

An Interpretable Machine Learning Framework: Improving Diagnostic Efficiency In Early Chronic Disease Detection

Sam Jackson S^{1*}, Christina Magneta S²

¹dept. Artificial Intelligence and Machine Learning Karunya Institute of Technology And Sciences
Coimbatore, India samodc04@gmail.com.

²dept. Artificial Intelligence and Machine Learning Karunya Institute of Technology And Sciences Coimbatore, India
christinarvs@gmail.com

Abstract

Early identification of chronic diseases continues to be a major challenge in healthcare, particularly when conditions progress silently during their initial stages. Chronic Kidney Disease (CKD) exemplifies this issue, as declining renal function often remains clinically unnoticed until the disease reaches an advanced and less manageable phase. Delayed diagnosis significantly restricts treatment options and increases the likelihood of irreversible complications, highlighting the need for reliable and clinically meaningful decision-support systems. This study presents an interpretable machine learning-based diagnostic framework aimed at enhancing early detection of chronic diseases, with a focused application to CKD. The proposed framework follows a systematic and reproducible learning pipeline that combines comprehensive data preprocessing, ensemble learning, and model interpretability. A Random Forest classifier is adopted as the primary predictive model due to its ability to effectively manage heterogeneous clinical features and capture non-linear relationships within medical data. To mitigate the class imbalance commonly encountered in clinical datasets, the Synthetic Minority Oversampling Technique (SMOTE) is incorporated during the training process. Model performance is optimized using stratified cross-validation and evaluated through clinically relevant metrics to ensure diagnostic reliability beyond overall accuracy. To promote transparency and foster clinical confidence, SHapley Additive exPlanations (SHAP) are employed to generate both global and patient-specific interpretations of model predictions. These explanations provide insight into feature importance and decision rationale, enabling alignment between model outcomes and established clinical understanding. The results demonstrate that the proposed framework offers a balanced combination of predictive performance and interpretability, making it a promising tool for early CKD diagnosis and broader clinical decision support applications.

Keywords - Chronic Kidney Disease; early disease detection; interpretable machine learning; explainable artificial intelligence; Random Forest classification; clinical decision support systems; healthcare analytics

How to cite this article: Sam Jackson S, Christina Magneta S. An Interpretable Machine Learning Framework: Improving Diagnostic Efficiency In Early Chronic Disease Detection. *Int J Drug Deliv Technol.* 2026;16(16s): 279-284; DOI: 10.25258/ijddt.16.16s.30

I. INTRODUCTION

Chronic diseases continue to pose a substantial burden on modern healthcare systems due to their prolonged progression, high treatment costs, and significant impact on patient quality of life. Among these conditions, Chronic Kidney Disease (CKD) is particularly critical, as it often develops gradually and remains asymptomatic during its early stages. Consequently, a large proportion of patients are diagnosed only after considerable and irreversible renal damage has already occurred. Early detection of CKD is therefore essential to enable timely clinical intervention, slow disease progression, and reduce the risk of severe complications. In routine clinical practice, CKD diagnosis primarily relies on laboratory-based indicators such as serum creatinine concentration, blood urea levels, and estimated glomerular filtration rate (eGFR). While these parameters are clinically validated and widely used, their interpretation is highly dependent on physician expertise and clinical context. Variations in diagnostic experience, combined with increasing patient volumes, can lead to delayed or inconsistent diagnosis across healthcare settings. Furthermore, large-scale population screening using conventional diagnostic approaches is often time-consuming and resource-

intensive, particularly in regions with limited medical infrastructure. ML algorithms can predict disease onset, classify patient risk levels, and support personalized treatment plans by learning from historical health records, wearable sensors, and genetic information [2,3]. Recent advances in machine learning (ML) have demonstrated significant potential in supporting early disease detection by identifying complex and non-linear patterns within clinical data that may not be apparent through traditional statistical methods. ML-based diagnostic models have shown promising predictive performance across various healthcare applications, including chronic disease analysis. However, despite their effectiveness, many of these models function as black-box systems[5], offering limited insight into how predictions are generated. This lack of transparency presents a major challenge for clinical adoption, where interpretability, trust, and accountability are essential for decision-making. Recent studies have illustrated the effectiveness of ML algorithms in predicting various disorders, including polycystic ovary syndrome (PCOS), heart disease, thyroid dysfunction, and arrhythmias [4–7]. Manju et al[1] utilized SHAP for evaluating a decision tree-based EAI (DT-EAI) approach aimed at improving diagnostic

