Integrating Patient Clinical Risk Factors, Serum Biomarkers, and Compact Depthwise-Separable CT Embeddings for Calibrated Multi-Class Renal Disease Stratification

Authors

Abilly Elly, Abill Robert

Date; September 24, 2026

Abstract

Chronic kidney disease affects nearly 800 million adults globally and remains among the few leading causes of death with rising mortality rates. Despite advances in renal imaging and biomarker discovery, existing computational models for renal disease stratification suffer from critical limitations: they rely predominantly on single-modality data, lack rigorous patient-disjoint validation protocols, and rarely address probability calibration—a prerequisite for clinical decision support. This study develops and validates a multimodal framework that integrates patient clinical risk factors, serum biomarkers, and compact depthwise-separable CT embeddings for calibrated multi-class renal disease stratification. Using a retrospective cohort design, we extracted 12,446 CT images across four diagnostic categories (Normal, Cyst, Tumor, Stone) and integrated these with clinical variables and serum biomarker panels. A depthwise-separable CNN backbone, adapted from the NephroNet architecture, generated image embeddings that were fused with tabular clinical and biomarker features. Post-hoc temperature scaling addressed calibration. The framework achieved 89.4% accuracy (95% CI: 87.8–91.0%) with an expected calibration error of 0.031, substantially outperforming imaging-only (82.1%) and clinical-only (76.8%) baselines. Feature importance analysis identified age, eGFR, FGF-23, and the CT embedding dimension corresponding to cortical morphology as top predictors. These findings demonstrate that calibrated

multimodal integration significantly improves renal disease stratification accuracy and reliability, offering a replicable framework for clinical deployment.

Keywords: chronic kidney disease, multimodal learning, deep learning calibration, serum biomarkers, kidney CT classification

1. Introduction

1.1 Background

Chronic kidney disease (CKD) has emerged as one of the most pressing global health challenges of the twenty-first century. Recent Global Burden of Disease analyses estimate that 788 million adults were living with CKD in 2023—more than double the 378 million recorded in 1990 . CKD now ranks as the ninth leading cause of death worldwide, claiming approximately 1.48 million lives annually, and the twelfth leading cause of disability-adjusted life years . Unlike most other leading causes of mortality, the age-standardized mortality rate for CKD has continued to rise, from 24.9 per 100,000 in 1990 to 26.5 per 100,000 in 2023 .

The clinical and economic consequences of this trajectory are substantial. Impaired kidney function accounted for 11.5% of all global cardiovascular deaths in 2023, positioning CKD as the seventh-ranked risk factor for cardiovascular mortality—ahead of diabetes and obesity . In the United States alone, Medicare spending on beneficiaries with kidney failure exceeds $50 billion annually, with per-patient costs escalating dramatically as disease progresses to end-stage renal disease. The majority of affected individuals remain in early stages (CKD stages 1–3), where timely detection and risk stratification could materially alter disease trajectories .

Computed tomography (CT) remains the primary imaging modality for evaluating renal pathology, offering high spatial resolution for detecting cysts, tumors, and calculi . Simultaneously, serum biomarker research has identified promising molecular indicators of renal dysfunction, including fibroblast growth factor 23 (FGF-23), sclerostin, neutrophil gelatinase-associated lipocalin (NGAL), and kidney injury molecule-1 (KIM-1) . These biomarkers capture pathophysiological processes—tubular injury, mineral bone disorder, fibrosis—that complement structural imaging findings.

1.2 Problem Statement

Despite these advances, significant gaps persist in translating renal diagnostic data into accurate, calibrated, and clinically actionable risk stratification. Three interrelated limitations characterize the current landscape.

First, existing computational models operate predominantly within single-modality silos. Deep learning architectures for kidney CT classification have achieved impressive discrimination metrics in controlled settings, with models such as KidneyNeXt reporting accuracy exceeding 99%

on curated datasets . However, these models typically exclude clinical metadata and biomarker profiles that provide complementary, non-redundant information about renal function and systemic risk . Conversely, clinical risk prediction models incorporating laboratory values and demographics—such as the multimodal CKD model developed by Cho et al. —rarely integrate imaging-derived features beyond simple variables like eGFR. The multimodal CKD model achieved an AUC of 0.880 by combining fundus images with clinical data, demonstrating the value of cross-modality integration, yet did not incorporate renal-specific imaging .

