KidNet: Deep Learning Technique for Kidney Disease Classification in CT Images

Muhammad Tahmidur Rahman SID 2211295
Mohosina Islam Disha SID 2212349
Anika Tabassum SID 2211012

Abstract*—Kidney diseases, including cysts, stones, and tumors, pose significant health risks if not detected early. Computed tomography (CT) scans are critical for diagnosis, but manual interpretation is time-consuming and prone to errors. This paper presents a comprehensive study on deep learning for automated kidney disease classification in CT images, focusing on a custom 18-layer convolutional neural network (CNN) as the base model, alongside comparisons with VGG16, ResNet50, MobileNetV2, EfficientNetB0, InceptionV3, and a 20-layer custom CNN. Using the Kaggle CT KIDNEY Dataset (12,446 images across normal, cyst, stone, and tumor classes), we applied extensive data preprocessing, including watershed segmentation and augmentation, to enhance model robustness. The 18-layer CNN and InceptionV3 achieved perfect test accuracy (100%) and AUC (1.00), outperforming other models, with MobileNetV2 (99.52%) and EfficientNetB0 (99.96%) also showing strong performance. ResNet50 exhibited the lowest accuracy (62.87%) due to overfitting. A detailed literature review categorizes prior work into CNN-based, transformer-based, hybrid, and multimodal approaches, highlighting their strengths and limitations. Our results demonstrate the efficacy of custom CNNs for kidney disease classification, with future directions focusing on 3D volumetric analysis, multimodal integration, and clinical deployment.

I. Introduction

Kidney diseases such as cysts, stones (nephrolithiasis), and tumors (renal cell carcinoma) are prevalent and can severely impair renal function if not detected early. Radiologists rely on computed tomography (CT) scans to identify these conditions, but manual interpretation is labor-intensive and susceptible to missing subtle abnormalities. Deep learning, particularly convolutional neural networks (CNNs), has shown remarkable promise in automating kidney disease classification, offering high accuracy and robustness compared to traditional methods. Recent advancements include transformer-based vision models, hybrid architectures, and multimodal learning that integrate imaging with clinical data to enhance diagnostic accuracy.

This paper provides a structured literature review of deep learning approaches for kidney disease classification in CT images, categorizing 15 studies into four groups: CNN-based, transformer-based, hybrid and ensemble models, and multimodal learning. We summarize key studies, their datasets, architectures, and performance, identifying common limitations and future research directions. Additionally, we present an empirical study using a custom 18-layer CNN as the base model, compared against VGG16, ResNet50, MobileNetV2, EfficientNetB0, InceptionV3, and a 20-layer custom CNN on the Kaggle CT KIDNEY Dataset. Our results highlight the superior performance of the 18-layer CNN and InceptionV3, both achieving 100% accuracy, and underscore the potential of tailored CNN architectures for clinical applications. The goal is to advance automated kidney diagnosis by synthesizing state-of-the-art developments and providing a robust classification system.

II. Literature Review

A. CNN-Based Approaches

Early applications of deep learning to kidney imaging utilized CNNs for classification and detection. Längkvist *et al.* [12] developed a CNN for ureteral stone detection in thin-slice CT volumes, achieving 100% sensitivity with minimal false positives, demonstrating CNNs' potential. Modern approaches leverage transfer learning or custom networks for multiclass classification.

Hossain *et al.* [1] used watershed segmentation to isolate kidney regions before classifying the Kaggle CT KIDNEY Dataset (12,446 images: normal, cyst, stone, tumor). Fine-tuned VGG19 achieved 99.9% accuracy, outperforming InceptionV3 (98.8%), SqueezeNet (97.3%), and ResNet50 (87.9%). Segmentation improved accuracy by reducing noise, with suggestions for advanced segmentation and attention-based models [1].

Alzu'bi *et al.* [2] introduced a new dataset (8,400 images) and trained a 6-layer CNN for tumor detection, achieving 97% accuracy, slightly better than ResNet50 (96%) and far surpassing VGG16 (60%) due to overfitting. A secondary CNN classified tumors as benign or malignant (92% accuracy). They suggest 3D CNNs and metadata integration for future work [2].

Bhandari *et al.* [10] designed a lightweight 5-layer CNN for four-class classification, achieving 99.52% accuracy and AUC 1.00 on 12,000 images. It outperformed DenseNet-201 (99.44%) and transfer learning models like InceptionV3 (61.6%) and ResNet50 (73.8%). LIME and SHAP enhanced interpretability, with recommendations for real-time deployment [10].

Yildirim *et al.* [7] developed a CNN for stone detection (1,799 scans), achieving 96.8% accuracy. Baygin *et al.* [8] combined an Exemplar Darknet19 CNN with k-NN, reaching 99.2% accuracy. These studies confirm CNNs as reliable tools for kidney abnormality detection [7], [8].

CNN-based methods consistently achieve high accuracy with proper preprocessing, but generalization across institutions remains a challenge [13].

