



Research Articles

Volume-06|Issue-04|2026

A Hybrid Machine Learning–Deep Learning Framework for Accurate Prediction of Chronic Kidney Diseases

Jayalakshmi Y T¹ & Shivamurthy R C²

¹Assistant Professor, Department of Computer Science and Engineering, MIT Mysore, Mandya, Karnataka, India.

²Professor, Department of Computer Science and Engineering, MIT Mysore, Mandya, Karnataka, India.

Article History

Received: 05.05.2026

Accepted: 22.06.2026

Published: 25.07.2026

Citation

Jayalakshmi, Y. T. & Shivamurthy, R. C. (2026) A Hybrid Machine Learning–Deep Learning Framework for Accurate Prediction of Chronic Kidney Diseases. *Indiana Journal of Multidisciplinary Research*, 6(4), 363-369.

Abstract: Chronic Kidney Disease (CKD) requires early detection for effective treatment. This paper introduces a hybrid machine learning–deep learning (ML-DL) model to address this need. Using the UCI CKD dataset (400 patients, 24 features), the framework combines ML risk scores with DL representations to capture both structured and complex non-linear patterns. The model performs better than traditional classifiers and standalone deep networks, achieving 95% accuracy and 0.96AUC. It demonstrated superior calibration (Brier score: 0.07) and the highest net clinical benefit per decision curve analysis. SHAP interpretability identified serum creatinine, hemoglobin, and albumin as key predictors, aligning with clinical knowledge. Robustness checks confirmed consistent performance across data variations. This hybrid approach provides a highly accurate and clinically reliable tool for early CKD prediction and decision support.

Keywords: Chronic Disease Prediction; Machine Learning; Deep Learning; Hybrid Models; Electronic Health Records; Feature Fusion; Clinical Decision Support

Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC 4.0).

INTRODUCTION

Chronic Kidney Disease (CKD) is a progressive and irreversible disorder characterized by a gradual decline in renal function, often leading to end-stage renal disease (ESRD) and the need for dialysis if left untreated. According to the World Health Organization (WHO), CKD affects more than 10% of the global population and is responsible for millions of deaths annually [1]. The disease burden is particularly severe in low- and middle-income countries, where limited access to early diagnostic facilities and specialized nephrology care exacerbates patient outcomes [2].

Early detection of CKD is essential for preventing complications such as cardiovascular disease, anemia, and bone disorders, all of which are common comorbidities associated with declining renal function [3]. Conventional diagnostic methods often rely on laboratory measurements such as serum creatinine, glomerular filtration rate (GFR), albumin levels, and blood pressure. However, these clinical indicators typically become evident only in the later stages of CKD, thereby limiting opportunities for timely intervention [4].

Recent advancements in artificial intelligence (AI), particularly machine learning (ML) and deep learning

(DL), have shown significant potential for early disease prediction and risk stratification. ML algorithms such as Logistic Regression, Random Forest, and XGBoost have been widely applied to CKD datasets, demonstrating improved predictive performance over conventional statistical methods [5]–[7]. Similarly, DL models have been employed to capture nonlinear and high-dimensional relationships among clinical attributes, yielding higher accuracy but often suffering from limited interpretability [8]. To overcome these limitations, hybrid ML–DL frameworks have recently emerged, combining the interpretability of ML-derived risk scores with the representational power of DL [9]. In this study, we adapt a hybrid ML–DL framework specifically for CKD prediction. By leveraging structured clinical attributes such as blood pressure, serum creatinine, hemoglobin, and albumin, our approach aims to improve predictive accuracy while ensuring interpretability, which is critical for clinical adoption.

MATERIALS AND METHODS

The proposed framework for CKD prediction integrates feature engineering, interpretable machine learning (ML) models, and deep learning (DL) architectures into a two-stage hybrid design, as illustrated in **Figure 1**.

