Pertanika J. Sci. & Technol. 34 (S3): 101 - 124 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

GenCKD-Net: A Lightweight Multilayer Perceptron Pipeline for Chronic Kidney Disease Classification Using the UCI CKD Dataset

Karimulla Sha Shaik[1]*, Konda Hari Krishna[1], and K. K. Baseer[2]

[1]Department of Computer Science and Engineering, School of Computing, Mohan Babu University, Sree Sainath Nagar, 517102 Tirupati, Andhra Pradesh, India

[2]Department of Computer Science and Systems Engineering, GITAM School of Computer Science and Engineering, GITAM (Deemed to be a University), NH 207, Nagadenehali, Doddaballapur Taluk, 561203 Bengaluru, Karnataka, India

ABSTRACT

Binary classification of chronic kidney disease (CKD) using standard clinical measurements represents a relevant machine-learning application, but such estimates may not generalise to independent clinical populations. GenCKD-Net, a lightweight multilayer perceptron pipeline for binary CKD classification using the publicly available UCI Chronic Kidney Disease dataset. The dataset contains 400 records and 24 predictor variables, including 250 CKD cases (62.5%, majority class) and 150 non-CKD controls (37.5%, minority class). The dataset was divided by stratified sampling into training (64%), validation (16%), and held-out test (20%) subsets. Missing-value imputation, categorical encoding, and numerical standardisation were fitted using the training data only, while SMOTE was applied only to the transformed training subset. GenCKD-Net uses a compact two-hidden-layer MLP with 128 and 64 neurons and is implemented in scikit-learn with warm-start training across 10 single-epoch fitting cycles. Classification threshold selection utilised validation data, prioritising specificity, and was not selected from the test set. On the held-out test set (n=80), GenCKD-Net achieved ROC-AUC = 0.998, average precision = 1.000, accuracy = 0.975, precision = 1.000, recall = 0.960, and F1-score = 0.980. Beginning with baseline classifiers, all architectures achieved comparable results, indicating that achieving optimal discrimination on this benchmark can be accomplished by many different architectures. The results indicate that GenCKD-Net is a research-stage tool with potential for use in benchmarking and direct clinical application. To translate into a clinical setting, further validation

ARTICLE INFO

Article history:
Received: 03 August 2026
Accepted: 29 September 2026
Published: 20 November 2026

DOI: https://doi.org/10.47836/pjst.34.S3.05

E-mail addresses:
karimullashah@gmail.com (Karimulla Sha Shaik)
konda.harikrishna@mbu.asia (Konda Hari Krishna)
kkamalap@gitam.edu.in (K. K. Baseer)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

would include independent cohort verification, repeated cross-validation, a calibration study, and prospective studies.

Keywords: Average precision, baseline comparison, chronic kidney disease, class imbalance, machine learning, multilayer perceptron, ROC-AUC, SMOTE

INTRODUCTION

Managing the growing problem of chronic kidney disease (CKD), which has become a serious challenge for nephrology, is another issue because it requires the identification of those with CKD who could potentially need additional evaluation. The health care system generates an immense amount of clinical information, including laboratory test results and comorbidities, that can be used to automate patient classification and risk assessment. While there have been many recent reviews of the use of artificial intelligence (AI) and machine learning in chronic kidney disease (CKD) that indicate potential applications of these computing methods, they also highlight several challenges, including maintaining high levels of data quality, ensuring that users can understand the results produced by the models, and validating the output through clinical trials (Gogoi & Valan, 2025; Sabanayagam et al., 2025; Sanmarchi et al., 2023).

Motivation: Burden of CKD and Need for Early Detection

Automated detection of CKD through clinical data is an important public health issue. Therefore, this paper utilises the standard UCI benchmark for identifying patients with CKD from routine clinical data, employs a binary approach to classify patients as having CKD or not, and does NOT attempt to model longitudinal disease progression or outcomes.

Challenges: Class Imbalance and Model Interpretability

Clinical classification problems often involve class imbalances. The UCI dataset used in this research has approximately 62.5% of the data classified as CKD and approximately 37.5% as not having CKD. Although increasing the amount of representation of a minority class will allow you to better utilise SMOTE, care must be taken when using SMOTE to prevent leaking information into your test set by applying it only to your training data (Burra et al., 2025).

Figure 1 uses a wheel-style layout to show common CKD symptoms—fatigue, poor appetite, nausea, swelling (oedema), frequent urination, sleep disturbances, itching, and shortness of breath. Because these signs are often nonspecific and can appear early or worsen as kidney function declines, noticing them together with routine lab results can prompt earlier evaluation and referral.

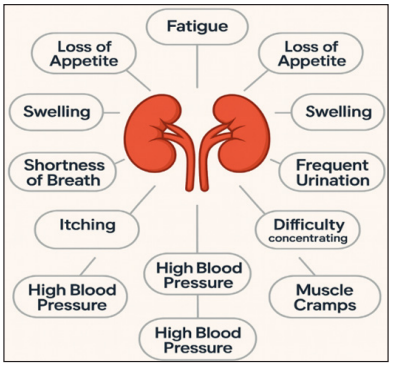


Figure 1. Common symptoms of Chronic Kidney Disease (CKD)

Contributions

GenCKD-Net is a lightweight multilayer perceptron pipeline for reproducible CKD classification on the UCI CKD benchmark. The contribution is methodological and empirical rather than architectural: the study evaluates a compact MLP within a leakage-controlled preprocessing and training workflow and compares its performance directly with established machine-learning baselines under the same evaluation framework.

Figure 2 presents the conceptual GenCKD-Net workflow used in this study: routine tabular clinical features undergo leakage-controlled preprocessing, training-only SMOTE, and classification by a compact 128→64 MLP. Imaging, wearables, explainability, and deployment components are not part of this experiment.

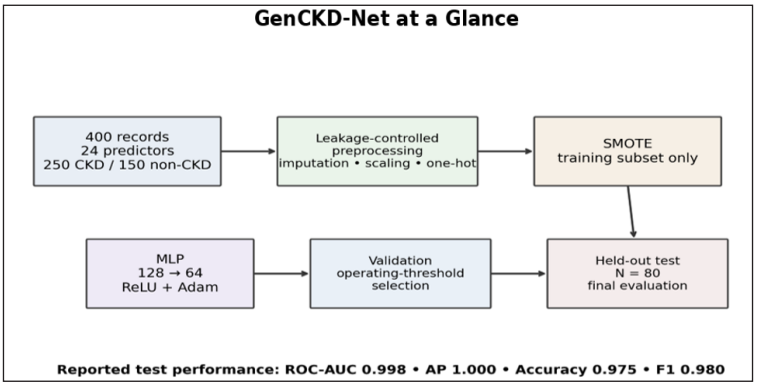


Figure 2. GenCKD-Net: Leakage-controlled tabular CKD classification

Related Work

Computational analysis of kidney disease classification has long employed a variety of traditional machine learning approaches. The most common have included Support Vector Machines (SVMs), Logistic Regression (LR), Random Forests (RFs), and Gradient Boosted Models (GBMs). These methods are well represented as very successful at predicting disease based on the collected data. However, many recent reviews have highlighted variability in data collection and validation procedures across studies, making it difficult to compare results from one study to another (Akter et al., 2023; Gogoi & Valan, 2025; Iliyas et al., 2025; Sanmarchi et al., 2023).