*Author for Correspondence: samodc04@gmail.com

efficiency in early chronic disease detection, with a specific focus on CKD. The framework integrates ensemble-based learning with explainable artificial intelligence techniques to achieve a balance between predictive accuracy and model transparency. Advanced models such as Mamba capsule routing [8], Transformers [9], and Capsule Networks [10] have an improved capacity to simulate intricate linkages and temporal patterns, whilst CNNs and other approaches can capture spatial information. In particular, SHapley Additive exPlanations (SHAP) are employed to enhance interpretability by quantifying the contribution of individual clinical features to model predictions at both global and instance levels. This enables clinicians to better understand model behavior and assess whether predictive outcomes align with established medical knowledge. By emphasizing interpretability alongside performance, the proposed framework seeks to bridge the gap between advanced ML methodologies and practical clinical deployment. This conceptual focus motivates the selection of explainable techniques within the framework and supports the development of trustworthy, data-driven decision-support systems for early CKD diagnosis..

II. METHODOLOGY

A. Dataset Compilation Strategy

The effectiveness of any clinical decision-support framework is strongly dependent on the quality and preparation of the underlying data. In this study, a publicly available Chronic Kidney Disease (CKD) dataset was used, containing a combination of demographic variables, biochemical test results, and physiological measurements commonly employed in renal function assessment. The dataset reflects real-world clinical characteristics, including heterogeneous feature types, missing entries, and class imbalance, making it suitable for evaluating the robustness of diagnostic models under practical conditions. Table 1 summarizes the key characteristics of the CKD dataset used in this study, including feature composition, class distribution, and the presence of missing values. During the data curation stage, non-standard representations of missing values were first identified and standardized by replacing them with null entries to ensure uniformity across all attributes. Records lacking valid diagnostic labels were removed to maintain the integrity of the target outcome.

Attribute Category	Description
Dataset source	Public CKD dataset (Kaggle)
Total instances	200
CKD-positive cases	128
CKD-negative cases	72
Numerical features	Serum creatinine, blood urea, eGFR, etc.
Categorical features	Hypertension, diabetes, appetite, etc.
Missing values	Present
Class imbalance	Yes

Table 1. Summary of the CKD Dataset

The disease status variable was then encoded into a binary format, representing CKD-positive and CKD-negative instances. Feature-level preprocessing was subsequently performed to construct a reliable and clinically meaningful input space. Attributes containing only missing values were excluded, as they did not contribute diagnostically relevant information. For numerical features, missing values were imputed using the median to reduce sensitivity to extreme values and preserve the underlying distribution of clinical measurements. Categorical attributes were imputed using the most frequent category and encoded numerically to enable compatibility with machine learning algorithms. In addition to preprocessing, exploratory analysis was conducted to examine the distribution of key clinical variables and to assess the degree of class imbalance within the dataset. This analysis informed subsequent modeling decisions, particularly the need for imbalance mitigation strategies during training. Overall, the dataset preparation process was designed to balance data completeness, clinical relevance, and statistical integrity,

providing a robust foundation for the proposed diagnostic framework.

B. Algorithmic Framework Selection

The choice of the learning algorithm was guided by three primary considerations: predictive reliability, interpretability, and computational efficiency. In clinical diagnostic settings, models must not only achieve strong performance but also remain stable in the presence of noisy and correlated variables, which are common in real-world medical data. Based on these requirements, a Random Forest classifier was selected as the core predictive model. Random Forest is an ensemble-based approach that combines the outputs of multiple decision trees trained on bootstrap samples of the data. This aggregation mechanism enables the model to capture complex and non-linear relationships among clinical attributes while reducing the risk of overfitting associated with single-tree models. Additionally, Random Forest demonstrates robustness to multicollinearity and missing-value imputation effects, making it well-suited for

structured healthcare datasets. From an interpretability perspective, tree-based ensemble models offer a natural advantage when paired with post-hoc explainability methods. Feature importance measures and local explanation techniques can be applied effectively to Random Forest models, allowing clinicians to examine how individual attributes influence predictions. This compatibility with explainable artificial intelligence techniques makes Random Forest particularly appropriate for healthcare applications, where transparency and trust are critical for clinical adoption. Debal and Sitote [12] also show how LR and recursive feature elimination (RFE) feature selection methods can predict CKD with an accuracy of 99.80% and 82.56% for both binary and multiclass respectively.

