

Feasibility of chronic kidney disease prediction using bilateral kidney ultrasound and foundation model representations

Nghiem Vo

Chonnam National University

Anh-Tien Nguyen

Justus-Liebig-Universität Gießen

Thong N. H. Vo

Chonnam National University

Ga-Yeon Yang

Chonnam National University Medical School

Su Hyun Song

Chonnam National University Medical School and Hospital

Binh D. Le

Saint Paul Hospital

Luu-Ngoc Do

Chonnam National University Medical School and Hospital

Thuong-Khanh Tran

GITPS-Global IT Professional Services Finland Oy

Chang Seong Kim

Chonnam National University Medical School and Hospital

Ilwoo Park

`ipark@jnu.ac.kr`

Chonnam National University Medical School and Hospital

Anne-Christin Hauschild

Justus-Liebig-Universität Gießen

Article

Keywords: Chronic Kidney Disease, Ultrasonography, Foundation Model, Deep Learning, Glomerular Filtration Rate, Albuminuria.

Posted Date: August 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10161117/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Feasibility of chronic kidney disease prediction using bilateral kidney ultrasound and foundation model representations

Nghiem Vo^{1,+}, Anh-Tien Nguyen^{2,3,+}, Thong N. H. Vo^{1,4}, Ga-Yeon Yang⁵, Su Hyun Song⁶, Binh D. Le⁷, Luu-Ngoc Do⁸, Thuong-Khanh Tran⁹, Chang Seong Kim^{6,*}, Ilwoo Park^{1,10,11,*}, and Anne-Christin Hauschild^{2,*}

¹Department of Artificial Intelligence Convergence, Chonnam National University, Korea

²Institute for Predictive Deep Learning for Medicine and Healthcare, Giessen University, Germany

³Department of Medical Informatics, University Medicine Gottingen, Germany

⁴Department of Radiology, Cho Ray Hospital, Viet Nam

⁵Department of Medicine, Chonnam National University Medical School, Korea

⁶Department of Internal Medicine, Chonnam National University Medical School and Hospital, Korea

⁷Department of Urology, Saint Paul Hospital, Viet Nam

⁸Chonnam National University, Korea

⁹GITPS-Global IT Professional Services Finland Oy

¹⁰Department of Radiology, Chonnam National University Medical School and Hospital, Korea

¹¹Department of Data Science, Chonnam National University, Korea

*laminion@hanmail.net, ipark@jnu.ac.kr, anne-christin.hauschild@uni-giessen.de

⁺These authors contributed equally to this work

ABSTRACT

Chronic kidney disease (CKD) is a major global health burden, yet early diagnosis remains challenging and conventional kidney ultrasonography is limited by operator dependence and low sensitivity to subtle textural changes. We investigated the feasibility of an ultrasound foundation model for CKD classification from bilateral kidney ultrasound images, with reference labels defined according to the Kidney Disease: Improving Global Outcomes (KDIGO) 2024 guidelines. In this retrospective single-center study, 13,245 kidney ultrasound images from 1,484 patients (1,128 with CKD and 356 without) were analyzed. The Universal Ultrasound Foundation Model (USFM) was compared with five fine-tuned deep learning architectures (VGG16, ResNet-50, DenseNet-121, EfficientNet-B0, and Vision Transformer) over 10 random seeds, with performance assessed at the sample and examination levels. USFM outperformed the baseline models and showed greater robustness across repeated experiments. At the examination level, it achieved a balanced accuracy of $74.49 \pm 4.63\%$, an F1-score of $73.59 \pm 3.60\%$, and an area under the receiver operating characteristic curve of $84.80 \pm 1.98\%$, with comparable results at the sample level. These findings demonstrate the potential of ultrasound foundation models to support CKD diagnosis from bilateral kidney ultrasound images.

Keywords: Chronic Kidney Disease, Ultrasonography, Foundation Model, Deep Learning, Glomerular Filtration Rate, Albuminuria.

1 Introduction

Chronic kidney disease (CKD) is a major global public health problem that affects a substantial proportion of the adult population worldwide, with reported prevalence ranging from about 6.8% to 14.4% across countries¹. In Korea, national data shows a prevalence of around 6.5% in 2020 and a persistent burden in recent years, particularly among older adults^{2,3}. CKD is especially concerning because it is associated with increased mortality, impaired health-related quality of life and considerable healthcare costs^{4,5}.

Kidney Disease: Improving Global Outcomes (KDIGO)⁶ defines CKD as abnormalities of kidney structure or function present for at least three months, with implications for health. This definition encompasses a broad spectrum of diagnostic criteria rather than relying on a single laboratory measurement. In clinical practice, CKD is often inferred from a sustained reduction in glomerular filtration rate (GFR), most commonly operationalized as an estimated GFR (eGFR) below 60 mL/min/1.73 m² for at least 3 months. However, KDIGO also recognizes persistent markers of kidney damage even when GFR is preserved, including albuminuria, histological evidence of kidney disease on biopsy, structural abnormalities detected by imaging and a history of kidney transplantation, among others. Together, these criteria highlight that CKD diagnosis requires integration of both functional and structural evidence over time.

Early identification of CKD is clinically important because timely intervention may delay disease progression and reduce the risks of complications, end-stage kidney disease and death^{6,7}. However, early diagnosis remains challenging. Serum creatinine and eGFR are widely used in routine practice, but may remain within the normal range until kidney function is substantially impaired and can also be influenced by non-renal factors such as age, sex and muscle mass^{8,9}. Albuminuria is an important marker of kidney damage, but urinalysis can be affected by transient conditions and requires repeated assessment to confirm persistence⁶. Kidney biopsy provides detailed histopathologic information, but it is invasive and not suitable as a routine screening tool¹⁰. In light of this, kidney ultrasonography is a non-invasive, radiation-free, and readily available imaging modality¹¹ that provides rapid assessment of kidney size, cortical echogenicity, hydronephrosis, and structural abnormalities. Nevertheless, conventional kidney ultrasonography has important limitations in CKD assessment, including limited sensitivity to subtle textural changes and early-stage disease, as well as operator dependence and inter-observer variability in the evaluation of renal size, cortical echogenicity, and structural abnormalities, which may affect CKD classification¹².

