

Deep attention transformer models for chronic kidney disease stage prediction

Subashini N. J., Venkatesh Kaliyamurthy

Department of Networking and Communications, SRM Institute of Science and Technology, Chennai, India

Article Info

Article history:

Received Oct 27, 2025

Revised Jul 30, 2026

Accepted Aug 12, 2026

Keywords:

AttentionNet

Chronic kidney disease

Feature ablation

Residual multilayer perceptron

TabTransformer

ABSTRACT

Chronic kidney disease (CKD) is a growing health problem that demands accurate stage prediction for timely clinical intervention. Conventional machine learning (ML) models, although widely used, often struggle to capture complex, non-linear relationships in clinical data. To address this limitation, this study proposes a deep learning (DL) framework based on two tabular architectures: attention network (AttentionNet) and tabular transformer (TabTransformer). AttentionNet assigns higher importance to clinically relevant features through an attention mechanism, while TabTransformer uses self-attention to model dependencies among clinical variables. These models were evaluated against baseline approaches, including residual multilayer perceptron (ResMLP), wide and deep (WnD), and a standard multilayer perceptron (MLP). The dataset comprised 1,446 patient records with 26 validated features after leakage removal, spanning five CKD stages. Preprocessing involved cleaning, categorical encoding, normalization, and evaluation using 10-fold cross-validation. Results showed that ResMLP and TabTransformer consistently achieved the highest accuracy and F1-scores, both exceeding 99%, with minimal variation across folds. Ablation analysis revealed the critical role of glomerular filtration rate (GFR) and serum creatinine (SC), as their removal led to notable performance degradation. High area under the curve (AUC) values further confirmed model robustness. Overall, DL-based tabular models offer a reliable foundation for CKD stage prediction and clinical decision support systems.

This is an open access article under the CC BY-SA license.



Corresponding Author:

Subashini N. J.

Department of Networking and Communications, SRM Institute of Science and Technology

Kattankulathur, Chennai, Tamil Nadu, India

Email: suba.satheesh@gmail.com

1. INTRODUCTION

Chronic kidney disease (CKD) is a major global public health concern, affecting more than 850 million people worldwide, with recent global estimates indicating a steadily rising prevalence across both developed and developing regions [1], [2]. The disease is progressive in nature and, if not identified in its early stages, can lead to renal failure and the need for dialysis or transplantation. Early detection of CKD stages is therefore crucial for timely medical intervention and improved patient outcomes.

Conventional clinical assessment methods and statistical approaches have been applied for CKD stage prediction. Traditional machine learning (ML) algorithms such as decision trees, support vector machines, and logistic regression have demonstrated some predictive capability [3], [4]. However, these models often face challenges in handling complex non-linear interactions among multiple clinical variables, resulting in reduced generalization and inconsistent performance across diverse patient groups [5].

Deep learning (DL) methods provide an opportunity to overcome these limitations by automatically capturing intricate feature interactions without manual engineering. Recent advances in DL for tabular medical data have shown significant improvements in classification and diagnosis tasks [6], [7]. Building upon these developments, this study proposes a DL framework using two modern architectures, attention network (AttentionNet) and tabular transformer (TabTransformer), along with baseline networks such as residual multilayer perceptron (ResMLP), wide and deep (WnD), and multilayer perceptron (MLP).

The contributions of this work are fourfold: i) development of a leakage-safe preprocessing pipeline with 26 validated clinical features; ii) implementation of two advanced DL models, AttentionNet and TabTransformer, tailored for CKD stage prediction; iii) comprehensive evaluation through 10-fold cross validation (CV) and ablation studies to assess the clinical importance of glomerular filtration rate (GFR) and serum creatinine (SC); iv) comparative analysis with baseline models to demonstrate the robustness and reliability of the proposed approach. The remainder of this paper is organized as follows. Section 2 reviews existing literature on CKD prediction and tabular DL methods. Section 3 explains the dataset, preprocessing pipeline, and model architectures. Section 4 presents the experimental setup, results, and discussion. Section 5 concludes with key insights, limitations, and future research directions.

2. RELATED WORKS

As in the paper [8], classical ML on tabular laboratory and clinical variables has delivered strong baselines for CKD prediction, with boosted trees and careful feature pruning reaching near-ceiling accuracy on benchmark datasets. Aoki *et al.* [9] showed that compact risk models trained on large, real-world laboratory streams can anticipate a 30% decline in estimated GFR within 5 years, achieving a receiver operating characteristic (ROC)-area under the curve (AUC) near 0.85 and highlighting eGFR slope and albuminuria as dominant signals. Sanmarchi *et al.* [10] surveyed clinical CKD applications of ML and noted three persistent gaps: limited external validation, sparse attention to fairness, and interpretability constraints. Chen *et al.* [11] set out a protocol to standardize evaluation across CKD prognosis and diagnosis studies, addressing inconsistent reporting that complicates head-to-head comparisons. Khalid *et al.* [12] synthesized progress on predicting CKD progression and emphasized pragmatic choices in variable selection and validation that improve transportability. Complementing this, Pan and Tong [13] performed a meta-analysis indicating that artificial intelligence models can match or exceed conventional risk equations for CKD prognosis, while still needing better calibration across populations. Beyond generic progression, Ruget *et al.* [14] reported time-to-event modeling for renal replacement therapy and stage transition, illustrating how survival-aware objectives add clinical utility. Feature selection for medical tabular data has advanced on two fronts. Battistella *et al.* [15] proposed graphical ensembling to stabilize feature subsets and improve downstream accuracy, while reviews on imbalanced medical data [16] cataloged sampling and algorithmic remedies that reduce bias in minority classes without sacrificing precision. Broader surveys of metaheuristic feature selection framed reproducible evaluation and robust subset discovery for clinical datasets [17].

Transformer-based methods have moved from images to sequential electronic health records (EHRs) and tabular data. Yang *et al.* [18] introduced TransformEHR, a generative encoder–decoder pretraining approach that improved precision–recall on multiple clinical tasks, signaling the value of sequence-aware representations for longitudinal care trajectories. In oncology EHR, Veisi *et al.* [19] applied a transformer to primary-care pathways and reported AUROC above 0.92 for early breast cancer detection, demonstrating that attention over temporal codes can outperform regression baselines. For tabular health data generation and augmentation, Kang *et al.* [20] combined transformers with GANs to synthesize heterogeneous medical tables more faithfully than prior conditional tabular generative adversarial network (CTGAN)-style models, suggesting a route to address data scarcity and privacy. Finally, domain-wide perspectives continue to underline unresolved issues—dataset shift, imbalance, and transparent reasoning—across CKD prediction studies [21], [22]. Cardiovascular risk cohorts [23], which achieved good transparency but did not evaluate feature ablation robustness. More recently, categorical boosting (CatBoost) with nature-inspired optimization reached accuracy near 99% while ranking SC and GFR as top features, further supporting the current findings [24].

