

Chronic Kidney Disease Prediction Using Federated Learning

Assistant Professor Mrs.Suje.S.A, Chinmaya.S, Harini.S

Department of AI and DS,
Nandha Engineering College (Affiliated to Anna University, Chennai) Erode, India

Abstract- Chronic kidney disease (CKD) is a global health challenge, affecting millions of individuals and often leading to kidney failure when not detected early. The application of machine learning (ML) for CKD prediction has gained significant attention, enabling timely diagnosis using clinical data. This paper explores various ML techniques used for CKD prediction, focusing on preprocessing challenges such as missing data, data imbalance, and feature selection. Additionally, the paper discusses the emerging role of Federated Learning (FL), a decentralised approach to ML that allows for privacy-preserving collaborative model training across institutions.

Index Terms- ChronicKidney Kidney Disease (CKD), Machine Learning (ML), Random Forest, AdaBoost Classification

I. INTRODUCTION

Chronic kidney disease (CKD) is a prevalent condition that, when left untreated, can progress to life-threatening stages. Traditional diagnostic approaches rely on biochemical tests such as serum creatinine measurements, which are costly and time-consuming. The integration of machine learning (ML) techniques into healthcare systems can assist in early CKD prediction based on routine clinical and demographic data.

The complexity of CKD data, which often suffers from missing values, imbalanced classes, and patient privacy concerns, necessitates the use of advanced techniques. Federated Learning (FL) provides a decentralised framework for training ML models across institutions without sharing sensitive patient data. Additionally, ensemble learning techniques such as Random Forest and XGBoost have proven effective in handling noisy, high-dimensional datasets, making them suitable for medical applications.

II. DATA PREPROCESSING AND CHALLENGES

1. Data Privacy and Federated Learning

One of the primary concerns in healthcare ML applications is patient data privacy. Regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) enforce strict data-sharing protocols, making it difficult to create centralized predictive models across multiple healthcare institutions. Federated Learning (FL) addresses this issue by enabling the decentralized training of ML models, allowing institutions to collaboratively improve a global model without exchanging

sensitive data. Figure 1 presents a diagram of the FL workflow, showing how local models are trained at each institution and aggregated into a global model at a central server.

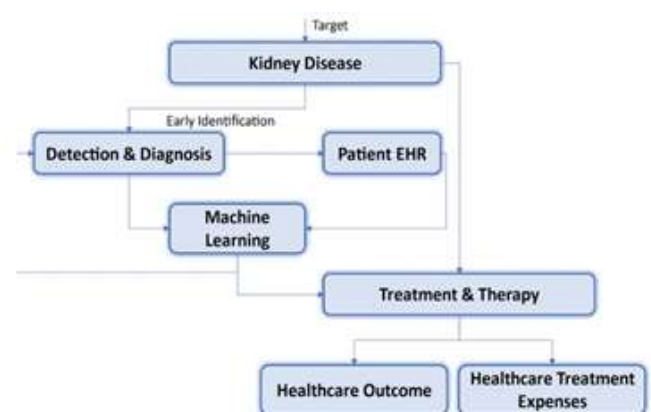


Figure 1: Federated Learning Workflow for CKD Prediction

2. Imbalanced Data

Imbalanced data can lead to biased models that fail to accurately predict CKD-positive cases. To address this, techniques such as the Synthetic Minority Oversampling Technique (SMOTE) and Random Undersampling (RUS) can be applied. Techniques balance the dataset by either oversampling the minority class (CKD patients) or undersampling the majority class (non-CKD patients).

3. Handling Missing Data

Medical datasets frequently contain missing values due to incomplete patient records or irregularities in data entry. Handling missing data in CKD prediction, XGBoost has proven particularly effective due to its ability to focus on

correcting and building reliable ML models. Techniques such as K-Nearest Neighbors (KNN) Imputation and Multiple Imputation by Chained Equations (MICE) are commonly used to fill missing values and maintain dataset integrity. Figure 2 illustrates the impact of different imputation techniques on dataset completeness.



Figure 2: Example of Data Distribution Before and After applying SMOTE

III. MACHINE LEARNING TECHNIQUES FOR CKD PREDICTION

1. Random Forest (RF)

Random Forest (RF) is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) from the individual trees. RF is robust to overfitting and performs well with noisy datasets. Its ability to handle missing data and high-dimensional feature spaces makes it an ideal choice for CKD prediction.

