chronic Kidney Disease, XGBoost, Gradient Boost, Extra tree, mode imputation, embedded feature selection.

I. Introduction

Chronic kidney disease (CKD) is a long-term medical disorder that causes damage and malfunction of the kidney. The important organ kidney generates urine and eliminates waste from circulation. A number of substances in the blood are regulated by the kidneys - aiding in the production of red blood cells by hormones and producing an active form of vitamin D that the body can use. All bodily functions are hindered when the kidneys suffer damage, leading to a deadly illness. With a rapidly increasing number of patients, CKD is presently the leading cause of death, accounting for 1.7 million deaths per year [1]. Most people with CKD experience tiredness, a slower rate of metabolism, acute pain, and swollen ankles and legs. The course of this chronic illness cannot be stopped or slowed down until the patient's sole options are dialysis or surgery. In Bangladesh, 35,000-40,000 new cases of renal failure each year. More than 20,000 people with CKD die away each year because dialysis or kidney transplants are not affordable to them [2].

The fact that there is no specific remedy for an incurable condition makes diagnosing and treating it extremely difficult [3]. Only observation of patients, interviews and other conventional methods are not that much reliable. Because of their heavy reliance on several biological parameters, conventional approaches for determining the presence of chronic renal disease are not always trustworthy [4]. Computerized diagnostics is required to assist the diagnostic judgments of doctors and radiologists due to the rising number of patients with chronic kidney disease, the scarcity of specialists for detection and treatment, and the high costs, especially in developing countries [5]. Machine learning has demonstrated high efficiency in delivering solutions for preliminary diagnosis in a range of medical fields by utilizing computational methodologies.

This paper makes a remarkable contribution by finding the relevant features through preprocessing of the raw dataset and applying machine learning techniques for the prediction of chronic kidney disease. The main contributions of this thesis are as follows:

1) Performing data augmentation using SMOTE to balance the dataset and enhance training.

2) Applying advanced techniques of feature selection combining filter, wrapper, and embedded methods.

3) Optimizing hyperparameters to improve accuracy, reduce error rates and minimize training time.

The literature review of the recent works is covered in section II. Section III describes the detailed methodology whereas section IV contains results and discussions. Lastly, conclusion about the whole work is represented in section V.

II. Literature Review

Machine learning algorithms have made the lives of people easier in all sectors including the medical field. These days, the use of ML models makes the identification of CKD more accurate and straightforward. Many researchers explored the vast area of different machine learning (ML) algorithms in the prediction of CKD.

The conceptual IoT framework for the deep learning-based early diagnosis of chronic kidney disease (CKD) is presented by Ali et al. [6]. Employing the Anova-F feature selection method, they constructed a multi-layer perceptron (MLP) classifier. With a 99% accuracy rate and cost-effective calculation time, the suggested solution performs better than competing machine learning and deep learning techniques. But the proposed framework is complex and high technical support is required to implement it in real life. Again, more validation of diverse data is required for generalizability.

Extreme gradient boost, or XGB, is the most accurate of six explainable machine learning methods that Dharmarathne et al. [7] devised to overcome the "black box" character of typical machine learning predictions. Its accuracy is 97.5%. Partial Dependency Plots (PDP) and Shapley Additive Explanations (SHAP) were used in the study to explain the reasoning behind the forecasts. They also introduced a graphical interface for easy detection. The study's limitations are using a small dataset directly, less attention to medical history and morbidity profiles and more training time.

Gogoi et al. [8] proposed a method to detect early CKD that is both effective and secure. They used two feature selection algorithms named genetic algorithm and bat algorithm which can solve complex optimization problems. The highest accuracy of 98.75% is achieved by logistic regression model with the genetic algorithm. They aimed to increase the confidentiality of the information. But the model is less reliable because it worked on a small dataset, and augmentation was not done.

Pal et al. [9] used different ML methods for CKD prediction using categorical attributes, non-categorical attributes and lastly combining both. The combined attributes achieved the best classification accuracy on Random Forest classifier with an accuracy of 94%. They did not use the specific data preprocessing, or feature selection methods like firefly optimization, chi-square for the UCI dataset.

