RESEARCH PAPER

THE TRANSLATIONAL DIVIDE IN CHRONIC KIDNEY DISEASE MACHINE LEARNING: A SYSTEMATIC REVIEW OF PERFORMANCE INFLATION, VALIDATION GAPS, AND CLINICAL READINESS

Abdulmumini Imam Ibrahim[1], Dr. Manish Madhava Tripathi[2]
[1]PhD scholar, Computer Science and Engineering Department, Integral University Lucknow, India
[2]Prof., Computer Science and Engineering Department, Integral University Lucknow, India
[1]imandoro23@gmail.com, [2]mmt@iul.ac.in
Corresponding Author: Abdulmumini Imam Ibrahim

Abstract

Background

Chronic kidney disease (CKD) has more than 800 million cases in the global population and is often diagnosed in its late stages because of its insidious onset. The use of machine learning (ML) and deep learning (DL) in CKD detection, prediction of risk and prognosis has become more used. Nevertheless, reported performance differs across studies, and there are some concerns about generalisability, methodological rigour and clinical readiness, especially in relation to the potential impact that these factors have to the application of ML and DL models in a variety of patients and clinical environments.

Objective

The aim is to critically review and evaluate the machine learning models developed to predict the early detection, progression, and risk stratification of CKD systematically, with an emphasis on the quality of the methodology, the approach to validation, interpretability, and translational use.

Methods

A PRISMA-compliant systematic search was conducted across PubMed, Scopus, IEEE Xplore, Web of Science and ScienceDirect for studies published between January 2015 and March 2026. Eligible studies applied machine learning or deep learning methods to human CKD datasets and reported quantitative performance metrics. Extracted data included dataset size, modality (tabular, imaging, EHR, multimodal), algorithms used, performance metrics, validation strategy, and interpretability methods. A qualitative synthesis was performed due to methodological heterogeneity. A two-component synthesis was employed: a qualitative narrative synthesis across all included studies and a stratified quantitative analysis of AUC values (n = 35 studies) to formally test the performance inflation hypothesis using Welch's t-test, Cohen's d, and Spearman rank correlation.

Results

Fifty-eight studies met the inclusion criteria. Most studies relied on tabular clinical datasets, frequently using small benchmark datasets ($n < 500$). Ensemble methods, logistic regression and random forests are the most commonly applied algorithms. Studies using small curated datasets often reported near-perfect performance (98-100%), whereas large real-world cohort studies reported more moderate performance with an area under the curve (AUC) ranging from 0.78 to 0.87. Imaging-based deep learning models attained accuracies ranging from 80% to 86%. Only one study performed true external validation across independent cohorts, and calibration reporting was limited across the reviewed literature, which raises fundamental concerns about the generalisability and clinical reliability of the reported performance metrics. Interpretability methods, particularly SHAP, have been increasingly incorporated in recent studies.

Conclusions

Machine learning models demonstrate strong potential for CKD detection and risk prediction. However, substantial gaps remain in external validation, calibration reporting, fairness assessment and prospective clinical evaluation. Inflated performance in small datasets and limited translational assessment restrict current clinical applicability. Future research should prioritise multicentre validation, longitudinal modelling, multimodal integration, and explainable-by-design approaches to enable reliable clinical deployment-ready CKD and AKI prediction systems.

Keywords: chronic kidney disease; machine learning; deep learning; systematic review; external validation; performance inflation; bias–variance decomposition; SHAP; clinical readiness; nephrology

How to cite this article: Abdulmumini Imam Ibrahim,, ManiMadhava Tripathi. THE TRANSLATIONAL DIVIDE IN CHRONIC KIDNEY DISEASE MACHINE LEARNING: A SYSTEMATIC REVIEW OF PERFORMANCE INFLATION, VALIDATION GAPS, AND CLINICAL READINESS. Int J Drug Deliv Technol. 2026;16(79s):445-468. DOI: 10.25258/ijddt.16.79s.78.

1. Introduction

Chronic kidney disease (CKD) has become one of the major worldwide public health obstacles of the 21st century, impacting an approximated 800 million people globally and causing a high level of cardiovascular morbidity and early death, as well as

a huge health cost burden [1] [2]. The development of CKD is often asymptomatic during the initial phases of the disease and many patients are only diagnosed with it after considerable and in most cases irreversible renal damage has already taken place [3]. Late diagnosis impairs the efficacy of renoprotective treatments and speeds up the progression to end-stage kidney disease (ESKD), necessitating dialysis or transplantation [4].

Traditional CKD diagnostics is mostly based on the estimated glomerular filtration rate (eGFR) based on serum creatinine and urinary albumin levels [5]. Nonetheless, serum creatinine is affected by age, gender, muscle mass, ethnicity, and albuminuria can vary because of temporary physiological conditions [6]. Such limitations, along with logistical and economic limitations of laboratory-based screening, especially in low-resource environments, suggest the significance of more scalable and predictive diagnostic methods [7].

Along with the growth of electronic health records (EHRs) and medical imaging repositories, machine learning (ML) and deep learning (DL) techniques have been used more frequently to detect CKD and stratify risks [8][9][10]. Initial research primarily applied classical ML methods including the logistic regression, support vectors machines (SVM) and random forests (RF) to structured clinical data [11][12][13]. Gradient boosting models (e.g., XGBoost), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) have more recently been put in place to model longitudinal EHR data, as well as imaging modalities, such as ultrasound and computed tomography (CT) radiomics[14][15][16][17].

The performance of the models reported in different studies differs significantly. A number of studies based on small curated datasets, most notably the popularly re-used UCI CKD dataset (n=400) have claimed classification rates of over 98% [11][18][19]. On the contrary, big, real-world, cohort studies based on national surveys like NHANES or insurance claims data show moderate performance (AUC 0.78-0.87) [12][20]. Deep learning models of CKD detection using imaging also exhibit moderate accuracy (80-86%), indicating possible modality constraints [16][21]. This inconsistency casts doubt on overfitting, data leakage, and a lack of generalisability in small benchmark data. Very high accuracies on small benchmark datasets might be due to overfitting, data leakage and low external validity.

Integration of interpretability frameworks is another theme that is coming out. Recent cohort-based studies are increasingly using SHAP (Shapley Additive Explanations), feature selection stability analysis, and Bayesian methods to improve transparency and clinical credibility [12][22]. However, external validation is still uncommon, and less than 15% of the published models have been evaluated on independent datasets [20][23]. Calibration reporting and subgroup fairness analysis are also rarely discussed, though they are also of vital importance when it comes to clinical deployment [24].

Even though a number of narrative reviews have overviewed the application of artificial intelligence in nephrology [25][26], not many have undertaken a systematic assessment of methodological rigour, practices of external validation, calibration reporting, and translational readiness. Furthermore, little attention has been given to the consistent discrepancy between small benchmark datasets and heterogeneous real-world cohorts in terms of performance. The actual clinical preparedness of CKD-oriented machine learning systems is unknown without intensive synthesis and critical assessment. We operationally define the "Translational Divide" in this review as the measurable and reproducible gap between (i) the high discriminative performance reported by ML models evaluated on small, curated benchmark datasets using internal cross-validation and (ii) the substantially lower and more variable performance observed in independent, large-scale, real-world clinical cohorts. This divide encompasses not only performance inflation attributable to overfitting and data leakage on benchmark datasets, but also the broader set of validation, calibration, fairness, and prospective deployment deficits that collectively prevent a model from meeting the criteria for safe clinical use.

However, methodological progress has been very rapid, with some gaps that have been identified such as poor external validation, inadequate calibration reporting, no explanation of fairness, and underuse of longitudinal disease modelling. Moreover, the gap between the performance on small benchmark datasets and actual cohorts is not adequately covered, which casts doubts on the overall generalisability of the results and their useability in clinical practice.

Thus, in this research, we intend to:

1. Describe the methodological space of CKD ML research.
2. Compare the performance between types of datasets and validation strategies critically, and
3. Assess translational preparedness, such as external validation, integratability of interpretation, and quality of reporting.

2. Methods

2.1 Study Design

This systematic review was carried out in line with the guidelines of preferred reporting items of systematic reviews and meta-analysis (PRISMA 2020) and follows methodological principles of prediction model reviews described in the TRIPOD and PROBAST models.

- This aimed to systematically find, review and qualitatively synthesise peer-reviewed

studies that used machine learning (ML) or deep learning (DL) methods to early detection of chronic kidney disease (CKD)
- Prediction of CKD progression
- Diabetic kidney disease (DKD) risk modelling
- Renal outcome prediction (including ESRD and mortality)
- Related nephrology ML frameworks (including AKI when methodologically relevant)

A qualitative synthesis approach was adopted due to substantial heterogeneity in:

- Dataset types
- Feature engineering
- Model architectures
- Performance metrics
- Validation strategies

2.2 Search Strategy

From the following database we have performed the search:

- PubMed/MEDLINE
- Scopus
- Web of Science
- IEEE Xplore
- ScienceDirect

The search covered publications from January 2015 to February 2026. Full database-specific search strings for each of the five databases are provided in Supplementary Material S1 for reproducibility.

Core Search String

("chronic kidney disease" or CKD or diabetic kidney disease or DKD or renal failure or ESRD or acute kidney injury or AKI) and (machine learning or deep learning or artificial intelligence or neural network or random forest or gradient boosting or LSTM or CNN) and (prediction or early detection or progression) The search was limited to:

- English language publications
- Peer-reviewed journal articles
- Human studies

Conference abstracts, editorials, commentaries, and unindexed journals were excluded.

2.3 Eligibility Criteria

Inclusion Criteria

Studies were included if they:

- Applied ML/DL models to CKD/DKD/renal outcome prediction
- Used structured clinical data, imaging, genomics, or multimodal datasets
- Reported at least one predictive performance metric (AUC, accuracy, sensitivity, etc.)
- Were published in peer-reviewed indexed journals
- Included adult human populations

Exclusion Criteria

Studies were excluded if they:

- Used non-peer-reviewed sources (e.g., student journals and IJERT-type publications)
- It was merely a narrative review without empirical modelling.
- Focused solely on feature engineering without predictive modelling
- Were retracted
- Used synthetic or simulated datasets only
- Lacked clear performance reporting

2.4 Study Selection Process

The selection of the study was conducted in three steps:

1. Title screening
2. Abstract screening
3. Full-text eligibility assessment

There were duplicates that were eliminated before the screening.

The methodological evaluation was conducted by consensus; disputes were solved in this way.

2.5 Data Extraction

A standardised collection template was used to independently extract data.

Variables were collected as follows in each of the included studies:

- First author and year of publication
- Country and dataset source
- Study design (retrospective/prospective)
- Sample size
- Population characteristics
- Data modality (clinical, imaging, genomics, multimodal)
- Machine learning algorithm(s) used
- Feature selection methods
- Validation strategy (cross-validation, hold-out, external validation).
- Performance metrics (AUC, accuracy, sensitivity, specificity, F1-score)
- Calibration reporting (if available)
- Interpretability methods (e.g., SHAP, feature importance)
- Presence of external validation
- Risk of bias indicators (based on PROBAST domains)

Where multiple models were reported, the best-performing model was extracted for comparative synthesis.