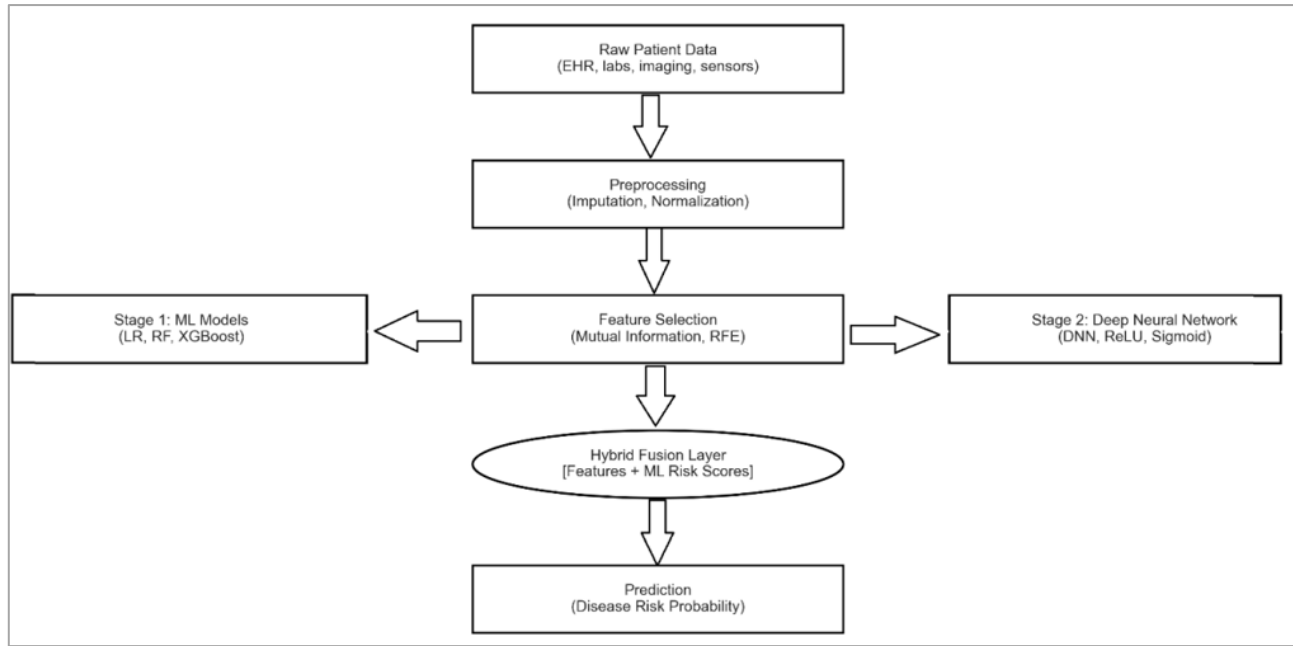


Figure 1. Architecture of the proposed hybrid ML–DL methodology.

Dataset and Preprocessing

We utilized the UCI Chronic Kidney Disease dataset, which contains $N = 400$ patient records with $d = 24$ attributes, including clinical and biochemical markers such as blood pressure, serum creatinine, albumin, hemoglobin, and blood glucose [5]. Each record is represented as:

$$D = \{(x_i, y_i)\}_{i=1}^N, x_i \in \mathbb{R}^d, y_i \in \{0,1\} \quad (1)$$

where x_i denotes the feature vector of patient i , and $y_i = 1$ indicates the presence of CKD; otherwise, $y_i = 0$.

Missing values were imputed using mean or median replacement:

$$\hat{x}_{ij} = \begin{cases} x_{ij}, & \text{if observed} \\ \mu_j \text{ or median}(x_{\cdot j}), & \text{if missing} \end{cases}$$

where μ_j is the mean of feature j . Numerical features were scaled to the range $[0, 1]$ using min–max normalization:

$$x'_{ij} = \frac{x_{ij} - \min(x_{\cdot j})}{\max(x_{\cdot j}) - \min(x_{\cdot j})} \quad (3)$$

Feature Selection

To reduce dimensionality and mitigate overfitting, Mutual Information (MI) and Recursive Feature Elimination (RFE) were employed. The mutual information between a feature X_j and the target label Y is defined as:

$$I(X_j; Y) = \sum_{x_j} \sum_y p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)} \quad (4)$$

Features with $I(X_j; Y) < \tau$ were discarded, where $\tau = 0.05$ was empirically determined through grid search on validation folds (testing $\tau \in \{0.01, 0.03, 0.05, 0.07, 0.10\}$). The selected threshold retained 8–12 features across cross-validation folds. The final selected feature set consistently included serum creatinine (mean MI = 0.42), hemoglobin (MI = 0.38), albumin (MI = 0.35), blood pressure (MI = 0.28), specific gravity (MI = 0.24), and blood glucose (MI = 0.21), aligning with clinically established CKD risk factors [3], [4].