Overall, recent work converges on three insights: i) compact feature sets anchored in routinely collected labs remain competitive [8], [9]; ii) stability-oriented feature selection and imbalance handling improve reliability on medical datasets [15], [16]; and iii) transformer architectures that encode care timelines or column-wise relations offer measurable gains, especially when pretraining and synthetic augmentation are used judiciously [22]–[25]. The present study builds on these directions by pairing clinically grounded features with attention-based modeling and clear reporting to enhance transportability and interpretability. Recent transformer-based models such as survival transformer for electronic health

records (STRAFE) and kidney failure deep learning (KFDeep) primarily focus on longitudinal or time-to-event CKD prediction using sequential EHR data [26]. In contrast, this study formulates CKD stage prediction as a multi-class classification problem using static, leakage-controlled clinical features, with explicit emphasis on robustness and clinical feature importance through ablation analysis.

3. METHOD

The proposed work aims to develop a robust DL framework for CKD stage prediction using clinically validated features. The workflow begins with preprocessing and leakage control, followed by feature handling and dataset preparation. Models including AttentionNet and TabTransformer are then trained and evaluated against baseline networks such as ResMLP, WnD, and a standard MLP. Ablation studies further assess the clinical impact of removing specific biomarkers such as GFR and SC. The complete pipeline is illustrated in Figure 1, which depicts the architecture from data input to final model evaluation.

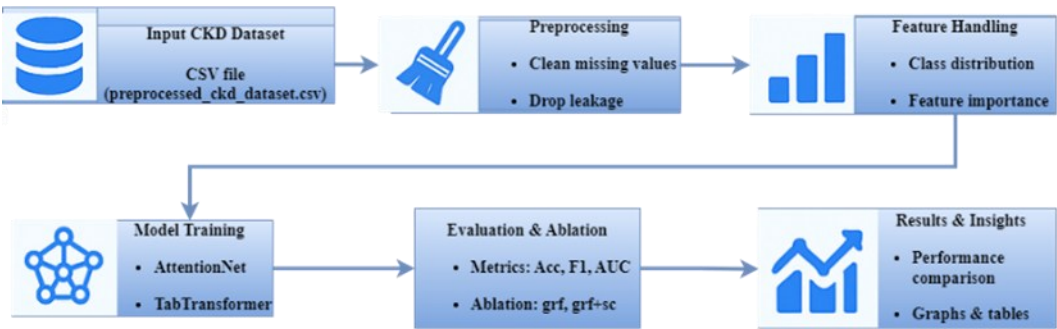


Figure 1. Proposed architecture for CKD stage prediction

3.1. Dataset and features

The study was carried out on a CKD dataset consisting of 1,446 patient records. Each record contained 27 clinical variables including both laboratory and categorical attributes, along with a five-class disease stage label. After leakage checks, 26 valid features were retained for training. The data were split into training, validation, and test sets as shown in Table 1. The stage-wise distribution of cases is provided in Table 2, showing a moderate imbalance, with stage 1 containing the highest samples (390) and stage 3 the lowest (222).

Table 1. Dataset and split sizes with feature and class counts

Split	Samples (n)	Features (p)	Classes (k)
Train	867	26	5
Validation	289	26	5
Test	290	26	5

Table 2. Stage-wise sample counts (original labels)

Stage	Count
1	390
2	262
3	222
4	325
5	247

3.2. Preprocessing

Preprocessing ensures that the dataset is leakage-free, normalized, and ready for training. To ensure reliability of the modeling process, preprocessing was performed in three steps. Leakage control: exploratory analysis revealed that the column class (cls) showed leakage, as it contained information directly correlated with the target [27]. This feature was dropped before model training. Details are shown in Table 3. Normalization: all numeric features were normalized to zero mean and unit variance to reduce scale differences. Encoding: categorical features such as hypertension (HTN), diabetes mellitus (DM), and anemia (ANE) were encoded into binary representations for compatibility with neural network inputs. The stepwise process is shown in Figure 2, which highlights how leakage features were removed, numeric variables normalized, and categorical variables encoded.

Table 3. Leakage checks and remediation

Column	Issue	Action
cls	Potential label leakage	Dropped before training

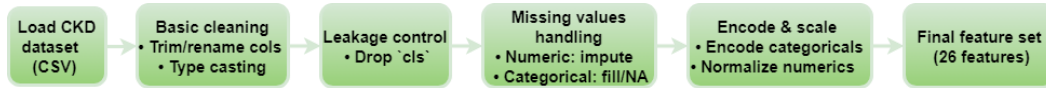


Figure 2. Preprocessing pipeline including leakage control, normalization, and encoding

3.3. Proposed models

3.3.1. AttentionNet

The AttentionNet incorporates an attention mechanism to focus on clinically significant features. The computational flow is shown in Figure 3, where feature embeddings are processed, attention weights are applied, and the aggregated representation is passed to the classifier [28]. The AttentionNet model is a feed-forward neural architecture augmented with an attention mechanism. The network computes weighted representations of hidden units using the relation (1). Where, α_{ij} is the attention coefficient, e_{ij} is the attention score, and h_j is the hidden representation. This mechanism allows the model to emphasize clinically relevant features such as GFR and SC during training.

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_k \exp(e_{ik})}, c_i = \sum_j \alpha_{ij} h_j \quad (1)$$

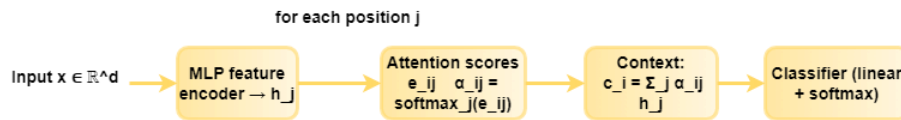


Figure 3. Schematic of AttentionNet highlighting feature weighting through attention mechanism

3.3.2. TabTransformer

The TabTransformer adapts self-attention layers to capture dependencies between features. The process is represented in Figure 4, which shows embedding layers, transformer blocks, and classification layers [29]. TabTransformer adapts the transformer encoder for tabular datasets. Its self-attention layers capture dependencies between categorical and numeric variables. The forward pass is expressed as in (2) and (3).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

$$H^{(l+1)} = LayerNorm(H^{(l)} + MultiHead(H^{(l)})) \quad (3)$$

Here Q, K, V represent query, key, and value matrices and d_k is the scaling factor.

