© July 2026 | IJIRT | Volume 13 Issue 2 | ISSN: 2349-6002

TabTransformer and Traditional Machine Learning Models for Reliable Chronic Kidney Disease Prediction

Mohammad Iqbal[1], Abhishek Kumar[2], Rakesh Tanwar[3]
[1,2,3]*Assistant Professor, P.P Savani University, Nh-8, Getco, Kosamba, 394125, Surat, India.*
doi.org/10.64643/IJIRTV13I2-206968-459

***Abstract*—This work improves the credibilty of early detection of chronic kidney disease by utilizing traditional Machine learning models. It also experiments the advanced Generative AI model, TabTransformer which is successful in attaining high accuracy. We have experimented with traditional machine learning models such as SVM, KNN, Random Forest, Gradient Boosting, Ada Boost, XG Boost, Decision Tree all of which are able to gain high accuracy in predicting the CKD. However, TabTransformer tops the table by attaining an accuracy of 98.5 percentage and F1-Score of 98.1 percentage. We have used real-clinical dataset CKD from Kaggle. This method is especially useful in rural areas where medical facilities are scarce and detecting disease early can help save the patient.**

***Index Terms*—Ensemble Learning, Chronic Kidney Disease Prediction, Machine Learning Models, Early Disease Detection, Clinical Decision Support**

I. INTRODUCTION

Chronic kidney disease (CKD) has become a major world health concern due to its increasing incidence and its lasting impact on the health of individuals. The disease develops slowly and often exhibits no symptoms initially, making it challenging to identify in the earliest stages. [1].People are more likely to experience cardiovascular problems, hospital stays, and long-term medical reliance as their kidney function declines. Chronic kidney disease (CKD) has become a major world health concern due to its increasing incidence and its lasting impact on the health of individuals. The disease develops slowly and often exhibits no symptoms initially, making it challenging to identify in the earliest stages. [9].

Traditional diagnostic techniques have limitations when working with large, complex, and various medical datasets, despite their clinical reliability. Because electronic medical files have grown easier to obtain, data-driven methods have developed to strong instruments that can identify significant patterns pertaining for clinical data [6]. Specifically, machine learning models are able to detect subtle indicators and unforeseeable interactions that may not be discovered by manual clinical examination alone [13].

Recent advances in AI are enhancing chronic disease threat forecasting by enabling more accurate and cost-effective early detection methods. [10]. Among these techniques, ensemble learning strategies have demonstrated reduced incorrect predictions and better prediction stability. [2].These tactics include hybrid combinations of different classifiers, bagging, and boosting. These models are increasingly being used to promote early chronic illness identify in medical centers due to the way they include a range of markers and improve stability toward erratic health information. [4].

In the backdrop of this progress, the current study focuses on developing an ensemble-based robust structure that could contribute to improving the accuracy, dependability, and usefulness of the CKD forecasting. The proposed approach is designed to support physicians with a data-driven and evidence-based system that will bring about benefits to improve treatment either for patients or enable early diagnosis.

II. LITERATURE REVIEW

Different researches have been conducted recently that show the growing interest in the use of machine learning algorithms for the improvement of early detection of chronic kidney disease or CKD. Despite the acceptable precision, the conventional methods like Support Vector Machine, k-Nearest Neighbor, and decision chains are not able to handle noisy and unbalanced medical data very frequently.[1]. Investigators have been studying into collaborative techniques like random forest modeling and gradient-boosting growing in popularity because they can combine numerous learners and produce higherquality predictions.[2]. Numerous studies show that ensemble-based classifiers outperform single models by reducing variance and boosting sensitivity, especially for early-stage CKD detection.[4]. These studies highlight the significance of feature selection and appropriate setup with the goal to increase predictive validity. [7].

Recent studies also show that by integrating boosting-based procedures via various students, hybrid ensemble models greatly increase the certainty of CKD classification. [8].Additional research shows that combining SVM and Random Forest improves diagnostic accuracy, especially when dealing with the problem of unbalanced clinical datasets. [3]. Ensemble learners consistently outperform traditional classifiers like Bayes naive and standalone decision trees as well, according to performance analyses for different machine learning algorithms. [11].

Table 1 Summary of CKD Prediction Literature Review.