C. Custom Model Architecture Design

The proposed model architecture was configured with careful consideration of both predictive capability and interpretability. In clinical applications, overly complex decision structures can hinder generalization and reduce clinician confidence in model outputs. To address this, key hyperparameters of the Random Forest model including maximum tree depth, minimum samples required for node splitting, and minimum leaf size were deliberately constrained. These settings limit excessive model complexity while maintaining sufficient expressive power to capture relevant clinical patterns. Feature subsampling was incorporated during tree construction to promote diversity among individual learners within the ensemble. This mechanism enhances robustness and stability by reducing correlation between trees, thereby improving overall model generalization. Such design choices are particularly important when working with heterogeneous clinical features that may exhibit varying degrees of correlation. Class imbalance, a common challenge in medical diagnosis datasets, was explicitly addressed using the Synthetic Minority Oversampling Technique (SMOTE). By generating synthetic instances of the minority class within the training data, SMOTE improves the model's sensitivity to CKD cases without altering the original distribution of the test set. Banu and Begum [11] combined with KNN imputation predicted CKD with 99.75% accuracy. To ensure an unbiased evaluation, oversampling was strictly applied only to the training subset, thereby preventing information leakage. In addition to architectural design, the classification decision threshold was adjusted to prioritize recall over precision. This choice reflects clinical priorities, where failing to identify a CKD-positive case may delay intervention and lead to adverse patient outcomes. To further examine the classification behavior of the proposed framework, a confusion matrix was generated on the test dataset. The confusion matrix provides a detailed breakdown of correctly and incorrectly classified instances, enabling systematic analysis of false positives and false negatives, which are critical considerations in clinical screening contexts.

D. Optimization Strategies and Continuous Learning

Model optimization was directed toward improving diagnostic reliability while maintaining reasonable computational efficiency. Stratified k-fold cross-

validation was employed to obtain stable and unbiased performance estimates, ensuring that the class distribution was preserved across all folds. This strategy is particularly important in imbalanced clinical datasets, where conventional validation schemes may produce misleading results. During model tuning, evaluation objectives prioritized recall and F1-score rather than overall accuracy. This emphasis aligns with clinical screening requirements, where the ability to correctly identify CKD-positive cases is more critical than maximizing aggregate classification performance. By focusing on these metrics, the optimization process was guided toward minimizing false-negative predictions, which carry significant clinical risk. The proposed framework was designed with modularity to support continuous learning and future adaptation. As new clinical data become available, the model can be retrained or updated without altering the core pipeline structure. This capability is essential in real-world healthcare environments, where patient demographics, clinical practices, and diagnostic patterns may evolve over time. Such adaptability ensures that the framework remains robust, scalable, and clinically relevant during extended deployment.

E. Validation and Generalization

Comprehensive validation was performed using an independent test set to assess the generalization capability of the proposed framework. Model performance was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). This multi-metric evaluation provides a balanced assessment of diagnostic effectiveness while capturing trade-offs between sensitivity and specificity, which are critical in clinical screening tasks. To enhance transparency and support clinical interpretability, SHapley Additive exPlanations (SHAP) were integrated into the validation framework. SHAP analysis enables both global and instance-level interpretation of model predictions by quantifying the contribution of individual clinical features to classification outcomes. Global explanations highlight the most influential attributes across the dataset, while instance-level explanations provide insight into patient-specific predictions. By explicitly linking predictive outcomes to clinically meaningful indicators, the proposed framework facilitates informed decision-making and strengthens clinician trust in model outputs. This combination of rigorous validation and explainability supports the safe and effective application of the framework in real-world CKD screening scenarios.

III. RESULTS AND DISCUSSION

The performance of the proposed interpretable machine learning framework was evaluated using standard classification metrics to assess its effectiveness in the early detection of Chronic Kidney Disease (CKD). Given the clinical context, greater emphasis was placed on recall and F1-score, as failure to identify affected individuals may delay intervention and lead to adverse health outcomes. Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false

negatives, respectively. Based on these quantities, evaluation metrics were computed using standard definitions. The model achieved a high overall accuracy, indicating strong predictive capability across the dataset. However, accuracy alone is not sufficient for assessing medical diagnostic systems, particularly in the presence of class imbalance. Therefore, additional performance measures were examined to provide a more clinically meaningful evaluation. A precision score of 100% indicates that all instances predicted as CKD-positive were correctly classified, reflecting the absence of false-positive predictions. From a clinical perspective, this reduces the likelihood of unnecessary follow-up tests and

minimizes patient anxiety caused by incorrect diagnoses. The recall value of 84.62% demonstrates that the majority of actual CKD cases were successfully identified by the model. This outcome is especially important for screening applications, where missed diagnoses represent a significant clinical risk. The use of class imbalance handling techniques and probability threshold adjustment contributed to improved sensitivity by reducing false-negative predictions. The F1-score of 91.67% reflects a strong balance between precision and recall, indicating that the framework maintains reliable classification performance while effectively identifying disease cases.