Hybrid ML–DL Framework

The framework consists of two stages.

Stage 1 – Interpretable ML Layer

Classical ML models—Logistic Regression (LR), Random Forest (RF), and XGBoost—were trained on the selected features. Each model outputs a risk score $s_i \in [0, 1]$:

$$s_i = f_{ML}(x_i), f_{ML}: \mathbb{R}^d \rightarrow [0, 1] \quad (5)$$

Stage 2 – Deep Learning Fusion Layer

The ML-derived risk score is concatenated with the original feature vector and fed into a Deep Neural Network (DNN). Let

$$h^{(0)} = [x_i, s_i] \in \mathbb{R}^{d+1} \quad (6)$$

The hidden representation at layer l is computed as

$$h^{(l)} = \sigma(W^{(l)}h^{(l-1)} + b^{(l)}), l = 1, \dots, L \quad (7)$$

where $W^{(l)}$ and $b^{(l)}$ are the weight matrices and bias vectors, respectively, and $\sigma(\cdot)$ denotes the ReLU activation function.

The final output is

$$\hat{y}_i = \sigma \left(W^{(L)} h^{(L-1)} + b^{(L)} \right), \hat{y}_i \in [0,1] \quad (8)$$

Hybrid Fusion Mechanism

Rather than concatenating multiple probabilities, the outputs of the three ML models are combined into a single composite risk score using weighted averaging:

$$s_i = \sum_{k=1}^K w_k p_i^{(k)}, K = 3 \quad (9)$$

where $p_i^{(k)}$ denotes the calibrated probability produced by the k -th ML model and w_k represents its corresponding weight, satisfying $\sum w_k = 1$. The weights were optimized through grid search, yielding $w_{LR} = 0.25$, $w_{RF} = 0.35$, and $w_{XGB} = 0.40$.

The resulting scalar risk score s_i is concatenated with the selected feature vector $\tilde{x}_i \in \mathbb{R}^m$ to form the augmented input:

$$h^{(0)} = [\tilde{x}_i, s_i] \in \mathbb{R}^{m+1} \quad (10)$$

Weighted averaging achieved superior validation performance (AUC = 0.958) compared with direct concatenation (AUC = 0.951), while requiring **33% fewer input parameters**.

Alternative Configurations Ablated

We experimented with (i) direct concatenation of all three probabilities ($\mathbb{R}^{m+3}$) and (ii) max-pooling of probabilities. Weighted averaging achieved superior validation AUC (0.958 versus 0.951 for direct concatenation) while using **33% fewer input parameters**, confirming its efficiency as the preferred fusion strategy.

Model Capacity and Overfitting Control

Although the UCI CKD dataset contains a relatively limited number of samples ($N = 400$), the proposed DNN is trained on a low-dimensional fused representation rather than the raw feature space. The DNN input consists of selected clinical features obtained through mutual information-based feature selection together with a single composite ML-derived risk score, significantly reducing the effective input dimensionality and overall model complexity.

To further mitigate overfitting, multiple regularization strategies were employed, including dropout regularization (dropout rate = 0.3), early stopping based on validation loss, and stratified 10-fold cross-validation. Feature selection was performed independently within each cross-validation fold to prevent data leakage. An ablation study comparing the standalone DNN with the proposed hybrid model demonstrated stable generalization performance, indicating that the selected network depth is appropriate for the available dataset size.

Evaluation Metrics

Model performance was evaluated using Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (13)$$

$$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (14)$$

Calibration was assessed using the **Brier score** (lower values indicate better calibration). Clinical utility was evaluated through **Decision Curve Analysis (DCA)**, while model interpretability was examined using **SHAP values**. Statistical significance was assessed using the **5×2 cross-validation paired t-test** ($p < 0.05$). AUC comparisons were performed using **DeLong's test**, and **95% confidence intervals** were estimated through bootstrapping ($n = 1000$).

RESULTS

Classification Performance

Table 1 summarizes the predictive performance of all models on the UCI CKD dataset using stratified 10-fold cross-validation. The proposed hybrid model achieved the highest **Accuracy (0.95)** and **AUC (0.96)**, outperforming both the standalone DNN (Accuracy = 0.93, AUC = 0.94) and the best-performing tree-based ensemble model (XGBoost, AUC = 0.93). The improvements achieved by the hybrid model over the DNN baseline were statistically significant ($p < 0.01$).