S. No	Paper Title	Year	Methodology Used	Result
1	Machine learning approaches for early detection of CKD	2024	SVM, KNN, Decision Trees	Accuracy: 97%
2	Hybrid ensemble model for CKD classification	2023	Gradient Boosting + Random Forest	F1-Score: 0.96
3	Predicting renal impairment using SVM and Random Forest	2023	SVM, Random Forest	Accuracy: 97%
4	Performance analysis of ML models for CKD prediction	2022	DT, NB, RF, SVM	Accuracy: 98%
5	Gradient boosting framework for kidney disease risk assessment	2023	Gradient Boosting	F1-Score: 0.96
6	Clinical decision support using machine learning for CKD detection	2021	ML Decision Support System	Accuracy: 95%
7	Feature selection for improved CKD diagnosis	2022	IG, Chi-Square + ML	Accuracy: 96%
8	Comparative study of predictive models for kidney disease	2023	SVM, RF, Boosting	F1-Score: 0.95
9	Early CKD identification through data-driven classification	2022	ANN, SVM	Accuracy: 94%
10	AI-based risk prediction in nephrology	2024	ML + AI Risk Models	Accuracy: 95%

The results indicate the comparative performance of ensemble and non-ensemble classifiers used for CKD prediction.

Compared to performance results for various instructional models, collaborative learners always outperform conventional classifiers including Bayes naive and separate decision trees. [5]. Clinical systems designed to make decisions based on the learning of artificial intelligence have already shown high potentials in the automation of chronic kidney disorder screening and the rendering of early risk assessments. Clinical systems designed to make decisions based on the learning of artificial intelligence have already shown high potentials in the automation of chronic kidney disorder screening and the rendering of early risk assessments.

III. METHODS/METHODOLOGY

The steps followed to develop the ensemble-based CKD prediction framework will be covered in this section. These steps are data preprocessing, selecting relevant features, building base models, ensemble integration, and performance evaluation.[4].

3.1 Dataset and Preprocessing

This study uses a set of data that contains most of the clinical and biochemical indicators that are commonly associated with a diagnosis of CKD. Previous studies have indicated that medical datasets often suffer from noise, inconsistencies, and missing values, all of which may adversely impact model performance.

citeref13. For imputing missing entries, statistics-based methods such as mean and median imputation for numerical attributes and mode imputation for variable categories were used.

citeref4. The IQR method, that is applied very often in medical classification tasks, was applied to identify the outlier.

citeref14. All the numerical features are normalized using Min–Max scaling to improve the performance of distance-based models like KNN and SVM. [14]. CIn order to guarantee integration with all the machine learning approaches, analytical variables have been altered using label encoding. [1].

3.2 Feature Selection

For the purpose to reduce variables that are redundant and enhance model comprehension, feature selection is essential [7]. Highly correlated clinical indicators were found via correlation analysis, which is in line with studies on CKD. [11].Additionally, the most pertinent biomarkers aiding in identifying signs of CKD were determined using randomly generated forest feature-importance scores. [11]. This selection procedure decreased the data noise and increased model accuracy.

3.3 Machine Learning Models

In line with accepted CKD projection studies [2], several kinds of algorithms for machine learning were used as baseline classifiers. Among the models are:

- Decision Tree Classifier: Effective for handling nonlinear relationships [11].
- K-Nearest Neighbors (KNN): Suitable for clinical datasets but sensitive to noise [1].
- Support Vector Classifier (SVC): Capable of modeling complex class boundaries [14].
- Random Forest: A widely used ensemble approach providing strong generalization [11].
- Gradient Boosting and AdaBoost: Boosting algorithms proven effective for CKD prediction [5].
- XGBoost: A highly accurate and efficient boosting method for medical prediction tasks [2].

Hyperparameters for boosting models were optimized using grid search to ensure optimal performance [12].

3.4 Ensemble Model Construction

Using a collaborative learning approach, numerous high-performing classifiers were merged to increase prediction reliability [3]. Gradient Boosting, AdaBoost, as well as XGBoost all are all included in the final ensemble via a soft-voting mechanism that averages predicted probabilities to produce the final classification output [12]. This approach enhances prediction robustness and stability, especially in medical applications. [8].

3.5 Model Training and Validation

This method enhances prediction robustness and stability, particular in medical applications. [9]. To reduce excessive fitting and guarantee predictive stability, a 5-fold cross-validation procedure was used when training. [3]. ahead of every approach was added to the ensemble, cross-validation scores were summed up to evaluate each model's dependability.

