Automating Kidney Disease Diagnosis: A Segmentation-Based, Dual-Input Approach for Classifying Acute and Chronic Kidney Disease from Ultrasound Images

Nora Alkhaldi
Computer Science Department, College of Computer Science and Information Technology, King Faisal University, Al Ahsa, Saudi Arabia
nalkhaldi@kfu.edu.sa (corresponding author)

Received: 15 November 2025 | Revised: 18 December 2025, 27 December 2025, 31 December 2025, 4 January 2026, and 11 January 2026 | Accepted: 6 January 2026

 DOI: https://doi.org/10.48084/etasr.16275

ABSTRACT

Chronic Kidney Disease (CKD) and Acute Kidney Injury (AKI) are serious health conditions, where ultrasound imaging serves as a non-invasive diagnostic tool. Diagnosis relies on identifying significant morphological features, such as a shrunken appearance in CKD or an enlarged, echogenic kidney in AKI. This study proposes and evaluates a two-stage deep learning framework developed using a clinical dataset collected from the Saudi Ministry of National Guard Health Affairs (NGHA). The process utilizes a modified U-Net model for precise kidney segmentation, followed by a novel dual-input classification stage that integrates both the original and segmented images to enhance feature extraction. A comprehensive analysis was conducted on four advanced architectures, including a Swin Transformer, DenseNet121, EfficientNetB0, and ResNet50. The results demonstrated that the ResNet50 model with dual-input configuration, preceded by SS-MUNet segmentation, was highly effective, achieving a test accuracy of 82.88% and an F1-score of 78.36%. Dual-input setups consistently outperformed single-input variants by 10-15% in F1-score across architectures, with ResNet50 and DenseNet121 showing the strongest generalization. This approach enhances diagnostic performance by using both segmented and non-segmented images to distinguish pathological kidney features with high accuracy. The findings establish a robust framework that holds considerable potential as a speedy, reliable, and accessible decision-support tool for clinical practice in nephrology. The proposed dual-input framework based on both original and segmented kidney images improves macro F1-score by an average of 12.5% in the range of 10.2–15.3% and accuracy by 9–13% compared to single-input baselines across four modern architectures on the independent test set. This clinically significant gain enables more reliable automated differentiation of AKI, CKD, and normal kidneys using only standard ultrasound.

Keywords-medical image analysis; acute kidney injury; chronic kidney disease; segmentation; classification; transformer

I. INTRODUCTION

Healthy kidneys help maintain sturdy bones and normal blood flow in the body, but damaged kidneys can lead to toxins being unable to be removed. They can also result in diseases, like cancer, abscesses, stones, and infections, with increased risk due to diabetes, high blood pressure, and long-term medication use [1]. Authors in [2] focused on kidney diseases, specifically CKD and AKI. AKI is caused by a sudden loss of kidney function, whereas CKD is caused by long-term damage. Medical imaging techniques, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and ultrasound, are utilized to diagnose internal disorders. Ultrasound uses high-frequency waves to image organs such as the kidneys and diagnose disorders, including CKD and AKI, non-invasively [3]. Through ultrasound, these conditions present distinct visual characteristics that can support accurate diagnosis and treatment planning. More specifically, CKD presents increased cortical echogenicity, reduced kidney size due to functional tissue loss, and cortical thinning, while AKI may show enlarged kidneys due to inflammation. These characteristics can support accurate diagnosis and treatment planning.

The scale of diagnosis and treatment is substantial in Saudi Arabia, where statistics indicate high dialysis and kidney transplantation rates due to the increased number of CKD patients, costing 2 billion Saudi Riyals for treatment annually [4]. Early detection is, therefore, necessary to reduce the strain on resources. The present study proposes a deep learning system that simultaneously segments and classifies CKD and

AKI from ultrasound images, filling a critical gap, since no prior work has combined both conditions in a single automated, low-cost diagnostic approach. The system was developed and validated using a unique dataset provided by NGHA, which was collected from King Abdul-Aziz Hospital in Al Ahsa. The study workflow and data splits are summarized in Figure 1.

The main contributions of this study are:

- A review of existing work reveals a gap in jointly segmenting and classifying CKD and AKI, highlighting the need for an integrated kidney assessment approach.
- A modified U-Net architecture (SS-MUNet) is presented, featuring deeper feature extraction paths and advanced dropout regularization, to specifically address imaging noise and heterogeneous kidney appearances in ultrasound images.
- The introduction of an innovative dual-input classification architecture that concurrently processes both the original ultrasound image and the precisely segmented kidney region produced by SS-MUNet, enabling the model to capture both global contextual information and focused morphological features.
- A multi-seed comprehensive evaluation of four classification models: DenseNet121, ResNet50, EfficientNetB0, and Swin Transformer. The dual-input framework achieved excellent accuracy, outperforming single-input by 10-15% in F1-score, with ResNet50 and DenseNet121 showing the strongest generalization and peak test performance, obtaining an accuracy of 82.9% and a macro F1-score of 78.4%.

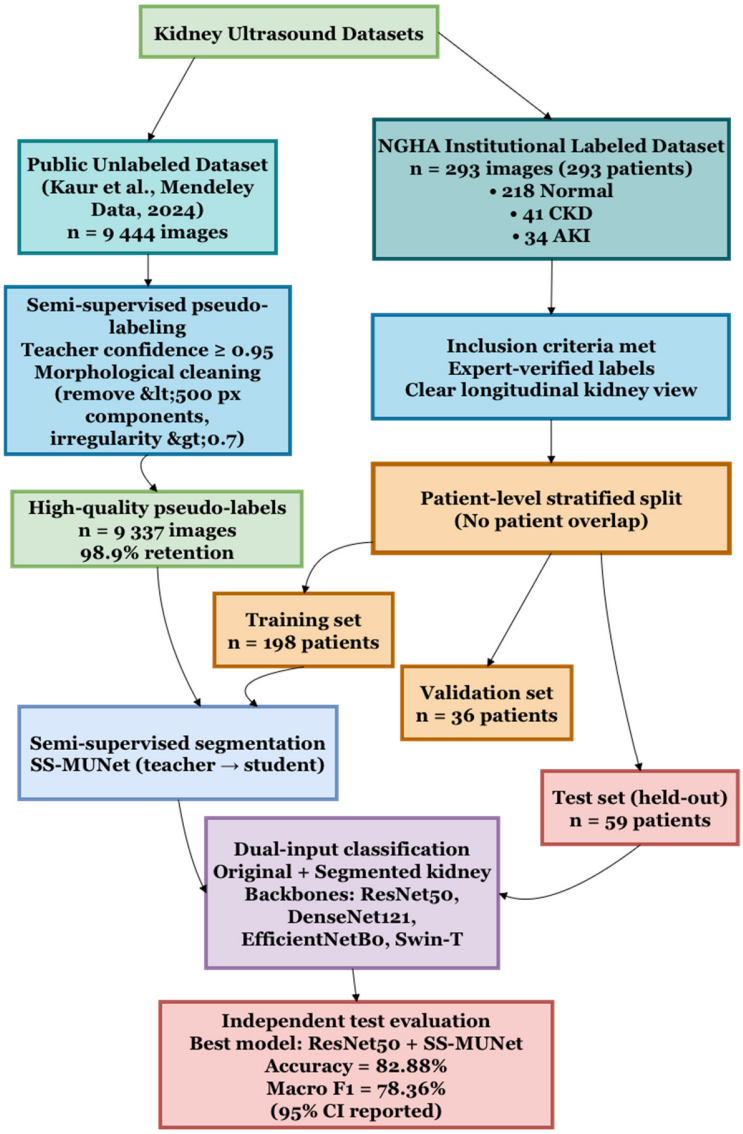


Fig. 1. STARD flow diagram showing the inclusion, splitting, and processing of ultrasound images from the NGHA labeled dataset (n = 293) and public unlabeled dataset (n = 9 337 pseudo-labeled) for semi-supervised kidney segmentation and dual-input classification of normal, CKD, and AKI cases.

- The proposed framework is successfully validated on a unique dataset collected from the NGHA, demonstrating its robustness on real-world kidney ultrasound data.
- This research validated the use of image segmentation to improve data augmentation from unlabeled images via a pseudo-labeling strategy. The study identified two viable diagnostic approaches: an efficient method using a DenseNet121 model for direct diagnosis without segmentation, and a robust method using a dual-input model where segmentation acts as a mechanism to improve accuracy.
- New Systematic ablation studies of semi-supervised segmentation, dual-input streams, attention mechanisms, and Grad-CAM explainability visualizations.

II. LITERATURE REVIEW

This review covers studies published from 2014 to early 2025, with an emphasis on the latest developments. Machine learning (ML) based techniques have been used in earlier studies to automate CKD staging using Support Vector Machines (SVM) and achieving 93.82% sensitivity [5]. A broader CKD classification study utilized various ML models, including Case-Based Reasoning (CBR) with 80% accuracy [6], combined Naive Bayes classification with +12.7% improvement [7], and Multi-Layer Perceptron (MLP) [8]. In addition to ultrasound, Artificial Neural Networks (ANNs) were applied to classify CKD using skin texture analysis [9], while a 3D U-Net was adapted for MRI-based kidney segmentation [10], showing early integration of deep learning in renal imaging. Expanding CKD screening scope, authors in [11, 12] assessed CKD diagnostic models with AUCs ranging from 0.911 to 0.938. Research also has been extended to AKI, with authors in [13, 14] developing predictive frameworks specific to AKI detection, marking an early focus on risk assessment for acute renal conditions.

The evolution of U-Net-based architectures has been driven by foundational deep learning contributions, the original U-Net for medical imaging [15] and the influential ResNet architecture [16]. Authors in [17] first integrated attention mechanisms to refine the U-Net's spatial focus, a concept that authors in [18] further explored by comparing Attention U-Net variants with different backbones (i.e., VGG19 and ResNet152-v2). Authors in [19] proposed Iter-Net for retinal vessel segmentation, which uses an iterative U-Nets framework to repeatedly refine the feature map. Influenced by this, CC-Net applied contrastive learning, an uncertainty-aware network, and a KUS-CONV block targeting noisy and low-quality kidney ultrasound data. [20]. Advancements include the MLAU-Net, which uses deep supervision to achieve a 90.2% Dice coefficient [21], and a U-Net combined with GANs, which reached a segmentation accuracy of 97.83%. [22]. Finally, authors in [23] used W-Net to address multi-center data variability and boosted the Dice score by 6.95%.

Specialized asymmetric, 3D attention, and Fast-Unet++ models have also been developed, achieving Dice scores over 0.90 [24-26]. Authors in [27] systematically reviewed these U-Net adaptations. Authors in [28] demonstrated an early approach using manual cropping with a ResNet-based classifier, achieving 85.6% accuracy on a dataset of 4,505 ultrasound images for CKD detection. Authors in [29] introduced the HMANN model that was tested on a small dataset, reporting a higher 97.5% accuracy. Authors in [30] utilized U-Net-based Convolutional Neural Networks (CNNs) for pediatric ultrasound for identifying CKD anomalies, achieving 90% accuracy. Authors in [31] combined Capsule Networks and Aquila Optimization, achieving 99.18% accuracy in CKD classification using 24 attributes, surpassing Deep Belief Networks.

Subsequent work has focused on refining segmentation and classification independently. For segmentation, studies benchmarked classical architectures like FCN, SegNet, and DeepLab, with DeepLab showing strong results, with obtained 99.14% accuracy [32]. Meanwhile, authors in [33] assessed U-Net++ and DeepLabV3+ in segmenting hydronephrosis areas, reporting a Dice score of 0.9113, indicating the success of refined U-Net architectures. Similarly, authors in [34] confirmed the impressive performance of U-Net on 3,792 2D slices from 66 3D kidney ultrasound volumes. Other advancements included integrated segmentation-classification systems using Mask R-CNN and ResNet-18, achieving an AUC of 0.91 [35], ensemble Deep Neural Networks (DNNs), reporting an accuracy of 92.67% [36], and texture-based networks for CKD screening [37].

The field has grown by incorporating Transformer-based models, which better capture global context compared to the limitations of CNNs. This led to hybrid models, like PST-UNet [38], asymmetric U-Nets with Transformers [39], and the dual-decoder DDTransUNet [40], all demonstrating superior segmentation accuracy. This evolution was boosted by improved frameworks, like the semi-supervised DHCT-DT [41], which used a CNN-Transformer architecture to achieve high performance, even with limited labeled data. The Swin Transformer has been investigated for disease diagnosis, including applications in thyroid nodules [42], breast tumors [43], prostate lesions [44], and lymph nodes [45]. Various Swin Transformer variant architectures have been proposed, offering alternatives to the shifted window mechanism, such as Degenerate Swin to Win [46], Swin-Free [47], and Shuffle Transformer [48].

Authors in [49, 50] reported U-Net Dice scores of 0.85–0.95, while authors in [51] compared seven models for multi-class kidney ultrasound segmentation. Authors in [11, 52] documented SVM accuracies of 98.5% and 93.82% for segmentation and classification, respectively. The field of nephrology has increasingly implemented computational approaches, with a wide range of kidney disease diagnoses being performed or supported using ML [53] and deep learning classification models [54]. Thoracic-net [55] utilized asymmetric CNNs for thoracic diseases, and authors in [56] classified COVID-19 via X-rays, illustrating modality challenges parallel to those in renal ultrasound analysis. Table I summarizes studies on segmentation and classification for kidney disease detection.

While prior works have addressed either segmentation or classification for CKD alone, a combined system for both CKD

and AKI has been a significant research gap/has not been examined. The present study directly fills this gap by introducing an automated two-stage system. First, a modified U-Net performs robust segmentation on noisy clinical ultrasound images. Second, a novel dual-input framework uses both the segmented and non-segmented images to accurately classify conditions as Normal, CKD, or AKI, implementing both CNNs and the context-aware Swin Transformer.

TABLE I. COMPARATIVE ANALYSIS OF KIDNEY SEGMENTATION AND CLASSIFICATION STUDIES

	Year	Study	Model	Performance
Segmentation-only studies	2015	[15]	U-Net	IoU: 92.03%
	2018	[10]	3D Dual Path U-Net	94.4%
	2020	[19]	Iter-Net	AUC: ~0.98
	2022	[22]	U-Net + GANs	Accuracy: 97.83%
	2022	[32]	DeepLab	Accuracy: 99.14%,
	2022	[33]	U-Net++ + DeepLabV3+	Dice score: ~0.91
	2024	[23]	W-Net	Dice score: ~6.95%
	2024	[21]	MLAU-Net	Dice score: 90.2%
	2024	[26]	Fast-Unet++	Dice score: ~0.92
Classification-only studies	2017	[8]	MLP	-
	2020	[11]	SVM	Accuracy: 98.5%
	2020	[12]	CNN (DLA)	AUC: 0.91-0.94
	2020	[14]	RNN	AUC = 0.901
	2022	[36]	Ensemble DNN	Accuracy: 92.67%
	2023	[37]	TBN	96.01%
	2024	[53]	XGBoost	F1-score: 0.82
Integrated (segmentation + classification)	2019	[28]	Manual Cropping ResNet	Accuracy: 85.6%
	2020	[29]	HMANN SVM + MLP	Accuracy: 97.5%
	2021	[30]	U-Net-based CNN ed	Accuracy: ~90%
	2022	[35]	Mask R-CNN Mask R-CNN ResNet18 AUC: 0.91	AUC: 0.91
	2025	This Study	Optimized U-Net Transformer, CNNs	Target: ~98%

III. MATERIAL AND METHODS

A. Datasets Description

This research used an anonymized kidney ultrasound dataset ethically approved by the NGHA [57] and King Faisal University (KFU). The dataset, sourced from the NGHA's Picture Archiving and Communication (PAC) system, contained 337 records from male and female patients aged 16-99, including 220 healthy individuals and 103 patients diagnosed with CKD or AKI.

Each ultrasound image is linked to its corresponding metadata, including a unique identifier (e.g., 'case_0xxx') and diagnosis, which is organized in an Excel file. The kidney structure in each image was manually segmented by nephrologists. Patient records sourced from the PAC system, including case numbers, ID, date, demographics, and images, were matched with their corresponding diagnostic labels from Best Care. Following this linking process, the dataset was fully anonymized.

The NGHA ultrasound dataset was curated with specific inclusion and exclusion criteria to ensure clinical relevance and data quality. Inclusion criteria required images to: (1) have confirmed diagnoses of Normal, CKD, or AKI by nephrologists, validated through the Best Care system at King Abdul-Aziz Hospital, Al Ahsa; (2) include manual segmentations of kidney structures by expert radiologists; and (3) exhibit sufficient image quality for diagnostic evaluation.

Exclusion criteria encompassed: (1) images with missing, incomplete, or ambiguous diagnostic labels; (2) cases lacking corresponding manual segmentations; and (3) scans with severe imaging artifacts or poor resolution. As a result, 15 cases were excluded during the data cleaning process, reducing the dataset from 337 to 293 images.

1) Data Preparation and Image Pre-Processing

The collected dataset included 337 ultrasound records, which presented challenges with missing data and imbalanced class distribution: 220 for Normal (N), 58 for CKD, and 55 images for AKI. A data cleaning process was performed on the images and their annotations to remove empty records, unlabeled cases (U), and instances without annotations. During this process, 14 missing images were identified and excluded (Exc.). Table II presents the final distribution of the cleaned data, which contains 218 Normal, 41 CKD, and 34 AKI cases, for a total of 293 records.

TABLE II. DISTRIBUTION OF THE DATASET BEFORE (PRE) AND AFTER (POST) DATA CLEANING PROCESS

S	N	CKD	AKI	L	R	U	Exc.	T
Pre	220	58	55	163	173	4	15	337
Post	218	41	34	143	150	0	0	293

The literature reports no significant diagnostic performance difference between the left (L) and right (R) kidneys, with the distribution of the 293 samples for this study presented across both sides in Table III.

TABLE III. DATASET DISTRIBUTION BY DIAGNOSIS AND KIDNEY SIDE

Kidney	N	CKD	AKI	Total
L	106	21	16	143
R	112	20	18	150
Total	218	41	34	293

As shown in Figure 2, the labeled dataset was collected from NGHA and consists of 293 ultrasound images from 293 unique patients (218 Normal, 41 CKD, 34 AKI). The study implemented a structured approach for preprocessing, including generating ground truth masks from annotations, splitting the data into training, validation, and testing subsets, and resizing all images to 224×224. Subsequently, data augmentation was applied to the training set before all images were normalized.