AI/ML in CKD Screening and Prognosis; Effects of Class Imbalance; Limits of Black-Box DL

A large body of CKD modelling relies on routinely collected EHR and laboratory data to flag disease and forecast progression. Recent work shows that benchmark datasets can yield strong performance, but differences in validation design and population hinder direct comparisons across studies. This study, therefore, emphasises baseline comparability and cautious interpretation rather than claims of clinical superiority. The broader evidence base includes systematic reviews of incident CKD prediction in community electronic health records (Haris et al., 2024) and AI-based CKD prognosis (Pan & Tong, 2024), both of which underscore the importance of validation design and population context. (Miller & Dwyer, 2025; Nadarajah et al., 2023; Su et al., 2022)

Explainable AI for Tabular Health Data (Permutation, SHAP, Counterfactuals)

Comprehensible model decisions facilitate clinical acceptance and appropriate utilisation, and under what conditions may they fail? However, permutation importance, SHAP, and counterfactual explanations are not experimentally evaluated in this study; they are identified as possible future extensions rather than reported findings.

Explainability can support model auditing by identifying which input variables contribute to predictions and by revealing potential instability or subgroup-specific behaviour. However, GenCKD-Net did not experimentally evaluate permutation importance, SHAP, counterfactual explanations, or clinician-facing interfaces. These methods are therefore treated as future analyses that should be performed only after the predictive model has been evaluated across repeated splits and independent cohorts (Lekadir et al., 2025; Rosenbacke et al., 2024). Reviews of AI in kidney disease and nephrology also emphasise the challenges of interpretability, clinical integration, and validation (Ali et al., 2022; Wu et al., 2024; Singh et al., 2025).

Multimodal approaches using imaging, wearables, point-of-care measurements, or omics may provide additional information for CKD research. In this study, however, only tabular clinical variables from the UCI CKD benchmark were evaluated. Multimodal fusion is therefore considered future work requiring separate data, validation, and ablation studies.

Multimodal approaches are increasingly being investigated to characterise renal disease by combining tabular clinical variables with imaging, longitudinal sensor measurements, pathology, or molecular data. Imaging-based studies have explored automated extraction of renal structural and pathological characteristics, while wearable and point-of-care technologies provide potentially useful longitudinal signals. However, multimodal models require appropriately aligned patient-level data, modality-specific preprocessing, adequate sample sizes, and independent validation. These requirements are substantially different from those of this study, which is restricted to tabular clinical variables. Therefore,

multimodal fusion is considered a future extension rather than a demonstrated capability of GenCKD-Net (Han et al., 2024; Puri et al., 2025; Zhang et al., 2024). Imaging-focused CKD studies provide a distinct evidence stream from the tabular benchmark used here; examples include kidney MRI classification and segmentation (Islam et al., 2024) and broader retinal-image AI screening research (Fang et al., 2025; Sághy et al., 2025; Zhang, 2025)

MATERIALS AND METHODS

Dataset

The data for this study were based on the Chronic Kidney Disease Data Set available in the UCI Repository. The data set is identified as #336 by Rubini et al. (2015); it has 400 samples with 24 features or predictors, which are a combination of demographic information about patients, their vital signs, urinalysis, blood biochemistry tests, and various blood test parameters that can be used to assess kidney function and other medical conditions related to kidney disease. Of these, there were 250 CKD (chronic kidney disease) cases, accounting for 62.5% of all records, and 150 non-CKD (non-chronic kidney disease) cases, accounting for 37.5%. The dataset description and benchmark provenance are consistent with the UCI CKD dataset record (Rubini et al., 2015); this study does not treat the dataset as a longitudinal cohort.

Table 1 summarises the predictor schema used in the UCI CKD benchmark, including numerical laboratory/vital measurements and categorical clinical and urine-related attributes. The dataset contains serum creatinine, blood urea, haemoglobin, electrolyte measurements, blood pressure, and categorical findings such as albumin, sugar, red blood cell status, pus cell status, and oedema. Several features contained missing entries; numerical features were imputed with the median value from the training set, while categorical features were imputed with the most frequent category, both imputed values fitted only on the training data. Numerical variables were standardised using training-set statistics, and categorical variables were one-hot encoded before training-only SMOTE was applied.

Pre-Processing

Careful preprocessing was used to prevent potential "information leakage" during testing/validation by preventing information from being carried over from the training set. The entire data set was split by stratified partitioning, creating three subsets: a training subset (64% of the data), a validation subset (16%), and a testing subset (20%). All pre-processed transformation models were fit exclusively on the training subset. After that, these pre-processed transformation models were applied to both the validation and test subsets, and no further fitting was performed. Numerical missing values were replaced using the median

Table 1
Dataset schema and feature summary

id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc
0	48	80	1.02	1	0	-	normal	notpresent	notpresent	121	36	1.2	-	-	15.4	44	7800	5.2
1	7	50	1.02	4	0	-	normal	notpresent	notpresent	-	18	0.8	-	-	11.3	38	6000	-
2	62	80	1.01	2	3	normal	normal	notpresent	notpresent	423	53	1.8	-	-	9.6	31	7500	-
3	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111	2.5	11.2	32	6700	3.9
4	51	80	1.01	2	0	normal	normal	notpresent	notpresent	106	26	1.4	-	-	11.6	35	7300	4.6
5	60	90	1.015	3	0	-	-	notpresent	notpresent	74	25	1.1	142	3.2	12.2	39	7800	4.4
6	68	70	1.01	0	0	-	normal	notpresent	notpresent	100	54	24	104	4	12.4	36	-	-
7	24	-	1.015	2	4	normal	abnormal	notpresent	notpresent	410	31	1.1	-	-	12.4	44	6900	5
8	52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.9	-	-	10.8	33	9600	4
9	53	90	1.02	2	0	abnormal	abnormal	present	notpresent	70	107	7.2	114	3.7	9.5	29	12100	3.7
10	50	60	1.01	2	4	-	abnormal	present	notpresent	490	55	4	-	-	9.4	28	-	-
11	63	70	1.01	3	0	abnormal	abnormal	present	notpresent	380	60	2.7	131	4.2	10.8	32	4500	3.8
12	68	70	1.015	3	1	-	normal	present	notpresent	208	72	2.1	138	5.8	9.7	28	12200	3.4
13	68	70	-	-	-	-	-	notpresent	notpresent	98	86	4.6	135	3.4	9.8	-	-	-
14	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6
15	40	80	1.015	3	0	-	normal	notpresent	notpresent	76	162	9.6	141	4.9	7.6	24	3800	2.8
16	47	70	1.015	2	0	-	normal	notpresent	notpresent	99	46	2.2	138	4.1	12.6	-	-	-