In this context, artificial intelligence has considerable potential to improve the analysis of kidney ultrasound images by capturing imaging patterns that may not be readily discernible through conventional visual assessment¹³. However, most existing AI studies for ultrasound-based CKD detection have relied on a single metric, most commonly eGFR, to define CKD labels^{14,15}. Such an approach does not fully reflect the breadth of the KDIGO framework and may therefore limit the clinical relevance of the resulting models. More recently, foundation models have been increasingly explored in medical imaging because they can learn generalizable visual representations from large-scale data and adapt effectively to downstream tasks with relatively limited task-specific training data^{16,17}. In particular, ultrasound foundation models offer the opportunity to leverage representations tailored to the unique characteristics of ultrasound imaging¹⁸. Therefore, this study aims to investigate the feasibility of an ultrasound foundation model for CKD classification using bilateral kidney ultrasound images in accordance with the KDIGO 2024 criteria, and to compare its performance against conventional deep learning architectures. To our knowledge, this is the first study to apply an ultrasound foundation model for classifying CKD.

2 Material and method

2.1 Study design

This retrospective single-center study was conducted to develop and internally validate a CKD prediction model based on an ultrasound foundation model using kidney ultrasound images. All data were obtained from the Nephrology Department of Chonnam National University Hospital between January 2020 and December 2021. The study protocol was approved by the institutional review board of Chonnam National University Hospital (CNUH-2025-443). All methods were performed in accordance with the relevant guidelines and regulations. The requirement for informed consent was waived.

Patients who underwent kidney ultrasound examination and had corresponding laboratory data available in the institutional records were considered for inclusion. Ultrasound images and related laboratory results were collected from the institutional Picture Archiving and Communication System (PACS) and electronic medical records. A total of 1,703 patients corresponding to 1,752 ultrasound examinations were initially identified. Because a single patient could undergo multiple ultrasound examinations and each examination could include multiple images for each kidney, the dataset was organized hierarchically at the patient, examination, and image levels.

Several exclusion criteria were applied. Specifically, 123 examinations from kidney transplant recipients, 91 examinations for which the interval between the ultrasound examination and the corresponding laboratory results exceeded 3 months, and 30 examinations without left or right kidney images were excluded. After these exclusions, the final dataset comprised 1,484 patients, 1,508 examinations and 13,245 kidney ultrasound images. The dataset was divided into training, validation and test sets, accounting for 60%, 20% and 20% of the total number of examinations, respectively. To prevent data leakage, all images and examinations from the same patient were strictly assigned to the same dataset split, ensuring patient-level separation across training, validation and testing. A flowchart of the patient selection process is presented in [Figure 1](#).

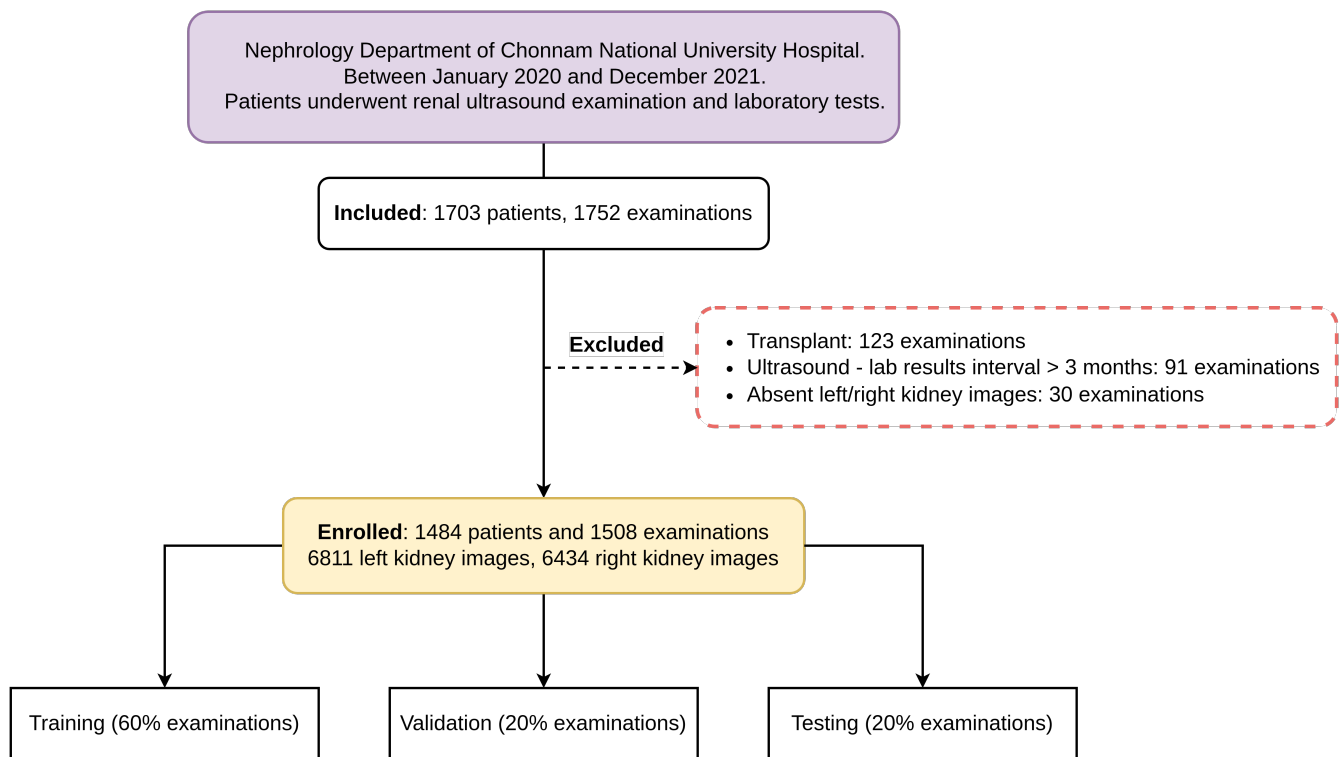


Figure 1. Flowchart of patient selection.

