

Hybrid U-Net Architectures with ResNet50 and VGG19 for Accurate CT-Based Kidney Disease and Stone Segmentation



Dishad Abdalrahman Mahmood, Muhammad Amin Muhammadali

Department of Computer Science, Faculty of Science, Soran University, Kurdistan Region, Iraq

ABSTRACT

Kidney illness is a major worldwide health issue requiring prompt and precise diagnosis for optimal management. This paper presents a comprehensive evaluation of hybrid deep learning (DL) architectures that integrate U-Net with ResNet50 and VGG19 for the automatic segmentation of kidney stones and renal disorders from computed tomography (CT) images. We assembled a dataset of 118 individuals from a private hospital, comprising 13,035 kidney-specific CT scans, while also using the publicly accessible Kaggle Kidney Stone Segmentation Dataset. Three experimental situations were established: (1) Concurrent segmentation of kidney disease and stones, (2) segmentation of kidney stones alone, and (3) segmentation of kidney disease exclusively. The hybrid U-Net + ResNet50 model attained superior performance in stone-only segmentation, with an F1-score of 0.8653, an IoU of 0.7626, and an accuracy of 0.9998 at a resolution of 256×256 . The U-Net + VGG19 model exhibited strong performance in all situations, attaining an F1-score and DC of 0.8663 for stone segmentation. Both models demonstrated exceptional generalization ability when evaluated on external datasets. The findings indicate that hybrid architectures markedly improve segmentation accuracy compared to conventional methods, providing dependable automated tools for clinical kidney pathology evaluation while ensuring computational efficiency with average processing durations below 0.05 s per scan.

Index Terms: Kidney stone, Renal disorders, DL, U-Net, VGG19, ResNet50.

1. INTRODUCTION

Kidney disease is a significant public health challenge to society today. Early diagnosis and awareness significantly reduce mortality. Disorders that interfere with the kidneys' typical function are referred to as kidney diseases [1]. Kidney disease classification requires dividing kidney images or patient data into multiple illness sorts, including chronic kidney disease (CKD) stages, polycystic kidney disease,

nephrolithiasis, and renal tumors. A kidney stone is a solid formation that can result in damage to the kidneys, intense pain, and reduced quality of life due to urinary system blockages [2].

Numerous imaging techniques are employed in medical diagnosis, including sonography, computed tomography (CT), magnetic resonance imaging (MRI), and X-rays; however, these methods are not without issues, as they can be time-consuming, subjective, and susceptible to human error. The increase in incidence, coupled with technological advancements, imposes a substantial financial strain on healthcare facilities for managing kidney stone disease (KSD), with an estimated global expenditure of USD 5.3 billion in 2014, rendering it the second most expensive condition [3].

Access this article online

DOI:10.21928/uhdjst.v9n2y2025.pp231-250

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Copyright © 2025 Mahmood and Muhammadali. This is an open access article distributed under the Creative Commons Attribution Non-Commercial No Derivatives License 4.0 (CC BY-NC-ND 4.0)

Corresponding author's e-mail: Dr. Muhammad Amin Muhammadali, Computer Science Department, Faculty of Science, Soran University, Kurdistan Region, Iraq. E-mail: muhammad.amin@soran.edu.iq

Received: 13-06-2025

Accepted: 21-09-2025

Published: 13-10-2025

The literature on modern-day machine learning techniques in healthcare progressively focuses on their use in renal disease, particularly in the study of renal pathologies, which are common and affect a significant percentage of the population, leading to various complications, such as death in some instances [4]. Deep learning (DL)-based image segmentation is now firmly recognized as an effective method in the field of image segmentation. It has been widely used to delineate homologous areas as primary and fundamental elements of the diagnostic and therapeutic process [5].

The automated detection of renal abnormalities represents a critical objective in clinical practice, with medical imaging modalities including ultrasound (US), MRI, and CT serving as primary diagnostic tools [6]. The integration of whole slide images from histological samples in digital pathology algorithms for computer-aided assessment has gained significant momentum in recent years [7]. Beyond detection, CT imaging enables precise determination of stone dimensions and anatomical positioning, thereby facilitating comprehensive risk assessment for spontaneous stone passage and informing decisions regarding surgical intervention [8]. Images CT consistently provide the most accurate diagnosis. Conventionally, US has lower sensitivity and specificity than CT [9]. The difficulties of interpreting the challenges of complex image data in medical images can be addressed by applying ML and DL concepts [10]. Clinicians can therefore use DL approaches to automatically diagnose renal disorders. However, improving performance in the identification and challenge of renal disease remains difficult [11], [12].

An important element of artificial intelligence (AI) techniques is the training of appropriate images available to the public. Researchers can now use various datasets, including CT, US, and MRI. However, each dataset differs from the others in terms of quantity, illumination variation, dimensions, and image quality, which may require fine-tuning. That is why authors Kaur and Singh believe that image fusion plays a significant role in different computer vision applications. However, designing an efficient image fusion technique is still a challenging task [13]. The objective of the research is to augment diagnostic precision, alleviate the strain for radiologists, and boost the early identification of renal disorders using automated image segmentation. The following are major contributions mentioned:

1. Novel Hybrid Architecture Development U-Net with ResNet50 and U-Net with VGG19
2. Comprehensive Multi-Scenario Evaluation Framework
3. Clinical Dataset Creation and Expert Annotation

4. Outstanding Performance Achievement
5. Clinical Translation and Practical Implementation.

Although developments, including Transformer-based designs like Swin-Transformers, have produced cutting-edge outcomes in medical picture segmentation, their clinical utility is limited by substantial computing demands and the necessity for exceptionally large annotated datasets. These restrictions make it hard to use them in everyday clinical practice, especially in hospital settings where resources are limited. Hybrid U-Net models, on the other hand, are a better choice because they keep the original U-Net decoder's speed and reliability while adding powerful encoders such as ResNet50 and VGG19 to make feature extraction better. This hybrid design immediately solves the gap between accuracy and efficiency, giving models that are not only competitive with more sophisticated approaches but also practicable for real-world clinical integration.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive review of related work in kidney pathology segmentation and DL applications in medical imaging. Section 3 details the methodology, including research design, data collection, preprocessing procedures, proposed hybrid U-Net architectures, model training protocols, and evaluation metrics. Section 4 presents the experimental results and provides an in-depth discussion of the performance evaluation across three distinct scenarios, comparing U-Net+ResNet50 and U-Net+VGG19 models under various configurations. Finally, Section 5 concludes the study with key findings, contributions, and future research directions for advancing automated kidney disease diagnosis through DL techniques.

2. RELATED WORK

The authors, Huang *et al.*, created a computer-aided diagnostic system for Kidney Ureter Bladder (KUB) imaging to help physicians correctly diagnose urinary tract stones [14]. Whereas Yildirim *et al.*, proposed an automatic detection system for kidney stones (stone or no stone) that uses coronary imaging CT, using DL techniques [15]. Moreover, Fitri *et al.*, using a convolutional neural network (CNN), have created an autonomous method for classifying urinary stones into the three categories based on micro-CT images [16]. Furthermore, Zhao *et al.*, introduced a multi-scale supervised 3D U-Net (MSS U-Net) for the segmentation of kidneys and renal tumors from CT scans [17].

However, researchers Alqahtani *et al.*, say sigmoid functions enhance prediction accuracy for binary outcomes. Finally, they apply classification using the proposed modified Extreme Gradient Boosting (XGBoost) for kidney stone prediction. The loss functions are modified to enhance the model's learning effectiveness and classification accuracy. Evaluate the proposed approach through internal comparison with the decision tree (DT) and Naive Bayes (NB) [18].

and Yang *et al.*, performed a retrospective analysis of the medical records of 358 patients who received shock wave lithotripsy for urinary stones (kidney and upper urinary tract stones), which includes patient demographic characteristics and urinary stone characteristics as depicted by non-contrast CT images. They used an 80% training set and a 20% test set to predict success, primarily using decision tree-based ML methods, including Random Forest (RF), XGBoost, and Light Gradient Boosting Method (LightGBM) [19].

However, Daniel *et al.* used computer learning, specifically 2D CNN, to accurately separate left and right kidneys from T2-weighted (MRI) data. The data set consisted of 30 HC volunteers and 30 CKD patients. The model was trained on 50 manually outlined HC and CKD kidney sections. The model was further evaluated using 50 test data sets, consisting of data from 5 healthy controls and 5 patients with CKD, each scanned 5 times in a session to facilitate comparison between the microscopic CNN and manual segmentation of the kidney [20].

Ma *et al.* enrolled 468 patients with kidney, bladder, and urinary stones at multiple sites at Peking Union Medical College Hospital. Urine metabolite profiling was used to discover markers for KSD using ML techniques. The total number of patients with renal stones was 148 (34.02%), bladder stones 34 (7.82%), and multisite stones 163 (34.83%). According to their analysis, the RF algorithm had the best prediction accuracy, with area under the curve (AUC) values for kidney stones of 0.809, urethral stones of 0.99, and multisite stones of 0.775 [21]. Aksakalli *et al.*, assessed multiple ML techniques, including DT, RF, support vector classifier, multilayer perceptron, K-nearest neighbors, NB (Bernoulli NB), and deep neural network utilizing CNN. The collection comprises 221 kidney X-ray images acquired from the Department of Urology at Atatürk University. The trials indicate that the DT yields the most favorable classification results. This method achieves the greatest F1 score, with a success rate of 85.3%, utilizing the S+U sampling technique [22].

Researchers, Fitri *et al.*, have devised an automated technique to categorize urinary stones into three categories utilizing micro-CT images through CNN. A total of 2,430 images were obtained from *in vitro* micro-CT scans of urinary stones in various patients. The validation accuracy of the devised method utilizing a CNN with optimized hyperparameters was 0.9852. The trained CNN algorithm attained a test accuracy of 0.9959 [16]. Furthermore, Parakh *et al.* used unenhanced CT scans of the abdomen and pelvis in 535 adults with suspected KSD. CNN's cascading model has a high accuracy AUC of 0.954 in detecting urinary tract stones on unenhanced CT scans [12].

Nonetheless, Elton *et al.* utilized a dataset of 91 CT colonography (CTC) images, including manually annotated kidney stones, alongside 89 CTC scans devoid of kidney stones. 50% of the data were allocated for training, while the remaining 50% was designated for testing. As an external validation set, 6185 patients' CTC scans from a separate institution were employed. A three-dimensional U-Net model was employed for kidney segmentation. A 13-layer CNN classifier was used to differentiate kidney stones from false-positive regions. The system attained an area under the receiver operating characteristic curve of 0.95 on an external validation set, with an AUC of 0.95, sensitivity of 0.88, and specificity of 0.91 at the Youden index [23].

Furthermore, Blau *et al.*, present a fully automated approach for renal cyst diagnosis, underpinned by a strong segmentation of the kidneys executed by a fully CNN. The evaluation of performance was conducted on 52 randomly selected abdominal CT scans from a genuine radiological process, which included more than 70 cysts annotated by a proficient radiologist. The program identified 59 out of 70 cysts (true-positive rate = 84.3%) while generating an average of 1.6 false positives per case [24].