value of the training data set. Numerical feature values were scaled using the mean and standard deviations of the training data set. Categorical missing values were replaced using the most common category from the training data set. Then, each categorical variable was one-hot encoded. A SMOTE-transformed version of the training matrix was created, after which no sampling was performed on either the validation or test matrices. Equations 1 to 5 formalise the numerical and categorical preprocessing operations described here; the fitted transformations are learned from the training subset and then applied unchanged to validation and test data.

Data normalisation and imputation stabilise the heterogeneous clinical variables before training:

$$x' = \frac{x - \text{Median}(X)}{\sigma_X} \quad [1]$$

$$x_{\text{imp}} = \begin{cases} x, & \text{if not missing} \\ \text{Median}(X), & \text{if missing} \end{cases} \quad [2]$$

Categorical variables are represented with one-hot encodings:

$$O_{i,j} = \begin{cases} 1, & \text{if } x_i = \text{category}_j \\ 0, & \text{otherwise} \end{cases} \quad [3]$$

Imputation of categorical attributes utilises mode frequency:

$$c_{\text{imp}} = \arg\max_k \ \text{freq}(c_k) \quad [4]$$

The complete preprocessing pipeline is described as

$$\mathbf{X}_{proc} = f_{\text{onehot}}\left(f_{\text{zscale}}\left(f_{\text{impute}}(\mathbf{X})\right)\right) \quad [5]$$

Imbalance Mitigation

Because CKD is the majority class (62.5%) in our dataset, class imbalance was addressed by applying SMOTE only to the minority class. SMOTE was applied only to the transformed training data after the train/validation/test split. The validation and test sets retained their original class distributions and were not resampled. The minority-class definition and training-only resampling procedure form the basis of the leakage-control protocol; Equation 9 specifies the synthetic-sample interpolation, and Equation 10 specifies the resulting training objective.

The non-CKD cases form the minority class (37.5%). Conventional SMOTE generates synthetic non-CKD examples by interpolation in the transformed feature space. Because SMOTE was applied after one-hot encoding, categorical-aware resampling, such as SMOTENC, remains a methodological sensitivity analysis for future work.

The preprocessing and imbalance-mitigation pipeline was implemented in leakage-controlled order: first, split the data using stratified sampling; then fit imputation, encoding, and scaling on the training subset only; transform the validation and test subsets without refitting; and finally apply SMOTE only to the transformed training matrix.

Algorithm 1: Preprocessing and Imbalance Mitigation

Input: Dataset D with numeric features N, categorical features C, labels Y
Output: Pre-processed training, validation, and test matrices

1: Begin
2: Split D into Train, Validation, and Test using stratified sampling
3: Fit numerical imputation, scaling, categorical imputation, and one-hot encoding on Train only.
4: Transform Validation and Test using the fitted training transformations
5: For each numeric feature n μN, replace missing values with the training-set median.
6: Standardise n' = (n − μN) / σN using training-set μN and σN
7: One-hot encode categorical variables using the training-fitted encoder
8: Apply SMOTE only to the transformed training matrix.
9: Keep validation and test unchanged.
10: Return processed train, validation, and test sets
11: End

Model Architecture

GenCKD-Net is a compact MLP with two hidden layers (128 → 64, ReLU). The model uses Adam optimisation and produces a positive-class probability score. The implementation does not include a separately calibrated probability layer; therefore, the output is treated as a model score/probability estimate rather than evidence of formal calibration. The model operations are summarised in Equations 11 and 12, which define the output mapping and the parameter count expression for the compact MLP.

The predictive component of GenCKD-Net uses a Multilayer Perceptron (MLP) with two hidden layers (128 and 64 neurons), implemented using scikit-learn. The compact architecture is intended as a lightweight benchmark model for tabular CKD classification.

Non-linear activations learn higher-level CKD representations:

$$h^{(1)} = \sigma\left(W^{(1)}x + b^{(1)}\right) \quad [9]$$

$$h^{(2)} = \sigma\left(W^{(2)}h^{(1)} + b^{(2)}\right) \quad [10]$$

The sigmoid function decodes output as disease probability:

$$\hat{y} = \frac{1}{1 + e^{-(W^{(3)}h^{(2)} + b^{(3)})}} \quad [11]$$

Total trainable parameters are

$$P = \sum_{l=1}^{L} (d_{l-1} \cdot d_l + d_l) \quad [12]$$

Training Protocol

The datasets were segregated by function. Training data were used to update parameters after training-only SMOTE was applied. Validation data were used to select the operating threshold. The test data remained untouched until the final evaluation. No early-stopping procedure was used in the implementation.

The model was optimised using binary cross-entropy through the scikit-learn MLP implementation with the Adam optimiser. Warm-start fitting was used with 1 iteration per fitting pass, and the fitting procedure was repeated for 10 recorded passes. The validation subset was used for operating-threshold selection; the held-out test subset was not used for model fitting or threshold optimisation.

The network minimises a cross-entropy objective as shown in Equations 13 and 14:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad [13]$$

Adaptive moment updates are performed as

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad [14]$$

Algorithm 2: GenCKD-Net Model Training Protocol. The implementation uses warm-start fitting with one iteration per fitting pass. Training and validation metrics are recorded for 10 passes; no automatic early-stopping criterion is applied in this experiment.

Implementation Details and Explainability Scope

This study evaluates only the classification pipeline and benchmark performance. Explainability mechanisms such as permutation importance, SHAP, and counterfactual analysis were not experimentally evaluated or presented as results.

