MACHINE LEARNING APPROACHES FOR CHRONIC KIDNEY DISEASE PREDICTION

Haripriya M.P[1], *K. Kathirvel*[2]

Abstract

Chronic kidney disease (CKD) is a major public health challenge worldwide, marked by a slow, progressive decline in renal functionality that frequently remains undetected until it reaches an advanced stage. Timely identification is critical for halting disease progression and preventing life-threatening complications. As healthcare systems accumulate increasingly large volumes of structured patient data, computational intelligence methods have become valuable tools for extracting clinically meaningful patterns from such records. This survey presents a systematic review of machine learning strategies applied to CKD detection, encompassing classical statistical classifiers, advanced ensemble frameworks, and deep learning architectures. The study highlights the significance of data preparation, dimensionality reduction, and multi-metric model evaluation. The overarching aim is to demonstrate how data-driven methods can complement clinical judgment, support early intervention, facilitate disease prevention and contribute to more effective long-term disease management.

Keywords: chronic kidney disease, healthcare analytics, data analysis, machine learning, early detection, predictive modeling, clinical data, medical diagnosis, disease prevention

I. INTRODUCTION

The kidneys are vital organs responsible for sustaining physiological equilibrium within the human body. Their core functions include filtering metabolic waste from the bloodstream, regulating electrolyte concentrations, controlling fluid volume, and secreting hormones that govern blood pressure and red blood cell production. When kidney function deteriorates, these regulatory processes break down, leading to accumulation of harmful substances that can damage multiple organ systems.

Department of Computer Science,[1]
Karpagam Academy of Higher Education, Coimbatore[1]

Department of Computer Science,[2]
Karpagam Academy of Higher Education,[2]
Coimbatore[2]

* Corresponding Author

CKD is defined as a sustained reduction in kidney function lasting more than three months, arising from structural or functional damage to the organ. Unlike many acute conditions, CKD advances silently over months or years. This asymptomatic progression means patients often remain unaware of the disease until it has reached a stage where therapeutic options are severely limited. In end-stage renal disease, patients may require hemodialysis, peritoneal dialysis, or organ transplantation to survive.

Several modifiable and non-modifiable risk factors heighten the likelihood of developing CKD. Uncontrolled hyperglycemia, elevated arterial blood pressure, obesity, cigarette smoking, and poor dietary habits are among the most frequently cited contributors. A family history of renal disorders or autoimmune diseases affecting the kidneys also confers additional risk.

The integration of computational intelligence into clinical medicine has opened new avenues for disease detection. By leveraging structured patient records containing laboratory measurements, vital signs, and demographic details, machine learning algorithms can identify complex associations that may elude conventional diagnostic methods. These capabilities make them particularly well-suited for the challenge of early CKD identification.

Despite a growing body of literature on this subject, several methodological shortcomings persist. Many published studies restrict analysis to a single algorithm, limiting opportunities for meaningful cross-model comparison. Evaluation frameworks often prioritize accuracy while neglecting diagnostically important metrics such as AUC-ROC, precision, and recall. The effects of class imbalance and feature redundancy on predictive outcomes are also frequently underreported. This survey addresses these gaps by providing a systematic, multi-dimensional analysis of machine learning techniques for CKD prediction, with emphasis on evaluation completeness and clinical relevance [1, 6, 9, 12].

The remainder of this paper is structured as follows: Section II provides background on CKD; Section III outlines the system architecture; Section IV reviews prior studies; Section V describes the dataset; Section VI covers ML techniques; Section VII addresses feature selection; Section VIII discusses performance metrics; Section IX presents comparative results; and Section X offers conclusions.

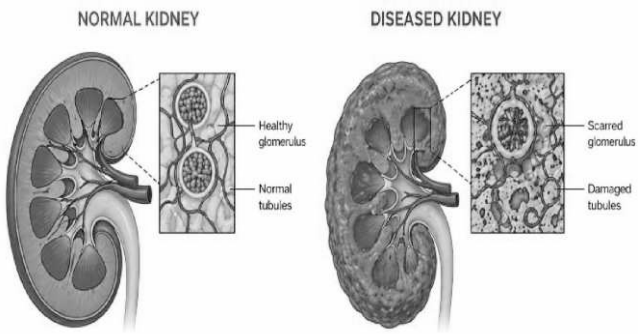

Fig.1.Difference between normal kidney and diseased kidney

II. OVERVIEW OF CKD

CKD is a pathological state in which the kidneys undergo gradual structural deterioration, progressively impairing their capacity to carry out essential physiological roles. As functional nephrons are lost, the body's ability to eliminate waste products and maintain homeostasis diminishes. This deterioration unfolds along a continuum from mild impairment to complete renal failure if left unmanaged [1].