Nonetheless, the authors Xiong *et al.* posited that ultrasonography is extensively utilized in the diagnosis of kidney tumors due to its widespread acceptance, affordability, and absence of radiation exposure. Consequently, they introduced a novel technique for segmenting renal tumors in US images utilizing the adaptive subregional diffusion level set model (ASSLSM). In comparison to conventional US segmentation techniques, ASSLSM demonstrates superior accuracy in renal tumor segmentation. The test yielded a Hausdorff distance (HD) of (8.75 ± 4.21) mm, a mean absolute distance of (3.26 ± 1.69) mm, and a dice index (Dice) of 0.93 ± 0.03 [25].

To create their study sample, Gaikar *et al.*, examined the Pathology and Image Archiving and Communication System (PACS) database. Produced a set of MP-MRI scan data for individuals with kidney masses. T1W-NG image 3D volumes for 108 patients and T2W, T1W-IP, T1W-OP, T1W-PRE, and T1W-CM image 3D volumes for 50 patients made up the dataset. They also created a method based on TL to enhance kidney segmentation on the dataset. Two stages were taken to apply the created kidney segmentation algorithm to various mp-MRI data. Using a DL-based attention U-Net model, the kidney segmentation was first identified on T1W-NG images. In the subsequent phase, the pretrained T1W-NG kidney segmentation model was fine-tuned to distinguish kidneys in T2W, T1W-IP, T1W-OP, T1W-PRE, and T1W-CM MRI sequences. Increased average DSC T1W-IP from 83.64% to 85.42%, T1W-OP from 79.35% to 83.66%, T1W-PRE from 82.05% to 85.94%, T1W-CM from 85.65% to 87.64%, and T2W climbed from 87.19% to 89.90% as a result of the TL technique [26]. Table 1 summarizes the findings of published investigations. CT-based models, such as RDA-UNET, exhibit enhanced performance, whereas DeepLab, ResNet50, and Swin Transformer thrive across many modalities and tasks.

3. METHODOLOGY

The standard imaging modalities for assessing nephrolithiasis and renal pathology include US [32], CT [33], MRI [34], and KUB X-ray imaging [35]. This study used medical CT scan data to segment kidney stones and kidney disease after sequencing and image preparation procedures. Furthermore, the DL models used for training and testing are discussed.

3.1. Research Design

In the first step, a dataset of the kidney 3D images is collected. After collecting the data, the DL models are used to detect the kidney images. The results of the segmentation of kidney stones and kidney disease of the models are compared with several other methods based on the standard performance criteria. Thereafter, the optimal model is trained and evaluated using a publicly accessible dataset from Kaggle, which is explained diagrammatically in Fig. 1.

3.2. Data Collection and Preprocessing

The first objective was to obtain realistic data on kidney disease. It was achieved to collect CT scan data. CT scans were obtained from 118 patients at a private hospital in Ranya, consisting of subjects: 34 normal healthy, 29 with kidney stones only, 23 with kidney disease, and 32 with

kidney stones and kidney disease, as shown in Table 2. Also, utilizing the publicly available Kidney Stone Segmentation Dataset from Kaggle.

The DICOMDIR of the 118 patients was then converted into images, using software (diVision Lite) to Joint Photographic Group (JPG), which contained 49,463 images. We isolated only 13,035 images dedicated to kidneys under the supervision of a nephrologist and urologist. Of these, there are 1131 CT scans of kidney stones and 1584 CT scans of kidney disease. Its data sets are shown in Table 3.

A preprocessing strategy was devised to prepare the dataset for efficient model training, encompassing numerous essential phases. Initially, all images were scaled to a standardized resolution of 512×512 pixels to maintain uniformity throughout the collection. All images were subsequently converted from 8-bit to 24-bit format to preserve greater color and detail information. Noise was mitigated by Gaussian filtering, which smooths images by averaging pixel values within the near vicinity. Furthermore, brightness was augmented to improve the visibility of essential features with linear brightness adjustment. The processed images were sorted into a separate folder for improved accessibility and efficient use in later modeling steps.

A specialized open-source Python application was utilized for annotating medical CT scans. The labeling process was executed through `labelImg.py`, a commonly employed annotation tool incorporated within the Anaconda environment. CT scans illustrating nephropathy (Fig. 2) and kidney stones (Fig. 3) were meticulously examined and documented under the direct supervision of a qualified nephrologist to guarantee clinical precision. Every image was carefully annotated, and the labels were stored in three distinct formats: JavaScript Object Notation (JSON), TXT, and Extensible Markup Language (XML). For the purpose of creating precise segmentation masks unique to each annotated image, the chosen formats were meant to improve compatibility with later processes. In medical image analysis, the manual way ensured better annotations, which are necessary for supervised learning models to work.

The JSON file serves as the primary reference for producing the appropriate segmentation mask. Automation of this procedure was achieved with Python scripts, which guarantee uniform and accurate extraction of mask regions. Fig. 4 shows an example of this procedure.

TABLE 1: Summary of published investigation findings comparing model performance across different approaches

Study	Method	Objective	challenge limitations	Result
Fu <i>et al.</i> [27]	RDA-UNET model	Automatic segmentation of renal cysts	Selecting the best model for renal cyst segmentation in medical image CT	Left kidney DSC=96.25 Precision=96.34 Recall=96.88 Right kidney DSC=94.22 Precision=95.34 Recall=94.61
Goyal <i>et al.</i> [28]	R-CNN	Automatic segmentation of the kidneys	The automatic segmentation of the kidneys in MRI	Dice=0.890 IoU=0.816
Sharma <i>et al.</i> [29]	InceptionV3	To analyze images of renal pathologies, including stones, tumors, and cysts.	For optimal performance in diagnosing kidney disorders, including cysts, tumors, and stones, by CT imaging.	Accuracy=97.76 Recall of=98.28 F1 score=0.95
Göker [30]	ResNet101	Classifying kidney stone illnesses	Compare transfer learning models from CT images	Overall accuracy=0.981 Recall=0.991 Specificity=0.974 Precision=0.966 F1-measure=0.978 Kappa statistic=0.962
Islam <i>et al.</i> [31]	Swin VGG16 ResNet50 Inceptionv3	Renal cyst, tumor, and stone detection.	To achieve the best model in the diagnosis of renal diseases, such as cysts and tumors, using CT imaging.	Swin Accuracy=99.30% VGG16 Accuracy=98.20% ResNet50 Accuracy=61.60% Inceptionv3 Accuracy=73.80%

MRI: Magnetic resonance imaging, CT: Computed tomography

TABLE 2: Comprehensive distribution of patients by renal health status

Patients total	Healthy	Only has kidney stones	Only has kidney disease	Kidney stones and kidney disease
118	34	29	23	32

From the collected dataset, three distinct datasets were ultimately constructed, each consisting of CT scans and their corresponding segmentation masks. The first dataset included images and masks for both kidney disease and kidney stones. The second dataset focused exclusively on kidney stone cases, while the third contained only instances related to renal illness. All datasets were randomly split into three subsets: 70% for training, 20% for validation, and 10% for testing. The entire process of dataset construction and splitting was automated using Python to ensure consistency and efficiency. The datasets are available at this link in Kaggle: <https://kaggle.com/datasets/b65b0abc924a6aa91f894ac91839c6cb52894b4092767770c5dd34e7eb09d210>

Furthermore, the study utilized the publicly accessible Kidney Stone Segmentation Dataset from Kaggle (<https://www.kaggle.com/datasets/bemorekkg/kidney-stone-segmentation-dataset>) to train and assess the efficacy of the

TABLE 3: The distribution of CT scans by kidney condition

Total CT scans	CT scans of the kidneys	CT scans of kidney stones	CT scans of kidney disease	CT scans contain kidney stones and kidney disease
49463	13035	1131	1584	308

CT: Computed tomography

suggested segmentation algorithm. This dataset includes CT scans of kidney stones together with related segmentation masks developed by the Segment Anything Model (SAM) utilizing YOLO bounding box annotations. The CT images are supplied in Joint Photographic Experts Group (JPEG) format, while the segmentation masks are presented as binary Portable Network Graphics (PNG) images. The dataset consists of 923 image-mask pairings, randomly assigned to training (70%, 646 images + 218 rotated images), validation (20%, 184 images), and testing (10%, 93 images). Using the same model and evaluation subsets, Fig. 5 shows examples of Cagle images and masks.

3.3. Proposed Methods

DL is achieving success and garnering interest across various disciplines, including computer vision, speech recognition,

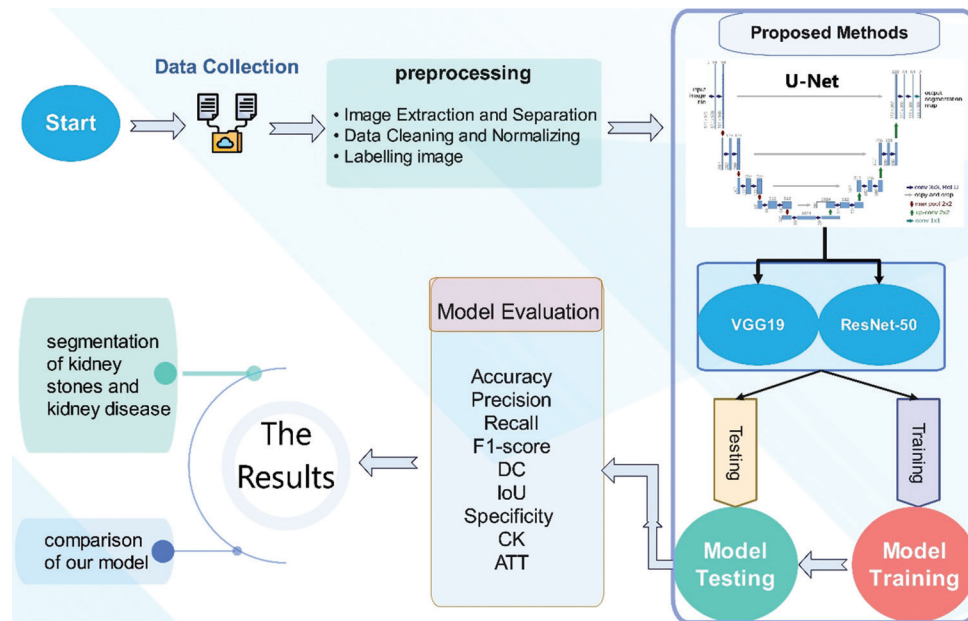


Fig. 1. The proposed structure.

