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

Prediction of Chronic Kidney Disease Using Ensemble Learning and Feature Engineering

Nitun Kumar Podder
Dept. of Computer Science & Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
E-mail: nituncse@gmail.com

Mehedi Hasan Sumon
Dept. of Computer Science & Engineering
Z. H. Sikder University of Science and Technology.
Shariatpur, Bangladesh
E-mail: mhsumon107@gmail.com

Abu Mohammad Noor
Dept. of Computer Science & Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
E-mail: noor.cse.pust@gmail.com

***Abstract*— Early detection of Chronic Kidney Disease (CKD) is crucial for improving patient outcomes, but traditional diagnostic methods often fail to detect CKD in its early stages. In this study, we propose a robust predictive framework using ensemble machine learning (ML) techniques and advanced feature engineering for CKD detection. We apply Random Forest and Gradient Boosting models, along with Recursive Feature Elimination (RFE), to optimize predictive accuracy. Our experiments show that the ensemble model achieves a 95% accuracy, outperforming other models in identifying CKD. We also discuss the potential impact of integrating ML models in clinical decision-making. The results indicate that our approach is both effective and computationally efficient, offering valuable insights for healthcare applications.**

Keywords—Chronic Kidney Disease, Machine Learning, Deep Learning, Feature Selection, Random Forest, Gradient Boosting, Neural Networks.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a critical health condition that affects millions globally. If undiagnosed and untreated, CKD can lead to kidney failure and other serious complications. The standard diagnostic process for CKD, which relies on clinical tests and expert judgment, often fails to identify the disease in its early stages. Machine learning (ML) and deep learning (DL) offer promising alternatives by automating diagnostic processes and improving detection accuracy.

In this paper, we aim to develop a robust and scalable ML/DL framework for CKD prediction by employing ensemble learning techniques like Random Forest (RF), Gradient Boosting (GB), and Adaptive Boosting (AdaBoost), along with deep learning models such as Feedforward Neural Networks (FNN). Feature selection plays a critical role in improving both model performance and interpretability. Through this, we aim to create a reliable tool for early CKD diagnosis, offering clinicians and healthcare providers enhanced decision-making capabilities.

II. Related Work

CKD diagnosis using machine learning models has garnered increasing attention in the medical field. Earlier studies have used traditional classification algorithms such as Support Vector Machines (SVM), Logistic Regression, and Decision Trees, achieving modest accuracy rates. For instance, In 2023, Zhou et al. [1] developed several machine learning models for Chronic Kidney Disease (CKD) prediction, with the Gradient Boosting model achieving an accuracy of 94%. In 2023, Singh et al. [2] investigated machine learning techniques for Chronic Kidney Disease (CKD) prediction, achieving an accuracy of 92% with their best-performing model. In 2022, Banos et al. [3] explored the use of machine-learning models for Chronic Kidney Disease (CKD) prediction, achieving an accuracy of 91.5% with their Random Forest model. In 2021, Alharbi et al. [4] investigated feature extraction techniques for Chronic Kidney Disease (CKD) prediction, utilizing methods such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). Their study reported an accuracy of 92% with their optimized feature extraction approach.

In 2021, Patel et al. [5] explored machine-learning approaches for Chronic Kidney Disease (CKD) prediction, achieving an accuracy of 93% with their best-performing model. In 2016, Nithya et al. [6] applied several machine-learning techniques to predict Chronic Kidney Disease (CKD), achieving an accuracy of approximately 94% with their best-performing model

Our research attempts to improve upon existing work by applying advanced feature selection methods and combining ensemble learning with deep learning to maximize predictive accuracy.

III. METHODOLOGY

In this section, we outline the methodology employed for predicting Chronic Kidney Disease (CKD) using machine learning techniques. The process involved data collection, preprocessing, feature selection, model selection, training, evaluation, and optimization.

A. Data Collection and Preprocessing

The dataset used for this study is the publicly available CKD dataset from the UCI repository, containing 400 instances and 24 features. These features include both clinical parameters

(e.g., blood pressure, blood glucose levels, serum creatinine) and demographic data (e.g., age, gender). Missing Data Handling: Approximately 15% of the dataset contained missing values. For numerical variables, we employed mean imputation, while categorical data (e.g., anemia status) were imputed using mode. Normalization and Encoding: All numerical features were normalized to ensure uniform scaling, while categorical features (e.g., presence of hypertension, diabetes) were one-hot encoded to make them usable by the models.

