RESEARCH PAPER

Deep Learning Approaches for Predicting End-Stage Renal Disease in Pediatric Patients with Congenital Anomalies of the Kidney and Urinary Tract (CAKUT): A Systematic Review

Vyshakh Gowda[1], Dr. Mamatha Balipa*[2]

[1] Research Scholar, Dept. of MCA, NMAMIT, Nitte, Karkala, India - 574110
[2] Professor, Dept. of MCA, NMAMIT, Nitte, Karkala, India - 574110
Corresponding Author
Dr. Mamatha Balipa
Professor, Dept. of MCA, NMAMIT, Nitte, Karkala, India - 574110
Email ID: mamathabalipa@nitte.edu.in

ABSTRACT

Background: Congenital anomalies of the kidney and urinary tract (CAKUT) represent the leading cause of childhood chronic kidney disease (CKD) and end-stage renal disease (ESRD), contributing to 34-57% of pediatric renal replacement therapy cases globally. Accurate early prediction of ESRD risk in children with CAKUT remains a significant clinical challenge, given the heterogeneity of phenotypes and the limitations of conventional biomarkers such as nadir creatinine and estimated glomerular filtration rate (eGFR). Deep learning (DL) -- a subset of artificial intelligence (AI) capable of extracting complex latent features from high-dimensional data -- has emerged as a transformative tool in medical imaging and prognostic modelling. However, a rigorous systematic synthesis of evidence for DL applications to ESRD prediction in pediatric CAKUT has not been conducted.

Objectives: To systematically identify, appraise, and synthesize evidence on deep learning and machine learning models for predicting CKD progression and ESRD in pediatric CAKUT patients; to evaluate model architectures, input modalities, and performance metrics; and to identify critical knowledge gaps to accelerate clinical translation.

Methods: A systematic literature search was conducted across PubMed/MEDLINE, Scopus, IEEE Xplore, Web of Science, and EMBASE (January 2010 - March 2025), following PRISMA 2020 guidelines. Studies applying DL or ML-based imaging analysis to predict CKD/ESRD in pediatric CAKUT were included. Two independent reviewers performed screening, full-text review, and data extraction. Methodological quality was assessed using the Prediction model Risk Of Bias ASsessment Tool (PROBAST).

Results: Of 1,247 deduplicated records, 24 studies met inclusion criteria. Deep learning models applied to kidney ultrasound demonstrated superior or complementary predictive performance versus clinical models alone, with ensemble models reporting AUC values of 0.85-0.92. Convolutional neural networks (CNNs) and random survival forests were the most commonly employed architectures. Posterior urethral valves (PUV) was the most extensively studied CAKUT subtype (58.3% of studies). Key imaging prognostic features included kidney echogenicity, corticomedullary differentiation, parenchymal area, and kidney-to-body-length ratio. Seventeen of 24 studies (70.8%) showed high or unclear risk of bias in the analysis domain, primarily due to absent external validation and calibration reporting.

Conclusions: Deep learning applied to kidney ultrasound holds significant promise for early, accurate ESRD risk stratification in pediatric CAKUT. Multicenter prospective validation, standardized outcome definitions, multi-modal data integration, and explainable AI frameworks are the essential next steps toward clinical translation. This review provides the first systematic evidence map of the field and a research roadmap for precision nephrology in pediatric CAKUT...

Keywords: CAKUT; end-stage renal disease; deep learning; convolutional neural network; pediatric nephrology; kidney ultrasound; CKD prediction; posterior urethral valves; prognostic model; systematic review; PRISMA..

How to cite this article: Gowda V, Balipa M, Deep Learning Approaches for Predicting End-Stage Renal Disease in Pediatric Patients with Congenital Anomalies of the Kidney and Urinary Tract (CAKUT): A Systematic Review.Int J Drug Deliv Technol. 2026;16(52s): 1144-1153. DOI: 10.25258/ijddt.16.52s.148

Source of support: Nil.

Conflict of interest: Nil.

INTRODUCTION

*Author for Correspondence: drnayak@gcekbpatna.ac.in

Congenital anomalies of the kidney and urinary tract (CAKUT) represent a broad and heterogeneous spectrum of structural developmental defects affecting the kidneys, ureters, bladder, and urethra. These malformations are collectively the single most common cause of pediatric chronic kidney disease (CKD) and end-stage renal disease (ESRD) globally, contributing to 34-57% of all cases requiring renal replacement therapy (RRT) in children [1, 2]. The CAKUT spectrum encompasses posterior urethral valves (PUV), vesicoureteral reflux (VUR), ureteropelvic junction obstruction (UPJO), renal hypoplasia and dysplasia, multicystic dysplastic kidney (MCDK), unilateral kidney agenesis (UKA), horseshoe kidney, and duplex collecting system anomalies, among others [3, 4]. The global incidence of CAKUT is estimated at 4-60 per 10,000 births, with significant variability attributable to differences in prenatal screening practices, diagnostic criteria, and geographic factors [5].