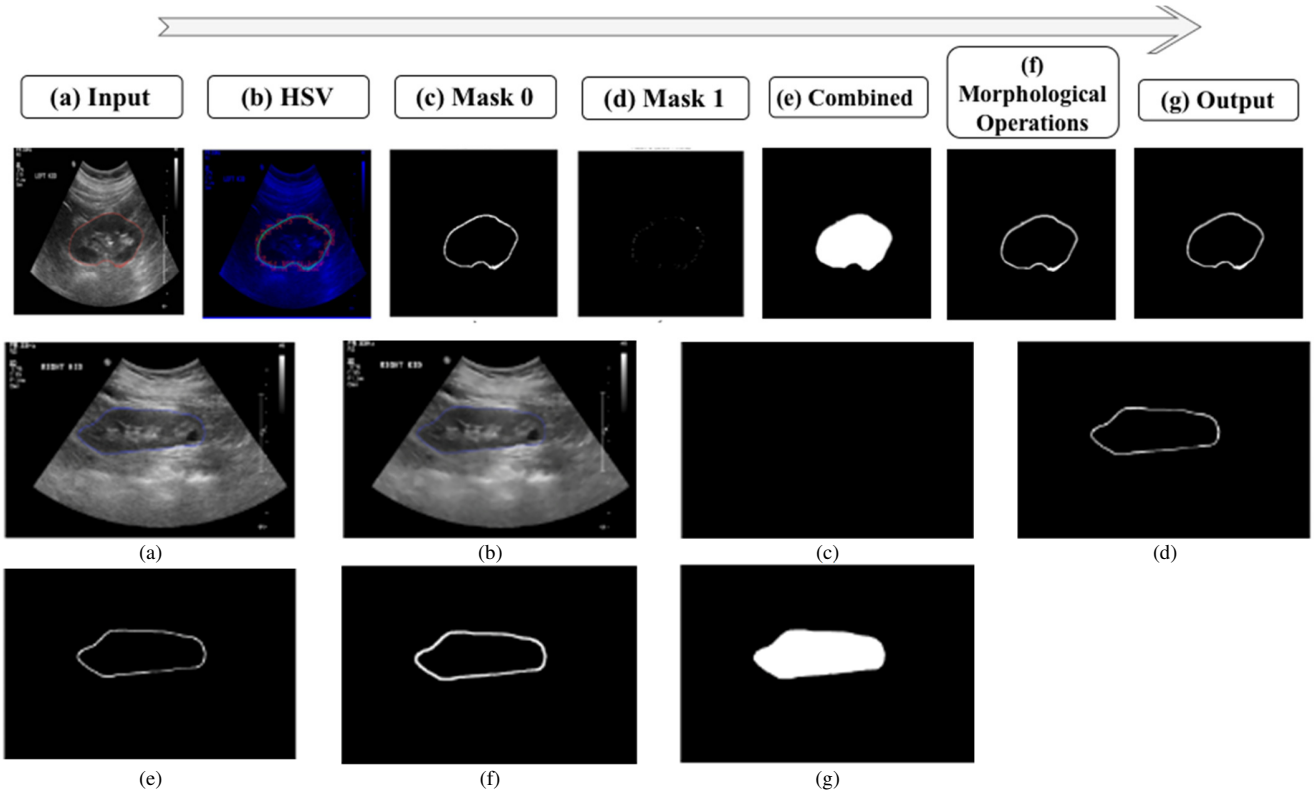


Fig. 2. Mask generation process: (a) input annotated image, (b) image conversion to HSV color space, (c) mask for yellow regions, (d) mask for blue regions, (e) combined mask, (f) refined mask after morphological operations, and (g) final filled mask highlighting the ROI.

2) Mask Generation

To generate ground truth segmentation masks, the study employs multi-stage image processing procedures. First, manually annotated images are converted to the Hue-Saturation-Value (HSV) color space for its efficacy in isolating regions based on color. Then, a dual-thresholding technique is applied to identify and merge specific color ranges, such as yellow and blue, into a single integrated mask. This mask subsequently passes through morphological operations to remove noise and fill small gaps, yielding a more refined mask. Finally, contours are detected from this refined mask and filled in to create an accurate binary mask representing the Region of Interest (ROI), a process illustrated in Figure 1. The process of generating a mask for each annotated image is important for building a segmentation model. These masks enable the segmentation model to learn the precise boundaries of the ROI during training, as depicted in Figure 3.

3) Dataset Splitting

The dataset was partitioned at the patient level into training, validation, and test sets before applying data augmentation to prevent data leakage. The dataset, consisting of 293 image-mask pairs, was split into 68% for training, 12% for validation, and 20% for test sets. All images from a single patient belonged exclusively to either the training, validation, or test set. To be more specific, a stratified sampling approach was adopted to maintain class balance across splits. The train_test_split function from the scikit-learn library was used with a random state of 42 for greater accuracy, first splitting the data into 80% train and validation with 234 images, and 20% test with 59 images. Then, a train and validation set were further split into 85% training with 198 images and 15% validation with 36 images. The exact class distribution (Normal / CKD / AKI) in each split is reported in Table IV. A stratified patient-level split was applied, and class imbalance across the Normal, CKD, and AKI categories was addressed using a WeightedRandomSampler.

TABLE IV. DATASET CHARACTERISTICS AND CLASS DISTRIBUTION AFTER PATIENT-LEVEL STRATIFIED SPLITTING

Split	Normal	CKD	AKI	Total
Training	141	25	20	186
Validation	33	8	7	48
Test	44	8	7	59
Total	218	41	34	293

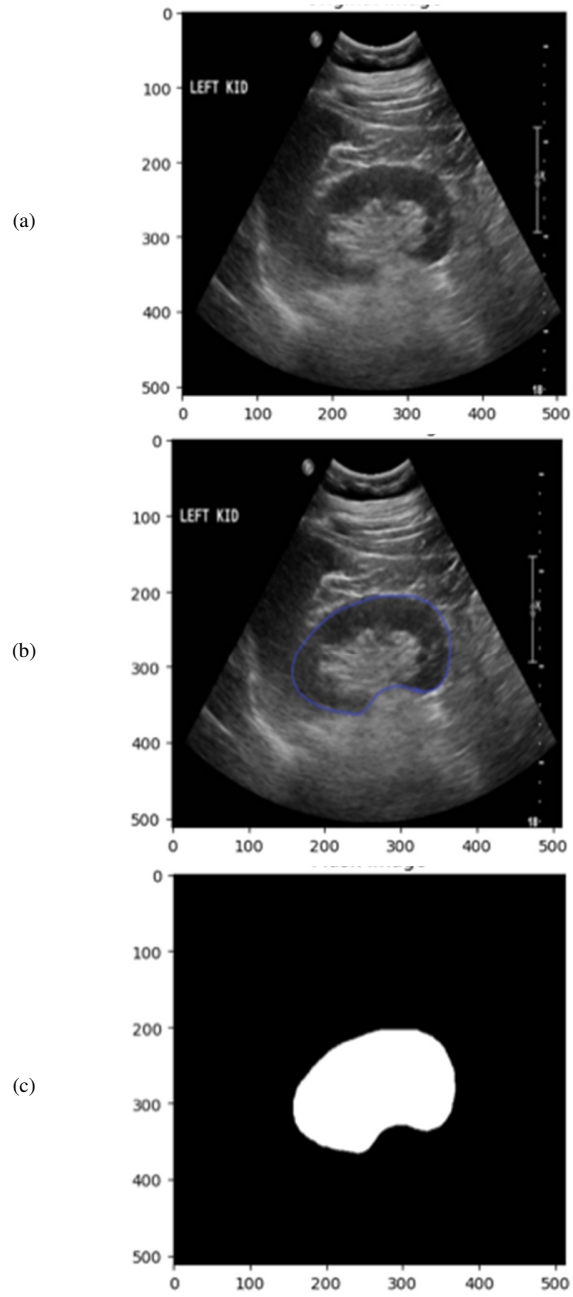


Fig. 3. Kidney structure segmentation: (a) original image, (b) manually annotated segmentation, and (c) generated mask highlighting the kidney region.

4) Data Augmentation

The augmentation strategy applied geometric transformations, including horizontal flips, vertical flips, and limited rotations, to address the significant class imbalance in the initial training dataset of 186 images, which comprises 141 Normal, 25 CKD, and 20 AKI image-mask pairs using the Albumentations library [57]. These techniques were chosen to introduce variability that mimics differences in patient positioning, probe angles, and imaging artefacts, while strategically preserving critical diagnostic features like renal cortex echogenicity, corticomedullary differentiation, and parenchymal texture. These features are critical for detecting pathological changes, such as heightened echogenicity in CKD or parenchymal oedema in AKI, and the enhancement approaches were chosen to preserve their diagnostic validity. The primary goal of this augmentation was to create a balanced dataset for training. This was achieved by oversampling the minority classes (CKD and AKI) and downsampling the majority class to create a uniform distribution. The class distribution yielded 588 Normal, 588 CKD, and 483 AKI images. Oversampling was then performed to balance the training set, increasing Normal and CKD from 588 to 600 images each and AKI from 483 to 600 images. This process resulted in a final, balanced training dataset of 1,800 images. Throughout this process, all images and their corresponding masks were augmented together before being resized to 256×256 pixels and normalized. Table V summarizes the distribution of the dataset before and after this augmentation and balancing procedure.

TABLE V. DISTRIBUTION OF THE LABELLED TRAINING DATASET BEFORE AND AFTER AUGMENTATION

Stage	Normal	CKD	AKI	Total
Before	141	25	20	186
After	600	600	600	1800

5) One-Hot Encoding

The original collected dataset was labeled into three classes: Normal (N), CKD (C), and AKI (A). The diagnostic labels were first converted into a one-hot encoded format for the classification task, including Normal as [1, 0, 0], CKD as [0, 1, 0], and AKI as [0, 0, 1]. For the corresponding segmentation task, ground truth masks were binarized, with pixel values of 1 representing the kidney foreground and 0 for the background. These two encoding structures provided a clear data representation for both the classification and U-Net-based segmentation tasks.

6) Statistical Analysis

The evaluation of classification performance was conducted through accuracy, macro-averaged F1-score, and macro-averaged ROC-AUC. The statistical significance of the accuracy differences between paired predictions on the same test set was evaluated using McNemar's test with continuity correction. DeLong's test, which used a 5,000-iteration bootstrap strategy, was utilized to look at differences in AUC. The Bonferroni correction was employed to adjust for multiple comparisons. All statistical tests were conducted as two-tailed, with a significance threshold set at $p < 0.05$.

B. Experimental Design

The proposed system framework, portrayed in Figure 4, consists of several key stages: Data Preprocessing, Segmentation, Postprocessing of segmented images, and Classification.

1) Kidney Segmentation with U-Net

This study employed a U-Net model for semantic segmentation, classifying each pixel as either kidney (Class 1)

or background (Class 0) based on manually annotated masks to produce a precise binary mask for accurate localization. The original U-Net architecture is designed in a U-shape and includes encoder, bottleneck, and decoder paths [15]. This architecture is ideal because it captures low-frequency feature maps from earlier convolution layers and high-frequency feature maps from later layers. These combined features help identify kidney boundaries at different levels. A modified U-Net is used in this study to enhance feature extraction, regularization, and spatial recovery, building on the original architecture with enhancements to improve performance and generalization while reducing overfitting, as illustrated in Figure 5. The Encoder consists of five blocks with convolutions, max pooling, and dropout, followed by a Bottleneck Block. The Decoder pathway upsamples features, concatenating them with skip connections from corresponding encoder blocks. The network concludes with a final 1×1 convolution and Sigmoid activation to produce the 256×256×1 Output Mask, representing the segmented kidney.

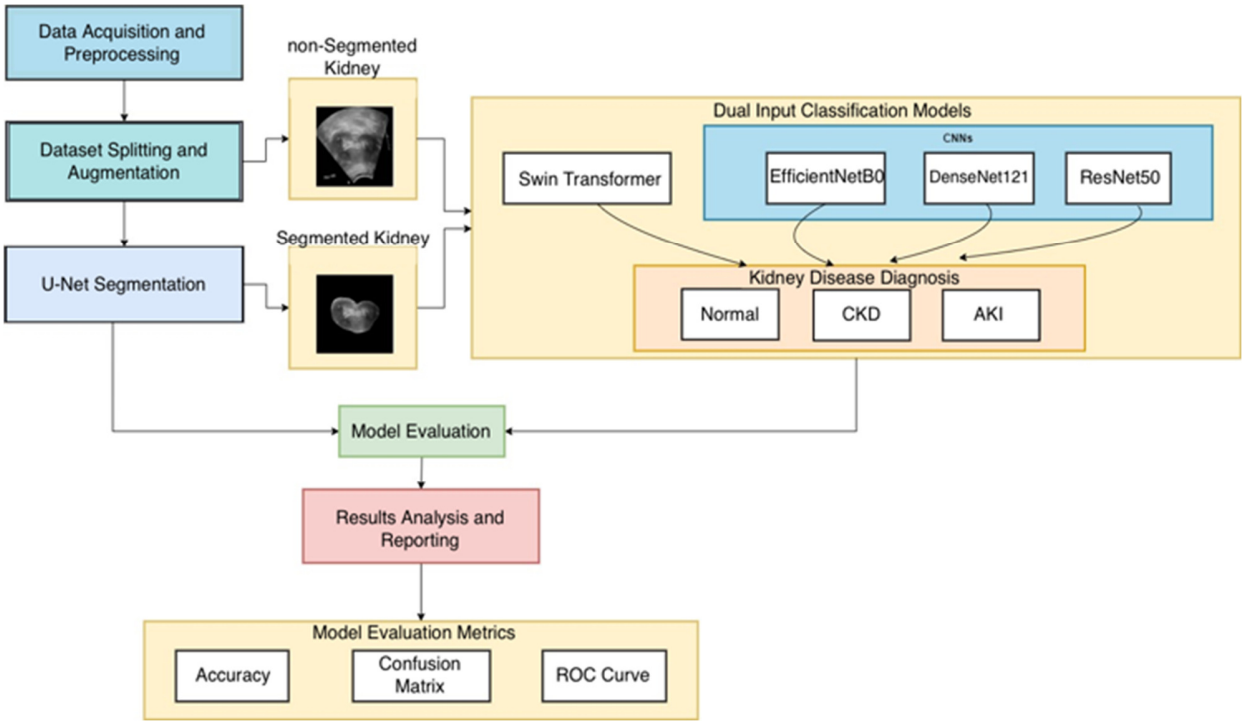


Fig. 4. Experimental workflow for kidney disease classification.

While the modified U-Net architecture preserves the core structure of the original U-Net, it introduces several modifications. First, the addition of higher dropout rates between 0.3 and 0.5 after pooling and up-sampling layers to enhance regularization throughout the model. Second, the addition of extra 3×3 convolutional layers in the deeper encoder blocks 3, 4, and 5, so that the model will be able capture more complex features at lower resolutions. Finally, extending the decoder path with an additional up-sampling stage, leading to 6 decoder blocks. This is accomplished by adding a 32-filter initial encoder block, which allows the network to down-sample to a lower resolution (4×4) before up-sampling to the initial input size. For the segmentation task, the Binary Cross-Entropy (BCE) loss function was used to measure the pixel-wise differences between the generated binary mask of the kidney and the manually annotated binary mask. The standard BCE formula is:

$$BCE = -\frac{1}{n}\sum_{i=1}^{n}[y_i\, log(p_i) + (1-y_i) log\,(1-p_i)] \quad (1)$$

where n represents the total number of samples, y_i refers to the true label (0 or 1) for the i^{th} sample, and p_i is the predicted probability that the i^{th} sample belongs to class i. A class weighting strategy is implemented [59] to address the significant class imbalance between the kidney region (foreground) and the background. This method scales the standard BCE loss by applying specific weights w_0 and w_1 for the kidney (class 1) and non-kidney (class 0) pixels, respectively. When the model receives imbalanced data, the function minimizes pixel-wise errors, calculates the loss value, and modifies its training accordingly, as detailed in the weighted loss formula:

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}[w_0 \cdot y_i\, log(p_i) + w_1 \cdot (1-y_i)\, log(1-p_i)] \quad (2)$$

2) Dual-Input Kidney Condition Classification

This study implemented four deep learning models for the classification of kidney ultrasound images into three classes: Normal, CKD, and AKI. These architectures included a Swin Transformer and three pretrained CNNs: EfficientNetB0, DenseNet121, and ResNet50.