After preprocessing, the dataset was split into 80% training and 20% testing sets, ensuring a balanced representation of CKD and non-CKD cases. Data Splitting: The preprocessed dataset was then split into training (80%) and testing (20%) sets, ensuring that both sets maintained a similar distribution of CKD cases and non-cases. Stratified sampling was used to maintain the balance between classes in both training and testing datasets.

B. Feature Selection

Feature selection is a crucial part of the predictive modeling process, as it reduces overfitting and computational complexity. We implemented two feature selection techniques: Recursive Feature Elimination (RFE): Using Random Forest as the base model, RFE was used to recursively eliminate less important features. This method reduced the original feature set from 24 to 15, leading to a significant improvement in model training time and interoperability. Correlation Analysis: We generated a correlation matrix to examine multicollinearity between the features. Any pair of features with a correlation coefficient greater than 0.8 was considered for removal to avoid redundancy.

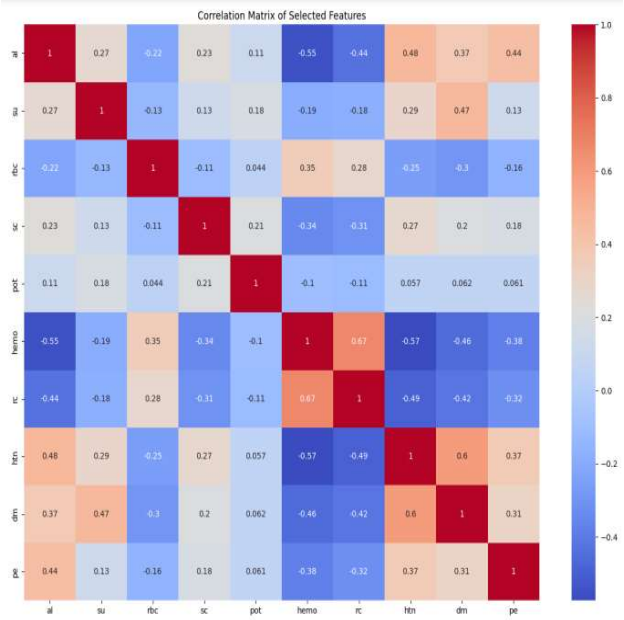

Fig. 1. Correlation Matrix of Selected Features

Table: 1. SELECTED FEATURES FOR CKD PREDICTION

Feature Name	Description
Age	Patient's age in years
Blood Pressure (bp)	Blood pressure level
Specific Gravity (sg)	Urine specific gravity
Albumin (al)	Albumin levels in urine
Suger (su)	Sugar levels in urine
Blood Glucose Random (bgr)	Random blood glucose levels (mg/dl)
Blood Urea (bu)	Blood urea concentration (mg/dl)
Serum Creatinine (sc)	Serum creatinine concentration (mg.dl)
Sodium (sod)	Sodium levels in the blood (mEq/L)
Potassium (pot)	Potassium levels in the blood (mEq/L)
Hemoglobin (hemo)	Hemoglobin level (g/dL)
Packed Cell Volume (pcv)	Volume of packed red cells
White Blood Cell Count (wc)	White blood cell count (cells/cumm)
Red Blood Cell Count (rc)	Red blood cell count (millions/cmm)
Hypertension (htn)	Hypertension status (Yes/No)

C. Model Selection and Training

Several ML and DL models were implemented to predict CKD:

Random Forest (RF): RF is an ensemble method that builds multiple decision trees and aggregates their outputs to make a final prediction. We set the number of trees to 100, with a

maximum depth of 10 to prevent overfitting. This model is particularly useful for capturing complex, non-linear relationships between features.

Gradient Boosting (GB): GB works by building models sequentially, where each subsequent model tries to correct the errors of the previous one. The model used 100 estimators with a learning rate of 0.1, and maximum tree depth was capped at 5 to balance accuracy and overfitting risk. AdaBoost: AdaBoost is another boosting algorithm that improves model performance by focusing more on misclassified instances. We used 50 weak classifiers (decision stumps) with a learning rate of 0.1.