Algorithm 2: GenCKD-Net Model Training

Input: Preprocessed training data Dtrain, number of fitting passes E = 10Output: Trained model parameters Θ*

1: Begin
2: Initialise model parameters and configure Adam
3: For pass = 1 to E do
4: Fit the MLP for 1 iteration using Dtrain.
5: Record training loss and accuracy
6: Evaluate validation loss and accuracy without parameter updates
7: End For
8: Return final model parameters Θ*.
9: End

No explainability module was executed as part of this experiment. Such methods are retained for future analyses after repeated internal validation and independent external evaluation.

Permutation importance is identified as a future global interpretation method; no permutation-importance results are reported.

SHAP is identified as a future patient-level explanation method; no SHAP values or feature-attribution results are reported.

Counterfactual analysis is identified as a future sensitivity analysis; no counterfactual results are reported.

Permutation-based feature significance as shown in Equations 16 to 18:

$$I_j = \frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{Acc}(D) - \mathrm{Acc}\left(D_{\pi(j)}\right)\right) \quad [16]$$

Shapley additive explanations are

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left(f(S \cup \{j\}) - f(S)\right) \quad [17]$$

The counterfactual sample search is formulated as

$$x^* = \arg\min_{x} \|x' - x\| s.t. f(x') \neq f(x) \quad [18]$$

Algorithm 3: Explainability Integration Module — Future Work

Algorithm 3 is a conceptual future work procedure for global and patient-level explainability. It was not executed in this study, and its outputs are not interpreted as reported experimental findings.

Algorithm 3: Conceptual Future Explainability Module

Input: Frozen trained model Θ* and an independently selected analysis dataset
Output: Candidate explanation summaries for future validation

1: Begin
2: Compute permutation importance on an analysis dataset
3: Compute SHAP values for representative samples
4: Generate counterfactual candidates subject to clinically plausible constraints
5: Store explanation outputs for qualitative and quantitative audits.
6: Validate explanation stability before any clinical interpretation
7: Return future explanation summaries
8: End

Multimodal Extension (Future Work)

This research focuses only on tabular UCI CKD data. Imaging, wearables/point-of-care measurements, omics, multimodal fusion, and clinical decision support deployment were not included in these experiments. These experiments are considered future research directions.

Figure 3 depicts the actual tabular experimental pipeline: preprocessing is fitted on the training subset, SMOTE is applied only to the training data, the MLP is trained using warm-start fitting, and validation is used to select the operating threshold before final test evaluation.

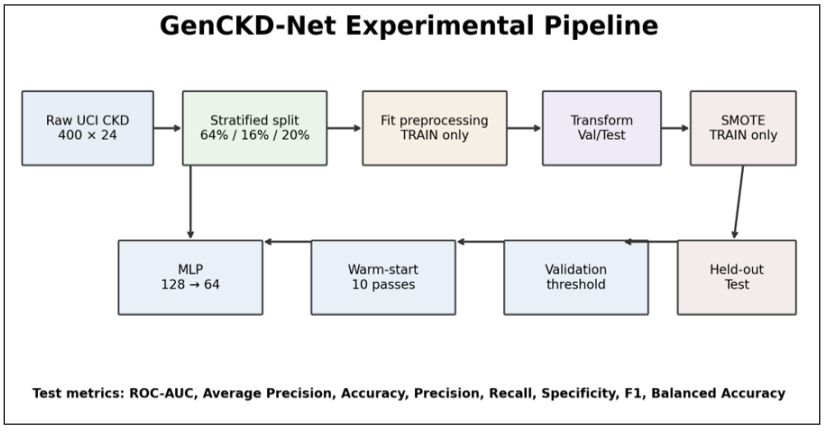


Figure 3. GenCKD-Net experimental pipeline

Evaluation Metrics

An extensive number of evaluation metrics were used to assess the GenCKD-Net system as follows: GenCKD-Net's accuracy was measured using a variety of metrics that assessed its ability to make accurate predictions (log-loss); evaluated GenCKD-Net's discrimination capability independent of threshold through the area under the receiver operating characteristic curve (ROC-AUC); provided the precision-recall average (PR-AP) as an additional metric to evaluate GenCKD-Net's performance when there is a substantial disparity in the amount of positive and negative outcome data; and utilised the confusion matrix (which provides true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN)) to increase the operational interpretability of the system. Additionally, we measured the time required to process each training epoch to assess the cost of training the GenCKD-Net system. Upon testing the GenCKD-Net system on our held-out test set, we found that GenCKD-Net achieved the following metrics: AUC = 0.998 (95% CI: 0.994-1.000), PR-AP ≈ 1.00, Accuracy = 0.975, Precision = 1.00, Recall = 0.96, F1 ≈ 0.98, Specificity = 1.00, and BA ≈ 0.98 with no FP and two FN. These results provide evidence that GenCKD-Net has clinical implications related to how a physician would prioritise clinical criteria to arrive at a diagnosis. For example, if the goal of GenCKD-Net is population screening, then increased recall would be clinically advantageous; however, based on our current findings, we believe that additional fine-tuning of the classification threshold may provide even greater opportunities to minimise missed cases while maintaining near-perfect precision. If the purpose of GenCKD-Net is to serve as a diagnostic tool to assist clinicians with confirming disease status, the absence of false positives provides substantial reassurance. Lastly, the results reported herein were derived from a single well-defined split of the data; therefore, our future research will include nested cross-validation, external validation, and calibration curves to establish the reliability and generalisability of our models across large datasets. All the performance metrics described above were evaluated using a set of metrics. These basic metrics included accuracy (Equation 19), precision (Equation 20), recall (Equation 21), and specificity (Equation 22). The harmonic mean of the first three was denoted as the F1 score (Equation 23), and balanced accuracy (Equation 24) was defined as the average of the recall and specificity to account for class imbalance. Threshold-independent evaluation metrics were provided by ROC-AUC (Equation 25) and Precision-Recall Average Precision (Equation 26). Probabilistic calibration was evaluated using log-loss (Equation 27), and the confusion matrix formulation (Equation 28) was provided to ensure operational clarity. Collectively, these metrics demonstrated GenCKD-Net's ability to discriminate and perform reliably under class imbalance. Figures 7–9 report the threshold-independent ROC and precision–recall curves and the threshold-dependent confusion matrix, respectively;

Figure 10 reports per-epoch wall-clock fitting time. These visualisations should be interpreted alongside the numerical results in Table 4, while Table 5 explicitly records that external validation was not performed.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad [19]$$

$$\text{Prec} = \frac{TP}{TP + FP} \quad [20]$$

$$\text{Rec} = \frac{TP}{TP + FN} \quad [21]$$

$$\text{Spec} = \frac{TN}{TN + FP} \quad [22]$$

$$F1 = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}} \quad [23]$$

$$BA = \frac{\text{Rec} + \text{Spec}}{2} \quad [24]$$

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx \quad [25]$$

$$AP = \sum_n (R_n - R_{n-1})P_n \quad [26]$$

$$LL = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad [27]$$

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad [28]$$

Figure 4 distinguishes the experimentally validated GenCKD-Net core from proposed future extensions. The validated component comprises leakage-controlled preprocessing, training-only SMOTE, a compact 128→64 MLP, validation-based operating-threshold selection, and held-out test evaluation. Formal calibration, explainability, multimodal fusion, and prospective deployment are not reported results.