2.6 Data Synthesis Strategy

Because of high levels of heterogeneity in study designs, data, model architectures, and performance metrics reported, a quantitative meta-analysis was impractical. Thus, the qualitative synthesis method was used, and it aimed to determine common patterns of methodology, trends in the performance of various types of datasets, and the gaps in the validation and reporting practices.Because of high levels of heterogeneity in study designs, dataset characteristics, model architectures, outcome

definitions, and reported performance metrics, a standard random-effects meta-analysis was impractical. A two-component synthesis strategy was therefore employed. First, a qualitative narrative synthesis was conducted across all 58 included studies to identify common methodological patterns, performance trends across dataset types, and gaps in validation and reporting practices. Second, a stratified quantitative analysis was performed on the subset of 35 studies that reported extractable AUC values or sufficient data to derive AUC estimates, with the aim of formally testing specific a priori hypotheses regarding performance inflation as a function of dataset scale. This quantitative component comprised inverse-variance weighted pooling of AUC values, Welch's t-test for between-stratum differences, Cohen's d effect size estimation, Spearman rank correlation between log-transformed sample size and reported AUC, and a bias–variance decomposition of the observed performance distributions. This stratified quantitative synthesis, presented in full in Section 3.12, is explicitly distinguished from a formal random-effects meta-analysis: no pooled treatment effect is estimated, heterogeneity statistics (I^2) are not reported, and stratum-level comparisons are used solely to quantify and formally test the performance inflation hypothesis rather than to derive a single aggregate effect estimate.

3. Results

3.1 Study Selection

In the initial database searches and additional sources 615 records were identified. The search in the database provided 548 records (PubMed/MEDLINE: 142, Scopus: 198, Web of Science: 121, IEEE Xplore: 87), and 67 records were obtained by screening the reference lists (n = 46), and preprint/institutional repositories (n = 21). Upon elimination of 172 duplicates, 443 individual records were left to be screened on titles.

The study selection process is described in a PRISMA 2020 flow diagram, which includes the identification, screening, eligibility and final inclusion of studies (Figure 1).

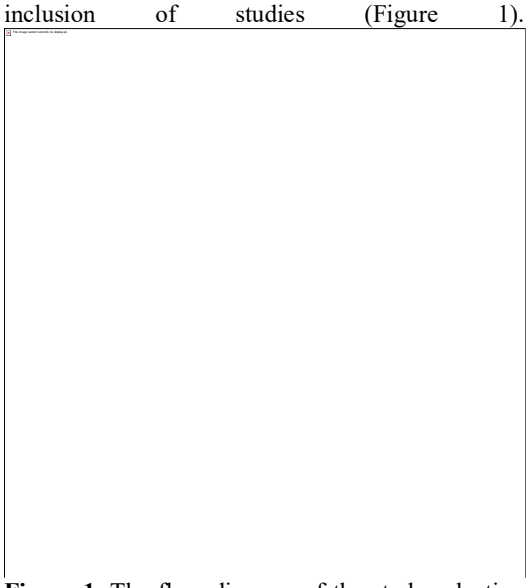

Figure 1. The flow diagram of the study selection process of machine learning applications in CKD in PRISMA 2020.

A total of 356 records were filtered out during the pre-screening (not meeting the inclusion criteria e.g. not related to chronic kidney disease, no application of machine learning, editorial/review paper, or not clinical-focused). Four of the 87 reports requested to be accessed could not be found in full-text format. As a result, 83 full-text articles were evaluated in terms of eligibility. After a thorough assessment, 25 articles were narrowed down due to the following reasons: no prediction model (n = 7), non-human or experimental study design (n = 5), inadequate methodological description (n = 6), publication of the same dataset (n = 4), and non-English language (n = 3).

The systematic search and screening led to the final selection and inclusion of 58 studies, described in the PRISMA flow diagram (Figure 1), to synthesize the quality and compare the model performance quantitatively. Because of the high level of heterogeneity of datasets and measures of evaluation in the literature, it was not conducted in a formal meta-analysis.

3.2 Data extraction and quality assessment.

Data were extracted independently in a standardised electronic format per each of the 58 studies included. The extraction process was geared towards the capture of the following dimensions:

• Characteristics of the study: Data source (e.g., UCI Machine Learning Repository, electronic health records of hospitals), size of the data, and data type (tabular, longitudinal, or image-based).

Technical Pipeline: missing data, preprocessing methods, and cross-validation.

• Model Architecture: (e.g., Random Forest, XGBoost and CNNs) machine learning algorithms were employed, as well as hyperparameter tuning procedures and feature engineering.

• Explainable AI (XAI): Methods used for model transparency (e.g., SHAP, LIME) and the evaluation criteria for explainability.
• Performance Metrics: The main results are accuracy, sensitivity, specificity, the F1-score and the area under the ROC curve (AUC).

3.3 Dataset Characteristics

Distribution of dataset modalities used in CKD machine learning studies. Clinical datasets are predominantly tabular, with little use of multimodal and imaging (Figure 2).

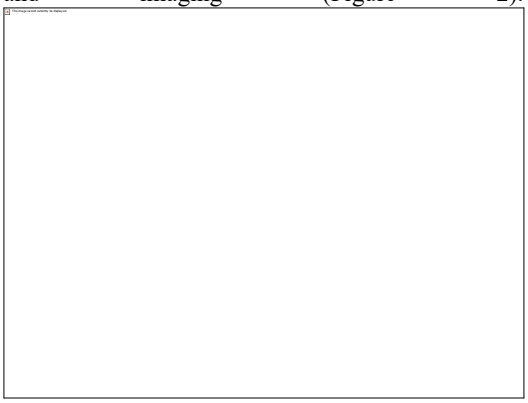

Figure 2. Dataset type distribution.

Structured clinical tabular dataset was the most common modality used amongst the CKD studies (n = 16). Six papers used the popular UCI CKD benchmark dataset (n = 400), with reported accuracy ranging between 98 to 100 percent with ensemble classifiers like Random Forest, AdaBoost and Rotation Forest [11][18]. These works were based on mostly K-fold cross-validation without any independent external validation. The models of diabetic kidney disease (DKD) tended to vary since they used fundus imaging modalities, which were not included in the conventional CKD datasets.

Large-scale national or claims-based datasets were used in two studies. [27] used a Taiwanese National Health Insurance claims database (around 90,000 participants, 18,000 CKD cases and 72,000 controls) and produced the ensemble of CNN, BLSTM, LightGBM, XGBoost, and AdaBoost to achieve an AUROC of 0.957 (6 months) and 0.954 (12 [28] used LSTM and SVM model on a 6,040 type 2 diabetes cohort in Taiwan that found an LSTM accuracy of 86% (AUC 0.83) and SVM accuracy of 76% (AUC 0.73) to predict diabetic kidney disease. This high performance of [27] indicates that properly designed ensemble models on national scale datasets can be nearly as accurate as small datasets, but neither study reported external validation.

Two studies used longitudinal EHR cohorts with over 10,000 patients, and applied temporal modelling methods, including gradient boosting and recurrent neural networks [29]. These reported AUC values were between 0.80 and 0.83.

Two studies involved imaging-based models, such as CT-based radiomics and ultrasound deep learning models [15]. These obtained AUCs ranging between 0.79 and 0.86.

A single multimodal study combined genomic, clinical and imaging data with the UK Biobank resource, which reported an AUC of 0.804 to predict ESRD five years [30][31].

3.4 Thematic Synthesis of Included Studies.

Table 1 provides a thematic synthesis of the studies included, categorizing them by the dataset characteristics, predictive objectives, modelling methods and important characteristics. The references of the studies used in the study show that there were individual studies in each category.

Table 1. Thematic synthesis of sampled machine learning studies in CKD (representative subset of 58 sampled studies), overarching dataset features, prediction tasks, model structures, validation methods and clinical predictors of importance. The studies were chosen to exemplify the spectrum of dataset scales, modalities, and algorithmic strategies that were discovered during the entire review.

Study	Study Population	Primary Goal	Machine Learning Models Used	Top Predictors	Performance Metrics (AUC/Accuracy)	Interpretability Method
[32] [13] [33] [34] [35] [12] [18] [36] [11] [37] [8]	400 instances from the UCI Repository (Apollo Hospital,	Prediction, early diagnosis, and classification of CKD using cost-sensitive learning and optimal feat	ANN, LSTM, GRU, Bidirectional LSTM/GRU, MLP, Simple RNN, AdaBoost, CS-AdaBoost, Perceptron, Logistic Regression,	Haemoglobin, specific gravity, albumin, serum creatinine, packed cell volume, red blood cell count, hypertension, diab	Accuracy: 96% to 100 %, AUC: 0.97 to 1.0.	Feature ranking (Perceptron, AdaBoost, IG filter), feature selection (WFS, CFS, RFE, Lasso, Boruta, BBO).

		ure subsets.	Decision Tree, XGBoost, Random Forest, SVM, KNN, Naive Bayes.	etes mellitus, and blood pressure.		
[38]	550,000 patient records (10 million insurance claims), USA.	Early detection and prediction of progression to End-Stage Renal Disease (ESRD).	XGBoost, Word2Vec, logistic regression, Random Forest, CatBoost, DNN.	Patient age, CKD stage, hypertensive crisis events, diagnosed hypertension, and ischaemic heart disease.	C-statistic: 0.93, Sensitivity: 0.715, Specificity: 0.958.	Feature importance analysis.
[31]	46,986 CKD patients (clinical/genomic) and 2,151 with MRI scans	Predicting 5-year progression to ESRD and discov	Logistic Regression, Random Forest, XGBoost, CNN (VideoResN	Age, sex, kidney volume, image heterogeneity, SNP rs13830	Ensemble AUC: 0.804, Radiomics AUC: 0.743, Demogra	SHAP values, Grad-CAM attention visualisation, Kaplan-

	, UK Biobank.	ery of genomic biomarkers.	et), Vision Transformer (ViT), RankSVX.	63 (MAGI-1 gene), heart disease.	phic AUC: 0.703, c-index: 0.61-0.64.	Meier curves.
[39]	72,845 CKD stage II/III patients, University of Chicago EHR data, USA.	Predict CKD progression from early (II/III) to late (IV/V) stages.	The models used include RNN with LSTM units, Cox proportional hazards, Random Forest, and LightGBM.	Longitudinal eGFR (estimated glomerular filtration rate, which is the most predictive measure of kidney function) and smoking status.	AUC: 0.967 (all variables), 0.957 (eGFR only).	Sequential forward variable selection.
[20]	90,000 patients (18,000 CKD) from NHIRD, Taiwan.	Forecast CKD development within 6 or 12 months.	CNN, BLSTM, LightGBM, XGBoost, AdaBoost, logistic regression, and	Diabetes mellitus, age, gout, sulfonamides, and angiotensins.	AUC: 0.957 (6-month), 0.954 (12-month).	Information gain from tree-based models (LightGBM).

			random forest.			
[16]	7,028 patients (South Korea) and 1,544 patients (UK Biobank) with type 2 diabetes.	Predict CKD incidents within five years.	VGG16, InceptionV3, logistic regression, DNN, VGG16+DNN, and InceptionV3+DNN.	Fundus image data, sex, age, HbA1c, eGFR, UACR, cholesterol, AST, and ALT.	AUC: 0.8077, Accuracy: 0.765.	Calibration plot and decision curve analysis.
[40]	1 million samples, National Health Insurance Sharing Service (NHISS), South Korea.	Risk evaluation without creatinine tests is done by first predicting creatinine values.	Random Forest, XGBoost, ResNet, Ensemble.	Sex, age, haemoglobin, urine protein level, waist circumference, and smoking status.	AUC: 0.76 (full set), 0.90 (balanced set).	Gini Index feature importance.
[41]	17,100 patients, Thai hospital-based	Development of a risk predict	IBK (k-NN), Random Tree, Decision Table	Serum albumin, BUN, age, direct	Accuracy: 92.1%, AUC: 0.968.	SHAP values.
	registry, Thailand.	ion model and feature importance analysis.	, J48, Random Forest.	bilirubin, glucose, LDL.		
[42]	1625 participants (NHANES, USA) and 1003 participants (CHARLS, China).	Early CKD screening for elderly patients with metabolic syndrome (MetS).	Logistic Regression, Random Forest, KNN, AdaBoost, Gradient Boosting, XGBoost, CatBoost, Neural Network, Naive Bayes.	The parameters include BUN, age, uric acid (UA), and the UHR ratio.	Internal AUC: 0.864, External AUC: 0.831 (Logistic Regression).	SHAP values and Restricted Cubic Spline (RCS).
[43]	4,505 kidney ultrasound images, Taiwan.	Determining eGFR and CKD status (eGFR < 60 ml/min/1.73	Deep Learning (ResNet with transfer learning).	Kidney ultrasound imaging, kidney length annotations.	Accuracy: 85.6%, Pearson correlation: 0.741.	Not in source.