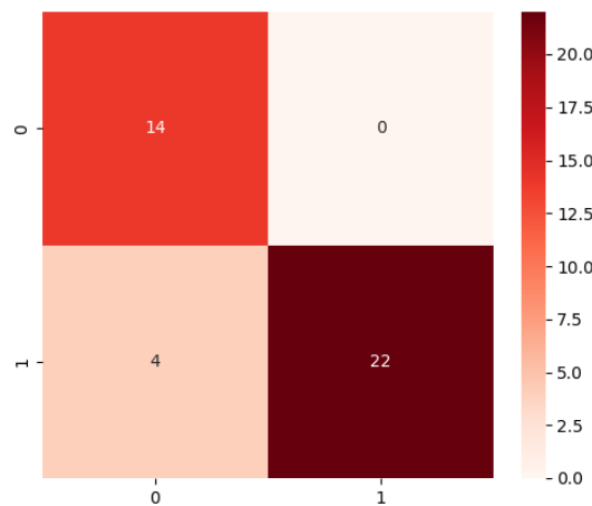


Fig. 1 Confusion matrix illustrating the classification performance of the proposed Random Forest-based framework for CKD detection.

As shown in Fig. 1, the model correctly classified 22 CKD-positive cases and 14 CKD-negative cases, while misclassifying four CKD-positive instances as non-CKD. Notably, no false-positive predictions were observed, indicating high precision. The limited number of false negatives reflects the effectiveness of the recall-oriented optimization strategy, which is critical in clinical screening scenarios where missed CKD diagnoses may delay timely intervention. Achieving this balance is essential in healthcare applications, as both excessive false positives and false negatives can have serious practical and clinical consequences. Beyond quantitative performance, model interpretability was examined using SHapley Additive exPlanations (SHAP). The SHAP interaction analysis reveals that both systolic blood pressure limit and diastolic blood pressure contribute interactively to CKD classification rather than acting as isolated predictors. Higher blood pressure values exhibit stronger interaction effects, as reflected by increased

SHAP interaction magnitudes, indicating elevated contribution toward CKD-positive predictions. Conversely, lower blood pressure values are associated with reduced interaction influence. This behavior aligns with established clinical understanding, where combined blood pressure abnormalities are known to increase renal risk. The interaction-based explanation further demonstrates that the model captures clinically meaningful feature dependencies instead of relying on individual attributes alone. The SHAP analysis showed that CKD predictions were influenced by multiple clinical attributes rather than being dominated by a single feature. This observation aligns with established medical understanding of CKD as a multifactorial condition affected by a combination of demographic, biochemical, and physiological factors. Such behavior increases confidence that the model’s decision-making process is consistent with clinical reasoning.

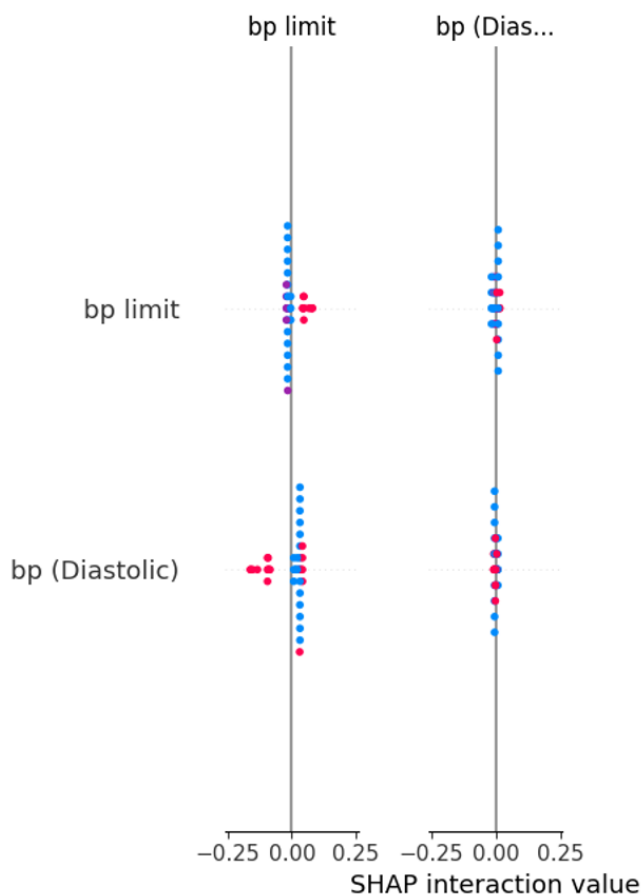


Fig. 2 SHAP interaction summary plot illustrating the interaction effects between blood pressure limit and diastolic blood pressure on CKD prediction

Fig. 2 presents the SHAP interaction summary, highlighting how blood pressure-related features jointly influence CKD predictions in the proposed model. Overall, the results demonstrate that the proposed framework effectively balances diagnostic performance with interpretability. By integrating ensemble learning with explainable artificial intelligence techniques, the system provides transparent and clinically meaningful predictions, supporting its potential use as a reliable decision-support tool for early chronic disease detection.