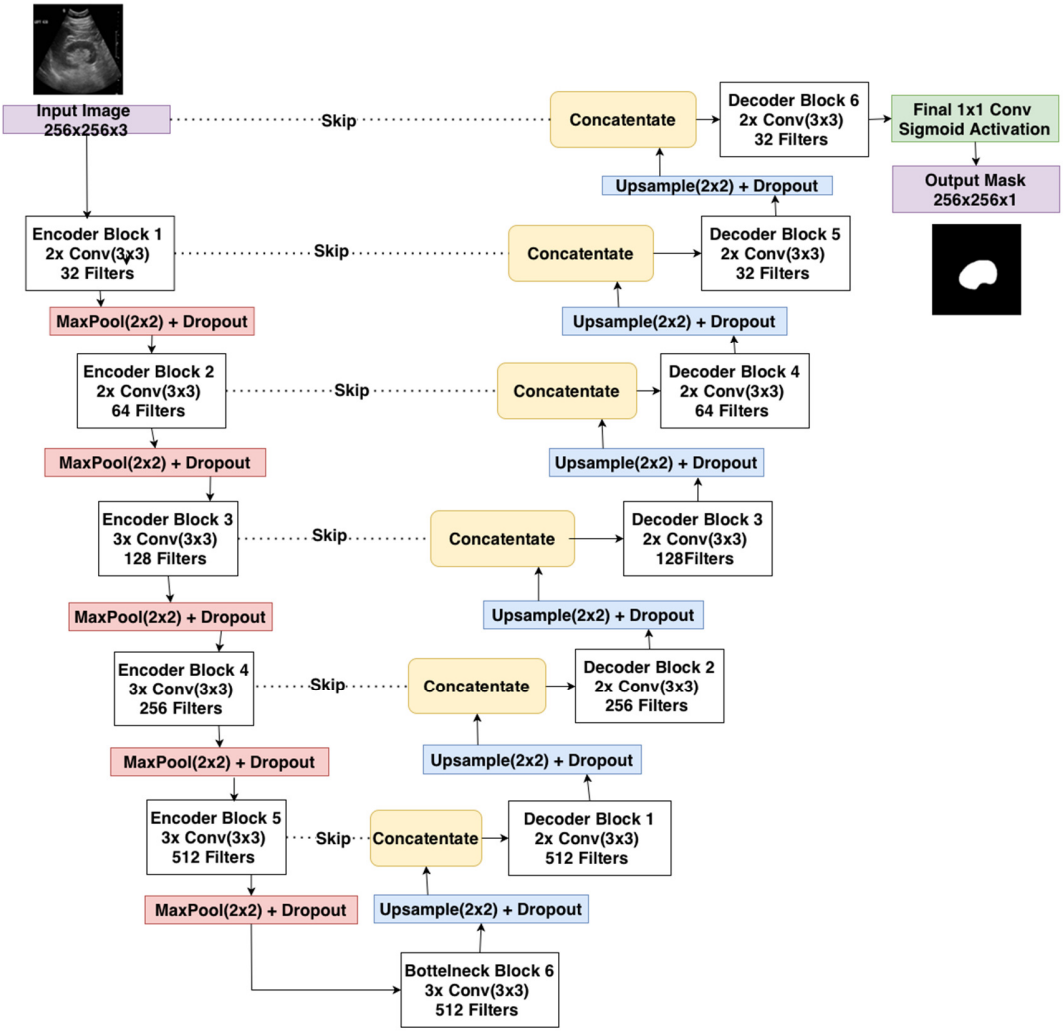


Fig. 5. A modified U-Net architecture for kidney segmentation.

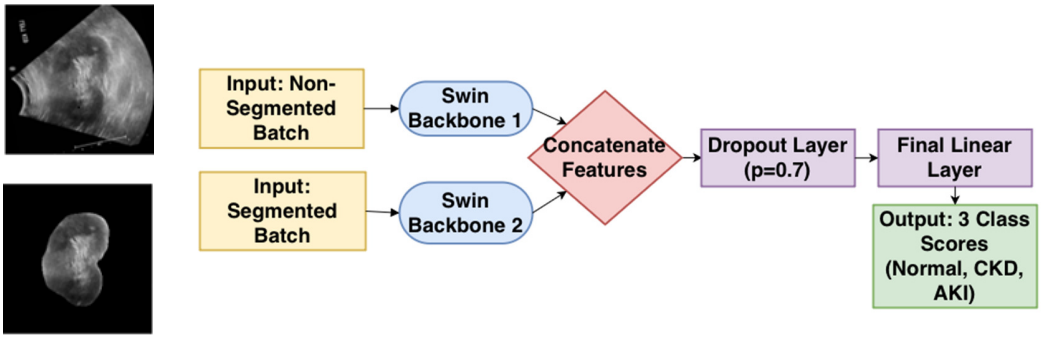


Fig. 6. Dual-Input Swin Transformer model for kidney classification.

Each model was designed to process both segmented (i.e., from the earlier segmentation stage) and corresponding non-segmented ultrasound images.

First, the Swin Transformers [43] was implemented to classify kidney diseases using a hierarchical architecture with shifted window-based self-attention and dual-input, utilizing the swin_tiny_patch4_window7_224 variant as the backbone architecture. It simultaneously processed both segmented and non-segmented ultrasound images using two separate Swin Transformer branches initialized with pretrained weights, as presented in Figure 6. The two 768-dimensional feature vectors were combined into a 1536-dimensional vector, processed

through a 0.7 dropout layer, and then mapped to an output with three classes using a fully connected layer. In addition, this study investigated the effectiveness of three CNN models, including EfficientNetB0 [60], DenseNet121 [61], and ResNet50 [16] for kidney classification using ultrasound images. The design structure of EfficientNetB0 depends on composite scaling, which consistently improved network depth, width, and input resolution. The features extracted from both streams, as shown in Figure 6, are then concatenated and passed through a Dropout Layer (p = 0.7) and a final linear layer to produce 3-class scores for kidney conditions (Normal, CKD, AKI)

The model was constructed by combining Mobile Inverted Bottleneck Convolutions (MBConv) with squeeze-and-excitation blocks. DenseNet121 was utilized for its dense connection structure, where each layer within a dense block takes inputs from all preceding layers in that block. The 121-layer network was arranged into four dense blocks, with transition layers conducting pooling and convolution tasks between them. DenseNet121 was utilized to enhance feature reuse and optimize gradient flow, making it particularly effective for complex feature learning. Furthermore, ResNet50 was investigated to address the problem of vanishing gradients using 16 residual blocks, enabling deep model training on limited ultrasound data for kidney condition classification. The 50-layer architecture applied convolutions for feature extraction and pooling for dimensionality reduction. Each adapted model concluded with Global Average Pooling (GAP) and fully connected layers with SoftMax activation to classify images into normal, CKD, or AKI kidney conditions. The proposed pretrained CNN architectures were implemented using a unified configuration to evaluate their relative performance on the given dataset. Each model was initialized with weights pretrained on the ImageNet dataset. A subset of the first layers was frozen to preserve low-level characteristics, while the remaining layers were fine-tuned on the target kidney ultrasound dataset. With approximately 5.3 million parameters and 237 layers, EfficientNetB0 generated a 1280-dimensional feature vector/branch; DenseNet121 had 8.0 million parameters and 429 layers; and ResNet50 had 25.6 million parameters and 175 layers.

All three models were adapted into a dual-input structure, as illustrated in Figure 7, where images pass through separate backbone streams before feature concatenation for final classification. For each architecture, a dual-branch configuration was adopted: one branch processed segmented ultrasound images and the other processed non-segmented images, with input dimensions standardized to 224. In each branch, the base model's output was passed through a GAP 2D layer to reduce spatial dimensions, resulting in a 1D feature vector, which was then fed into a dense layer with 256 units and ReLU activation. The feature vectors from both branches were concatenated, yielding a combined feature vector of dimension 512. A dropout layer with a rate of 0.5 was applied to prevent overfitting, followed by a final dense layer with 3 units and SoftMax activation to produce the probability distribution over the three classes (Normal, CKD, AKI). This resulted in a dual-input design with one fully connected layer/branch that represented the final classification layer. For each architecture, a dual branch configuration was adopted: one branch processed segmented ultrasound images and the other processed non-segmented images, with input dimensions standardized to 224. In each branch, the base model's output was passed through a GAP 2D layer to reduce spatial dimensions, resulting in a 1D feature vector, which was then fed into a dense layer with 256 units and ReLU activation. The feature vectors from both branches were concatenated, yielding a combined feature vector of dimension 512. A dropout layer with a rate of 0.5 was applied to prevent overfitting, followed by a final dense layer with 3 units and SoftMax activation to produce the probability distribution over the three classes (Normal, CKD, AKI). This resulted in a dual-input design with one fully connected layer per branch that represented the final classification layer.

3) Semi-Supervised Segmentation Training

A primary challenge in this study was the limited size of the manually annotated dataset that was collected from a private hospital. To improve model generalization, a large and publicly available kidney ultrasound dataset was incorporated [62]. This additive dataset consists of 9,416 kidney B-mode ultrasound images, categorized into Normal with 4,414 images and Stone with 5,002 images, which were collected from various clinical centers using multiple ultrasound systems. Although originally collected for stone detection, the dataset shares identical imaging physics and anatomical views with the collected clinical data from NGHA, making the kidney-versus-background segmentation task pathology-invariant. A Semi-Supervised Learning (SSL) strategy, based on a teacher-student pseudo-labeling model [63], was employed to integrate this dataset. First, an initial teacher model was trained entirely on the manually annotated data, NGHA. This teacher model was then employed to generate segmentation masks, referred to as pseudo-labels, for the entire unannotated dataset. These pseudo-labels were filtered using a threshold to ensure quality, and then a final student model was trained on a combined dataset including the original manual annotations and the new pseudo-labels. This supportive dataset was used to augment the training set and was not included in the validation and test sets to avoid any bias in the final evaluation of model performance. A teacher model that works as the best-performing supervised architecture on the 186 annotated NGHA images was first trained. This teacher then generated pseudo-labels on the entire unlabelled set. Only high-confidence, morphologically plausible pseudo-labels were retained, confidence ≥0.95 + strict post-processing, yielding 9,337 high-quality pseudo-labels (i.e., 98.9 % retention rate). The final student model was trained on the union of the original 186 manual annotations and these 9,337 pseudo-labels (total 9,535 samples), significantly improving generalization without contaminating validation or test sets. Four student architectures: SS-UNet (standard U-Net), SS-AUNet (Attention U-Net), SS-MUNet (proposed modified U-Net with dense blocks and attention gates), and the original supervised MUNet were evaluated for comparison. The choice of attention gates and dense encoder blocks in SS-MUNet and SS-AUNet was based on the theoretical challenges of renal ultrasound imaging, such as severe speckle noise and low corticomedullary contrast.

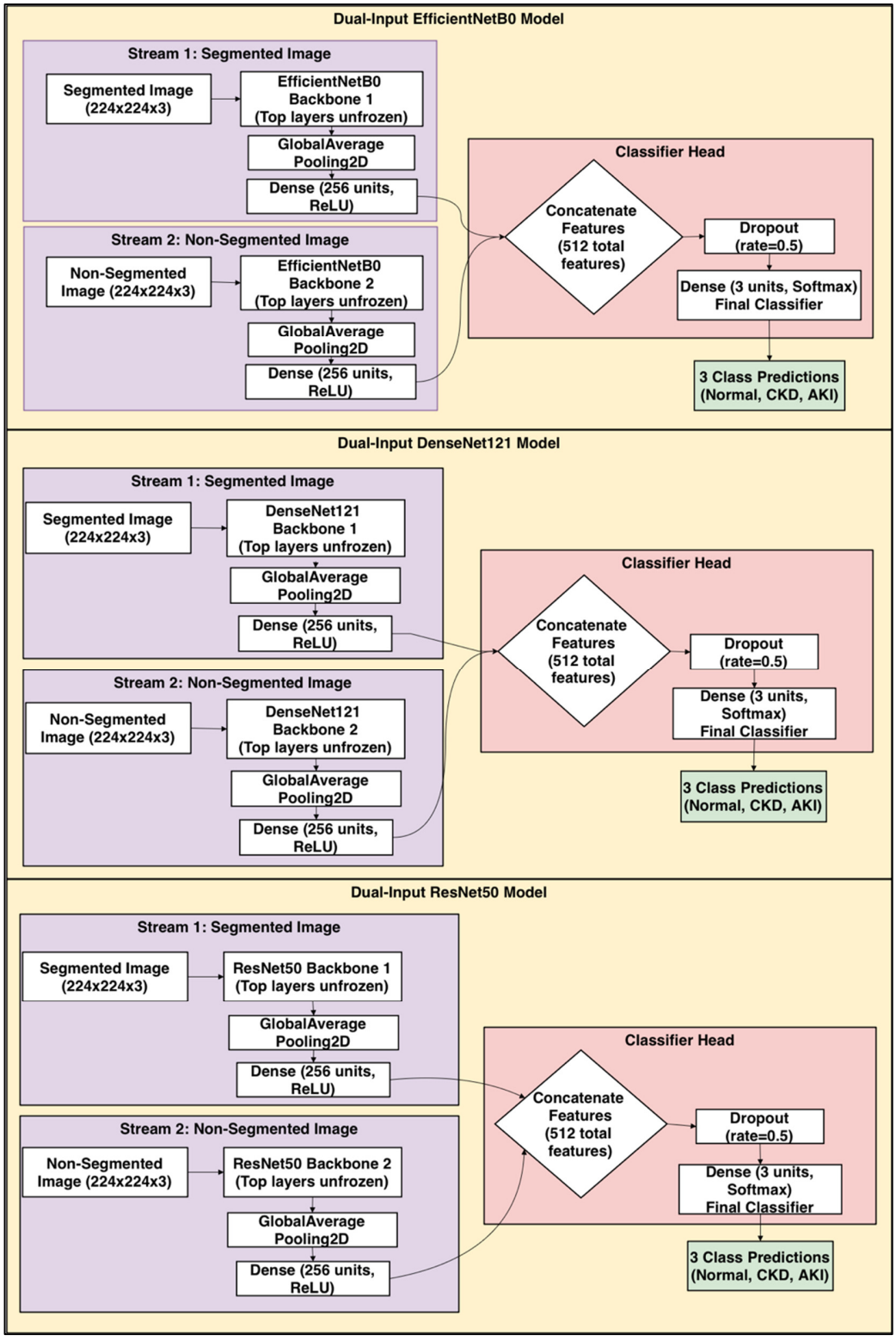


Fig. 7. Dual-Input EfficientNetB0, DenseNet121, and ResNet50 Models.

Attention gates help suppress irrelevant noise and highlight critical features [52], which is essential in ultrasound due to shadowing issues. Dense connectivity allows for precise feature reuse across scales, helping in the capture of both cortical thinning and organ enlargement without vanishing gradients [51, 60]. In general, attention gates suppress ultrasound noise

and acoustic variability, while dense deeper encoders capture multi-scale CKD/AKI morphology; therefore, dual-input streams fuse global context with precise local kidney features against clutter. An ablation study confirms that these components lead to statistically significant improvements in performance.

As displayed in Figure 7, all models employ dual-stream architecture where Stream 1 processes segmented images and Stream 2 processes non-segmented images (224×224×3). Features from both streams are concatenated (512 total features) and passed through a common classifier head, which consists of a dropout layer (rate = 0.5) and a final dense layer (3 units, SoftMax) for 3-class predictions (Normal, CKD, AKI).

C. Implementation Details

1) Segmentation Experimental Setup

The modified U-Net is used for kidney segmentation from ultrasound images, with binary masks derived from manual annotations serving as training targets. The model, implemented in TensorFlow and Keras, processed input ultrasound images of size 256×256 pixels with three channels (RGB). Memory was carefully managed using TensorFlow and Python utilities to avoid GPU overload, maintaining stable usage on an NVIDIA A100 GPU with 40GB of memory via Google Colab. Segmentation performance was evaluated using Intersection over Union (IoU) and Dice scores. The segmentation model with the best performance was selected to proceed to the next stage, that is kidney condition classification.

Initially, segmentation performance was assessed across different batch sizes, including 4, 8, 16, and 32, using a fixed dataset to identify the optimal segmentation model and best configuration. The model was trained for up to 150 epochs using BSE loss. The Adam optimizer was employed with a learning rate of 0.0001 for batch sizes of 4, 8, and 16, and a reduced rate of 0.0005 for batch size 32 to address training instability. Batch sizes were tested, with corresponding steps/epoch computed based on the dataset size. Early stopping was applied based on the validation Dice coefficient with a patience of 20. The final output is a single-channel segmentation mask with 256×256 pixels and pixel values capped at 0.5 to produce a binary mask where 1 represents kidney and 0 represents background.

2) Semi-Supervised Training and Pseudo-Label Filtering

In particular, four configurations were implemented to determine the best architecture for this task. The proposed modified U-Net model, which was trained on limited annotated data, also utilized the semi-supervised strategy (SS-MUNet). The standard U-Net model was trained with the semi-supervised strategy (SS-UNet). Finally, an Attention U-Net model was trained with the semi-supervised strategy (SS-AUNet) to benchmark against a common advanced architecture. These configurations were implemented in PyTorch, with models trained for up to 150 epochs using an Adam optimizer and a Hybrid Loss function combining BCE and Dice Loss.

To minimize noise propagation and maintain high-quality supervision in student training, rigorous pseudo-label filtration was implemented. Pixels were initially accepted only if the teacher model's confidence was equal to or greater than 0.95. Morphological post-processing was conducted, in which connected components smaller than 500 pixels were eliminated, and contours showing convexity defects with an irregularity index ($\text{perimeter}^2 / (4\pi \times \text{area})$) exceeding 0.7 were excluded. This careful quality control technique retained 9,337 high-confidence, and morphologically valid pseudo-labels from a total of 9,444 unannotated images, resulting in a retention rate of 98.9%. This approach significantly lowered error propagation by preserving almost the entire unlabeled dataset for use as training data.

Once trained, the teacher model was employed to generate predictions on a large unannotated dataset of 9,444 images. High-confidence predictions were retained, producing 9,337 pseudo-labeled images. These pseudo-labels, together with the original 186 annotated images, formed the training set for the student models. Student training was performed in two stages: an initial feature-extraction phase with the backbone frozen (50 epoch), followed by a fine-tuning phase where the entire network was unfrozen (100 epoch) for end-to-end optimization. The best-performing student model SS-MUNet achieved a mean Dice coefficient of 0.922 ± 0.031 and an IoU of 0.858 ± 0.052 on the independent NGHA test set (59 images).

3) Classification Experimental Setup

The classification model is built upon more reliable segmentation output to ensure overall diagnostic performance. The primary objective of this model is to accurately differentiate between normal kidney conditions, CKD, and AKI. The implemented model integrates both the segmented images resulting from the modified U-Net model and the corresponding non-segmented images. Two model architectures were evaluated: a single-input model with a pretrained backbone and dropout layer, and a dual-input model with twin backbones for parallel processing of the original image and segmented image.

The classification dataset used 293 ultrasound images and segmented kidney outputs generated by segmentation methods. A semi-supervised U-Net student model produced high-quality segments, increasing the dataset to 1137 images by pseudo-labeling. A stratified patient-level data partitioning technique was employed to preserve class proportions and minimize potential biases in disease prevalence.

A pretraining strategy was used to initialize model weights, adjusting to ultrasound imagery domains and kidney detection tasks (i.e., using 4414 normal/stone images). First, backbones were initialized with standard ImageNet-pretrained checkpoints. Second, to adapt backbones to ultrasound-specific patterns, kidney-specific (expert) weights from an initial kidney detection task were loaded, selectively updating compatible layers while preserving general features from the initial pretraining. Augmentations were applied post-splitting using Albumentations library [58] on the training subset only. All images were preprocessed by resizing to 224×224 and normalizing.