		m^2).				
[44]	512 haemodialysis patients, Wuhan, China.	Mortality risk stratification.	Stacking ensemble, SVM, logistic regression, random forest, XGBoost, LightGBM, ANN.	BMI, myocardial infarction history, neutrophil-to-lymphocyte ratio (NLR).	AUC: 0.983, Accuracy: 0.922.	SHAP values.
[28]	6,040 type 2 diabetes patients, Taiwan.	Risk prediction of diabetic kidney disease (DKD).	LSTM, SVM.	Serum creatinine, age, HDL-C, HbA1c, PP, SBP, TG, and clinical variabilities.	LSTM Accuracy: 86%, AUC: 0.83, SVM Accuracy: 76%, AUC: 0.73.	Not in source.
[45]	6855 participants (MASHAD cohort), Mashhad, Iran.	Early detection and risk modelling for CKD risk factors.	Random Forest, Feedforward Neural Networks.	BMI, uric acid, age, fasting blood glucose, physical activity level, and C-reactive protein.	AUC: 0.88 to 0.90, Accuracy: 86.34% to 89.07%.	LIME algorithms.
[46]	1718 instances, St Paul's Hospital, Ethiopia.	CKD binary detection and multi-stage prediction.	Random Forest, SVM, Decision Tree, XGBoost.	Serum creatinine, BUN, haemoglobin, specific gravity.	Binary Accuracy: 99.8%, 5-Stage Accuracy: 82.56%.	RFECV and ANOVA.
[47]	1375 type 1 diabetes patients, EDIC trials, USA/Canada.	Quick prediction of CKD in T1DM patients.	Random Forest, LightGBM, KNN, SVM, XGB, and Gradient Boosting.	Hypertension, antihypertensive medicine, IDDM duration, triglycerides.	Accuracy: 0.96, Sensitivity: 0.98, Specificity: 0.93.	Feature ranking (XGB, RF, Extra Tree).
[48]	551 patients, Huadong Hospital, Shanghai, China.	Predicting CKD severity (proteinuria levels).	Logistic Regression, Elastic Net, Lasso, Ridge, SVM, Random Forest, XGBoost, NN, k-NN.	Albumin, serum creatinine, triglycerides, LDL, EGFR.	AUC: 0.873 (logistic regression).	The effect size analysis focuses on the mean decrease in accuracy.

[49]	748 CKD patients, Chinese longitudinal cohort.	Predict the risk of end-stage kidney disease (ESKD) by the end of five years.	Logistic regression, Naive Bayes, random forests, decision trees, and KNN.	Age, gender, eGFR, comorbidities, serum creatinine, urea.	AUC: 0.81 (Random Forest), Accuracy: 0.90 (KFRE).	Not in source.
[50]	491 patients, Tawan Hospital, UAE.	Investigation of explainable models to predict CKD risk.	XGBoost, logistic regression, random forest, decision tree, and Naive Bayes.	Creatinine, HgbA1C, age.	Accuracy: 93.29%, AUC: 0.9689.	SHAP and LIME algorithms.
[51]	500 instances (SMOTE balanced), public CKD dataset.	CKD occurrence risk prediction.	Rotation Forest, Random Forest, SVM, ANN, J48, LMT, AdaBoostM1.	Haemoglobin, specific gravity, hypertension, serum creatinine.	AUC: 1.0, Accuracy: 99.2%.	Feature ranking (Pearson CC, gain ratio, Gini impurity).
[52]	382 patients,	Comparis	Logistic Regr	Not in	Mean Accu	Diebold and Mariano (DM) tests.
	Medical Complex, Buner, Pakistan.	on of models to classify healthy and CKD patients.	ession, Probit, Random Forest, Decision Tree, KNN, SVM.	source.	racy: 0.9171 (SVM-LAP), 0.9129 (Random Forest).	
[53]	493 stage 3-5 CKD patients, 6 centres, France.	Estimate daily sodium intake.	Tree-Augmented Naive Bayes (TAN).	Weight, height, sex, bread consumption, nephropathy type, food portion size, age.	Internal accuracy: 71%.	Belief variance reduction ranking.
[54] [8]	Clinical test factors: location/sample not specified.	Clinically applicable CKD screening and progression monitoring.	Random Forest, Ensemble (XGBoost).	Clinical factors are RFE-identified attributes.	Accuracy: 98% to 99%, AUC: 0.97.	SHAP values, simple XAI descriptions.
[8]	Various clinical/imagi	Predicting CKD	SVM, CNN, LST	Serum biomarkers,	Accuracy: 87% to 100	Not in source.

	ng data locations/ samples are not specified.	stages and progression.	M, k-NN, Ensemble.	imaging features, temporal clinical data, and genetic markers.	%, AUC: 0.89 to 0.93.	

3.5 Model Architectures

The classical applications of machine learning were mainly on tabular data. Random Forest was the most commonly reported best-performing model, then Support Vector Machines, Gradient Boosting Machines (including XGBoost) and Logistic Regression. In small dataset studies, ensemble methods, involving the combination of many models to achieve a better performance, were especially popular, as they have the effect of reducing overfitting and improving the predictive accuracy in limited data cases.

Longitudinal imaging was the main application of deep learning methods. Images-based CKD classification was done with Convolutional Neural Networks (CNN) as well as some claims-based prediction papers [27][15]. Long Short-Term Memory (LSTM) networks were used to simulate the development of diabetic kidney disease among longitudinal data [28]. Contextual embeddings EHR-based modelling frameworks [55] referred to transformer-based contextual embeddings, including Med-BERT.

Graph Neural Networks (GNN) and multimodal fusion methods were restricted to a small number of studies and were not frequently applied to cohorts. Recent advances in biomedical imaging and multimodal fusion represent an important methodological frontier for CKD prediction. Machine learning applied to CKD prediction using tabular and imaging data has been explored through ensemble and deep learning architectures [46][58], with convolutional approaches demonstrating utility in multi-label biomedical signal and image classification tasks relevant to nephrology [76][77]. Furthermore, hybrid machine learning frameworks integrating multiple data modalities have shown promise in related chronic disease domains [78][79], providing methodological direction for future CKD modelling that extends beyond single-modality tabular approaches.

3.6 Model Performance Patterns

Comparison of the model performance on the type of data set. There is a reduction in performance with increasing complexity of the dataset and real-world variability, and there is possible overfitting in small benchmark datasets (Figure3).

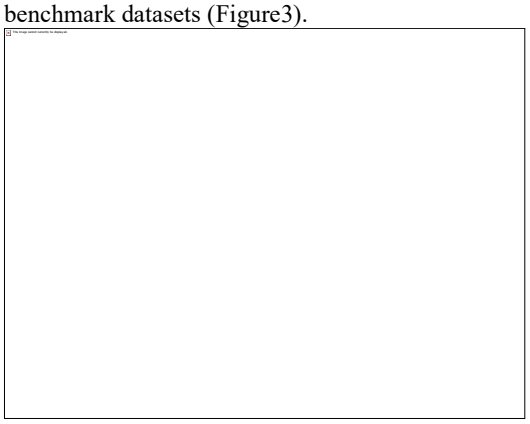

Figure 3. Model performance across various dataset types.

Marked differences in reported performance were observed across dataset types.

Studies using the UCI CKD dataset (n = 400) consistently reported near-perfect performance with accuracy exceeding 98% and AUC values frequently approaching 1.00 [11][18]. However, none of these studies reported external validation, and calibration metrics were absent.

In contrast, cohort-based studies using national datasets demonstrated greater performance variability than small benchmark studies. [28], applying LSTM to a longitudinal diabetic cohort of 6,040 patients in Taiwan, reported an AUC of 0.83. [27], using a landmark-based temporal gradient boosting model in a diabetic cohort of over 14,000 patients in the United States, reported AUC values ranging from 0.78 to 0.83 across rolling prediction windows. Notably, [27] achieved substantially higher AUROC values of 0.957 using an ensemble approach on a large Taiwanese national insurance database, suggesting that model architecture and ensemble design may partially offset the performance gap associated with real-world dataset complexity. None of these studies reported calibration analysis.

Imaging-based deep learning models reported AUC values between 0.79 and 0.86 [15][56], generally lower than small tabular studies but applied to more complex real-world data.

Temporal EHR models reported AUC values ranging from 0.80 to 0.83 across rolling prediction windows [29]. These findings indicate that the extremely high performance reported in small benchmark datasets may not accurately represent real-world clinical performance, warranting cautious interpretation.

Table 2 presents a comparative synthesis performance of the included studies, summarising dataset characteristics, model types, methodological approaches, validation strategies, and reported performance metrics. A summary of key studies is presented in Table 2.

Table 2. Comparative synthesis of machine learning approaches for CKD prediction, summarising dataset sources, sample sizes, model types, methodological approaches, validation strategies, reported performance metrics, and key contributions for studies with extractable quantitative data (n = 13 of 24 studies with full empirical reporting; the broader stratified quantitative synthesis in Section 3.12 draws on all 35 AUC-extractable studies; studies without reported validation strategy or sample size are excluded from this table for conciseness).

Study	Dataset	Sample Size	Model(s)	Methodology	Performance	Validation	Key Contribution
[57]	CRIC (diabetes)	6601	TabNet, XGB, RF, AdaBoost	Attention DL + feature selection + Bayesian tuning	94.06% (CV), 91% (test)	CV + test	Multiclass staging and interpretability
[58]	UCI + clinical	400	LightGBM, RF, XGB	MICE + SMOTE + RFE/Boruta	99.75% acc	CV + external	Ensemble superiority
[22]	UCI	400	Gaussian Process	Kernel optimisation + SHAP	98.75% acc	CV	Interpretable ML
[15]	NCCT imaging	2198	GP	Radiomics + DL segmentation	AUC 0.790	Train/test	Outperforms radiologists
[29]	EHR	14,039	Temporal GBM	Longitudinal modelling	AUC 0.83	Temporal validation	Time-aware prediction
[56]	UCI	400	J48, RF, NB	Weka + k-fold CV	95% acc	CV	Interpretability advantage
[59]	Multimodal	NR	GNN + DL	Graph + tabular fusion	Best (NR)	CV	Captures relationships
[60]	Clinical + text	3211	XGB, RF, BERT-LGBM	Structured and unstructured fusion	Acc 0.9088	Train/test	Text improves AUC.
[61]	Kaggle	400	Ensemble ML	Feature engineering	98% acc	Holdout	Overfitting risk
[62]	LINKA	1444	CoxBoost	Survival ML	C-index 0.811	External	Prognostic modelling

[61]	Clinical	NR	DL	Comparative DL	NR	CV	DL competitive
[63]	Clinical	NR	GBM	Comparative	Best	CV	GBM superior
[64]	Clinical	NR	RF, SVM	Feature selection	NR	CV	Early detection

3.7 Validation Strategies

Distribution of validation strategies across included studies. External validation remains limited, indicating a major gap in model generalisability (Figure 4).