IV. CONCLUSION

This study proposed an interpretable machine learning framework aimed at improving diagnostic efficiency in the early detection of chronic diseases, with a specific focus on Chronic Kidney Disease. The framework integrates ensemble-based learning with systematic data preprocessing, class imbalance handling, and SHAP-based explainability to achieve reliable predictions while maintaining model transparency. Experimental results indicate that the proposed approach delivers strong diagnostic performance without sacrificing interpretability, addressing a key limitation of many black-box machine learning models in healthcare. By providing both accurate predictions and clear explanations of feature contributions, the framework supports informed clinical decision-making and promotes trust and accountability in data-driven diagnostic systems. Future work will focus on extending the framework to multi-disease diagnostic

scenarios and conducting prospective clinical validation to further evaluate its effectiveness in real-world healthcare settings. Such extensions will help assess scalability, adaptability, and long-term clinical impact.

V. REFERENCES

1. Manju VN and Aparna N. Decision Tree-Based Explainable AI for Diagnosis of Chronic Kidney Disease. In: 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2023. pp. 947–52.
2. Tarafdar, R., Soumik, M. S., & Venkateswaranaidu, K. (2025, May). Applying Artificial Intelligence for Enhanced Precision in Early Disease Diagnosis from Healthcare Dataset Analytics. In 2025 3rd International Conference on Data Science and Information System (ICDSIS) (pp. 1–7). IEEE.
3. Manik, M. M. T. G., Saimon, A. S. M., Miah, M. A., Ahmed, M. K., Khair, F. B., Moniruzzaman, M., & Bhuiyan, M. M. R. (2021). Leveraging AI-powered predictive analytics for early detection of chronic diseases: A data-driven approach to personalized medicine. *Nanotechnology Perceptions*, 17(3), 269–288.
4. Kumar, A., Singh, J. & Khan, A. A. A comprehensive machine learning framework with particle swarm optimization for improved polycystic ovary syndrome (pcos) diagnosis. *Eng. Res. Expr.*

- 6(3), 035233. <https://doi.org/10.1088/2631-8695/ad76f9> (2024).
5. Sharma, A. et al. A systematic review on machine learning intelligent systems for heart disease diagnosis. *Arch. Computat. Meth. Eng.* <https://doi.org/10.1007/s11831-025-10271-2> (2025).
 6. Dhanka, S. et al. Advances in machine learning and deep learning for hormonal disorder diagnosis: An exhaustive review on PCOS, thyroid, and optimization techniques. *Arch. Comput. Meth. Eng.* <https://doi.org/10.1007/s11831-025-10380-y> (2025).
 7. Kumar, A., Singh, J., Khan, A.A.: Arrhythmia detection using machine learning: A study with uci arrhythmia dataset. In: Bhateja, V., Patel, P., Tang, J. (eds.) *Evolution in Computational Intelligence*, pp. 217–226. Springer (2025). https://doi.org/10.1007/978-981-96-2124-8_16
 8. Zhang, D.; Cheng, L.; Liu, Y.; Wang, X.; Han, J. Mamba capsule routing towards part-whole relational camouflaged object detection. *Int. J. Comput. Vis.* 2025, 133, 7201–7221. [CrossRef]
 9. Xia, C.; Chen, H.; Han, J.; Zhang, D.; Li, K. Identifying children with autism spectrum disorder via transformer-based representation learning from dynamic facial cues. *IEEE Trans. Affect. Comput.* 2024, 16, 83–97. [CrossRef]
 10. Liu, Y.; Cheng, D.; Zhang, D.; Xu, S.; Han, J. Capsule networks with residual pose routing. *IEEE Trans. Neural Netw. Learn. Syst.* 2025, 36, 2648–2661. [CrossRef] [PubMed]
 11. Banu A, Begum Z. Chronic Disease Diagnosis Using Machine Learning Algorithm. *J Emerg Technol Innov Res* 2023;10:765-72.
 12. Debal AD, Sitote MT. Chronic kidney disease prediction using machine learning techniques. *J Big Data* 2022. doi: 10.1186/s40537-022-00657-5.