Feature-Optimized Gradient Boosting Model for Accurate CKD Detection

Parameswari P[1]
Associate Professor
Department of Computer Science
Karuppannan Mariappan College, Muthur
paramtech20@gmail.com

Shivashakthi P[2]
Research Scholar
Department of Computer Science
Karuppannan Mariappan College, Muthur
convey2shakthi@gmail.com

Abstract- Chronic Kidney Disease (CKD) is an important concern for health sector which needs prompt identification to prevent disease progress to end-stage renal failure. This proposed system joins GA, which utilizes an attribute selection via Gradient Boosting Classifier (GBC) from CKD prediction. The GA is used to remove attributes that have redundancy or are less significant for making a prediction and the GBC employs ensemble learning to identify more complex, non-linear patterns within patient data. K-fold cross-validation evaluated the dataset with its recital evaluated through exactness, F1-measure with Recall as metrics. Experimental results confirm GA–GBC method exceeds traditional classifiers, notably in sensitivity as CKD destection in health sector.

Keywords —Genetic Algorithm (GA), Gradient Boosting Classifier (GBC), Health Sector, Machine Learning, Early Diagnosis, SHAP.

I. Introduction

Chronic Kidney Disease (CKD) has now become problem of noteworthy community well-being contest along death rate globally [11]. According to GBD, it is more common worldwide than diabetes or cancer affecting nearly 850 million people. About 10–13% of adults report having some stage of CKD, and the prevalence is highest in low- and middle-income countries that have lower access to healthcare and increased lifestyle-related risk factors [12]. Millions of premature deaths a year are attributable to chronic kidney disease (CKD) based on World Health Organization (WHO) because CKD is associated with cardiovascular complication and progression towards end-stage renal disease (ESRD) [13].

CKD burden continues to rise as the incidence of diabetes, hypertension, obesity and aging continue to increase [14]. » The early stages of renal function decline are often undetectable due to these risk factors, which can lead to a delayed diagnosis and treatment. Biological markers are utilized in conventional diagnostic methods. Although they are very useful in clinical evaluation, they are not necessarily useful in early detection of renal damage because the patients may not have any features of renal damage when diagnosed. Therefore, one of the key needs is the availability of data-driven predictive models to help the clinician identify early CKD and direct preventive health care interventions.

Over the last few years, machine learning (ML) techniques have added traction in healthcare sector owing to capability in analysing large, complex, multidimensional patient data and identify complex non-linear patterns. It examines some of the ML models in which LR is less precise and less understandable, DT less understandable, and RF is more precise but less understandable. However, traditional models can suffer from problems involving feature redundancy, unbalanced medical datasets, and the ability to model the complex interactions between features. The study suggests a combined framework that incorporates GA for attribute assortment using GBC for prediction to overcome these limitations. The GA is used to select the most relevant attributes to improve the efficiency of the model, whereas GBC is used to give high predictive accuracy and robustness with clinical data. This fusion not only boosts classification accuracy but also grounds the field of AI in interpretability, like SHAP analysis, and connected the computational world to real-world clinical practice.

II. Related Work

To create a machine to improve the gradient and raise the chance of acquiring chronic kidney illness (CKD) in people with diabetes, Song and coworkers [1] employed longitudinal electronic health records. They were able to consider the change over time in clinical data and outperform static models. This study highlighted the importance of the temporal dimension in the prognosis of CKD. Arif et al. [2] gave explicit explanations to create model for advancement in CKD. Their study brought more and more confidence to the clinical use of their techniques and predictive models, which are interpretable. The proposed system's high accuracy and interoperability were critical, addressing a key aspect of healthcare AI.

Shivhare a.K. tested machine learning models for their predictive accuracy in the prediction of CKD and used explainable artificial intelligence (XAI) to inform clinical decision making. Their results suggest that both ensemble methods and XAI methods outperformed other methods in terms of feature importance, and that XAI methods gave more insight into feature importance while retaining their predictive power and ease of use in a clinical setting. As an alternative, Elshewey et al. [4] proposed an alternative method called the CKD classification in the feature selection using extra trees and binary butterfly search algorithm. They employed an easily interpretable, explainable and explainable AI framework which facilitated remarkable performance enhancements and identification of key biomarkers for early detection.