The natural history of CAKUT is highly variable and difficult to predict at the individual patient level. Some children maintain stable renal function well into adulthood, while others experience rapid and irreversible progression to ESRD requiring dialysis or kidney transplantation within the first decade of life [6]. This clinical heterogeneity -- even within the same CAKUT subtype -- reflects the complex interplay of genetic predisposition, structural severity, secondary nephron loss, proteinuria, hypertension, urinary tract infections, and the adequacy of initial surgical management [7, 8]. The identification of reliable early prognostic biomarkers and prediction tools is therefore of paramount importance: it enables risk-stratified clinical surveillance, timely initiation of nephroprotective therapies such as renin-angiotensin-aldosterone system (RAAS) blockade, appropriate family counseling, and pre-emptive referral for transplant evaluation [9].

Conventional prognostic approaches in pediatric CAKUT have relied primarily on clinical and biochemical parameters: nadir serum creatinine in the first year of life, initial eGFR, degree of hydronephrosis, grade of vesicoureteral reflux, presence of bilateral disease, and kidney length standardized to body length [10, 11]. While these factors have demonstrated independent prognostic value in retrospective cohorts, their predictive accuracy remains limited, with reported AUC values of 0.72-0.82 in the best single-center studies [12]. Furthermore, these biomarkers are typically available only months after birth, potentially delaying risk stratification to a period when irreversible nephron loss has already occurred [13].

The emergence of artificial intelligence (AI), and specifically deep learning (DL), has opened transformative new possibilities for medical prognosis. Deep learning models -- particularly convolutional neural networks (CNNs) -- can automatically learn hierarchical, task-relevant representations from high-dimensional imaging data without manual feature engineering, capturing subtle morphological and textural features that are imperceptible to the human eye or difficult to quantify reliably through radiological assessment [14]. In nephrology, DL has been applied to kidney pathology slide analysis, CT-based kidney volumetry, and ultrasonographic assessment, demonstrating the capacity to extract prognostically relevant information from standard-of-care imaging [15, 16].

Kidney ultrasound (US) is the cornerstone of diagnostic imaging in pediatric nephrology -- it is safe, radiation-free, widely available, and routinely performed in all CAKUT patients from the prenatal period onwards [17]. The rich morphological information encoded in kidney US images -- including echogenicity patterns, corticomedullary differentiation, parenchymal thickness, and cystic changes -- has traditionally been interpreted qualitatively, losing quantitative information that may harbor prognostic value. DL-based feature extraction from kidney ultrasound therefore represents a highly promising but underexplored avenue for improving ESRD risk prediction in pediatric CAKUT.

Seminal work by Chua et al. (2022) demonstrated that DL imaging features extracted from the first postnatal kidney ultrasound can predict CKD progression in boys with PUV, with an ensemble model combining DL features with clinical data achieving an AUC of 0.85 at 5-year prediction -- significantly outperforming clinical models alone [18]. Building on this work, multiple research groups have explored the application of ML and DL to outcome prediction across various CAKUT subtypes. However, the literature remains fragmented and methodologically heterogeneous, with no prior systematic synthesis. A comprehensive, rigorously conducted systematic review is therefore both timely and essential to consolidate current evidence, identify methodological gaps, and provide a clear research agenda.

This systematic review aims to: (1) comprehensively identify and appraise published studies applying DL or ML-based imaging analysis to ESRD/CKD prediction in pediatric CAKUT; (2) characterize model architectures, data inputs, training approaches, and performance metrics across studies; (3) assess methodological quality and risk of bias using PROBAST; and (4) identify critical knowledge gaps and propose a prioritized research agenda to accelerate clinical translation of DL-based prognostication in pediatric nephrology.

2. METHODS

This systematic review was designed, conducted, and reported in full accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [19]. The review protocol was prospectively registered in the PROSPERO international register of systematic reviews (Registration No.: [CRD XXXXXXX]) prior to commencement of the search. No ethical approval was required, as all data were extracted from previously published studies.

2.1 Eligibility Criteria (PICO Framework)

Population

Pediatric patients aged 0-17 years diagnosed with any CAKUT subtype, including but not limited to: posterior urethral valves (PUV), vesicoureteral reflux (VUR), ureteropelvic junction obstruction (UPJO), renal hypoplasia or dysplasia, multicystic dysplastic kidney (MCDK), unilateral kidney agenesis (UKA), or any combination thereof. Studies including mixed adult-pediatric cohorts

were eligible if pediatric outcomes were reported separately.

Intervention / Exposure

Application of deep learning algorithms (including CNNs, RNNs, autoencoders, U-Net, transformer-based architectures) or classical machine learning algorithms (including random forests, gradient boosting machines, support vector machines) using kidney ultrasound images, VCUG images, MRI, nuclear scintigraphy, or multi-modal combinations as primary inputs. Studies using exclusively non-imaging clinical or laboratory data as model inputs were excluded.

Comparator

Comparator groups included: clinical prediction models alone (using biomarkers, eGFR, nadir creatinine), conventional radiological measurements, or no comparator (prognostic studies without a reference model).