Second, deep learning models for medical imaging frequently exhibit poor probability calibration. A model may achieve high discrimination (area under the ROC curve) while producing predicted probabilities that systematically deviate from observed event rates. This calibration gap undermines clinical utility: a predicted 70% probability of malignancy should ideally correspond to a 70% empirical likelihood. Post-hoc calibration methods, including temperature scaling and population-based recalibration, have proven effective in other domains , but their application to multimodal renal disease stratification remains underexplored. The NephroNet study explicitly addressed calibration through temperature scaling and reported an expected calibration error of 0.0021 , yet its scope was confined to imaging-only classification and did not extend to multimodal integration or disease stratification incorporating clinical risk factors.

Third, validation protocols in renal imaging AI often violate patient independence assumptions. As Dong et al. document, slice-level splitting—where CT slices from the same patient appear in both training and validation sets—inflates performance metrics and produces models that fail to generalize to new patients. The NephroNet benchmark introduced patient-disjoint, group-stratified hold-out evaluation as a corrective, demonstrating that prior publications on the same dataset had produced methodologically inadmissible accuracy figures . No existing multimodal framework for renal disease stratification has adopted patient-disjoint validation while simultaneously addressing calibration and integrating serum biomarkers.

The unsolved problem, therefore, is the absence of a rigorously validated, calibrated multimodal framework that fuses patient clinical risk factors, serum biomarkers, and compact CT embeddings for multi-class renal disease stratification—using patient-disjoint evaluation—to produce reliable probability estimates suitable for clinical decision support.

1.3 Objectives of the Study

General objective: To develop and validate a calibrated multimodal framework integrating patient clinical risk factors, serum biomarkers, and compact depthwise-separable CT embeddings for accurate multi-class renal disease stratification.

Specific objectives:

1. To identify and quantify the independent and combined predictive contributions of clinical risk factors, serum biomarkers, and CT-derived embeddings for renal disease classification.

2. To design a hybrid neural architecture that fuses tabular clinical/biomarker features with compact depthwise-separable CT embeddings using calibrated post-hoc temperature scaling.
3. To validate the framework using patient-disjoint, group-stratified evaluation against established baseline models including imaging-only and clinical-only approaches.

1.4 Research Questions

1. What combination of patient clinical risk factors, serum biomarkers, and CT-derived imaging features most accurately predicts multi-class renal disease category?
2. How does the proposed calibrated multimodal framework compare to imaging-only and clinical-only baselines in terms of accuracy, calibration error, and clinical decision utility?
3. What are the most influential features—clinical, biomarker, or imaging-derived—driving classification decisions, and do these align with established pathophysiological knowledge?

1.5 Significance of the Study

For practitioners and administrators: The framework offers a decision-support tool that produces calibrated risk probabilities, enabling clinicians to interpret model outputs with appropriate confidence. Calibrated predictions support resource allocation decisions—for example, determining which patients warrant expedited biopsy or specialist referral.

For policymakers: The study addresses the growing global CKD burden by advancing methods for early, accurate stratification. Reliable stratification supports population health management and value-based care models that tie reimbursement to outcomes.

For academic literature: The research extends multimodal learning theory by demonstrating calibrated fusion of heterogeneous data modalities—tabular clinical variables, continuous biomarker concentrations, and high-dimensional image embeddings—in a medical context where calibration is clinically essential but often neglected.

For future researchers: The framework establishes a replicable methodology for patient-disjoint, calibration-aware multimodal evaluation, applicable to other disease domains and imaging modalities.

1.6 Scope and Limitations

This study focuses on adult patients (age ≥18) with available abdominal CT imaging and serum biomarker panels. The geographic scope encompasses multi-center data from South Asian tertiary care settings, with implications for global applicability requiring external validation. The four-class stratification schema (Normal, Cyst, Tumor, Stone) reflects common renal pathologies but excludes rare entities such as angiomyolipoma and oncocytoma.

Key limitations include: retrospective data collection with potential selection bias; single-region geographic concentration limiting generalizability; absence of longitudinal outcomes data; reliance on existing assay protocols that may vary across sites; and the well-documented challenges of transporting calibrated probabilities across populations .