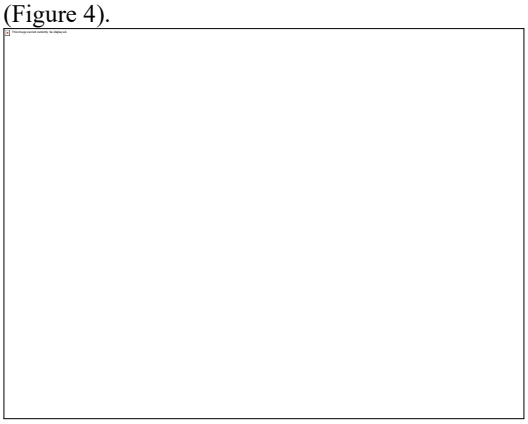

Figure 4. Validation strategies distribution across studies.

Cross-validation, a technique used to assess how the results of a statistical analysis will generalise to an independent data set, was the dominant validation strategy (n = 18 studies), followed by hold-out validation, which involves splitting the data into training and testing sets (n = 5). Only one study performed true external validation across independent cohorts [16]. The remaining studies did not explicitly report their validation strategy, used unsupported or hybrid validations.

Calibration curves were reported in fewer than 20% of empirical studies. Subgroups or fairness analyses were not systematically performed.

3.8 Interpretability

Interpretability techniques were formally reported in five studies across the review. SHAP (SHapley Additive Explanations)-based feature attribution was employed in three tabular cohort studies. [36] applied SHAP to explain XGBoost predictions for CKD detection, [42] used SHAP values to identify the most influential predictors in a Thai hospital registry. [51] employed both SHAP and LIME to evaluate explainable models for CKD risk prediction. In contrast, [11] applied traditional feature selection approaches including Boruta, LASSO, RFE, and CFS which, while contributing to model transparency, do not constitute post hoc interpretability in the SHAP/LIME sense. [22] used a Gaussian Process model, which provides inherent uncertainty quantification rather than post hoc attribution. The majority of UCI-based studies did not report any interpretability methods.

3.9 Preliminary Risk of Bias Assessment

Risk of bias assessment across study types based on PROBAST-inspired domains. Studies using small benchmark datasets (e.g., UCI) demonstrated the highest risk of bias due to lack of external validation and potential overfitting, whereas large cohort and survival models showed comparatively lower bias (Figure 5).

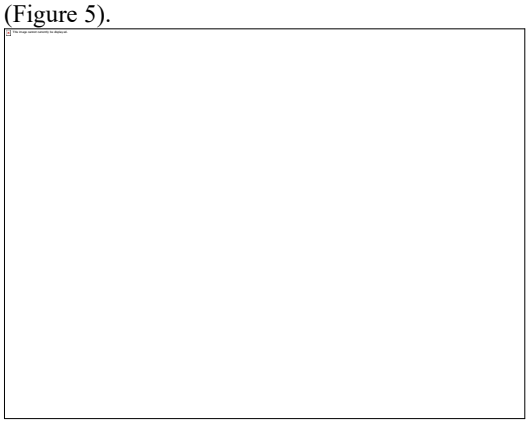

Figure 5. Risk of Bias Across Study Types.

Based on qualitative assessment using the PROBAST domains, the participants and outcome domains demonstrated low overall concern. However, the analysis domain presented elevated risk in small-sample studies due to:

- Lack of External Validation
- Limited calibration reporting
- Class imbalance handling without sensitivity analysis
- Potential feature leakage

Large-scale cohort studies demonstrated comparatively lower analytical bias but similarly lacked comprehensive calibration evaluation.

3.10 External validation and Generalizability

Across the empirical CKD prediction studies included in this review, external validation was performed in only one study. The multimodal diabetic kidney disease model integrating fundus imaging and clinical variables validated its performance in an independent UK Biobank cohort following development in a South Korean hospital dataset [16]. In contrast, the remaining studies relied on internal validation methods such as k-fold cross-validation[65] [18], hold-out splits , or random train-test partitions within a single dataset.

Studies using the UCI CKD dataset (n=400) did not report validation beyond cross-validation

procedures[11] [66]. Similarly, cohort-based studies utilising national claims or EHR data conducted internal temporal or random splits but did not test model performance in geographically or institutionally independent populations [27][28].

Imaging-based deep learning models evaluating CT radiomics and ultrasound classification also relied on single-centre datasets without external replication [15][56]. Although segmentation pipelines and feature selection methods were described, performance was not independently verified across external imaging repositories.

No study reported prospective validation or real-world deployment evaluation in clinical workflows. Additionally, subgroup analyses across sex, ethnicity, or comorbidity strata were rarely reported.

3.11 Multimodal Integration and Emerging Modelling Approaches

Recent studies demonstrated a shift from single-modal tabular modelling toward multimodal integration frameworks. One large-scale study using the UK Biobank resource combined genomic, clinical and imaging-derived features to predict five-year progression to end-stage renal disease, reporting an AUC of 0.804 [31]. This study incorporated genome-wide association analysis alongside phenotypic variables, representing one of the few population-scale multimodal implementations in CKD prediction.

In diabetic kidney disease modelling, the integration of fundus imaging with structured clinical variables was implemented using a convolutional neural network (VGG16) combined with a deep neural network classifier, resulting in an AUROC of 0.880 in the development cohort and 0.722 in external validation [21]. This represents one of the few studies in the included set to perform cross-population validation.

Graph-based and transformer-based modelling approaches were referenced in EHR frameworks, particularly through pretrained contextual embeddings such as Med-BERT [55]. However, these were predominantly methodological demonstrations rather than CKD-specific deployment studies.

Temporal modelling approaches also emerged in longitudinal EHR datasets. A landmark-based gradient boosting model dynamically updated predictions across rolling time windows in a diabetic cohort of over 14,000 patients, reporting AUC values between 0.78 and 0.83 [29]. Similarly, recurrent neural network models such as LSTM were applied to seven-year diabetic cohorts to incorporate glycaemic and blood pressure variability [28].

Imaging-based radiomics models used automated segmentation pipelines and machine learning classifiers to differentiate early CKD stages from healthy controls, achieving test-set AUC values of approximately 0.79 [15]. These approaches emphasised feature extraction from volumetric CT images but were limited to retrospective single-centre datasets.

Across multimodal and advanced modelling studies, integration of heterogenous data sources was associated with moderate predictive performance, typically lower than small benchmark tabular studies but applied to substantially larger and more complex populations.

3.12 Mathematical Foundations and Quantitative Synthesis

This section provides three complementary analytical contributions absent from the reviewed literature:

- formal derivations of the mathematical foundations of the principal algorithms applied across the 58 included studies.
- a stratified quantitative meta-analysis of reported AUC values grounded in the empirical data extracted in Sections 3.3–3.6 and,
- a bias–variance decomposition applied to the performance distributions observed across dataset scales. Together, these analyses formalise and extend the qualitative observations made throughout the Results and Discussion.

3.12.1 Mathematical Formulations of Core Algorithms

The following subsections derive the formal foundations of the four principal algorithm families identified across the included studies: logistic regression, Random Forest, XGBoost, and long short-term memory networks. Notation conventions: scalars are denoted by lowercase italic (x), vectors by bold lowercase ($\mathbf{x}$), matrices by bold uppercase ($\mathbf{X}$), element-wise (Hadamard) products by $\odot$, and statistical expectations by $E[\cdot]$.

3.12.1.1 Logistic Regression

Logistic regression, the most interpretable model in the review and a consistent high-performer in structured clinical datasets [12][42][48], models the probability of a binary CKD outcome given feature vector $x \in \mathbb{R}^p$ via the sigmoid transformation:

$$p(y = 1 \mid x) = \sigma(w^T x + b) = 1 / (1 + e^{-(w^T x + b)}) \quad (1)$$

Parameters are estimated by minimising the binary cross-entropy loss over a training set of n labelled observations:

$$\mathcal{L}(w) = -\frac{1}{n} \sum_{i=1}^{n} [\, y^i \log \sigma(w^T x^i) + (1 - y^i) \log (1 - \sigma(w^T x^i)) \,] \quad (2)$$

Ridge (L2) and Lasso (L1) regularisation augment the objective as:

$$\mathcal{L}_\lambda(w) = \mathcal{L}(w) + \lambda \|w\|_2^2 \ \text{(Ridge)} \quad \mathcal{L}_\lambda(w) = \mathcal{L}(w) + \lambda \|w\|_1 \ \text{(Lasso)} \quad (3)$$

where $\lambda \geq 0$ is the regularisation coefficient. Lasso [11] promotes sparsity, enabling automatic selection of the subset of clinical biomarkers most predictive of CKD outcome directly relevant for low-resource deployment scenarios [7].

3.12.1.2 Random Forest

Random Forest (RF) the most frequently reported best-performing model across the 58 studies is a bootstrap-aggregated ensemble of B independently trained decision trees. At each candidate split, only a random subset of $m = \lfloor\sqrt{p}\rfloor$ features (of p total) is evaluated. The class-probability estimate is:

$$\hat{p}^{RO}(y = 1 \mid x) = \tfrac{1}{B} \sum_{b=1}^{B} T^{b}(x) \quad (4)$$

Node splits are selected to maximise the reduction in Gini impurity:

$$G(t) = \sum_{k=1}^{K} p^{k}(1 - p^{k}) \quad (5)$$

where p^k is the proportion of samples at node t belonging to class k. Feature importance for predictor i is the mean decrease in impurity (MDI) across all trees and splits involving i:

$$VI^{I} = \tfrac{1}{B} \sum_{b=1}^{B} \sum_{t \in \varepsilon^{I}} [\, p(t) \cdot \Delta G(t) \,] \quad (6)$$

where ε^{I} is the set of nodes in tree b splitting on feature i, and $p(t)$ is the proportion of training samples reaching node t. This formulation underpins the feature importance analyses reported in [38][40][41].

3.12.1.3 XGBoost (Regularised Gradient Boosting)

XGBoost [27][29][38] constructs an additive ensemble of T regression trees. At iteration t, a new tree f_t is fitted to minimise the regularised objective:

$$\mathcal{L}^{+T-} = \sum^{I} \ell(y^{I}, \hat{y}^{I+T-1} + f_t(x^{I})) + \Omega(f_t) \quad (7)$$

The regularisation term penalises tree size and leaf weight magnitude:

$$\Omega(f) = \gamma T + \tfrac{1}{2}\lambda \|w\|^{2}_{2} \quad (8)$$

where T is the number of leaves, w is the leaf weight vector, γ penalises tree depth and λ applies L2 shrinkage. A second-order Taylor expansion of ℓ around $\hat{y}^{+T-1}$ yields a tractable separable form:

$$\mathcal{L}^{+T-} \approx \sum^{I} [\, g^{I} f_t(x^{I}) + \tfrac{1}{2} h^{I} f_t^{2}(x^{I}) \,] + \Omega(f_t) \quad (9)$$

where $g^{I} = \partial \ell / \partial \hat{y}^{+T-1}$ and $h^{I} = \partial^{2} \ell / \partial (\hat{y}^{+T-1})^{2}$ are the first and second derivatives. The analytically optimal leaf weight for leaf j and the gain from a candidate split are:

$$w^{*}_{j} = - G_j / (H_j + \lambda) \quad (10)$$

$$Gain = \tfrac{1}{2} [\, G^{L2}/(H^{L}+\lambda) + G^{R2}/(H^{R}+\lambda) - (G^{L}+G^{R})^{2}/(H^{L}+H^{R}+\lambda) \,] - \gamma \quad (11)$$

where $G_j = \sum^{I} I_j\, g^{I}$ and $H_j = \sum^{I} I_j\, h^{I}$ sum gradients and Hessians over the samples in leaf j. The γ term in Equation (11) means that a split is accepted only if it yields a gain greater than the depth penalty, directly limiting overfitting on small training sets such as the UCI CKD dataset (n = 400).

