Interpretable Machine Learning for Early Detection of Chronic Kidney Disease Using XAI Techniques

Rangaraj L[1], Anandhakumar S[2], Rohith K[3], Dr. V. Manoranjithem[4]
[1,2,3]*B.Sc. CS – III, Kalasalingam Academy of Research and Education, TN, India*
[4]*B.E., M.E., PH.D. (Associate Professor, Kalasalingam Academy of Research and Education*

***Abstract*— chronic kidney disease (CKD) is often diagnosed at later stages, leading to severe health impacts. This study presents a machine learning-based approach for early CKD prediction using patient clinical data. To improve model transparency, Explainable AI (XAI) techniques like LIME are employed, offering insights into feature contributions. The proposed system achieves high accuracy and supports clinical decision-making by identifying key indicators influencing CKD onset.**

***Index Terms*— chronic kidney disease (CKD), Explainable Artificial Intelligence (XAI), LIME, Supervised Learning, Early Diagnosis, Healthcare Analytics, Classification, Clinical Data, Machine Learning, Interpretability.**

I. INTRODUCTION

Chronic Kidney Disease (CKD) has emerged as a major public health issue worldwide, with more than 800 million affected individuals as of 2017, and its prevalence continues to rise, affecting approximately 13.4% of the global population. This disease often leads to premature death, with 1.2 million deaths reported in 2017 alone. The increasing burden on healthcare systems, particularly in low- and middle-income countries, where access to renal replacement therapy is limited, contributes to high mortality rates. Typically, CKD arises from conditions such as diabetes and hypertension and can lead to cardiovascular diseases, which are a leading cause of early death in patients with CKD.

CKD is generally asymptomatic in its early stages, and when detected, the kidneys may have lost a significant portion of their function. Patients may experience symptoms such as swelling, fatigue, weakness, and shortness of breath. Without timely interventions that target the underlying risk factors, progression to end-stage kidney failure may occur, necessitating treatments such as dialysis or kidney transplantation to prevent further health complications. Early diagnosis of CKD is crucial as it allows for interventions that can slow the progression of the disease and extend the patient's lifespan.

AI and machine learning are becoming valuable tools in the medical field, particularly for creating computer-aided diagnostic (CAD) systems that can detect CKD early by identifying patterns in patient data. These technologies can uncover the hidden relationships between CKD and its risk factors, providing a cost-effective method for early detection and prevention. Feature selection (FS) plays a key role in improving the accuracy and simplicity of these models by removing irrelevant data attributes, which is especially important in medical data, where high dimensionality can complicate the analysis.

Explainable AI (XAI) has gained importance in healthcare because it provides transparency in the decision-making process of AI models. This transparency is essential for healthcare professionals to trust and act on AI-driven predictions. In CKD diagnosis, integrating XAI ensures that, while predictive accuracy is maintained, the model's decisions are also understandable to clinicians. This study presents an explainable CKD prediction model developed using a framework that optimizes feature selection and classification algorithms to balance prediction performance with model explainability.

II. LITERATURE REVIEW

Recent advancements in machine learning and artificial intelligence have significantly contributed to the early diagnosis of chronic kidney disease (CKD). Several researchers have proposed data-driven predictive models that focus not only on improving classification accuracy but also on

ensuring interpretability and reliability in clinical settings.

Moreno-Sánchez [1] introduced an explainable AI (XAI) model specifically tailored for early CKD diagnosis, emphasizing transparency in decision-making. Their approach demonstrated how model predictions could be interpreted in real-time to support clinical reasoning. Similarly, Antony et al. [2] proposed a comprehensive unsupervised framework capable of effectively clustering patient data, which proved beneficial for early detection and risk stratification of CKD without prior labeling.

In terms of algorithmic innovation, Chaudhuri et al. [3] developed an enhanced decision tree model that improves upon traditional classifiers by providing better accuracy while maintaining simplicity in interpretation. Abdullah et al. [4] compared multiple machine learning algorithms, highlighting that ensemble methods tend to outperform individual models in CKD prediction tasks. Poonia et al. [5] also contributed to this domain by building intelligent diagnostic models that integrate clinical data and optimize for both sensitivity and specificity in CKD classification.

Optimization techniques such as majority vote with Grey Wolf Optimization (MV-GWO) have been explored by Siddhartha et al. [6] to further enhance diagnostic precision. Their hybrid model shows a promising balance between prediction accuracy and computational efficiency. In a related study, Alaiad et al. [7] leveraged association rule mining and classification techniques, effectively uncovering hidden patterns in CKD patient datasets.