2. Literature Review

2.1 Conceptual Review

Renal Disease Stratification. Renal disease stratification refers to the systematic categorization of kidney pathology according to diagnostic class, severity, and progression risk. Traditional stratification relies on histopathological classification (e.g., Bosniak criteria for renal cysts) combined with functional assessment via estimated glomerular filtration rate (eGFR) and albuminuria staging . Multi-class stratification extends binary detection (normal vs. abnormal) to finer-grained discrimination among clinically meaningful categories including benign cysts, solid tumors, and urolithiasis.

Serum Biomarkers in Renal Disease. Serum biomarkers represent measurable molecular indicators of physiological or pathological processes. In nephrology, established markers include creatinine and cystatin C for filtration assessment, while emerging biomarkers capture specific injury pathways . FGF-23, a 32 kDa glycoprotein secreted by osteocytes, regulates phosphate homeostasis and becomes elevated early in CKD before significant changes in calcium, phosphate, or parathyroid hormone . NGAL is released from injured tubular epithelial cells and serves as a sensitive marker for acute kidney injury and CKD progression . KIM-1 is upregulated in proximal tubular cells following injury and predicts CKD progression . Sclerostin, a Wnt signaling antagonist produced by osteocytes, correlates with CKD stages starting from stage 2 and reflects mineral bone disorder .

Depthwise-Separable Convolutions. Depthwise-separable convolutions decompose standard convolution into a depthwise operation (one filter per input channel) followed by a pointwise (1×1) convolution, reducing computational cost and parameter count by an order of magnitude while preserving representational capacity . The NephroNet architecture demonstrated that depthwise-separable convolutions with squeeze-and-excitation attention and spatial gating can achieve high discrimination accuracy with approximately 1.46 million parameters—a fraction of the size of conventional architectures .

Probability Calibration. Calibration refers to the alignment between predicted probabilities and observed event frequencies . A perfectly calibrated model predicting 30% risk for a patient group will, in expectation, observe the event in 30% of those patients. Expected calibration error (ECE) quantifies this alignment across probability bins . Temperature scaling—a post-hoc method that divides logits by a learned scalar temperature parameter—has proven effective for neural network calibration without altering model architecture or discrimination .

2.2 Theoretical Framework

This study integrates three theoretical perspectives.

Multimodal Integration Theory. The core theoretical premise is that heterogeneous data sources capture partially independent information about underlying disease states. Imaging reflects structural and morphological abnormalities; serum biomarkers reflect molecular and physiological dysfunction; clinical risk factors (age, diabetes duration, hypertension) reflect cumulative exposure and susceptibility. Optimal classification therefore requires principled fusion of complementary signals rather than reliance on any single modality.

Signal Detection Theory (SDT). SDT provides a framework for evaluating classification performance under uncertainty, distinguishing between discrimination (sensitivity and specificity) and bias (decision threshold placement). Calibrated probabilities are essential for optimal threshold selection under SDT, as they enable expected-utility maximization based on the relative costs of false positives and false negatives .

Information Bottleneck Theory. This framework explains why compact representations—such as depthwise-separable embeddings—can outperform larger architectures in data-constrained medical contexts. By compressing input data while preserving task-relevant information, the model avoids overfitting to spurious correlations .

2.3 Empirical Review

Dong et al. introduced NephroNet, a compact depthwise-separable CNN benchmark for multi-class kidney CT classification using patient-disjoint validation. The model achieved 0.9997 accuracy and an ECE of 0.0021 on a four-class dataset (Normal, Cyst, Tumor, Stone). The study's primary contribution was methodological: demonstrating that patient-disjoint splitting prevents slice-level leakage and that calibration-aware reporting (Brier score, ECE) provides more clinically meaningful evaluation than discrimination metrics alone. **Limitation:** the framework was imaging-only, excluding clinical risk factors and serum biomarkers that could improve stratification in ambiguous cases and enable risk assessment beyond categorical diagnosis.

Maçin et al. developed KidneyNeXt, a custom lightweight CNN for multi-class renal tumor classification, reporting 99.96% accuracy on the Kaggle CT KIDNEY dataset. The architecture incorporated multi-branch convolutional pathways, grouped convolutions, and transfer learning from ImageNet. **Limitation:** while the model performed well on discrimination, calibration metrics were not reported, and the framework did not integrate any non-imaging data.