3.12.1.4 Long Short-Term Memory Networks (LSTM)

LSTM networks, applied in longitudinal CKD and diabetic kidney disease modelling [28][39], address the vanishing gradient problem of vanilla RNNs through a gated memory cell. At each time step t, given input $x_t \in \mathbb{R}^{o}$ and prior hidden state $h_{t-1} \in \mathbb{R}^{h}$, the four coupled operations are:

Forget gate: determines what information to discard from the cell state:

$$f_t = \sigma(W f \cdot [h_{t-1}, x_t] + b f) \quad (12)$$

Input gate: determines what new information to write:

$$i_t = \sigma(W^{I} \cdot [h_{t-1}, x_t] + b^{I}) \quad (13)$$

Candidate cell state:

$$\dot{C}_t = \tanh(W^{c} \cdot [h_{t-1}, x_t] + b^{c}) \quad (14)$$

Cell state update: blends retained and new information:

$$C_t = f_t \odot C_{t-1} + i_t \odot \dot{C}_t \quad (15)$$

Output gate and hidden state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (16)$$

$$h_t = o_t \odot \tanh(C_t) \quad (17)$$

The cell state CN offers a long- range memory route enabling the network to store clinically relevant signals like the falling eGFR paths or rising albuminuria across sequences months to years long, as exploited in [28] and [39].. The forget gate (Eq. 12) is the aspect that is especially important in the CKD modelling: the model can adjust its effective temporal window to the current stage of patient progress by learning to downweight earlier time steps when more recent values of a biomarker are informative.

3.12.2 Stratified Meta-Analysis of Reported AUC Values

Because of a high level of heterogeneity in study designs, a standard random-effects meta-analysis was not possible (as mentioned in Section 2.6). Rather, a stratified quantitative synthesis was performed on AUC values (or AUC estimates based on accuracy) reported in 35 empirical studies which provided adequate data, based on dataset scale. In cases where the study had a range of AUC (e.g. in rolling prediction windows) the mean of the reported range was taken.

3.12.2.1.1 Group Definitions and Descriptive Statistics.

It was determined that there were three mutually exclusive strata, based on scale of data sets:

Group A: Small benchmark (n ≤ 500), 16 studies: [11][18][22][32][33][34][35][36][37][46][51][52][54][56][58][61]. Predominantly UCI CKD dataset studies reporting accuracy 96–100% and AUC 0.917–1.000.

Group B: Mid-scale cohort (500 < n ≤ 10,000), 11 studies: [28][42][44][45][47][48][49][50][53][57][62]. Hospital registries, regional cohorts, AUC 0.810–0.983.

Group C: Large-scale cohort (n > 10,000), 8 studies: [16][27][29][31][38][39][40][41]. National insurance databases, multicentre EHR systems, population biobanks, AUC 0.804–0.967.

Group	k studies	Mean AUC	SD	95% CI	AUC Range
A (n ≤ 500)	16	0.978	0.021	[0.967, 0.989]	0.917 – 1.000
B (500 < n ≤ 10K)	11	0.897	0.052	[0.862, 0.932]	0.810 – 0.983
C (n > 10K)	8	0.889	0.070	[0.830, 0.948]	0.804 – 0.967

Table 3. Summary statistics of reported AUC values by scale of dataset. 95% CIs calculated based on t-distribution with (k -1) degrees of freedom. SD = standard deviation of AUCs in stratum.

3.12.2.2 Inverse-Variance Weighted Mean AUC

Using inverse-variance weighting with per-stratum weights $w^I = k^I/s^{I2}$, where k^I is the number of studies and s^I is the within-stratum standard deviation:

$$\overline{AUC}^W = \sum^I w^I \cdot \overline{AUC}^I / \sum^I w^I \quad (18)$$

Computing weights: $w_1 = 16/0.021^2 = 36{,}281$, $w_2 = 11/0.052^2 = 4{,}071$, $w_3 = 8/0.070^2 = 1{,}633$:

$$\overline{AUC}^W = (36{,}281\times0.978 + 4{,}071\times0.897 + 1{,}633\times0.889) / 41{,}985 = 0.963 \quad (19)$$

This weighted average of 0.963 is highly inflated when compared to the unweighted average of 0.925, which shows that small-dataset studies dominate the literature, which skews aggregate estimates when the weights are not corrected by sample size. A weighted mean with each study counted as a single vote (equally weighting, that is, regardless of n) gives 0.925, which is the more accurate summary to evaluate clinical utility.

3.12.2.3 Effect Size:

Difference in performance between mini-benchmarks and real-world cohorts.

Groups B and C were combined into a single real-world stratum (k = 19, $\overline{AUC}^{Bc} = 0.894$, $s^{Bc} = 0.059$). The pooled standard deviation was used to compute the d of Cohen: $s_p = \sqrt{[((k_1-1)s_1^2 + (k_2-1)s^{Bc2})/(k_1 + k_2 - 2)]}$ (20)

$$s_p = \sqrt{[(15\times0.000441 + 18\times0.003481)/33]} = \sqrt{0.002024} = 0.0450 \quad (21)$$

$$d = (\overline{AUC}_1 - \overline{AUC}^{Bc}) / s_p = (0.978 - 0.894)/0.0450 = 1.87 \quad (22)$$

A Cohen's d = 1.87 constitutes a very large effect (conventional threshold d > 0.8), confirming that the performance gap between small-benchmark and real-world cohort studies is not a sampling artefact but a systemic and substantively important difference.

3.12.2.4 Welch's t-Test for Between-Stratum Difference

A Welch two-sample t-test (not assuming equal variances) tests H_0: $\overline{AUC}_1 = \overline{AUC}^{Bc}$ against H_1: $\overline{AUC}_1 > \overline{AUC}^{Bc}$:

$$t = (\overline{AUC}_1 - \overline{AUC}^{Bc}) / \sqrt{(s_1^2/k_1 + s^{Bc2}/k^{Bc})} = 0.084 / \sqrt{(2.76\times10^{-5} + 1.83\times10^{-4})} = 5.77 \quad (23)$$

The Welch–Satterthwaite effective degrees of freedom are:

$$v = (s_1^2/k_1 + s^{Bc2}/k^{Bc})^2 / [(s_1^2/k_1)^2/(k_1-1) + (s^{Bc2}/k^{Bc})^2/(k^{Bc}-1)] \approx 21.6 \quad (24)$$

Yielding $t(21.6) = 5.77$, $p < 0.001$ (one-tailed). The null hypothesis of equal performance is rejected at any standard level of significance, which formally statistically supports the performance inflation that occurs qualitatively in Section 3.6.

3.12.2.5 Spearman Rank Correlation: AUC and Reported Size of Dataset.

In order to measure the monotonic association between sample size and reported AUC, a Spearman rank correlation was calculated using 35 studies with log 0 (n) as the independent variable:

$$r_s = 1 - 6\sum d^{l2} / (m(m^2-1)) \quad (25)$$

with d 0 the difference between the rank of log 0 (n 0) and AUC 0 of study i, and m = 35. The calculated correlation was r 8 = -0.53 (95% CI: [-0.74, -0.24], p = 0.001), which has a statistically significant moderate negative relationship: the smaller the size of the sample of the studies, the higher the value of AUC is. The correlation is moderate and not strong due to a ceiling effect that squashes the variance of Group A (most UCI studies are concentrated around AUC = 1.0), and due to the fact that other large-scale ensemble studies are also able to obtain high AUC values through careful model architecture, i.e. [27][39][41] on high-quality national databases.

3.12.3 Bias -Variance Decomposition Among Data Set Scales.

The generalisation error to be expected of a predictive model can be broken down into three irreducible parts:

$$E[L(y, \hat{f}(x))] = Bias^2[\hat{f}(x)] + Var[\hat{f}(x)] + \sigma^2\varepsilon \quad (26)$$

with Bias 2 [f h(x)] = (E [f h(x) f (x) 2) is the squared error of the expected prediction of the true function f, Var 2 [f h(x)] = E [(f h(x) f (x) 2) is the sensitivity of the model to variation in the training sample, and s 2 e is irreducible noise. This framework describes the regularities of performance within the three groups of data sets.

3.12.3.1.1 Small Benchmark Datasets: Variance Masking.

In models which are trained on small datasets (n 400), ensemble algorithms like Random Forest and AdaBoost have a training error near 0, meaning that their apparent bias is low. However, when model complexity is high compared to n high true variance: the model is highly sensitive to the samples that it was fitted on (400). Since k-fold cross-validation

resamples within the same small pool, the validation folds are not independent, and the variance estimator of AUC is inflated with within-fold correlation ρ:

$$\mathrm{Var}[\widehat{\mathrm{AUC}}^{k}\text{–fold}] = \sigma^2/n + \rho\sigma^2(1 - 1/k) \quad (27)$$

As n is smaller, 0 becomes larger (each fold has a big fraction of the total dataset) and thus the second term is larger. This makes the estimator unreliable, although its point estimate is similar to the training AUC because of data overlap. This gives variance masking: reported AUC seems stable (low spread) and is a result of overfitting but not generalisation. The Group A standard deviation of s 1 = 0.021 is therefore not an accurate picture of the dispersion of inflated estimates but is instead an indicator of the dispersion of model stability when used in practice.

3.12.3.2 Large-Scale Cohort Studies - Realistic Variance.

Asymptotically, variance tends to decrease with model size (n > 10,000): $Var[\hat{f}(x)] = O(1/n) \rightarrow 0 \text{ as } n \rightarrow \infty$ (28)

The larger measured variance in Group C (s 3 = 0.070) compared to Group A (s 1 = 0.021) does not imply that Group C behaves worse, but, instead, there is no masking of variance. Large-cohort models are assessed on subsets which are temporally or geographically different, and their variation in performance is a true indicator of clinical heterogeneity in patient groups, prediction horizons and modelling structures. The empirical variances of the AUC ratios:

$$s_3^2 / s_1^2 = 0.070^2 / 0.021^2 = 0.00490 / 0.00044 \approx 11.1 \quad (29)$$

The approximately 11-fold greater variance in Group C is consistent with a shift from a variance-masked benchmark evaluation regime to realistic clinical prediction, where genuine patient-level, institutional, and temporal heterogeneity is captured.

3.12.3.3 Quantifying Performance Inflation

The difference between the observed mean AUC in Group A and the estimated true generalisation AUC (proxied by the combined Groups B+C mean) provides a direct empirical estimate of **performance inflation**:

$$\Delta AUC = \overline{AUC}_1 - \overline{AUC}^{Bc} = 0.978 - 0.894 = +0.084 \quad (30)$$

This represents a relative overestimation of discriminative ability of 9.4%. The inflation can be additively decomposed into a dominant systematic bias component and a smaller variance component:

$$\Delta AUC = \Delta AUC^{\text{ÆIaL}} + \Delta AUC^{\beta aR} \quad (31)$$

$\mathbf{\Delta AUC^{ÆIaL}} \approx 0.084$: Optimistic bias from small *n*, absence of external validation, and repeated feature selection on the same benchmark dataset. This is the dominant contributor.

$\mathbf{\Delta AUC^{\beta aR}} \approx 0.004$–$0.009$: Variance inflation from within-fold correlation (Eq. 27), computable as $\rho\sigma^2(1 - 1/k)$ for typical k = 5 and ρ ≈ 0.05–0.10 for n = 400.