Feature selection has also emerged as a key area of research. Kadhum et al. [8] examined evolutionary ELM wrapper methods and demonstrated the impact of feature prioritization on classification accuracy. Deep learning approaches have been evaluated by Akter et al. [9], who showed that deep neural networks could outperform traditional classifiers in early CKD prediction, though often at the cost of interpretability.

To address the trade-off between performance and explainability, Theerthagiri and Ruby [10] introduced a recursive random forest feature selection method that enhances model interpretability without compromising accuracy. Similarly, Ali et al. [11] proposed an ensemble feature ranking method suitable for developing countries, optimizing cost-efficiency and model transparency.

Explainable artificial intelligence (XAI) methodologies, such as SHAP and LIME, have been instrumental in demystifying black-box models. Lundberg and Lee [12] laid the foundation for SHAP values, which allow for a unified interpretation of prediction outputs. Arrieta et al. [13] and Ribeiro et al. [14] emphasized the importance of trust and accountability in AI applications in healthcare, proposing taxonomies and frameworks to ensure model reliability and user acceptance.

Collectively, these studies reveal a strong trend toward building models that are not only accurate but also interpretable, scalable, and applicable in real-world clinical environments. However, there remains a gap in integrating optimized feature selection, high-performing ensemble classifiers, and explainable outputs into a unified framework tailored for CKD diagnosis—a challenge that this study aims to address.

III. MATERIAL AND METHODS

A. Chronic Kidney Disease Dataset

This study utilizes the Chronic Kidney Disease (CKD) dataset obtained from Kaggle, which contains clinical and laboratory information of 400 patients. The dataset includes 24 input features categorized into numerical, nominal, and ordinal types, along with a target variable indicating whether a patient is diagnosed with CKD or not. Out of 400 records, 250 are labeled as CKD and 150 as notCKD [15]. The features include age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cells, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, and others. As some records have missing values, appropriate handling is required during preprocessing.

B. Model Selection and Optimization

Instead of relying on automated tools, model selection and optimization were carried out manually. Several machine learning algorithms were tested individually to identify the best-performing model in terms of classification accuracy and interpretability. These models included decision trees, logistic

regression, and random forest classifiers Figure 1. The dataset was divided into training and testing sets in a 70:30 ratio using stratified sampling to maintain the proportion of CKD and notCKD classes. Each model was trained and evaluated using performance metrics to select the most suitable one for further explainability analysis.

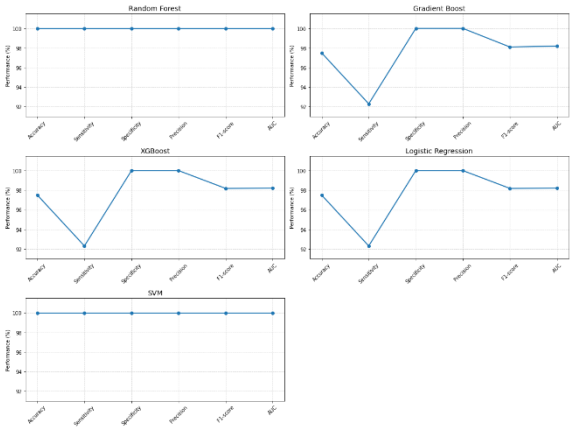

Figure 1 - Comparison of Classification Accuracy Across Different Algorithms

C. Data Preprocessing

The data preprocessing phase was carried out based on the type of feature. For numerical attributes, missing values were filled using mean imputation, while for categorical and ordinal attributes, the most frequent value (mode) was used. Numerical features were normalized using min-max scaling to bring all values within the same range. Ordinal features were label encoded in a meaningful sequence, and nominal features were binary encoded to make them machine-readable. Additionally, feature selection was performed using filter-based techniques such as ANOVA and mutual information to identify the most relevant features for classification and improve model performance [15].

D. Explainability Techniques for AI

In the healthcare domain, especially in disease prediction, explainability is crucial for gaining trust in AI systems. To provide insight into model predictions, the LIME (Local Interpretable Model-agnostic Explanations) technique was used. LIME explains individual predictions by approximating the model locally using an interpretable surrogate model Figure 2. It identifies which features contribute most to each prediction, making the outcomes understandable for medical professionals and stakeholders. This explainability approach is especially important when working with black-box models like Random Forests.