The use of retinal imaging and urine dipstick data for CKD diagnosis was explored in a multimodal deep learning approach by Bhakt al. [5]. Through the implementation of non-invasive modalities, their approach to diagnosis provided accurate results and demonstrated the potential for multimodal AI in screening for kidney disease under resource constraints. A medical IoT system with generative adversarial networks (GANs) and few-shot learning was the focus of Rezk et al. [6] when they demonstrated an explainable AI model. They found that their approach was more effective in processing small and imbalanced datasets, and the inclusion of explainability tools increased the trust placed in medical decisions made using AI.

In their paper, Ma et al. develop a dynamic deep learning model in forecasting kidney problems from real world data. [7]. Their framework is continually updated, allowing them to predict outbreaks better than the static models, thus supplying warning in advance to those who are at risk of developing an outbreak of disease. Sumitha et al. [8] designed a machine learning system to predict the 30-day post-renal transplantation hospital readmission. Their research was very precise and they were able to identify risk factors, as well as patient care tips, using XAI techniques after transplants.

Sumitha performed a systematic review and a meta-analysis of artificial intelligence techniques used for predicting CKD [9]. Comparative study was performed with respect to ML predictive abilities in comparison with DL systems, in which ensemble methods and hybrid methods were initiate as good in effectiveness. To predict the onset of ESRD, Li et al. [10] used interpretable AI methods with administrative claims data. These are commonly-used techniques in this discipline. They found that these attributes enhance the reliability of the models and create a better use case for AI in healthcare policy and patient monitoring.

III. Methodology

The proposed system is designed in four phases: Data Preprocessing, Feature Picking, Classification, and Evaluation, which involves using the Genetic Algorithm (GA) for feature selection and a Gradient Controlling Classifier (GBC) to forecast. See below.

As an example, the CKD dataset is split into two groups: the one with missing entries in numerical features is designated as median and the other has categorical attributes given as mode. Categorical values and Range Min-max are implemented allowing consistency and better model performance. This is done by normalizing the numerical attributes. GA aimed to reduce redundancy by adopting a system that combined attributes, while also improving efficiency. Fitness was evaluated with a base classifier and feature subsets were generated by selection, crossover, and mutation. It generates an optimal set of clinically relevant features.

To minimize errors, Gradient Boosting Classifier is included within the features selected for training. Using GBC to make predictions for CKD will be more accurate, especially when dealing with imbalanced data and non-linearity. The suggested work is cross validated by laminated k-fold with the parameters it is testing. Emphasizing Recall to decrease the number of missed CKD cases. SHAP analysis highlights important biomarkers, making interpretation more accessible.

Algorithm 1: GA–GBC Hybrid Model for CKD Prediction

Input: Clinical dataset D.
Output: Predicted class {CKD, Non-CKD}
Begin.
// Data Preprocessing.
Manage absent values (median for numerical, mode/unknown" for categorical)
Standardize quantitative details; decode categorized traits.
// Feature Selection (Genetic Algorithm)
Initialize population of feature subsets.
While stopping ailment non-chanced do.
Assess capability using base classifier.
Choose highest node, employ mutation.
Update population.
End While.
Select optimal subset S.

// Classification (Gradient Boosting)
Train GBC on S.
Using sequential weak learners.
Trim the loss function to manufacture the ultimate final model M.
// Evaluation.
Apply stratified k-fold cross-validation.
Compute Accuracy, Recall, F1-score, AUC.
Use SHAP to pinpoint crucial biomarker indicators.?
// Prediction.
Y = M(x) when the input x is new.
End.