2.2 Reference standard for CKD classification

The reference standard for CKD classification was defined to align as closely as feasible with the KDIGO 2024 definition of CKD. According to KDIGO 2024, CKD is defined by abnormalities of kidney structure or function present for at least 3 months, with implications for health. Because not all KDIGO 2024 diagnostic components were consistently available in this retrospective dataset, the ground-truth labels were operationalized using the subset of KDIGO 2024-concordant criteria available in our cohort. Specifically, these criteria included eGFR, calculated using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) creatinine equation¹⁹; albuminuria was assessed using the urine albumin-to-creatinine ratio (UACR)

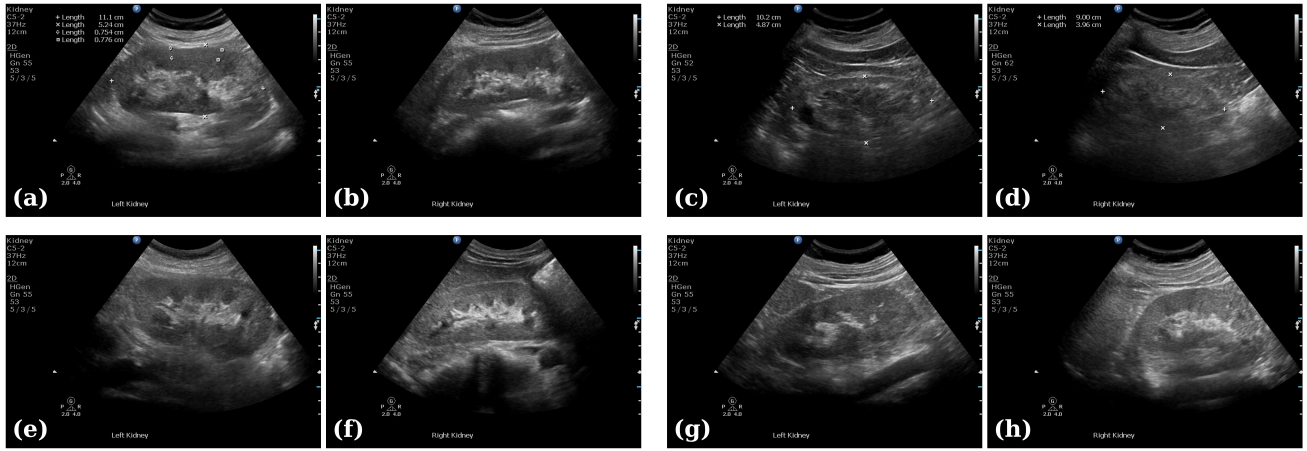


Figure 2. Representative kidney ultrasound images obtained from different patients in our dataset. Panels (a) and (b) show the left and right kidneys of a non-CKD patient with an eGFR of approximately 100 mL/min/1.73m². Panels (c) and (d) show the left and right kidneys of a CKD patient with an eGFR of approximately 7 mL/min/1.73m². Panels (e) and (f) present the left and right kidneys of another CKD patient with an eGFR of approximately 25 mL/min/1.73m². Panels (g) and (h) show the left and right kidneys of a CKD patient with a preserved eGFR of 97 mL/min/1.73m², but with an urine albumin-to-creatinine ratio fluctuating around 600 mg/g.

and histologic evidence of kidney disease when available. Chronicity was determined using longitudinal clinical data over a minimum interval of 3 months.

Because laboratory markers such as serum creatinine, eGFR, and UACR reflect overall renal function, CKD status could not be assigned separately to each kidney. Even when functional asymmetry between the two kidneys might have been present, the available laboratory results did not permit reliable attribution of renal function to either kidney individually. Therefore, all ultrasound images acquired during the same examination were assigned the same ground-truth label.

2.3 Ultrasound acquisition and data preprocessing

Renal ultrasound imaging was conducted using a Philips iU22 ultrasound system (Philips Ultrasound, Bothell, WA, USA). Examinations were performed by board-certified nephrologists with support from a trained sonographer using a low-frequency curvilinear transducer (2-5 MHz). Standard renal or abdominal imaging presets were applied, and acquisition parameters such as imaging depth, focal zone, and overall gain were adjusted to optimize visualization of the kidneys. Patients were positioned predominantly in the lateral decubitus position, with the liver and spleen used as acoustic windows for the left and right kidneys, respectively. Standard longitudinal and transverse renal ultrasound images were acquired during routine clinical practice.

For model development, only midsagittal images of the kidneys were used. Ultrasound images (600 x 800 pixels) contained patient identifiers and acquisition metadata in the header region. To ensure anonymization and prevent information leakage, the header was removed by cropping, resulting in images of 540 x 800 pixels. This process eliminated identifiable information while preserving the renal field of view. Examples of preprocessed images are shown in [Figure 2](#).

2.4 Statistical analysis of patient characteristics

For statistical analysis of patient characteristics, each patient was represented by a single examination. In patients with multiple examinations, the most recent examination was used for the analysis to avoid duplicate counting. The eGFR value was obtained from the laboratory test closest to the ultrasound examination date, representing a single time point. Continuous variables were presented as median and interquartile range (IQR), and compared between CKD and non-CKD groups using the Mann-Whitney U test, while categorical variables were compared between groups using the chi-square test. Statistical significance was defined as a two-sided p -value < 0.05.

2.5 Model development

In this study, we employed Monte Carlo cross-validation²⁰ (MCCV) on the dataset collected from Chonnam National University Hospital (CNUH-CKD). To provide reliable and fairness comparison, we randomly split the CNUH-CKD dataset into training and testing sets multiple times using 10 different random seeds, as described in **Appendix A. 1**. For each seed, the model was independently trained on the corresponding training set and evaluated on the held-out testing set.