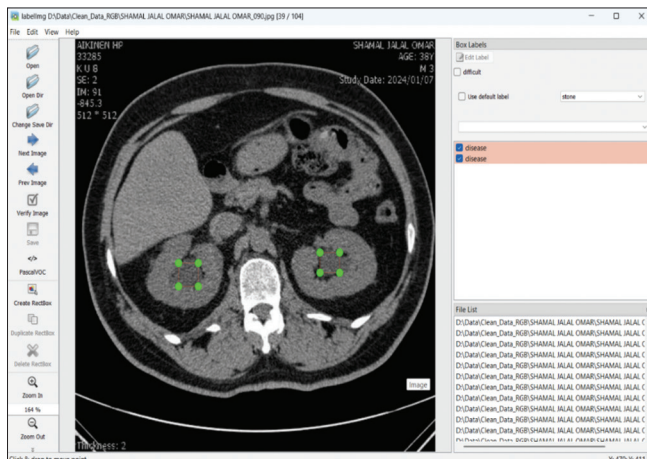


Fig. 2. Labelling kidney disease.

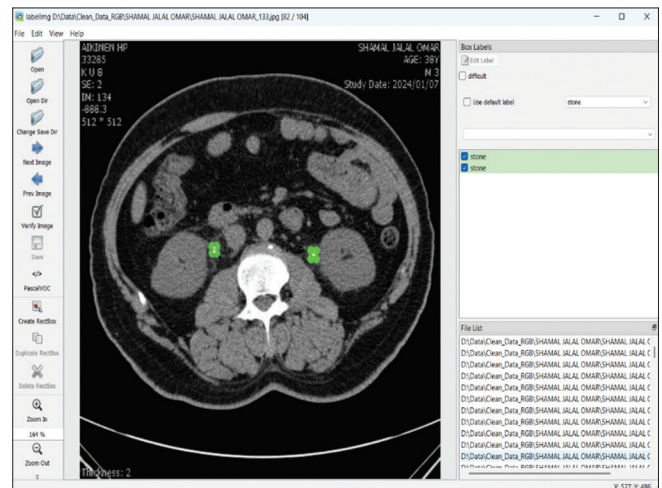


Fig. 3. Labelling a kidney stone.

natural language processing, and gaming. DL techniques generate a correspondence between raw inputs and target outputs (e.g., image classifications) [36]. The use of DL-based medical image analysis in computer-aided detection (CAD) offers decision support to clinicians, enhancing the accuracy and efficiency of diagnostic and treatment processes while stimulating new research and development initiatives in CAD [37].

3.3.1. U-Net model

U-Net, as shown in Fig. 6, is a CNN architecture designed for image segmentation tasks, introduced by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in 2015. The network

derives its name from its U-shaped architecture, comprising a contracting path (encoder) and an expanded path (decoder) [38].

The U-Net is a CNN designed for biological image segmentation. The architecture, featuring an encoder-decoder framework, exhibits remarkable stability and can achieve accurate segmentation with a reduced number of training images. The network consists of three 3×3 convolutional layers, with a maximum of two 2×2 layers following each pooling layer, utilizing the ReLU activation function. A 1×1 convolutional layer is appended at

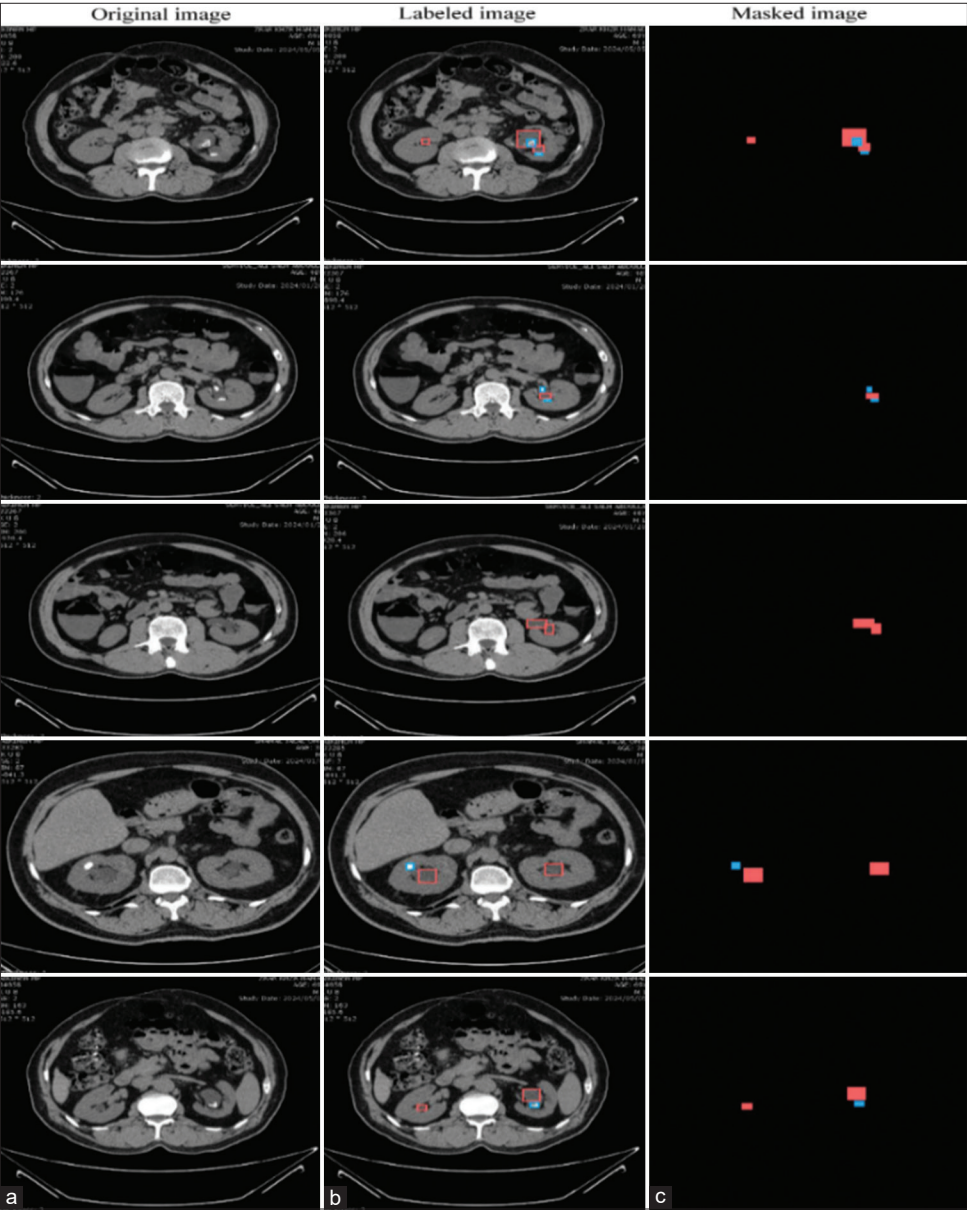


Fig. 4. (a) Some original images; (b) labels the images; (c) creates masks from the JSON file; the red color represents kidney disease, and the blue color represents kidney stones.

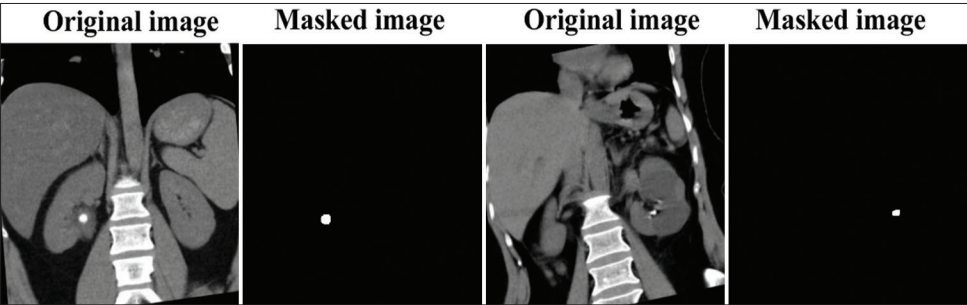


Fig. 5. A sample dataset on kidney stones and masks is available on Kaggle.

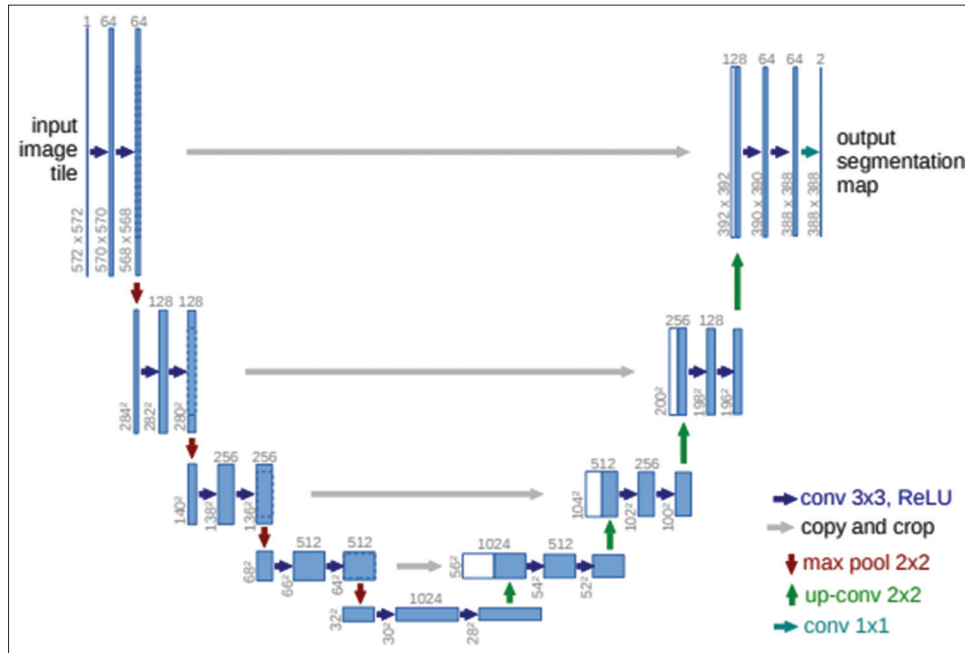


Fig. 6. U-Net design (example for 32×32 pixels at the lowest resolution) [39].

the conclusion. The U-Net comprises a contracting path and a symmetric expanding path, which are utilized for accurate localization and feature extraction. The U-Net design relies significantly on data augmentation methods. The network is capable of segmenting a 512×512 image on a contemporary GPU in under one second. The U-Net has been effectively utilized for medical image segmentation; nevertheless, 3D convolution is recommended to fully leverage the spatial information in 3D images, including CT and MRI scans [39].

In the domain of binary classification, the binary cross-entropy loss function is one of the most widely used metrics to evaluate the performance of a model. It measures the dissimilarity between the predicted probabilities and the actual binary labels (0 or 1), penalizing incorrect predictions more heavily as they diverge from the ground truth. Mathematically, the binary cross-entropy loss is defined as:

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (1)$$

where y_i represents the actual label and $\hat{y}_i$ signifies the expected probability for the i_{th} sample. This function guarantees that the loss converges to zero with accurate forecasts and escalates markedly with inaccuracies. Moreover, it is essential in gradient-based optimization by offering explicit and comprehensible feedback throughout model training.

