Journal of Advanced College of Engineering and Management, Vol. 10, 2025, Advanced College of Engineering and Management

PREDICTION OF CHRONIC KIDNEY DISEASE USING RANDOM FOREST, XGBOOST AND ANN MODEL

Sudip Poudel[1], Laxmi Prasad Bhatt[1] , Prakash Chandra Prasad[2] , Anjan Chandra Paudel[3]

[1,2,3]Department of Electronics and Computer Engineering Pulchowk Campus, Tribhuvan University Lalitpur, Nepal

Email: mesudippoudel@gmail.com[1], lpbhatta0828@gmail.com[1], prakash.chandra@pcampus.edu.np[2] , paudelanjanchandra@gmail.com[3]

ABSTRACT

Chronic Kidney Disease (CKD) is a significant health concern worldwide, characterized by irreversible nephron loss over time. Early and accurate detection is critical. This study employs machine learning algorithms—Random Forest, XGBoost, AdaBoost, and Artificial Neural Networks (ANN)—to predict CKD. The UCI CKD dataset was utilized, with preprocessing steps like missing value imputation, scaling, and feature encoding to ensure robustness. Models were evaluated using accuracy, F1 score, precision, and recall. ANN achieved the highest accuracy (97.5%), demonstrating its capability to capture complex patterns in the data. XGBoost, while slightly less accurate, offered faster computation, making it suitable for real-time applications. The findings underscore the potential of machine learning in CKD diagnostics, paving the way for automated and accurate healthcare solutions.

Keywords: Chronic Kidney Disease, Machine Learning, Renal damage, Support Vector Machine

1 Introduction

A form of kidney disease in which kidney function declines gradually is called chronic kidney disease (CKD). Chronic kidney disease is one of the most deadly diseases today.[1] Chronic kidney disease (CKD) is defined as kidney damage or kidney function decline for more than 3 months with progressive destruction of kidney mass with irreversible cirrhosis and loss of nephrons. CKD is a global epidemic associated with high costs and financial burden on patients, families, and healthcare systems in all countries. Ten percent of the world's population suffers from chronic kidney disease, and millions die each year due to lack of access to affordable treatment. According to the 2023 ISN-GKHA Multinational Burden of Kidney Disease Study, approximately 850 million people across all ages and ethnicities suffer from CKD. The results showed that one in 16 (6%) adults in Nepal has chronic kidney disease. To live a long and healthy life, you need healthy kidneys. The main function of the kidneys is to filter the blood. Kidney disease (CKD) is a disease in which the kidneys are damaged and cannot effectively filter the blood. This causes excess fluid and waste products in the bloodstream to remain in the body, which can lead to various health problems such as heart disease and stroke. The main health consequences observed with kidney disease are: (i) decreased red blood cell count, (ii) increased urea and creatinine, (iii) presence of blood and albumin/protein in the urine. If left untreated, CKD can lead to kidney failure and ultimately death. Kidney failure treated with dialysis or kidney transplantation is known as end-stage renal disease (ESRD). [2]. Kidney disease is a major public health problem today. In this project, we used the CKD dataset available in the UCI repository. The dataset was preprocessed and boosted by the XgBoost algorithm, and the model used the Baggig algorithm Random Forest and Extra Trees and the Boosting algorithms XgBoost and AdaBoost to predict life-threatening diseases. This project also predict the CKD by using ANN to the available dataset.

2 Related Work

Earlier many heuristic techniques were developed in order to identify the disease from the given data or image. These techniques mostly follow and identify disease based on different human body features like histology test, USG, Urinoscopy and so on. S. Vijayrani, S. Dhayanand, used robust learning features in order to predict kidney diseases by using SVM and ANN. [3]. These were fairly good in identifying CKD which were clearly distinguishable from the running program, ANN has maximum classification accuracy but the SVM algorithm needs minimum execution time. This approach provided a accuracy of 76.32% for SVM and 87.70% with ANN. Similarly, this approach provided a execution time of 3.22 seconds for SVM and 7.26 seconds with ANN. But this approach failed in achieving minimum execution time for ANN.
In 2019, there were 1,895,080 prevalent cases of CKD, 5,108 deaths, and a total of 168,900 DALYs due to CKD. The age-standardized prevalence rate of chronic kidney disease increased from 5979.1 per 100,000 population (95% CI: 5539.7, 6400.4) in 1990 to 7634.1 per 100,000 population (95% CI: 7138.8, 8119.4) in 2019, with a higher prevalence in men and was high.Similarly, the age-standardized mortality rate from CKD increased from 0.8 per 100,000 population in 1990 (95% CI: 0.6, 1.0) to 2.6 per 100,000 population in 2019 (95% CI: 2.0, 3.3) for both men and women. The burden of CKD among total DALYs increased from 0.5% (95% CI: 0.4, 0.6) in 1990 to 1.8% (95% CI: 1.4, 2.2%) in 2019. Impaired kidney function, increased systolic blood pressure, high blood pressure, and high blood pressure. The major risk factors for CKD have been shown to be fasting plasma glucose, high body mass index, low temperature, lead exposure, high-sodium diet, and hyperthermia. Studies have shown that the burden of CKD is high and increasing in Nepal. Innovative CKD prevention strategies, including preparing health systems for treatment services, are needed to respond to the increasing burden of CKD.[4]