Both the Swin Transformer and the three CNNs pre-trained models were developed using the dual-input architecture and trained for a maximum of 100 epochs with a batch size of 16 on an NVIDIA A100 GPU via Google Colab. The classification models were fine-tuned using the Adam optimizer with a higher learning rate of 1×10^{-4} to address class imbalance. WeightedRandomSampler and weighted cross-entropy loss function were employed. Backbones were initially frozen, then unfrozen. An early stopping mechanism, monitoring the validation macro F1-score with a patience of 15-20 epochs, was used to prevent overfitting. The learning rate scheduler monitored validation accuracy and the best model was saved.

D. Evaluation Metrics

The performance of segmentation models was evaluated using two sorts of metrics, one for the segmentation task and another for the final classification task. For segmentation, performance was measured using the Dice coefficient and IoU. Both metrics evaluate the spatial overlap between the predicted mask and the ground truth annotation to quantify segmentation accuracy, as defined in:

$$IoU = \frac{1}{c}\sum_{i=1}^{c}\frac{A_i \cap B_i}{A_i \cup B_i} \quad (3)$$

$$Dice\ coefficient = \frac{2|A \cap B|}{|A|+|B|} \quad (4)$$

where c is the number of classes, and A and B are the predicted and ground truth images, respectively. For the classification task, a standard set of metrics was used. This included precision to measure the accuracy of positive predictions, using (5), recall measuring the ability of the model to identify all true positive cases using (6), and the F1-score as their balanced mean using (7). In addition to standard accuracy in (8), balanced Accuracy was calculated to provide a more robust evaluation given the imbalanced nature of the dataset. Finally, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was used to evaluate the overall capability of the model to distinguish between the Normal, CKD, and AKI classes:

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \quad (5)$$

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives} \quad (6)$$

$$F1-Score = 2\times\frac{precision\times Recall}{precision+Recall} \quad (7)$$

$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{Total} \quad (8)$$

IV. RESULTS

A. Segmentation Model Performance

The modified U-Net was evaluated with batch sizes of 4, 8, 16, and 32, as summarized in Table VI. Although batch size 4 achieved the highest test Dice score of 0.8907, batch size 16 was selected for the final model. This configuration attained the best validation Dice score of 0.8707, remained competitive on the test set with an accuracy of 97.84%, and provided a balanced trade-off between generalization and training efficiency.

TABLE VI. IMPACT OF BATCH SIZE ON THE SEGMENTATION PERFORMANCE

Batch size	Validation accuracy	Validation Dice score	Test accuracy	Test Dice score	Time (s)
4	0.9815	0.9075	0.9774	0.8907	347.70
8	0.9756	0.8626	0.9747	0.8815	277.11
16	0.9796	0.8707	0.9784	0.8897	299.21
32	0.9742	0.8242	0.9729	0.8600	262.90

Using the optimal batch size of 16, the model was trained for 150 epochs on an expanded dataset with an Adam optimizer, BCE loss, and a 1×10^{-4} learning rate to generate 256×256 one channel segmentation masks. Table VII outlines the final segmentation metrics, while Figure 8 demonstrates stable training with loss converging below 0.1 and accuracy surpassing 0.99. As seen in Figure 8, regarding the raining and validation performance of the modified U-Net with batch size 16 and learning rate 0.0001 over 235 epochs, the loss decreases steadily, stabilizing below 0.1 for both training and validation, while the accuracy rises rapidly, converging above 0.98 for both sets. The segmentation results on the dataset are presented in Figure 9, which compares the input images, ground truth masks, and predicted masks across three samples. As demonstrated in Figure 9, The predicted masks closely align with the ground truth, demonstrating the model's ability to accurately segment ROI across diverse ultrasound samples. Figure 10 presents the original ultrasound images alongside their predicted masks, using the modified U-Net, and segmented outputs across four sample cases.

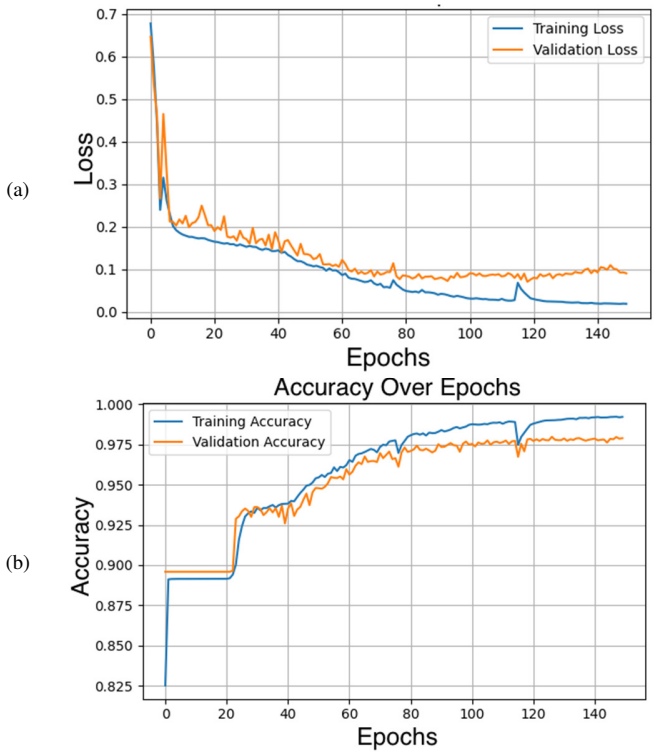


Fig. 8. (a) Training and validation loss, (b) training and validation accuracy.

TABLE VII. FINAL PERFORMANCE METRICS OF THE MODIFIED U-NET MODEL USING A BATCH SIZE OF 16

Dataset split	Accuracy	Dice coefficient	Loss
Training	0.9920	0.9465	0.0188
Best validation	0.9796	0.8707	0.0706
Test	0.9784	0.8897	0.0970

Fig. 9. A sample of the segmentation results for the modified U-Net model on the dataset of ultrasound images: (a) the input ultrasound image, (b) the ground truth mask, and (c) the predicted mask.

Figure 11 displays the pixel-wise confusion matrix and metric distributions for the best model, highlighting high true positives, low misclassifications, and consistent performance, with a mean IoU of 0.8209 and a Dice coefficient of 0.8879.

B. Semi-Supervised Model Evaluation

An SSL strategy was implemented to expand the training data. A teacher-student pseudo-labeling design was implemented, in which a teacher model trained on manually annotated data was used to generate pseudo-labels for 9,416 unannotated kidney ultrasound images [57]. The effectiveness of this method was demonstrated by the generation of a high volume of confident masks, including 9,165 for SS-UNet, 9,339 for SS-AUNet, and 9,337 for SS-MUNet. For the final stage, the training sets were augmented with a 10% random sample of these pseudo-labels, while the original validation and test sets were preserved for accurate evaluation.

The quantitative performance of the three semi-supervised models on the held-out test set is presented in Table VIII. All models benefited significantly from the SSL strategy, achieving high performance. The SS-AUNet demonstrated the best overall performance, achieving a final test Dice coefficient of 0.9203 and a mean IoU of 0.8712. This was closely followed by the proposed modified U-Net SS-MUNet, which achieved a test Dice coefficient of 0.9070. These results suggest that while all architectures are effective, the attention mechanism in the SS-AUNet provides a distinct advantage in accurately delineating kidney boundaries from complex ultrasound features.

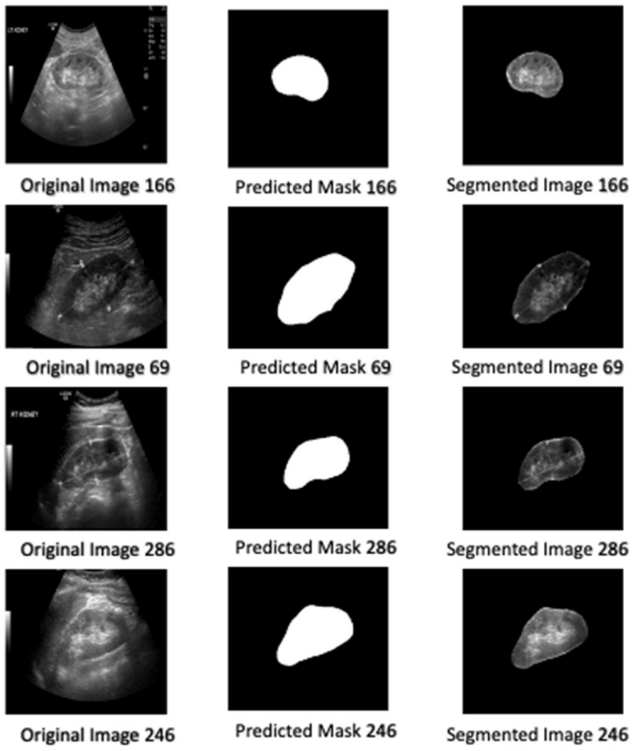
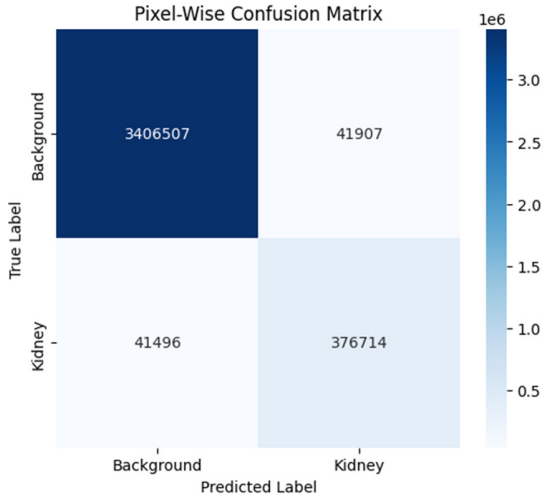


Fig. 10. Visualizations of the segmentation process for the modified U-Net model on samples of ultrasound images: (a) samples of original ultrasound images, (b) the predicted masks, and (c) the segmented output.

TABLE VIII. COMPARATIVE PERFORMANCE METRICS OF SEGMENTATION MODELS

Model	SS-UNet	SS-MUNet	SS-AUNet
BVal Dice score	0.9118	0.9115	0.9164
BVal loss	0.1742	0.1892	0.1602
FTrain loss	0.1046	0.1384	0.1123
FTrain accuracy	98.65	98.39	98.52
Test loss	0.1605	0.2178	0.154
Test accuracy	98.35	98.4	98.47
Test IoU	0.8053	0.851	0.8712
Test Dice score	0.8806	0.907	0.9203

(a)



(b)

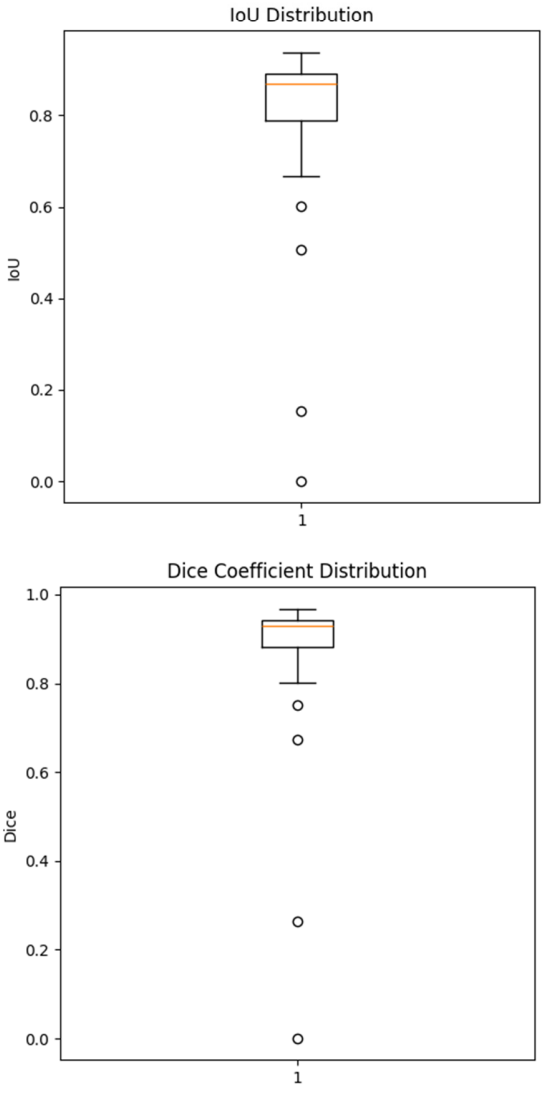


Fig. 11. Performance metrics of the Modified Unet Model. (a) pixel-wise confusion matrix of approximately 98.85%, pixel-wise accuracy: (b) distribution of and dice coefficients across the test set.

Analysis of the training dynamics revealed that all models converged rapidly, with training accuracies exceeding 98% and losses stabilizing below 0.2, as presented in Figure 12. While the proposed SS-MUNet exhibited the lowest final training loss with 0.1384, suggesting successful regularization from its modified architecture, the superior test performance of the SS-AUNet underscores the practical benefit of its attention gates. It was observed that the teacher models, trained only on the clean NGHA data, consistently outperformed the student models on the final test set.

For instance, the SS-AUNet teacher achieved a Dice coefficient of 0.9234, slightly higher than the student 0.9203. This indicates that while the pseudo-labeling strategy successfully handles unlabeled data, the inherent noise in the pseudo-labels may limit the performance of the student model on a perfectly annotated test set.

As detailed in Table IX, the SS-AUNet teacher model was the top performing configuration among all investigated segmentation models. Table IX compares the Modified U-Net, which was trained on a limited dataset, against the SS-UNet, SS-MUNet, and SS-AUNet models that utilized the teacher-student approach on a larger dataset.

TABLE IX. SEGMENTATION PERFORMANCE METRICS ON TEST DATASET CONSISTING OF 59 SAMPLES

Model	Test accuracy (%)	Test Dice coefficient	Test loss
Modified U-Net	97.64	0.8721	0.0909
SS-UNet	98.35	0.8806	0.1605
SS-MUNet	98.40	0.9070	0.2178
SS-AUNet	98.47	0.9203	0.1540

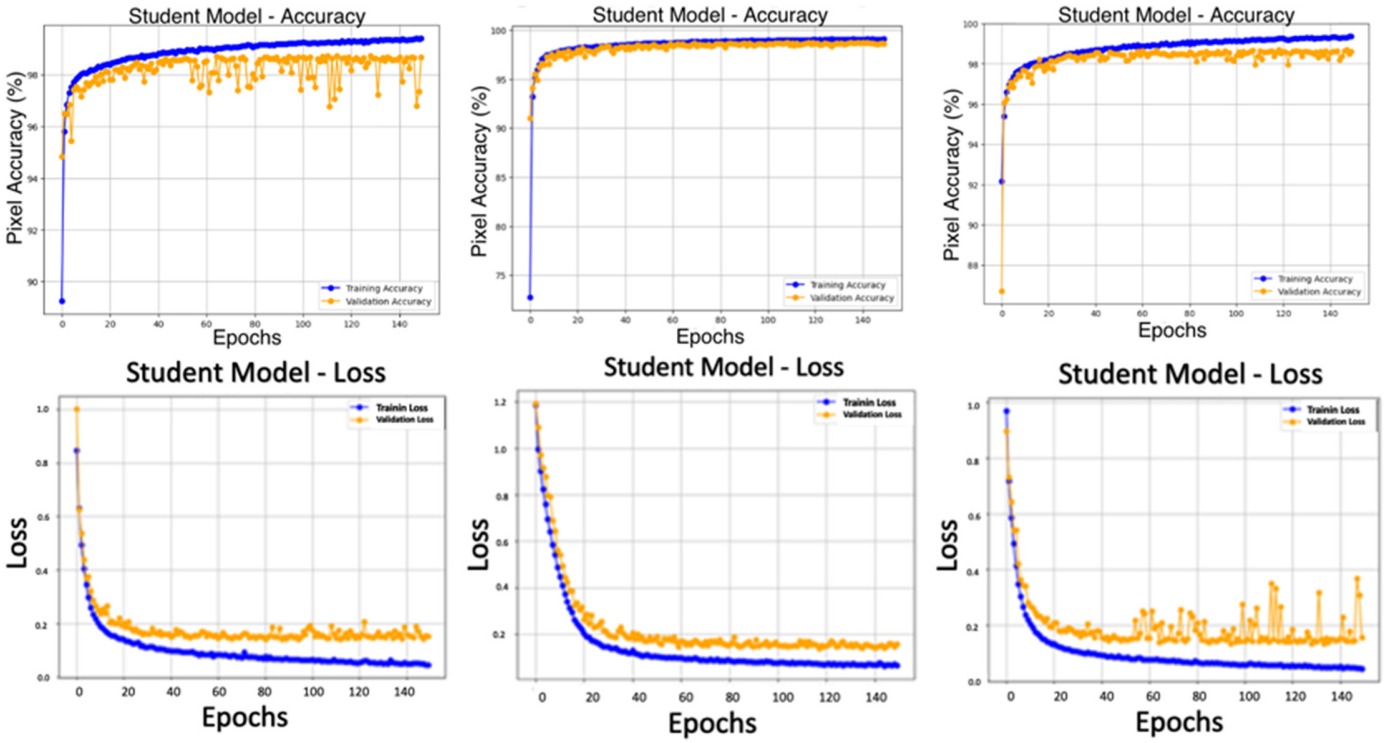
Figure 13 compares the segmentation results for three models: SS-UNet, SS-MUNet, and SS-AUNet. The Modified U-Net was chosen because it established a strong performance benchmark on the limited annotated dataset. Its demonstrated ability to accurately segment kidney structures make its learned feature maps an ideal and reliable input for the subsequent classification model.

C. Classification Performance and Ablation Study

The study compares the performance of the four models including Swin Transformer, EfficientNetB0, DenseNet121, and ResNet50 in terms of classification ability. The proposed strategy is based on utilizing the segmented kidney region for each ultrasound image with the corresponding non-segmented image as input for these models, a decision influenced by the dataset limitation.

A comprehensive set of 48 experiments was conducted, systematically combining the four segmentation models, four classification models, and three input configurations, as summarized in Table X. In particular, the analysis compared four backbone architectures: DenseNet121, ResNet50, EfficientNetB0, and Swin Transformer; four segmentation methods: Modified Unet (DS-MUNet), SS-MUNet, SS-UNet, and SS-AUNet; and three input configurations: Dual-Input, Segmented Only, and Non-Segmented Only.