3.3.2. Comparison of hybrid U-Net methods

Hybrid U-Net methodologies augment the conventional U-Net design by substituting the encoder with superior DL models, such as ResNet50 and VGG19. These adjustments aim to enhance feature extraction and increase the model's accuracy, especially in intricate tasks such as medical image segmentation. By retaining the original U-Net decoder, these hybrid models sustain spatial detail reconstruction while leveraging enhanced feature representations. Comparative assessments generally quantify performance through metrics such as segmentation pixel-wise accuracy, precision, recall, Dice coefficient (DC), and Intersection over Union (IoU), with findings frequently indicating that hybrid models surpass the original U-Net in both precision and generalization. In this study, advanced segmentation frameworks were used by combining VGG19 and ResNet50 with U-Net to accurately segment kidney stones and related kidney diseases from medical images. This approach enhances the automated analysis of kidney conditions, supporting more effective diagnosis and treatment planning.

3.4. Model Training and Evaluation

To thoroughly assess the efficacy of the hybrid models utilized on the three specified scenarios, an extensive array of performance metrics was applied, including accuracy, precision, recall, F1-score, DC, IoU, specificity, Cohen's kappa (CK), and average test time (ATT). The selection of these measures was guided by their ability to evaluate several dimensions of model

performance, particularly in medical image segmentation. Accuracy provides a broad assessment of overall correctness by juxtaposing the total number of accurate predictions against all predictions produced; precision denotes the ratio of accurately predicted positives to the total predicted positives, hence regulating the false-positive rate, whereas recall measures the ratio of correctly detected actual positives, emphasizing the model's sensitivity. Their harmonic mean, the F1-score, offers a balanced metric, especially advantageous in cases of uneven class distribution. The DC and IoU function as metrics for spatial overlap between predicted and ground truth segmentation masks, essential for assessing model performance at the pixel level. Specificity enhances recall by evaluating the model's capacity to accurately detect negative instances, hence minimizing false positives. CK provides a chance-corrected metric for assessing agreement between forecasts and actual outcomes, particularly significant in multi-class tasks. The ATIT serves as a pragmatic metric for the model's computational efficiency, underscoring its viability for real-world applications where processing speed is critical. These measures collectively establish a comprehensive framework for evaluating the segmentation accuracy, reliability, and efficiency of hybrid models in complicated, clinically relevant datasets.

3.5. Ethical Considerations

This research followed ethical guidelines to make sure that data and technology were used in a responsible way. To protect patient privacy, all CT scan data were made anonymous. To avoid bias and overfitting, a diverse dataset was employed. The AI models were created just for research and are not meant to be used in clinical settings without more testing. The whole process was guided by ethical principles to make sure that it was fair, safe, and respectful of human dignity.

4. RESULTS AND DISCUSSION

This section delineates the experimental setup, preprocessing section, training procedure, and assessment of the suggested DL models. The emphasis is on evaluating several hybrid U-Net topologies to determine the most efficient model for the particular segmentation problem. Various models, U-Net with ResNet50 and VGG19, were trained and evaluated utilizing the prepared datasets. According to the evaluation standards described in Section 3.4, the models were assessed using various performance metrics.

4.1. Experimental Setup Recap

This study evaluated the performance of DL models for kidney CT scan segmentation using U-Net + ResNet50 and

U-Net + VGG19 architectures. Experiments were conducted in a consistent environment with an NVIDIA GTX 1080 Ti GPU, an Intel Core i7 CPU, and 16 GB RAM, using Python 3, TensorFlow 2.x, Keras, and OpenCV on Windows 10 through Anaconda and Jupyter Notebook.

4.2. Model Performance Evaluation

The models were assessed utilizing various performance indicators. All models were trained and validated under uniform experimental settings to guarantee an equitable comparison. The experiments were conducted using different image resolutions (128×128 and 256×256), batch sizes (8 and 24), and epoch settings (100, 150, and 200) with a learning rate of 0.0001 and the Adam optimizer, and a 5-fold cross-validation (CV) strategy.

This section presents a quantitative assessment of each DL model employed for segmenting CT scans associated with kidney stones and renal illnesses. The models were assessed and trained over three distinct scenarios within the dataset to gauge their segmentation efficacy. The scenarios were as follows: In the first scenario, the data set consisted of CT scans with their respective segment masks for kidney disease with kidney stones. The dataset was split into 70% for training (1675 scans), 20% for validation (479 scans), and 10% for testing (240 scans). In the second scenario, the dataset consisted exclusively of renal stone-related CT scans and masks. We used data augmentation to enhance the training set. The training set consisted of 1582 CT scans (70%), the validation set included 226 scans (20%), and the test set included 114 scans (10%). The third scenario focused the dataset exclusively on CT scans and masks associated with renal disease. The data set consisted of 1108 scans from the training set (70%), 317 from the validation set (20%), and 159 from the test set (10%). Table 4 provides a summary of all three scenarios along with their corresponding distributions. Samples were randomly assigned to each category to guarantee unbiased assessment.

The study thoroughly assessed two hybrid models, U-Net + ResNet50 and U-Net + VGG19, for the multiclass segmentation of renal disease and kidney stones, utilizing CT scans. To ensure equitable comparisons and minimize residual learning effects, each model was trained from inception using varying image resolutions (256×256 and 128×128), batch sizes (8 and 24), and training epochs (100, 150, and 200), as shown in Table 5. The U-Net + ResNet50 used deep feature extraction to make boundaries more accurate, whereas the U-Net + VGG19 used its deep hierarchical structure to keep spatial features. There were three tests for each model:

TABLE 4: Splitting of the training, validation, and testing in three diagnostic scenarios

n	Data sets	Train (70%)	Validation (20%)	Test (10%)	Total of images
1	CT scans and masks related to kidney disease and kidney stones.	1675	479	240	2394
2	Only CT scans and kidney stone-related masks.	791×2 rotation	226	114	1922
3	Only CT scans and masks for kidney disease.	1108	317	159	1584

CT: Computed tomography

TABLE 5: Experimental configuration parameters for U-Net-based model architectures

Models	Image size	Batch size	Epochs
U-Net+ResNet50	• 256×256	• 8	• 100
	• 128×128	• 24	• 150
			• 200
U-Net+VGG19	• 256×256	• 8	• 100
	• 128×128	• 24	• 150
			• 200

the first scenario was segmentation of kidney disease and kidney stones, the second scenario was only kidney stone segmentation, and the third scenario was only kidney disease segmentation. Figs. 7 and 8 show the input images, ground truth masks, and anticipated outputs for all configurations. This methodical approach yielded significant insights into the advantages and drawbacks of each hybrid architecture across various training scenarios.

Figs. 7 and 8 display segmentation outcomes produced by two models: U-Net combined with ResNet50 and U-Net combined with VGG19. Each image comprises the actual CT scans, ground truth masks, and anticipated segmentation masks. Subfigures (a-c) present findings for a picture dimension of 256×256 with a batch size of 8 at 100, 150, and 200 epochs, respectively. Subfigures (d-f) present findings for an image dimension of 128×128 with a batch size of 8, whereas subfigures (g-i) depict outcomes for the identical image dimension with a batch size of 24. These visualizations illustrate the models' efficacy across various training configurations and segmentation contexts.

4.3. Discussion

The evaluation criteria were used across two different U-Net-based models, hybrid models comprising U-Net with each model, ResNet50 and VGG19, for three separate scenarios. This multi-scenario methodology provides a detailed examination of how model performance fluctuates based on task complexity and data characteristics.

4.3.1. Hybrid U-Net + ResNet50

The U-Net+ResNet50 model was evaluated across multiple configurations of batch size, image dimensions, and training

epochs. Initial testing with 256×256 images and batch size 8 showed optimal performance at 150 epochs (accuracy: 0.9992, precision: 0.8514, recall: 0.786, F1-score: 0.8174), with performance declining at 200 epochs due to overfitting. Reducing image size to 128×128 maintained robust segmentation performance while significantly decreasing inference time from 0.0275s to 0.0149s. The best overall results were achieved using 128×128 images, a batch size of 24, and 200 epochs, yielding an F1-score of 0.8238 and an IoU of 0.7000 with an inference time of 0.0157s. This setup achieved an optimal balance between segmentation accuracy and computational efficiency, demonstrating that the model works well with less computing power by optimizing parameters in Scenario 1.

In the second scenario, the hybrid U-Net + ResNet50 model was evaluated exclusively on CT kidney stone images with consistent accuracy (0.9998) and specificity (0.9999) across all configurations. The model was tested across three parameter combinations: 256×256 image resolution with a batch size of 8, 128×128 resolution with a batch size of 8, and 128×128 resolution with a batch size of 24, each trained for 100, 150, and 200 epochs. Optimal performance was achieved using 256×256 images, a batch size of 8, and 200 epochs, yielding a precision of 0.879, a recall of 0.8521, an F1-score/DC of 0.8653, an IoU of 0.7626, CK of 0.8652, and an average training time of 0.0259 s. Reducing image dimensions to 128×128 generally decreased performance metrics, while increasing batch size from 8 to 24 showed mixed results with slightly improved precision but reduced recall.

For the third scenario, focusing on renal disease segmentation using the hybrid U-Net + ResNet50 model, experiments were conducted with varying training epochs (100, 150, 200) and image resolutions (256×256 and 128×128) at different batch sizes. At 256×256 resolution with a batch size of 8, optimal performance was achieved after 100 epochs, yielding an accuracy of 0.9987, a precision of 0.8142, a recall of 0.7951, an F1-score of 0.8046, and an IoU of 0.673. Extended training to 150 and 200 epochs showed gradual performance degradation across most metrics. When image resolution was reduced to 128×128 , similar patterns emerged