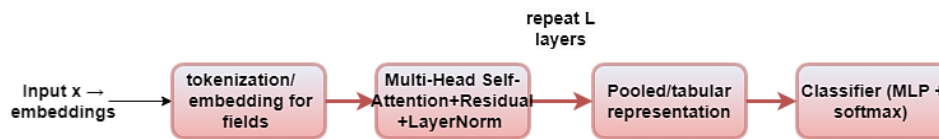


Figure 4. TabTransformer architecture adapted for CKD stage prediction

AttentionNet and TabTransformer were selected due to their suitability for moderate-sized clinical tabular datasets and their ability to model feature interactions without heavy pretraining requirements. Compared to architectures such as attentive interpretable tabular learning (TabNet) and feature tokenizer + transformer (FT-transformer), which often involve higher computational complexity and extensive hyperparameter tuning, the chosen models provide a balanced trade-off between performance, interpretability, and training stability. This makes them more practical for real-world clinical deployment scenarios where computational efficiency and reproducibility are critical.

3.4. Baseline models

To assess comparative performance, three widely used baselines were implemented. ResMLP introduces skip connections into standard MLPs for improved gradient flow. Multilayer perceptron baseline

(MLPBaseline) is a conventional fully connected neural network. WnD combines wide linear models with deep feature extractors, effective for tabular prediction.

3.5. Evaluation setup

Model evaluation was carried out with 10-fold CV to ensure robust generalization. The stepwise flow for CV is shown in Figure 5, illustrating how the dataset was partitioned into training and validation folds, and how metrics were aggregated across folds. Metrics included accuracy (Acc), F1-score, and ROC-AUC. ROC-AUC was obtained by integrating the true positive rate (TPR) against the false positive rate (FPR).

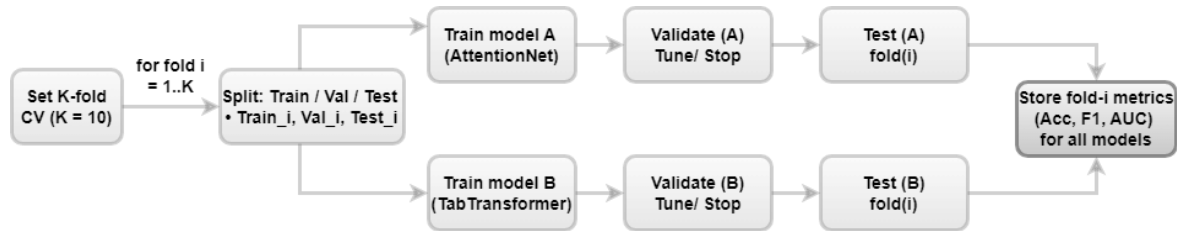


Figure 5. The 10-fold cross-validation framework for CKD stage prediction

4. RESULTS AND DISCUSSION

4.1. Overall performance

The average outcomes of the main models are presented in Table 4. AttentionNet reached an accuracy of 99.10% and an F1 of 99.10%, while TabTransformer performed similarly with an accuracy of 98.97% and an F1 of 98.96%. Both models achieved AUC values close to 1.000, confirming their reliability.

A broader comparison across all models is given in Table 5. ResMLP stood out with the best mean accuracy of 99.72% and F1 of 99.72%, followed by TabTransformer at 99.24%. This shows that DL methods designed for tabular data can outperform classical MLP baselines.

Table 4. Leakage-safe k-fold performance summary (mean $\pm$ std)

Model	Acc (mean $\pm$ std)	F1 (mean $\pm$ std)	AUC (mean $\pm$ std)
AttentionNet	99.10 $\pm$ 0.82	99.10 $\pm$ 0.82	0.99986 $\pm$ 0.00018
TabTransformer	98.97 $\pm$ 1.61	98.96 $\pm$ 1.62	0.99989 $\pm$ 0.00021

Table 5. Model comparison (10-fold CV, REMOVE FEAT=[])

Model	Acc mean (%)	Acc std	F1 mean (%)	F1 std	AUC mean (%)	AUC std
ResMLP	99.72	0.58	99.72	0.58	100.00	0.00
TabTransformer	99.24	1.24	99.23	1.25	100.00	0.01
MLPBaseline	99.17	0.85	99.17	0.85	99.99	0.02
WideAndDeep	98.69	1.15	98.69	1.15	99.95	0.08
AttentionNet	98.62	1.30	98.62	1.31	99.95	0.05

4.2. Per-fold cross validation

The detailed per-fold results are summarized in Table 6. The outcomes remained stable across all 10-folds, with accuracy values typically above 98%. The folds with lower values were rare and usually involved the AttentionNet, which still maintained a strong performance overall. This consistency highlights the robustness of the models under different splits of the data.

4.3. Ablation studies

The ablation outcomes are shown in Table 7. When GFR was excluded, ResMLP still achieved an accuracy above 99%, while AttentionNet dropped to about 91.8%. This indicates that attention-based models depend more strongly on GFR compared to ResMLPs. Notably, this represents a performance reduction of more than 7% for AttentionNet and over 5% for WnD models, clearly emphasizing the critical clinical importance of GFR in CKD stage prediction. Further results with both GFR and SC removed are given in Tables 8 and 9. ResMLP again maintained top performance, while WnD suffered the largest losses. This demonstrates the clinical importance of GFR and SC in predicting CKD stages.

Table 6. Per-fold results (10-fold CV, REMOVE FEAT=[])

Model	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
MLPBaseline	Acc	1.0	0.9931	0.9862	0.9931	0.9724	0.9862	1.0	1.0	0.9931	0.9931
	F1	1.0	0.9931	0.9863	0.9931	0.9722	0.9863	1.0	1.0	0.993	0.993
	AUC	1.0	0.9998	0.9999	1.0	0.9995	1.0	1.0	1.0	1.0	1.0
ResMLP	Acc	0.9862	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9861
	F1	0.9863	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9861
	AUC	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WideAndDeep	Acc	0.9931	0.9862	0.9724	0.9931	1.0	0.9655	1.0	0.9861	0.9931	0.9792
	F1	0.9931	0.9862	0.9725	0.9931	1.0	0.9653	1.0	0.9862	0.993	0.9791
	AUC	0.9999	0.9984	0.9996	0.9999	0.9998	0.9977	1.0	1.0	1.0	0.9998
AttentionNet	Acc	0.9931	0.9931	0.9655	1.0	0.9724	0.9724	1.0	0.9792	1.0	0.9861
	F1	0.9931	0.9931	0.9653	1.0	0.9722	0.9724	1.0	0.9793	1.0	0.9861
	AUC	0.9997	0.9991	0.9989	1.0	0.9994	0.9989	1.0	0.999	1.0	0.9996
TabTransformer	Acc	1.0	1.0	1.0	1.0	0.9724	1.0	1.0	0.9722	0.9792	1.0
	F1	1.0	1.0	1.0	1.0	0.9719	1.0	1.0	0.9722	0.979	1.0
	AUC	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9997	0.9999	1.0