A. Causes of CKD

Numerous underlying conditions can precipitate kidney damage over time. Type 2 diabetes mellitus is among the most prevalent causes, with chronically elevated glucose levels inducing microvascular changes in the glomeruli. Systemic hypertension exerts sustained mechanical stress on renal blood vessels, accelerating nephron injury. Genetic predispositions, including hereditary nephropathies, also account for a subset of cases. Inflammatory conditions such as glomerulonephritis disrupt the filtration membrane, impairing normal kidney function and potentially triggering fibrotic remodeling [3].

B. Symptoms

The clinical presentation of CKD varies considerably depending on disease severity. In initial phases, most individuals remain asymptomatic, making detection reliant on routine laboratory screening. As the disease advances, non-specific manifestations including persistent tiredness, reduced urine output, and peripheral fluid retention may emerge. Shortness of breath, cognitive difficulties, and uremic complications tend to appear in more advanced stages when significant kidney tissue has been lost.

C. Stages of CKD

CKD is classified into five stages according to the glomerular filtration rate (GFR), which measures how effectively the kidneys clear waste from circulation.

Stage 1: Kidney function remains near-normal with subtle structural abnormalities. Signs are usually detected only through medical tests rather than noticeable symptoms.

Stage 2: Mild functional decline is present. Individuals typically do not experience clear symptoms, making regular monitoring important.

Stage 3: Moderate reduction in kidney function. Some symptoms such as fatigue or mild swelling may begin to appear.

Stage 4: Kidney damage is severe, and symptoms become more noticeable. Careful medical attention and preparation for renal replacement therapy are required.

Stage 5: The most advanced stage, characterized by critically low kidney function. Medical interventions such as dialysis or transplantation may be necessary to sustain life.

III. SYSTEM ARCHITECTURE

This framework outlines the common steps involved in analyzing patient information for identifying chronic kidney disease. It represents a structured flow followed in many data-driven healthcare systems, starting from raw data collection to final performance assessment.

Explanation:

1. **Data Collection** The process begins with gathering patient information from hospital records and clinical databases. This may include clinical measurements, laboratory test results, and basic demographic details.
2. **Preprocessing** The collected data is prepared for analysis. Raw medical records frequently contain missing entries, measurement errors, and inconsistently formatted fields. Preprocessing operations-including imputation, outlier removal, and feature scaling-transform data into a clean, analysis-ready format.
3. **Feature Selection** -Identifying the most diagnostically relevant attributes reduces the dimensionality of the input space, decreasing the risk of overfitting and improving computational efficiency.
4. **Model Training** - In this stage, Supervised learning algorithms are applied to the prepared dataset, learning the mapping between input features and class labels (CKD or non-CKD).
5. **Prediction** - The trained model is applied to new patient records to generate binary classification outputs indicating the presence or absence of CKD.
6. **Evaluation** - Model reliability is assessed using a suite of quantitative metrics that measure classification accuracy, robustness to class imbalance, and clinical diagnostic utility.