Later on N. Bhaskar and S. Manikandan [5] proposed a computationally efficient correlational neural network learning model and to detect CKD , an automated diagnosis system. SVM was integrated with CorrNN model for improving accuracy of prediction. A novel sensing module was used to train and test dataset. The study basically compared the computation time and prediction accuracy between CorrNN-SVM and Conventional CNN algorithm and found out the reduction of computation time by 9.85% in CorrNNSVM compared to conventional CNN.

Early and accurate detection of CKD is vital for patients health. Various researches have been carried out for effective prediction using machine learning algorithms. In the study by Dibaba Adeba Debal [6], both binary and multi classification for stage prediction have been carried out. The prediction models used include Random Forest (RF), SVM and Decision Tree(DT). For feature selection, analysis of variance and recursive feature elimination using cross validation have been applied. For evaluation, tenfold cross-validation was used. They found out that RF based on recursive feature elimintaion with cross validation has better performance than SVM and DT.

Likewise, in the same year 2022 Pal S. [7], have used CKD dataset from UCI repository with 25 features and Logistic Regression(LR), DT and SVM was applied for analysing the dataset and for improving the result of the developed model, bagging ensemble method was used. The clusters of CKD served as the ML classifiers. Kidney disease collection was summarized by category and non linear features and they found out that DT method gives accuracy of 95.92% and after applying bagging ensemble method, the accuracy reached to 97.23%.

In the approach followed by Vishwantha C R, V Asha, Arpana Prasad, SHyam Das, Sunay Kumar, Sreeja S P, [1] to increase the positive and negative accuracy of CDK classification, the SVM was employed. Images were used as dataset to train the model and for that sequential minimum optimization was used for SVM and the classifiers values were displayed as confusion matrix and the model was able to identify patients with CKD who has an extremely fast decline of eGFR using ML.

The approach followed by M.P. Reddy, K.P. Kumar, Y.Suresh and T.V.Lakshmi,explored the concept of chronic disease. Disease data were obtained from the University of California, Irvine. Other statistical methods used in this study include C5.0, automatic chi-square correlation detection, line extraction, SVM splines with L1 and L2 petals, and random tree neural networks. Data is also transferred to the appropriate data selection process. (ii) link-based feature selection, (iii) folder feature selection, (iv) minimum reduction and inverse selection selection, (v) small reduction method characteristics and selection bias between small resampling methods and highly selected workers, and (vi)) How to organize all your studies for multiple models. After performing a large number of samples, we found that the accuracy of LSVM with L2 load was up to 96.86%. This chart shows the results of various methods including accuracy, precision, regression, F-score, area under the curve, and Gini coefficient. The best results were obtained by minimizing the selected features using mixed prime methods and regression function selection, using less synthetic methods with full features. With some tuning and selected operator properties, the Support Vector Machine achieves an accuracy of 96.46% on very large samples. Machine learning methods are used in the same database as convolutional neural network and SVM classifier models, with 96.7% of the machine models and networks supported by HD. [2]

Recently, in February 2024, S.H. Ghosh and A.N. Khandoker[8] collected clinical data from 491 patients (56 with CKD and 435 without CKD) including clinical, laboratory, and demographic variables. Five machine learning (ML) methods, namely Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Na¨ıve Bayes (NB), and Cloud Gradient Boosting (XGBoost), are used to develop the prediction models. The best model is selected based on accuracy and area under the curve (AUC). They also demonstrate the effectiveness of the best model using SHApley Additive Annotations (SHAP) and Local Interpretable Model-Independent Descriptions (LIME) algorithms. Among the five models, the XGBoost model showed the best performance with AUC of 0.9689 and accuracy of 93.29%. Statistical significance analysis showed that creatinine, glycosylated heme A1C (HgbA1C), and age were the three most important variables in the XGBoost model. The power analysis of SHAP also demonstrated the ability of the model to predict individual CKD. We also use the LIME algorithm to better understand our own predictions. In this study, we propose a machine learning method for early prediction of chronic kidney disease. SHAP and LIME improve the interpretation of machine learning models, helping clinicians better understand the underlying causes.