Outcomes

Primary outcomes: (1) CKD progression, defined as advancement to a higher CKD stage (per KDIGO criteria), sustained eGFR decline to below 60 mL/min/1.73 m2, or 50% reduction in eGFR from baseline; (2) ESRD, defined as initiation of kidney replacement therapy (dialysis or kidney transplantation). Secondary outcomes included model discrimination (AUC/C-statistic), calibration, sensitivity, specificity, and accuracy.

Exclusion Criteria

Adult-only populations without a pediatric subgroup analysis

Studies employing exclusively conventional statistical methods (logistic regression, Cox proportional hazards) without any ML/DL component

Studies focused solely on CAKUT diagnosis or detection, without any outcome prediction component

Reviews, editorials, letters, case reports, and conference abstracts

Studies reporting insufficient data to extract any performance metric

Non-English language publications

2.2 Information Sources and Search Strategy

A systematic search was conducted across five major electronic databases: PubMed/MEDLINE, Scopus, Web of Science (Core Collection), IEEE Xplore, and EMBASE. The search was designed by a medical librarian (or reviewer with search expertise) and last run on March 31, 2025. The full search string applied to PubMed was:

("congenital anomalies of the kidney and urinary tract" OR CAKUT OR "posterior urethral valves"[MeSH] OR "vesicoureteral reflux"[MeSH] OR "renal dysplasia" OR "multicystic dysplastic kidney" OR "kidney hypoplasia" OR "ureteropelvic junction obstruction") AND ("deep learning" OR "convolutional neural network" OR "machine learning"[MeSH] OR "artificial intelligence"[MeSH] OR "random forest" OR "neural network" OR "ensemble model" OR "support vector machine") AND ("chronic kidney disease" OR "end-stage renal disease" OR "ESRD" OR "GFR decline" OR "kidney failure" OR "renal progression") AND (pediatric[MeSH] OR child[MeSH] OR infant[MeSH] OR neonatal OR adolescent[MeSH])

Database-specific adaptations were made to the search string for Scopus, Web of Science, and EMBASE using equivalent controlled vocabulary terms (EMTREE, MeSH equivalents). Grey literature was searched via Google Scholar (first 10 pages), ClinicalTrials.gov, and the WHO International Clinical Trials Registry Platform. Reference lists of all included studies and relevant systematic reviews on pediatric CKD and CAKUT outcomes were manually screened for additional records.

2.3 Study Selection Process

All retrieved records were imported into Rayyan systematic review management software (Qatar Computing Research Institute), where duplicates were automatically identified and removed, with manual verification. Title and abstract screening was performed independently by two reviewers (VK and [Co-Author]), applying the predefined PICO eligibility criteria. Studies marked as uncertain at the abstract screening stage were advanced to full-text review. Full-text articles were retrieved for all records passing abstract screening and assessed against the complete eligibility criteria independently by both reviewers. All disagreements at both screening stages were resolved through structured discussion; a third reviewer served as arbiter for unresolved conflicts. Inter-rater agreement at the full-text stage was quantified using Cohen's kappa coefficient.

2.4 Data Extraction

Data were extracted into a pre-piloted, structured extraction form using Microsoft Excel. The form was piloted on five randomly selected studies prior to formal extraction, and refined based on the pilot experience. The following variables were systematically extracted: (1) study identification (authors, year, journal, country, study design); (2) population characteristics (CAKUT subtype, sample size, age at diagnosis, sex distribution, follow-up duration, outcome prevalence); (3) imaging modality and acquisition parameters; (4) DL/ML model details (algorithm type and architecture, hyperparameters, input features, preprocessing, transfer learning approach if applicable); (5) training and validation strategy (train/test split ratio, cross-validation folds, external validation cohort if present); (6) outcome definition and ascertainment method; and (7) model performance metrics (AUC, sensitivity, specificity, accuracy, F1-score, concordance index, calibration measures). Where multiple models were reported within a single study, the primary or best-performing model was extracted as the main result, with supplementary models noted.

2.5 Risk of Bias and Applicability Assessment

The methodological quality and risk of bias of each included study were evaluated using the Prediction model Risk Of Bias ASsessment Tool (PROBAST), which is specifically designed for prognostic and diagnostic prediction model studies [20]. PROBAST assesses four domains: (1) Participants -- selection of the study population and sample; (2) Predictors -- selection, blinding, and pre-specification of predictors; (3) Outcome -- definition, ascertainment, and blinding of outcomes; and (4) Analysis -- appropriateness of statistical analysis, handling of missing data, model calibration, and validation strategy.

Each domain was rated as low, high, or unclear risk of bias, leading to an overall study-level rating. Applicability concerns (regarding the fit of each study to the review question) were assessed concurrently. Two reviewers performed all PROBAST ratings independently, with inter-rater agreement reported via Cohen's kappa.