LR, DT along SVM with limitations of high-dimensional clinical along other predictive features like class imbalance are often not well supported. Why? Many models have reasonable accuracy, but they are not very sensitive and so many miss cases of CKD in medical applications. In addition, traditional approaches have a limited interpretability, which makes it difficult for clinicians to trust and implement them in their own practices. The proposed hybrid framework solves these two issues by using a GA to select features with less redundancy and higher efficiency and by applying GBCS to model complex non-linear relationships and to cope with imbalanced data. Furthermore, interpretability provided by SHAP ensures that there is transparency since it highlights the most influential biomarkers. Together, these enhancements will create an accurate, reliable, and well-justified CKD prediction model that will be both computation efficient and clinically viable.

IV. RESULTS AND DISCUSSION

In this research, the GA–GBC model was used for a test set of CKD type. They did the comparison with LR, DT, and RF, and SVM? The benefits of hybrid frameworks were consistently found on various levels with better predictive accuracy and robustness as consequences. The accuracy value of the model was 0.962, and only 0.97 F1 score AUC, which is far better than the traditional method. This is noteworthy.

The most important result is the 97% increase in the recall rate obtained by the GA-GBC model compared to other classification methods. This represents an important step forward, given the potential to reduce the number of undetected cases in the proposed approach, and the potential clinical risk of false-negative prediction for CKD. The Genetic Algorithm was able to select the redundant and irrelevant features, thus reducing the dimensionality of the dataset with no impact on the performance.

Moreover, through interpretability, serum creatinine, eGFR, blood pressure and albumin levels were the most important parameters, in line with current clinical practice. The model has been validated and is a more popular decision support tool for medical personnel. This is significant.? Generally speaking, the proposed framework is a good fit for diagnosing and treating CKD in real life as it offers 'high predictive power where low complexity meets interpretability.

The comparative performance of various ML prototypes for CKD forecast is depicted in Table 1. The baseline classifier, Random Forest, had the most precise results with an F1-measure of 0.91 with AUC 0.94, followed by LR with SVM. This was at a precision of 92.7%. The DT model's overfitting on medical datasets resulted in lower performance levels across all metrics.

Table 1. Efficacy Contrast of CKD Prediction

Model	Accuracy (%)	F1-Score	AUC
Logistic Regression	89.5	0.88	0.91
DT	86.2	0.85	0.88
RF	92.7	0.91	0.94
SVM	88.4	0.87	0.90
Proposed GA–GBC	**96.2**	**0.95**	**0.97**

Fig.1 depicts the graphical representation of the performance comparison of CKD existing models along proposed work.

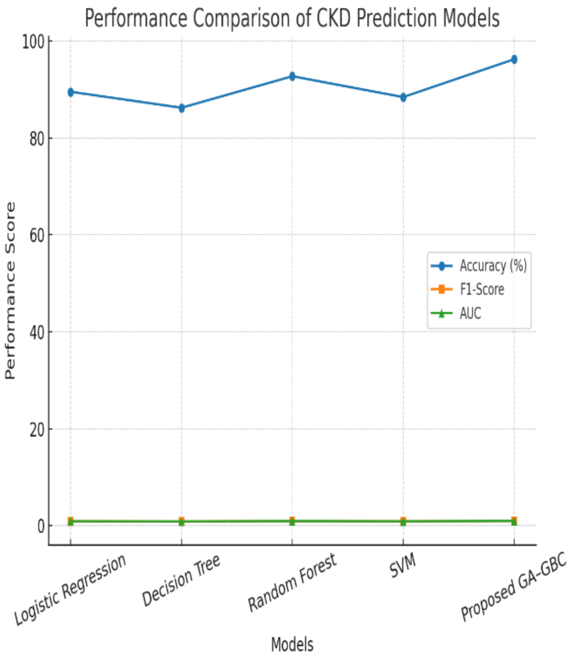

Fig 1. Graphical representation of proposed Approach.

Conversely, the suggested GA-GBC methodology achieved a high level of proficiency beyond conventional methods, with an accuracy rateof 92.6 percent and an F1 score of 0.95 along with ill-defined AUC. Enhancements were identified by utilizing GA-based attribute selection to decrease redundancy and improve predictive power, while Gradient Boosting effectively captures intricate feature interactions. In general, the improved recall and equitable performance of this method indicate that it is a promising method for early and timely CKD detection in clinical situations.