Table 7. Ablation study (10-fold CV, REMOVE FEAT=['grf'])

Model	Acc mean (%)	Acc std	F1 mean (%)	F1 std	AUC mean (%)	AUC std
TabTransformer	99.17	0.79	99.17	0.79	99.97	0.04
ResMLP	99.03	0.74	99.04	0.74	99.96	0.04
MLPBaseline	97.30	1.32	97.29	1.34	99.84	0.18
WideAndDeep	94.06	2.67	94.08	2.63	99.34	0.43
AttentionNet	91.84	3.53	91.87	3.45	98.79	0.66

Table 8. Per-fold results (10-fold CV, REMOVE FEAT=['grf'])

Model	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
MLPBaseline	Acc	0.9724	0.9793	0.9931	0.9586	0.9517	0.9655	0.9722	0.9931	0.9722	0.9722
	F1	0.9724	0.9792	0.9931	0.9583	0.9506	0.9654	0.9717	0.9931	0.9724	0.9725
	AUC	0.9993	0.9994	1.0	0.9965	0.9942	0.999	0.9993	1.0	0.9979	0.998
ResMLP	Acc	0.9862	0.9931	0.9931	0.9862	1.0	0.9793	0.9792	0.9931	1.0	0.9931
	F1	0.9862	0.9931	0.9931	0.9862	1.0	0.9795	0.9793	0.9931	1.0	0.9931
	AUC	0.9991	0.9991	1.0	0.9997	1.0	0.9993	0.9994	0.9999	1.0	0.9995
WideAndDeep	Acc	0.9103	0.9172	0.9103	0.9448	0.9586	0.9448	0.9444	0.9861	0.9722	0.9167
	F1	0.9123	0.9167	0.9112	0.9448	0.9579	0.9452	0.944	0.9862	0.9724	0.9174
	AUC	0.9881	0.9891	0.9932	0.9886	0.9935	0.998	0.9953	0.9999	0.9978	0.9907
AttentionNet	Acc	0.9586	0.9241	0.8621	0.8759	0.9103	0.9103	0.9444	0.9444	0.9653	0.8889
	F1	0.9588	0.9244	0.8652	0.8765	0.9087	0.9109	0.9438	0.9436	0.9654	0.8899
	AUC	0.995	0.9835	0.9863	0.9785	0.9855	0.9814	0.9925	0.9954	0.9972	0.984
TabTransformer	Acc	0.9931	0.9931	0.9931	1.0	1.0	0.9862	0.9792	1.0	0.9931	0.9792
	F1	0.9931	0.9931	0.9931	1.0	1.0	0.9863	0.9793	1.0	0.993	0.9791
	AUC	0.9997	0.9993	0.9999	1.0	1.0	1.0	0.9989	1.0	0.9998	0.9992

Table 9. Top-2 models by accuracy under different ablation scenarios

Scenario	Top-1	Top-2
FULL (no removal)	ResMLP	MLPBaseline
NO_GRF (remove=['grf'])	ResMLP	MLPBaseline
NO_GRF_SC (remove=['grf','sc'])	ResMLP	TabTransformer

4.4. Confusion matrices

The confusion matrices highlight how each model classified CKD stages under different scenarios. With all features, ResMLP presented in Figure 6 and MLPBaseline presented in Figure 7 produced clean diagonals, while AttentionNet presented in Figure 8 and TabTransformer presented in Figure 9 also performed strongly with only minor overlap between stages 2 and 3. When GFR was removed, ResMLP presented in Figure 10 and MLPBaseline presented in Figure 11 continued to separate stages clearly. AttentionNet presented in Figure 12 showed a small rise in errors, while TabTransformer presented in Figure 13 remained stable.

The toughest case was removal of both GFR and SC. ResMLP as shown in Figure 14, still produced a strong diagonal, MLPBaseline, as shown in Figure 15, lost some sharpness, and AttentionNet as shown in Figure 16 displayed more confusion in the mid-stages. TabTransformer as shown in Figure 17, held up better, keeping most predictions on the diagonal.

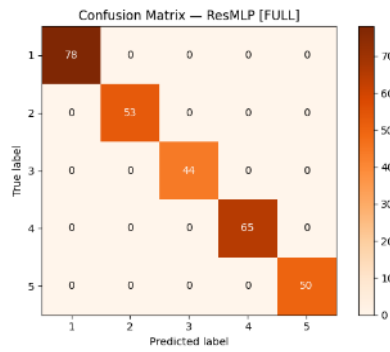


Figure 6. Confusion matrix—ResMLP [FULL]

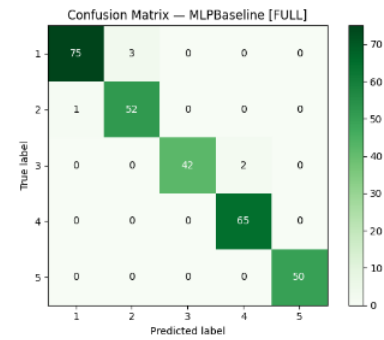


Figure 7. Confusion matrix—MLPBaseline [FULL]

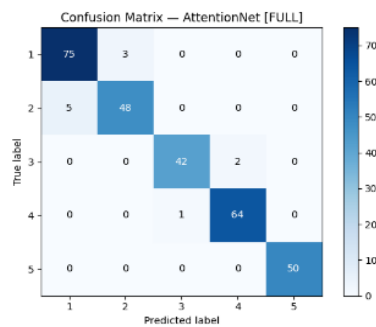


Figure 8. Confusion matrix—AttentionNet [FULL]

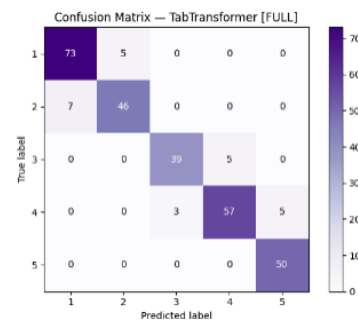


Figure 9. Confusion matrix—TabTransformer [FULL]

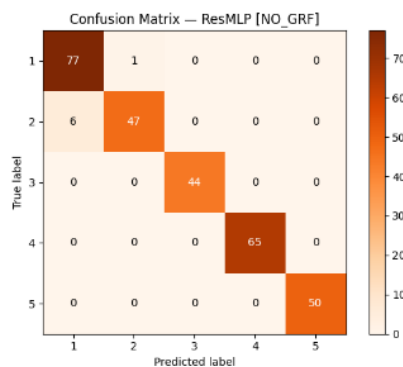


Figure 10. Confusion matrix—ResMLP [NO_GRF]