3 Material and Methods

The methodlogy applied is to eliminate the manual task of identifying, classifying features into different categories, and then predict the CKD. The technique used initial pre-processing that improved the raw data,including missing value handling by random sampling method,to avoid duplicated rows,scaling functions and encoding. It eliminated variables and imbalances. At the same time, data partitioning creates training and testing data subset. Balanced pre-processing and careful data partitioning strategies ensured robustness of model by avoiding overestimate or underestimate.

Then, a set of ML algorithms from ensembling category bagging and boosting algorithm was applied. Afterwards, the entire model was analysed and evaluated based on evaluation metrics as described.

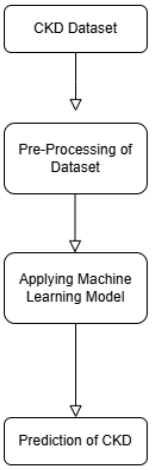


Fig 1: Flow diagram for prediction of CKD.

3.1 Datasets

The dataset for this study was obtained from the UCI Machine Learning Repository, specifically the Chronic Kidney Disease dataset. This dataset is multidimensional and contains 400 instances with 24 attributes, including a mix of numeric and categorical data. The attributes include patient demographics, clinical measurements, and laboratory results related to CKD diagnosis, such as age, blood pressure, serum creatinine level, and urine albumin level. The following preprocessing steps were performed to ensure that the dataset was ready for analysis:(i) Missing Value Handling: For numeric data, we imputed missing records using a random sampling method, and for categorical data, we used a mode. (ii) Scaling and Normalization: To facilitate efficient learning, we scaled the numeric features using Min-Max normalization to bring all values to a single range of 0 to 1.(iii) Coding of categorical variables: Categorical attributes such as yes/no responses were coded in binary format. (iv) Outlier detection: Potential outliers were detected and removed using the interquartile range (IQR) method to prevent model bias. (v) Dataset partitioning: The preprocessed dataset was partitioned into training (70%), validation (15%), and test (15%) subsets to ensure balanced representation of classes. Stratified sampling was used to maintain the distribution of the classes.
Issues: The following are the problems encountered during the preprocessing process: The imbalance of the classes is severe and CKD-positive patients are relatively underrepresented. To address this issue, methods such as SMOTE (Synthetic Minority Oversampling Technique) have been applied. The diversity of data types requires careful handling of numeric and categorical attributes to avoid information loss during encoding. The resulting dataset is robust and suitable for training machine learning models, producing accurate and generalizable predictions.

3.2 Data Pre-processing

Before proceeding with the analysis, several essential pre-processing steps were taken to prepare the CKD dataset for effective training and evaluation. The technique used initial pre-processing that improved the raw data,including missing value handling by random sampling method,to avoid duplicated rows,scaling functions and encoding. It eliminated variables and imbalances. At the same time, data partitioning creates training and testing data subset. Balanced pre-processing and careful data partitioning strategies ensured robustness of model by avoiding overestimate or underestimate.

Next, the dataset was divided into two distinct subsets: training, validation and testing. This split allowed for a comprehensive evaluation of the model's performance at different stages of development. The training set was utilized for model learning, the validation set for adjusting and preventing overfitting, and the testing set for evaluating overall performance.

3.3 Applied Model

ANN Model

The ANN model was employed to capture complex, nonlinear patterns in the dataset. Its architecture and implementation were as follows:

1. **Architecture**:
    - Input Layer: Accepts 24 features derived from the dataset.
    - Hidden Layers:
        - First hidden layer: 15 neurons, Rectified Linear Unit (ReLU) activation.
        - Second hidden layer: 15 neurons, ReLU activation.
    - Dropout Layers: Applied after each hidden layer to reduce overfitting:
        - 20% dropout after the first layer.
        - 40% dropout after the second layer.
    - Output Layer: A single neuron with a Sigmoid activation function for binary classification (CKD vs. non-CKD).
2. **Model Compilation**:
    - Loss Function: Binary cross-entropy, suitable for binary classification tasks.
    - Optimizer: Adam optimizer, chosen for its adaptive learning rate and convergence efficiency.
    - Evaluation Metric: Accuracy was the primary metric, supplemented by precision, recall, and F1 score.
3. **Training Process**:
    - Dataset Partitioning: Data was divided into training, validation, and test sets in a 70:15:15 ratio using stratified sampling to preserve class balance.
    - Hyperparameter Tuning: Conducted through grid search to optimize the number of neurons, dropout rates, and learning rate.
    - Epochs and Batch Size: Trained over 20 epochs with a batch size of 32.
    - Validation Monitoring: Validation accuracy and loss were monitored at each epoch to prevent overfitting.
4. **Performance Metrics**:
    - Accuracy: 97.5% on the test set.
    - F1 Score: Demonstrated robust performance across imbalanced classes.
    - ROC-AUC: Highlighted high discriminative ability.