We adopted a pooled unilateral training strategy. Each examination in our dataset includes ultrasound images of both kidneys, and we treated each kidney image independently, pooling left and right kidney images into a single training set without distinguishing laterality. This approach used the contralateral kidney images as a form of natural data augmentation, thus doubling the number of training instances and potentially encouraging the model to learn generalized, side-invariant renal representations.

2.5.1 Baseline models

To establish a comparative benchmark for CKD classification, we trained four widely adopted deep convolutional neural network (CNN) architectures, namely ResNet-50²¹, DenseNet-121²², EfficientNet-B0²³, VGG16²⁴ alongside ViT16²⁵, a transformer-based model. All baseline models were initialized with pre-trained weights on the ImageNet²⁶ dataset. For each baseline architecture, the original fully connected classification head was replaced with a custom layer matching the number of CKD classes in our study. All models were fully fine-tuned using the Adam optimizer for CNN architectures and AdamW for ViT16, with the learning rates of 3×10^{-4} for ResNet-50, DenseNet-121 and 1×10^{-4} for EfficientNet-B0, VGG16. For ViT16, a warmup learning rate of 1×10^{-5} was used before transitioning to a base learning rate of 1×10^{-4} .

2.5.2 Ultrasound foundation model

To evaluate the efficacy of ultrasound-specific foundation models for CKD, we applied the Universal Ultrasound Foundation Model¹⁸ (USFM) to the CNUH-CKD dataset. The model was utilized as a robust feature extractor, where the foundational ultrasound encoder processed the input scans to generate high-dimensional embeddings, which were then mapped to CKD predictions via a task-specific classification head. By training on the mixed pool of left and right kidneys, the USFM could bridge its universal ultrasound representations with the specific variations present in our clinical data.

2.6 Performance evaluation

2.6.1 Performance metrics

The models were trained and validated on individual kidney images as independent samples; no left-right pairing was performed during training or validation. Pairing was applied only at test time as a late-fusion strategy to integrate integrate bilateral kidney information into a single prediction per sample.

Test-time pair construction. Within each examination in the test set, left-right image pairs were constructed as follows. For an examination containing n_L left and n_R right kidney images, $\min(n_L, n_R)$ pairs were formed by randomly pairing each image on the smaller side, without replacement, with a distinct image on the opposite side; any surplus images on the larger side were not used. This ensured that no image contributed to more than one test sample. To assess the robustness of the evaluation to this random pairing, pair construction was repeated independently for each of the 10 Monte Carlo cross-validation runs, and all reported metrics represent the mean and standard deviation across these runs.

Sample-level prediction. For each test sample i , the trained model produced a class-probability vector for the left kidney image and another for the right kidney image. Let $\mathbf{p}_{i,L} \in \mathbb{R}^C$ and $\mathbf{p}_{i,R} \in \mathbb{R}^C$ denote these two vectors, where C is the number of classes. The sample-level probability vector was obtained by averaging the two vectors elementwise:

$$\hat{\mathbf{p}}_i^{\text{sample}} = \frac{1}{2} (\mathbf{p}_{i,L} + \mathbf{p}_{i,R}). \quad (1)$$

The sample-level predicted class was then obtained as the argmax over classes of this averaged probability vector:

$$\hat{y}_i^{\text{sample}} = \arg \max_{c \in \{1, \dots, C\}} \hat{p}_{i,c}^{\text{sample}}. \quad (2)$$

Examination-level prediction. Because each examination may contain multiple left-right pairs, predictions were further aggregated across all samples belonging to the same examination. For each examination e , let $S_e = \{i \mid \text{exam}(i) = e\}$ denote the set of sample indices belonging to examination e . The examination-level probability vector was obtained by averaging the sample-level probability vectors within S_e :

$$\hat{\mathbf{p}}_e^{\text{exam}} = \frac{1}{|S_e|} \sum_{i \in S_e} \hat{\mathbf{p}}_i^{\text{sample}}. \quad (3)$$

The examination-level predicted class was then obtained as:

$$\hat{y}_e^{\text{exam}} = \arg \max_{c \in \{1, \dots, C\}} \hat{p}_{e,c}^{\text{exam}}. \quad (4)$$

Table 1. Baseline characteristics of CKD and non-CKD cohorts. Age and eGFR are presented as median [interquartile range], while Sex is presented as counts (percentages).

Variable	CKD (n=1128)	non-CKD (n=356)	p-value
Age (years)	66 [54-76]	58 [44-67]	< 0.001
Sex, n (%)			< 0.001
Male	720 (63.8%)	190 (53.4%)	
Female	408 (36.2%)	166 (46.6%)	
eGFR (mL/min/1.73 m ²)	37 [19-57]	95.5 [78-106]	< 0.001

Abbreviations: eGFR, estimated glomerular filtration rate.

Evaluation metrics. Model performance was reported at two levels: the sample level, where each sample consisted of one paired left-right image from the same examination, and the examination level, where predictions were aggregated across all pairs belonging to the same examination. At both levels, each model was evaluated using five metrics: balanced accuracy (B-ACC), F1-score (F1), area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. All metrics are reported as the mean and standard deviation across the 10 Monte Carlo cross-validation runs.

2.6.2 Decision Curve Analysis

Decision curve analysis (DCA) was performed to assess the potential clinical utility of each model across different threshold probabilities. Unlike conventional discrimination metrics such as AUC or balanced accuracy, DCA evaluates whether using a prediction model would provide a higher clinical benefit than default strategies, such as treating all patients or treating no patients. The net benefit was calculated across a range of threshold probabilities, where the threshold probability represents the level of predicted risk at which a clinical decision would be made.

For each model, DCA was conducted at both the sample and examination levels. At the sample level, the predicted probability was obtained from the late-fusion probability vector generated by averaging the left and right kidney predictions. At the examination level, the predicted probability was obtained by averaging all sample-level probability vectors belonging to the same examination. The resulting net benefit curves were compared with two reference strategies: treating no patients, which has a net benefit of zero, and treating all patients, which assumes that all cases are classified as positive.