Prior to using principle component analysis (PCA), a sophisticated feature engineering technique, Islam et al. [10] first used the original dataset for CKD prediction. By the end of the study, the best subset of criteria to detect CKD was reduced to 30% from the original list of 25 variables. XGBoost outperformed the other 12 algorithms used in this research with an accuracy of 98.3%. But they did use the small dataset without any augmentation. If more data or large dataset could be used, the generalization of the method would increase.

A hybrid strategy combining KNN, tensor factorization, Adaptive Neuro-Fuzzy Inference System for missing values handling, and a new Adaptive Weight Dynamic Butterfly Optimization Algorithm for feature selection for early CKD diagnosis was proposed by Saroja and colleagues. The new NWCNN (Novel Weight Convolution Neural Network) classifier performs better than other ML and DL models with an accuracy of 99.04%. But they did not show a proper training time evaluation in their work [11].

III. Methodology

The approach starts with some preprocessing and moves on to the classification techniques. Preprocessing of the raw data obtained from the repository improves the accuracy and suitability of the outcome for medical applications. The detailed method is shown in Fig. 1.

A. Dataset Collection

The CKD dataset used for this research conduction is collected from an open-source UCI Machine Learning Repository [12]. Among 400 samples, the overall number of CKD data is 250, whereas the total number of non-CKD data is 150. All of them contain 24 features.

B. Data Preprocessing

Before building a model, the dataset must be preprocessed to remove undesirable noise and anomalies. After the data collection, it must be cleaned and prepared for the model construction. The preprocessing starts from data cleaning to feature selection, augmentation and handling of missing values.

C. Handling Missing Values

The dataset's missing values may have an effect on machine learning models' performance. The k-nearest Neighbor (kNN) mechanism was utilized to handle the missing values of numerical attributes taking into account a predetermined number of nearest neighbors (in this case 5), and mode imputation for categorical attributes. It replaces missing values with the most frequent (mode) category, which is easy to compute and implement

D. Feature Encoding and Scaling

For every category, one-hot encoding generates a binary column that indicates whether the category is present or not. This is an effective way to handle non-ordinal categorical features. ANN, KNN, and SVM are a few machine learning