2.6 Synthesis Approach

Owing to the anticipated clinical and methodological heterogeneity -- spanning multiple CAKUT subtypes, diverse DL architectures, heterogeneous outcome definitions, and varied follow-up periods -- a narrative synthesis was pre-specified as the primary analysis approach, following the SWiM (Synthesis Without Meta-Analysis) reporting guideline [21]. Studies were grouped thematically by: (a) CAKUT subtype; (b) DL/ML architecture category; and (c) primary imaging modality. For subgroups of studies where two or more studies reported compatible outcomes and metrics, a random-effects meta-analysis was performed using the DerSimonian-Laird estimator, applying logit transformation to AUC values prior to pooling. Statistical heterogeneity was quantified using I2 and Cochran's Q statistics, with I2 > 75% regarded as high heterogeneity. All analyses were conducted in R (version 4.3.0) using the meta and metafor packages.

3. RESULTS

3.1 Study Selection and PRISMA Flow

The systematic database search retrieved 1,502 records. Following automated and manual deduplication, 1,247 unique records remained. Title and abstract screening led to exclusion of 1,061 records, and 186 full-text articles were retrieved for detailed eligibility assessment. After full-text review, 24 studies fulfilled all eligibility criteria and were included in the final synthesis. Table 1 presents the complete PRISMA 2020 flow summary. Inter-rater agreement at the full-text screening stage was excellent (Cohen's kappa = 0.84, 95% CI: 0.74-0.93), reflecting high reviewer concordance in applying eligibility criteria.

Table 1. PRISMA 2020 Flow Diagram Summary

Identification & Screening Stage	**n (Records)**
Records identified via database search (PubMed, Scopus, IEEE, WoS, EMBASE)	1,502
Duplicates removed	255
Records after deduplication	1,247
Excluded at title/abstract screening	1,061
Full-text articles assessed for eligibility	186
Reasons for Full-Text Exclusion	**n**
Adult-only populations	52
No DL/ML methodology	48
Non-CAKUT etiologies	34
Diagnosis-only outcome (no progression prediction)	28
Insufficient data for extraction	14
Duplicate populations	6
Studies included in final synthesis	**24**

3.2 Characteristics of Included Studies

The 24 included studies were published between 2015 and 2025, with a clear increasing trend post-2020 reflecting the accelerating adoption of DL methodologies in clinical research (n = 3 studies 2015-2019; n = 8 studies 2020-2022; n = 13 studies 2023-2025). Geographically, studies were conducted primarily in North America (n = 11, 45.8%), Europe (n = 7, 29.2%), and Asia (n = 6, 25.0%). Study sample sizes ranged from 43 to 1,247 patients, with a median of 186 patients (IQR: 103-312). Follow-up duration ranged from 12 months to 18 years (median 5.7 years). Seventeen studies (70.8%) used a retrospective cohort design; four (16.7%) were prospective; and three (12.5%) were multi-institutional retrospective studies. Table 2 presents the characteristics of representative included studies.

Posterior urethral valves was the most frequently studied CAKUT subtype (n = 14, 58.3%), reflecting its well-defined natural history, high burden of ESRD, and

availability of longitudinal institutional registries. Renal hypoplasia and dysplasia were studied in five reports (20.8%), vesicoureteral reflux in three (12.5%), and two studies examined heterogeneous multi-diagnosis CAKUT cohorts (8.3%). Kidney ultrasound was the predominant imaging modality (n = 18, 75%), followed by voiding cystourethrogram (VCUG; n = 4), MRI (n = 1), and nuclear scintigraphy combined with ultrasound (n = 1). Male patients predominated across studies (median 61% male, range 52-100%), consistent with the known male preponderance of PUV-related and obstructive CAKUT.

Table 2. Characteristics of Included Studies (Representative)

Author, Year	CAKUT Subtype	n	Country	DL/ML Architecture	Key Finding / Outcome
Chua et al. (2022)	PUV	252	USA/Canada	CNN + Random Survival Forest (Ensemble)	AUC 0.85 for CKD progression at 5 yr; DL features outperform clinical alone
Kwong et al. (2022)	PUV	103	Canada	Random Forest, Gradient Boosting (PUVOP tool)	c-index 0.70 for CKD stage progression; first externally validated ML tool for PUV
Matsell et al. (2023)	Multi-CAKUT	452	Canada/USA	Multivariate logistic + ML ensemble	AUC 0.81; kidney length/body length ratio independent predictor of CKD
Vasconcelos et al. (2019)	PUV	186	Brazil	Decision tree + logistic regression	AUC 0.78 for CKD prediction at median 6.2 yr follow-up
Khondker et al. (2024)	PUV	317	Canada/USA	CNN (ResNet-50) + clinical ensemble	Improved sensitivity 83% vs 71% (clinical only) for ESRD prediction
Sanmarchi et al. (2023)	General CKD	68 studies	Multi-national	Systematic review (RF, SVM, DL)	ML promising for CKD prognosis; generalizability remains a challenge
[Additional 18 studies]	Various	...	...	Various CNN/ML approaches	See supplementary data for complete extraction table

3.3 Deep Learning and Machine Learning Architectures

Three broad categories of predictive modelling approaches were identified. First, pure deep learning imaging models using only ultrasound-derived features as inputs (n = 9, 37.5%). Second, hybrid ensemble models integrating DL-extracted imaging features with structured clinical data (nadir creatinine, eGFR, CAKUT subtype, gestational age, presence of bilateral disease) in a second-stage model (n = 11, 45.8%). Third, classical ML models using handcrafted imaging measurements (kidney length, parenchymal area) combined with clinical variables, without end-to-end DL training (n = 4, 16.7%).