3 Results

3.1 Patient characteristics

The cohort comprised 910 male patients (61.3%) and 574 female patients (38.7%), with a median age of 64 [52–75] years and a median eGFR of 47 [23-82] mL/min/1.73 m². According to the reference standard, 1128 patients (76.0%) were classified as having CKD while 356 patients (24.0%) were classified as non-CKD. With regard to patients who underwent multiple examinations, no patient exhibited discordant ground truth labels across different examinations.

Baseline characteristics of the CKD and non-CKD cohorts are presented in [Table 1](#). Patients in the CKD group were significantly older than those in the non-CKD group (median [IQR]: 66 [54–76] vs 58 [44–67] years, $p < 0.001$). eGFR was markedly lower in the CKD group compared to the non-CKD group (median [IQR]: 37 [19–57] vs 95.5 [78–106] mL/min/1.73m², $p < 0.001$). A significant difference in sex distribution was also observed between groups ($p < 0.001$).

3.2 Sample-level performance

As shown in [Table 2](#), USFM outperformed all baseline models in most evaluation metrics, achieving the highest B-ACC (73.51 ± 3.68%), F1 (72.78 ± 2.89%), AUC (83.76 ± 2.12%) and specificity (61.88 ± 9.26%). VGG16 ranked second in B-ACC (71.78 ± 4.21%), F1 (70.84 ± 3.22%), AUC (82.22 ± 1.53%), and specificity (59.54 ± 12.82%), albeit with a substantially larger parameter count of 134.3M compared to 39K for USFM. EfficientNet-B0 recorded the highest sensitivity (89.85 ± 1.66%) but the lowest specificity (44.02 ± 6.56%) and B-ACC (66.93 ± 2.87%), reflecting a tendency to over-predict the positive class. As visualized in [Figure 3](#) (a), the USFM ROC curve dominates the upper-left quadrant with the highest mean AUC across 10 runs.

3.3 Examination-level performance

[Table 3](#) presents the examination-level evaluation results, where consistent patterns were observed across models. USFM obtained the highest B-ACC (74.49 ± 4.63%), F1 (73.59 ± 3.60%), AUC (84.80 ± 1.98%) and specificity (63.20 ± 10.78%), with VGG16 ranking second in most metrics, while ViT16 ranked second in F1-score and sensitivity. EFFICIENTNET-B0

Table 2. Comparison of performance metrics (%) at the sample level across 10 Monte Carlo cross-validation runs. **Bold:** best results, underline: second-best results.

Model	Params	B-ACC	F1	AUC	Sensitivity	Specificity
ResNet-50 ²¹	23.5M	69.74 $\pm$ 2.40	69.92 $\pm$ 1.65	80.19 $\pm$ 1.63	86.29 $\pm$ 3.26	53.19 $\pm$ 7.66
DenseNet-121 ²²	6.9M	69.94 $\pm$ 2.56	70.29 $\pm$ 1.68	81.53 $\pm$ 1.50	87.29 $\pm$ 4.81	52.58 $\pm$ 9.24
EfficientNet-B0 ²³	4M	66.93 $\pm$ 2.87	68.22 $\pm$ 2.75	80.37 $\pm$ 2.01	89.85 $\pm$ 1.66	44.02 $\pm$ 6.56
VGG16 ²⁴	134.3M	<u>71.78 $\pm$ 4.21</u>	<u>70.84 $\pm$ 3.22</u>	<u>82.22 $\pm$ 1.53</u>	84.02 $\pm$ 5.03	<u>59.54 $\pm$ 12.82</u>
ViT16 ²⁵	85.8M	70.47 $\pm$ 3.08	70.64 $\pm$ 2.16	81.77 $\pm$ 2.01	86.67 $\pm$ 3.86	54.27 $\pm$ 9.46
USFM ¹⁸	39K	73.51 $\pm$ 3.68	72.78 $\pm$ 2.89	83.76 $\pm$ 2.12	85.13 $\pm$ 3.44	61.88 $\pm$ 9.26

Abbreviations: B-ACC, balanced accuracy; F1, F1-score; AUC, area under the receiver operating characteristic curve.