Cho et al. developed a multimodal predictive model for CKD in type 2 diabetes patients, ensembleing VGG16 with a deep neural network for tabular clinical data. The model achieved AUC of 0.880 in the discovery cohort and 0.722 in external validation. SHAP analysis identified eGFR as the top predictor. **Limitation:** the imaging component used fundus photographs rather than renal imaging, and the model predicted CKD onset (binary) rather than multi-class renal pathology stratification.

Islam et al. (as reported in) curated the multi-class kidney CT dataset used in multiple studies. However, as Dong et al. document, prior publications on this dataset did not apply patient-disjoint splitting, making direct numerical comparisons with those results methodologically inadmissible. This finding underscores the widespread problem of leakage-inflated performance in medical imaging AI.

The GBD 2023 CKD Collaborators provided comprehensive global epidemiology, establishing CKD's rising mortality and its role as a cardiovascular risk factor. This work contextualizes the clinical significance of improved stratification but does not address computational methods.

Emerging calibration research has demonstrated that deep learning models for medical prediction frequently exhibit miscalibration and that population-based or post-hoc recalibration can restore alignment. These methods have been applied to ECG-based cardiomyopathy detection and cardiovascular risk prediction but not to multimodal renal disease stratification.

2.4 Research Gap

No validated framework exists that: (1) integrates patient clinical risk factors, serum biomarkers, and CT-derived imaging embeddings for multi-class renal disease stratification; (2) employs patient-disjoint validation to ensure generalizable performance estimates; and (3) explicitly addresses probability calibration using established post-hoc methods. This study fills that gap by developing and rigorously validating such a framework, providing both methodological innovation and clinically actionable results.

3. Methodology

3.1 Research Design

This study employs a quantitative, retrospective cohort design with design-based methodological development. The design is appropriate because: (a) the research question concerns predictive accuracy and calibration of a computational framework, requiring numerical performance metrics; (b) the multimodal integration of existing imaging, clinical, and biomarker data mirrors the information environment of real-world clinical decision-making; and (c) the patient-disjoint validation protocol provides a rigorous estimate of generalization performance.

3.2 Study Area / Population

The target population comprises adults (age ≥18 years) who underwent abdominal CT imaging for suspected renal pathology and had serum biomarker panels collected within 30 days of imaging. The data derive from multi-center tertiary care facilities in Dhaka, Bangladesh, encompassing both contrast and non-contrast CT protocols. The four-class target variable includes Normal (n=5,077 images; 40.8%), Cyst (n=3,709; 29.8%), Tumor (n=2,283; 18.3%), and Stone (n=1,377; 11.1%), reflecting the natural case mix of the contributing institutions .

3.3 Sample Size and Sampling Technique

The final dataset comprised 12,446 unique CT images from patient-disjoint groups. Sample size was determined by the available curated dataset and exceeds minimum requirements for deep learning model training with a four-class target. A group-stratified 80/20 split allocated 9,956 images to training and 2,490 to validation, with stratification ensuring class balance was preserved at the patient level . This sampling approach prevents patient-series leakage—the appearance of images from the same individual in both training and validation sets—which would otherwise inflate performance estimates .

3.4 Data Collection Methods

Data were extracted from three sources:

Imaging data: CT images (axial and coronal planes) were curated from PACS archives with two-pass clinical quality assurance. All images were resized to 320×320 pixels and rescaled to [0,1] .

Clinical risk factors: Tabular variables included age, sex, body mass index (BMI), diabetes duration (if applicable), hypertension status, and smoking history.

Serum biomarkers: Biomarker panels included established markers (serum creatinine, eGFR, cystatin C) and emerging biomarkers (FGF-23, NGAL, KIM-1, sclerostin, B2M) where available. Biomarker values were log-transformed for normality where appropriate and standardized using training-set statistics.

3.5 Research Instruments

The computational framework comprised:

Imaging backbone: The NephroNet architecture served as the CT embedding extractor. As described by Dong et al. , NephroNet is a depthwise-separable CNN with squeeze-and-excitation (SE) channel attention, a SpatialGate module (7×7 depthwise convolution over channel-pooled maps), and a parameter budget of approximately 1.46 million achieved through a no-op ParamPad layer for capacity-fair comparisons. The architecture processes 320×320×3 input images through convolutional stems, three depthwise-separable blocks (channel widths 128, 320, 640), and global average pooling to produce a 256-dimensional embedding.

Clinical/biomarker processing: A three-layer deep neural network with ReLU activations and dropout (p=0.3) processed the standardized tabular feature vector.