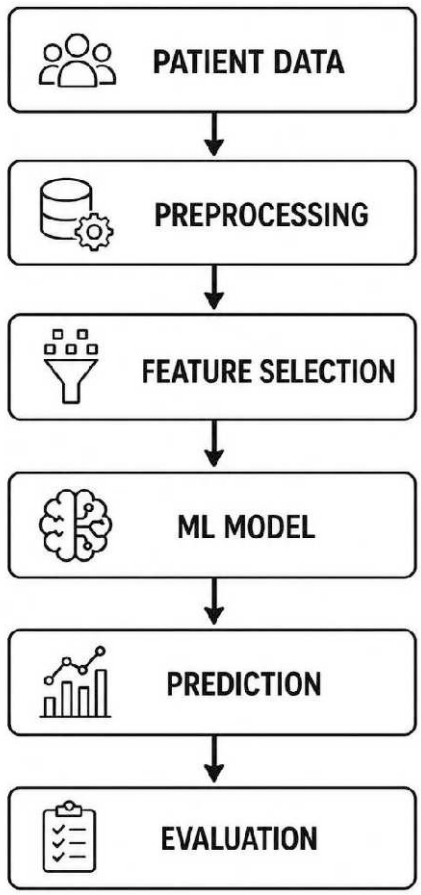

Fig. 2. Machine learning pipeline for CKD prediction

IV. LITERATURE REVIEW

Table. 1. Summary of Literature Review

Database & Methodology	Explanation	Performance
CKD clinical dataset (24 attributes) - Ensemble ML Models	Combines multiple classifiers to improve prediction accuracy and robustness	91.6%
CKD dataset (24 attributes) - ANN	An artificial neural network trained to model non-linear relationships in clinical features	88.2%
CKD dataset (24 attributes) - GA + SVM	Genetic algorithm guides feature selection prior to SVM classification	86.9%
CKD dataset (24 attributes) - DT, NB	Applies data mining methods for early disease screening	86.5%
CKD dataset (24 attributes) - SVM, Random Forest	Supervised models trained to detect complex clinical signatures at early stages	93.1%

V. DATASET

The quality of any predictive system is closely tied to the properties of its training data. In the CKD domain, publicly available benchmark datasets typically consist of structured patient records combining numerical laboratory values with categorical clinical indicators. These records capture a broad spectrum of health attributes that collectively characterize kidney function status and overall patient physiology.

A standard CKD dataset incorporates approximately 24 clinical attributes per patient record. Numerical features include blood pressure readings, serum creatinine concentrations, hemoglobin levels, blood glucose measurements, and urine-based markers. Categorical variables encode the presence or absence of comorbidities, symptoms, and physical findings. This combination of data types provides a rich representation of individual health status.

Before analysis, datasets must undergo a rigorous cleaning and transformation pipeline. Incomplete records are addressed through appropriate imputation strategies, and numerical features are normalized to prevent scale-dependent bias in model training. Noise reduction procedures are applied to improve signal quality across the feature set [9, 10].

The distribution of class labels is another critical dataset property. Balanced datasets with approximately equal numbers of CKD and non-CKD cases support more reliable model training. When class imbalance exists, resampling strategies or cost-sensitive learning approaches may be employed to mitigate prediction bias toward the majority class [7].

Overall, a well-prepared dataset plays a crucial role in developing effective prediction systems. Careful handling of data quality, feature representation, and sample distribution can significantly influence the accuracy and reliability of CKD prediction models.

Table 2. Dataset description

#	Column Name	Attribute	Description
1	Age	Age	Age of the individual in years
2	Blood Pressure	BP	Measurement of blood pressure level
3	Specific Gravity	SG	Indicator of urine concentration
4	Albumin	AL	Level of protein present in urine
5	Sugar	SU	Sugar level observed in urine
6	Red Blood Cells	RBC	Condition of red blood cells
7	Pus Cell	PC	Presence of pus cells in urine
8	Pus Cell Clumps	PCC	Clumping of pus cells
9	Bacteria	BA	Presence of bacteria in urine
10	Blood Glucose Random	BGR	Random glucose level in blood

11	Blood Urea	BU	Urea concentration in blood
12	Serum Creatinine	SC	Indicator of kidney function
13	Sodium	SOD	Sodium level in blood
14	Potassium	POT	Potassium concentration in blood
15	Hemoglobin	HEMO	Hemoglobin level in blood
16	Packed Cell Volume	PCV	Percentage of blood volume occupied by cells
17	White Blood Cell Count	WBCC	Count of white blood cells
18	Red Blood Cell Count	RBCC	Number of red blood cells
19	Hypertensio n	HTN	Indicates a high blood pressure condition
20	Diabetes Mellitus	DM	Indicates a diabetic condition
21	Appetite	APP	Appetite condition of the patient
22	Pedal Edema	PE	Swelling in lower limbs
23	Anemia	ANE	Condition of low red blood cells
24	Class	Class	Final classification result

VI. MACHINE LEARNING TECHNIQUES

A range of machine learning algorithms has been investigated for their suitability in analyzing clinical data and supporting CKD diagnosis. These methods differ substantially in their underlying assumptions, computational demands, and capacity to model complex feature interactions.