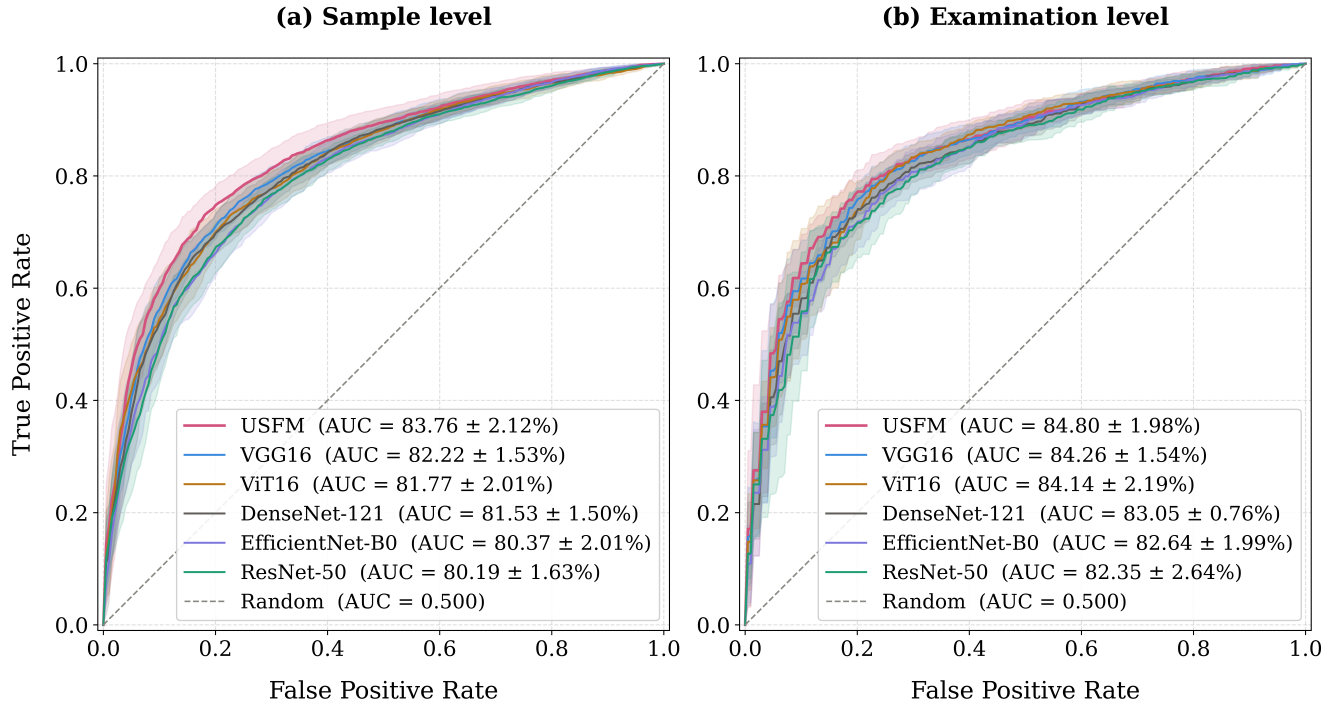


Figure 3. Receiver Operating Characteristic (ROC) curves for CKD classification on the CNUH-CKD dataset across models at the (a) **Sample level** and (b) **Examination level**. Solid lines represent the mean performance across 10 random seeds, while shaded regions indicate the standard deviation.

Table 3. Comparison of performance metrics (%) at the examination level across 10 Monte Carlo cross-validation runs. **Bold:** best results, underline: second-best results.

Model	Params	B-ACC	F1	AUC	Sensitivity	Specificity
ResNet-50 ²¹	23.5M	70.42 ± 3.37	70.70 ± 2.55	82.35 ± 2.64	87.42 ± 2.86	53.42 ± 8.78
DenseNet-121 ²²	6.9M	70.47 ± 3.95	70.79 ± 2.74	83.05 ± 0.76	87.92 ± 4.15	53.02 ± 11.13
EfficientNet-B0 ²³	4M	68.59 ± 2.89	70.15 ± 2.08	82.64 ± 1.99	91.17 ± 2.22	46.02 ± 6.34
VGG16 ²⁴	134.3M	<u>73.14 ± 4.51</u>	72.31 ± 3.28	<u>84.26 ± 1.54</u>	85.84 ± 4.74	<u>60.45 ± 13.04</u>
ViT16 ²⁵	85.8M	71.82 ± 3.95	<u>72.33 ± 3.21</u>	84.14 ± 2.19	<u>89.04 ± 3.41</u>	54.61 ± 10.52
USFM ¹⁸	39K	74.49 ± 4.63	73.59 ± 3.60	84.80 ± 1.98	85.79 ± 3.38	63.20 ± 10.78

Abbreviations: B-ACC, balanced accuracy; F1, F1-score; AUC, area under the receiver operating characteristic curve.

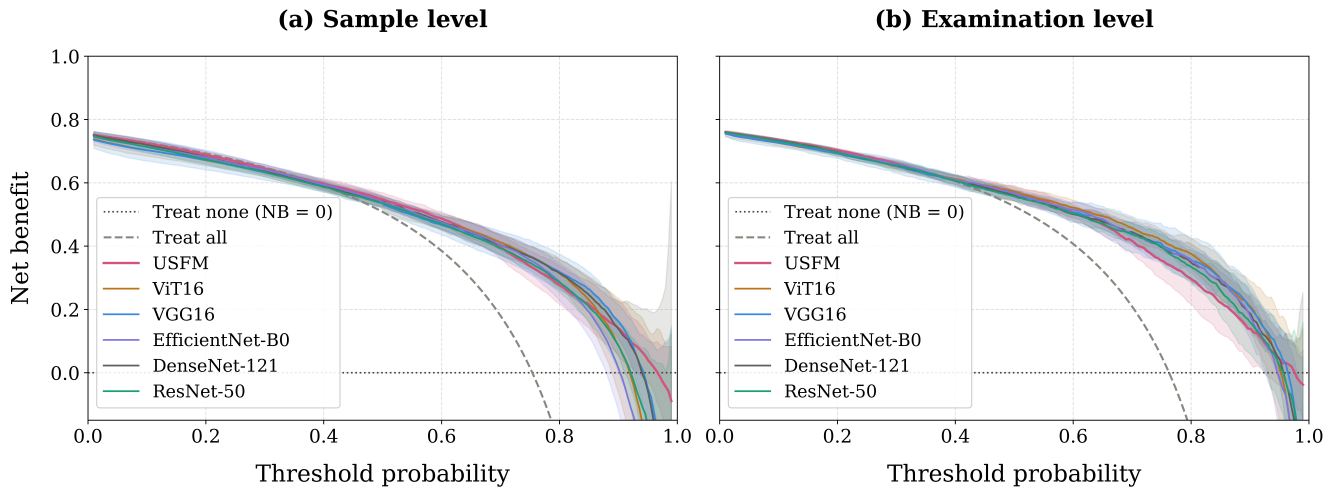


Figure 4. Decision curve analysis of the evaluated models at the (a) **Sample level** and (b) **Examination level**. Net benefit was calculated across a range of threshold probabilities and compared with two reference strategies: Treat all and Treat none. Shaded regions represent variability across 10 Monte Carlo cross-validation runs.

similarly recorded the highest sensitivity ($91.17 \pm 2.22\%$) alongside the lowest specificity ($46.02 \pm 6.34\%$). The AUC curves for all models at the examination level are visualized in Figure 3 (b).

Notably, performance metrics across all models generally improved from sample-level to examination-level evaluation. This trend suggests that averaging bilateral predictions provides a more stable assessment by reducing the impact of image-level variability inherent to ultrasound acquisition. At the examination level, USFM maintained a sensitivity of 85.79% while achieving the highest specificity among all evaluated models, suggesting a balanced trade-off between minimizing false positives and retaining true CKD cases.