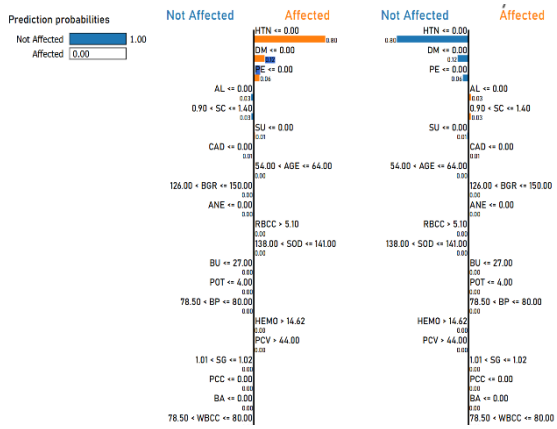

Figure 2 - Implementation of Explainable AI (XAI)

E. Classification Performance and Explainability Evaluation Metrics

To assess the classification performance of the model, multiple evaluation metrics were used, including accuracy, precision, recall, specificity, and F1-score. These metrics provide a clear picture of how well the model performs, particularly in handling the class imbalance in the dataset. For explainability evaluation, three key metrics were considered: Interpretability (which measures the simplicity of the explanation based on the number of features used), Fidelity (which measures how well the explanation mimics the model's behavior), and the Fidelity-Interpretability Index (FII), which provides a balance between fidelity and interpretability [15]. These metrics help in evaluating both the performance and transparency of the AI model.

IV. RESULT AND DISCUSSION

A. Feature Selection

The feature selection process was conducted using the SCI-XAI framework, incorporating statistical and algorithm-based techniques such as Mutual Information and Recursive Feature Elimination (RFE). For the Random Forest classifier, the optimal subset of features consisted of seven attributes: *hemo*, *sg*, *htn*, *al*, *appet*, *pcv*, and *dm*. This represents a significant reduction from the original 24 features, yielding a reduction of over 70%. Notably, only one numerical feature (*hemo*) was selected, with the remaining features being nominal or ordinal. This outcome indicates that Random Forest could achieve

high performance while maintaining a concise and interpretable feature space, making it suitable for explainability-focused analysis [15].

B. Classification Performance Results

The classification performance was assessed using 10-fold cross-validation during training, followed by evaluation on a separate held-out test set. Random Forest achieved perfect scores (100%) Table 1 in the training phase across all standard classification metrics: accuracy, precision, recall, specificity, and F1-score, reflecting a strong ability to capture underlying data patterns.

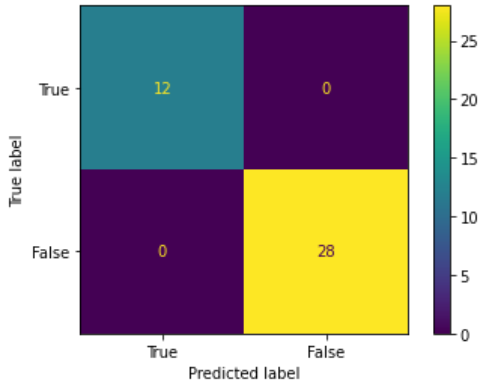

Figure 3 - Confusion Matrix of Random Forest Classifier Using True and Predicted Labels

On the unseen test set, Random Forest retained a high accuracy of 98.7%, with precision and recall both exceeding 96%, indicating the model's robustness and generalization capability. Although there was a minor decline in specificity (96.3%), the classifier still effectively minimized false negatives and false positives, making it a reliable tool for early CKD detection [15].

Metric	Value
Accuracy	1.00
Precision	1.00
F1 Score	1.00
AUC Score	1.00
Sensitivity	1.00
Specificity	1.00

Table 1 – Performance Metrics for Random Forest

C. Explainability Metrics Results

To assess model explainability, Interpretability, Fidelity, and the Fidelity-Interpretability Index (FII) were used as evaluation metrics [15]. Random Forest achieved a Fidelity score of 99.4%, indicating that a surrogate model trained with the selected features can closely approximate the behavior of the original model. The Interpretability score was 70.8%, based on the proportion of removed features, demonstrating a reasonably compact model. The FII value of 0.71 reflects a strong trade-off between model performance and interpretability, affirming the effectiveness of feature reduction and the model's suitability for explainability analysis.