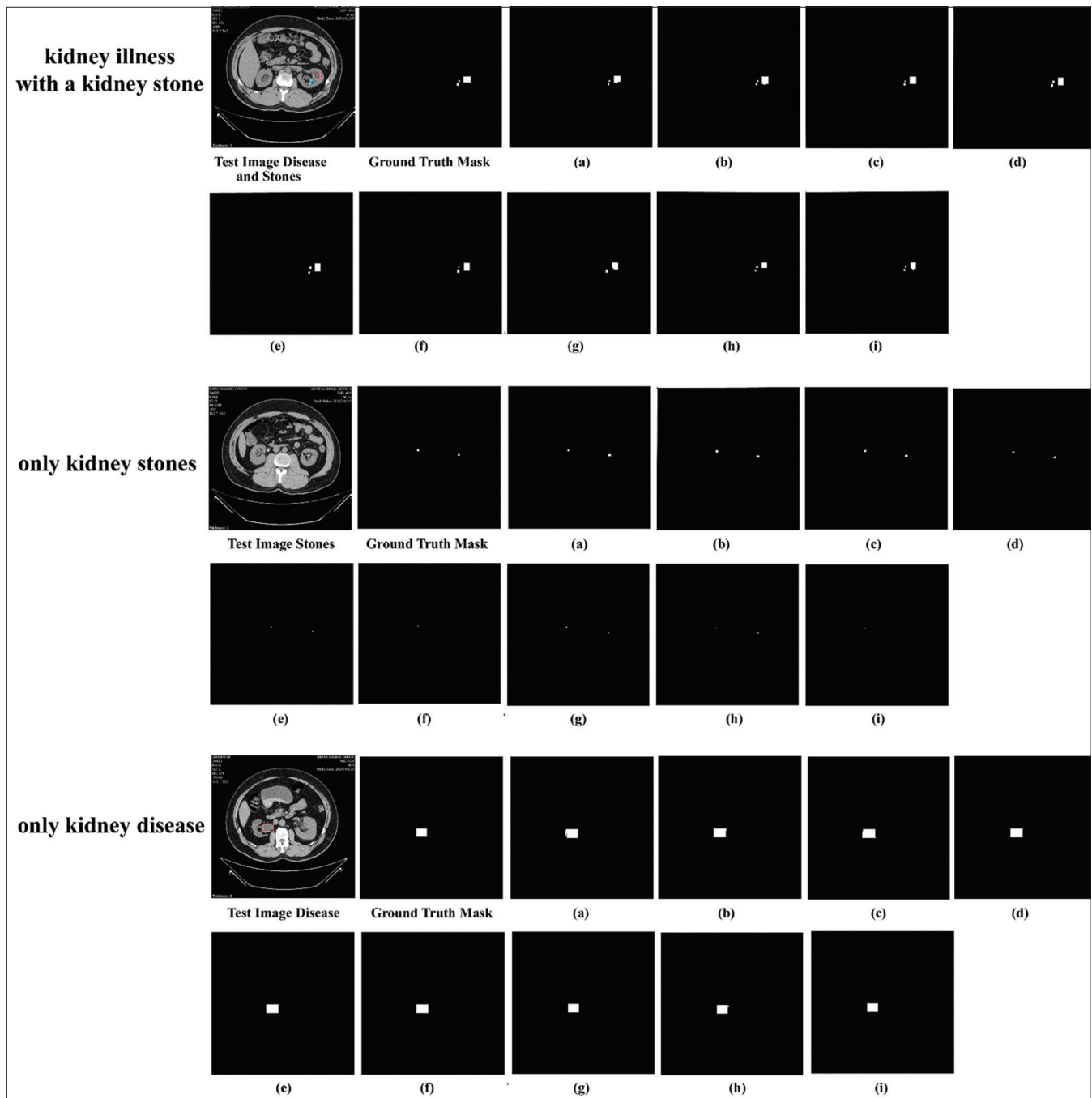


Fig. 7. Illustrates segmentation results for three scenarios: (top) renal disease with a kidney stone, (middle) a kidney stone only, and (bottom) renal disease only. Each set displays the CT test image, the ground truth mask, and the predicted masks ((a)–(i)) generated by the proposed U-Net + ResNet50 model.

with peak performance at 100 epochs, albeit overall metrics were slightly lower. Increasing the batch size from 8 to 24 while maintaining a 128×128 resolution produced marginal improvements but failed to surpass the original configuration. The analysis demonstrated that optimal segmentation quality

was achieved using the hybrid model trained at 256×256 resolution for 100 epochs with a batch size of 8.

Fig. 9 presents the optimal performance of assessment metrics attained by the hybrid U-Net+ResNet50 model for

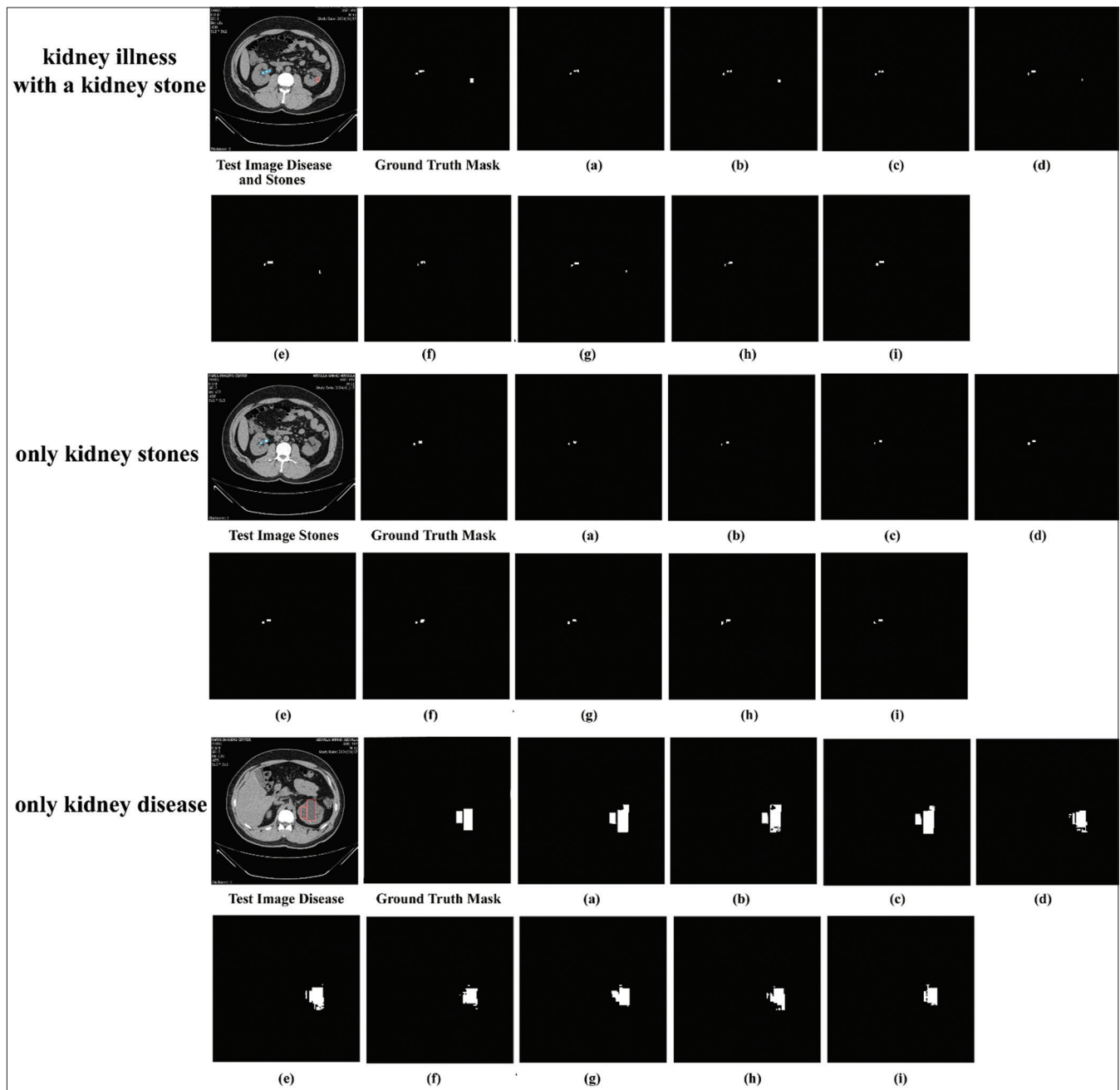


Fig. 8. Illustrates segmentation results for three scenarios: (Top) renal disease with a kidney stone, (middle) a kidney stone only, and (bottom) renal disease only. Each set displays the CT test image, the ground truth mask, and the predicted masks (a-i) generated by the proposed U-Net + VGG19 model.

segmenting renal disease and kidney stones across all three experimental scenarios, depicted in chart form. This graphic depiction highlights the model's efficacy across several parameters, including batch size, image size, and training epochs.

Fig. 10 illustrates the CM for the three U-Net + ResNet50 combinations across various circumstances. Scenario 2 (utilizing a 256×256 input size) exhibits the most equitable performance, recording the fewest false positives (655) and false negatives (826), underscoring its proficient

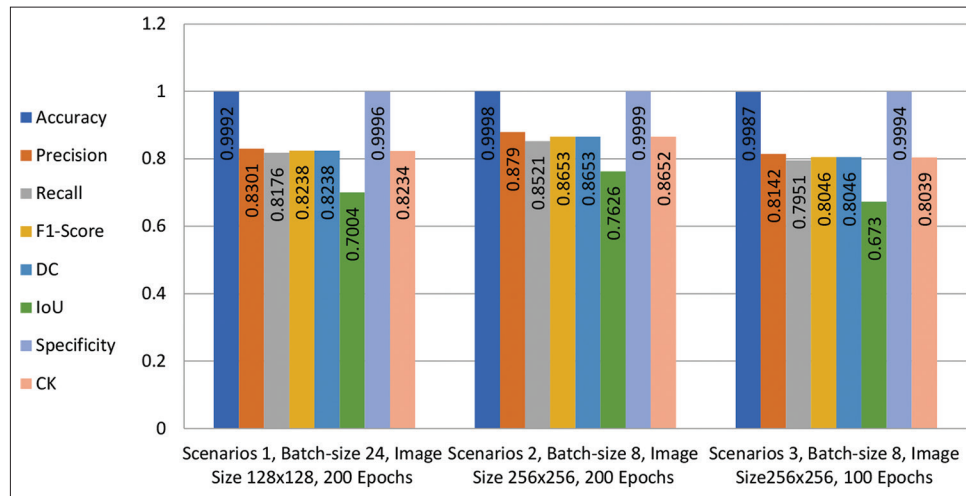


Fig. 9. Presents the highest evaluation criteria for the hybrid U-Net+ResNet50 model in all three scenarios.

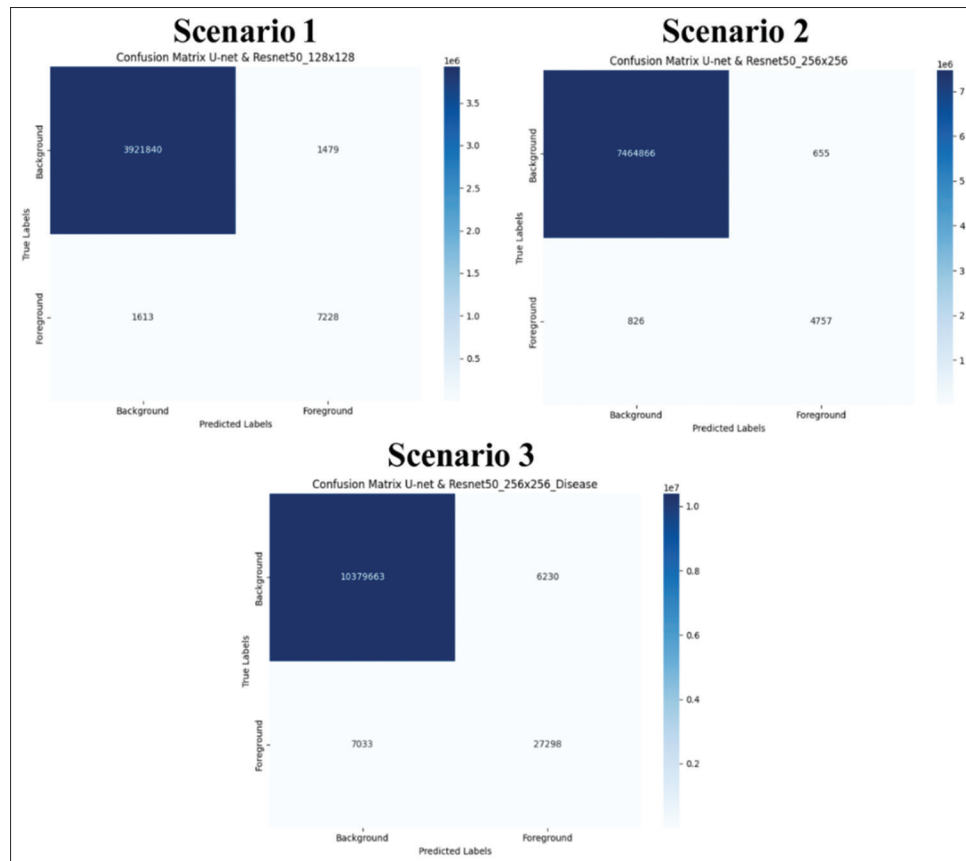


Fig. 10. Depicts the confusion matrices for the three distinct U-Net+ResNet50 model configurations employed in each scenario.

differentiation between background and foreground classes. Scenario 1 demonstrates robust segmentation performance but exhibits a marginally elevated incidence of misclassifications, suggesting potential enhancements in

precision. Conversely, Scenario 3 shows the highest incidence of false positives (6,230), suggesting a propensity for over-segmentation, although it yields a substantial number of real positives. These findings underscore the enhanced

reliability of Scenario 2 in producing accurate and consistent segmentation results.