The hyperparameters used in the RF model are shown below, and Figure 3 compares the accuracy, precision, and recall of RF with other traditional models like Decision Trees (DT) and Logistic Regression (LR).

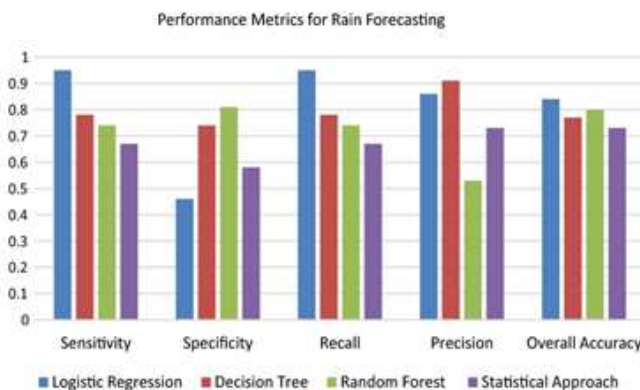


Figure 3: Performance Comparison of Random Forest, Decision Trees, and Logistic Regression

2. XGBoost

XGBoost, short for eXtreme Gradient Boosting, is a scalable and efficient version of gradient boosting that iteratively improves model performance by optimising for previously misclassified instance errors and handling missing data. When combined with data balancing techniques like SMOTE, XGBoost delivers highly accurate predictions.

3. Ensemble Learning

Ensemble learning methods combine the predictive power of multiple models to improve overall accuracy and robustness. Random Forest and XGBoost are both ensemble methods, but they approach the problem differently—Random Forest uses bagging (bootstrap aggregation), while XGBoost uses boosting. By combining these two approaches, ensemble learning can achieve superior results in CKD prediction. Figure 4 illustrates the overall accuracy improvements achieved through ensemble learning.

IV. MODEL EVALUATION AND PERFORMANCE METRICS

In this model, we applied several machine learning models to predict Chronic Kidney Disease (CKD) using a comprehensive dataset consisting of both categorical and numerical features. The models evaluated include K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Random Forest (RF), AdaBoost, Gradient Boosting (GB), XGBoost, CatBoost, and LightGBM (LGBM). The primary objective was to assess the performance of these models in accurately predicting CKD and to select the best-performing model based on several evaluation metrics.

1. Evaluation Metrics

To evaluate the models' performance, we used a set of standard classification metrics:

- **Accuracy:** Accuracy measures the overall rate of correct predictions but may be limited by potential class imbalances in CKD datasets.
- **Precision, Recall, and F1-Score:** Precision reflects the model's ability to avoid false positives, while recall measures its capacity to identify true CKD cases. The F1-Score balances these metrics, providing a single performance value, especially valuable for imbalanced datasets.
- **AUC-ROC:** Evaluates the model's ability to distinguish between CKD and non-CKD cases across thresholds.
- **Recall (sensitivity):** Measures the model's ability to identify actual CKD cases, with a high value indicating few false negatives.
- **F1-Score:** The harmonic mean of precision and recall, particularly useful for imbalanced datasets.

2. Accuracy

Accuracy is a fundamental metric that calculates the proportion of correctly predicted instances (both CKD and non-CKD) relative to the total number of predictions made. While useful, accuracy can be misleading in the presence of imbalanced datasets—common in CKD prediction where there are far more non-CKD cases than CKD-positive cases.

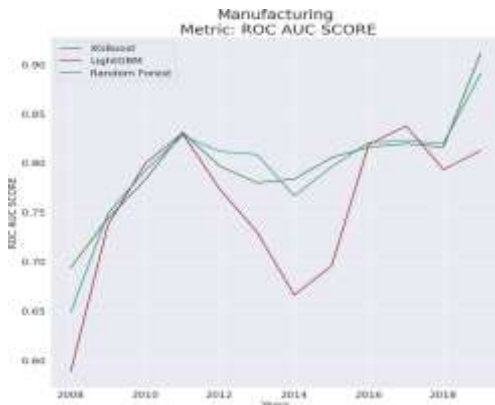


Figure 4: Comparison of XGBoost vs. Random Forest in Terms of AUC-ROC and F1-Score

3. Precision

Precision measures the ratio of true positive predictions (CKD cases) to the total predicted positive cases, reflecting the model's accuracy in identifying CKD cases without overpredicting.