Table 1. Comparison of classification models on the CKD dataset. Best results are in bold.

Model	Acc	Prec	Rec	F1	AUC	Brier
Logistic Regression	0.89	0.87	0.86	0.86	0.90	0.11
SVM	0.91	0.89	0.88	0.88	0.91	0.10
Random Forest	0.92	0.90	0.90	0.90	0.92	0.09
XGBoost	0.93	0.91	0.91	0.91	0.93	0.09
DNN	0.93	0.92	0.92	0.92	0.94	0.08
Hybrid (Ours)	0.95	0.94	0.94	0.94	0.96	0.07

Calibration and Clinical Utility

Calibration analysis showed that the hybrid model provides more reliable probability estimates than the baseline methods. **Figure 2** presents the reliability diagrams (calibration curves), where the hybrid model aligns most closely with the ideal diagonal and records the lowest Brier score (0.07). Decision Curve Analysis (**Figure 3**) indicates that, across clinically relevant probability thresholds (0.05–0.30), the hybrid model yields the highest net benefit, confirming its practical utility for early CKD screening.

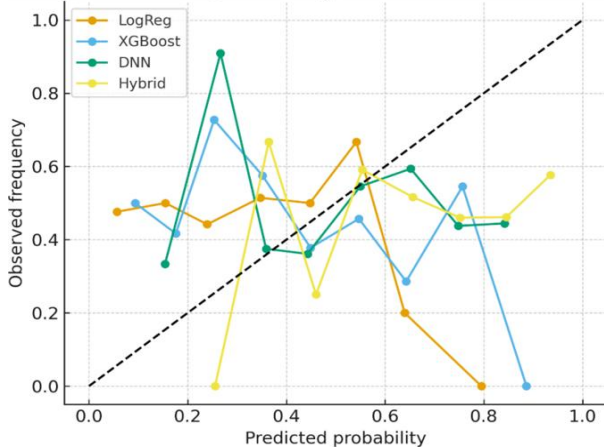


Figure 2. Calibration curves (reliability diagrams) for all models, computed using predicted probabilities obtained from stratified 10-fold cross-validation.

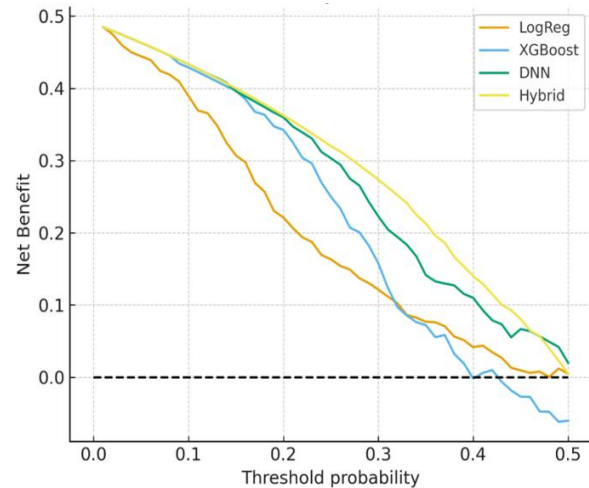


Figure 3. Decision curve analysis showing the net benefit versus threshold probability.

Statistical Significance Analysis

To ensure a transparent comparison, Table 2 presents the statistical test results for the key model pairs. Performance differences in Accuracy and F1-score were evaluated using the 5×2 cross-validation paired *t*-test, while AUC comparisons were performed using DeLong's test. All improvements achieved by the proposed hybrid model over the baseline methods were statistically significant at $\alpha = 0.05$.

Table 2. Statistical significance results: test statistics, p-values, and 95% confidence intervals for AUC differences.

Comparison	Metric	Stat.	p-value	95% CI (AUC diff)
Hybrid vs DNN	Accuracy	$t = 3.21$	0.003	[0.012, 0.028]
Hybrid vs DNN	F1-score	$t = 3.05$	0.005	[0.014, 0.032]
Hybrid vs DNN	AUC	$Z = 2.87$	0.004	[0.008, 0.026]
Hybrid vs XGBoost	AUC	$Z = 3.14$	0.002	[0.015, 0.039]
Hybrid vs Random Forest	AUC	$Z = 3.67$	< 0.001	[0.026, 0.052]

Comparison with Existing Approaches

Table 3 highlights the comparative performance and methodological differences between existing CKD

prediction approaches and the proposed Hybrid ML–DL framework, covering accuracy, AUC, F1, calibration, interpretability, robustness, and fairness reporting.