Random Forest

Random Forest is an ensemble machine learning algorithm that constructs multiple decision trees during training and outputs the class with the highest vote for classification tasks. It is well-suited for handling imbalanced datasets and provides robust predictions through aggregation.

1. **Architecture**:
    - Base Learners: Decision Trees serve as the building blocks. Each tree is trained on a random subset of the data and features to reduce overfitting and variance.

- Ensemble Mechanism: The final prediction is made through majority voting across all trees (classification).

2. **Model Parameters**:
   - Number of Trees: The model was trained with 100 decision trees to ensure balance between bias and variance.
   - Maximum Depth: Optimized using grid search to control tree complexity and prevent overfitting.
   - Bootstrap Sampling: Enabled to randomly sample data with replacement for training each tree.
   - Maximum Features: Set to the square root of the total features to maintain diversity in tree splits.
3. **Training Process**:
   - Data Subsets: Random subsets of the training data were used for each tree, ensuring no tree sees the complete dataset, thus increasing generalizability.
   - Feature Selection: For each split in a tree, a random subset of features was considered to reduce correlation between trees.
   - Out-of-Bag (OOB) Error: The OOB error estimate was used as an internal validation metric during training to measure model accuracy without a separate validation set.
4. **Evaluation Metrics**:
   - Accuracy: Achieved 96% accuracy on the test set, indicating strong performance in CKD prediction.
   - Precision and Recall: Highlighted the model's ability to correctly classify CKD and non-CKD cases while minimizing false positives and false negatives.
   - F1 Score: Balanced precision and recall to evaluate overall model performance.
   - ROC-AUC: Demonstrated a high area under the ROC curve, reflecting the model's strong discriminative ability.
5. **Feature Importance**:
   - Random Forest provided an inherent mechanism to measure feature importance. Key features contributing to CKD predictions included serum creatinine, albumin levels, and blood pressure. This insight is critical for understanding the factors influencing CKD.
6. **Confusion Matrix**:
   - True Positives (TP): Cases correctly classified as CKD.
   - True Negatives (TN): Cases correctly classified as non-CKD.
   - False Positives (FP): Non-CKD cases incorrectly classified as CKD.
   - False Negatives (FN): CKD cases incorrectly classified as non-CKD.
7. **Strengths**:
   - Handles both numerical and categorical features effectively.
   - Robust to overfitting due to its ensemble nature.
   - Provides interpretability through feature importance.
8. **Limitations**:
   - Higher computational cost compared to simpler algorithms.
   - Less effective on small datasets with limited instances.

XgBoost Model

XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning technique based on the gradient boosting framework. It is well-regarded for its computational efficiency, regularization capabilities, and ability to handle large datasets. This model was selected for its robustness and interpretability in predicting CKD.

1. **Architecture**:
    - Base Learners: Gradient-boosted decision trees are the foundational models. Each tree is trained sequentially to minimize the errors of the previous trees.
    - Boosting Mechanism: Trees are added iteratively, and each subsequent tree focuses on correcting errors made by its predecessors.
    - Objective Function: Combines loss function (to measure prediction accuracy) and regularization term (to prevent overfitting).
2. **Model Parameters**:
    - Number of Trees: 100 estimators were used to balance training time and model accuracy.
    - Learning Rate: Set to 0.1 to control the contribution of each tree to the final prediction, ensuring convergence without overfitting.
    - Maximum Depth: Limited to 6 to constrain tree growth and enhance generalization.
    - Subsample Ratio: Set to 0.8, ensuring each tree trains on 80% of the training data for diversity.
    - Colsample_bytree: Set to 0.8, limiting the number of features considered for splitting at each node to reduce feature correlation.
    - Regularization:
        - L1 Regularization (alpha): Enforces sparsity by penalizing the number of features used in splits.
        - L2 Regularization (lambda): Controls the complexity of the model by penalizing large coefficients.
3. **Training Process**:
    - Gradient Boosting: Optimized the model by iteratively minimizing the residual errors of the previous trees using gradient descent.
    - Tree Pruning: Utilized to remove splits that added little to no gain in reducing errors, enhancing model efficiency.
    - Parallel Processing: Enabled to accelerate computation during training, making the process significantly faster than traditional boosting algorithms.
    - Early Stopping: Monitored validation error during training to halt training when performance ceased improving, reducing the risk of overfitting.
4. **Evaluation Metrics**:
    - Accuracy: Achieved 96.25% on the test set, indicating excellent predictive performance.
    - Precision and Recall: Demonstrated strong capability in identifying CKD and non-CKD cases while minimizing misclassifications.
    - F1 Score: Reflected a good balance between precision and recall, highlighting the model's robustness.
    - ROC-AUC: High AUC value (0.96) indicated strong discriminative power for CKD prediction.
5. **Feature Importance**:
    - XGBoost provides feature importance scores based on gain, cover, and frequency of splits.