High precision is crucial in clinical settings to ensure that the number of false positives is minimized, reducing unnecessary anxiety or treatment for patients misdiagnosed as having CKD.

4. Recall (Sensitivity)

Recall, also known as sensitivity or true positive rate, indicates how well the model can identify actual CKD cases from the dataset. A high recall value means that the model is effective at minimizing false negatives, which is critical for early CKD detection.

5. F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when dealing with imbalanced datasets, as it ensures that both false positives and false negatives are accounted for equally.

6. Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

The AUC-ROC curve is a graphical representation of a model's ability to distinguish between the positive (CKD) and negative (non-CKD) classes. The area under the curve (AUC)

provides a single metric to evaluate the model's performance across all classification thresholds. A perfect model would have an AUC of 1, whereas an AUC of 0.5 suggests a model with no discriminative ability.

AUC-ROC Curve: A curve plotting the true positive rate (recall) against the false positive rate (FPR) across various thresholds. The closer the curve is to the top-left corner, the better the model's performance.

V. CKD PREDICTION

Federated Learning Privacy Preserving

Federated Learning (FL) is a revolutionary approach that allows decentralized training of predictive models across different healthcare institutions without compromising data privacy. In this framework, each participating institution trains a local model on its patient data and then shares model updates (rather than the data itself) with a central server. The central server aggregates these updates to form a global model, which is then shared back with the participating institutions for further training.

This approach ensures that patient data never leaves the local institution, thereby preserving privacy while benefiting from the insights generated across multiple institutions.

Federated Learning Architecture

In traditional machine learning setups, all patient data must be consolidated into a single repository for model training. However, with Federated

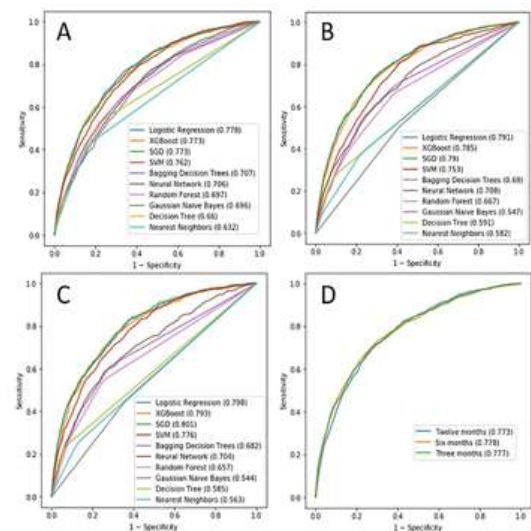


Figure 5 shows ROC curves comparing models like Random Forest and XGBoost and simpler models like Logistic Regression to illustrate how advanced ensemble models outperform in terms of AUC-ROC.

Learning local models are trained on datasets at each institution. These models do not exchange raw data; instead, they send model updates (such as weights and gradients) to a central server, where they are aggregated to form a global model. This global model is then distributed back to the participating institutions for further refinement.

In the context of healthcare, Federated Learning offers several advantages over traditional machine learning approaches. It facilitates collaborative model development across multiple institutions without violating patient privacy or requiring data transfer.

Benefits of Federated Learning in CKD Prediction

- **Data Security:** Since only model parameters are shared, there is a reduced risk of sensitive information leakage.
- **Data Privacy:** Patient data remains within the local institution, ensuring compliance with privacy laws while still allowing for collaborative model training.

Future Research Directions

Addressing Data Heterogeneity

In Federated Learning, each participating institution may have different data types, feature distributions, and levels of noise. This heterogeneity can impact the quality and accuracy of the aggregated global model. Future research should focus on developing methods that adapt to non-IID (non-independent and identically distributed) data, such as personalized Federated Learning models. These methods would adjust the global model to better fit the specific characteristics of local data.