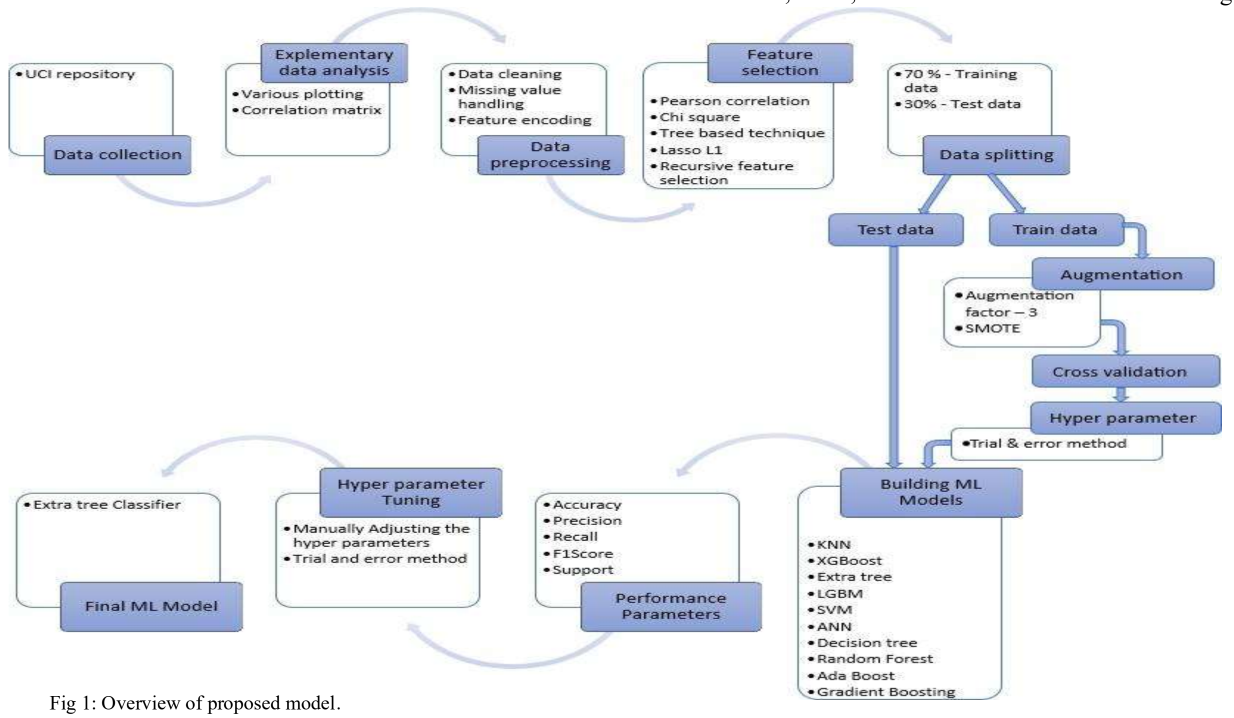

Fig 1: Overview of proposed model.

techniques that are impacted by feature scaling. In this study, the Min-Max scaling method was used to rescale features to a specific range, normally between 0 and 1.

E. Feature Selection

To select the best suitable features, three different approaches such as the filter method, the wrapper approach and embedded approach were introduced. Pearson correlation and Chi-Square techniques were employed in the filter approach. The Recursive Feature Elimination methodology has been used for the wrapper method. Two different kinds of techniques have been used in the embedded method. There are two types of methods: tree-based techniques and Lasso L1.

The top 13 features which are supported by most of the feature selection techniques are selected for the further process of this research. Selected features are 'hemoglobin', 'albumin', 'hypertension', 'red blood cell count', 'blood glucose random', 'packed cell volume', 'sodium', 'specific gravity', 'Peda edema', 'appetite', 'diabetes mellitus', 'blood urea', 'serum creatinine'. These features provide the highest accuracy of the models.

F. Data Splitting and Augmentation

The amount of data reserved for training is higher when a 70:30 split ratio is used. To provide consistent data shuffles and splits, the random state parameter is used, making it possible to compare different models or iterations with confidence. The data augmentation factor was initially set to 3, meaning each sample in the training set would generate 3 additional synthetic samples. So, a total of 280 + (280×3) = 1120 samples were generated. SMOTE further augments the dataset by generating synthetic samples to balance the classes. After SMOTE, the total number of samples becomes 1392, ensuring a balanced representation of both classes for effective model training and evaluation.

G. Machine Learning Algorithms

10 ML algorithms are used for predicting chronic kidney disease such as K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Ada Boost, Gradient Boosting, XGBoost, Extra Trees, Light Gradient-Boosting Machine (LGBM), Support Vector Machine (SVM), and Artificial Neural Networks (ANN). Every algorithm has its own distinct qualities and performance trade-offs.

1) KNN: In order to predict the class of a new data point, KNN computes its distance from each other point in the training set. It then selects the nearest k neighbors depending on this distance measurement. The class is decided by a majority voting among the k closest neighbors of the new data point.

2) Random Forest: The first stage entails creating a parameter grid with various hyperparameter combinations. Next, this parameter grid is extensively searched by using GridSearchCV, which applies 5-fold cross-validation.

3) SVM: Support vector machines (SVMs) are a type of supervised machine learning technique to identify the optimal hyperplane in an N-dimensional space for optimizing the distance between every class in order to classify data.

4) Extra Trees: It is essentially the same as a Random Forest Classifier; the principle of which the forest's decision trees are constructed is the only thing that makes it different. Each Decision Tree in the Extra Trees is based on the training sample. From the feature set at each test node, the decision trees are given a random sample of k features, and they are needed to select the optimal features based on a set of mathematical conditions. This approach of random feature selection results in a number of de-correlated decision trees.

5) ANN: It functions similarly to the human brain. The network processes information in a forward pass, with each neuron receiving inputs, adding up their weights, and using an activation function to produce an output.

The hyperparameters were selected on a continuous trial-and-error basis for all the models. Table I shows the hyperparameter selection.