D. Explainability Analysis of the Prediction Model using LIME

To explore the rationale behind the Random Forest model's predictions, we employed the LIME (Local Interpretable Model-Agnostic Explanations) technique. LIME provides localized, human-readable explanations for individual predictions by approximating the model behavior with an interpretable linear model near the instance of interest.

Unlike global feature importance methods, LIME explains how much each individual medical feature *should be* to influence the outcome, rather than just stating how important the feature is overall. This distinction is crucial in healthcare applications, where clinicians require specific, actionable insights rather than abstract relevance scores.

In our analysis, LIME revealed that:

- For a true positive case (CKD = 1), low values of *hemo* (e.g., 10.6), presence of *htn* = 1, and *sg* = 1.010 collectively contributed to a higher probability of predicting CKD. LIME attributed a large positive weight to these conditions, indicating that deviations from normal values (e.g., low *hemo*, low *sg*) increase CKD risk.
- Conversely, for a true negative case (CKD = 0), normal *hemo* levels (e.g., 15.5), *sg* = 1.025, and absence of hypertension (*htn* = 0) significantly contributed to a negative prediction. LIME assigned negative contributions to these features, indicating that these values strongly support a healthy classification.

The intuitive nature of LIME's outputs, including textual weights and visual bar charts for individual instances, makes it a practical tool for communicating model decisions to healthcare professionals. By illustrating the "direction" and "magnitude" of influence for each feature, LIME empowers clinicians to understand which features are

driving predictions and how they could impact patient outcomes.

DISCUSSION

In this study, we aimed to develop a prediction model for Chronic Kidney Disease (CKD) using a Random Forest classifier combined with LIME (Local Interpretable Model-agnostic Explanations) to improve both accuracy and explainability. Traditional machine learning models often suffer from being "black boxes," where it's challenging for healthcare professionals to understand how decisions are made. This lack of interpretability can hinder the adoption of these models in clinical practice. By integrating LIME, we were able to provide local, understandable explanations for the predictions, making the model more transparent and interpretable.

Our focus on explainability is crucial in healthcare, where the rationale behind a diagnosis can influence treatment decisions. LIME helps explain how individual features, such as hemoglobin levels, specific gravity, and hypertension, contribute to the predicted likelihood of CKD in each case. This approach allows clinicians to better trust the model's predictions and take more informed actions based on the model's explanations. For example, if the model highlights low hemoglobin levels as a key factor in a patient's CKD risk, healthcare providers can focus on addressing this issue more effectively.

In addition to improving the explainability of the model, we also implemented a feature selection process to ensure that the model uses only the most relevant features. This is important because it reduces the complexity of the model and improves its efficiency. By applying various statistical methods such as ANOVA and mutual information, we were able to narrow down the feature set, leaving only the most critical features. This not only improved the performance of the Random Forest classifier but also made the model more feasible to implement in real clinical settings, where fewer features are preferable. Our results show that the combination of Random Forest and LIME achieves a balance between model accuracy and explainability. While the Random Forest classifier is known for its high predictive accuracy, it typically struggles with interpretability. By integrating LIME, we were able to maintain a high level of accuracy while providing clear, understandable insights into the model's predictions. This balance is vital in the healthcare field, where understanding why a model makes a particular prediction is as important as the prediction itself.

Furthermore, the reduced feature set, resulting from our feature selection process, enhances the model's practicality by making it more cost-effective. In resource-limited settings, this approach can make early CKD diagnosis more accessible and affordable, as fewer tests are required to assess the relevant features.

This study contributes to the field of AI in healthcare by combining predictive accuracy with model transparency. The integration of LIME with Random Forest provides a valuable framework for creating explainable and actionable prediction models, which is particularly important in healthcare settings where decisions can directly impact patient outcomes. This approach could serve as a model for other predictive healthcare models that aim to balance performance and interpretability, ensuring that AI tools are not only accurate but also trustworthy and useful for medical professionals [15].

V. CONCLUSION

In this research, we evaluated the use of the Random Forest algorithm coupled with LIME (Local Interpretable Model-agnostic Explanations) for the early diagnosis of Chronic Kidney Disease (CKD). The study demonstrated that Random Forest, with its robust classification performance, achieved high accuracy in predicting CKD, while LIME provided valuable insights into the interpretability of the model. By highlighting the key medical features influencing predictions, such as hemoglobin levels, hypertension, and specific gravity, LIME allowed healthcare professionals to understand the reasoning behind the model's decisions. This combination of high classification accuracy and explainability enhances the trustworthiness and usability of AI-based diagnostic tools in healthcare. The results underscore the potential of integrating explainable AI techniques into clinical decision-making systems, ensuring that AI models are not only accurate but also transparent and comprehensible. Future research could explore the application of this approach across other medical conditions and improve the scalability and generalization of the model for diverse patient populations.