4.3.2. Hybrid U-Net + VGG19

In Scenario 1, the hybrid U-Net+VGG19 architecture was systematically evaluated by varying batch size, image resolution, and training epochs to assess their impact on kidney disease and kidney stone segmentation. The model demonstrated robust performance across all configurations, with optimal results achieved using 256×256 image resolution, a batch size of 8, and 200 training epochs. Under these conditions, the model achieved an accuracy of 0.9992, an F1-score of 0.8193, an IoU of 0.6939, and a CK of 0.8189, while maintaining computational efficiency with an average training time of 0.0310 seconds. Alternative configurations with reduced image resolution (128×128) and increased batch size (24) demonstrated comparable performance, with F1-scores ranging from 0.8066 to 0.8189, indicating the model's stability across different parameter combinations. These findings confirm the robustness and effectiveness of the U-Net+VGG19 hybrid model for kidney pathology segmentation.

In Scenario 2, the U-Net+VGG19 hybrid model was trained from scratch on kidney stone CT scans using various configurations of image dimensions, batch sizes, and training epochs to systematically evaluate the impact of each parameter on model performance. The optimal configuration, achieved with a 256×256 image size, batch size of 8, and 150 epochs, delivered the best balance of segmentation accuracy (F1-score = 0.8663), robustness (CK = 0.8662), and spatial overlap (IoU = 0.7641). While reducing image size to 128×128 improved processing time (~ 0.0123 s), it compromised segmentation quality with lower F1-scores peaking at 0.8439. Increasing the batch size to 24 showed minimal performance variations, and extending training beyond 150 epochs led to overfitting, with decreased recall and F1-scores, indicating that the 150-epoch threshold represents the optimal training duration for this model configuration.

In Scenario 3, the hybrid U-Net + VGG19 model was trained exclusively on CT scans of diseased kidneys to evaluate segmentation performance with reduced class diversity. The model demonstrated sensitivity to hyperparameters, achieving optimal results with a batch size of 8, image dimensions of 256×256 , and 150 training epochs, yielding a consistent accuracy of 0.9987, an F1-score of 0.8032, and an IoU of 0.6712. Reducing the image size to 128×128 improved computational efficiency (ATT ~ 0.0120 seconds), but slightly compromised segmentation metrics. In contrast, increasing

the batch size to 24 maintained accuracy but reduced the F1-score. Training beyond 150 epochs showed diminishing returns with potential overfitting, indicating that the optimal configuration balances spatial accuracy, overlap consistency, and computational cost. The results confirm the model's robust generalization capabilities in simplified classification contexts.

Fig. 11 illustrates a comparative chart that emphasizes the superior performance of the hybrid U-Net + VGG19 model in segmenting kidney stones and renal illnesses across the three experimental scenarios. This visual summary illustrates how critical parameters, such as batch size, image resolution, and training epochs, affect the model's efficacy, providing a concise understanding of its segmentation abilities across various configurations.

Fig. 12 shows the confusion matrices for the U-Net+VGG19 model across the three scenarios, providing an extensive overview of its segmentation efficacy in various medical imaging contexts. In Scenario 2 (stone alone), the model has good precision and recall, evidenced by the low counts of false positives (691) and false negatives (789), indicating a robust capacity to identify different stone features. Scenario 1 (stone + disease) and Scenario 3 (disease only) have consistently elevated accurate negative rates, indicating dependable background classification. Both situations indicate an increase in false negatives (6,674 and 6,643, respectively), suggesting a minor reduction in sensitivity when illness characteristics are evident, particularly in more intricate or nuanced pathological areas. The model demonstrates strong performance, highlighting the versatility and reliability of the U-Net + VGG19 architecture in addressing various image segmentation difficulties in the biomedical field.

Table 6 reveals distinct performance patterns across the hybrid architectures and segmentation scenarios. Both models achieve their highest F1-scores (0.8653 and 0.8663) in stone-only segmentation, indicating that isolated kidney stone identification represents the least complex task among the three scenarios. The U-Net+VGG19 model demonstrates superior consistency, maintaining F1-scores above 0.80 across all scenarios, while U-Net+ResNet50 shows more variable performance with a notable drop to 0.8046 in disease-only segmentation. Configuration analysis reveals that optimal performance typically requires higher resolution (256×256) images, contradicting the common assumption that computational efficiency necessarily improves outcomes. The processing time differences are substantial, with U-Net+ResNet50 achieving nearly twice the speed in the

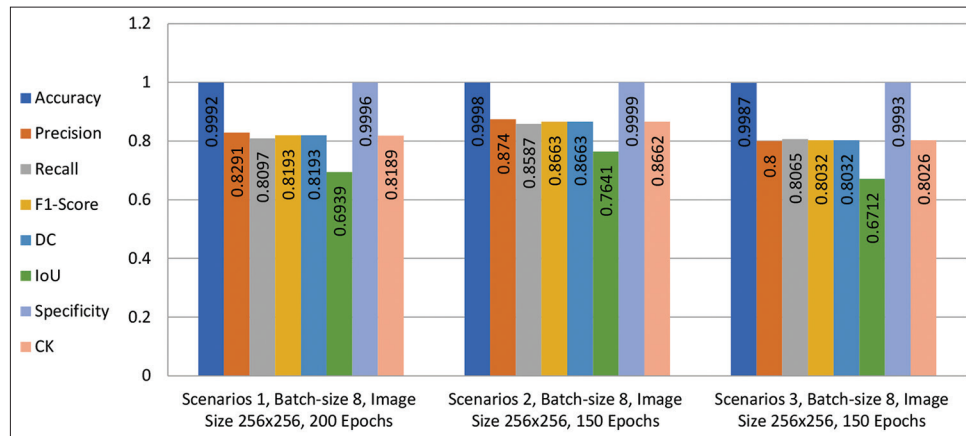


Fig. 11. Presents the highest evaluation criteria for the hybrid U-Net + VGG19 model in all three scenarios.

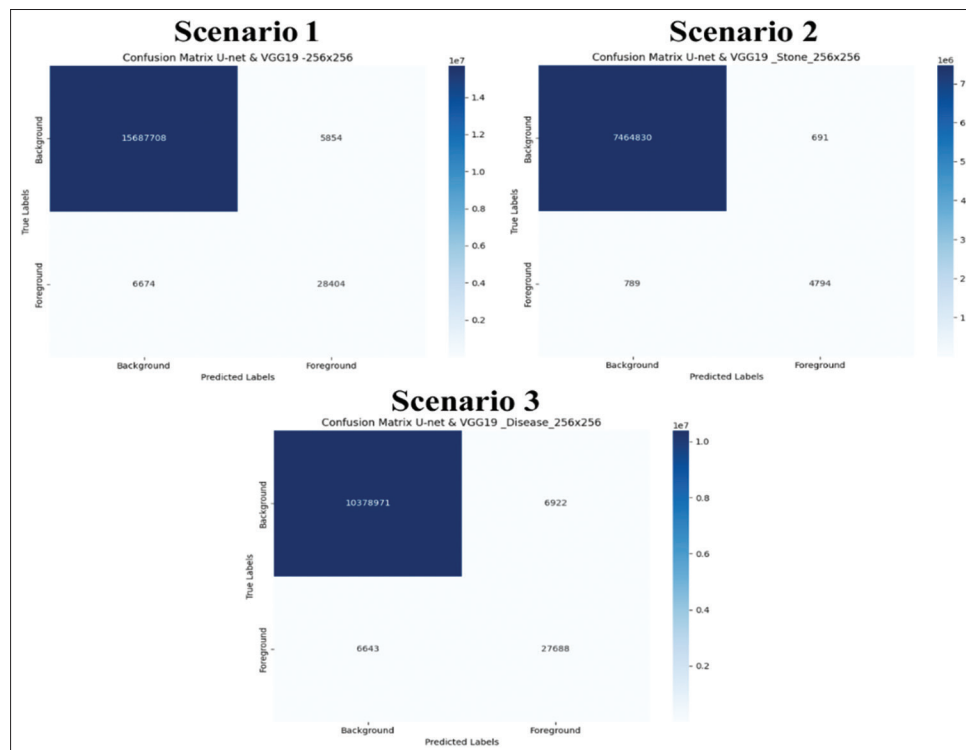


Fig. 12. Depicts the confusion matrices for the three distinct U-Net + VGG19 model configurations employed in each scenario.

TABLE 6: Optimal performance summary by model and scenario

Model	Scenario	Batch size	Image size	Epochs	Accuracy	Precision	Recall	F1-score	IoU	ATT (s)
U-Net+ResNet50	Disease+Stone	24	128×128	200	0.9992	0.8301	0.8176	0.8238	0.7004	0.0157
U-Net+ResNet50	Stone Only	8	256×256	200	0.9998	0.879	0.8521	0.8653	0.7626	0.0259
U-Net+ResNet50	Disease Only	8	256×256	100	0.9987	0.8142	0.7951	0.8046	0.673	0.0302
U-Net+VGG19	Disease+Stone	8	256×256	200	0.9992	0.8291	0.8097	0.8193	0.6939	0.031
U-Net+VGG19	Stone Only	8	256×256	150	0.9998	0.874	0.8587	0.8663	0.7641	0.0343
U-Net+VGG19	Disease Only	8	256×256	150	0.9987	0.8	0.8065	0.8032	0.6712	0.0296

disease+stone scenario (0.0157s vs. 0.0310s), suggesting that ResNet50's residual connections provide computational advantages beyond accuracy improvements. The IoU values consistently lag behind F1-scores by approximately 0.10-0.15 points, indicating that while both models achieve good pixel-wise classification, precise boundary delineation remains challenging in kidney pathology segmentation.