TABLE I. HYPERPARAMETERS OF THE ALGORITHMS

Algorithm	**Hyperparameters Tuning**
KNN	algorithm: 'auto', n_neighbors: 3, weights: 'distance'
Decision Tree	criterion: 'gini', max_depth: 10, min_samples_leaf: 2, min_samples_split: 5
Random Forest	max_depth: None, max_features: 'log2', min_samples_leaf: 1, 'min_samples_split': 2, n_estimators: 20
AdaBoost	algorithm: 'SAMME.R', learning_rate: 1.0, n_estimators: 15
Gradient Boosting	learning_rate: 1.0, max_depth: 5, min_samples_split: 6, n_estimators: 15
XGBoost	gamma: 0, learning_rate: 1.0, max_depth: 5, min_child_weight: 1, n_estimators: 50
Extra Tree	max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100
LGBM	colsample_bytree: 0.6, learning_rate: 0.2, max_depth: 7, min_child_samples: 5, n_estimators: 25, subsample: 0.6
SVM	C: 10, gamma: 'auto', kernel: 'rbf'
ANN	activation: 'relu', alpha: 0.001, hidden_layer_sizes: (15), learning_rate: 'constant', max_iter: 30, solver: 'sgd'

IV. RESULT AND DISCUSSION

To find out which one provides the best accuracy, 10 ML models are evaluated on the dataset. Every classifier's output has been examined using a range of evaluation criteria, and the 10-fold cross-validation technique has been applied to verify the output's accuracy against overfitting. Table II below shows the performance metrics. Though some classifier shows the highest accuracy of 99.17%. Extra Tree outperformed others with less training time and an error rate.

TABLE II. PERFORMANCES OF DIFFERENT MODELS

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**CV**	**Training Time (s)**
KNN	0.9167	0.92	0.92	0.92	0.69	1.238
Decision Tree	0.9667	0.98	0.97	0.97	0.95	0.010
Random Forest	0.9917	0.99	0.99	0.99	0.98	0.086
AdaBoost	0.9667	0.97	0.97	0.97	0.97	0.084
Gradient Boosting	0.9917	0.99	0.99	0.99	0.98	0.163
XGBoost	0.9917	0.99	0.99	0.99	0.97	0.085
Extra Tree	0.9917	0.99	0.99	0.99	0.99	0.078
LGBM	0.9833	0.98	0.98	0.98	0.97	0.040
SVM	0.8000	0.82	0.80	0.78	0.62	16.49
ANN	0.9750	0.98	0.97	0.98	0.97	0.002

Table III shows that feature selection and data augmentation are essential techniques in machine learning that improve model performance by reducing overfitting and enhancing robustness to uncertain input. Tree-based models are more interpretable, simpler to understand, flexible with both numerical and categorical features, and robust to noise and outliers. In contrast, boosting models are less interpretable (black box), have complex architectures, limited flexibility with data types, and are sensitive to noise.

TABLE III. Comparison of Model Accuracy Before and After Feature Selection and Data Augmentation

Model	Without FS & DA	With FS & DA
KNN	0.7300	0.9167
Decision Tree	0.9580	0.9667
Random Forest	0.9670	0.9917
AdaBoost	0.9500	0.9667
Gradient Boosting	0.9580	0.9917
XGBoost	0.9500	0.9917
Extra Tree Classifier	0.9670	0.9917
LGBM	0.9580	0.9833
SVM	0.6500	0.8000
ANN	0.9500	0.9750

Fig. 2 shows that optimization of model parameters helps to reduce the error rate and improve the accuracy as well. Training time and low error rate are a crucial part of using the models in real life. In this case, our model shows a good performance reducing the error rate and training time.

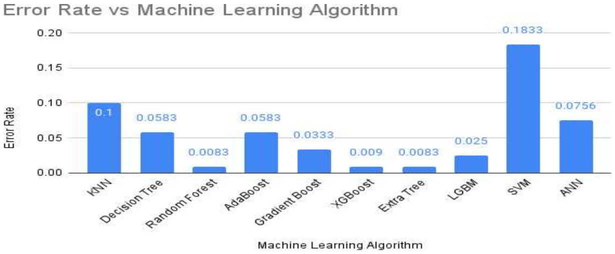

Fig. 2. The error rate for all the ML models.

The confusion matrix and ROC curve are also added for a better interpretation of the result.