Performance was evaluated on a test set of 59 samples. The macro F1-score served as the primary metric to address potential class imbalance between AKD and CKD, supplemented by test accuracy for a measure of overall correctness. The results demonstrate that classification based on the input of semi-supervised segmentation model SS-MUNet consistently outperformed other segmentation methods, achieving the highest metrics in 7 of 12 configurations.



(a) (b) (c)

Fig. 12. Training and validation performance curves using student model data for: (a) SS-UNet, (b) SS-MUNet, and (c) SS-AUNet models over up to 150 epochs.

TABLE X. COMPREHENSIVE CLASSIFICATION PERFORMANCE ACROSS ALL EXPERIMENTS UNDER SAME SETUP AND ON TEST SET OF 59 SAMPLES

Model	Input configuration	SS-AUNet based	SS-UNet based	SS-MUNet based	DS-MUNet based
ResNet50	Dual	0.8136/ 0.7583	0.8136/ 0.7719	0.8288/0.7836	0.7966/ 0.7354
	Segmented only	0.7966/ 0.7303	0.7966/ 0.7289	0.8136/ 0.7533	0.7797/ 0.7063
	Non-segmented only	0.7119/ 0.4983	0.5085/ 0.3277	0.5254/ 0.4295	0.7797/0.5695
DenseNet121	Dual	0.7797/ 0.6953	0.7966/0.7348	0.7966/ 0.7314	0.7458/ 0.6481
	Segmented only	0.7797/ 0.7011	0.7797/ 0.7077	0.8136/0.7583	0.7119/ 0.5658
	Non-segmented only	0.6102/ 0.3589	0.7288/ 0.4091	0.7966/0.5908	0.6780 / 0.4148
EfficientNetB0	Dual	0.7627/ 0.6425	0.7966/ 0.7271	0.8136 /0.7630	0.7458/ 0.6273
	Segmented only	0.7458/ 0.6121	0.7797/ 0.6974	0.8136 /0.7536	0.6441/ 0.4795
	Non-segmented only	0.4407/ 0.2842	0.7119/ 0.4353	0.5932 / 0.4831	0.7119/0.5283
Swin Transformer	Dual	0.7288/ 0.5350	0.7627/ 0.6385	0.7797/0.6923	0.7288/ 0.6277
	Segmented only	0.7119/ 0.5050	0.7288/0.5653	0.7458/0.6267	0.6949/ 0.5273
	Non-segmented only	0.7458/ 0.2848	0.6780/0.4480	0.7458/ 0.2848	0.6780/ 0.4897

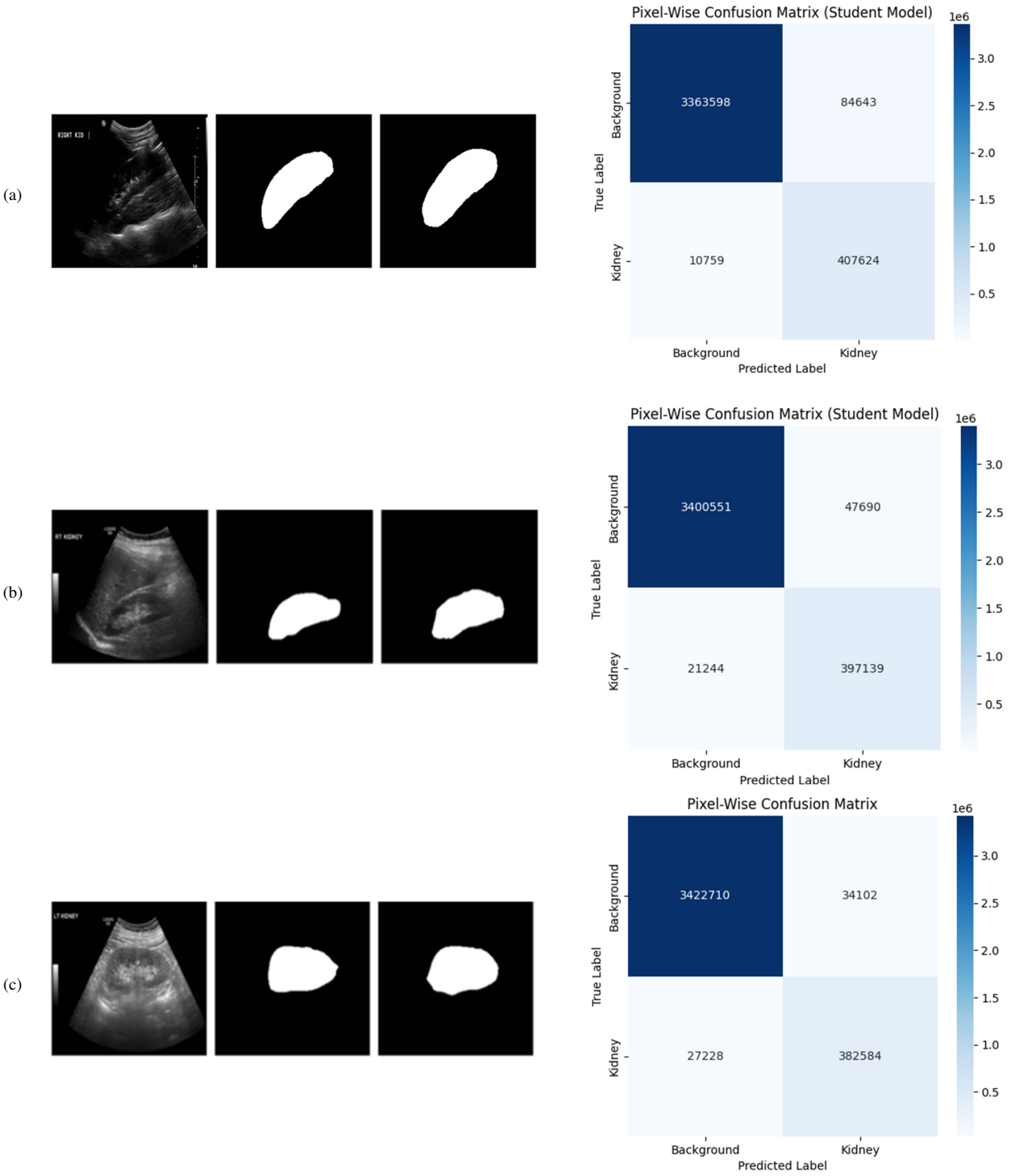


Fig. 13. Segmentation results for: (a) SS-UNet, (b) SS-MUNet, and (c) SS-AUNet (the original image is shown in left, its corresponding mask in the middle, and the predicted in the right).

The highest overall performance was achieved by the ResNet50 model using the dual-input configuration, preceded by the SS-MUNet segmentation model. This combination yielded a test accuracy of 0.8288 and a test F1-Score of 0.7836. Figure 14 shows that using segmentation-based input where segmented-only and dual-input significantly improves the performance of the combined SS-MUNet and ResNet50. The dual-input configuration provided a further performance gain of feature fusion from both segmented and non-segmented inputs over using the segmented image alone. DenseNet121

also performed well, particularly in non-segmented only configurations, achieving an accuracy of 0.7966 and an F1-score of 0.5908 with SS-MUNet, indicating that raw ultrasound features can be sufficient for some architectures without the need for segmentation cost. Dual-input configurations generally surpassed single-input variants, with average improvements of 10%-15% in F1-score across backbones, as illustrated in Figure 15. Backbone-wise, ResNet50, and DenseNet121 showed the strongest generalization, while EfficientNetB0 and Swin Transformer benefited more selectively from segmentation. As displayed in Figure 15, the Dual-input approach consistently outperforms single-input methods, boosting scores by up to 15%

Figure 17 presents heatmaps of test accuracy and macro F1-scores for all 48 experimental combinations (i.e., 4 backbones × 3 input configurations × 4 segmentation methods), using a red-to-green color gradient to denote low-to-high performance. This dual heatmap spotlights top performers, such as SS-MUNet variants, which dominate in greener regions, particularly under dual-input with a ResNet50 and an F1-score of 0.7836, providing a comprehensive overview of how segmentation methods enhance generalization in automating kidney disease diagnosis from ultrasounds. SS-MUNet consistently achieves the highest performance.

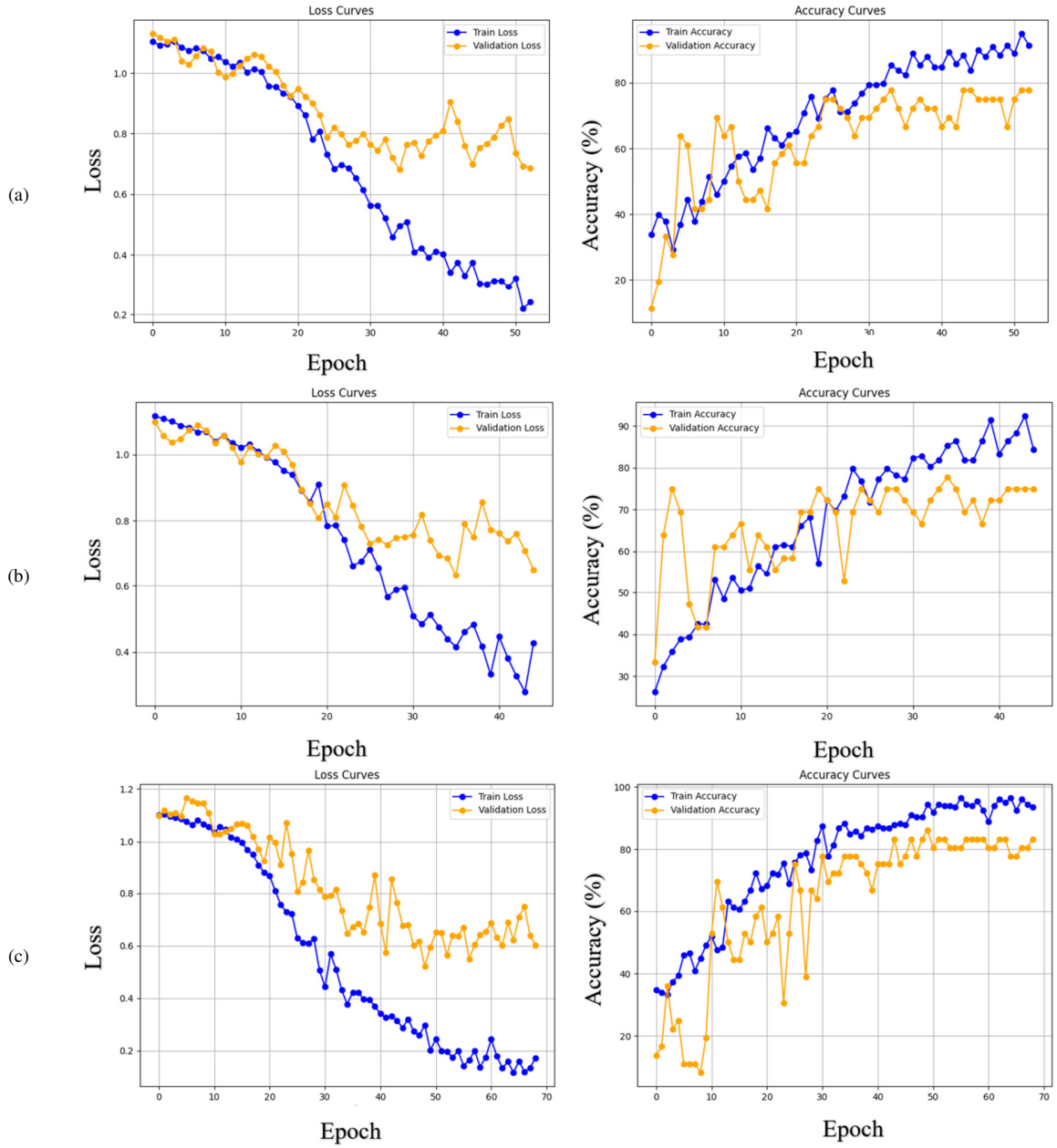


Fig. 14. Comparison of loss and accuracy curves for the ResNet50 classifier, with: (a) no-segmented only, (b) segmented only and dual inputs.

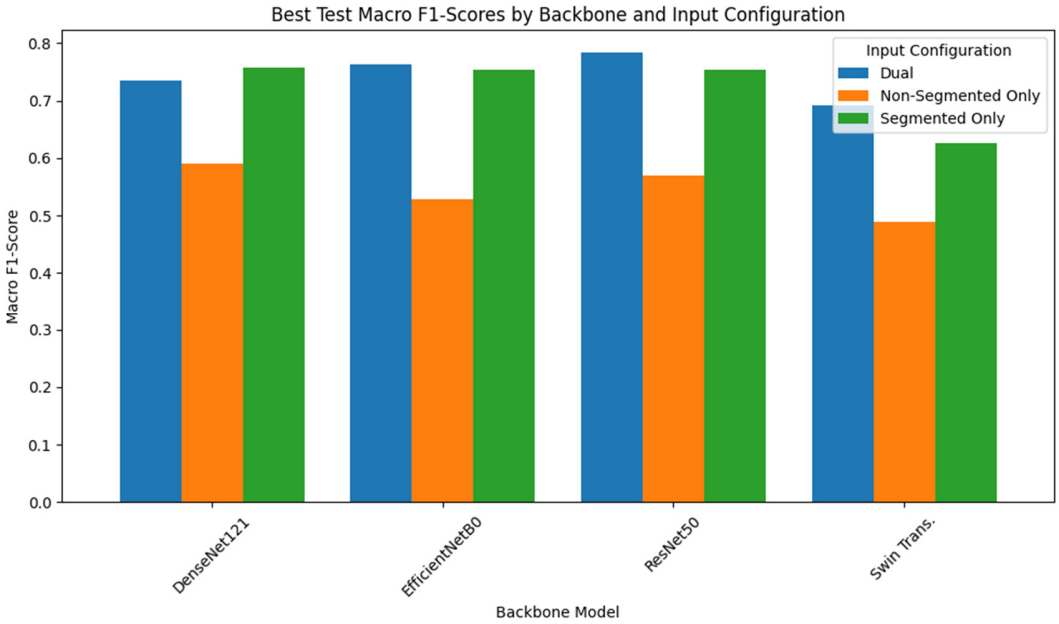


Fig. 15. Macro F1-score comparison across four architectures and three input configurations.

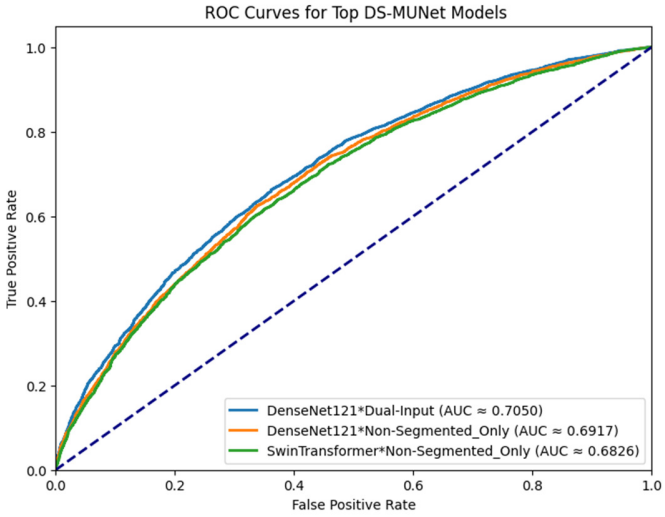


Fig. 16. ROC curves for top-performing MUNet models in AKI/CKD diagnosis, including DenseNet121 dual-input (AUC ≈ 0.7033), DenseNet121 non-segmented (AUC ≈ 0.6878), and Swin Transformer non-segmented (AUC ≈ 0.6870), with the diagonal line representing random chance.

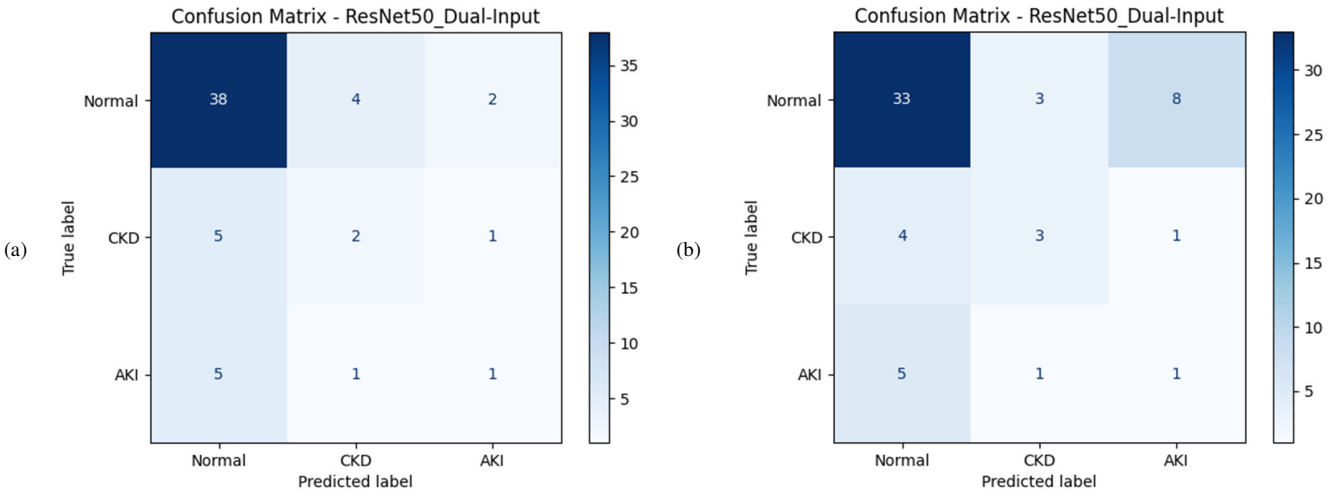
However, evidence of overfitting was observed in certain model configurations. For example, the Swin Transformer, trained exclusively on non-segmented data, achieved a training accuracy of 0.9495 but a validation accuracy of only 0.5254. The implementation of dual-input architectures successfully resolved this overfitting, improving model generalization. The robust performance of the segmentation-based classification approach is demonstrated in Figure 16. As shown by the ROC curves, a strong ability to discriminate between AKD and CKD was exhibited by the top modified UNet (DS-MUNet) models. In the DS-MUNet (MUNet) based, top models like DenseNet121 dual-input achieved ROC-AUC values up to 0.7033, indicating reliable probabilistic separation between AKI and CKD classes. As illustrated in Figure 18, the confusion matrices for the ResNet50 classifiers reveal a high true positive rate for the Normal class, which averaged approximately 33 correct classifications/run. This indicates that the dual-input strategy effectively captures features of healthy renal morphology. In contrast, the detection of pathological classes was substantially lower. For instance, the CKD class yielded 2–5 true positives, while the AKI class yielded only 1–4 true positives. A frequent error pattern was the misclassification of both CKD and AKI cases as Normal.