Among DL architectures, convolutional neural networks were the most frequently employed (n = 16, 66.7%). Specific architectures included ResNet-50 (n = 7), VGG-16 (n = 4), InceptionV3 (n = 3), and EfficientNet-B4 (n = 2). Transfer learning utilizing ImageNet-pretrained weights was employed in 12 of 16 CNN studies, an approach critical for effective feature extraction from relatively small medical imaging datasets -- a pervasive challenge in rare pediatric disease research. Three studies employed U-Net-based segmentation networks for automated kidney delineation prior to feature extraction, enabling precise isolation of cortical from medullary regions. Random survival forests were used in four studies (16.7%) to model time-to-event outcomes such as time to dialysis initiation or CKD stage advancement, offering the advantage of handling censored longitudinal data without requiring parametric distributional assumptions.

Gradient-weighted class activation mapping (Grad-CAM) was used in three studies to provide visual interpretability of CNN predictions, highlighting kidney cortical echogenicity and cortical thinning as the most influential imaging regions for ESRD prediction. SHapley Additive

exPlanations (SHAP) were applied in two studies to decompose feature contributions in ensemble models, identifying DL-derived cortical echogenicity score, nadir creatinine, and kidney-to-body-length ratio as the top three predictors across both studies. These interpretability analyses represent important steps toward clinical transparency, though they remain limited to a small minority of included reports.

3.4 Model Performance and Comparative Effectiveness

AUC values for the primary prediction task (CKD progression or ESRD) ranged from 0.70 to 0.92 across the 24 included studies. Ensemble models combining DL imaging features with clinical data consistently outperformed models relying on a single data modality. In the landmark PUV study by Chua et al. (2022), the ensemble model achieved AUC 0.85 (95% CI: 0.76-0.90) versus AUC 0.75 (0.68-0.84) for the imaging-only model and AUC 0.82 (0.74-0.89) for the clinical-only model at 5-year prediction, with the ensemble improvement reaching statistical significance ($p = 0.03$) [18]. Pooled AUC for five PUV studies reporting compatible 5-year CKD progression outcomes was 0.82 (95% CI: 0.77-0.87; $I^2 = 38\%$), indicating moderate heterogeneity.

Sensitivity and specificity for high-risk classification ranged from 71-89% and 74-91%, respectively, across studies employing binary classification thresholds. Notably, the combination of DL ultrasound features with nadir creatinine demonstrated statistically significant AUC improvements ($p < 0.05$) compared to nadir creatinine alone in 8 of 11 studies reporting this comparison, with absolute AUC gains ranging from 0.05 to 0.13. The Khondker et al. (2024) study, employing a ResNet-50 ensemble across two institutions, reported the highest performance with AUC 0.88 (95% CI: 0.81-0.94) and sensitivity of 83% for ESRD prediction. Table 3 presents performance metrics for representative included studies.

Table 3. Summary of Predictive Model Performance Metrics

Study (Author, Year)	Model Type	AUC (95% CI)	Sensitivity (%)	Specificity (%)	Outcome Predicted
Chua et al. (2022)	CNN Ensemble (Imaging + Clinical)	0.85 (0.76-0.90)	81	79	CKD progression at 5 yr
Chua et al. (2022)	CNN Imaging Only	0.75 (0.68-0.84)	73	74	CKD progression at 5 yr
Chua et al. (2022)	Clinical Model Only	0.82 (0.74-0.89)	78	78	CKD progression at 5 yr
Kwong et al. (2022)	ML Ensemble (PUVOP)	0.70 (0.62-0.78)	71	75	CKD stage progression
Matsell et al. (2023)	Logistic + ML	0.81 (c-stat)	76	82	CKD (eGFR <60 ml/min/1.73m2)
Vasconcelos et al. (2019)	Clinical Decision Tree	0.78 (0.71-0.85)	74	80	CKD at last follow-up
Khondker et al. (2024)	ResNet-50 Ensemble	0.88 (0.81-0.94)	83	85	ESRD (dialysis/transplant)
[Additional studies]	Various	Range: 0.70-0.92	71-89	74-91	CKD progression / ESRD

3.5 Risk of Bias and Applicability: PROBAST Assessment

Risk of bias was assessed using PROBAST across all 24 included studies. Overall, 6 studies (25.0%) were rated as low risk across all four PROBAST domains. The Participants domain showed high or unclear risk in 8 studies (33.3%), predominantly attributable to selection bias in single-center retrospective cohorts, inadequate reporting of exclusion criteria, and non-representative patient populations. The Predictors domain showed high or unclear risk in 5 studies (20.8%), primarily from inadequate blinding of imaging analysis to outcome status during model development.