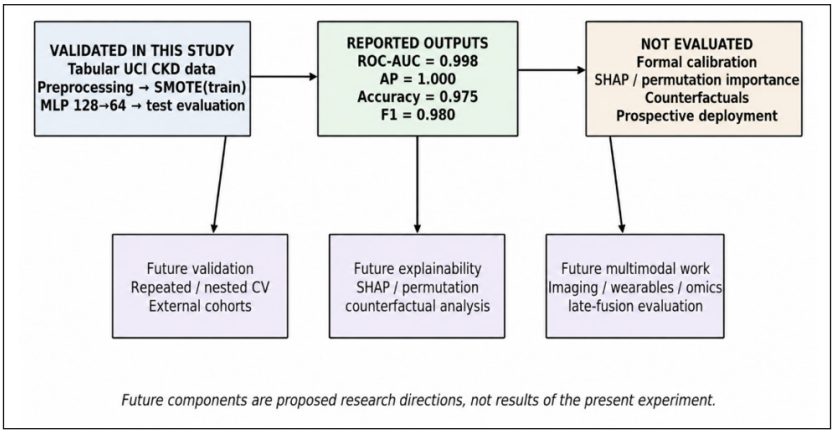


Figure 4. GenCKD-Net: Validated core and future extensions

RESULT AND DISCUSSIONS

Predictive Performance

For the test cohort, GenCKD-Net produced ROC-AUC = 0.998 and PR-AP = 1.000. At the validation-selected operating threshold, accuracy was 0.975, precision was 1.000, recall was 0.960, and the F1-score was 0.980. These results indicate high benchmark performance, but they should not be interpreted as evidence of clinical generaliability.

Learning Dynamics

Training and validation accuracy increased during the recorded warm-start fitting passes, while the recorded loss values decreased over the observed sequence. These observations describe optimisation behaviour on the split and do not establish generalisation or deployment readiness.

Complexity & Runtime

Each recorded fitting pass took approximately 0.2–1.6 s, with a mean of approximately 0.6 s in the implementation. These measurements indicate modest computational requirements for the evaluated benchmark dataset. They should not be interpreted as evidence of production deployment performance or routine clinical retraining.

Error Analysis

GenCKD-Net produced two false negatives and no false positives on the 80-record test cohort at the selected operating threshold. These counts are specific to this held-out split. Because formal patient-level explainability was not performed, case-specific physiological explanations are omitted.

Figure 5 shows that training and validation accuracy increased during the recorded fitting passes. The curves are presented as descriptive evidence of training behaviour; no automatic early-stopping criterion was implemented.

Figure 6 shows that training and validation loss decreased during the recorded fitting passes. Decreasing loss indicates progress in optimisation but does not establish formal probability calibration or generalisation.

Figure 7 is the ROC curve for the held-out test cohort and should be read as a summary of discrimination, not as evidence of external clinical validity.

The ROC curve yields an ROC-AUC of 0.998 in the test evaluation. This threshold-independent result should be interpreted as split-specific benchmark performance.

Figure 8 presents the precision–recall curve for the held-out test cohort; its interpretation is conditional on the observed test-set class distribution.

The precision-recall curve corresponds to AP = 1.000 for the test evaluation. The result should be interpreted in the 80-record held-out test cohort.

Figure 9 presents the held-out test confusion matrix at the selected operating threshold. The confusion matrix reports TP = 48, TN = 30, FP = 0, and FN = 2, yielding a precision of 1.00 and a recall of 0.96 at the selected cutoff. The absence of false positives was observed only on this held-out test split.

Figure 10 reports the recorded wall-clock fitting time per pass; it does not establish deployment latency or clinical readiness.

Figure 10 shows the observed wall-clock time for each recorded fitting pass. These measurements describe computational behaviour for the benchmark experiment and should not be interpreted as evidence of clinical deployment readiness.

Table 2 shows the stratified 64%/16%/20% train/validation/test split. The original class proportions were retained in all three subsets, while SMOTE was applied only to the transformed training data. Validation data were used for operating-threshold selection and remained otherwise untouched; the test subset was reserved for final evaluation. The resulting test confusion matrix contained TP = 48, TN = 30, FP = 0, and FN = 2.

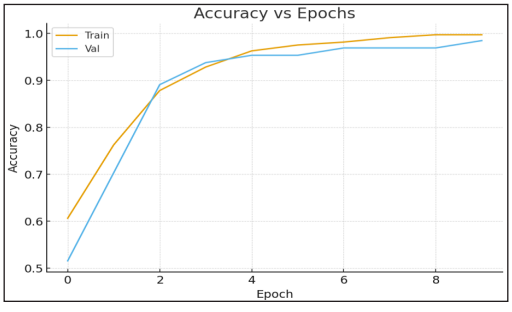


Figure 5. Accuracy vs. Epochs (Train/Validation)

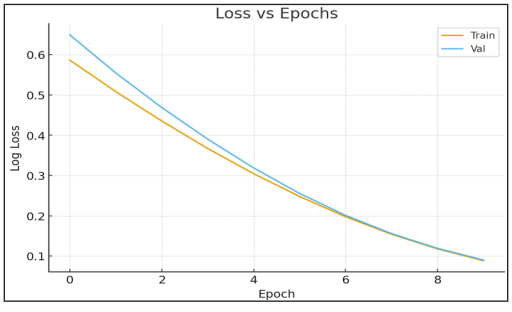


Figure 6. Loss vs. Epochs (Train/Validation)