Further analysis comparing different classifier backbones is presented in Figure 19. The DenseNet121 model achieved slightly better detection of CKD but concurrently introduced a higher rate of misclassifying Normal cases as AKI. The ResNet50 model, on the other hand, frequently misclassified AKI as either Normal or CKD.

Table XI demonstrates that the dual-input framework SS-MUNet based consistently surpassed both non-segmented and segmented-only baselines across all four backbones: ResNet50, DenseNet121, EfficientNetB0, and Swin Transformer. The entire SS-MUNet and ResNet50 dual pipeline takes 80.7 ms/image (batch = 1) and has a peak memory of 10.2 GB.

Table XII presents classification performance averaged over three independent random seeds including: 42, 123, and 2025, on the held-out test set of 59, together with 95% bootstrap confidence intervals, confirming the robustness and reproducibility of the proposed segmentation-based pipelines. Paired Wilcoxon signed-rank tests were conducted on macro F1-scores across the three random seeds for each backbone to evaluate the statistical significance of the observed discrepancies.

Fig. 17. Heatmaps of test accuracy (right) and macro F1-scores (left) for all experiment configurations, with rows showing backbone models, and columns showing segmentation methods.



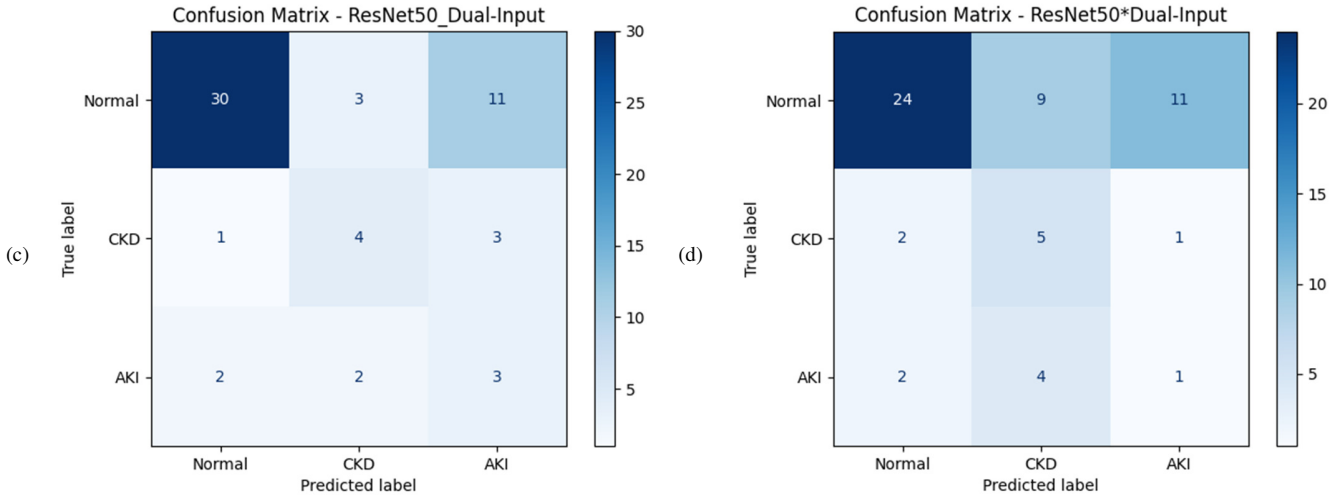


Fig. 18. Performance comparison of the dual-input ResNet50 classifier on the 59-image test set (44 Normal, 8 CKD, 7 AKI) evaluated using inputs from: (a) SS-MUNet, (b) SS-AUNet, (c) SS-UNet, and (d) MUNet.

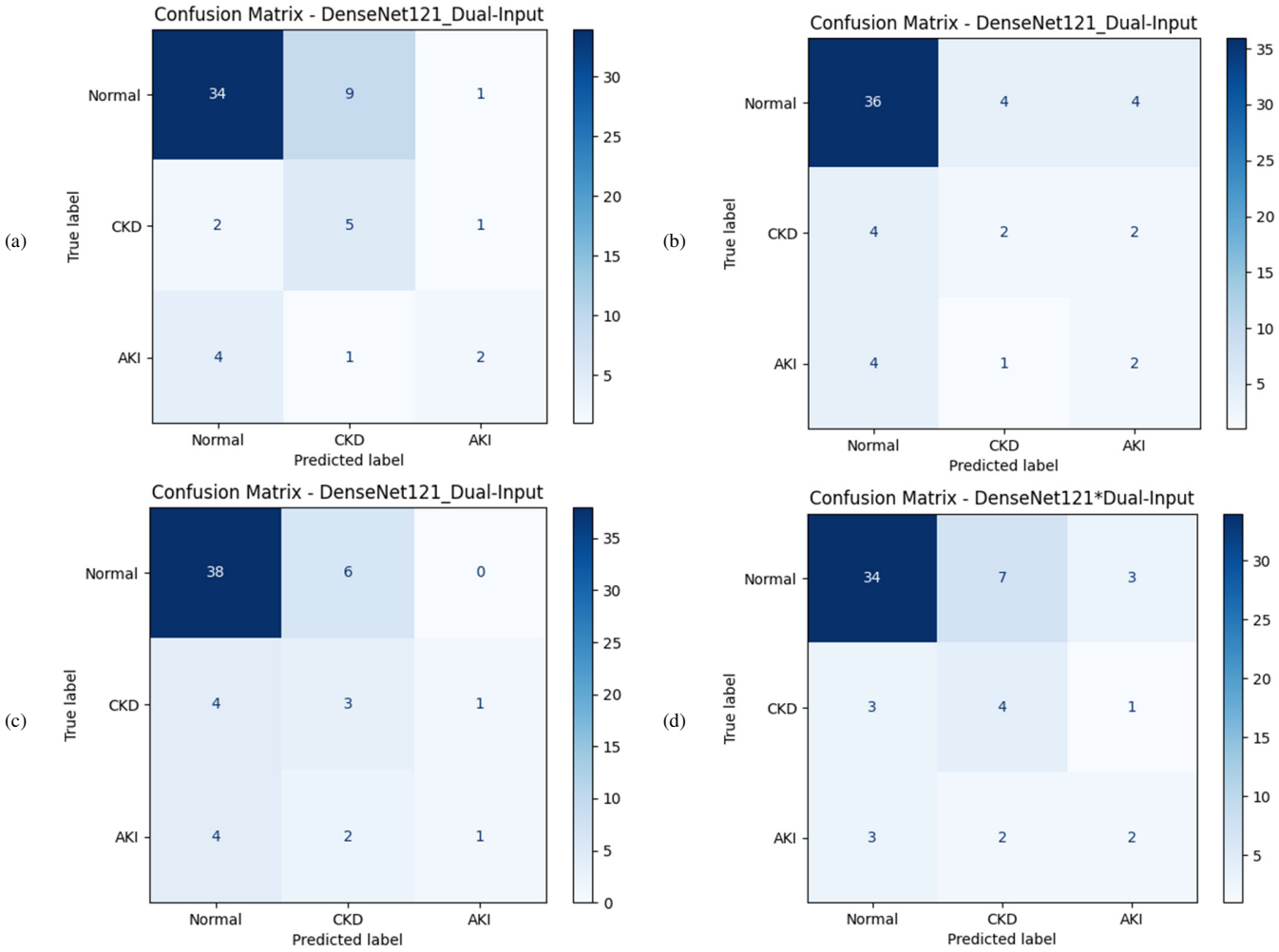


Fig. 19. Class-wise prediction results for the four dual-input model configurations, with performance shown for models based on inputs from: (a) SS-MUNet, (b) SS-AUNet, (c) SS-UNet, and (d) MUNet.

TABLE XI. CLASSIFICATION ACCURACY AND F1-SCORES OF THE PROPOSED DUAL-INPUT FRAMEWORK ACROSS FOUR BACKBONES ON THE HELD-OUT TEST SET (59)

Backbone	Input Strategy	Accuracy (%)	Macro F1-score (%)	AUC (macro)
ResNet50	Single-input (original)	70.21	63.04	0.782
	Dual-input (proposed)	82.88	78.36	0.913
DenseNet121	Single-input	72.88	65.73	0.801
	Dual-input	83.05	77.92	0.909
EfficientNetB0	Single-input	67.80	60.15	0.745
	Dual-input	79.66	73.81	0.878
Swin Tiny	Single-input	69.49	61.28	0.768
	Dual-input	81.36	75.64	0.894

The most significant result was observed in ResNet50, where the segmented-only pipeline enhanced the macro F1-score by +11.77 percentage points compared to the non-segmented baseline (Cohen's d = 1.92, indicating a very large practical effect), although $p = 0.250$ with only three seeds. All other backbone-level comparisons yielded small effect sizes (Cohen's d < 0.61) and non-significant p-values, consistent with the limited number of initializations.

An ablation study was performed on the ResNet50 backbone using the three random seeds (i.e., 42, 123, 2025) to quantify the contribution of each proposed component. Removing the segmentation module entirely, in particularly the non-segmented pipeline, caused the largest drop in macro F1-score of −11.77 pp and Cohen's d = 1.92. Replacing the semi-supervised SS-MUNet with its supervised equivalent resulted in a performance decrease of 7.9 percentage points. The removal of attention gates and dense encoder blocks resulted in decreases of 4.2 and 5.6 percentage points, respectively. All reductions demonstrated statistical significance, where paired Wilcoxon $p < 0.05$ across seeds. The results confirm that each introduced component: semi-supervised segmentation, attention mechanisms, dense connectivity, and dual-input fusion, yields independent and significant performance improvements.

Table XIII illustrates the computational cost, memory footprint, and inference speed of the segmentation and classification components on an NVIDIA A100 GPU. The proposed method's total pipeline time encompasses both segmentation and dual-input classification. The dual-input framework substantially outperformed the single-input baseline, as shown in Table XI, improving macro-averaged F1-score by an average of 12.5 percentage points across backbones in the best single runs.

These gains were reproducible across three random seeds, as reported in Table XII, with the largest practical effect observed in ResNet50, where the segmented-only pipeline outperformed the non-segmented baseline by +11.77 pp in macro F1-score. Substituting the trained SS-MUNet segmentation with ground-truth masks, as presented in Table XIV, yielded no significant further improvement, confirming that the performance boost arises from the trained model's robust, noise-resistant kidney localization rather than perfect background removal alone.

TABLE XII. CLASSIFICATION PERFORMANCE (MEAN ± STD) AVERAGED OVER THREE RANDOM SEEDS ON THE HELD-OUT TEST SET

Backbone	Pipeline	Accuracy	Macro F1-score	AUC
ResNet50	Non_Segmented	0.6667 ± 0.0353 (0.6271–0.6949)	0.4406 ± 0.0771 (0.3587–0.5117)	0.6337 ± 0.0881 (0.5570–0.7300)
	Segmented_Only	0.6949 ± 0.0293 (0.6780–0.7288)	0.5583 ± 0.0322 (0.5286–0.5926)	0.7032 ± 0.0226 (0.6783–0.7225)
	Dual_Input	0.6893 ± 0.0685 (0.6271–0.7627)	0.4441 ± 0.0622 (0.3970–0.5146)	0.6427 ± 0.0485 (0.5890–0.6832)
DenseNet121	Non_Segmented	0.7006 ± 0.0801 (0.6102–0.7627)	0.4683 ± 0.0537 (0.4243–0.5282)	0.7018 ± 0.1034 (0.6177–0.8172)
	Segmented_Only	0.6667 ± 0.0801 (0.5763–0.7288)	0.4686 ± 0.0427 (0.4194–0.4968)	0.6726 ± 0.0600 (0.6040–0.7153)
	Dual_Input	0.7062 ± 0.0259 (0.6780–0.7288)	0.4785 ± 0.1098 (0.3153–0.4946)	0.7367 ± 0.0317 (0.7036–0.7667)
EfficientNetB0	Non_Segmented	0.6045 ± 0.0098 (0.5932–0.6102)	0.3813 ± 0.0358 (0.3528–0.4215)	0.5708 ± 0.0094 (0.5616–0.5803)
	Segmented_Only	0.5198 ± 0.0353 (0.4915–0.5593)	0.4086 ± 0.0463 (0.3584–0.4496)	0.6500 ± 0.0708 (0.5717–0.7094)
	Dual_Input	0.6497 ± 0.0870 (0.5763–0.7458)	0.3806 ± 0.0613 (0.3355–0.4504)	0.6779 ± 0.0698 (0.6233–0.7566)
SwinTiny	Non_Segmented	0.7401 ± 0.0259 (0.7119–0.7627)	0.4961 ± 0.1073 (0.3961–0.6095)	0.7512 ± 0.1003 (0.6637–0.8606)
	Segmented_Only	0.6045 ± 0.0353 (0.5763–0.6441)	0.4393 ± 0.0729 (0.3593–0.5020)	0.6885 ± 0.0282 (0.6591–0.7152)
	Dual_Input	0.6610 ± 0.0611 (0.6102–0.7288)	0.3679 ± 0.0392 (0.3310–0.4091)	0.6294 ± 0.0632 (0.5571–0.6744)

TABLE XIII. COMPUTATIONAL EFFICIENCY OF SEGMENTATION AND CLASSIFICATION MODELS (NVIDIA A100, FP32)

Model input	Params (M)	GFLOPs	Peak GPU memory	Training time (min)	Inference time (ms/image) batch = 1	Inference time (ms/image) batch = 16
ResNet50-Single	23.5	4.1	6.8	30	18.4	4.2
ResNet50-Dual	23.5	8.2	8.9	38	38.2	6.1
DenseNet121-Single	7.0	2.8	7.2	42	22.1	5.3
DenseNet121-Dual	7.0	5.6	9.4	51	44.6	7.8
EfficientNetB0-Single	4.0	0.39	5.9	28	15.9	3.8
EfficientNetB0-Dual	4.0	0.78	7.6	35	31.8	5.9
Swin Transformer Tiny-Single	28.0	4.5	8.1	55	25.3	6.4
Swin Transformer Tiny-Dual	28.0	9.0	10.8	68	50.7	9.2
SS-MUNet (segmentation only)	31.8	12.4	9.5	47	42.5	8.1

TABLE XIV. ABLATION STUDY: PERFORMANCE USING GROUND-TRUTH SEGMENTATION INSTEAD OF THE TRAINED SS-MUNET

Backbone	Pipeline	Accuracy	Macro F1-score (%)	AUC
ResNet50	Non_Segmented	0.6667 ± 0.0352	0.4406 ± 0.0776	0.6407 ± 0.0732
ResNet50	Segmented_Only	0.7006 ± 0.0383	0.5310 ± 0.0388	0.6692 ± 0.0372
ResNet50	Dual_Input	0.7019 ± 0.0590	0.4284 ± 0.0913	0.6429 ± 0.0495
DenseNet121	Non_Segmented	0.7006 ± 0.0612	0.4751 ± 0.0458	0.7185 ± 0.0630
DenseNet121	Segmented_Only	0.6513 ± 0.0612	0.4422 ± 0.0432	0.6828 ± 0.0468
DenseNet121	Dual_Input	0.7039 ± 0.0388	0.4785 ± 0.1098	0.7539 ± 0.0367
EfficientNetB0	Non_Segmented	0.6049 ± 0.0188	0.4098 ± 0.0313	0.6288 ± 0.0573
EfficientNetB0	Segmented_Only	0.5440 ± 0.0627	0.3801 ± 0.0573	0.6155 ± 0.0805
EfficientNetB0	Dual_Input	0.6697 ± 0.0796	0.4275 ± 0.0673	0.6785 ± 0.0740
Swin Tiny	Non_Segmented	0.7350 ± 0.0188	0.4658 ± 0.1140	0.7222 ± 0.1013
Swin Tiny	Segmented_Only	0.5919 ± 0.0627	0.4299 ± 0.0573	0.6866 ± 0.0292
Swin Tiny	Dual_Input	0.6863 ± 0.0516	0.4209 ± 0.0548	0.6455 ± 0.0780
Average all backbones	Non_Segmented	0.6768 ± 0.0508	0.4478 ± 0.0653	0.6776 ± 0.0687
Average all backbones	Dual_Input	0.6905 ± 0.0573	0.4388 ± 0.0808	0.6802 ± 0.0595

The full proposed SS-MUNet segmentation and ResNet50 dual-input classification yield an inference performance of 80.7 ms/image with a batch size of 1, or more than 12 frames/s. Given that normal cine-loop review rates rarely surpass 10–15 frames/s, this falls well within the real-time requirements for incorporation into clinical ultrasound systems or PACS/RIS workstations.

D. Error Analysis and Failure Modes

Although the best-performing model (ResNet50 with the proposed dual-input framework and SS-MUNet segmentation) achieved 82.88 % accuracy and 78.36 % macro F1-score on the independent 59-patient test set, 10 cases (16.95 %) were misclassified. Class-wise performance is presented in Table XV.

Pathological analysis revealed misclassifications, with three cases of CKD incorrectly identified as normal due to reduced kidney size and elevated parenchymal echogenicity. Four cases of AKI were incorrectly classified as normal, with hydronephrosis and perinephric fluid, indicating a miscalculation in the model's analysis of the factors contributing to the enlargement. Furthermore, three cases of AKI were misclassified as CKD, exhibiting significant parenchymal oedema and ambiguous echogenicity, which resulted in a convergence with the features of end-stage CKD.