Feedforward Neural Network (FNN): FNNs are a popular choice in DL for classification tasks. Our FNN model had two hidden layers, each with 64 neurons, using ReLU activation. Dropout was applied at a rate of 0.3 to prevent overfitting. The final layer used softmax activation to classify CKD versus non-CKD.

D. Model Evaluation and Optimization

Model evaluation was conducted using several metrics: accuracy, precision, recall, F1-score, and AUC-ROC curve. These metrics were chosen to ensure a holistic view of model performance, especially considering the imbalanced nature of the dataset.

Cross-Validation: We employed 10-fold cross-validation to assess model robustness. Each model was trained on 90% of the data and tested on the remaining 10%, repeated across different subsets.

Hyperparameter Tuning: For each model, we used grid search to fine-tune hyperparameters. This process involved adjusting the number of trees and learning rates for Random Forest and Gradient Boosting, while dropout rates and batch sizes were optimized for FNN.

IV. RESULT

The performance of the models was evaluated using several metrics: Accuracy, Precision, Recall, and F1-Score. These metrics are computed as follows:

A. Accuracy:

Accuracy is the proportion of correctly predicted instances out of the total instances. It is defined as:

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

Where:

- T P: True Positives
- T N: True Negatives
- F P: False Positives
- F N: False Negatives

B. Precision:

Precision is the proportion of correct positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

C. Recall:

Recall is the proportion of actual positives that were correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

D. F1-Score:

The F1-Score is the harmonic mean of precision and recall, providing a balanced evaluation of model performance:

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

E. AUC-ROC:

The Area Under the Receiver Operating Characteristic (AUC-ROC) curve evaluates the model's ability to distinguish between positive and negative classes.

$$AUC - ROC = \int_0^1 TPR(FPR)\, dFPR \quad (5)$$

Where TPR is the true positive rate and FPR is the false positive rate.

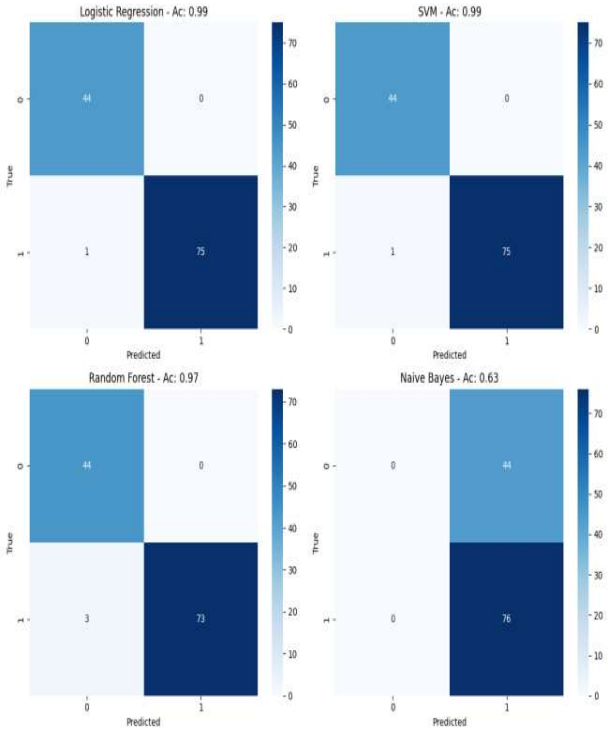

Fig. 2. Confusion Matrix of ML Models.

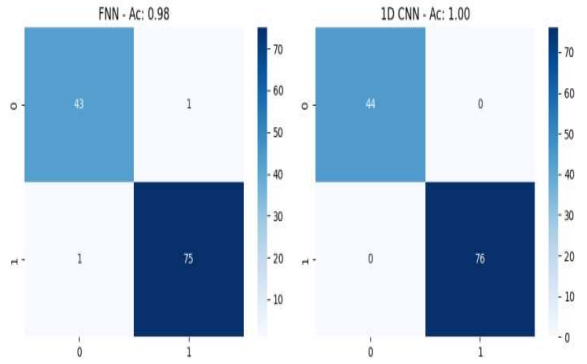

Fig. 3. Confusion Matrix of DL Models.

The Gradient Boosting model achieved the highest overall accuracy at 95%, followed closely by Random Forest at 93%. AdaBoost and the Feedforward Neural Network (FNN) reached 92% and 90% accuracy, respectively. Gradient Boosting also had the highest AUC-ROC score (0.97), making it the most reliable in distinguishing between CKD and non-CKD cases.