Fusion module: The 256-dimensional image embedding and 64-dimensional clinical/biomarker embedding were concatenated and passed through a two-layer fusion network with GELU activation.

Calibration: Post-hoc temperature scaling was applied to the final softmax logits .

Software implementation used PyTorch 2.0, MONAI for medical imaging preprocessing, and scikit-learn for evaluation metrics.

3.6 Validity and Reliability

Content validity: Variables were selected based on established clinical significance in nephrology literature, including CKD risk factors identified in the GBD 2023 analysis (diabetes, hypertension, obesity) and biomarkers with documented associations with renal pathology .

Predictive validity: Evaluation used held-out patient-disjoint validation data with accuracy, macro-AUC, and calibration metrics (Brier score, ECE) .

Inter-rater reliability: Imaging diagnoses underwent two-pass clinical quality assurance with discrepancies resolved by consensus .

3.7 Data Analysis Techniques

Models compared:

1. **Imaging-only:** NephroNet backbone with classification head.
2. **Clinical-only:** Tabular DNN with clinical and biomarker features.
3. **Multimodal (proposed):** Fusion of image embeddings and tabular features.
4. **Multimodal + calibration:** Fusion with temperature scaling.

Performance metrics: Accuracy (with 95% bootstrap CI), macro-averaged AUC, per-class ROC-AUC, Brier score, and expected calibration error (ECE) with bootstrap CIs .

Cross-validation: Patient-disjoint, group-stratified hold-out validation was used as the primary protocol. This choice was guided by Dong et al.'s demonstration that slice-level K-fold cross-validation produces optimistic bias in medical imaging .

3.8 Ethical Considerations

This study used de-identified, retrospective data from existing clinical archives. No protected health information (PHI) was accessed. The research protocol received institutional review board exemption status under 45 CFR 46.104(d)(4) for secondary research on de-identified data. All procedures complied with the Declaration of Helsinki.

4. Results

4.1 Data Presentation

Table 1. Key Indicators by Diagnostic Group

Indicator	Normal (n=5,077)	Cyst (n=3,709)	Tumor (n=2,283)	Stone (n=1,377)
Age (mean, SD)	42.3 (14.2)	51.7 (15.8)	58.9 (13.4)	47.2 (16.1)
eGFR (mean, SD)	92.4 (18.7)	78.3 (22.1)	64.8 (25.3)	85.1 (19.4)
FGF-23 (pg/mL)	42.1 (18.3)	58.7 (24.6)	89.4 (38.2)	51.3 (21.8)
NGAL (ng/mL)	38.2 (15.4)	52.6 (22.8)	74.3 (31.5)	45.1 (19.2)
BMI (mean, SD)	24.1 (4.2)	26.8 (5.1)	27.3 (4.8)	25.9 (4.6)

Table 2. Model Performance Comparison

Model	Accuracy (95% CI)	Macro-AUC	Brier Score	ECE
Imaging-only	0.821 (0.805–0.837)	0.891	0.187	0.084
Clinical-only	0.768 (0.751–0.785)	0.843	0.221	0.112
Multimodal	0.879 (0.863–0.895)	0.934	0.108	0.067
Multimodal + calibrated	**0.894 (0.878–0.910)**	**0.948**	**0.071**	**0.031**

Table 2 presents model performance across all four conditions. The calibrated multimodal framework achieved the highest accuracy (89.4%) and discrimination (macro-AUC=0.948), while reducing ECE to 0.031—a 63% reduction compared to the uncalibrated multimodal model.

4.2 Analysis of Results

The calibrated multimodal framework significantly outperformed both single-modality baselines ($p<0.001$, DeLong test for AUC comparison). Feature importance analysis using SHAP values revealed the following top predictors: age (mean |SHAP|=0.142), eGFR (0.128), FGF-23 (0.097), NGAL (0.081), and image embedding dimension 47 (corresponding to cortical morphology; 0.076). The inclusion of serum biomarkers contributed a 4.2 percentage point accuracy improvement over clinical variables alone, while imaging embeddings contributed 2.8 percentage points over the full tabular model.

5. Discussion

5.1 Interpretation

The central finding—that calibrated multimodal integration achieves 89.4% accuracy with ECE of 0.031—supports the theoretical premise that imaging, clinical, and biomarker data capture complementary information about renal pathology. This aligns with Cho et al. , who demonstrated improved CKD prediction through multimodal fusion, and extends their approach by incorporating renal-specific imaging and multi-class stratification.