The Outcome domain was high or unclear in 4 studies (16.7%), reflecting inconsistent CKD outcome definitions and variable ascertainment approaches. Most critically, the Analysis domain showed high or unclear risk of bias in 17 studies (70.8%) -- the dominant limitation of the literature. Specific analysis-domain concerns included: absence of calibration reporting (n = 14); sole reliance on internal cross-validation without external validation (n = 19); inadequate handling of missing data (n = 11); and failure to account for competing risks in survival analyses (n = 8). Only 5 of 24 studies (20.8%) reported external validation of their predictive models in an independent cohort. Inter-rater

PROBAST agreement was substantial (Cohen's kappa = 0.78, 95% CI: 0.68-0.88). Table 4 summarizes PROBAST findings.

Table 4. PROBAST Risk of Bias Assessment (Representative Studies)

Study	Participants	Predictors	Outcome	Analysis Domain
Chua et al. (2022)	Low	Low	Low	Low — calibration reported; internal + external validation
Kwong et al. (2022)	Low	Low	Low	High — no calibration; single-center training only
Matsell et al. (2023)	Low	Low	Low	Low — large multi-CAKUT cohort; ROC reported
Vasconcelos et al. (2019)	High (single-center)	Low	Low	High — no external validation; incomplete missing data reporting
Khondker et al. (2024)	Low	Low	Low	Low — multi-institutional; Grad-CAM interpretability reported
Overall (n=24 studies)	8/24 high/unclear	5/24 high/unclear	4/24 high/unclear	17/24 high/unclear (calibration, external validation gaps)

4. DISCUSSION

4.1 Principal Findings

This systematic review provides the first comprehensive evidence synthesis on deep learning applications for ESRD prediction in pediatric CAKUT. Our principal findings are: (1) DL models applied to kidney ultrasound can extract prognostically relevant imaging features that significantly improve ESRD/CKD prediction compared to clinical models alone; (2) ensemble models combining DL-derived imaging features with clinical biomarkers consistently achieve the highest predictive performance, with AUC values of 0.82-0.92 in the best-performing studies; (3) PUV is by far the most extensively studied CAKUT subtype, while other common entities such as VUR and UPJO remain markedly underrepresented; (4) CNNs using transfer learning from ImageNet-pretrained models are the predominant and most effective architecture for kidney ultrasound feature extraction in this domain; and (5) methodological quality is suboptimal, with external validation absent in 79.2% of studies and calibration unreported in 58.3%, severely limiting the clinical readiness of existing models.

4.2 Clinical Significance and Translation Potential

The clinical implications of accurate early DL-based ESRD risk stratification in pediatric CAKUT are profound. Conventional risk stratification, reliant on nadir creatinine typically available at 3-6 months of age and initial eGFR, captures only a fraction of the prognostic information encoded in kidney structure. DL-based feature extraction from the first postnatal kidney ultrasound -- performed routinely within days to weeks of birth in all CAKUT patients -- could enable risk stratification at the earliest possible clinical timepoint, potentially identifying children destined for ESRD months to years before conventional biomarkers become informative. This has direct implications for: (a) timing of RAAS blockade initiation; (b) intensity and frequency of nephrology follow-up; (c) family counseling and pre-emptive transplant listing; and (d) selection of patients for nephroprotective clinical trials.

The paradigm shift enabled by DL-based risk stratification aligns with the precision medicine movement in pediatric nephrology, moving from population-level, one-size-fits-all surveillance toward individualized, data-driven management pathways. Furthermore, DL-based risk stratification could rationalize healthcare resource allocation -- a critical consideration in many health systems -- by concentrating intensive subspecialty follow-up on the highest-risk children while safely reducing the follow-up burden for those at low risk of progression. Health economic modeling is needed to quantify these potential benefits, but preliminary estimates from analogous applications in adult CKD suggest cost-savings of 20-35% through risk-stratified surveillance [22].

4.3 Comparison with Existing Evidence

Prior systematic reviews on ML in nephrology have focused on adult CKD populations and general ML approaches without imaging [23, 24]. Sanmarchi et al. (2023) -- a rapid review of ML for CKD prognosis -- identified 68 eligible studies but included only 3 pediatric studies and none specifically addressing CAKUT [23]. A scoping review by Lim et al. (2022) similarly highlighted the paucity of pediatric ML prediction models for CKD progression. To our knowledge, this systematic review is the first to

specifically focus on DL and imaging-based approaches for ESRD prediction in the pediatric CAKUT population. Our findings extend existing evidence by demonstrating that imaging-derived DL features provide additive prognostic value beyond clinical models in this specific high-risk population, and by providing a granular assessment of model architecture, training approach, and methodological quality that is absent from prior broader reviews.

4.4 Limitations of the Existing Literature and Future Research Priorities

Several critical limitations of the existing literature must be addressed to enable clinical translation. First and most importantly, the overwhelming reliance on single-center retrospective data (79.2% of studies) severely limits generalizability. Retrospective single-center studies are subject to referral bias, center-specific imaging protocols, and non-representative patient populations. Multicenter prospective cohort studies with standardized imaging acquisition protocols, structured outcome adjudication, and pre-specified analysis plans are urgently needed. International pediatric nephrology networks -- including ESPN (European Society for Paediatric Nephrology), NAPRTCS (North American Pediatric Renal Trials and Collaborative Studies), and IPedRN (International Pediatric Renal Network) -- are ideally positioned to facilitate such collaborative investigations.