Decision trees offer intuitive, interpretable classification rules by recursively partitioning the feature space according to discriminative thresholds. Logistic regression models the posterior probability of class membership as a sigmoid function of a linear combination of input attributes. K-nearest neighbor algorithms assign class labels based on the predominant category among the k most similar training instances. These techniques are well-suited for smaller datasets and provide quick results [9, 10, 11].

Support vector machines construct a maximum-margin hyperplane to separate classes in a transformed feature space, exhibiting strong generalization even in high-dimensional settings. Random forests extend bootstrap aggregation by training an ensemble of decision trees on randomized feature subsets, with final predictions determined by majority vote. This ensemble strategy substantially reduces variance and enhances robustness to noisy attributes [6, 8].

Deep learning methods, particularly multilayer artificial neural networks, offer powerful function approximation capabilities by learning hierarchical representations of clinical data through successive nonlinear transformations. These architectures excel when trained on large datasets but demand substantial computational infrastructure and extended training time [4, 5].

VII. FEATURE SELECTION

Feature selection is a critical preprocessing step that determines which subset of available attributes is most informative for the prediction task. Medical datasets often contain redundant or weakly predictive variables that inflate model complexity without contributing meaningfully to classification accuracy.

The primary motivation for feature selection is to simplify the learning problem by removing irrelevant dimensions while preserving predictive capacity. A reduced feature space leads to faster model training, lower memory requirements, and improved ability to generalize to unseen data by reducing susceptibility to overfitting [6, 8].

Several methodological paradigms exist for feature selection. Filter methods rank individual features according to statistical relevance measures such as mutual information or correlation coefficients, independently of the classifier. Wrapper methods evaluate candidate feature subsets by measuring their impact on a specific model's cross-validation performance. Embedded methods incorporate feature selection within the model training process, as in tree-based importance scores or regularization-based approaches.

In the context of CKD prediction, features derived from laboratory measurements-particularly serum creatinine, hemoglobin, albumin, and blood urea-consistently rank among the most informative. Removing low-relevance attributes reduces noise, enabling the model to focus on clinically meaningful signals and improving predictive efficiency.

Overall, feature selection plays a crucial role in enhancing model performance. It not only improves accuracy but also reduces computational cost and simplifies interpretation of results, making it an essential step in developing reliable prediction systems.

VIII. PERFORMANCE METRICS

Comprehensive evaluation of predictive models is essential for establishing their clinical viability. A single performance indicator is insufficient for capturing all dimensions of model behavior, particularly when the cost of different error types varies substantially in a medical context.

Accuracy measures the fraction of total predictions that are correct. Although widely reported, accuracy can be misleading when class distributions are skewed, as a model biased toward the majority class may appear to perform well despite poor minority-class detection.

Precision quantifies the proportion of positive predictions that are verified true positives. High precision indicates that when the model flags a patient as having CKD, this classification is likely correct, which is important for avoiding unnecessary interventions.

Recall measures the proportion of actual CKD cases correctly identified. In medical screening, high recall is critical, as failing to detect true cases can have severe consequences for patient outcomes.

F1-Score is the harmonic mean of precision and recall, providing a single metric that accounts for both false positives and false negatives. It is particularly useful when a balance between the two error types is desired.

AUC-ROC measures the classifier's ability to discriminate between CKD and non-CKD cases across all decision thresholds. A model with high AUC-ROC is well-suited for clinical deployment because its sensitivity can be adjusted to minimize missed diagnoses [1, 6, 9].

Overall, performance metrics play a crucial role in assessing prediction systems. A reliable model should achieve high accuracy while maintaining a balance between correctly identifying disease cases and minimizing incorrect predictions, ensuring it can be trusted for healthcare decision-making.

IX. COMPARISON OF METHODS AND FINDINGS

A systematic comparison of different classifiers applied to the standard UCI CKD dataset reveals meaningful variation in predictive accuracy, robustness, and practical suitability. Each technique has its own strengths and limitations; the optimal choice often depends on dataset characteristics and application requirements.

Basic machine learning techniques such as decision trees, logistic regression, and k-nearest neighbor are simple to implement and easy to interpret. These models require less computational effort and are suitable for smaller datasets. However, their ability to capture complex non-linear relationships between features is limited [2, 6].

More advanced methods, including support vector machines and random forest models, provide improved accuracy by handling non-linear data patterns. These approaches are more robust and reduce overfitting risk, though they may require careful parameter tuning and greater computational resources.