Table 7 exposes the architectural trade-offs between the two hybrid models and their clinical implications. U-Net+VGG19 emerges as the superior choice for stone segmentation, achieving the highest F1-score and IoU values, which translates to more accurate stone boundary detection crucial for treatment planning and surgical guidance. However, U-Net+ResNet50 demonstrates faster processing across most scenarios, with its computational efficiency being particularly pronounced in complex multi-class segmentation tasks. The consistent performance metric reveals U-Net+VGG19's architectural stability, maintaining more uniform results across diverse pathological conditions, while U-Net+ResNet50 shows task-specific optimization that favors certain scenarios over others. From a clinical deployment perspective, the processing time differences, though measured in milliseconds, could compound significantly in high-throughput environments where thousands of scans require analysis daily. The performance gap between stone and disease segmentation across both models (F1-score difference of ~0.06) suggests that kidney stone identification benefits from clearer radiological contrast compared to the more subtle tissue changes associated with kidney disease, highlighting the inherent complexity differences in pathological feature recognition.

Fig. 13 presents a comprehensive comparison of F1-score versus inference time, clearly demonstrating the performance-efficiency trade-offs between models. The chart reveals that U-Net+ResNet50 (blue) consistently achieves faster inference times across all scenarios, while U-Net+VGG19 (red) provides marginally higher F1-scores

in certain configurations. Both models peak in Scenario 2 (kidney stones only), with minimal performance differences but significant efficiency variations, enabling clinicians to select the optimal model based on specific operational requirements.

The performance disparities between U-Net+ResNet50 and U-Net+VGG19 stem from their architectural attributes. ResNet50's residual connections enhance feature extraction for high-contrast structures like kidney stones, whereas VGG19's hierarchical depth more effectively maintains fine spatial details, resulting in marginally superior outcomes in complex pathological segmentation. In practical applications, U-Net+ResNet50 is advantageous for rapid and reliable stone identification, while U-Net+VGG19 delivers more uniform performance across various renal states, offering doctors adaptability in model selection based on the diagnostic context.

Table 8 provides a comparative examination of two hybrid models, U-Net integrated with ResNet50 and U-Net integrated with VGG19, across several scenarios. Scenario 2 of both models, utilizing a batch size of 8, an image dimension of 256×256 , and a substantial epoch count, attained the highest overall performance. Significantly, the U-Net + VGG19 model in Scenario 2 marginally surpassed its equivalent, attaining an F1-score of 0.8663 and an IoU of 0.7641. The results demonstrate that the U-Net combined with VGG19 is a competitive alternative to ResNet50 regarding segmentation accuracy and efficiency.

Furthermore, the comparative assessment of the proposed hybrid designs reveals significant differences that elucidate their relative performance in the experimental circumstances. The U-Net+ResNet50 model attained optimal performance in stone-only segmentation, attributable to the residual connections that improve gradient propagation and facilitate more efficient extraction of distinguishing characteristics from high-contrast areas. Conversely, the U-Net+VGG19

TABLE 7: Comparative performance analysis					
Performance aspect	Best Model	F1-Score	IoU	Processing Time (s)	Interpretation
Overall Performance	U-Net+VGG19 (Stone)	0.8663	0.7641	0.0343	Highest segmentation accuracy
Stone Segmentation	U-Net+VGG19	0.8663	0.7641	0.0343	Superior stone boundary detection
Disease Segmentation	U-Net+ResNet50	0.8046	0.673	0.0302	Better disease feature extraction
Computational Efficiency	U-Net+ResNet50 (Disease+Stone)	0.8238	0.7004	0.0157	Fastest processing time
Consistent Performance	U-Net+VGG19	0.8296*	0.7097*	0.0316*	Most stable across scenarios

*Average across all three scenarios

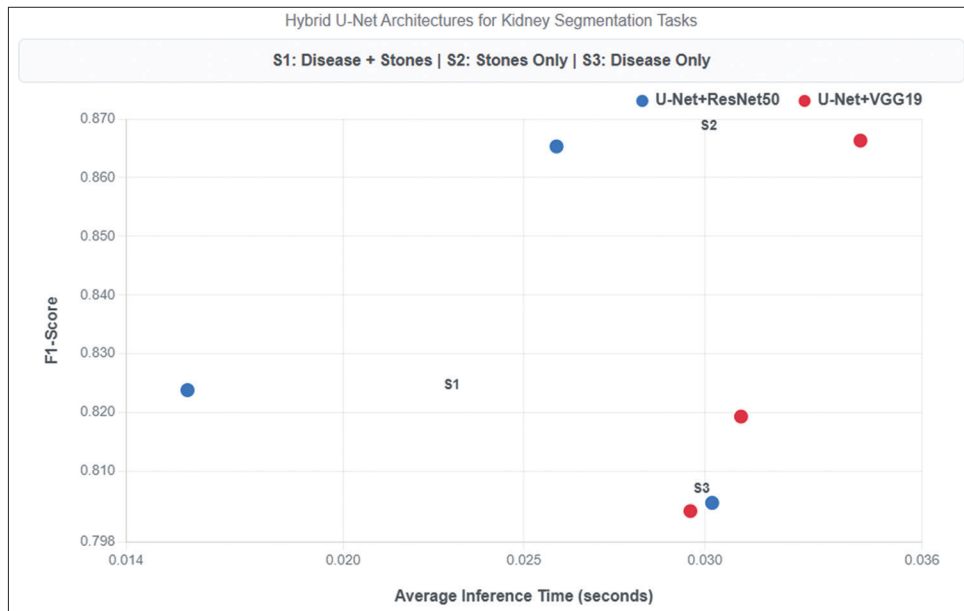


Fig. 13. Comparative chart of F1-score versus inference time for both models.

TABLE 8: A comparative analysis of the employed models across three scenarios, U-Net in conjunction with ResNet-50 and VGG-19

M	S	B-S	IS	Epo	Acc	Pre	Rec	F1-S	DC	IoU	Spe	CK	ATT
U-Net	1	24	128×128	200	0.9992	0.8301	0.8176	0.8238	0.8238	0.7004	0.9996	0.8234	0.0157 s
and	2	8	256×256	200	0.9998	0.879	0.8521	0.8653	0.8653	0.7626	0.9999	0.8652	0.0259 s
ResNet50	3	8	256×256	100	0.9987	0.8142	0.7951	0.8046	0.8046	0.673	0.9994	0.8039	0.0302 s
U-Net	1	8	256×256	200	0.9992	0.8291	0.8097	0.8193	0.8193	0.6939	0.9996	0.8189	0.031 s
and	2	8	256×256	150	0.9998	0.874	0.8587	0.8663	0.8663	0.7641	0.9999	0.8662	0.0343 s
VGG19	3	8	256×256	150	0.9987	0.8	0.8065	0.8032	0.8032	0.6712	0.9993	0.8026	0.0296 s

M: Model, S: Scenario, B-S: Batch-size, IS: Image Size, Epo: Epochs, Acc: Accuracy, Pre: Precision, Rec: Recall, F1-S: F1-Score, DC: Dice coefficient, Spe: Specificity, CK: Cohen's kappa, ATT: Average time test

model demonstrated unwavering stability throughout all three scenarios, a consequence presumably attributable to its hierarchical feature representation, which maintains intricate spatial information essential for precisely defining complicated renal diseases. The architectural differences explain the observed performance variances and have practical consequences for clinical practice. Enhanced segmentation of kidney stones and renal diseases might augment diagnostic reliability, enable precise quantification for treatment planning, and alleviate the manual workload on radiologists, hence promoting expedited and consistent clinical decision-making.

In addition, to evaluate the generalizability of the models to novel data, we re-examined the models, which excelled on the benchmark dataset, using the Kaggle dataset. The results showed that both hybrid models exhibited excellent

performance, demonstrating their robustness across various datasets. The U-Net + VGG19 hybrid model, trained with a batch size of 8, an image size of 256×256 , and for 150 epochs, attained an accuracy of 0.9997, a precision of 0.8495, a recall of 0.8502, an F1-score and DC of 0.8498, an IoU of 0.7389, a specificity of 0.9998, a CK of 0.8497, and an ATT of 0.0443 s. In the same way, the U-Net + ResNet50 hybrid model, employing the identical training configuration, achieved an accuracy of 0.9996, precision of 0.8499, recall of 0.8067, F1-score and DC of 0.8277, IoU of 0.7061, specificity of 0.9998, CK of 0.8276, and ATT of 0.0257 seconds. These results validate the efficacy and robustness of both models across various datasets.

Currently, the suggested hybrid U-Net models can be included in hospital (PACS) Picture Archiving and Communication Systems to autonomously segment kidney stones and

renal diseases from CT data. These outputs function as decision-support instruments, assisting radiologists in swiftly and reliably validating findings, while also supplying urologists with quantitative data for treatment formulation. Consequently, the models alleviate workload, enhance diagnostic reliability, and facilitate coordinated decision-making in authentic clinical workflows.

Several limitations should be acknowledged in this study. The relatively small dataset of 118 patients from a single institution, while supplemented with public data, may limit the generalizability of our findings across diverse populations, imaging protocols, and clinical settings. The lack of multimodal validation using other imaging modalities, such as MRI or US, restricts our ability to assess model performance across different diagnostic approaches commonly used in clinical practice. Furthermore, while our models demonstrate computational efficiency with processing times under 0.05 s per scan, the deployment of these DL architectures may pose significant challenges in resource-limited hospital environments where access to high-performance GPUs and adequate computational infrastructure is constrained. The memory requirements and initial setup costs for implementing such systems could create barriers for widespread clinical adoption, particularly in developing countries or smaller healthcare facilities. In addition, the requirement for specialized technical expertise to maintain and troubleshoot these AI systems may further limit their practical implementation in settings with limited IT support resources.

5. CONCLUSION

This research advances the field of automated medical image analysis by demonstrating that hybrid U-Net architectures can bridge the critical gap between segmentation accuracy and computational efficiency in kidney pathology detection. Our primary contribution lies in establishing a practical framework where established encoder architectures (ResNet50, VGG19) can be strategically integrated with U-Net decoders to achieve clinically viable performance without the prohibitive computational demands of transformer-based approaches. However, significant gaps remain in translating these technical achievements to diverse clinical environments, particularly regarding model robustness across different imaging protocols, patient demographics, and healthcare systems with varying computational resources. The geographical and institutional limitations of our validation

highlight the urgent need for federated learning approaches that can leverage multi-institutional datasets while preserving patient privacy. Future research should prioritize three critical directions: developing domain adaptation techniques to ensure model generalizability across different CT scanner types and imaging protocols, integrating explainable AI methodologies to build clinician trust and facilitate clinical adoption, and establishing standardized evaluation frameworks that can assess not only technical performance but also clinical utility in real-world healthcare workflows. The ultimate success of automated kidney pathology detection will depend on creating systems that complement rather than replace radiological expertise, requiring interdisciplinary collaboration between computer scientists, radiologists, and healthcare administrators to address both technical and implementation challenges.