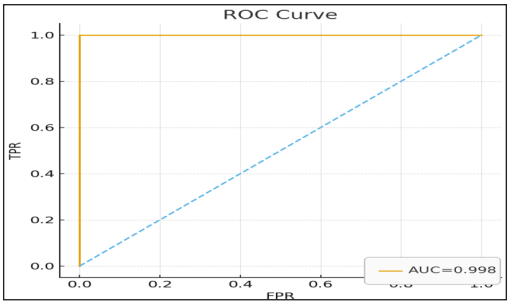


Figure 7. ROC curve

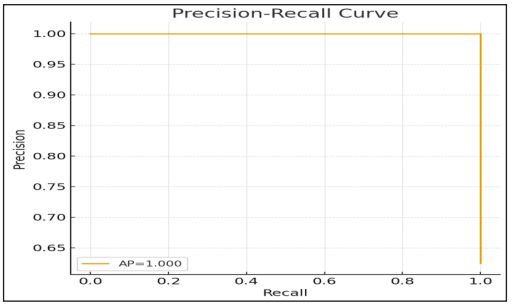


Figure 8. Precision–recall curve

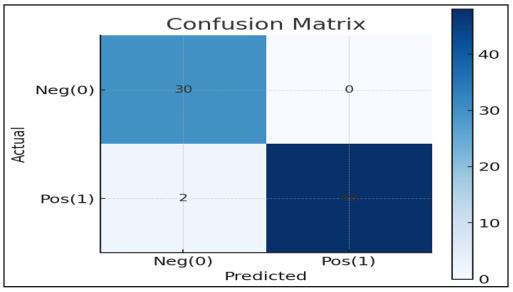


Figure 9. Confusion matrix

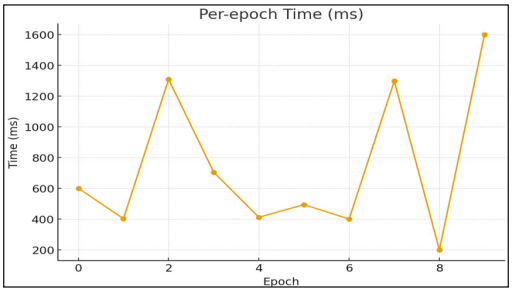


Figure 10. Per-Epoch Wall-Clock Time (ms)

Table 2
Train/ validation/ test split statistics

Split	N (rows)	CKD (n)	CKD (%)	Non-CKD (n)	Non-CKD (%)	SMOTE Used	Role/Purpose
Train	256	160	62.50%	96	37.50%	Yes.	Model fitting
Validation	64	40	62.50%	24	37.50%	No	Threshold selection
Test	80	50	62.50%	30	37.50%	No	Final held-out evaluation
Total	400	250	62.50%	150	37.50%	No	Original dataset

Table 3 summarises the held-out test performance of GenCKD-Net. The model achieved ROC-AUC = 0.998 and PR-AP = 1.000, with accuracy 0.975, precision 1.000, recall 0.960, F1-score 0.980, and balanced accuracy 0.980. These results are benchmark-specific and do not establish clinical readiness.

RESULT AND DISCUSSIONS

GenCKD-Net produced no false positives and two false negatives on the test cohort, corresponding to a precision of 1.00 and a recall of 0.96 at the selected operating point. These values describe only the 80-record evaluation and should not be interpreted as evidence that false positives will be absent in other populations.

Table 3
Overall test performance metrics

Metric	GenCKD-Net	Basis/note
Test set size (N)	80	TP + TN + FP + FN = 48 + 30 + 0 + 2
True Positives (TP)	48	Held-out test confusion matrix
True Negatives (TN)	30	Held-out test confusion matrix
False Positives (FP)	0	Held-out test confusion matrix
False Negatives (FN)	2	Held-out test confusion matrix
ROC-AUC	0.998	Threshold-free ROC metric
PR-AP	1.000	Threshold-free precision-recall metric
Accuracy	0.975	78/80 correct
Precision (PPV)	1.000	48/(48 + 0)
Recall (Sensitivity)	0.960	48/(48 + 2)
Specificity (TNR)	1.000	30/(30 + 0)
F1 Score	0.980	Harmonic mean of precision and recall
Balanced Accuracy	0.980	(0.960 + 1.000)/2

At the operating point, GenCKD-Net produced TN = 30, FP = 0, FN = 2, and TP = 48, corresponding to precision = 1.00 and recall = 0.96 on this 80-record test split. These values indicate strong thresholded benchmark performance but do not establish that false-positive or false-negative rates will remain the same in independent populations. Thresholds should be selected according to the intended use case and evaluated with independent validation, calibration assessment, and, where appropriate, decision-curve analysis before clinical application.

Learning Curve Patterns Suggest Generalisation for Tabular CKD

The training behaviour was consistently predictable in the recorded fitting passes. Validation accuracy reached approximately 0.98 by epochs 7–9, and runtime remained low. These observations indicate stable benchmark optimisation but do not by themselves demonstrate generalisation.

From Strong Offline Results to Reliable Bedside Tool: Explainability, Auditability, Workflow Fit

The current study should not be described as a trusted clinical tool. This experiment does not include formal explainability analysis, probability calibration assessment, external validation, or prospective clinical evaluation. These are important requirements for future translation.

Comparison of Existing vs. Proposed System

Published CKD prediction studies include conventional machine-learning models, deep-learning approaches, and multimodal systems, but their reported performance is difficult to compare directly because datasets, outcome definitions, preprocessing procedures, and validation designs differ (Gogoi & Valan, 2025; Iliyas et al., 2025; Sanmarchi et al., 2023). GenCKD-Net should therefore be interpreted as a compact benchmark model rather than as a demonstrated clinical or architectural improvement over existing systems. Its main methodological contribution is the leakage-controlled preprocessing and training-only resampling workflow combined with a small 128→64 MLP. The near-ceiling test metrics reported here remain specific to the UCI benchmark and require repeated and external validation before claims of transportability can be made.

Table 4 compares GenCKD-Net with the evaluated baseline systems on the same held-out test cohort; the values are benchmark estimates and should not be interpreted as independent validation.

The comparison should be interpreted cautiously because published CKD studies use different datasets, validation schemes, and outcome definitions. This study should not claim to outperform existing CKD systems without a matched experimental comparison.

Performance Evaluation

GenCKD-Net was evaluated on a held-out test set after training-only SMOTE and validation-based threshold selection. It achieved ROC-AUC = 0.998, PR-AP = 1.000, accuracy = 0.975, precision = 1.000, recall = 0.960, specificity = 1.000, and F1 ≈ 0.98. These results describe the selected split and should be interpreted as benchmark performance rather than as clinical readiness, as shown in Equations 29-37.

Table 4
GenCKD-Net vs. existing CKD systems: Performance & runtime metrics (test cohort)

Parameters	Published Comparator Range/Note	GenCKD-Net
ROC-AUC	0.998–1.000 reported on UCI benchmark	0.998
PR-AP	Near-perfect in selected reports	1.000
Accuracy	Varies by study and validation design	0.975
Precision	Not directly comparable across studies	1.000
Recall	Not directly comparable across studies	0.960
Specificity	Not consistently reported	1.000
F1	Varies by study	0.980
False Positives	Not consistently reported	0
False Negatives	Not consistently reported	2
Median seconds/epoch	Often not reported	~0.6

- **ROC-AUC:** How well the model ranks positives above negatives without picking a threshold.