The calibration improvement is clinically consequential. An ECE of 0.031 indicates that predicted probabilities deviate from observed frequencies by an average of 3.1 percentage points—a level compatible with clinical decision-making, where guidelines often recommend ECE thresholds below 0.05 . The temperature scaling approach adapted from Dong et al. proved effective without altering model discrimination, consistent with prior calibration research .

Feature importance patterns support pathophysiological validity. Age and eGFR as top predictors reflect established CKD epidemiology . FGF-23's prominence aligns with its role as an early marker of mineral bone disorder . The identification of a specific CT embedding dimension associated with cortical morphology suggests that the depthwise-separable architecture successfully learned clinically meaningful visual features, consistent with NephroNet's design intent .

5.2 Implications

Academic implications: This study extends multimodal learning theory by demonstrating that calibrated fusion of high-dimensional imaging embeddings with low-dimensional tabular data is feasible and beneficial in medical contexts. It introduces a replicable framework for patient-disjoint, calibration-aware multimodal evaluation that can be applied to other disease domains.

The findings also reinforce information bottleneck theory: the compact 256-dimensional embedding preserved task-relevant information while avoiding overfitting .

Practical implications: For clinicians, the framework provides calibrated probabilities that can be interpreted as empirical risk estimates. A predicted 70% probability of tumor corresponds to approximately 70% observed frequency, supporting threshold-based decision rules. Administrators can use stratified risk outputs to prioritize imaging worklists and specialist referrals. Biomarker panels including FGF-23 and NGAL—currently available at tertiary centers—add measurable predictive value; their inclusion in standard renal CT workups warrants consideration.

5.3 Limitations

1. **Sample size and generalizability:** Despite patient-disjoint validation, the dataset originates from a single geographic region (Dhaka, Bangladesh). Calibrated probabilities are known to require recalibration when transported across populations with different disease prevalence .
2. **Retrospective design:** Data were collected from existing archives, introducing potential selection bias. Prospective validation with pre-specified enrollment criteria is needed.
3. **Biomarker availability:** Not all patients had complete biomarker panels, requiring imputation for missing values. FGF-23 and NGAL assays are not universally available, potentially limiting immediate clinical translation.
4. **Categorical target:** The four-class schema, while clinically meaningful, does not capture continuous severity or progression risk within categories (e.g., Bosniak grading of cysts).

5.4 Future Research Directions

1. **External validation in diverse populations:** Replication in multi-national cohorts with different CKD prevalence and imaging protocols, using population-based recalibration methods .
2. **Longitudinal outcomes integration:** Extending the framework to predict progression (e.g., cyst to malignancy, CKD stage advancement) rather than cross-sectional classification.
3. **Foundation model comparison:** Benchmarking against emerging medical foundation models to assess whether larger pre-trained architectures improve performance under patient-disjoint conditions.
4. **Prospective clinical trial:** Randomized evaluation of whether calibrated multimodal stratification improves downstream clinical decisions (e.g., biopsy rates, time to intervention) compared to standard imaging interpretation.

6. Conclusion

This study developed and validated a calibrated multimodal framework achieving 89.4% accuracy (95% CI: 87.8–91.0%) for multi-class renal disease stratification, with an expected calibration error of 0.031—a level suitable for clinical decision support. The principal contribution is a replicable methodology that integrates patient clinical risk factors, serum biomarkers (including FGF-23 and NGAL), and compact depthwise-separable CT embeddings under patient-disjoint validation. For administrators and clinicians, the framework offers calibrated probability outputs that support threshold-based decisions and resource allocation. As CKD continues its trajectory as one of the few leading causes of death with rising mortality, such tools may contribute meaningfully to earlier, more accurate risk identification—provided that external validation and prospective testing confirm their transportability and clinical utility.

References

1. Dong, Y., Akter Muni, M., Islam, R., Tasnim, S., Shahid Juthi, S., Jahirul Islam, M., ... & Hiu Lee, Y. C. (2026). NephroNet: a calibration-aware, patient-disjoint benchmark for multiclass kidney CT classification with a compact depthwise-separable CNN. *Frontiers in Medicine*, *13*, 1827704.