6. ACKNOWLEDGMENT

The assistance of numerous individuals was essential to the accomplishment of this endeavor. Many thanks to my advisor, the College of Science, Soran University, and many thanks to the Department of Computer Science, as well as to my friend Jafar Majidpour, who helped me a lot, and my friend Dr. Dyari Awla Awla, a specialist in kidney and urology. And finally, thanks to my brother Karokh and my wife.

REFERENCES

- [1] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed and S. O. Olatunji. "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study". *Computers in Biology and Medicine*, vol. 109, pp. 101-111, 2019.
- [2] F. Ahmed, S. Abbas, A. Athar, T. Shahzad, W. A. Khan, M. Alharbi, M. A. Khan and A. Ahmed. "Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence". *Scientific Reports*, vol. 14, no. 1, p. 6173, 2024.
- [3] M. Akram, V. Jahrreiss, A. Skolarikos, R. Geraghty, L. Tzelves, E. Emilliani, N. F. Davis and B. K. Somani. "Urological guidelines for kidney stones: Overview and comprehensive update". *Journal of Clinical Medicine*, vol. 13, no. 4, p. 1114, 2024.
- [4] R. Magherini, E. Mussi, Y. Volpe, R. Furferi, F. Buonamici and M. Servi. "Machine learning for renal pathologies: An updated survey". *Sensors (Basel)*, vol. 22, no. 13, pp. 4989, 2022.
- [5] M. H. Hesamian, W. Jia, X. He and P. Kennedy. "Deep learning techniques for medical image segmentation: Achievements and challenges". *Journal of Imaging Informatics in Medicine*, vol. 32, no. 4, pp. 582-596, 2019.
- [6] Y. Wu and Z. Yi. "Automated detection of kidney abnormalities using multi-feature fusion convolutional neural networks". *Knowledge-Based Systems*, vol. 200, p. 105873, 2020.

- [7] T. A. Rashid, J. Majidpour, R. Thinakaran, M. Batumalay, D. A. Dewi, B. A. Hassan, H. Dadgar and H. Arabi. "NSGA-II-DL: Metaheuristic optimal feature selection with deep learning framework for HER2 classification in breast cancer". *IEEE Access*, vol. 12, pp. 38885-38898, 2024.
- [8] K. K. Patro, A. Jayaprakash, U. Rajendra Achary, M. Hammad, O. Yildirim and P. Pławiak. "Application of kronecker convolutions in deep learning technique for automated detection of kidney stones with coronal CT images". *Information Sciences*, vol. 640, p. 119005, 2023.
- [9] W. Brisbane, M. R. Bailey and M. D. Sorensen. "An overview of kidney stone imaging techniques". *Nature Reviews Urology*, vol. 13, no. 11, pp. 654-662, 2016.
- [10] D. A. Mahmood and S. A. Aminfar. "Efficient machine learning and deep learning techniques for detection of breast cancer tumor". *BioMed Target Journal*, vol. 2, no. 1, pp. 1-13, 2024.
- [11] S. Asif, M. Awais and S. U. R. Khan. "IR-CNN: Inception residual network for detecting kidney abnormalities from CT images". *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 12, no. 1, p. 35, 2023.
- [12] A. Parakh, H. Lee, J. H. Lee, B. H. Eisner, D. V. Sahani and S. Do. "Urinary stone detection on CT images using deep convolutional neural networks: Evaluation of model performance and generalization". *Radiology: Artificial Intelligence*, vol. 1, no. 4, p. e180066, 2019.
- [13] M. Kaur and D. Singh. "Fusion of medical images using deep belief networks". *Cluster Computing*, vol. 23, no. 2, pp. 1439-1453, 2020.
- [14] Z. H. Huang, Y. Y. Liu, W. J. Wu and K. W. Huang. "Design and validation of a deep learning model for renal stone detection and segmentation on kidney-ureter-bladder images". *Bioengineering*, vol. 10, no. 8, 970.
- [15] K. Yildirim, P. G. Bozdog, M. Talo, O. Yildirim, M. Karabatak and U. R. Acharya. "Deep learning model for automated kidney stone detection using coronal CT images". *Computers in Biology and Medicine*, vol. 135, p. 104569, 2021.
- [16] L. A. Fitri, A. Koudonas, G. Langas, S. Tsiakaras, D. Memmos, I. Mykoniatis, E. N. Symeonidis, D. Tsiptsios, E. Savvides, I. Vakalopoulos, G. Dimitriadis and J. De la Rosette. "Automated classification of urinary stones based on microcomputed tomography images using convolutional neural network". *Physical Medicine*, vol. 78, pp. 201-208, 2020.
- [17] W. Zhao, D. Jiang, J. Peña Queraltá and T. Westerlund. "MSS U-Net: 3D segmentation of kidneys and tumors from CT images with a multi-scale supervised U-Net". *Informatics in Medicine Unlocked*, vol. 19, p. 100357, 2020.
- [18] A. Alqahtani, S. Alsubai, A. Binbusayyis, M. Sha, A. Gumaei and Y. D. Zhang. "Optimizing kidney stone prediction through urinary analysis with improved binary particle Swarm optimization and extreme gradient boosting". *Mathematics*, vol. 11, no. 7, p. 1717.
- [19] S. W. Yang, Y. K. Hyon, H. S. Na, L. Jin, J. G. Lee, J. M. Park, J. Y. Lee, J. H. Shin, J. S. Lim, Y. G. Na, K. Jeon, T. Ha, J. Kim and K. H. Song. "Machine learning prediction of stone-free success in patients with urinary stone after treatment of shock wave lithotripsy". *BMC Urology*, vol. 20, no. 1, p. 88, 2020.
- [20] A. J. Daniel, C. E. Buchanan, T. Allcock, D. Scerri, E. F. Cox, B. L. Prestwich and S. T. Francis. "Automated renal segmentation in healthy and chronic kidney disease subjects using a convolutional neural network". *Magnetic Resonance in Medicine*, vol. 86, no. 2, pp. 1125-1136, 2021.
- [21] L. Ma, Y. Qiao, R. Wang, H. Chen, G. Liu, H. Xiao and R. Dai. "Machine learning models decoding the association between urinary stone diseases and metabolic urinary profiles". *Metabolites*, vol. 14, no. 12, p. 674.
- [22] I. Aksakalli, S. Kaçdioğlu and Y. S. Hanay. "Kidney x-ray images classification using machine learning and deep learning methods". *Balkan Journal of Electrical and Computer Engineering*, vol. 9, no. 2, pp. 144-151, 2021.
- [23] D. C. Elton, E. B. Turkbey, P. J. Pickhardt and R. M. Summers. "A deep learning system for automated kidney stone detection and volumetric segmentation on noncontrast CT scans". *Med Phys*, vol. 49, no. 4, pp. 2545-2554, 2022.
- [24] N. Blau, E. Klang, N. Kiryati, M. Amitai, O. Portnoy and A. Mayer. "Fully automatic detection of renal cysts in abdominal CT scans". *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, pp. 957-966, 2018.
- [25] X. Xiong, Y. Guo, Y. Wang, D. Zhang, Z. Ye, S. Zhang and X. Xin. "Kidney tumor segmentation in ultrasound images using adaptive sub-regional evolution level set models". *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi = Journal of Biomedical Engineering = Shengwu Yixue Gongchengxue Zazhi*, vol. 36, no. 6, pp. 945-956, 2019.
- [26] R. Gaikar, F. Zabihollahy, M. W. Elfaal, A. Azad, N. Schieda and E. Ukwatta. "Transfer learning-based approach for automated kidney segmentation on multiparametric MRI sequences". *Journal of Medical Imaging*, vol. 9, no. 3, p. 036001, 2022.
- [27] X. Fu, H. Liu, X. Bi and X. Gong. "Deep-learning-based CT imaging in the quantitative evaluation of chronic kidney diseases". *Journal of Healthcare Engineering*, vol. 2021, p. 3774423, 2021.
- [28] M. Goyal, J. Guo, L. Hinojosa, K. Hulse and I. Pedrosa. "Automated kidney segmentation by mask R-CNN in T2-weighted magnetic resonance imaging". In: *Medical Imaging 2022: Computer-Aided Diagnosis*, vol. 12033. Springer, Berlin, pp. 803-808, 2022.
- [29] G. Sharma, V. Anand, R. Chauhan, H. S. Pokhariya, S. Gupta and G. Sunil. "Transfer Learning Empowered Multi-Class Classification of Kidney Diseases: A Deep Learning Approach". In: *2024 2nd International Conference on Advancement in Computation and Computer Technologies (InCACCT)*, pp. 240-245, 2024.
- [30] H. Göker. "Transfer Learning-Based Classification of Kidney Stone From Computed Tomography Images". In: *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1-7, 2024.
- [31] M. N. Islam, M. Hasan, M. K. Hossain, M. G. R. Alam, M. Z. Uddin and A. Soylu. "Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography". *Scientific Reports*, vol. 12, no. 1, p. 11440, 2022.
- [32] F. J. N. Mactina and S. Neduncheliyan. "Multi-classification of kidney abnormalities in sonography using the LOA-MFO and long-term recurrent convolutional network". *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 13577-13612, 2024.
- [33] A. Odenrick, N. Kartalis, N. Voulgarakis, F. Morsbach and L. Loizou. "The role of contrast-enhanced computed tomography to detect renal stones". *Abdominal Radiology*, vol. 44, pp. 652-660, 2019.
- [34] M. S. B. Islam, M. S. I. Sumon, R. Sarmun, E. H. Bhuiyan and M. E. H. Chowdhury. "Classification and segmentation of kidney MRI images for chronic kidney disease detection". *Computers and Electrical Engineering*, vol. 119, p. 109613, 2024.
- [35] U. Kilic, I. Karabey Aksakalli, G. Tumuklu Ozyer, T. Aksakalli, B. Ozyer and S. Adanur. "Exploring the effect of image

- enhancement techniques with deep neural networks on direct urinary system X-Ray (DUSX) images for automated kidney stone detection". *International Journal of Intelligent Systems*, vol. 2023, no. 1, p. 3801485, 2023.
- [36] G. Chartrand, P. M. Cheng, E. Vorontsov, M. Drozdal, S. Turcotte, C. J. Pal and S. Kadoury, A. Tang. "Deep learning: A primer for radiologists". *Radiographics*, vol. 37, no. 7, pp. 2113-2131, 2017.
- [37] H. P. Chan, R. K. Samala, L. M. Hadjiiski and C. Zhou. "Deep learning in medical image analysis". *Deep learning in Medical Image Analysis: Challenges and Applications*. Berlin: Springer, pp. 3-21, 2020.
- [38] A. Abdelrahman and S. Viriri. "FPN-SE-ResNet model for accurate diagnosis of kidney tumors using CT images". *Applied Sciences*, vol. 13, no. 17, p. 9802, 2023.
- [39] 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: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*". Springer, pp. 234-241, 2015.