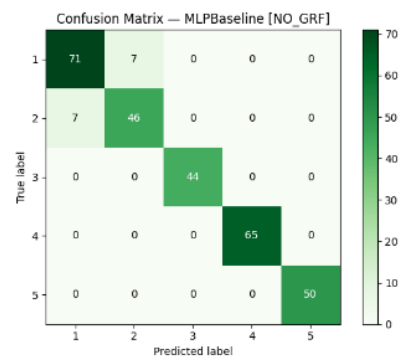


Figure 11. Confusion matrix—MLPBaseline [NO_GRF]

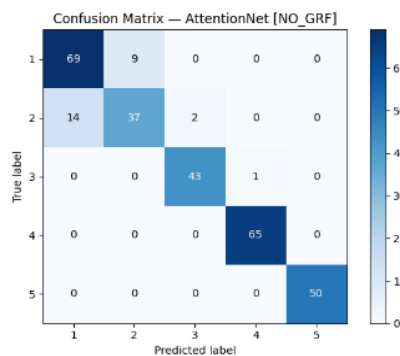


Figure 12. Confusion matrix—AttentionNet [NO_GRF]

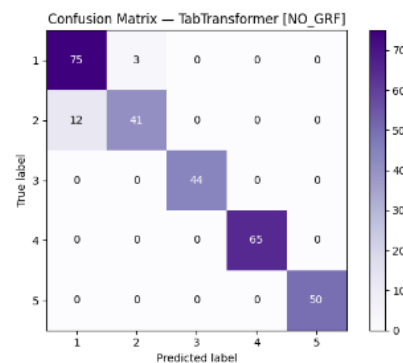


Figure 13. Confusion matrix—TabTransformer [NO_GRF]

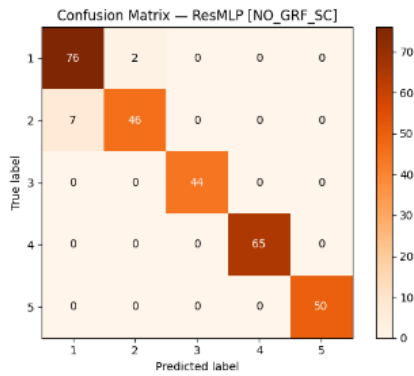


Figure 14. Confusion matrix—ResMLP [NO_GRF_SC]

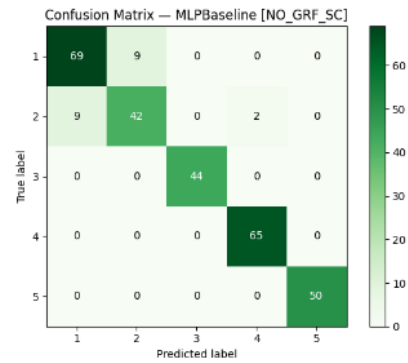


Figure 15. Confusion matrix—MLPBaseline [NO_GRF_SC]

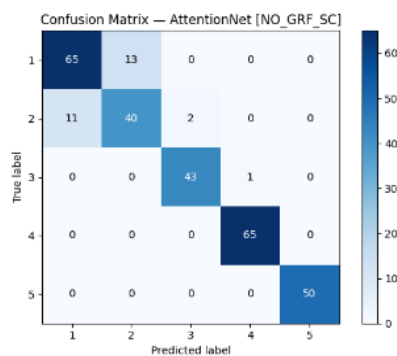


Figure 16. Confusion matrix—AttentionNet [NO_GRF_SC]

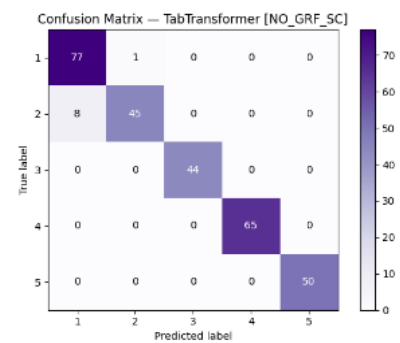


Figure 17. Confusion matrix—TabTransformer [NO_GRF_SC]

4.5. ROC curves

Figure 18 shows the macro-average ROC for the full feature set. ResMLP and MLPBaseline both reach AUC = 1.000. AttentionNet is at 0.999, WideAndDeep at 0.995, and TabTransformer at 0.994. The curves hug the top-left, indicating near-perfect separation. Figure 19 reports results after removing GFR. ResMLP remains strongest with AUC = 0.998. MLPBaseline is 0.996, TabTransformer 0.993, WideAndDeep 0.990, and AttentionNet 0.986. The small drop shows a mild dependency on GFR, but ranking stays similar. Figure 20 covers the hardest case with both GFR and SC removed. ResMLP still leads at AUC = 0.998 and TabTransformer follows at 0.997. MLPBaseline and WideAndDeep are 0.991, and AttentionNet is 0.987. Even with two key labs removed, overall discrimination stays high.

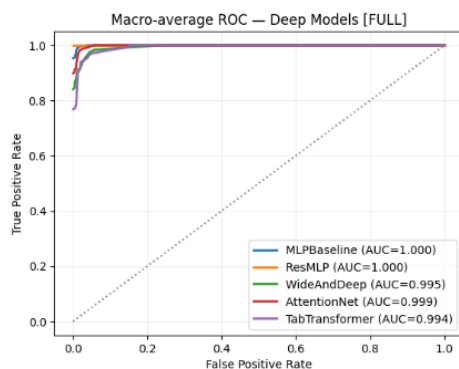


Figure 18. Macro-average ROC—deep models [FULL]

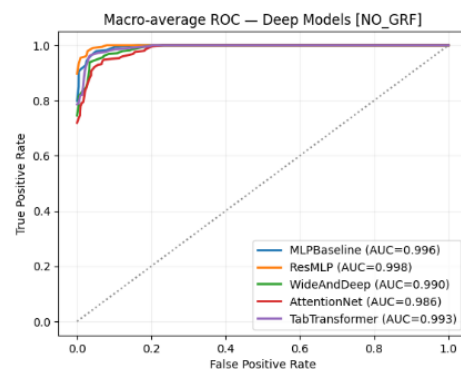


Figure 19. Macro-average ROC—deep models [NO_GRF]

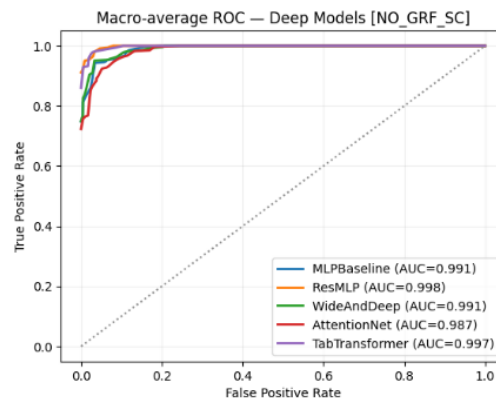


Figure 20. Macro-average ROC—deep models [NO_GRF_SC]