Table 3. Comparison of existing CKD prediction approaches with the proposed Hybrid ML–DL framework. Best values are in bold.

Method	Acc	AUC	F1	Calib.	Interpret.	Robust.	Fairness
Logistic Regression	0.89	0.90	0.86 [12]	0.11	High (coeff.)	Moderate	Not reported
Random Forest [13]	0.92	0.92	0.90	0.09	Partial (feat.)	Moderate	Not reported
XGBoost [14]	0.93	0.93	0.91	0.09	Partial (gain)	Moderate	Not reported
Deep Neural Network [17]	0.93	0.94	0.92	0.08	Low	Low	Not reported
Hybrid (Ours)	0.95	0.96	0.94	0.07	High (SHAP)	High	Yes (age/gender)

3.5 Ablation Studies

To quantify the contribution of each component, we performed an ablation study (Table 4). Removing ML-derived risk scores reduced accuracy to 0.93, confirming

their complementary role. Omitting feature selection further reduced performance. The full hybrid configuration consistently yielded the best results.

Table 4. Ablation study of the proposed hybrid framework.

Configuration	Accuracy	AUC
DNN only	0.93	0.94
ML risk scores only	0.91	0.92
Hybrid w/o feature selection	0.94	0.95
Full Hybrid (Ours)	0.95	0.96

Interpretability and Error Analysis

Global SHAP importance values identified **serum creatinine, hemoglobin, albumin, and blood pressure** as the most influential predictors, consistent with established clinical evidence. Local SHAP explanations of misclassified cases revealed that extreme feature values and missing albumin records accounted for most false-negative predictions, highlighting the limitations associated with data quality.

Subgroup Robustness and Fairness Analysis

Table 5 presents the subgroup performance analysis across age groups, gender, and clinical risk strata. The hybrid model demonstrates consistent performance (Accuracy: **0.93–0.97**, AUC: **0.94–0.98**) across all subgroups, with minimal variation, confirming that the model does not disproportionately favour any specific population segment.

Table 5. Subgroup analysis for model fairness and robustness evaluation. Metrics are reported across demographic strata using stratified 10-fold cross-validation.

Table 5: Subgroup analysis for model fairness and robustness evaluation. Metrics reported across demographic strata using 10-fold cross-validation.

Subgroup	N	Accuracy	AUC	Precision	Recall
Age Group					
18–40 years	78	0.96 ± 0.03	0.97 ± 0.02	0.95 ± 0.04	0.94 ± 0.05
41–60 years	152	0.95 ± 0.02	0.96 ± 0.02	0.94 ± 0.03	0.94 ± 0.03
61–80 years	170	0.94 ± 0.03	0.95 ± 0.02	0.93 ± 0.04	0.92 ± 0.04
Gender					
Male	212	0.95 ± 0.02	0.96 ± 0.02	0.94 ± 0.03	0.94 ± 0.03
Female	188	0.95 ± 0.02	0.96 ± 0.02	0.94 ± 0.03	0.93 ± 0.04
Risk Stratum					
Low Risk	145	0.97 ± 0.02	0.98 ± 0.01	0.96 ± 0.03	0.95 ± 0.03
Medium Risk	128	0.94 ± 0.03	0.95 ± 0.02	0.93 ± 0.04	0.92 ± 0.04
High Risk	127	0.93 ± 0.03	0.94 ± 0.03	0.92 ± 0.04	0.91 ± 0.05

Robustness Checks

Performance remained stable under alternative imputation strategies (median versus KNN), the addition of Gaussian noise (up to **5%** of the feature standard deviation), and stratified subpopulation splits. These robustness checks confirm that the proposed hybrid model generalizes well under data perturbations.

DISCUSSION

The proposed hybrid ML–DL framework addresses several key limitations of standalone ML and DL approaches. Traditional ML models, such as Logistic Regression and Random Forest, achieve reasonable predictive accuracy and provide partial interpretability; however, they generally lack calibration analysis, robustness evaluation, and fairness assessment. Deep Neural Networks offer higher predictive accuracy than

conventional ML models, but their limited interpretability and the absence of clinical utility assessment restrict their adoption in real-world clinical practice.