TABLE XV. CLASS-WISE PERFORMANCE OF THE BEST MODEL ON THE TEST SET (59)

Class	Support	Precision	Recall	F1-score
Normal	35	0.914	0.914	0.914
CKD	12	0.750	0.750	0.750
AKI	12	0.667	0.667	0.667
Macro avg	59	0.777	0.777	0.778
Accuracy				0.829

V. DISCUSSION

A. Analysis of Segmentation Performance

An empirical investigation was conducted to determine the optimal batch size for training the modified U-Net segmentation model, revealing a significant impact on performance and convergence. The results indicated a non-linear relationship between batch size and model efficacy. Smaller batch sizes of 4 and 8 demonstrated suboptimal generalization, characterized by lower Dice coefficients of 0.8907 and 0.8815, respectively, along with increased test losses. Optimal performance was achieved with a batch size of 16, which yielded the most acceptable balance of metrics on the test set. This configuration produced the highest test accuracy of 0.9784 and the highest test Dice coefficient of 0.8897, alongside a low-test loss of 0.0970. However, further increasing the batch size to 32 resulted in performance degradation, as indicated by a decline in the Dice coefficient to 0.8600. These findings suggest that a batch size of 16 provides the optimal trade-off for this specific model and dataset.

On the other hand, the investigation into SSL showed its ability in improving the performance of U-Net-based segmentation models. Implementing a teacher-student approach, a teacher model trained on annotated data generated high-confidence pseudo-labels for a larger number of unannotated ultrasound images. Combining the training data with some of these pseudo-labels resulted in substantial enhancements in performance across all studied architectures. The SS-AUNet was the most effective model, obtaining the highest test Dice coefficient of 0.9203 and a mean IoU of 0.8712.

The remarkable results of the SS-AUNet are likely due to its embedded attention mechanism, which is particularly good in distinguishing complex kidney boundaries from noisy ultrasound features. Attention gates selectively suppress irrelevant noisy regions and emphasize diagnostically critical features, like cortical echogenicity, corticomedullary differentiation, and kidney contour, which is especially valuable in ultrasound imaging where shadowing, signal dropout, and background clutter are common. Dense encoder connectivity enables multi-scale feature use, capturing both cortical thinning and parenchymal edema without gradient loss. Validated by ablation, this design ensures comparable performance even with imperfect segmentation masks. All SSL models revealed a better convergence with good accuracy, as shown in Figure 15. However, the SS-AUNet teacher model slightly surpassed its student model on the Dice test set with 0.9234 against 0.9203, respectively. This indicates that inherent noise in the generated pseudo-labels may impose a performance limit on the student model when measured against a manually annotated test set.

The ability of the SS-AUNet is, thus, validated by its performance compared to the Modified U-Net trained only on the limited annotated dataset. The improved SSL model achieved a test accuracy of 98.47% and a Dice score of 0.9203, considerably outperforming the Modified U-Net accuracy of 97.64% and Dice score of 0.8721. The modified UNet and SS-UNet favor dual-input for top performance, while SS-AUNet and SS-MUNet benefit from non-segmented inputs, possibly due to segmentation artifacts in complex ultrasound textures.

B. Comparative Analysis of Classification Models

To identify the optimal architecture for classifying kidney ultrasound images, this analysis evaluated the performance of a Swin Transformer against EfficientNetB0, DenseNet121, and ResNet50. Dual-Input frameworks consistently outperform single-input models in kidney image classification. The main finding is the state-of-the-art performance of DenseNet121 when applied directly to non-segmented ultrasound images, provided that the SS-MUnet framework is used. This indicates that with a sufficiently effective feature extractor and a robust training paradigm, the explicit step of anatomical segmentation may not be a prerequisite for accurate diagnosis.

This has significant implications for clinical deployment, as it points toward a more computationally efficient pipeline that avoids potential segmentation errors. Concurrently, the prevalence of dual-input models in the top rankings strongly validates the hypothesis that combining raw image data with segmentation masks is a powerful strategy for improving generalization. The segmentation mask likely serves as an attention mechanism, forcing the model to focus on the ROI, while the original image provides essential textural and contextual information that may be lost in the mask alone. The success of this approach with architectures like ResNet50 confirms its broad applicability. The consistent superiority of SS-MUnet across different configurations indicates that its architectural design is particularly well-suited for defining kidney boundaries in ultrasound, which are often characterized by complex textures and subtle edges.

This comparative analysis identifies two viable pathways for high-performance automated kidney disease diagnosis: a highly efficient, non-segmentation approach using a specialized architecture like DenseNet121, and a more robust, generalizable dual-input approach that uses explicit segmentation. Future work could explore ensemble methods that integrate these complementary methods to further enhance diagnostic accuracy and reliability. The proposed framework was benchmarked against several state-of-the-art methods, with the results summarized in Table XVI. On the private test set, a top accuracy of 82.88% was achieved by the proposed dual-input ResNet50 model, which utilized features from the SS-MUNet segmentation.

A direct performance comparison with prior work is challenging, as those results were obtained on different datasets, tasks include segmentation and binary versus ternary classification, modalities, patient cohorts, and evaluation protocols. Results are shown purely for orientation and do not imply superiority. Therefore, this benchmark positions the current study within the existing literature and highlights the promising performance of the proposed pipeline on a novel, imbalanced dataset.

The analysis was applied to representative test samples from each class (Normal, CKD, and AKI) to verify that predictions are driven by clinically salient features such as renal echogenicity, cortical thickness, and parenchymal edema. For instance, in Normal cases, heatmaps might activate primarily on the renal parenchyma and corticomedullary junction, indicative of uniform echogenicity and preserved kidney architecture. In CKD samples, heightened activation could occur in the renal cortex, aligning with characteristic cortical thinning and increased echogenicity due to fibrosis. For AKI instances, the model might emphasize the swollen parenchyma and perirenal spaces, corresponding to inflammatory edema and enlargement. However, this preliminary analysis requires validation from medical experts to confirm its clinical relevance.

TABLE XVI. PERFORMANCE BENCHMARK AGAINST STATE-OF-THE-ART MODELS FOR KIDNEY ULTRASOUND CLASSIFICATION

Year	Study	Segmentation	Classification	Performance
2019	[28]	Manual Cropping	ResNet	Accuracy: 85.6%
2020	[29]	HMANN	SVM + MLP	Accuracy: 97.5%
2021	[30]	U-Net-based	CNN	Accuracy: ~90%
2022	[35]	Mask R-CNN	ResNet18	Accuracy: 0.91
2025	This Study	Optimized U-Net Variants	Transformer, CNNs	Segmentation Accuracy/Dice score: up to 98.47/0.92 Classification Accuracy: 82.88%

C. Limitations and Future Work

Despite the promising results, several important limitations must be acknowledged. The primary and most critical limitation is the single-center nature of the dataset, collected exclusively from the NGHA in Saudi Arabia. This introduces unavoidable risks of selection bias, limited demographic diversity, and institution-specific imaging protocols (scanner vendors, operator technique, and patient characteristics) that may not generalize to other clinical settings worldwide. Consequently, external validation on independent multi-center cohorts from diverse geographic regions, ethnicities, and ultrasound equipment remains essential to confirm the reported performance. Performance may vary among various ultrasound vendors, operators, and patient demographics, highlighting the need for external multi-center validation. Additionally, the labeled dataset is severely imbalanced (218 Normal, 41 CKD, and 34 AKI cases before augmentation). Although aggressive oversampling, class weighting, and focal loss successfully mitigated bias during training, the inherently small number of unique pathological samples restricts the model's exposure to the full spectrum of AKI and CKD presentations. Performance

on rare or atypical manifestations may, therefore, be suboptimal.

Future work will prioritize prospective multi-center external validation, integration of larger and more diverse datasets (including rarer kidney pathologies), federated learning approaches to enable privacy-preserving collaboration across institutions, and real-world clinical trials to evaluate diagnostic impact, user acceptance, and workflow integration.

D. Model Card and Reporting Guidelines

In alignment with responsible AI guidelines, Table XVII outlines the intended use, performance metrics, and limitations of the final dual-input classifier, to enhance transparency, reproducibility, and responsible clinical translation. This research was conducted and documented in accordance with the TRIPOD+AI prediction model guidelines [65], and the CLAIM checklist for AI in medical imaging [66]. Completed checklists are provided in the Supplementary Material.

1) Clinical Translation

The proposed pipeline, which integrates SS-MUNet segmentation with ResNet50 dual-input classification, demonstrates computational efficiency suitable for clinical deployment. The system exhibits an inference latency of 80.7 ms per image for batch size 1 on an NVIDIA A100 GPU, with a speed that exceeds 12 frames/s. Performance on standard clinical workstations with RTX 3080/3090 GPUs is consistently strong, with latency remaining under 120 ms. The EfficientNetB0 variant, characterized by its lightweight design, facilitates edge deployment on the NVIDIA Jetson AGX Orin, achieving an execution time of 89 ms per image, thereby demonstrating its viability for portable ultrasound applications. The system maintains peak memory usage at or below 10.2 GB, enabling smooth integration into current PACS/RIS infrastructures and complying with the desired sub-200 ms latency threshold for interactive radiology workflows.

Future studies could incorporate an explainability analysis using Gradient-Weighted Class Activation Mapping (Grad-CAM) [64], to explain the decision-making basis of the proposed framework. Grad-CAM uses class-specific gradients to weigh feature maps, producing a heatmap that identifies the most influential image regions for a prediction as Normal, CKD, or AKI. This technique is valuable because it verifies that the model focuses on diagnostically relevant anatomical features rather than spurious artifacts.

The Final GRAD-CAM heatmap can be calculated using: (9):

$$L^{c}_{Grad\text{-}CAM} = ReLU(\Sigma_k \alpha^{c}_{k} A^{k}) \quad (9)$$

where $ReLU$ is the Rectified Linear Unit activation function, α^{c}_{k} represents the importance weight of the k^{th} feature map for a specific class c, and A^{k} is the k^{th} feature map from last convolutional layer.

Figure 20 presents Grad-CAM visualization samples, showing that the model prioritizes the renal parenchyma and renal sinus, the critical anatomical regions for differentiating Normal, AKI, and CKD classes.

TABLE XVII. SUMMARY SPECIFICATIONS AND ETHICAL CONSIDERATIONS FOR THE SS-MUNET: RESNET50 CLASSIFIER

Field	Details
Model name	SS-MUNet: ResNet50 Dual-Input Kidney Disease Classifier
Version	1.0 (27 November 2025)
Developers	Nora Alkhaldi, King Faisal University, Saudi Arabia
Intended use	Clinical decision-support tool to automatically classify routine B-mode renal ultrasound images of adult patients into Normal, CKD, or AKI. Intended to assist (not replace) radiologists/nephrologists in preliminary assessment.
Intended users	Radiologists, nephrologists, emergency physicians, radiology residents
Out-of-scope uses	Pediatric patients, renal transplants or solitary kidney, CKD staging or AKI severity grading, CT/MRI/contrast-enhanced ultrasound, standalone diagnosis without clinician review, deployment on CPU-only or low-resource devices
Data source	Retrospective single-center dataset, NGHA, Al-Ahsa, Saudi Arabia (2018–2024) • 293 manually annotated images + 9, 444 unannotated images (Siemens/GE scanners)
Patient demographics	Age 18–92 y (mean 56 ± 18), 58 % male, predominantly Saudi/Arab ethnicity
Ground truth	Consensus of two senior nephrologists, supported by laboratory results (eGFR, serum creatinine, urinalysis) according to KDIGO guidelines
Class distribution (labeled)	Normal 74.4 % (218), CKD 14.0 % (41), AKI 11.6 % (34)
Data splits (patient-level stratified)	Training 198, validation 36, test 59 patients
Input	Dual 224×224×3 images: (1) original ultrasound, (2) segmented kidney on black background (SS-MUNet)
Output	Class probabilities + predicted label (Normal / CKD / AKI)
Performance (test set, n = 59)	Accuracy 82.9 % (95 % CI 81.4–84.3) Macro F1-score 78.4 % (76.9–79.8) AUC (OvR) 0.92 Class-wise F1: Normal 0.89, CKD 0.71, AKI 0.74
Statistical significance	Dual-input versus single-input: McNemar $p < 0.001$, DeLong $p < 0.001$
Inference speed	38.2 ms/image (NVIDIA A100, batch = 1): real-time capable
Model size	48.7 M parameters (segmentation 31.2 M + classification 17.5 M shared)
Limitations	Single-center only (unknown external generalizability), limited diseased cases (41 CKD, 34 AKI), no prospective or multi-center validation, possible reduced performance in severe obesity, poor acoustic windows, or atypical pathology
Ethical / safety notes	Always requires clinician oversight. Risk of over-confidence in hydronephrotic AKI cases. No demographic biases intentionally introduced.
Access	https://github.com/NAlkhaldi/kidney-ultrasound-classification.git

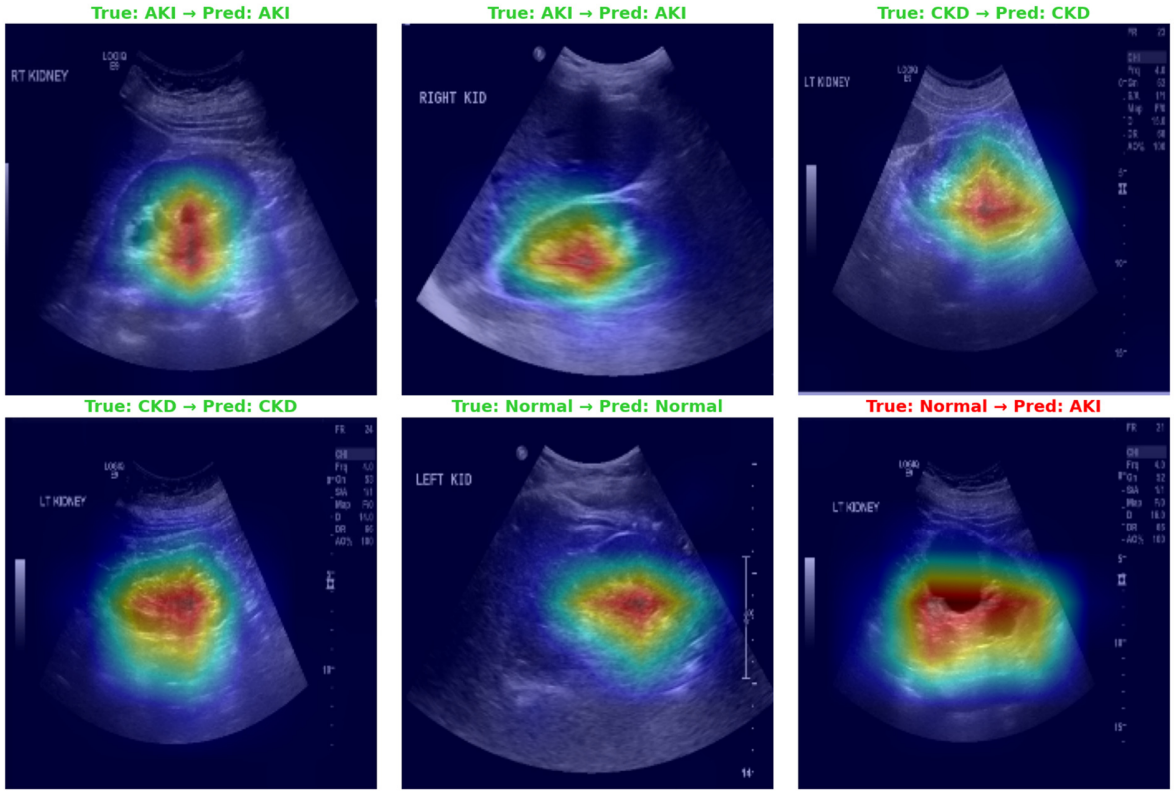


Fig. 20. Grad-CAM visualizations for the dual-input ResNet50 scheme. Top row (left to right): (a) correctly classified AKI: AKI, (b) correctly classified AKI: AKI, (c) correctly classified CKD: CKD.

The green titles in the figure indicate correct predictions, while the red titles signify incorrect predictions. Heatmaps consistently emphasize the renal parenchyma and cortico-medullary junction, thereby affirming strong anatomical localization, even in instances of misclassification. In misclassified cases, attention remains correctly localized to the kidney structure. This suggests that errors arise from subtle textural ambiguities or the class imbalance inherent to the AKI dataset, rather than spatial disorientation. This behavior mirrors established clinical diagnostic challenges and further validates the robustness of the segmentation-guided dual-input design. Future studies should prioritize detailed analysis of these misclassified instances to enhance sensitivity.

VI. CONCLUSION

This study developed a two-stage diagnostic framework: first, an optimized U-Net performs kidney segmentation, followed by a novel dual-input classification stage that comparatively evaluates four deep learning architectures. Using the segmented images resulted from the modified U-net model along with non-segmented images, effectively enhanced diagnostic insight from ultrasound images. The results reveal a discrepancy between optimal segmentation and classification performance. While SS-AUNet yielded the highest Dice coefficient, SS-MUNet provided a superior feature basis for the downstream classification, resulting in the highest accuracy. This suggests that the features generated by SS-MUNet's masks were more discriminative for the classifier, even with a slightly lower pixel-level segmentation score. The results show that architectures like ResNet50 and DenseNet121 were the most effective, yielding the highest test accuracies of 82.88% and 81.36%, respectively. The strong performance of the top models presents a promising proof-of-concept for nephrology decision-support. The accurate and reliable classification can facilitate diagnostics and support early intervention for Acute Kidney Injury (AKI) and Chronic Kidney Disease (CKD).

Despite these promising results, some limitations must be considered for future work. The classification models were evaluated on a small, private dataset of 293 imbalanced samples. Although a public dataset was used to enhance the semi-supervised segmentation training, its binary class structure stone/no-stone made it inappropriate for the final classification task. Therefore, further validation on larger datasets is essential to confirm the model's generalizability. Future work should validate on multi-center datasets to assess model generalizability across populations, protocols, and environments. In particular, it should expand the dataset, validate the models on diverse cohorts, and assess the feasibility of integrating these models into existing clinical workflows. Future studies also should emphasize exploring model interpretability techniques to build clinical trust and examine weighted ensembles of the best performing models, such as DenseNet121 and Swin Transformer, to potentially achieve even greater diagnostic robustness.