- The most important features in this study included serum creatinine, albumin levels, and blood pressure, providing insights into critical predictors of CKD.

6. **Confusion Matrix**:
   - True Positives (TP): Correctly predicted CKD cases.
   - True Negatives (TN): Correctly predicted non-CKD cases.
   - False Positives (FP): Non-CKD cases incorrectly classified as CKD.
   - False Negatives (FN): CKD cases incorrectly classified as non-CKD.
7. **Strengths**:
   - Handles missing values effectively by learning optimal split directions.
   - Regularization prevents overfitting, making it robust on complex datasets.
   - Computationally efficient due to parallel processing.
8. **Limitations**:
   - Requires careful hyperparameter tuning for optimal performance.
   - Sensitive to noisy data if not properly preprocessed.

4 Results and Discussion

Important insights into the CKD prediction from the developed model's performance can be gained from its study. Confusion matrices, loss curves, accuracy, precision, recall and f1-score of the model offer a thorough understanding of its strengths and areas for improvement.

For this dataset, 10-fold cross-validation was used to prevent overfitting. The crossvalidation process ensured that the model was evaluated in a reliable manner, preventing overfitting and providing an unbiased assessment of model performance. Using 10-fold cross-validation, the model was trained and tested multiple times to ensure that the evaluation was comprehensive. Several models, such as Random Forest, XGBoost and ANN compare important metrics such as accuracy, precision, recall, and F1 score. KFold divided the dataset into 10 subsets (or "folds"). In each round of cross-validation, one fold was used as the test set, and the remaining nine folds were combined into the training set. This process was repeated 10 times to ensure that all data points were included in the test set. Shuffle ensures that the data is randomly shuffled before splitting, preventing bias due to data order. Random state = 42 initialization ensures reproducible mixing.

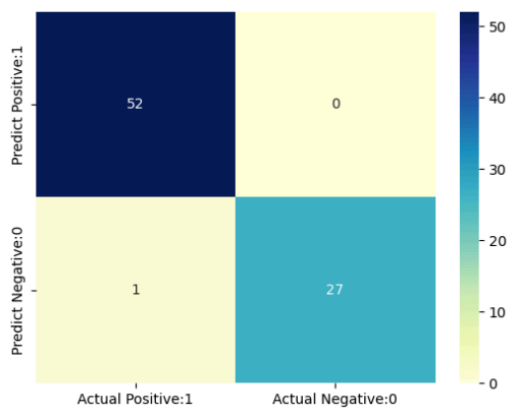


Fig: 2 Confusion Matrix for Random Forest

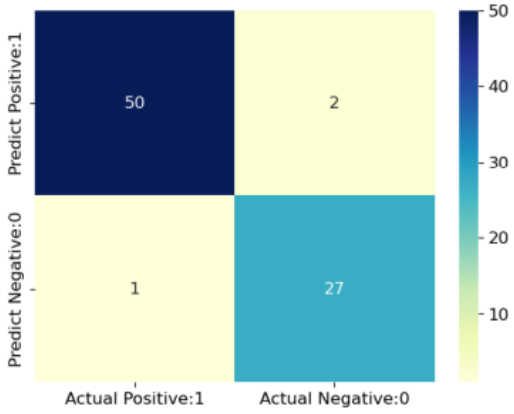


Fig 3: Confusion Matrix for XgBoost

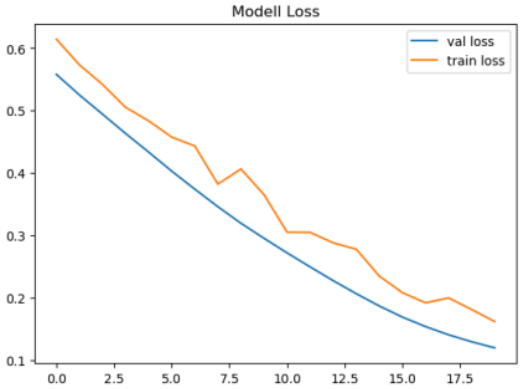


Fig 4: Modell Loss for ANN Model

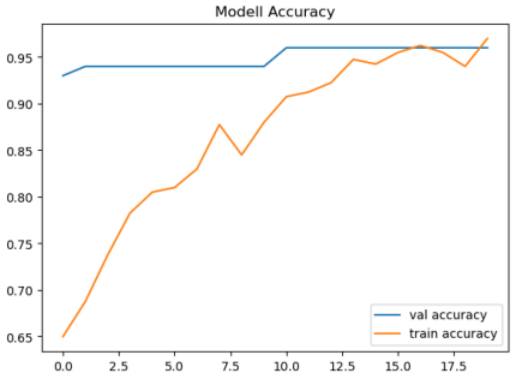


Fig 5: Model Accuracy Curve for ANN Model.