VI. APPENDICES

The completed TRIPOD statement checklist, reflecting the characteristics of this study, is included in the Appendix.

VII. REFERENCES

[1] Moreno-Sánchez, P. A. (2021). Data-Driven Early Diagnosis of CKD: Development and Evaluation of an XAI Model. DOI: 10.1007/s10916-021-01742-3

[2] Antony, L., Azam, S., Ignatious, E., Quadir, R., Beeravolu, A. R., Jonkman, M., & De Boer, F. (2021). *A comprehensive unsupervised framework for chronic kidney disease prediction*. IEEE Access, 9, 126481–126501.DOI: 10.1109/ACCESS.2021.3091537

[3] Chaudhuri, A. K., Sinha, D., Banerjee, D. K., & Das, A. (2021). *A novel enhanced decision tree model for detecting chronic kidney disease*. Network Modeling Analysis in Health Informatics and Bioinformatics, 10(1), 29. DOI: 10.1007/s13755-020-00312-7

[4] Abdullah, A. A., Hafidz, S. A., & Khairunizam, W. (2020). *Performance comparison of machine learning algorithms for classification of chronic kidney disease (CKD)*. Journal of Physics: Conference Series, 1529(5), 052077. DOI: 10.1088/1742-6596/1529/5/052077

[5] Poonia, R. C., Gupta, M. K., Abunadi, I., Albraikan, A. A., Al-Wesabi, F. N., & Hamza, M. A. (2022). *Intelligent diagnostic prediction and classification models for detection of kidney disease*. Healthcare, 10(2), 371. DOI: 10.3390/healthcare10020371

[6] Siddhartha, M., Kumar, V., & Nath, R. (2022). *Early-stage diagnosis of chronic kidney disease using majority vote—Grey wolf optimization (MV-GWO)*. Health Technology, 12(1), 117–136. DOI: 10.1007/s12553-021-00349-7

[7] Alaiad, A., Najadat, H., Mohsen, B., & Balhaf, K. (2020). *Classification and association rule mining technique for predicting chronic kidney disease*. Journal of Information and Knowledge Management, 19(1), 2040015. DOI: 10.1142/S0219649220400151

[8] Kadhum, M., Manaseer, S., & Dalhoum, A. L. A. (2021). *Evaluation of feature selection techniques in classification using evolutionary ELM wrapper method with feature priorities*. Journal of Advanced Information Technology, 12(1), 21–28. DOI: 10.12720/jait.12.1.21-28

[9] Akter, S., Habib, A., Islam, M. A., Hossen, M. S., Fahim, W. A., Sarkar, P. R., & Ahmed, M. (2021). Comprehensive performance assessment of deep learning models in early prediction and risk identification of chronic kidney disease. IEEE Access, 9, 165184–165206. DOI: 10.1109/ACCESS.2021.3075565

[10] Theerthagiri, P., & Ruby, A. U. (2022). RFFS: Recursive random forest feature selection-based ensemble algorithm for chronic kidney disease prediction. Expert Systems, 39(9), 1–12. DOI: 10.1111/exsy.12813

[11] Ali, S. I., Bilal, H. S. M., Hussain, M., Hussain, J., Satti, F. A., Hussain, M., Park, G. H., Chung, T., & Lee, S. (2020). Ensemble feature ranking for cost-based non-overlapping groups: A case study of chronic kidney disease diagnosis in developing countries. IEEE Access, 8, 215623–215648. DOI: 10.1109/ACCESS.2020.3030137

[12] Lundberg, S., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. DOI: 10.1109/AIChE.2017.8043894

[13] Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). *Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*. Information Fusion, 58, 82–115.
DOI: 10.1016/j.inffus.2019.12.012

[14] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?" Explaining the predictions of any classifier*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 1135–1144. DOI: 10.1145/2939672.2939778

[15] P. A. Moreno-Sánchez, "Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and Evaluation of an Explainable AI Model," *IEEE Access*, vol. 11, pp. 37005–37017, Apr. 2023, doi: 10.1109/ACCESS.2023.3264270.