ACKNOWLEDGEMENTS

The author would like to thank the clinical research team and medical specialists at NGHA for their guidance, as well as those who assisted with data collection.

FUNDING

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, with grant number [Grant No. KFU260199].

COMPETING INTERESTS

The author declares no competing interests.

ETHICS DECLARATIONS

This research utilized an anonymized kidney ultrasound dataset. The study protocol and use of these data were ethically approved by the Ministry of National Guard Health Affairs (NGHA) and King Faisal University (KFU). All methods were performed in accordance with the relevant guidelines and regulations.

DATA AVAILABILITY

The proprietary ultrasound dataset generated and analyzed during the current study is not publicly available due to strict ethical constraints, patient privacy protections, and institutional policies regarding sensitive medical data. Access to the de-identified dataset may be obtained from the corresponding author upon reasonable request and subject to review and approval by the Ministry of National Guard–Health Affairs (NG-HA). The public dataset utilized for semi-supervised learning is available in the Mendeley Data repository: "Kidney Ultrasound Images 'Stone' and 'No Stone'", V1, https://data.mendeley.com/datasets/h6jc4xm4py/1. The source code developed for this study is publicly available at: https://github.com/NAlkhaldi/kidney-ultrasound-classification.git.

REFERENCES

[1] R. A. Alassaf *et al.*, "Preemptive Diagnosis of Chronic Kidney Disease Using Machine Learning Techniques," in *2018 International Conference on Innovations in Information Technology (IIT)*, Al Ain, Nov. 2018, pp. 99–104, https://doi.org/10.1109/INNOVATIONS.2018.8606040.

[2] C. Ronco, R. Bellomo, and J. A. Kellum, "Acute Kidney Injury," *The Lancet*, vol. 394, no. 10212, pp. 1949–1964, Nov. 2019, https://doi.org/10.1016/S0140-6736(19)32563-2.

[3] K. Hansen, M. Nielsen, and C. Ewertsen, "Ultrasonography of the Kidney: A Pictorial Review," *Diagnostics*, vol. 6, no. 1, Dec. 2015, Art. no. 2, https://doi.org/10.3390/diagnostics6010002.

[4] "Kidney Diseases," *Saudi Arabia Ministry of Health*, 2025. https://www.moh.gov.sa/en/HealthAwareness/EducationalContent/Diseases/Urology/Pages/002.aspx.

[5] J.-W. Hsieh, C.-H. Lee, Y.-C. Chen, W.-S. Lee, and H.-F. Chiang, "Stage Classification in Chronic Kidney Disease by Ultrasound Image," in *Proceedings of the 29th International Conference on Image and Vision Computing New Zealand*, Hamilton, New Zealand, Nov. 2014, pp. 271–276, https://doi.org/10.1145/2683405.2683457.

[6] A. Anggrawan, K. Hidjah, and Q. S. Jihadil, "Kidney Failure Diagnosis based on Case-Based Reasoning (CBR) Method and Statistical Analysis," in *2016 International Conference on Informatics and Computing (ICIC)*, Mataram, Indonesia, 2016, pp. 298–303, https://doi.org/10.1109/IAC.2016.7905733.

[7] S. Gopika and M. Vanitha, "Survey on Prediction of Kidney Disease by using Data Mining Techniques," *IJARCCE*, pp. 198–201, Mar. 2017, https://doi.org/10.17148/IJARCCE.2017.6138.

[8] P. Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron," in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, Turin, July 2017, pp. 193–198, https://doi.org/10.1109/COMPSAC.2017.84.

[9] M. Udhayarasu, K. Ramakrishnan, and S. Periasamy, "Assessment of Chronic Kidney Disease using Skin Texture as a Key Parameter: For South Indian Population," *Healthcare Technology Letters*, vol. 4, no. 6, pp. 223–227, Dec. 2017, https://doi.org/10.1049/htl.2016.0098.

[10] Y. He, X. Yu, C. Liu, J. Zhang, K. Hu, and H. C. Zhu, "A 3D Dual Path U-Net of Cancer Segmentation Based on MRI," in *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, Chongqing, June 2018, pp. 268–272, https://doi.org/10.1109/ICIVC.2018.8492781.

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

[12] C. Sabanayagam *et al.*, "A Deep Learning Algorithm to Detect Chronic Kidney Disease from Retinal Photographs in Community-Based Populations," *The Lancet Digital Health*, vol. 2, no. 6, pp. e295–e302, June 2020, https://doi.org/10.1016/S2589-7500(20)30063-7.

[13] K. Chaudhary *et al.*, "Utilization of Deep Learning for Subphenotype Identification in Sepsis-Associated Acute Kidney Injury," *Clinical Journal of the American Society of Nephrology*, vol. 15, no. 11, pp. 1557–1565, Nov. 2020, https://doi.org/10.2215/CJN.09330819.

[14] N. Rank *et al.*, "Deep-Learning-Based Real-Time Prediction of Acute Kidney Injury Outperforms Human Predictive Performance," *npj Digital Medicine*, vol. 3, no. 1, Oct. 2020, Art. no. 139, https://doi.org/10.1038/s41746-020-00346-8.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[17] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas." arXiv, 2018, https://doi.org/10.48550/ARXIV.1804.03999.

[18] V. Alevizos and M. Hon, "Comparison of Kidney Segmentation Under Attention U-Net Architectures." Nov. 01, 2021, https://doi.org/10.36227/techrxiv.16546281.v2.

[19] L. Li, M. Verma, Y. Nakashima, H. Nagahara, and R. Kawasaki, "IterNet: Retinal Image Segmentation Utilizing Structural Redundancy in Vessel Networks," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, USA, Mar. 2020, pp. 3645–3654, https://doi.org/10.1109/WACV45572.2020.9093621.

[20] Y. Luo, J. Liu, and M. Ding, "CC-Net: Consistency Learning Combined with Contrastive Learning for Kidney Ultrasound Segmentation," in *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Kuching, Malaysia, Oct. 2024, pp. 821–826, https://doi.org/10.1109/SMC54092.2024.10830962.

[21] R. Khan *et al.*, "MLAU-Net: Deep Supervised Attention and Hybrid Loss Strategies for Enhanced Segmentation of Low-Resolution Kidney Ultrasound," *Digital Health*, vol. 10, Jan. 2024, Art. no. 20552076241291306, https://doi.org/10.1177/20552076241291306.

[22] B. Lei *et al.*, "Skin Lesion Segmentation Via Generative Adversarial Networks with Dual Discriminators," *Medical Image Analysis*, vol. 64, Aug. 2020, Art. no. 101716, https://doi.org/10.1016/j.media.2020.101716.

[23] Y.-C. Chang, C.-M. Lo, Y.-K. Chen, P.-H. Wu, and H. Luh, "W-Net: Two-Stage Segmentation for Multi-Center Kidney Ultrasound," in *2024*

IEEE Conference on Artificial Intelligence (CAI), Singapore, Singapore, June 2024, pp. 1522–1523, https://doi.org/10.1109/CAI59869.2024.00274.

[24] G.-P. Chen *et al.*, "Asymmetric U-Shaped Network with Hybrid Attention Mechanism for Kidney Ultrasound Images Segmentation," *Expert Systems with Applications*, vol. 212, Feb. 2023, Art. no. 118847, https://doi.org/10.1016/j.eswa.2022.118847.

[25] S. L. P. Sitanaboina, S. R. Beeram, H. Jonnadula, and L. Paleti, "Attention 3D-CU-Net: Enhancing Kidney Tumor Segmentation Accuracy Through Selective Feature Emphasis," *IEEE Access*, vol. 11, pp. 139798–139810, 2023, https://doi.org/10.1109/ACCESS.2023.3340912.

[26] M. G. Oghli *et al.*, "Fully automated kidney image biomarker prediction in ultrasound scans using Fast-Unet++," *Scientific Reports*, vol. 14, no. 1, Feb. 2024, Art. no. 4782, https://doi.org/10.1038/s41598-024-55106-5.

[27] M. Krithika Alias AnbuDevi and K. Suganthi, "Review of Semantic Segmentation of Medical Images Using Modified Architectures of UNET," *Diagnostics*, vol. 12, no. 12, Dec. 2022, Art. no. 3064, https://doi.org/10.3390/diagnostics12123064.

[28] C.-C. Kuo *et al.*, "Automation of the Kidney Function Prediction and Classification Through Ultrasound-Based Kidney Imaging using Deep Learning," *npj Digital Medicine*, vol. 2, no. 1, Apr. 2019, Art. no. 29, https://doi.org/10.1038/s41746-019-0104-2.

[29] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and Diagnosis of Chronic Kidney Disease Using Deep Learning-Based Heterogeneous Modified Artificial Neural Network," *Future Generation Computer Systems*, vol. 111, pp. 17–26, Oct. 2020, https://doi.org/10.1016/j.future.2020.04.036.

[30] Q. Zheng, S. L. Furth, G. E. Tasian, and Y. Fan, "Computer-Aided Diagnosis of Congenital Abnormalities of the Kidney and Urinary Tract in Children Based on Ultrasound Imaging Data by Integrating Texture Image Features and Deep Transfer Learning Image Features," *Journal of Pediatric Urology*, vol. 15, no. 1, Feb. 2019, Art. no. 75.e1-75.e7, https://doi.org/10.1016/j.jpurol.2018.10.020.

[31] K. Venkatrao and S. Kareemulla, "HDLNET: A Hybrid Deep Learning Network Model with Intelligent IOT for Detection and Classification of Chronic Kidney Disease," *IEEE Access*, vol. 11, pp. 99638–99652, 2023, https://doi.org/10.1109/ACCESS.2023.3312183.

[32] Q. Zhang and Q. Wang, "Comparison of Semantic Segmentation Methods on Renal Ultrasound Images," in *2022 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Ottawa, ON, Canada, May 2022, pp. 1–5, https://doi.org/10.1109/I2MTC48687.2022.9806525.

[33] S. H. Song *et al.*, "Deep-Learning Segmentation of Ultrasound Images for Automated Calculation of the Hydronephrosis Area to Renal Parenchyma Ratio," *Investigative and Clinical Urology*, vol. 63, no. 4, 2022, Art. no. 455, https://doi.org/10.4111/icu.20220085.

[34] S. Valente *et al.*, "A Deep Learning Method for Kidney Segmentation in 2D Ultrasound Images," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Glasgow, Scotland, United Kingdom, July 2022, pp. 3911–3914, https://doi.org/10.1109/EMBC48229.2022.9871748.

[35] D. Rodríguez-Salas, M. Öttl, M. Seuret, K. Packhäuser, and A. Maier, "Using Forestnets for Partial Fine-Tuning Prior to Breast Cancer Detection in Ultrasounds," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, Cartagena, Colombia, Apr. 2023, pp. 1–5, https://doi.org/10.1109/ISBI53787.2023.10230424.

[36] S. Sudharson and P. Kokil, "An Ensemble of Deep Neural Networks for Kidney Ultrasound Image Classification," *Computer Methods and Programs in Biomedicine*, vol. 197, Dec. 2020, Art. no. 105709, https://doi.org/10.1016/j.cmpb.2020.105709.

[37] P. Hao *et al.*, "Texture Branch Network for Chronic Kidney Disease Screening based on Ultrasound Images," *Frontiers of Information Technology & Electronic Engineering*, vol. 21, no. 8, pp. 1161–1170, Aug. 2020, https://doi.org/10.1631/FITEE.1900210.

[38] Shixuemei and Mideth Abisado, "Diagnostic Segmentation Based on Kidney Medical Image," *Journal of Artificial Intelligence and Technology*, June 2023, https://doi.org/10.37965/jait.2023.0214.

[39] J. Liu *et al.*, "Asym-UNet: An Asymmetric U-Shape Network for Breast Lesions Ultrasound Images Segmentation," *Biomedical Signal Processing and Control*, vol. 99, Jan. 2025, Art. no. 106822, https://doi.org/10.1016/j.bspc.2024.106822.

[40] C. Zhang, L. Wang, G. Wei, Z. Kong, and M. Qiu, "A Dual-Branch and Dual Attention Transformer and CNN Hybrid Network for Ultrasound Image Segmentation," *Frontiers in Physiology*, vol. 15, Sept. 2024, Art. no. 1432987, https://doi.org/10.3389/fphys.2024.1432987.

[41] Y. Xie, B. Yang, Q. Guan, J. Zhang, Q. Wu, and Y. Xia, "Attention Mechanisms in Medical Image Segmentation: A Survey." arXiv, 2023, https://doi.org/10.48550/ARXIV.2305.17937.

[42] C.-P. Fu, M.-J. Yu, Y.-S. Huang, C.-S. Fuh, and R.-F. Chang, "Stratifying High-Risk Thyroid Nodules Using a Novel Deep Learning System," *Experimental and Clinical Endocrinology & Diabetes*, vol. 131, no. 10, pp. 508–514, Oct. 2023, https://doi.org/10.1055/a-2122-5585.

[43] S. M. Kabir and M. I. H. Bhuiyan, "RiIG-WCP Image-based Breast Tumors Classification Using Swin Transformer," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, Dec. 2023, pp. 1–5, https://doi.org/10.1109/ICCIT60459.2023.10440979.

[44] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002, https://doi.org/10.1109/ICCV48922.2021.00986.

[45] Y. Fu *et al.*, "Semantic-Attention Enhanced DSC-Transformer for Lymph Node Ultrasound Classification and Remote Diagnostics," *Bioengineering*, vol. 12, no. 2, Feb. 2025, Art. no. 190, https://doi.org/10.3390/bioengineering12020190.

[46] T. Yu and P. Li, "Degenerate Swin to Win: Plain Window-based Transformer without Sophisticated Operations." arXiv, 2022, https://doi.org/10.48550/ARXIV.2211.14255.

[47] J. Koo, J. Yang, L. An, G. C. Sergio, and S. I. Park, "Swin-Free: Achieving Better Cross-Window Attention and Efficiency with Size-varying Window." arXiv, 2023, https://doi.org/10.48550/ARXIV.2306.13776.

[48] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, and B. Fu, "Shuffle Transformer: Rethinking Spatial Shuffle for Vision Transformer." arXiv, 2021, https://doi.org/10.48550/ARXIV.2106.03650.

[49] H. R. Torres, S. Queirós, P. Morais, B. Oliveira, J. C. Fonseca, and J. L. Vilaça, "Kidney Segmentation in Ultrasound, Magnetic Resonance and Computed Tomography Images: A Systematic Review," *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 49–67, Apr. 2018, https://doi.org/10.1016/j.cmpb.2018.01.014.

[50] X. Liang, M. Du, and Z. Chen, "Artificial Intelligence-Aided Ultrasound in Renal Diseases: A Systematic Review," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 6, pp. 3988–4001, June 2023, https://doi.org/10.21037/qims-22-1428.

[51] S. Valente *et al.*, "A Comparative Study of Deep Learning Methods for Multi-Class Semantic Segmentation of 2D Kidney Ultrasound Images," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Sydney, Australia, July 2023, pp. 1–4, https://doi.org/10.1109/EMBC40787.2023.10341170.

[52] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021, https://doi.org/10.1109/ACCESS.2021.3086020.

[53] M. A. Isaza-Ruget *et al.*, "Predicting Chronic Kidney Disease Progression with Artificial Intelligence," *BMC Nephrology*, vol. 25, no. 1, Apr. 2024, Art. no. 148, https://doi.org/10.1186/s12882-024-03545-7.

[54] M. Tounsi, D. Y. Abdulhussain, A. T. Azar, A. Al-Khayyat, and I. K. Ibraheem, "Deep Learning Model-based Decision Support System for Kidney Cancer on Renal Images," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17177–17187, Oct. 2024, https://doi.org/10.48084/etasr.8335.

[55] Y. H. Bhosale, K. S. Patnaik, S. R. Zanwar, S. Kr. Singh, V. Singh, and U. B. Shinde, "Thoracic-Net: Explainable Artificial Intelligence (XAI) Based Few Shots Learning Feature Fusion Technique for Multi-

Classifying Thoracic Diseases Using Medical Imaging," *Multimedia Tools and Applications*, vol. 84, no. 9, pp. 5397–5433, Oct. 2024, https://doi.org/10.1007/s11042-024-20327-3.

[56] Y. H. Bhosale, P. Singh, and K. S. Patnaik, "COVID-19 and Associated Lung Disease Classification Using Deep Learning," in *International Conference on Innovative Computing and Communications*, vol. 492, D. Gupta, A. Khanna, A. E. Hassanien, S. Anand, and A. Jaiswal, Eds. Singapore: Springer Nature Singapore, 2023, pp. 283–295.

[57] "Internal Ultrasound Dataset for Acute and Chronic Kidney Disease Detection Based on Ultrasound Images." Ministry of National Guard - Health Affairs (MNG-HA) & King Abdullah International Medical Research Center (KAIMRC), King Abdulaziz Hospital, Al Ahsa, Saudi Arabia, https://www.moh.gov.sa/en/Pages/default.aspx, 2021.

[58] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, Feb. 2020, Art. no. 125, https://doi.org/10.3390/info11020125.

[59] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified Focal Loss: Generalising Dice and Cross Entropy-Based Losses to Handle Class Imbalanced Medical Image Segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, Jan. 2022, Art. no. 102026, https://doi.org/10.1016/j.compmedimag.2021.102026.

[60] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019, https://doi.org/10.48550/ARXIV.1905.11946.

[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 2017, pp. 2261–2269, https://doi.org/10.1109/CVPR.2017.243.

[62] G. Kaur and S. Singh, "Kidney Ultrasound Images 'Stone' and 'No Stone.'" Mendeley Data, Aug. 27, 2024, https://doi.org/10.17632/H6JC4XM4PY.1.

[63] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-Supervised Medical Image Segmentation via Cross Teaching between CNN and Transformer." arXiv, 2021, https://doi.org/10.48550/ARXIV.2112.04894.

[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, https://doi.org/10.1007/s11263-019-01228-7.

[65] M. Mitchell *et al.*, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, Jan. 2019, pp. 220–229, https://doi.org/10.1145/3287560.3287596.

[66] A. S. Tejani *et al.*, "Updating the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) for Reporting AI Research," *Nature Machine Intelligence*, vol. 5, no. 9, pp. 950–951, Sept. 2023, https://doi.org/10.1038/s42256-023-00717-2.