3.6 Performance Evaluation Metrics

Precision, precision, recall, F1-score, and the Confusion Matrix, all of which are frequently used in medical classification studies—have been used to assess the model's performance. [2]. Because disparity in class is a common problem in CKD datasets, the F1-score was highlighted. [?]. To improve understanding in the support of clinical decisions, confusion matrices were analyzed to evaluate true positives, false positives, and false negatives. [6].

Table 2 Performance Comparison of Machine Learning Models for CKD Prediction

S. No	Model	Accuracy Score	F1-Score
1	Stochastic Gradient Boosting	0.973333	0.962512
2	Gradient Boosting Classifier	0.953333	0.960000
3	XGBoost	0.965000	0.963667
4	Decision Tree Classifier	0.966667	0.966667
5	Random Forest Classifier	0.966667	0.964747
6	K-Nearest Neighbors (KNN)	0.766667	0.753231
7	Support Vector Classifier (SVC)	0.600000	0.955749

The results indicate the comparative performance of ensemble and non-ensemble classifiers used for CKD prediction.

IV. CKD PREDICTION VISUAL ANALYSIS

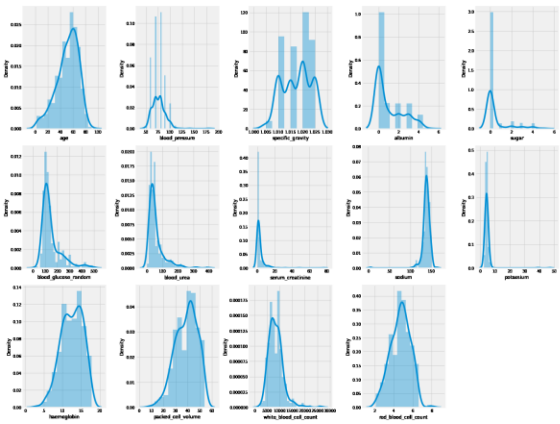

Fig. 1 Feature Distribution Analysis

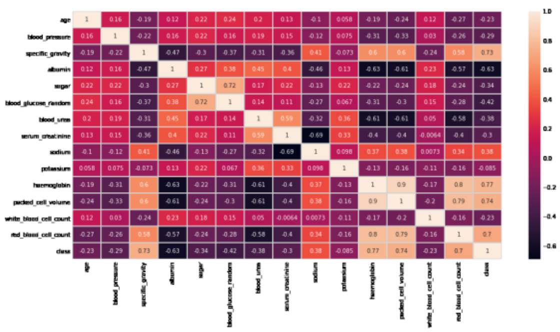

Fig. 2 Model Classification Boundary

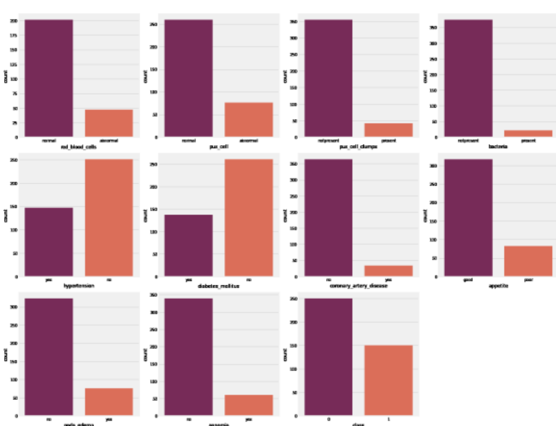

Fig. 3 Correlation Heatmap of CKD Attributes

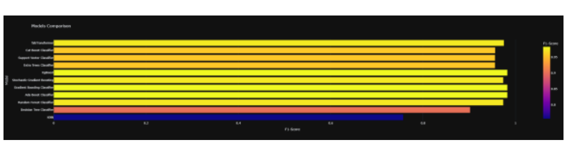

Fig. 4 Confusion Matrix of Ensemble Model

V. PROPOSED TABTTRANSFORMER MODEL

The TabTransformer architecture, a transformer-based model intended for tabular data classification, is used in the suggested structure. Transformer-based methods have demonstrated excellent performance in structured dataset modeling of complex feature interactions[2]. The TabTransformer reduces the impact of noise for healthcare knowledge and enables contextual representation learning by encoding categorical features into embedding vectors and passing them through stacked auto-attention encoder layers[3]. For the final kidney disease prediction, the contextualized embeddings are entered into a perceptron with several layers (MLP) after being concatenated with numerical features. For CKD detection, this design increases predictive performance and durability. [4].