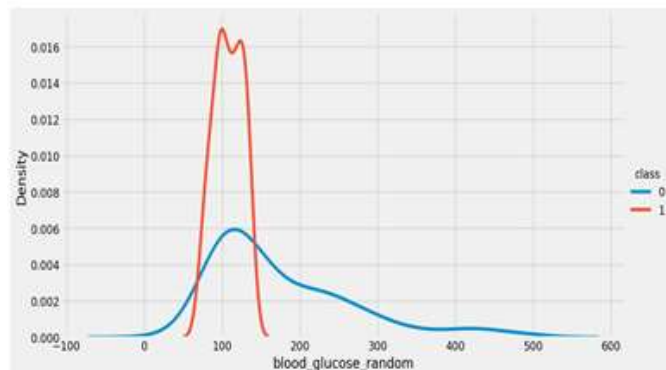


Figure 6: It shows a density plot comparing blood glucose levels for two different classes, likely related to chronic kidney disease (CKD) prediction.

Enhancing Model Interpretability

One of the main criticisms of ensemble learning methods, particularly Random Forest and XGBoost, is their lack of interpretability. In healthcare, it is essential that predictive

models not only perform well but also provide insights that clinicians can understand and trust. Future work should explore the development of explainable AI (XAI) techniques, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), to make the predictions from complex clinical settings to improve early CKD detection and patient outcomes.

Reducing Computational Costs in Federated Learning

Federated learning requires significant multi-model resources due to the need for continuous model updates between local institutions and the central aggregator. For real-world adoption in resource-constrained environments (e.g., smaller hospitals), future research should focus on reducing the computational and communication costs associated with FL.

Leveraging Transfer Learning for CKD Prediction

Transfer learning can be a valuable tool for improving CKD prediction models, particularly when data from one institution is scarce. By leveraging pre-trained models from other domains (e.g., cardiovascular disease prediction), CKD models can be fine-tuned with smaller CKD-specific datasets, thus reducing the need for large amounts of labeled data and accelerating the model-building process.

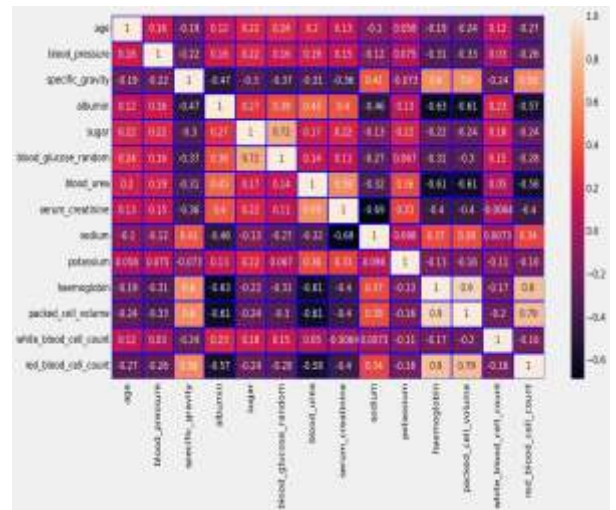


Figure 7: The image shows a correlation matrix of medical parameters for predicting CKD, highlighting the relationships between biomarkers and kidney function.

Integrating Multi-modal Data for CKD Prediction

Future research could explore the integration of data, combining clinical records, laboratory results, imaging data (e.g., renal ultrasounds), and genomic data. By leveraging these diverse data sources, more comprehensive and learning techniques make chronic kidney disease classification, predictive models accurate and generalizable, benefiting a wide range of healthcare institutions. We could develop accurate CKD prediction models by employing deep learning

(DL) architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to process multi-modal data, thereby enhancing prediction capabilities.

VI. CONCLUSION

The integration of Federated Learning, Ensemble Learning, Random Forest, and XGBoost creates a robust framework for CKD prediction. Federated learning allows institutions to collaborate privately, while ensemble methods improve prediction accuracy by combining multiple models.

This approach offers a promising path forward for integrating machine learning into healthcare, where secure, efficient, and accurate predictive models are essential. This framework uses privacy-preserving federated learning, allowing institutions to work together on improving CKD prediction without sharing sensitive patient data.

Additional federated learning-specific metrics, such as communication efficiency and data distribution stability, underscore the model's ability to operate effectively across distributed datasets without compromising predictive performance. This study showcases the potential of federated learning in healthcare, where data privacy and secure sharing are paramount, enabling institutions to collectively develop accurate models without centralizing sensitive data.

Thus, this project marks a promising step toward data-private, decentralized healthcare AI, enhancing early CKD diagnosis and patient management.