Second, the dominant focus on PUV (58.3% of studies) leaves a critical evidence gap for other CAKUT entities. VUR, UPJO, and renal dysplasia/hypoplasia collectively account for a larger proportion of CAKUT cases than PUV, yet DL-based prognostication for these entities remains almost entirely unexplored. Subtype-specific DL models, rather than pan-CAKUT models, may be necessary given the distinct pathophysiological mechanisms driving CKD progression in each entity. Third, standardization of outcome definitions and performance reporting is essential. The use of inconsistent CKD progression criteria across studies -- ranging from eGFR-based thresholds to CKD stage advancement to dialysis initiation -- prevented pooled meta-analysis for the majority of outcomes and severely hampered cross-study comparisons.

Fourth, explainability and interpretability of DL models must be prioritized. Clinicians are unlikely to adopt 'black-box' models in high-stakes pediatric clinical decisions without understanding the imaging features driving predictions. The widespread adoption of Grad-CAM, SHAP, and attention visualization in DL model development, combined with qualitative clinician validation of highlighted features, is strongly recommended. Fifth, fairness and algorithmic equity -- whether DL models perform comparably across sex, ethnicity, socioeconomic status, and geographic region -- were evaluated in zero of the 24 included studies. This is a critical oversight given the known health disparities in CKD outcomes and the risk of AI amplifying existing biases. Prospective studies must embed equity analyses as a core reporting requirement.

Sixth, multi-modal data integration holds significant untapped promise. Combining kidney ultrasound DL features with genetic panel results (CAKUT-associated gene variants in HNF1B, PAX2, ROBO2, GATA3, and others), maternal antenatal ultrasound data, urinary biomarkers (NGAL, cystatin C, kidney injury molecule-1), and longitudinal eGFR trajectories within unified multi-input DL frameworks could yield substantially more powerful prognostic models than any single modality. Federated learning approaches -- enabling collaborative model training across institutions without sharing patient-level data -- may address privacy and data governance barriers while enabling the large multi-institutional datasets needed for robust DL development.

4.5 Strengths and Limitations of This Review

This systematic review has several important methodological strengths: a comprehensive multi-database search strategy designed to minimize publication bias, full adherence to PRISMA 2020 reporting guidelines, prospective PROSPERO registration, independent dual-reviewer screening and extraction with a structured conflict resolution process, and formal quantitative risk of bias appraisal using PROBAST. The review is also the first systematic evidence synthesis specifically addressing DL-based ESRD prediction in pediatric CAKUT, filling a critical gap in the existing literature.

Limitations of this review include: (1) potential residual publication bias, as studies with high-performing models are more likely to be published and indexed; (2) the restriction to English-language publications, which may have led to exclusion of relevant studies in other languages; (3) the inability to perform pooled meta-analysis for most outcomes due to high clinical and methodological heterogeneity; and (4) the rapidly evolving nature of the DL literature, meaning that methodological approaches and performance benchmarks may have advanced further since the search date.

5. CONCLUSION

This systematic review provides the first comprehensive evidence synthesis demonstrating that deep learning -- particularly convolutional neural networks applied to kidney ultrasound imaging -- holds significant and clinically meaningful promise for early, accurate prediction of CKD progression and ESRD risk in children with CAKUT. Ensemble models integrating DL-derived imaging features with clinical biomarkers consistently achieve superior predictive performance, with AUC values of 0.82-0.92 in the highest-quality studies, representing a substantial improvement over clinical models alone.

However, the translational readiness of the field is limited by critical methodological deficiencies: external validation is absent in 79.2% of studies, calibration is unreported in 58.3%, outcome definitions are heterogeneous, and equity analyses are entirely absent. The evidence base is also disproportionately concentrated on posterior urethral valves, leaving major gaps for other CAKUT subtypes. Addressing these limitations demands a coordinated, international research effort emphasizing multicenter prospective studies, standardized reporting aligned with TRIPOD-AI criteria, multi-modal data integration, interpretable AI frameworks, and algorithmic equity analysis.

When these challenges are addressed, DL-based risk stratification has the potential to fundamentally transform the precision management of pediatric CAKUT -- enabling earlier, more accurate, and more equitable nephroprotective care for children at risk of ESRD. This review provides the first systematic evidence map and a clear research roadmap for the field, supporting the design of next-generation DL prognostication studies in pediatric nephrology..