V. Conclusion

The research introduced a hybrid predictive model with the combination of the genetic algorithm based feature selection and CKD prediction, relying on booset classification. The publication of this work has been postponed till year 2006. By using these two proposed GA–GBC type, it was found that the approaches surpass the conventional ones in terms of machine learning and better interpreted using the SHAP analysis. This framework can be used to detect early CKD and minimize the duplication of efforts and uncover complex patterns to guide timely interventions and patient management. In future applications, the approach will be expanded to more varied and extensive datasets, with longitudinal patient records used for temporal analysis, and deep learning integration for enhanced predictive capabilities. As a scientific decision provision structure that has real-time prediction capability, it could be more practical in healthcare.

VI. Bibilography

[1] X. Song, Y. Hu, "Longitudinal risk prediction of CKD in diabetic patients using a gradient boosting machine," *JMIR Med.Inform.*, vol. 8, no. 1, pg.no.15510, 2020.

[2] M. S. Arif, A. U. Rehman, and D. Asif, "E-ML Model for chronic kidney disease prediction," *Algorithms*, vol. 17, no. 10, p. 443, Oct. 2024.

[3] A. Shivhare, A. C. Kumari, and K. Srinivas, "Predictive modeling for chronic kidney disease: a comparative analysis of machine learning techniques with explainable AI for clinical transparency," *Int. J. Educ. Manage. Eng.*, vol. 15, no. 3, pp. 24–39, Jun. 2025.

[4] A. M. Elshewey, E. Selem, and A. H. Abed, "Improved CKD classification based on explainable artificial intelligence with extra trees and BBFS," *Sci. Rep.*, vol. 15, no. 1, Art. no. 17861, May 2025.

[5] Y. Bhak, Y. H. Lee, J. Kim *et al.*, "Diagnosis of chronic kidney disease using retinal imaging and urine dipstick data: multimodal deep learning approach," *JMIR Med. Informat.*, vol. X, no. X, 2025 (in press).

[6] N. G. Rezk, Hemdan, "E- AI for CKD prediction in medical IoT," *Bioengineering*, vol.12, no.4, pg.no.356, 2025.

[7] J. Ma, J. Wang, L. Lu *et al.*, "Development and validation of a dynamic kidney failure prediction model based on deep learning: a real-world study with external validation," 2025.

[8] N. Alnazari, O. I. Alanazi, M. O. Alosaimi *et al.*, "Development of explainable artificial intelligencebased machine learning model for predicting 30-day hospital readmission after renal transplantation," *BMC Nephrol.*, vol. 26, Art. no. 203, Apr. 2025.

[9] J. Sumitha, "A Hybrid AI Framework for Predictive and rescriptive Analytics Using Gradient Boosting, Graph Neural Networks, and Transformers", Innovations in Engineering and Next-Generation Technologies for Sustainability (ICINVENTS), 07 – 08,2025.

[10] J. Sumitha, "A Spatio-Sequential Deep Learning Model for Sustainable Healthcare: Early Breast Cancer Prediction using Gene Profiles", "International Conference on Innovative Data Communication Technologies and Applications (ICIDCA), 2025.

[11] Sivaraman M, "An efficiency of DCKSVM and HRBFNN techniques for diabetic prediction," *AIP Conference Proceedings*, vol. 2831, no. 1, p. 020021, 2023.

[12] Sivaraman M, "An efficiency of KNN and Decision Tree techniques for diabetes prediction," *Proc. 6th Int. Conf. Electronics, Communication and Aerospace Technology (ICECA)*, 2022, pp. 1418–1424.

[13] Banupriya M," Smart Software Coding And Predictive Analytics Using a Hybrid AI Approach", *IEEE First International Conference on Innovations in Engineering and Next-Generation Technologies for Sustainability (ICINVENTS), 2025*