REFERENCES

1. S. Srivastava, R. K. Yadav, V. Narayan, and P. K. Mall, "An ensemble learning approach for Pharmaceutical Negative Results, vol. 10, pp. 2401–2409, Dec. 2022
2. S. Eddy, L. H. Mariani, and M. Kretzler, "Integrated multi-omics approaches to improve classification of chronic kidney disease," *Nature Rev. Nephrol.*, vol. 16, no. 11, pp. 657–668, Nov. 2020.
3. V. Giannouli and N. Syrmos, "Attitudes of younger and older adults towards kidney diseases in Greece," *Health Psychol. Res.*, vol. 7, no. 2, p. 8230, 2019.
4. Bilal Khan, Rashid Naseem, Fazal Muhammad, Ghulam Abbas, Sunghwan Kim, "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy," *IEEE Access* Volume: 8, 2020.
5. N. G. Ramadhan and A. Romadhony, "Preprocessing handling to enhance detection of type 2 diabetes mellitus based on random forest," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 7, pp. 223–228, 2021.
6. Shahinda Mohamed Mostafa Elkholy, Amira Rezk, Ahmed Abo El Fetoh Saleh on "Early Prediction of Chronic Kidney Disease Using Deep Belief Network," *IEEE Access*, Volume 9, Journal 2021.
7. Hamdi A. Al-Jamimi, "Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction," *IEEE Access*, Volume: 12 | Journal Article, 2024.
8. Pooja Yadav, S. C. Sharma, Shashikant P. Patole, and Rajesh Mahadeva, "Communicable Chronic Disease Using Stacking Classifier," *IEEE Access* 2023.
9. J. Jasmine Gabriel, L. Jani Anbarasi, Accurate Cardiovascular Disease Prediction: Leveraging Opt_hpLGBM With Dual-Tier Feature Selection, *IEEE Access*, 2024. *Computing Methodologies and Communication*
10. M. A. H. Saeed, "Diabetes type 2 classification using machine learning algorithms with up-sampling technique," *J. Electr. Syst. Inf. Technol.*, vol. 10, no. 1, pp. 1–10, Feb. 2023.
11. Y. Ren, H. Fei, X. Liang, D. Ji, and M. Cheng, "A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records," *BMC Med. Informat. Decis. Making*, vol. 19, no. S2, Apr. 2019.
12. Y. Khourdifi and M. Bahaj, "Heart disease diagnosis and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 242–252, 2019.
13. F. Pedregosa, G. Varoquaux, A. Gramfort et al. *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12:2825–2830, 2022.
14. Lisha Li, Kevin Jamieson, Afshin Rostamizadeh et al., *Massively parallel hyperparameter tuning*, 2020.
15. Abiodun Oguntimilehin, Martin Victor Inyang, Opani Aweh, and Oluwatoyin Abiola, "K-Nearest Neighbour-Based Chronic Kidney Disease Prediction System: A Case of Toxic Metals in Urine," 2024 International Conference on Science, Engineering, and Business for Driving Sustainable Development Goals (SEB4SDG), Conference Paper, *IEEE Access*, 2024.
16. Ritu Rani, Kanwarpartap Singh Gill, Deepak Upadhyay, and Swati Devliyal, "XGBoost-Driven Insights: Enhancing Chronic Kidney Disease Detection," 5th International Conference on Smart Electronics and Communication (ICOSEC), Conference Paper, *IEEE Access*, 2024.
17. Vishwanatha C R, V Asha, Arpana Prasad, Shyamal Das, Sunay Kumar, Sreeja S P, Support Vector Machine (SVM) and Artificial Neural Networks (ANN) based Chronic Kidney Disease Prediction", 2023, 7th International Conference on (ICCMC), Conference Paper, *IEEE Access*, 2023

18. Yixuan Zhao, "Comparison of Federated Learning Algorithms for Image Classification," 2nd International Conference on Data Analytics, Computing, and Artificial Intelligence (ICDACAI), 2023.
19. Pedro A. Moreno-Sánchez, Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and Evaluation of an Explainable AI Model, IEEE Access, Volume 11, 20.
20. L. J. Rubini and E. Perumal, "Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm," Int. J. Imag. Syst. Technol., vol. 30, no. 3, pp. 660–673, Sep. 2020.
21. Abdulkadir Korkmaz, Ahmad Alhonainy, and Praveen Rao, "An Evaluation of Federated Learning Techniques for Secure and Privacy-Preserving Machine Learning on Medical Datasets, IEEE Access, 2022.