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx \quad [29]$$

ROC-AUC = 0.998 indicates near-perfect ranking on this held-out benchmark test set.

- **PR-AP:** Threshold-free performance in precision-recall space, ideal for imbalanced data.

$$AP = \sum_n (R_n - R_{n-1})P_n \quad [30]$$

PR-AP = 1.000 indicates excellent precision-recall performance on this held-out test cohort.

- **Accuracy:** Fraction of all predictions that are correct.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad [31]$$

0.975 = 78/80, correct at the chosen cutoff.

- **Precision (PPV):** Of the flagged positives, how many are truly positive?

$$\text{Prec} = \frac{TP}{TP + FP} \quad [32]$$

0.96 = 48/50 CKD cases captured.

- **Specificity (TNR):** Of all true negatives, how many did the model correctly clear?

$$\text{Spec} = \frac{TN}{TN + FP} \quad [33]$$

1.00 indicates that 30 out of 30 controls were correctly classified as non-CKD.

- **F1-Score:** The F1 score is the harmonic mean of precision and recall, which balances both metrics.

$$F1 = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}} \quad [34]$$

F1 ≈ 0.98 reflects the harmonic mean of precision 1.00 and recall 0.96.

- **Balanced Accuracy:** This metric is the average of sensitivity and specificity, adjusted for class imbalance.

$$BA = \frac{\text{Rec} + \text{Spec}}{2} \quad [35]$$

$\approx 0.98 = (0.96 + 1.00)/2$.

- **Log-Loss (Cross-Entropy):** This metric penalises overconfident incorrect predictions and assesses the model's calibration.

$$LL = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad [36]$$

The recorded train/validation log-loss behaviour was smooth, but formal probability calibration was not assessed; therefore, smooth loss should not be interpreted as evidence of calibration.

- **Confusion Matrix:** Counts of outcomes by class to support thresholded metrics.

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad [37]$$

The confusion matrix confirms zero false positives and two false negatives on the held-out test cohort.

LIMITATIONS AND THREATS TO VALIDITY

Dataset Provenance and External Generalisability

External Validation Status

External validation was not performed in this study. All reported experimental results are based on the UCI CKD dataset and a single stratified train/validation/test split. Accordingly, no independent-cohort performance estimate is reported for GenCKD-Net. A future external-validation study should freeze the fitted preprocessing pipeline, model parameters, and operating threshold before evaluating an independently collected cohort.

A frozen-model external-validation protocol should retain the fitted preprocessing pipeline and model without retraining or threshold optimisation and then evaluate predefined discrimination, threshold-dependent, and calibration metrics on an independent cohort.

Table 5 documents that external validation was not performed in this study; therefore, no independent-cohort performance claim is made.

Table 5
External validation—Not performed in this study

Metric	Accuracy	Precision	Recall	F1-Score
GenCKD-Net (UCI held-out test)	0.975	1.000	0.960	0.980

No external validation results are available for this study. Any independent-cohort performance should be reported only after the frozen-model protocol has been executed and the results independently verified.

Future external validation should use independently collected cohorts with documented feature definitions, missingness patterns, prevalence, and outcome-label procedures so that any observed performance differences can be interpreted appropriately.

Because the UCI CKD benchmark is small and may differ from future target populations in patient characteristics, prevalence, laboratory measurement, coding, and missingness patterns, performance may change under a dataset shift.

Why Nested Cross Validation and Calibration Matter When Your Sample Size is Small

The headline metrics were obtained from a relatively small dataset with only 80 held-out test records. Performance estimates can therefore have substantial sampling variability, and near-ceiling benchmark values may overstate expected performance in broader populations. Repeated or nested cross-validation, sensitivity analyses across random seeds and split assignments, subgroup analyses where feasible, and explicit calibration assessments are appropriate safeguards for future work.

Tabular-only focus today; multimodal needs prospective proof

This study deliberately focuses on tabular data. Extending to imaging, wearables/POCT, or omics is valuable, but these modalities require separate datasets, preprocessing, fusion strategies, and validation.

- A safer expansion plan involves prospective and staged validation.
- Start with late fusion: retain the tabular MLP and train separate encoders for any new modality before combining representations in a prediction head.
- Validate each modality separately and in combination using discrimination, calibration, decision-curve analysis, clinically relevant thresholds, and modality-ablation studies.
- Track data provenance and modality-specific drift, and complete privacy and consent requirements for any data linkages.
- Only after late-fusion performance is stable across internal and external validation should more complex cross-modal architectures be considered.

Future Work

Future work should focus on repeated or nested cross-validation, categorical-aware resampling sensitivity analysis, formal probability calibration, subgroup analysis, independent external validation, and prospective evaluation. Explainability and multimodal extensions should be treated as components for future research rather than as experimentally demonstrated capabilities.

CONCLUSION

This study evaluated GenCKD-Net, a lightweight multilayer perceptron pipeline for binary CKD classification using the UCI Chronic Kidney Disease benchmark. The model achieved high performance on the held-out test set and was comparable to logistic regression, support vector machines, random forests, and LightGBM. The similarity between the models indicates that the study does not provide evidence of architectural superiority for the proposed MLP; instead, it demonstrates that a compact neural-network pipeline can achieve competitive benchmark performance under the leakage-controlled preprocessing and training-only resampling workflow.

ACKNOWLEDGEMENT

The authors thank Mohan Babu University and GITAM (Deemed to be a University) for their academic support.