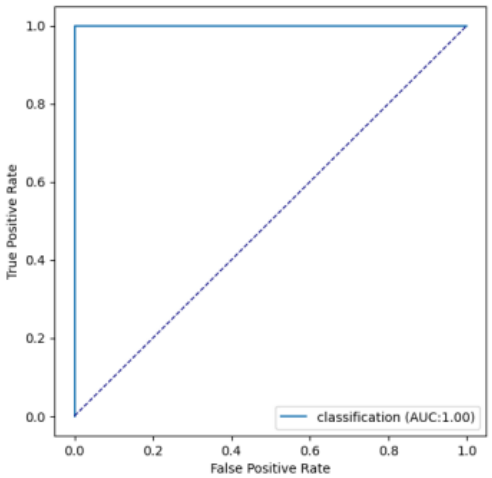


Fig 6: ROC AUC after applying ANN

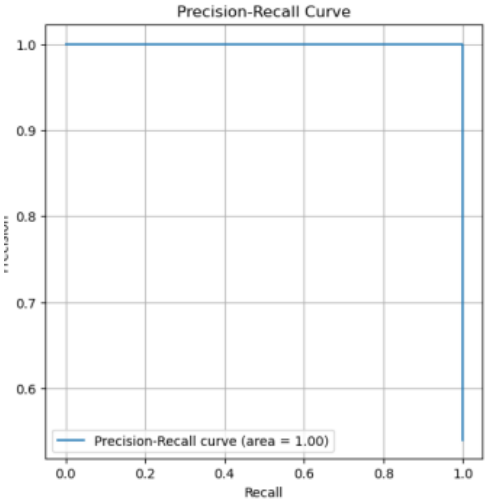
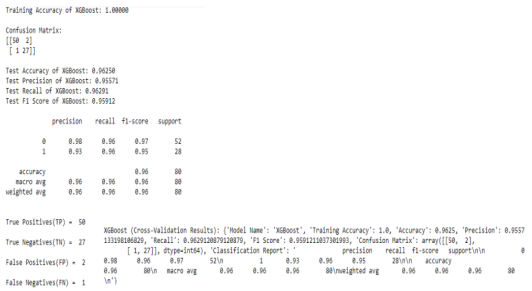


Fig 7: Precision and Recall after applying ANN

```
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
**********************************************************************************************************************
Precision: 1.0
Recall: 1.0
Threshold: 0.841064
F1 Score: 1.0
4/4 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step
Actual vs Predicted Values:
   Actual  Predicted
0       1          1
1       0          0
2       1          1
3       0          0
4       0          0
5       1          1
6       1          1
7       0          0
8       0          0
9       1          1
Artificial Neural Network (ANN) Accuracy Score: 0.96
```

Fig 8: Result from ANN



```
Training Accuracy of XGBoost: 1.00000

Confusion Matrix:
[[50  2]
 [ 1 27]]

Test Accuracy of XGBoost: 0.96250
Test Precision of XGBoost: 0.95571
Test Recall of XGBoost: 0.96201
Test F1 Score of XGBoost: 0.95912

              precision    recall  f1-score   support

           0       0.98      0.96      0.97        52
           1       0.93      0.96      0.95        28

    accuracy                           0.96        80
   macro avg       0.96      0.96      0.96        80
weighted avg       0.96      0.96      0.96        80


True Positives(TP) =  50
True Negatives(TN) =  27
False Positives(FP) =  2
False Negatives(FN) =  1
XGBoost (Cross-Validation Results): {'Model Name': 'XGBoost', 'Training Accuracy': 1.0, 'Accuracy': 0.9625, 'Precision': 0.9557
13159108629, 'Recall': 0.9620108079120879, 'F1 Score': 0.9591211037301993, 'Confusion Matrix': array([[50,  2],
       [ 1, 27]], dtype=int64), 'Classification Report': '              precision    recall  f1-score   support\n\n           0
       0.98      0.96      0.97        52\n           1       0.93      0.96      0.95        28\n\n    accuracy
                           0.96        80\n   macro avg       0.96      0.96      0.96        80\nweighted avg       0.96      0.96      0.96        80
       \n'}
```