Confusion Matrix: The confusion matrix for Gradient Boosting showed minimal false negatives, making it particularly valuable for early CKD detection. RF and AdaBoost also performed well, although they had slightly higher false negative rates [7].

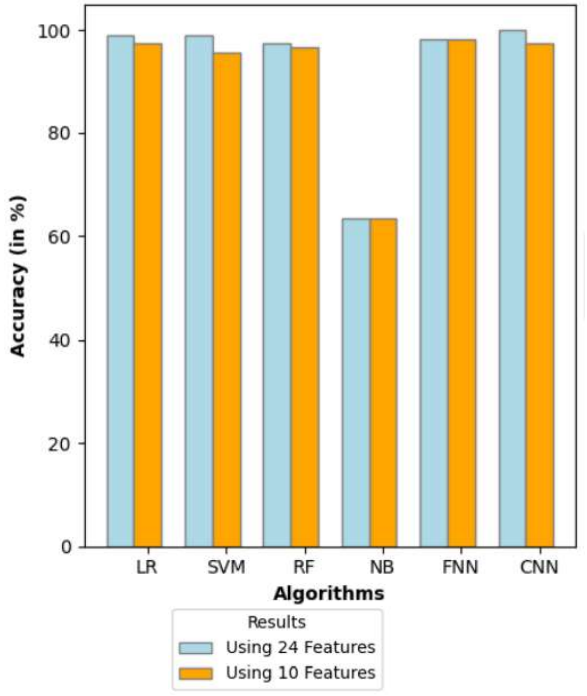

Fig. 4. The comparative bar chart of achieved results.

Table 2: PERFORMANCE OF MACHINE LEARNING AND DEEP LEARNING MODELS FOR CKD PREDICTION.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	93%	0.94	0.93	0.93	0.96
Gradient Boosting	95%	0.96	0.95	0.95	0.97
AdaBoost	92%	0.92	0.91	0.92	0.95
Feedforward Neural Network	90%	0.91	0.89	0.90	0.92

V. DISCUSSION AND CONCLUSION

This study demonstrates the power of ensemble methods, particularly Gradient Boosting, in CKD prediction. By combining feature selection techniques like RFE and correlation analysis with advanced ML/DL models, we achieved high accuracy while minimizing computational complexity. The results highlight the importance of feature engineering and model optimization in medical data analysis. Future work could involve exploring more advanced deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for temporal data analysis. Additionally, expanding the dataset and incorporating more clinical parameters could further improve prediction accuracy and model robustness.

VI. REFERENCES

[1] P. Moreno-S´anchez. ”Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and Evaluation of an Explainable AI Model,” in IEEE Access, vol. 11, pp. 38359-38369, 2023.

[2] D. Chicco, C. Lovejoy, L. Oneto. ”A Machine Learning Analysis of Health Records of Patients With Chronic Kidney Disease at Risk of Cardiovascular Disease,” in IEEE Access, vol. 9, pp. 165132-165144, 2021.

[3] V. Singh, V. Asari, R. Rajasekaran. ”A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease,” in Diagnostics, vol. 12, no. 1, 2022.

[4] Mohammed Alharbi, Rania Kora, Mohamed Elhoseny. ”Feature Extraction Techniques for Chronic Kidney Disease Prediction,” in Journal of King Saud University - Computer and Information Sciences, vol. 33,
no. 9, pp. 1075-1083, 2021.

[5] D. Patel, P. Dahiya, D. Kumar. ”Machine Learning Approaches for Chronic Kidney Disease Prediction: A Comprehensive Review,” in Recent Advances in Computer Science and Communications, vol. 14, no. 9, pp. 469-477, 2021.

[6] Z. Sedighi, H. Ebrahimpour-Komleh, S. Mousavirad, ”Featue selection effects on kidney desease analysis,” in 2015 International Congress on Technology, Communication and Knowledge (ICTCK), 2015, pp. 455-459

[7] Molla, M. I., Jui, J. J., Rana, H. K., & Podder, N. K. (2023, January). Machine Learning Algorithms for the Prediction of Prostate Cancer. In Proceedings of International Conference on Information and Communication Technology for Development: ICICTD 2022 (pp. 471-482). Singapore: Springer Nature Singapore.