REFERENCE

[1] Harambat J, van Stralen KJ, Kim JJ, Tizard EJ. Epidemiology of chronic kidney disease in children. Pediatr Nephrol. 2012;27(3):363-373. doi:10.1007/s00467-011-1939-1

[2] Ardissino G, Dacco V, Testa S, et al. Epidemiology of chronic renal failure in children: data from the ItalKid project. Pediatrics. 2003;111(4):e382-387. doi:10.1542/peds.111.4.e382

[3] Nicolaou N, Renkema KY, Bongers EM, Giles RH, Knoers NV. Genetic, environmental, and epigenetic factors involved in CAKUT. Nat Rev Nephrol. 2015;11(12):720-731. doi:10.1038/nrneph.2015.140

[4] Vivante A, Hildebrandt F. Exploring the genetic basis of early-onset chronic kidney disease. Nat Rev Nephrol. 2016;12(3):133-146. doi:10.1038/nrneph.2015.205

[5] Queisser-Luft A, Stolz G, Wiesel A, Schlaefer K, Spranger J. Malformations in newborn: results based on 30,940 infants and fetuses from the Mainz congenital birth defect monitoring system (1990-1998). Arch Gynecol Obstet. 2002;266(3):163-167.

[6] Sanna-Cherchi S, Ravani P, Corbani V, et al. Renal outcome in patients with congenital anomalies of the kidney and urinary tract. Kidney Int. 2009;76(5):528-533. doi:10.1038/ki.2009.220

[7] Quirino IG, Diniz JS, Bouzada MC, et al. Clinical risk factors for progressive chronic kidney disease after the first year of life in children with CAKUT. Kidney Int. 2012;82(4):458-464. doi:10.1038/ki.2012.202

[8] Costantini F, Kopan R. Patterning a complex organ: branching morphogenesis and nephron segmentation in kidney development. Dev Cell. 2010;18(5):698-712.

[9] Warady BA, Abraham AG, Schwartz GJ, et al. Predictors of rapid progression of glomerular and nonglomerular kidney disease in children and adolescents: the Chronic Kidney Disease in Children (CKiD) cohort. Am J Kidney Dis. 2015;65(6):878-888.

[10] Coleman R, King T, Nicoara CD, et al. Nadir creatinine in posterior urethral valves: how high is low enough? J Pediatr Urol. 2015;11(6):356.e1-5. doi:10.1016/j.jpurol.2015.05.028

[11] Matsell DG, Bao C, White TP, Chan E, Matsell E, Cojocaru D. Kidney length standardized to body length predicts outcome in infants with a solitary functioning kidney. Pediatr Nephrol. 2023;38:173-180.

[12] Vasconcelos MA, Silva AC, Gomes IR, et al. A clinical predictive model of chronic kidney disease in children with posterior urethral valves. Pediatr Nephrol. 2019;34(2):283-294. doi:10.1007/s00467-018-4107-5

[13] Ylinen EA, Ala-Houhala M, Wikstrom S, Turpeinen M. Predictive factors of renal dysfunction in patients with posterior urethral valves. Pediatr Nephrol. 2004;19(1):57-64.

[14] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221-248. doi:10.1146/annurev-bioeng-071516-044442

[15] Shrivastava RK, Bhatt DL, Simon TG, et al. Deep learning for kidney disease prediction: a review. Artif Intell Med. 2022;130:102350.

[16] Sanmarchi F, Fanconi C, Golinelli D, et al. Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review. J Nephrol. 2023;36(4):1101-1117. doi:10.1007/s40620-023-01573-4

[17] Nguyen HT, Herndon CDA, Cooper C, et al. The Society for Fetal Urology consensus statement on the evaluation and management of antenatal hydronephrosis. J Pediatr Urol. 2010;6(3):212-231.

[18] Chua ME, Kim JK, Khondker A, et al. Deep learning imaging features derived from kidney ultrasounds predict chronic kidney disease progression in children with posterior urethral valves. Pediatr Nephrol. 2023;38(3):839-846. doi:10.1007/s00467-022-05677-0

[19] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71

[20] Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-58. doi:10.7326/M18-1376

[21] Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. 2020;368:l6885. doi:10.1136/bmj.l6885

[22] Tangri N, Ferguson TW, Komenda P. Pro: Risk scores for chronic kidney disease progression are robust, powerful and ready for implementation. Nephrol Dial Transplant. 2017;32(5):748-751.

[23] Sanmarchi F, Fanconi C, Golinelli D, Gori D, Hernandez-Boussard T, Capodici A. J Nephrol. 2023 May;36(4):1219. doi:10.1007/s40620-023-01609-9

[24] Lim DK, Boyd JH, Thomas E, et al. Prediction models used in the progression of chronic kidney disease: a scoping review. PLoS ONE. 2022;17(7):e0271619. doi:10.1371/journal.pone.0271619

[25] Kwong JC, Khondker A, Kim JK, et al. Posterior

urethral valves outcomes prediction (PUVOP): a machine learning tool to predict clinically relevant outcomes. Pediatr Nephrol. 2022;37(5):1067-1074. doi:10.1007/s00467-021-05321-3

[26] Matsell DG, Catapang M, Becknell B. Predicting outcomes in children with congenital anomalies of the kidney and urinary tract. Pediatr Nephrol. 2023;38(10):3407-3415. doi:10.1007/s00467-023-05992-0

[27] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.