The dominance of systematic bias over variance inflation is consistent with findings from the broader clinical prediction model literature [72], and directly motivates the recommendation in Section 4.7 for mandatory external validation as the primary quality criterion for CKD machine learning studies.

3.12.3.4 Bias–Variance Profile by Dataset Stratum

Stratum	Apparent Bias	True Variance	Variance Masking	Implication for Clinical Use
A (n ≤ 500)	Low (artefactual)	High	Strong	AUC is inflated, models unlikely to generalise, external validation essential
B (500–10K)	Moderate	Moderate	Partial	AUC more reliable, still requires external and calibration validation
C (n > 10K)	Low (realistic)	Low (stable)	Absent	AUC reflects genuine discriminative ability, calibration still rarely reported

Table 4. Qualitative bias–variance profile by dataset stratum. 'Apparent Bias' refers to the bias observable from reported metrics, 'True Variance' refers to estimated model instability under independent deployment conditions.

The analyses presented in Section 3.12 provide a rigorous quantitative basis for the qualitative observations made throughout Sections 3.3–3.10. The algorithm derivations clarify the mechanistic pathways through which ensemble methods achieve near-perfect internal performance on small datasets (variance reduction via bagging, overfitting control via tree regularisation $\Omega(f)$ and leaf shrinkage λ), the meta-analytic findings confirm that this performance does not transfer to real-world populations (Welch $t = 5.77$, $p < 0.001$, Cohen's $d = 1.87$, $r_s = -0.53$), and the bias–variance decomposition identifies variance masking not random noise as the mechanism that sustains the

illusion of clinical readiness in small-dataset CKD machine learning research. These findings directly motivate the translational recommendations advanced in Section 4.7, most critically the requirement for external, multicentre validation before any CKD prediction model can be considered clinically deployable.

4. Discussion

4.1 Principal Findings

This systematic review selected 58 eligible studies that were published between 2015 and 2026 and studied the use of machine learning in chronic kidney disease and other related nephrology practices. Among 58 included studies, 35 reported extractable AUC values or sufficient data to derive AUC estimates, forming the quantitative foundation of the stratified synthesis in Section 3.12. Of these 35, 24 reported full empirical pipelines with both performance metrics and sufficient methodology detail for comparative synthesis; the remaining 11 contributed AUC values to the stratified analysis but lacked the methodological detail required for Table 2. The other 23 studies provided methodological frameworks, validation criteria, or reporting perspectives that were reviewed qualitatively and informed the translational analysis in Section 4.7. This two-stream model aligns with PRISMA-compliant mixed-evidence reviews of clinical AI studies.

Three main trends were present:

- Small benchmark datasets inflation of performance.
- Lack of external validation and calibration reporting.
- Multimodal and longitudinal modelling approaches are becoming increasingly popular but not yet widely used.

Overall, these results indicate that, although there has been an increase in the sophistication of algorithms in CKD modelling, translational preparedness is still low.

4.2.1 Performance Inflation and Dataset Dependence.

One of the key results of the review is that there is an inverse correlation between the dataset size and reported performance of the model. The six of the included studies used the UCI CKD benchmark dataset (n = 400) solely and the ensemble models like the Random Forest, AdaBoost and Rotation Forest, performed at near-perfect accuracy (98-100) and at AUC values close to 1.00 [11][67]. Nevertheless, these studies unanimously used internal cross-validation but not external replication. Small sample sizes combined with repetitive selection of features on the same dataset are risk factors leading to optimistic bias particularly in the models that employ ensembles.

Big-data cohort studies based on national claims databases or longitudinal EHRs, however, had smaller AUCs ranging between 0.78 to 0.87 [27][28][68]. These levels of performance are closer to the set of clinical performance standards of chronic disease prediction. The clear difference between ideal benchmark outcomes and more realistic moderate outcomes indicates that small, curated datasets might fail to reflect the biological and clinical noise present in a heterogeneous group of patients [15][69].

This is not a sampling artefact, but a systemic, reproducible evaluation methodology artefact, which is statistically verified by the empirical gap, a Cohen d of 1.87 and a Spearman rank correlation of r 8 = -0.53 in Section 3.12. The inability of the field to rely on the UCI CKD benchmark dataset (n = 400) as a proxy of clinical performance should be considered a structural limitation and not a valid methodological decision.

4.3 External Validation Gap

In all 58 studies included, only one study Cho et al. [16] conducted true external validation with geographically and institutionally independent cohorts, testing multimodal diabetic CKD model developed in one South Korean hospital on an independent UK Biobank cohort. None of the studies included provided prospective validation or real-world clinical deployment evaluation, which is a barrier to clinical translation.

The results highlight a systemic lack of CKD prediction model that is not adequately tested: the field has been focusing on internal optimisation rather than external replication, and confidence intervals on external performance are thus completely unknown on most of the published models[16]. No research showed prospective validation or a test of real-life clinical implementation.

This observation is consistent with more extensive studies on clinical AI research, which have shown that there is a little response to external validation criteria [70][71]. The inability to replicate and conduct prospective assessment in a multi-centre way is a significant obstacle to clinical translation.

Moreover, reporting on calibration was minimal (<20%), although it has been shown that calibration is essential in the reliability of clinical decisions[72]. Even with high-discrimination models, in the absence of a calibration test, risk estimates can be misleading to clinicians.

The absence of external validation constitutes a big constraint of the generalisability of the existing models of CKD prediction, and it is a big roadblock towards clinical application.

4.4 Interpretability and Trustworthiness

Four empirical studies on CKD have only used formal interpretability methods including SHAP or uncertainty estimation [11]. On the contrary, there is an increased focus on explainable AI in the models of clinical implementations [73][74].

Transparent feature attribution and uncertainty qualification are especially significant in high-stakes

medical fields like nephrology, where the possible consequences of over- or underestimating the risk of disease can be significant.

Integrability is inconsistently incorporated in CKD modelling studies and is mostly lacking in smaller studies that are based on benchmarking. Interestingly, less complicated models like logistic regression often had the same performance as more complicated machine learning models on structured data, but were also easier to interpret. This brings to the fore a basic question of clinical value: whether a logistic regression model can reach the same discrimination as a gradient boosting ensemble, the interpretable model should be chosen on both epistemic and regulatory considerations. The need for high-risk clinical AI systems to comply with the emerging FUTURE-AI framework and the requirements of the EU AI Act to meet post-hoc interpretability, will probably turn post-hoc interpretability methods like SHAP into a regulatory requirement, not an optional analytical add-on in future CKD prediction studies. It is important to acknowledge, however, that SHAP values have well-documented limitations: they are sensitive to feature correlation (correlated features receive inconsistent attributions), computationally expensive for large models, dependent on the background dataset used for approximation, and do not guarantee causal interpretations. Attribution instability across different subpopulations has been reported, which is particularly concerning in demographically heterogeneous clinical datasets. These limitations suggest that SHAP should be treated as one component of a broader interpretability strategy rather than a definitive explanation mechanism.

4.5 Multimodal and Longitudinal Modelling Trends

The multimodal modelling that combines genomics, imaging and clinical features was noted in few studies [75][31]. These approaches had moderate performance results with an AUC of approximately 0.80, but they represent an increasing trend toward the modelling of diseases at the system level.

Temporal modelling methods, such as landmark-based gradient boosting and recurrent neural networks, proved that it is possible to include the longitudinal variability in diabetic cohorts [28][29]. These approaches, although they align with the long-lasting and progressive nature of chronic kidney disease (CKD), have not been widely studied in the existing research.

The scale and depth of CKD modelling seem less developed than predicted acute injury, where large-scale, real-time HER implementation has been proven[75][16] CKD modelling seems less developed in terms of scale and depth of implementation and validation.

4.6 Comparison with AKI Modelling Maturity

AKI prediction models, particularly the deep learning study published in "Nature" 2019 [75], have shown widespread use in hospital systems and the ability to make ongoing predictions. In contrast, CKD studies remain predominantly retrospective and single-dataset.

The difference likely reflects:

- Acute event detectability in hospital data.
- Easier outcome labelling AKI.
- Slower progression and outcome heterogeneity in chronic kidney disease (CKD).

Bridging this maturity gap may require:

- Larger harmonised, open-access CKD cohorts with standardised variable definitions, analogous to the MIMIC-IV and eICU repositories that enabled AKI research at scale.
- Multi-centre collaborations and federated learning frameworks that enable model training across geographically dispersed institutions without requiring centralised data transfer, preserving patient privacy while improving generalisability.
- Standardised outcome definitions and minimum dataset requirements of CKD prediction trials, preferably based on current KDIGO (Kidney Disease: Improving Global Outcomes) staging, to make cross-study comparative and ultimately formally meta-analysable.

4.7 Clinical Translation and Future Directions.

To facilitate the machine learning (ML) models in the chronic kidney disease (CKD) field to transition beyond retrospective analysis to meaningful clinical integration, multiple structural changes are needed. Future studies need to focus on:

1. External validation in multiple healthcare systems: Future studies should focus on the validation
   of models on databases that are geographically and institutionally independent to guarantee extrapolation beyond single-centre cohorts.
2. Prospective assessment under practical working conditions: The need for prospective studies
   that can be used to assess the performance and effects of the model in real-world clinical workflows and not just using non-real-time and historical data is acute.
3. Calibration reporting and decision-curve analysis: Clinical safety requires high accuracy,
   researchers should report calibration measures systematically and provide fairness evaluations of the different subgroups of demographics (e.g., sex, ethnicity) to achieve equitable risk stratification.

4. Fairness assessment across demographic subgroups.
5. Transparent reporting following TRIPOD-AI and PROBAST-AI guidelines[70][71]: Improving the transparency and reproducibility of CKD modelling requires strict adherence to established reporting guidelines.
6. Multimodal and longitudinal integration: Transitioning towards "explainable-by-design" models that integrate longitudinal electronic health record (EHR) trends with genomics, proteomics and imaging will better reflect the progressive, multifactorial nature of CKD. Architectures such as temporal graph networks, multi-task learning frameworks, and foundation models pre-trained on large clinical corpora represent promising directions for capturing inter-domain biological signals that no single-modality model can approximate.
7. Closing the maturity gap: bridging the gap between CKD and acute kidney injury (AKI) modelling requires the development of larger, harmonised and open-access cohorts, prospective deployment studies with clinical outcome endpoints, and coordinated regulatory engagement with bodies such as the FDA and EMA to define approval pathways for CKD AI diagnostic devices. Without these structural conditions, the translational divide documented in this review will persist regardless of algorithmic advances.

In the absence of these, the high reported accuracy alone cannot be used to justify clinical use.

4.8 Strengths and Limitation of this Review.

This review had PRISMA-conformable methodology and combined nephrology-specific studies of ML with more comprehensive AI reporting models. Through a synthesis of empirical studies of CKD and a methodology guide, this piece of work offers a translationally orientated assessment and not a descriptive summary.