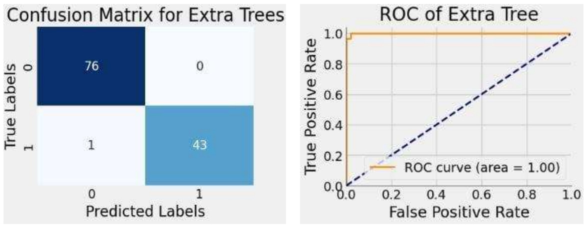

Fig. 3. The confusion matrix and ROC curve of Extra tree model.

Various authors used various models for CKD detection. Table IV represents the comparison of the previous work with the proposed method of extra tree classifier.

TABLE IV. Comparison with Previous Work

Reference	Year	Approaches	Accuracy
Pal et al. [9]	2023	Random Forest	94%
Islam et al. [10]	2023	XGBoost	98.3%
Saroja el al. [11]	2023	NWCNN	99.04%
Ali et al. [7]	2024	Anova-F + MLP	99%
Dharmarathne et al. [6]	2024	XGB	97.5%
Gogoi et al. [8]	2024	LR	98.75%
Proposed Model		**Extra Tree Classifier**	**99.17%**

Our proposed model performed better than others which worked on the same dataset.

V. Conclusion and Future Work

Over time, chronic kidney disease has created serious health issues for people around the world. To stop its pandemics, several preventive measures as well as a cure are required. For this, the early detection is necessary. In our proposed work, we have achieved great accuracy in CKD detection with an extra tree classifier and relevant feature extraction and selection methods. In this study, we used augmented data to train the model more efficiently resulting in better classification accuracy while other authors used the dataset directly. We also achieved a low error rate and less training time. In the future, we will try to use an authentic clinical dataset of any clinic. Then the impact of using deep learning models can be examined through analysis which may improve the result.

References

[1] H. Khalid, A. Khan and M. Z. Khan, "Machine learning hybrid model for the prediction of chronic kidney disease.," *Computational Intelligence and Neuroscience,* p. 9266889, 2023.

[2] M. R. Alam, "Kidney Disease – Bangladesh Perspective," *Bangladesh Journal of Medicine,* vol. 34, no. 2, pp. 179-180, 2023.

[3] M. Gollapalli, B. Saad and J. Alabdulk, "Detection of Chronic Kidney Disease Using Machine Learning Approach," in *14th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2022.

[4] D. Swain, U. Mehta and A. Bhatt, "A robust chronic kidney disease classifier using machine learning," *Electronics,* vol. 12, no. 1, p. 212, 2023.

[5] E. M. Senan, M. H. Al-Adhaileh and F. W. Alsaade, "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," *ournal of healthcare engineering,* vol. 2021, no. 1, p. 1004767, 2021.

[6] M. M. Ali, M. S. Islam, M. N. Uddin and M. A. Uddin, "A conceptual IoT framework based on Anova-F feature selection for chronic kidney disease detection using deep learning approach," *Intelligence-Based Medicine,* vol. 10, p. 100170, 2024.

[7] G. Dharmarathne, M. Bogahawaththa, M. McAfee and U. Rathnayake, "On the diagnosis of chronic kidney disease using a machine learning-based interface with explainable artificial intelligence.," *Intelligent Systems with Applications,* p. 200397, 2024.

[8] P. Gogoi and . J. . A. Valan, "Privacy-preserving predictive modeling for early detection of chronic kidney disease.," *Network Modeling Analysis in Health Informatics and Bioinformatics,* vol. 13, no. 1, p. 16, 2024.

[9] S. Pal, "Prediction for chronic kidney disease by categorical and non_categorical attributes using different machine learning algorithms," *Multimedia Tools and Applications,* vol. 82, no. 26, pp. 41253-41266, 2023.

[10] M. A. Islam, M. Z. H. Majumde and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," *Journal of pathology informatics,* vol. 14, p. 100189, 2023.

[11] T. Saroja and Y. Kalpana, "Hybrid missing data imputation and novel weight convolution neural network classifier for chronic kidney disease diagnosis," *Measurement: Sensors,* vol. 27, p. 100715, 2023.

[12] L. Rubini, P. Soundarapandian and P. Eswaran, "Chronic Kidney Disease," UC Irvine Machine Learning Repository, 7 February 2015. [Online].Available:https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease.%20[Accessed%2015%209%202024].. [Accessed 9 January 2025].