3.4 Decision curve analysis

Figure 4 presents the decision curve analysis of all models at both the sample and examination levels. Across both levels, all evaluated models achieved positive net benefit over a wide range of threshold probabilities, consistently outperforming the *Treat none* strategy and surpassing the *Treat all* strategy at thresholds above approximately 0.5. At the sample level, the net benefit curves of all models were closely clustered across low-to-moderate thresholds, indicating comparable clinical utility. At the examination level, a similar pattern was observed, although the curves showed greater separation at higher thresholds, where USFM maintained a competitive net benefit compared to other models.

4 Discussion

Early and accurate diagnosis of CKD remains challenging due to its silent progression and the complexity of current diagnostic criteria⁶. Kidney ultrasound has been increasingly utilized for CKD evaluation owing to its non-invasive nature and wide clinical availability¹⁰, as it captures morphological characteristics, such as kidney size and cortical echogenicity, which have been shown to correlate with eGFR in patients with CKD²⁷. However, differentiating CKD from non-CKD based on ultrasound

images has been reported to be challenging even for experienced radiologists and nephrologists²⁸, highlighting the potential role of AI in enhancing diagnostic accuracy. Although recent studies have explored AI-based approaches for CKD classification from renal ultrasound, several shortcomings remain. Many studies do not adhere to the KDIGO definition, relying solely on eGFR or lacking longitudinal confirmation^{14,15}. Others follow the diagnostic criteria but assess only a single kidney^{29,30}, thereby reducing the comprehensiveness of the evaluation and reducing their applicability in clinical practice. To address these gaps, we applied an ultrasound foundation model to classify CKD using bilateral kidney information. Specifically, CKD diagnosis incorporated multiple clinical parameters, including eGFR, UACR, and kidney ultrasonography findings. The presence of chronic kidney dysfunction was confirmed using repeated laboratory measurements obtained over a period of at least 3 months, including follow-up serum creatinine, eGFR, and albuminuria assessments, in accordance with current CKD diagnostic guidelines. To the best of our knowledge, this is the first study to investigate the utility of an ultrasound foundation model for CKD classification.

The experimental results demonstrate that USFM achieved the strongest overall performance in most evaluation metrics. At the sample level, USFM obtained the highest B-ACC, F1-score, AUC, and specificity, while maintaining competitive sensitivity. Similar trends were observed at the examination level, where USFM again achieved the highest B-ACC, F1-score, AUC, and specificity. This was further supported by the class activation maps in **Appendix A. 4**, where USFM demonstrated well-localized attention toward the renal cortex compared with other models. These findings indicate that ultrasound-specific foundation representations are particularly effective for CKD classification and can provide a better balance between sensitivity and specificity compared with the conventional CNN-based and ImageNet-pretrained ViT16. Importantly, USFM achieved competitive performance with only 39K trainable parameters, substantially fewer than large deep learning baselines such as VGG16 and ViT16. This suggests that the proposed approach is not only accurate but also computationally efficient. In addition, the examination-level evaluation improved overall diagnostic performance compared with the sample-level evaluation, indicating that aggregating predictions across bilateral kidney images via averaged softmax probabilities provides a more stable and representative assessment of renal status, reducing the impact of image-level variability inherent to ultrasound acquisition.

The utility of the proposed approach is further illustrated through representative cases shown in **Figure 2**. Panels (a) and (b) correspond to a non-CKD patient with an eGFR of approximately 100 mL/min/1.73 m², presenting with normal renal morphology on bilateral ultrasound. USFM correctly classified this patient as non-CKD, with an averaged bilateral prediction probability of 0.78 for the non-CKD class. Panels (c) and (d) depict a typical CKD case with a markedly reduced eGFR of approximately 7 mL/min/1.73 m², in which the ultrasound images demonstrated overt morphological changes consistent with advanced renal disease, including reduced kidney size, thinning of the renal cortex, increased parenchymal echogenicity and an irregular renal contour. USFM correctly identified this patient as CKD with high confidence, yielding a prediction probability of 0.90 for the CKD class. These two cases suggest that USFM performs reliably in straightforward presentations where clinical and imaging findings are concordant.

Beyond these typical presentations, the model's behavior in more diagnostically challenging cases reveals both its strengths and limitations. Panels (e) and (f) correspond to a CKD patient with an eGFR of approximately 25 mL/min/1.73 m², yet the bilateral kidney ultrasound images remain morphologically preserved with near-normal features, including relatively maintained kidney size, fairly thick renal cortex, smooth and continuous renal surface, normal renal parenchymal echogenicity, and clear corticomedullary differentiation, closely resembling the appearance of a normal kidney on ultrasound. Despite this imaging ambiguity, USFM nonetheless assigned a correct CKD prediction, with a bilateral prediction probability of 0.69 for the CKD class, indicating that the model may capture subtle imaging patterns that are not readily apparent on visual assessment. In contrast, panels (g) and (h) represent a particularly atypical case in which the patient was diagnosed with CKD based on a persistently elevated UACR over more than 3 months, despite a high eGFR of 97 mL/min/1.73 m² and bilaterally preserved renal morphology similar to that observed in panels (e) and (f). USFM misclassified this patient as non-CKD, with an averaged bilateral prediction probability of 0.46 for the CKD class. The diagnosis in this case was primarily supported by urinary ACR rather than eGFR, which may have a different subtle pattern and contributed to the model's misclassification.

While these findings demonstrate the promise of ultrasound foundation models for CKD classification, several limitations of this study should be acknowledged. First, both training and evaluation were performed on a single-center dataset, and the absence of external validation limits the assessment of the model's generalizability across different patient populations, ultrasound devices, and imaging protocols. Second, the retrospective design may have introduced selection bias due to the reliance on available and complete clinical records. This bias may have been compounded by the fact that kidney ultrasonography is generally performed in patients with suspected kidney disease rather than in the general population, resulting in a study cohort enriched for CKD cases and potentially limiting the generalizability of the findings.

Overall, this study highlights the potential of ultrasound foundation models as promising tools for CKD classification from kidney ultrasound images. Future work should prioritize prospective multi-center validation across diverse institutions, ultrasound devices, and imaging protocols to further confirm the robustness and broad clinical utility of the proposed approach.