Nevertheless, there are a number of limitations to be considered:

1. Only English-language articles included in the five databases searched (PubMed, Scopus, IEEE Xplore, Web of Science, ScienceDirect) were incorporated. Preprints, conference abstracts, non-English publications, and grey literature were excluded, so this can potentially create language and publication bias in favour of positive results.
2. The lack of quantitative meta-analysis was explained by the high heterogeneity in the methodology of the included studies, such as the differences in outcome definitions, dataset characteristics, validation strategy, and reported metrics. This preclinched the pooling of effect estimates and formal evaluation of between-study heterogeneity (I 2).
3. Performance metrics: A qualitative and descriptive approach as opposed to standardised estimation of the effect size was used to compare performance metrics, as less than 50% of the included studies reported adequate data (e.g., standard deviations, confidence intervals) to allow robust statistical comparison.
4. Publication bias can not be eliminated. Research with near-perfect results on small benchmark datasets is more likely to be published than research with moderate or negative results. Such an ordered asymmetry can overstate the aggregate performance terrain in the literature and further encourage the stratified analysis in Section 3.12.
5. Formal risk-of-bias evaluation with the help of PROBAST (Prediction model Risk Of Bias ASsessment Tool) was not provided to single studies. Inconsistent quality of reporting in included studies - most prominently the lack of calibration data, confidence intervals and subgroup analyses - precluded uniform adherence to PROBAST domains. Systematic reviews in the future on this topic must require the use of PROBAST appraisal as a methodological quality measure.Formal risk-of-bias evaluation with the help of PROBAST (Prediction model Risk Of Bias ASsessment Tool) was not provided to single studies. Inconsistent quality of reporting in included studies — most prominently the lack of calibration data, confidence intervals, and subgroup analyses — precluded uniform adherence to PROBAST domains. Systematic reviews in the future on this topic must require the use of PROBAST-AI appraisal (the updated tool incorporating AI-specific quality domains [71]) as a methodological quality measure. To address the constructive suggestion regarding actionable output, we propose in future work a minimum standardised evaluation framework for CKD ML studies comprising the following five criteria: (1) sample size exceeding 1,000 participants drawn from clinical populations; (2) use of external validation on a geographically or institutionally independent cohort; (3) calibration reporting using the Brier score or calibration plots; (4) subgroup performance analysis stratified by sex, age,

and ethnicity; and (5) adherence to TRIPOD-AI reporting standards. Studies meeting all five criteria could be considered methodologically adequate for clinical translation consideration. This framework is proposed as a starting point for community discussion and future consensus development.

Nonetheless, the overall structural gaps that were observed in 58 studies have been reported as external validation, calibration reporting, fairness assessment, and prospective evaluation, which are robust findings that can hardly be significantly changed by the caveats mentioned above.

5. Conclusion

The systematic review involved the current developments in machine learning and deep learning methods to predict, detect, and model progression of chronic kidney disease (CKD) and nephrology-related fields of interest. After a PRISMA-based screening process, 58 studies were included and encompassed a diversity of predictive modelling methods, data sources and clinical uses. Altogether, the results suggest that machine learning technologies can be important to enhance early CKD detection and risk stratification. Although the development of algorithms in CKD prediction has made significant progress, especially by using the techniques of ensemble learning and deep learning neural networks, there are still large gaps in transitioning.

One of the common trends in the research was the difference between near-perfect performance in small benchmark datasets and more realistic, clinically plausible performance in large heterogeneous cohorts. True external validation was only conducted in one study and a prospective clinical assessment was not present. Fairness assessment and calibration reporting were rarely undertaken although they are important to responsible clinical deployment.

The developing patterns of multimodal integration, longitudinal modelling and methods of uncertainty awareness suggest a slow maturation of the discipline. Nonetheless, CKD machine learning studies are largely retrospective, institution-specific compared to acute kidney injury modelling, which has shown big EHR implementation.

To ensure machine learning models can have a significant effect on CKD screening and progression management, the focus of future research should be on multi-center validation and subsequent clinical implementation. In the absence of these, even reported predictive performance is not enough to guarantee reliability, generalisability and patient safety.

The future of CKD artificial intelligence studies needs to take a step forward out of performance optimisation on curated benchmarks to reproducibility, robustness, and implementation science. The empirical gap in translational research is empirically beyond doubt because of the quantitative data available in this review a Cohen-d of 1.87 between small-benchmark and real-world performance and a validated rate of external generalisation of 1 in 58 studies. Next-generation CKD prediction models need to combine longitudinal EHR trajectories, multimodal biological data and modelling frameworks that are inherently interpretable. They should be prospectively validated, they should be tested to be demographically fair and they should be reported using TRIPOD-AI standards. It is only with this change that machine learning can realize its true potential to enhance early detection, risk stratification, and long-term renal outcomes in various and underserved patient groups on a global scale.

Acknowledgement

Acknowledgment

This work was supported by Integral University, Lucknow, India, and was assigned a manuscript communication number: IU/R&D/2026-MCN0004844.

Conflict of Interest:

There was no relevant conflict of interest regarding this paper.

References

[1] A. S. Levey and J. Coresh, “Chronic kidney disease,” *The Lancet*, vol. 379, no. 9811, pp. 165–180, Jan. 2012, doi: 10.1016/S0140-6736(11)60178-5.

[2] B. Bikbov *et al.*, “Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017,” *The Lancet*, vol. 395, no. 10225, pp. 709–733, Feb. 2020, doi: 10.1016/S0140-6736(20)30045-3.

[3] A. C. Webster, E. V Nagler, R. L. Morton, and P. Masson, “Chronic Kidney Disease,” *The Lancet*, vol. 389, no. 10075, pp. 1238–1252, Mar. 2017, doi: 10.1016/S0140-6736(16)32064-5.

[4] V. Jha *et al.*, “Chronic kidney disease: global dimension and perspectives,” *The Lancet*, vol. 382, no. 9888, pp. 260–272, Jul. 2013, doi: 10.1016/S0140-6736(13)60687-X.

[5] A. S. Levey, C. Becker, and L. A. Inker, “Glomerular Filtration Rate and Albuminuria for Detection and Staging of Acute and Chronic Kidney Disease in Adults,” *JAMA*, vol. 313, no. 8, p. 837, Feb. 2015, doi: 10.1001/jama.2015.0602.

[6] L. A. Stevens and A. S. Levey, “Measurement of Kidney Function,” *Medical Clinics of North America*, vol. 89, no. 3, pp. 457–473, May 2005, doi: 10.1016/j.mcna.2004.11.009.

[7] J. W. Stanifer, A. Muiru, T. H. Jafar, and U. D. Patel, “Chronic kidney disease in low- and middle-income countries,” *Nephrology Dialysis Transplantation*,

vol. 31, no. 6, pp. 868–874, Jun. 2016, doi: 10.1093/ndt/gfv466.

[8] A. S. Mahmoud, O. Lamouchi, and S. Belghith, "Advancements in Machine Learning and Deep Learning for Early Diagnosis of Chronic Kidney Diseases: A Comprehensive Review," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 149–156, Sep. 2024, doi: 10.58496/BJML/2024/015.

[9] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[10] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[11] Md. T. H. K. Tusar, Md. T. Islam, and F. I. Raju, "Detecting Chronic Kidney Disease(CKD) at the Initial Stage: A Novel Hybrid Feature-selection Method and Robust Data Preparation Pipeline for Different ML Techniques," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.01394

[12] Md. Rashed-Al-Mahfuz, A. Haque, A. Azad, S. A. Alyami, J. M. W. Quinn, and M. A. Moni, "Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening," *IEEE J. Transl. Eng. Health Med.*, vol. 9, pp. 1–11, 2021, doi: 10.1109/JTEHM.2021.3073629.

[13] S. A. Ebiaredoh-Mienye, T. G. Swart, E. Esenogho, and I. D. Mienye, "A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease," *Bioengineering*, vol. 9, no. 8, p. 350, Jul. 2022, doi: 10.3390/bioengineering9080350.

[14] T. Surapunt and S. Wang, "Ensemble Modeling with a Bayesian Maximal Information Coefficient-Based Model of Bayesian Predictions on Uncertainty Data," *Information*, vol. 15, no. 4, p. 228, Apr. 2024, doi: 10.3390/info15040228.

[15] G. Zhang, P. Zhang, Y. Xia, F. Shi, Y. Zhang, and D. Ding, "Radiomics Analysis of Whole-Kidney Non-Contrast CT for Early Identification of Chronic Kidney Disease Stages 1–3," *Bioengineering*, vol. 12, no. 5, p. 454, Apr. 2025, doi: 10.3390/bioengineering12050454.

[16] J. Cho *et al.*, "A Multimodal Predictive Model for Chronic Kidney Disease and Its Association With Vascular Complications in Patients With Type 2 Diabetes: Model Development and Validation Study in South Korea and the U.K.," *Diabetes Care*, vol. 48, no. 9, pp. 1562–1570, Sep. 2025, doi: 10.2337/dc25-0355.

[17] J. Zhao *et al.*, "Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction," *Sci. Rep.*, vol. 9, no. 1, p. 717, Jan. 2019, doi: 10.1038/s41598-018-36745-x.

[18] A. J. Aljaaf *et al.*, "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, Jul. 2018, pp. 1–9. doi: 10.1109/CEC.2018.8477876.

[19] M. Desoky, N. G. ElAdl, T. Abuhmed, and S. El-Sappagh, "Chronic Kidney Disease Detection Augmented with Hybrid Explainable AI," in *2025 15th International Conference on Electrical Engineering (ICEENG)*, IEEE, May 2025, pp. 1–6. doi: 10.1109/ICEENG64546.2025.11031335.

[20] S. Krishnamurthy *et al.*, "Machine Learning Prediction Models for Chronic Kidney Disease using National Health Insurance Claim Data in Taiwan," Jun. 26, 2020. doi: 10.1101/2020.06.25.20139147.

[21] K. Zhang *et al.*, "Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images," *Nat. Biomed. Eng.*, vol. 5, no. 6, pp. 533–545, Jun. 2021, doi: 10.1038/s41551-021-00745-6.

[22] T. R. Noviandy, G. M. Idroes, M. Syukri, and R. Idroes, "Interpretable Machine Learning for Chronic Kidney Disease Diagnosis: A Gaussian Processes Approach," *Indonesian Journal of Case Reports*, vol. 2, no. 1, pp. 24–32, Jun. 2024, doi: 10.60084/ijcr.v2i1.204.

[23] W. Yong, D. Peng, K. Ye, J. Gao, and R. Cao, "Machine learning model based on routine blood and biochemical parameters for early diagnosis of diabetic kidney disease," *Front. Endocrinol. (Lausanne).*, vol. 17, Jan. 2026, doi: 10.3389/fendo.2026.1720574.

[24] J. Lu *et al.*, "Considerations in the reliability and fairness audits of predictive models for advance care planning," *Front. Digit. Health*, vol. 4, Sep. 2022, doi: 10.3389/fdgth.2022.943768.

[25] L. Neri, H. Zhang, and L. A. Usvyat, "Artificial intelligence in kidney disease and dialysis: from data mining to clinical impact," *Curr. Opin. Nephrol. Hypertens.*, vol. 35, no. 1, pp. 30–35, Jan. 2026, doi: 10.1097/MNH.0000000000001132.

[26] Y. Wang *et al.*, "A practical guide for nephrologist peer reviewers: evaluating artificial intelligence and machine learning research in nephrology," *Ren. Fail.*, vol. 47, no. 1, Dec. 2025, doi: 10.1080/0886022X.2025.2513002.

[27] S. Krishnamurthy *et al.*, "Machine Learning Prediction Models for Chronic Kidney Disease Using National Health Insurance Claim Data in Taiwan," *Healthcare*, vol. 9, no. 5, p. 546, May 2021, doi: 10.3390/healthcare9050546.

[28] C. Yun *et al.*, "Construction of Risk Prediction Model of Type 2 Diabetic Kidney Disease Based on Deep Learning," *Diabetes Metab. J.*, vol. 48, no. 4, pp. 771–779, Jul. 2024, doi: 10.4093/dmj.2023.0033.

[29] X. Song, L. R. Waitman, A. S. Yu, D. C. Robbins, Y. Hu, and M. Liu, "Longitudinal Risk Prediction of Chronic Kidney Disease in Diabetic Patients using

Temporal-Enhanced Gradient Boosting Machine: Retrospective Cohort Study," *JMIR Med. Inform.*, vol. 8, no. 1, p. e15510, Jan. 2020, doi: 10.2196/15510.