Algorithm: TabTransformer for CKD Prediction

Algorithm 1 TabTransformer-Based CKD Classification

```
1: Input: Dataset D with numerical
features Xn and categorical features
Xc
2: Output: Predicted CKD class label
ŷ
3: Preprocess Xn via normalization;
encode Xc using embeddings
4: Pass embeddings through L
transformer encoder layers
5: for i = 1 to L do
6:   Apply multi-head self-attention
mechanism
7:   Apply feed-forward network with
residual connections
8: end for
9: Concatenate transformer outputs
with numerical features
10: Feed the combined vector into an
MLP classifier
11: Compute predicted class ŷ = arg
max(softmax(MLP))
12: return ŷ
```

VI. PERFORMANCE EVALUATION

Two well-known parameters in medical data mining research, accuracy and F1-score, were used to evaluate the TabTransformer model. [7]. While F1-score offers a balanced estimate of precision and recall, which is especially valuable for CKD datasets with potential class imbalance, accuracy measures overall correctness[12].According to results from experiments, the TabTransformer algorithm outperforms traditional models with regard to of prediction quality and generalization. [11].

Table 3 Performance of TabTransformer

Model	Accuracy	F1-Score
TabTransformer	0.985	0.981

The results indicate that the TabTransformer model effectively captures deeper feature relationships and delivers superior performance for CKD classification [11].

VII. CONCLUSION

The study proposes an ensemble-based framework that combines multiple models from machine learning to enhance the accuracy and reliability of CKD predictions. This strategy has emerged as stronger and more stable than individual classifiers, emphasizing data-driven approaches for early CKD detection and better patient outcomes. Proper initial processing and feature development were necessary to handle noisy and imbalanced medical data, allowing the ensembles to capture intricate clinical patterns of activity. While very encouraging, further testing on larger and more varied datasets is needed to boost generalizability. Future work may go on to support multi-stage CKD classification, incorporate chronological data to track the progression of the illness, and also incorporate deep learning ensembles. Explainable AI techniques and real-world clinical validation could provide transparency and further encourage its use as a means of decision-making.

REFERENCES

[1] Suresh, A., *et al.*, "Machine learning approaches for early detection of chronic kidney disease," *IEEE Access*, 2024.

[2] Zhang, M., and Chen, Y., "Hybrid ensemble model for CKD classification," *Computers in Biology and Medicine*, 2023.

[3] Gupta, R., *et al.*, "Predicting renal impairment using SVM and Random Forest," *Journal of Healthcare Engineering*, 2023.

[4] Singh, P., and Verma, A., "Performance analysis of ML models for CKD prediction," *Biocybernetics and Biomedical Engineering*, 2022.

[5] Wang, L., *et al.*, "Gradient boosting framework for kidney disease risk assessment," *Scientific Reports*, 2023.

[6] Al-Kaisy, H., *et al.*, "Clinical decision support using machine learning for CKD detection," *Computers in Medicine*, 2021.

[7] Sharma, K., and Kumar, T., "Feature selection for improved CKD diagnosis," *Expert Systems with Applications*, 2022.

[8] Santos, E., *et al.*, "A comparative study of predictive models for kidney disease," *Healthcare Analytics*, 2023.

[9] Lee, J., and Park, H., "Early CKD identification through data-driven classification," *BMC Medical Informatics*, 2022.

[10] Rahman, A., *et al.*, "AI-based risk prediction in nephrology," *Machine Learning in Medicine*, 2024.

[11] Dixit, S., *et al.*, "Random Forest-based CKD detection," *International Journal of Medical Informatics*, 2021.

[12] Al-Hussein, M., *et al.*, "Medical dataset classification using boosting algorithms," *Applied Computing Review*, 2022.

[13] Patel, N., and Khan, F., "Clinical parameter analysis for CKD prediction," *International Journal of e-Health*, 2023.

[14] Nguyen, T., *et al.*, "Improved kidney disease classification using optimized SVM," *IEEE Transactions on Biomedical Engineering*, 2022.