4.6. Model comparison (bar charts)

The bar plots compare Accuracy, F1, and AUC values for all deep models under the three experimental settings. In the FULL feature set, as shown in Figures 21 to 23, ResMLP reached the highest values, scoring 1.000 in both Accuracy and F1, and also achieving an AUC of 1.000. MLPBaseline followed with 0.979 for Accuracy and F1, and 0.999 for AUC. AttentionNet and TabTransformer recorded slightly lower values, while WnD was consistently the weakest of the group. When GFR was excluded, as shown in Figures 24 to 26, ResMLP still led with Accuracy and F1 around 0.976 and an AUC of 0.998. MLPBaseline stayed near 0.95 for Accuracy and F1 and 0.995 for AUC. TabTransformer and WnD showed stable performance, while AttentionNet declined more sharply, especially in AUC, where it dropped to 0.985. In the toughest case, with both GFR and SC removed as shown in Figures 27 to 29, ResMLP and TabTransformer both achieved 0.969 for Accuracy and F1, and near-perfect AUC values of 0.998 and 0.997. MLPBaseline remained above 0.93 for Accuracy and F1 with an AUC of 0.990, while AttentionNet stayed lowest but still respectable with Accuracy and F1 above 0.90 and an AUC of 0.986.

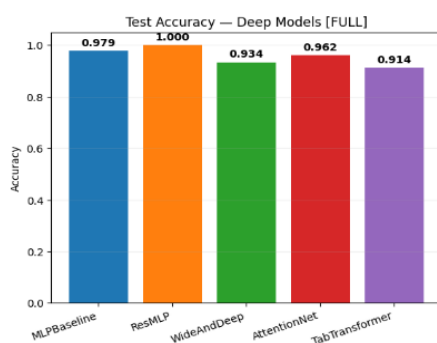


Figure 21. Accuracy of deep models [FULL]

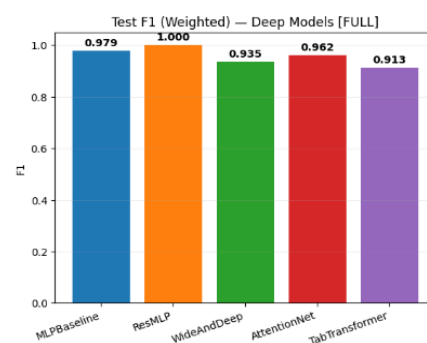


Figure 22. F1 of deep models [FULL]

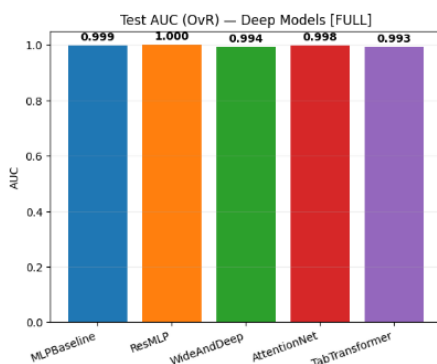


Figure 23. AUC of deep models [FULL]

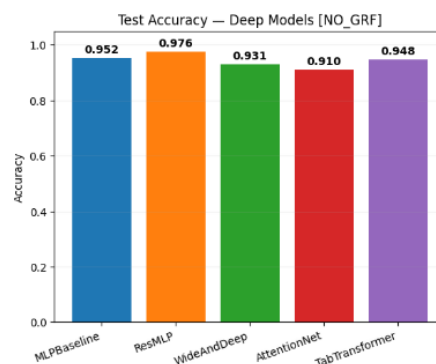


Figure 24. Accuracy of deep models [NO_GRF]

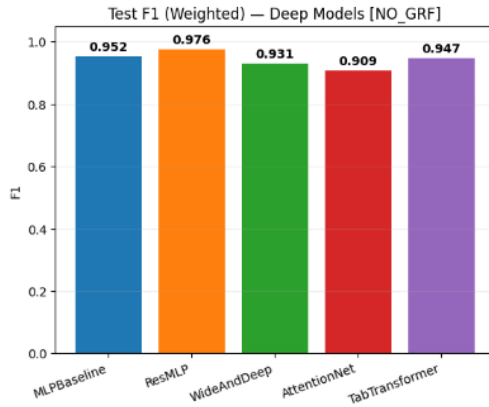


Figure 25. F1 of deep models [NO_GRF]

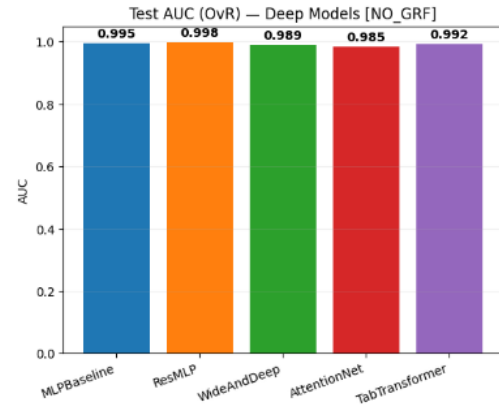


Figure 26. AUC of deep models [NO_GRF]

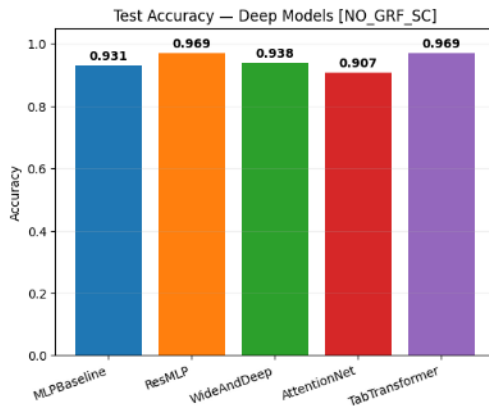


Figure 27. Accuracy of deep models [NO_GRF_SC]

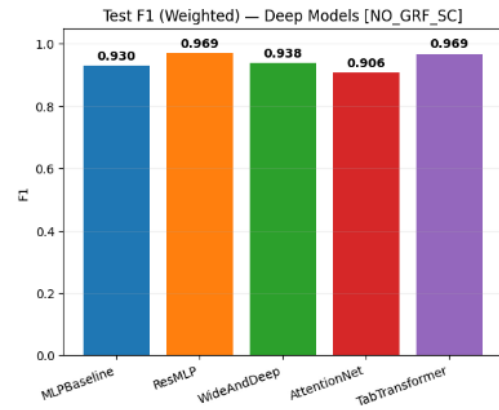


Figure 28. F1 of deep models [NO_GRF_SC]