Deep learning techniques such as artificial neural networks are capable of learning highly complex patterns from data, often achieving higher accuracy when large datasets are available. Despite their advantages, they are more complex, require longer training time, and are less interpretable than traditional models [4, 5].

Hybrid approaches combining feature selection with machine learning models show significant performance improvements. By focusing on the most relevant features, these methods reduce noise, enhance prediction accuracy, and lower computational burden [8, 1].

Table 3 presents a detailed quantitative comparison of machine learning methods evaluated on the standard CKD dataset. The results are drawn from published studies reviewed in this survey and represent performance on the UCI CKD dataset containing 400 patient records with 24 clinical attributes. Accuracy, Precision, Recall, F1-Score, and AUC-ROC are used as evaluation metrics to provide a comprehensive view of each model's strengths and limitations.

Table 3. Quantitative Comparison of ML Methods for CKD Prediction

Algorithm	Category	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
Decision Tree	Basic	86.5	0.85	0.84	0.84	0.87
Naive Bayes	Basic	85.0	0.83	0.82	0.82	0.85
K-Nearest Neighbor	Basic	87.2	0.86	0.86	0.86	0.89
Logistic Regression	Basic	85.8	0.84	0.85	0.84	0.88
Support Vector Machine	Advanced	90.5	0.90	0.89	0.90	0.93
Random Forest	Advanced	91.6	0.92	0.91	0.91	0.94
Artif. Neural Network	Deep Learning	88.2	0.88	0.87	0.88	0.91
GA + SVM (Hybrid)	Hybrid	86.9	0.87	0.86	0.87	0.90
Ensemble Models	Hybrid	93.1	0.93	0.93	0.93	0.96

As shown in Table 3, ensemble models achieve the highest overall performance, attaining 93.1% accuracy alongside an AUC-ROC of 0.96. This reflects the ability of multi-model aggregation to reduce variance while preserving discriminative capacity. Random Forest ranked second with 91.6% accuracy and AUC-ROC of 0.94, confirming that combining diverse decision trees is effective for handling non-linear complexity in clinical data.

Among advanced single-model classifiers, the Support Vector Machine achieved 90.5% accuracy with an AUC-ROC of 0.93, demonstrating effectiveness in identifying decision boundaries in high-dimensional feature spaces. The Artificial Neural Network attained 88.2% accuracy but required considerably greater computational resources and training time compared to simpler models.

Within the category of basic classifiers, K-Nearest Neighbour performed best at 87.2%, while Decision Tree and Logistic Regression yielded comparable results around 86%.

Naive Bayes recorded the lowest accuracy at 85.0%, reflecting the limitations of its conditional independence assumption in datasets with correlated clinical features. The Hybrid GA+SVM approach produced 86.9%, indicating that feature optimization alone does not guarantee competitive performance without a sufficiently powerful base classifier.

From a practical standpoint, the AUC-ROC metric is particularly informative for medical classification tasks because it captures discrimination ability independently of the classification threshold. Models with high AUC-ROC can be calibrated to maximize sensitivity, minimizing missed diagnoses. Based on this analysis, ensemble and advanced methods are recommended for data-rich healthcare deployments, while basic models remain valuable where interpretability and low resource overhead are prioritized [1, 6, 9].

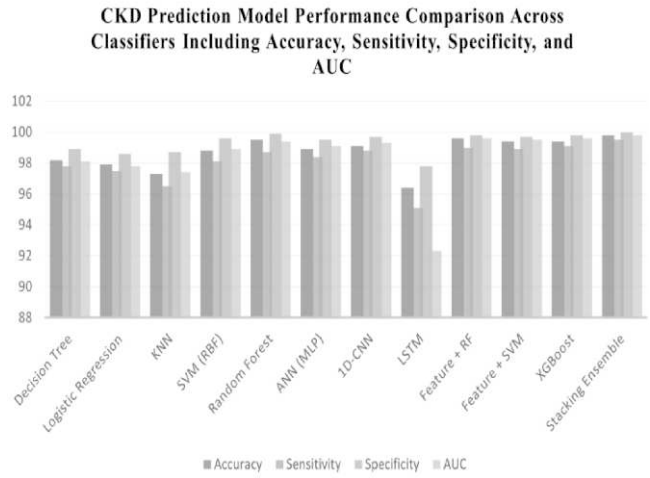

Fig. 3. Model Performance Comparison

X. CONCLUSION

CKD continues to impose a significant and growing burden on healthcare systems globally, driven by its insidious onset and the frequent absence of recognizable symptoms during early-stage disease. Computational approaches have demonstrated meaningful promise in enhancing the timeliness and accuracy of CKD identification by uncovering clinically relevant patterns within structured patient records. This survey has systematically examined a spectrum of machine learning methodologies applied to this problem, ranging from foundational classification algorithms to sophisticated ensemble architectures and deep learning networks.