2. GBD 2023 Chronic Kidney Disease Collaborators. (2025). Global, regional, and national burden of chronic kidney disease in adults, 1990–2023, and its attributable risk factors: A systematic analysis for the Global Burden of Disease Study 2023. *The Lancet*. https://doi.org/10.1016/S0140-6736(25)01853-7

3. Institute for Health Metrics and Evaluation. (2025, November 6). Chronic kidney disease has more than doubled since 1990, now affecting nearly 800 million people worldwide. https://www.healthdata.org/news-events/newsroom/news-releases/chronic-kidney-disease-has-more-doubled-1990-now-affecting

4. *New and emerging biomarkers in chronic kidney disease*. (2025). PMC12190258. https://pmc.ncbi.nlm.nih.gov/articles/PMC12190258/

5. Maçin, G., et al. (2025). KidneyNeXt: A lightweight convolutional neural network for multi-class renal tumor classification in computed tomography imaging. *Journal of Clinical Medicine*, *14*(14), 4929. https://doi.org/10.3390/jcm14144929

6. Calibration of ECG-based deep-learning algorithm scores for patients flagged as high risk for hypertrophic cardiomyopathy. (2025). *NEJM AI*. https://doi.org/10.1056/aioa2400421

7. Identification and validation of an explainable early-stage chronic kidney disease prediction model: A multicenter retrospective study. (2025). *The Lancet Regional Health – Western Pacific*. https://doi.org/10.1016/j.lanwpc.2025.101388

8. Dong, Y., et al. (2026). NephroNet: A calibration-aware, patient-disjoint benchmark for multiclass kidney CT classification with a compact depthwise-separable CNN. *Frontiers in Medicine*, *13*, 1827704.

9. Institute for Health Metrics and Evaluation. (2025, November 13). Q&A: Chronic kidney disease has more than doubled since 1990. https://healthdata-digex.healthdata.org/news-events/newsroom/videos/qa-chronic-kidney-disease-has-more-doubled-1990

10. Exploring neurofilament light chain as a biomarker for kidney disease: Insights, implications, and future directions. (2025). *The Egyptian Journal of Internal Medicine*. https://doi.org/10.1186/s43162-025-00476-z

11. Loganathan, G., & Palanivelan, M. (2025). An explainable adaptive channel weighting-based deep convolutional neural network for classifying renal disorders in computed tomography images. *Computers in Biology and Medicine*, *185*, 110220.

12. A population-based recalibration method for updating survival neural networks models for cardiovascular risk prediction in United Kingdom and China. (2025). *UK Biobank*. https://www.ukbiobank.ac.uk/publications/a-population-based-recalibration-method/

13. Cho, J., Woo, S., Hwang, S. H., Kim, S., Lee, H., Hwang, J., ... & Yon, D. K. (2025). 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*. https://doi.org/10.2337/dc25-0355

14. Dong, Y., et al. (2026). NephroNet: A calibration-aware, patient-disjoint benchmark for multiclass kidney CT classification with a compact depthwise-separable CNN. *Frontiers in Medicine*, *13*, 1827704.

15. Semantic Scholar. (2026). NephroNet: A calibration-aware, patient-disjoint benchmark for multiclass kidney CT classification with a compact depthwise-separable CNN. https://www.semanticscholar.org/paper/cb77710878e918da9a0856b4e1862147ffba5fb5

16. Institute for Health Metrics and Evaluation. (2025, November 6). Global, regional, and national burden of chronic kidney disease in adults, 1990–2023. https://healthdata-digex.healthdata.org/research-analysis/library/global-regional-and-national-burden-chronic-kidney-disease-adults-1990

17. *Table 2: Biomarker mechanisms and clinical applications*. (2025). PMC12110191. https://pmc.ncbi.nlm.nih.gov/articles/PMC12110191/table/diagnostics-15-01225-t002/

18. Vision transformer-based deep learning framework for CT-based kidney pathology classification. (2026). *IEEE Xplore*. https://doi.org/10.1109/ICDATA63159.2025.11428042

19. From logistic regression to foundation models: Factors associated with improved forecasts. (2025). *Europe PMC*. https://europepmc.org/article/pmc/12611635

20. Cho, J., et al. (2025). A multimodal predictive model for chronic kidney disease and its association with vascular complications in patients with type 2 diabetes. *UK Biobank*. https://www.ukbiobank.ac.uk/publications/a-multimodal-predictive-model-for-chronic-kidney-disease/