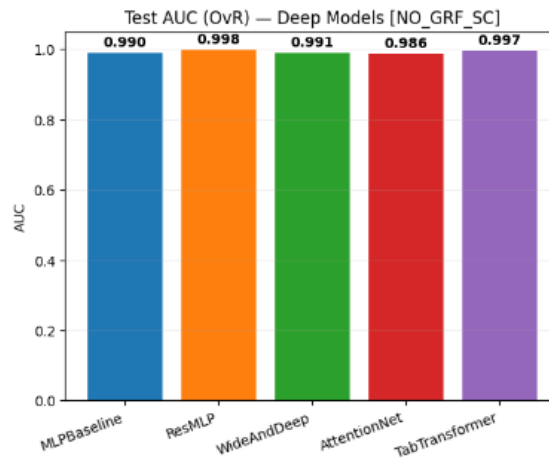


Figure 29. AUC of deep models [NO_GRF_SC]

When both GFR and SC were removed, all models experienced performance loss. ResMLP still achieved close to 0.97 accuracy and near-perfect AUC, making it the most resilient. TabTransformer followed, while AttentionNet dropped further, especially in the middle stages where overlap is greater. These findings confirm the clinical value of GFR and SC for CKD staging. The ablation study shows that

the proposed framework not only performs well overall but also aligns with established medical knowledge, strengthening its potential use in decision support. To further enhance interpretability, model-agnostic explanation methods such as Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) can be integrated in future extensions. SHAP enables global and local attribution of clinical features, while LIME provides instance-level explanations for individual predictions. These tools can improve transparency and clinician trust when deploying DL-based CKD stage prediction systems.

5. CONCLUSION

This study presented a DL-based framework for CKD stage prediction using AttentionNet and TabTransformer, alongside established baseline models. Across all experimental settings, ResMLP and TabTransformer consistently demonstrated strong and stable performance, achieving near-perfect accuracy, F1-score, and AUC when all features were available. Even under feature ablation scenarios, these models showed greater resilience compared to AttentionNet, indicating their robustness in handling partial clinical information. The ablation analysis further emphasized the clinical significance of GFR and SC in CKD staging. The pronounced performance degradation observed when these biomarkers were removed highlights their critical role in disease assessment and confirms that the proposed models align well with established medical understanding. Notably, attention-based models were more sensitive to the exclusion of these key indicators, underscoring differences in how model architectures utilize clinical features. Despite the promising results, certain limitations should be acknowledged. The dataset used in this study was obtained from a single clinical source and was moderate in size, which may limit generalizability across diverse populations. In addition, interpretability was primarily assessed through feature ablation, rather than through dedicated explanation techniques at the individual prediction level. Future research will therefore focus on validating the proposed framework using larger, multi-center datasets to assess robustness across demographic and clinical variations. Further work will also explore the integration of model-agnostic explainability methods to provide clinician-friendly insights into individual predictions. In addition, extending the current stage-based classification framework toward longitudinal and survival-aware modeling could offer a more comprehensive view of CKD progression. Ultimately, embedding such transparent and reliable predictive models into real-world clinical decision support systems will be essential for practical adoption and improved patient care.

ACKNOWLEDGMENTS

The authors would like to thank the management and staff of Ponjesly Super Speciality Hospital, Nagercoil, India, for their assistance in providing access to anonymized clinical data used in this study. The authors also acknowledge the support of the Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Chennai, for providing the research environment and technical resources necessary for this work.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Subashini N. J.	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Venkatesh Kaliyamurthy		✓				✓		✓		✓	✓	✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors state no conflict of interest.

INFORMED CONSENT

This study did not involve direct human participation or the use of personally identifiable data. The dataset consisted solely of anonymized numerical records from routine laboratory analyses. Therefore, informed consent was not required.

ETHICAL APPROVAL

The research did not involve any direct experimentation on human participants or animals. The anonymized dataset containing numerical values from blood and urine samples was obtained from Ponjesly Super Speciality Hospital, Nagercoil, India, with prior ethical approval from the Institutional Ethics Committee. The study complies with all relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [SNJ], upon reasonable request. The dataset was obtained under a confidentiality agreement with Ponjesly Super Speciality Hospital, and access is restricted to protect patient privacy.

REFERENCES

- [1] N. Liang, X. Ma, Y. Cao, T. Liu, J. Fang, and X. Zhang, "Mendelian randomization studies: opening a new window in the study of metabolic diseases and chronic kidney disease," *Endocrine, Metabolic and Immune Disorders-Drug Targets*, vol. 25, no. 6, pp. 442–457, Feb. 2024, doi: 10.2174/0118715303288685240808073238.
- [2] J. Chen *et al.*, "Global, regional, and national trends in chronic kidney disease burden (1990–2021): a systematic analysis of the global burden of disease in 2021," *Tropical Medicine and Health*, vol. 53, no. 1, Feb. 2025, doi: 10.1186/s41182-025-00703-x.
- [3] R. Kaur and S. Khattar, "Prediction of chronic kidney disease using machine learning algorithms," in *Soft Computing: Theories and Applications*, Singapore: Springer, 2025, pp. 201–211, doi: 10.1007/978-981-96-5958-6_18.
- [4] M. S. Arif, A. Mukheimer, and D. Asif, "Enhancing the early detection of chronic kidney disease: a robust machine learning model," *Big Data and Cognitive Computing*, vol. 7, no. 3, Aug. 2023, doi: 10.3390/bdcc7030144.
- [5] H. Iftikhar, M. Khan, Z. Khan, F. Khan, H. M. Alshanbari, and Z. Ahmad, "A comparative analysis of machine learning models: a case study in predicting chronic Kidney disease," *Sustainability*, vol. 15, no. 3, Feb. 2023, doi: 10.3390/su15032754.
- [6] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, Nov. 2017, doi: 10.1093/bib/bbx044.
- [7] S. Arık and T. Pfister, "TabNet: attentive interpretable tabular learning," in *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, May 2021, no. 8, pp. 6679–6687, doi: 10.1609/aaai.v35i8.16826.
- [8] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," *Journal of Pathology Informatics*, vol. 14, 2023, doi: 10.1016/j.jpi.2023.100189.
- [9] J. Aoki *et al.*, "CKD progression prediction in a diverse US population: a machine-learning model," *Kidney Medicine*, vol. 5, no. 9, Sep. 2023, doi: 10.1016/j.xkme.2023.100692.
- [10] F. Sanmarchi, C. Fanconi, D. Golinelli, D. Gori, T. H.-Boussard, and A. Capodici, "Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review," *Journal of Nephrology*, vol. 36, no. 4, pp. 1101–1117, Feb. 2023, doi: 10.1007/s40620-023-01573-4.
- [11] F. Chen, P. Kantagowit, T. Nopsopon, A. Chuklin, and K. Pongpirul, "Prediction and diagnosis of chronic kidney disease development and progression using machine-learning: protocol for a systematic review and meta-analysis of reporting standards and model performance," *PLoS ONE*, vol. 18, Feb. 2023, doi: 10.1371/journal.pone.0278729.
- [12] F. Khalid *et al.*, "Predicting the progression of chronic kidney disease: a systematic review of artificial intelligence and machine learning approaches," *Cureus*, May 2024, doi: 10.7759/cureus.60145.
- [13] Q. Pan and M. Tong, "Artificial intelligence in predicting chronic kidney disease prognosis. A systematic review and meta-analysis," *Renal Failure*, vol. 46, no. 2, Dec. 2024, doi: 10.1080/0886022X.2024.2435483.
- [14] M. A. I.-Ruget *et al.*, "Predicting chronic kidney disease progression with artificial intelligence," *BMC Nephrology*, vol. 25, no. 1, Apr. 2024, doi: 10.1186/s12882-024-03545-7.
- [15] E. Battistella, D. Ghiassian, and A. L. Barabási, "Improving the performance and interpretability on medical datasets using graphical ensemble feature selection," *Bioinformatics*, vol. 40, no. 6, Jun. 2024, doi: 10.1093/bioinformatics/btae341.
- [16] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artificial Intelligence Review*, vol. 57, no. 10, Sep. 2024, doi: 10.1007/s10462-024-10884-2.
- [17] L. Wang, H. Ni, R. Yang, V. Pappu, M. B. Fenn, and P. M. Pardalos, "Feature selection based on meta-heuristics for biomedicine," *Optimization Methods and Software*, vol. 29, no. 4, pp. 703–719, Jul. 2014, doi: 10.1080/10556788.2013.834900.
- [18] Z. Yang, A. Mitra, W. Liu, D. Berlowitz, and H. Yu, "TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records," *Nature Communications*, vol. 14, no. 1, Nov. 2023, doi: 10.1038/s41467-023-43715-z.