Data availability statement

The datasets generated and/or analysed during the current study are not publicly available due patient privacy and institutional regulations but are available from the corresponding author on reasonable request.

References

1. Cockwell, P. & Fisher, L.-A. The global burden of chronic kidney disease. *The lancet* **395**, 662–664 (2020).
2. Yoon, S.-Y. *et al.* National trends in the prevalence of chronic kidney disease among korean adults, 2007–2020. *Sci. reports* **13**, 5831 (2023).
3. Kim, Y. Trends in prevalence of chronic kidney disease, 2013–2023. *Public Heal. Wkly. Rep.* **18**, 478–479, DOI: [10.56786/PHWR.2025.18.10.3](https://doi.org/10.56786/PHWR.2025.18.10.3) (2025).
4. Go, A. S., Chertow, G. M., Fan, D., McCulloch, C. E. & Hsu, C.-y. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *New Engl. J. Medicine* **351**, 1296–1305 (2004).
5. Bikbov, B. *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* **395**, 709–733 (2020).
6. Stevens, P. E. *et al.* Kdigo 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney international* **105**, S117–S314 (2024).
7. Korsa, A., Tesfaye, W., Sud, K., Krass, I. & Castelino, R. L. Risk factor-based screening for early detection of chronic kidney disease in primary care settings: A systematic review. *Kidney Medicine* **7**, 100979 (2025).
8. Bufkin, K. B., Karim, Z. A. & Silva, J. Review of the limitations of current biomarkers in acute kidney injury clinical practices. *SAGE Open Medicine* **12**, 20503121241228446 (2024).
9. Liu, P. *et al.* Accounting for age in the definition of chronic kidney disease. *JAMA internal medicine* **181**, 1359–1366 (2021).
10. Chouhan, A. S., Kaple, M., Hingway, S. & Kaple, M. N. A brief review of diagnostic techniques and clinical management in chronic kidney disease. *Cureus* **15** (2023).
11. Luque-Garcia, J. & De Castro, M. L. Ultrasound: a powerful tool for leaching. *TrAC Trends Anal. Chem.* **22**, 41–47 (2003).
12. Armaly, Z. *et al.* An innovative ultrasound technique for early detection of kidney dysfunction: superb microvascular imaging as a reference standard. *J. Clin. Medicine* **11**, 925 (2022).
13. Zhou, S. K. *et al.* A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE* **109**, 820–838 (2021).
14. Kuo, C.-C. *et al.* Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. *NPJ digital medicine* **2**, 29 (2019).
15. Lee, S. *et al.* Machine learning–aided chronic kidney disease diagnosis based on ultrasound imaging integrated with computer-extracted measurable features. *J. Digit. Imaging* **35**, 1091–1100 (2022).
16. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
17. Azad, B. *et al.* Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689* (2023).
18. Jiao, J. *et al.* Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Med. image analysis* **96**, 103202 (2024).
19. Inker, L. A. *et al.* Estimating glomerular filtration rate from serum creatinine and cystatin c. *New Engl. J. Medicine* **367**, 20–29 (2012).
20. Xu, Q.-S. & Liang, Y.-Z. Monte carlo cross validation. *Chemom. Intell. Lab. Syst.* **56**, 1–11 (2001).
21. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
22. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
23. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114 (PMLR, 2019).

24. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
25. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
26. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).
27. Araújo, N. C., Rebelo, M. A. P., da Silveira Rioja, L. & Suassuna, J. H. R. Sonographically determined kidney measurements are better able to predict histological changes and a low ckd-epi egfr when weighted towards cortical echogenicity. *BMC nephrology* **21**, 123 (2020).
28. Zhao, L. *et al.* Quantifying ultrasound echogenicity difference for accurate chronic kidney disease diagnosis. *J. Nephrol.* **38**, 2353–2362 (2025).
29. Svrcek, P. T., Jang, J., Ge, C., Lee, H. & Kim, Y. H. Combined application of deep learning and conventional computer vision for kidney ultrasound image classification in chronic kidney disease: preliminary study. *Ultrasonography* **44**, 346–353 (2025).
30. Tian, S. *et al.* Deep learning radiomics based on ultrasound images for the assisted diagnosis of chronic kidney disease. *Nephrology* **29**, 748–757 (2024).

Acknowledgements

This research was funded by the Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-RS-2023-00256629) grant funded by the Korea government (MSIT); the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25430545); the NRF funded by the Korea government, MSIT (RS-2023-00217317); Chonnam National University (2025-1036-01); and Chonnam National University Hospital Biomedical Research Institute (BCRI25010 & BCRI26033).

This work is supported in part by funds from the German Ministry of Education and Research (BMBF) under grant agreements No. 01D2208A and No. 01KD2414A (project FAIRPaCT). The authors gratefully acknowledge the computing time granted by the KISSKI project. The calculations for this research were conducted with computing resources under the project *kisski-umg-fairpact-2*. The authors also acknowledge the computing time granted by the Resource Allocation Board and provided on the supercomputer Emmy/Grete at NHR-Nord@Göttingen as part of the NHR infrastructure. The calculations for this research were conducted with computing resources under the project *nim00014*. Anh-Tien Nguyen was a member of the Ph.D. program "Genome Science" – International Max Planck Research School.

We gratefully acknowledge support from the hessian.AI Service Center (funded by the Federal Ministry of Research, Technology and Space, BMFTR, grant no. 16IS22091) and the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. S-DIW04/0013/003)."

We gratefully acknowledge support the hessian. AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. S-DIW04/0013/003).

Author contributions statement

N.V, A-T.N., C.S.K., I.P., A-C.H. performed study concept and design; N.V, A-T.N., T.N.H.V., I.P. performed writing, review and revision of the paper; N.V, A-T.N. provided method developments, analysis and interpretation of data, and statistical analysis. N.V., T.N.H.V, G-Y.Y. provided material support and data collection. All authors reviewed and approved the manuscript.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [fmckdcnuhappendix.pdf](#)