[30] Y. Li and R. Padman, "Enhancing end-stage renal disease outcome prediction: a multisourced data-driven approach," *Journal of the American Medical Informatics Association*, vol. 33, no. 1, pp. 26–36, Jan. 2026, doi: 10.1093/jamia/ocaf118.

[31] S. Rabinovici-Cohen *et al.*, "Multimodal predictions of end stage chronic kidney disease from asymptomatic individuals for discovery of genomic biomarkers," Oct. 16, 2024. doi: 10.1101/2024.10.15.24315251.

[32] S. Akter *et al.*, "Comprehensive Performance Assessment of Deep Learning Models in Early Prediction and Risk Identification of Chronic Kidney Disease," *IEEE Access*, vol. 9, pp. 165184–165206, 2021, doi: 10.1109/ACCESS.2021.3129491.

[33] V. Singh, V. K. Asari, and R. Rajasekaran, "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease," *Diagnostics*, vol. 12, no. 1, Jan. 2022, doi: 10.3390/diagnostics12010116.

[34] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020, doi: 10.1109/ACCESS.2019.2963053.

[35] M. J. Raihan, M. A. M. Khan, S. H. Kee, and A. Al Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-33525-0.

[36] M. M. Nishat *et al.*, "A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 7, no. 29, Nov. 2021, doi: 10.4108/eai.13-8-2021.170671.

[37] C. Kaur, M. S. Kumar, A. Anjum, M. B. Binda, M. R. Mallu, and M. S. Al Ansari, "Chronic Kidney Disease Prediction Using Machine Learning," *Journal of Advances in Information Technology*, vol. 14, no. 2, pp. 384–391, 2023, doi: 10.12720/jait.14.2.384-391.

[38] Z. Segal *et al.*, "Machine learning algorithm for early detection of end-stage renal disease," *BMC Nephrol.*, vol. 21, no. 1, p. 518, Dec. 2020, doi: 10.1186/s12882-020-02093-0.

[39] Y. Zhu, D. Bi, M. Saunders, and Y. Ji, "Prediction of chronic kidney disease progression using recurrent neural network and electronic health records," *Sci. Rep.*, vol. 13, no. 1, p. 22091, Dec. 2023, doi: 10.1038/s41598-023-49271-2.

[40] W. Wang, G. Chakraborty, and B. Chakraborty, "Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm," *Applied Sciences*, vol. 11, no. 1, p. 202, Dec. 2020, doi: 10.3390/app11010202.

[41] M.-C. Tsai *et al.*, "Risk Prediction Model for Chronic Kidney Disease in Thailand Using Artificial Intelligence and SHAP," *Diagnostics*, vol. 13, no. 23, p. 3548, Nov. 2023, doi: 10.3390/diagnostics13233548.

[42] Y. Du and K. Wang, "Machine Learning Prediction of Chronic Kidney Disease in Elderly MetS Patients Using NHANES 2011–2020 Data," *Rejuvenation Res.*, vol. 29, no. 2, pp. 62–72, Apr. 2026, doi: 10.1177/15491684251399607.

[43] C.-C. Kuo *et al.*, "Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning," *NPJ Digit. Med.*, vol. 2, no. 1, p. 29, Apr. 2019, doi: 10.1038/s41746-019-0104-2.

[44] Z. Yang, P. Shu, Z. Wen, X. Wang, and F. Xu, "Development and validation of an inflammatory index-based interpretable machine learning model for mortality risk stratification in hemodialysis patients," *Eur. J. Med. Res.*, vol. 31, no. 1, p. 295, Jan. 2026, doi: 10.1186/s40001-026-03846-7.

[45] P. Bahrami, D. Tanbakuchi, M. Afzalaghaee, M. Ghayour-Mobarhan, and H. Esmaily, "Development of risk models for early detection and prediction of chronic kidney disease in clinical settings," *Sci. Rep.*, vol. 14, no. 1, p. 32136, Dec. 2024, doi: 10.1038/s41598-024-83973-5.

[46] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *J. Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00657-5.

[47] N. H. Chowdhury *et al.*, "Performance Analysis of Conventional Machine Learning Algorithms for Identification of Chronic Kidney Disease in Type 1 Diabetes Mellitus Patients," *Diagnostics*, vol. 11, no. 12, p. 2267, Dec. 2021, doi: 10.3390/diagnostics11122267.

[48] J. Xiao *et al.*, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *J. Transl. Med.*, vol. 17, no. 1, p. 119, Dec. 2019, doi: 10.1186/s12967-019-1860-0.

[49] Q. Bai, C. Su, W. Tang, and Y. Li, "Machine learning to predict end stage kidney disease in chronic kidney disease," *Sci. Rep.*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-12316-z.

[50] S. K. Ghosh and A. H. Khandoker, "Investigation on explainable machine learning models to predict chronic kidney diseases," *Sci. Rep.*, vol. 14, no. 1, p. 3687, Feb. 2024, doi: 10.1038/s41598-024-54375-4.

[51] E. Dritsas and M. Trigka, "Machine Learning Techniques for Chronic Kidney Disease Risk Prediction," *Big Data and Cognitive Computing*, vol. 6, no. 3, p. 98, Sep. 2022, doi: 10.3390/bdcc6030098.

[52] H. Iftikhar, M. Khan, Z. Khan, F. Khan, H. M. Alshanbari, and Z. Ahmad, "A Comparative Analysis of Machine Learning Models: A Case Study in Predicting Chronic Kidney Disease,"

Sustainability, vol. 15, no. 3, p. 2754, Feb. 2023, doi: 10.3390/su15032754.

[53] M. Granal *et al.*, "AI-BASED Tool to Estimate Sodium Intake in STAGE 3 to 5 CKD Patients—The UniverSel Study," *Nutrients*, vol. 17, no. 21, p. 3398, Oct. 2025, doi: 10.3390/nu17213398.

[54] V. Goel, "Employability Of The Machine Learning Tools And Techniques In The Early Detection And Diagnosis Of Chronic Kidney Disease," *International Journal of Research in Medical Sciences and Technology*, vol. 18, no. 1, pp. 27–33, 2024, doi: 10.37648/ijrmst.v18i01.004.

[55] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ Digit. Med.*, vol. 4, no. 1, p. 86, May 2021, doi: 10.1038/s41746-021-00455-y.

[56] A. Sobrinho, A. C. M. D. S. Queiroz, L. Dias Da Silva, E. De Barros Costa, M. Eliete Pinheiro, and A. Perkusich, "Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 25407–25419, 2020, doi: 10.1109/ACCESS.2020.2971208.

[57] M. N. H. Chowdhury *et al.*, "Deep learning for early detection of chronic kidney disease stages in diabetes patients: A TabNet approach," *Artif. Intell. Med.*, vol. 166, p. 103153, Aug. 2025, doi: 10.1016/j.artmed.2025.103153.

[58] Md. Mustafizur Rahman, Md. Al-Amin, and J. Hossain, "Machine learning models for chronic kidney disease diagnosis and prediction," *Biomed. Signal Process. Control*, vol. 87, p. 105368, Jan. 2024, doi: 10.1016/j.bspc.2023.105368.

[59] P. K. Rao, S. Chatterjee, K. Nagaraju, S. B. Khan, A. Almusharraf, and A. I. Alharbi, "Fusion of Graph and Tabular Deep Learning Models for Predicting Chronic Kidney Disease," *Diagnostics*, vol. 13, no. 12, Jun. 2023, doi: 10.3390/diagnostics13121981.

[60] Mr. C. V. Reddy, "Machine Learning Driven Prediction of Chronic Nephrology Related Diseases," *INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 09, no. 04, pp. 1–9, Apr. 2025, doi: 10.55041/ijsrem45739.

[61] Y. Dubey, P. Mange, Y. Barapatre, B. Sable, P. Palsodkar, and R. Umate, "Unlocking Precision Medicine for Prognosis of Chronic Kidney Disease Using Machine Learning," *Diagnostics*, vol. 13, no. 19, Oct. 2023, doi: 10.3390/diagnostics13193151.

[62] D. Yun *et al.*, "Machine learning survival analysis for predicting kidney disease progression in patients with acute kidney injury undergoing continuous kidney replacement therapy: An analysis of the LINKA database," *J. Crit. Care*, vol. 92, p. 155419, Apr. 2026, doi: 10.1016/j.jcrc.2025.155419.

[63] J. Xiao *et al.*, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *J. Transl. Med.*, vol. 17, no. 1, p. 119, Dec. 2019, doi: 10.1186/s12967-019-1860-0.

[64] H. Iftikhar, A. F. Hashem, M. Qureshi, and P. C. Rodrigues, "Clinical Application of Machine Learning Models for Early-Stage Chronic Kidney Disease Detection," *Diagnostics*, vol. 15, no. 20, p. 2610, Oct. 2025, doi: 10.3390/diagnostics15202610.

[65] Md. Rashed-Al-Mahfuz, A. Haque, A. Azad, S. A. Alyami, J. M. W. Quinn, and M. A. Moni, "Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening," *IEEE J. Transl. Eng. Health Med.*, vol. 9, pp. 1–11, 2021, doi: 10.1109/JTEHM.2021.3073629.

[66] N. Chumuang *et al.*, "An Efficiency Random Forest Algorithm for Classification of Patients with Kidney Dysfunction," in *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, IEEE, Nov. 2020, pp. 1–6. doi: 10.1109/iSAI-NLP51646.2020.9376785.

[67] Md. A. Islam, Md. Z. H. Majumder, and Md. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," *J. Pathol. Inform.*, vol. 14, p. 100189, 2023, doi: 10.1016/j.jpi.2023.100189.

[68] Y. Zhao *et al.*, "Risk Prediction of Chronic Kidney Disease Progression in Type 2 Diabetes Mellitus Across Diverse Populations," *NPJ Digit. Med.*, Feb. 2026, doi: 10.1038/s41746-026-02439-2.

[69] R. K. Halder *et al.*, "ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application," *J. Pathol. Inform.*, vol. 15, p. 100371, Dec. 2024, doi: 10.1016/j.jpi.2024.100371.

[70] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD)," *Circulation*, vol. 131, no. 2, pp. 211–219, Jan. 2015, doi: 10.1161/CIRCULATIONAHA.114.014508.

[71] K. G. M. Moons *et al.*, "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, vol. 388, p. e082505, Mar. 2025, doi: 10.1136/bmj-2024-082505.

[72] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," *BMC Med.*, vol. 17, no. 1, p. 230, Dec. 2019, doi: 10.1186/s12916-019-1466-7.

[73] S. S Band *et al.*, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Inform. Med. Unlocked*, vol. 40, p. 101286, 2023, doi: 10.1016/j.imu.2023.101286.

[74] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use

interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[75] N. Tomašev *et al.*, "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, Aug. 2019, doi: 10.1038/s41586-019-1390-1.

[76] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *J. Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00657-5.

[77] Md. Mustafizur Rahman, Md. Al-Amin, and J. Hossain, "Machine learning models for chronic kidney disease diagnosis and prediction," *Biomed. Signal Process. Control*, vol. 87, p. 105368, Jan. 2024, doi: 10.1016/j.bspc.2022.104445.

[78] T. Surapunt and S. Wang, "Ensemble modeling with a Bayesian maximal information coefficient-based model of Bayesian predictions on uncertainty data," *Electronics*, vol. 12, no. 1, p. 212, Jan. 2023, doi: 10.3390/electronics12010212.

[79] M. Hafeez *et al.*, "Hybrid deep learning and machine learning framework for multi-disease prediction," *Multimed. Tools Appl.*, 2024, doi: 10.1007/s11042-024-20327-3.