- [19] F. Veisi, "Enhancing breast cancer detection: FT-transformer vs. traditional models," in *International Conference on Artificial Intelligence, Computer, Data Sciences, and Applications, ACDSA 2025*, 2025, pp. 1–7, doi: 10.1109/ACDSA65407.2025.11166464.
- [20] H. Y. J. Kang, M. Ko, and K. S. Ryu, "Tabular transformer generative adversarial network for heterogeneous distribution in healthcare," *Scientific Reports*, vol. 15, no. 1, Mar. 2025, doi: 10.1038/s41598-025-93077-3.
- [21] I. I. Ilyas, S. Boukari, and A. Y. Gital, "Recent trends in prediction of chronic kidney disease using different learning approaches: a systematic literature review," *Journal of Medical Artificial Intelligence*, vol. 8, pp. 62–62, Dec. 2025, doi: 10.21037/jmai-24-256.
- [22] A. P. Kengne, C. George, and O. I. Ameh, "Predicting the outcomes of chronic kidney disease in older adults," *BMJ*, vol. 385, Apr. 2024, doi: 10.1136/bmj.q749.
- [23] Z. Zhao, R. Birke, and L. Y. Chen, "TabuLa: harnessing language models for tabular data synthesis," in *Advances in Knowledge Discovery and Data Mining*, Singapore: Springer, 2025, pp. 247–259, doi: 10.1007/978-981-96-8186-0_20.
- [24] M. E. Haque, S. M. J. Islam, J. Maliha, M. S. H. Sumon, R. Sharmin, and S. Rokoni, "Improving chronic Kidney disease detection efficiency: fine-tuned CatBoost and nature-inspired algorithms with explainable AI," in *2025 IEEE 14th International Conference on Communication Systems and Network Technologies, CSNT 2025*, Mar. 2025, pp. 811–818, doi: 10.1109/CSNT64827.2025.10968421.
- [25] N. G. Rezk, S. Alshathri, A. Sayed, and E. E. D. Hemdan, "Explainable AI for chronic kidney disease prediction in medical IoT: integrating GANs and few-shot learning," *Bioengineering*, vol. 12, no. 4, Mar. 2025, doi: 10.3390/bioengineering12040356.
- [26] M. Zisser and D. Aran, "Transformer-based time-to-event prediction for chronic kidney disease deterioration," *Journal of the American Medical Informatics Association*, vol. 31, no. 4, pp. 980–990, Apr. 2024, doi: 10.1093/jamia/ocae025.
- [27] G. A. L.-Ramirez, A. A.-Zavala, and C. V.-Rosales, "Exploratory data analysis for path loss measurements: unveiling patterns and insights before machine learning," *IEEE Access*, vol. 12, pp. 62279–62295, 2024, doi: 10.1109/ACCESS.2024.3394904.
- [28] Z. Rahman, R. Zhang, and J. A. Bhutto, "A symmetrical approach to brain tumor segmentation in MRI using deep learning and threefold attention mechanism," *Symmetry*, vol. 15, no. 10, Oct. 2023, doi: 10.3390/sym15101912.
- [29] X. Liu, L. Su, J. Gao, Q. Cui, and B. Wang, "An interpretable analytical prediction method for heterogeneous tabular data of electricity load based on TabTransformer," *Journal of Computational Methods in Sciences and Engineering*, Aug. 2025, doi: 10.1177/14727978251371208.

BIOGRAPHIES OF AUTHORS



Mrs. Subashini N. J.     is a research scholar in the Department of Networking and Communications at SRM Institute of Science and Technology, Kattankulathur, Chennai, India. She received her B.E. degree in Computer Science and Engineering and M.E. degree in Computer Science and Information Technology. She is currently pursuing her Ph.D. in the field of machine learning for biomedical applications. Her research interests include artificial intelligence, machine learning, and biomedical systems. She has published several papers in these areas, contributing valuable insights to the academic community. She can be contacted at email: suba.satheesh@gmail.com.



Dr. Venkatesh Kaliyamurthy     received his Ph.D. degree in Computer Science and Engineering and is currently an associate professor in the Department of Networking and Communications at SRM Institute of Science and Technology, Kattankulathur, Chennai, India. His areas of expertise include networking, cybersecurity, artificial intelligence, machine learning, and SD-WAN technologies. He has published numerous papers in reputed international journals and conferences and has guided several research scholars. His current research focuses on intelligent network optimization and security in distributed systems. He can be contacted at email: venkatek2@srmist.edu.in.