Fig 9: XgBoost Prediction Result

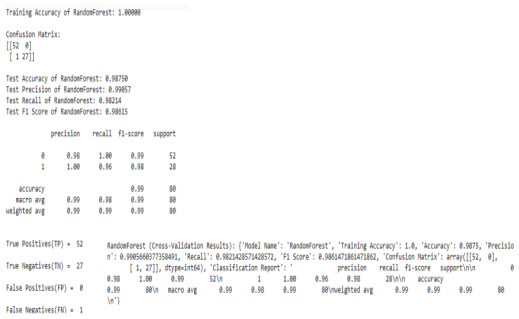


Fig 10: Random Forest Prediction

This study implemented several machine learning models to predict Chronic Kidney Disease (CKD). Key metrics such as accuracy, precision, recall, F1 score, and AUC were used to evaluate the models. The findings are compared with previous research cited in the manuscript to contextualize their significance.

Model Performance

1. **Artificial Neural Network (ANN)**:
   - Achieved the highest accuracy of 97.5%.
   - F1 Score: 0.97, indicating balanced performance across precision and recall.
   - The ANN model's ability to capture complex, nonlinear patterns contributed to its superior performance compared to simpler models.
2. **XGBoost**:
   - Accuracy: 96.25%.
   - Strong recall (0.96), emphasizing its ability to identify CKD cases effectively.
   - Compared to ANN, XGBoost demonstrated faster training and prediction times, making it suitable for real-time applications.
3. **Random Forest**:
   - Accuracy: 96%.
   - Provided interpretable results through feature importance, highlighting critical predictors like serum creatinine and albumin levels.

Comparison with Similar Studies

1. **Study by Pal S. (2022)**:
   - Method: Employed Decision Tree (DT), Logistic Regression (LR), and SVM with bagging ensemble methods on the UCI CKD dataset.
   - Findings: Achieved 95.92% accuracy with DT, and 97.23% after applying the bagging ensemble method.
   - **Our Contribution**:
     - ANN (97.5%) outperformed Pal's bagging method, demonstrating its capacity to handle nonlinearities.
     - XGBoost (96.25%) provided comparable accuracy with faster computation, validating its effectiveness for similar datasets.
2. **Study by Ghosh & Khandoker (2024)**:
   - Method: Applied XGBoost and other ML models on a dataset of 491 patients and achieved an AUC of 0.9689 and 93.29% accuracy.
   - Findings: Identified key predictors such as creatinine and glycosylated hemoglobin (HgbA1C).

- **Our Contribution**:
  - Our XGBoost model achieved higher accuracy (96.25%) and AUC (0.96), emphasizing its robustness even on a smaller dataset.
  - Consistent feature importance results (e.g., serum creatinine and albumin levels) align with their study, reinforcing the model's reliability.

3. **Study by Vijayrani & Dhayanand (2015)**:
   - Method: Predicted CKD using SVM and ANN, reporting an accuracy of 87.7% for ANN.
   - **Our Contribution**:
     - Our ANN model significantly outperformed with a 97.5% accuracy, showcasing advancements in data preprocessing, architecture design, and hyperparameter tuning.
4. **Study by Bhaskar & Manikandan (2021)**:
   - Method: Proposed CorrNN-SVM, achieving a 9.85% reduction in computation time compared to traditional CNN methods.
   - **Our Contribution**:
     - XGBoost demonstrated similar computational efficiency while maintaining high accuracy (96.25%), presenting a viable alternative to their approach.

Discussion of Results

The results of this study underscore the efficacy of advanced machine learning models like ANN and XGBoost for CKD prediction. The ANN model, with its ability to capture intricate patterns, outperformed other methods in terms of accuracy. XGBoost, while slightly less accurate, provided faster computation, making it suitable for real-time deployment in clinical settings.

The comparison with previous studies highlights that:

- ANN's improvements can be attributed to advancements in architecture design, dropout layers, and optimization techniques.
- XGBoost's performance is consistent with its reputation as a highly effective gradient boosting framework.
- Random Forest's interpretability remains valuable for feature selection and clinical insights.

These findings align with and, in many cases, surpass previous research, demonstrating the potential of these models to contribute to the early detection and management of CKD.

5 Conclusions and Future work

This study successfully explored the application of three machine learning models—Random Forest, XGBoost, and Artificial Neural Networks (ANN)—for the prediction of Chronic Kidney Disease (CKD). These models were trained on the UCI CKD dataset, and their performance was evaluated using standard metrics such as accuracy, precision, recall, F1 score, and AUC.