REFERENCES

Akter, S., Ahmed, M., Al Imran, A., Habib, A., Ul Haque, R., Rahman, M. S., Hasan, M. R., & Mahjabeen, S. (2023). CKD.Net: A novel deep learning hybrid model for effective, real-time, automated screening tool towards prediction of multi stages of CKD along with eGFR and creatinine. *Expert Systems with Applications, 223*, Article 119851. https://doi.org/10.1016/j.eswa.2023.119851

Ali, S. I., Jung, S. W., Bilal, H. S. M., Lee, S.-H., Hussain, J., Afzal, M., Hussain, M., Ali, T., Chung, T., & Lee, S. (2022). Clinical decision support system based on hybrid knowledge modeling: A case study of chronic kidney disease-mineral and bone disorder treatment. *International Journal of Environmental Research and Public Health, 19*(1), Article 226. https://doi.org/10.3390/ijerph19010226

Burra, L. R., Tumuluru, P., Bonam, J., Raju, S. H., Nalajala, S., & Borra, S. P. R. (2025). Innovative application of conditional deep convolutional generative adversarial networks to enhance chronic kidney disease diagnosis with uneven datasets. In R. Nidhya, M. Kumar, S. Karthik, R. Anand, & S. Balamurugan (Eds.), *Smart factories for Industry 5.0 transformation* (pp. 71-87). Wiley. https://doi.org/10.1002/9781394200467.ch4

Fang, P., Wu, Y., He, Y., Li, H., Guan, Z., Wang, X., Chen, T., & Shen, J. (2025). Research progress on AI-assisted screening and prediction of systemic diseases based on retinal images. *The Visual Computer, 41*(12), 9509-9537. https://doi.org/10.1007/s00371-025-03977-8

Gogoi, P., & Valan, J. A. (2025). Machine learning approaches for predicting and diagnosing chronic kidney disease: Current trends, challenges, solutions, and future directions. *International Urology and Nephrology, 57*(4), 1245-1268. https://doi.org/10.1007/s11255-024-04281-5

Han, S., Yamamoto, S., Jung, C.-Y., Jin, D. Y., Lee, T., & Kim, J.-S. (2024). Wearable sensors for monitoring chronic kidney disease. *Communications Materials, 5*(1), Article 153. https://doi.org/10.1038/s43246-024-00606-0

Haris, M., Raveendra, K., Travlos, C. K., Lewington, A., Wu, J., Shuweidhi, F., Nadarajah, R., & Gale, C. P. (2024). Prediction of incident chronic kidney disease in community-based electronic health records: A systematic review and meta-analysis. *Clinical Kidney Journal, 17*(5), Article sfae098. https://doi.org/10.1093/ckj/sfae098

Iliyas, I. I., Boukari, S., & Gital, A. Y.-U. (2025). Recent trends in prediction of chronic kidney disease using different learning approaches: A systematic literature review. *Journal of Medical Artificial Intelligence, 8*, Article 62. https://doi.org/10.21037/jmai-24-256

Islam, M. S. B., Sumon, M. S. I., Sarmun, R., Bhuiyan, E. H., & Chowdhury, M. E.-H. (2024). Classification and segmentation of kidney MRI images for chronic kidney disease detection. *Computers and Electrical Engineering, 119*, Article 109613. https://doi.org/10.1016/j.compeleceng.2024.109613

Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., … Starmans, M. P. A. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ, 388*, Article e081554. https://doi.org/10.1136/bmj-2024-081554

Miller, Z. A., & Dwyer, K. (2025). Artificial intelligence to predict chronic kidney disease progression to kidney failure: A narrative review. *Nephrology, 30*(1), Article e14424. https://doi.org/10.1111/nep.14424

Nadarajah, R., Mohammad, H., Younsi, T., Romer, E., Raveendra, K., Travlos, C., Wahab, A., Nakao, Y. M. N., Nakao, K., Wu, J., & Gale, C. P. (2023). A systematic review and meta-analysis of prediction models for incident heart failure and chronic kidney disease in community-based electronic health records [Conference abstract]. *European Heart Journal, 44*(Supplement_2), Article ehad655.2951. https://doi.org/10.1093/eurheartj/ehad655.2951

Pan, Q., & Tong, M. (2024). Artificial intelligence in predicting chronic kidney disease prognosis. a systematic review and meta-analysis. *Renal Failure, 46*(2), Article 2435483. https://doi.org/10.1080/0886022x.2024.2435483

Puri, V., Priyadarshini, I., Kataria, A., Rani, S., & Min, H. (2025). Privacy-first ML for chronic kidney disease prediction: Exploring a decentralized approach using blockchain and IPFS. *IEEE Access, 13*, 43178-43189. https://doi.org/10.1109/access.2025.3548645

Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. *JMIR AI, 3*, Article e53207. https://doi.org/10.2196/53207

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). *Chronic kidney disease* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5G020

Sabanayagam, C., Banu, R., Lim, C., Tham, Y. C., Cheng, C.-Y., Tan, G., Ekinci, E., Sheng, B., McKay, G., Shaw, J. E., Matsushita, K., Tangri, N., Choo, J., & Wong, T. Y. (2025). Artificial intelligence in chronic kidney disease management: A scoping review. *Theranostics, 15*(10), 4566-4578. https://doi.org/10.7150/thno.108552

Sághy, E., Elsharkawy, M., Moriarty, F., Kovács, S., Wittmann, I., & Zemplényi, A. (2025). A novel machine learning methodology for the systematic extraction of chronic kidney disease comorbidities from abstracts. *Frontiers in Digital Health, 7*, Article 1495879. https://doi.org/10.3389/fdgth.2025.1495879

Sanmarchi, F., Fanconi, C., Golinelli, D., Gori, D., Hernandez-Boussard, T., & Capodici, A. (2023). Predict, diagnose, and treat chronic kidney disease with machine learning: A systematic literature review. *Journal of Nephrology, 36*(4), 1101-1117. https://doi.org/10.1007/s40620-023-01573-4

Singh, P., Goyal, L., Mallick, D. C., Surani, S. R., Kaushik, N., Chandramohan, D., & Simhadri, P. K. (2025). Artificial intelligence in nephrology: Clinical applications and challenges. *Kidney Medicine, 7*(1), Article 100927. https://doi.org/10.1016/j.xkme.2024.100927

Su, C.-T., Chang, Y.-P., Ku, Y.-T., & Lin, C.-M. (2022). Machine learning models for the prediction of renal failure in chronic kidney disease: A retrospective cohort study. *Diagnostics, 12*(10), Article 2454. https://doi.org/10.3390/diagnostics12102454

Wu, C.-C., Islam, M. M., Poly, T. N., & Weng, Y.-C. (2024). Artificial intelligence in kidney disease: A comprehensive study and directions for future research. *Diagnostics, 14*(4), Article 397. https://doi.org/10.3390/diagnostics14040397

Zhang, M., Ye, Z., Yuan, E., Lv, X., Zhang, Y., Tan, Y., Xia, C., Tang, J., Huang, J., & Li, Z. (2024). Imaging-based deep learning in kidney diseases: Recent progress and future prospects. *Insights into Imaging, 15*(1), Article 50. https://doi.org/10.1186/s13244-024-01636-5

Zhang, S. (2025). Advances in multi-omics integration and AI-driven precision medicine for diabetes. In *Proceedings of the 6th International Symposium on Artificial Intelligence for Medical Sciences (ISAIMS)* (pp. 96-102). Association for Computing Machinery. https://doi.org/10.1145/3777577.3777592