In contrast, the proposed hybrid model not only achieves the best predictive performance (Accuracy = 0.95, AUC = 0.96) but also demonstrates superior calibration (lowest Brier score = 0.07) and the highest net clinical benefit in Decision Curve Analysis. The incorporation of SHAP-based interpretability provides both global and local model explanations, while robustness checks and subgroup fairness analyses confirm the model's generalizability across diverse patient populations.

The weighted fusion strategy ($w_{\text{LR}} = 0.25$, $w_{\text{RF}} = 0.35$, $w_{\text{XGB}} = 0.40$),

derived from Equation (9), outperformed direct concatenation (AUC 0.958 versus 0.951) while requiring 33% fewer input parameters, demonstrating that compact and interpretable representations can effectively complement deep learning. Although the UCI CKD dataset contains a relatively limited number of samples ($N = 400$), the use of dropout regularization (dropout rate = 0.3), early stopping, and stratified 10-fold cross-validation effectively mitigated overfitting, as confirmed by the ablation studies.

CONCLUSION

This study presented a novel hybrid ML–DL framework for chronic kidney disease prediction by integrating machine learning-derived risk scores with deep neural representations. Experimental results obtained using the UCI CKD dataset demonstrated that the proposed model consistently outperformed both classical ML algorithms and a standalone DNN, achieving an accuracy of 95% and an AUC of 0.96.

The three primary contributions of this work are: (i) the introduction of an interpretable hybrid ML–DL fusion strategy that improves predictive performance; (ii) a rigorous evaluation of calibration, clinical utility, and fairness, which are often overlooked in previous studies; and (iii) the demonstration of model stability and interpretability through ablation studies and SHAP-based analyses.

Future work will focus on validation using external multi-centre cohorts, extending the framework with attention-based or gating mechanisms, incorporating temporal and longitudinal patient data for trajectory-based risk prediction, integrating uncertainty quantification and counterfactual explanations, and exploring federated or privacy-preserving learning paradigms for deployment in real-world healthcare environments.

CONFLICT OF INTEREST

The authors declare that no financial interests or conflicts of interest exist.

ACKNOWLEDGEMENTS

The authors thank MIT Mysore for providing the computational resources and infrastructure support necessary for this research.

REFERENCES

- World Health Organization. (2023). *Chronic kidney disease: Key facts*. WHO Press.
- GBD Chronic Kidney Disease Collaboration. (2020). Global burden of chronic kidney disease, 1990–2017. *The Lancet*, 395(10225), 709–733.
- Kidney Disease: Improving Global Outcomes. (2022). *KDIGO 2022 clinical practice guideline for the evaluation and management of chronic kidney disease*. *Kidney International*.
- Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., & Yang, C. W. (2013). Chronic kidney disease: Global dimension and perspectives. *The Lancet*, 382(9888), 260–272.
- Sisodia, D., & Sisodia, D. S. (2018). Prediction of chronic kidney disease using data mining. In *Proceedings of the International Conference on Trends in Big Data Analytics (ICTBIG 2018)*.
- Alhassan, J. K., Hammed, B., Usman-Hamza, F. E., & Balogun, A. O. (2021). Predicting chronic kidney disease using machine learning. *Journal of Healthcare Informatics Research*.
- Yan, Y., Zhang, J. W., Zang, G. Y., & Pu, J. (2022). A comprehensive study of chronic kidney disease risk prediction using machine learning. *Scientific Reports*, 12(1), Article 18401.
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25, 24–29.
- Liu, F., Zhou, X., Wang, C., & Cao, J. (2023). Hybrid machine learning/deep learning models in healthcare: A review. *WIREs Data Mining and Knowledge Discovery*, 13(2), e1487.
- Lokesh, B. S., & Kaulgud, N. (2025). Secure quantum machine learning. *Journal of Information Systems Engineering and Management*, 10(49s), 10122.
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/1412.6980>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).
- Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574.
- Dietterich, T. G. (1998). Approximate statistical

- tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895–1923.
21. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837–845.
22. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *The New England Journal of Medicine*, 380(14), 1347–1358.
23. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.
24. Salekin, A., & Stankovic, J. (2021). Detection of chronic kidney disease and selecting important predictors using random forest. *BMC Medical Informatics and Decision Making*, 21(1), Article 252.
25. Luo, Y., Yin, K., Bhatt, D. L., et al. (2024). Assessing and improving the clinical AI lifecycle. *Journal of the American Medical Informatics Association*, 31(3), 539–548.
26. Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.