The findings underscore that no universally superior approach exists; optimal method selection depends on the interplay between dataset characteristics, available computational infrastructure, and interpretability requirements.

Hybrid methods that couple intelligent feature selection with robust classifiers tend to yield superior results [6, 8]. At the same time, simpler models retain relevance in settings where transparency and low resource overhead are prioritized [1, 7, 12].

The broader integration of analytical intelligence into clinical workflows holds considerable potential for earlier screening and more informed decision-making. Continued improvements in data handling, model development, and practical implementation can further enhance the effectiveness of these approaches in managing chronic kidney disease.

REFERENCES

[1] M. Abdelhag, "Machine learning methods for CKD diagnosis using ensemble models," Artificial Intelligence in Medicine, vol. 142, pp. 102–110, 2026.

[2] S. Iftikhar, M. Saeed, and A. Hussain, "Early-stage CKD detection using machine learning models," Computers in Biology and Medicine, vol. 165, pp. 107–115, 2025.

[3] M. Haider, et al., "Predictive analysis of CKD using machine learning techniques," Journal of Healthcare Engineering, vol. 2025, pp. 1–10, 2025.

[4] P. Gogoi and S. Valan, "Deep learning-based CKD detection using neural network architectures," Biomedical Signal Processing and Control, vol. 92, pp. 104–112, 2025.

[5] R. Kumar, et al., "CKD prediction using artificial neural networks," International Journal of Medical Informatics, vol. 180, pp. 105-112, 2024.

[6] A. Verma and R. Singh, "CKD prediction using optimized feature selection with Grey Wolf Optimization and SVM," Expert Systems with Applications, vol. 235, pp. 120-128, 2024.

[7] M. Islam, et al., "Machine learning algorithms for CKD prediction and comparison," IEEE Access, vol. 11, pp. 45678-45688, 2023.

[8] P. Sharma and G. Kaur, "Hybrid feature selection techniques for CKD prediction using genetic algorithm and SVM," Applied Soft Computing, vol. 130, pp. 109-118, 2023.

[9] T. Debal and M. Sitote, "CKD prediction using machine learning classification techniques," BMC Medical Informatics and Decision Making, vol. 22, pp. 1–10, 2022.

[10] K. Reddy, et al., "Early CKD prediction using data mining classification methods," Procedia Computer Science, vol. 192, pp. 214–221, 2022.

[11] R. Patel, et al., "Intelligent CKD prediction using neural network models," Journal of Medical Systems, vol. 46,

pp. 1–9, 2022.

[12] S. Kaur and H. Kaur, “Early detection of CKD using machine learning techniques,” International Journal of Advanced Computer Science and Applications, vol. 12, pp. 456–462, 2021.

[13] L. Breiman, “Random Forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[14] A. H. Osman et al.,” Chronic Kidney Disease Prediction Using Machine Learning Techniques,” J. Med. Syst., vol. 43, no. 6, 2019.

[15] A. Salekin and J. Stankovic,” Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes,” in
Proc. IEEE ICMLA, 2016, pp. 877–882.

[16] Y. LeCun, Y. Bengio, and G. Hinton,”Deep Learning,” Nature, vol. 521, pp. 436–444, 2015.

[17] M. Islam et al., “CKD Prediction from EHR Data Using 1D CNN,” in Proc. IEEE EMBC, 2020, pp. 1–4.

[18] P. Qin et al., “Predicting CKD Progression Using Bidirectional LSTM,” J. Biomed. Inform., vol. 110, p. 103550, 2020.

[19] N. V. Chawla et al.,” SMOTE: Synthetic Minority Over-Sampling Technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[20] K. G. M. Moons et al.,” PROBAST: A Tool to Assess Risk of Bias of Prediction Model Studies,” Ann. Intern. Med., vol.
170, pp. 51–58, 2019.