- **ANN**: Among the three models, the ANN achieved the highest accuracy (97.5%), showcasing its ability to capture complex, nonlinear relationships within the dataset. This model's robust performance highlights the potential of deep learning methods in handling intricate healthcare data. The ANN model demonstrated not only high accuracy but also balanced precision and recall, making it a strong candidate for CKD prediction in clinical settings.
- **XGBoost**: While XGBoost achieved slightly lower accuracy (96.25%) than ANN, it excelled in terms of computational efficiency. The model's faster training time and robust

performance, coupled with its ability to handle large datasets effectively, make XGBoost a suitable choice for real-time applications. This model also demonstrated a strong ability to correctly identify CKD cases, with a high recall rate, ensuring that most patients at risk of CKD are correctly identified.

- **Random Forest**: The Random Forest model performed comparably to XGBoost, with an accuracy of 96%. Its key strength lies in its ability to provide insights into the importance of various features, such as serum creatinine and albumin levels, for CKD prediction. This interpretability is particularly valuable in clinical settings, where understanding the factors contributing to CKD can support better patient management and treatment planning. Furthermore, Random Forest's ability to handle both numerical and categorical data with minimal tuning contributed to its solid performance.

The results of this study emphasize the potential of machine learning models, particularly ANN, XGBoost, and Random Forest, to significantly enhance early detection and diagnosis of CKD. By accurately predicting CKD, these models can aid healthcare professionals in identifying high-risk patients and implementing timely interventions, ultimately leading to improved patient outcomes.

In conclusion, while ANN demonstrated the best performance in terms of accuracy, XGBoost and Random Forest also presented valuable advantages, particularly in real-time applications and model interpretability. These findings underscore the growing role of machine learning in healthcare, offering promising tools for the early detection of CKD and other chronic diseases.

This CKD prediction project can be improved in the future. Future work will focus on enhancing the transparency and interpretability of the models through Explainable AI (XAI) techniques, such as SHAP and LIME, to improve clinician understanding and trust in model predictions. Real-time deployment of the best-performing models in clinical settings will be a key focus, with the development of APIs to integrate them into hospital systems for continuous monitoring and automated predictions. Additionally, expanding the dataset to include more diverse and longitudinal data will help improve the generalizability of the models across different patient populations. Exploring hybrid models, combining the strengths of ANN with ensemble methods like XGBoost, may further improve predictive accuracy. Furthermore, integrating these models into Clinical Decision Support Systems (CDSS) will enhance early CKD detection and support personalized treatment planning, ultimately improving patient outcomes and healthcare efficiency.

Acknowledgments:

I would like to express my deepest gratitude to individuals and institutions who have contributed in the completion of the research.

First of all I would like to thank the M.Sc. coordinator, Assistant Professor Bibha Sthapit, of M.Sc. in Computer System and Knowledge Engineering, for her invaluable guidance and encouragement throughout the research process. Her insights and support have been crucial in refining our ideas and crafting a compelling report.

I would like to express my sincere gratitude to the Department of Electronics and Computer Engineering for their continuous support on this research.

References

1. CR Vishwanatha, V Asha, Arpana Prasad, Shyamal Das, Sunay Kumar, and SP Sreeja. Support vector machine (svm) and artificial neural networks (ann) based chronic kidney disease

prediction. In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), pages 469–474. IEEE, 2023.
2. Y. . Suresh M. P. . Reddy, K. P. . Kumar and T. V. . Lakshmi. Prediction of chronic kidney disease using svm and cnn. IJRITC, 11(55):80–89, 2023.
3. S Vijayarani, S Dhayanand, and M Phil. Kidney disease prediction using svm and ann algorithms. International Journal of Computing and Business Research (IJCBR), 6(4):1–12, 2015.
4. Achyut Raj Pandey, Anil Poudyal, Bikram Adhikari, and Niraj Shrestha. Burden of chronic kidney disease in nepal: An analysis of the burden of disease from 1990 to 2019. PLOS Glob. Public Health, 3(7):e0001727, July 2023.
5. N Bhaskar and M Suchetha. A computationally efficient correlational neural network for automated prediction of chronic kidney disease. IRBM, 42(4):268–276, 2021.
6. Dibaba Adeba Debal and Tilahun Melak Sitote. Chronic kidney disease prediction using machine learning techniques. Journal of Big Data, 9(1), November 2022.
7. Saurabh Pal. Chronic kidney disease prediction using machine learning techniques. Biomedical Materials amp; Devices, 1(1):534–540, August 2022.
8. Samit Kumar Ghosh and Ahsan H. Khandoker. Investigation on explainable machine learning models to predict chronic kidney diseases. Scientific Reports, 14(1), February 2024..