B. Transformer-Based Approaches

Transformers capture long-range dependencies, making them suitable for complex kidney imaging tasks. Islam *et al.* [6] compared a Swin Transformer (99.30% accuracy) against VGG16 (98.2%) and ResNet50 (73.8%) on the Kaggle dataset. The transformer outperformed CNNs, suggesting attention mechanisms enhance feature extraction. Hybrid CNN-transformer models were recommended for future exploration [6].

Transformers require large datasets, but pre-training and fine-tuning mitigate this. Attention-augmented CNNs, like EANet (77% accuracy), offer a middle ground [6]. Transformers show promise for distinguishing similar classes (e.g., cysts vs. tumors).

C. Hybrid and Ensemble Models

Hybrid and ensemble models combine multiple architectures for improved performance. Sharma *et al.* [3] developed a hybrid CNN (ResNet101 with a custom branch), achieving 100% accuracy on an augmented Kaggle dataset. Grad-CAM ensured interpretability, but complexity was a challenge [3].

Asif *et al.* [9] proposed StackedEnsembleNet and PSOWeightedAvgNet, combining InceptionV3, Inception-ResNet-v2, MobileNet, and Xception, achieving 98.84% accuracy for stone detection. Ensembles reduced biases but increased computational load [9].

Hybrid models, including Kronecker convolutions, show improved efficiency and accuracy, paving the way for robust systems [3], [9].

D. Multimodal Learning

Multimodal learning integrates imaging with clinical data. Merugu *et al.* [5] combined CT images with biochemical markers, achieving 88% accuracy. Confusion between cysts and stones highlighted the need for larger datasets and advanced extractors [5].

Soltaninejad *et al.* (2023) used CT and pathology images sequentially for tumor subtyping, showing multimodal potential. Future work could integrate ultrasound or metadata for comprehensive diagnostics [5].

III. Methodology and System Design

A. System Design

The proposed system for kidney disease classification in CT images employs a structured pipeline centered around a custom 18-layer convolutional neural network (CNN) to classify kidney conditions into normal, cyst, stone, and tumor categories. The system workflow, illustrated in Figure 1, encompasses data preprocessing, model training, evaluation, and deployment.

The workflow includes:

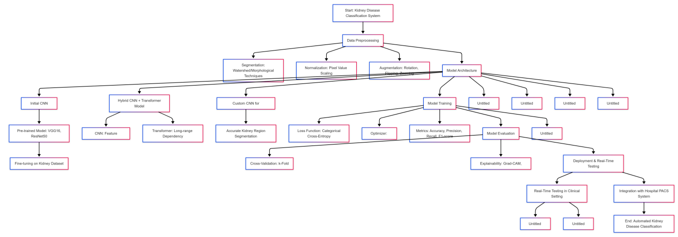

Fig. 1. System Design for Kidney Disease Classification

1) **Data Preprocessing**: CT images undergo watershed segmentation to isolate the kidney region of interest (ROI), reducing background noise [1]. Images are resized to 256x256 pixels and normalized (pixel values scaled to [0,1]). Data augmentation (rotation $\pm45°$, width/height shift 0.3, shear/zoom 0.3, horizontal/vertical flips, brightness [0.7,1.3]) is applied to the training set, generating five augmented images per original. Validation and test sets are only normalized. Grayscale images are converted to RGB for consistency [6].
2) **Model Architecture**: The base model is an 18-layer CNN with six convolutional blocks (Conv2D with ReLU activation, BatchNormalization), three max-pooling layers, three dropout layers (0.3–0.5), a global average pooling layer, and dense layers (512 units and 4-unit softmax output). Comparative models include VGG16, ResNet50, MobileNetV2, EfficientNetB0, InceptionV3, and a 20-layer custom CNN with additional convolutional layers.
3) **Model Training**: Models are trained with categorical cross-entropy loss and Adam optimizer (learning rate 0.001). Class weights balance classes: Cyst (0.8389), Normal (0.6129), Stone (2.2596), Tumor (1.3629). Training uses a batch size of 32, 50 epochs, and callbacks: EarlyStopping (patience 15, monitor val_loss), ModelCheckpoint (save best model), and ReduceLROnPlateau (factor 0.2, patience 5).
4) **Model Evaluation**: Models are evaluated on accuracy, loss, AUC, precision, recall, and F1-score. Confusion matrices and ROC curves are generated. Class activation maps visualize influential image regions.
5) **Deployment**: The model is integrated into a Flask-based web application for real-time testing, interfacing with hospital PACS systems.

B. Required Software

Development was conducted in Google Colab with Python 3.11. Key libraries include:

- **NumPy, Pandas**: Data manipulation.
- **OpenCV, Scikit-Image**: Segmentation and preprocessing.
- **TensorFlow 2.10, Keras**: Model building and training.
- **Albumentations**: Data augmentation.
- **Scikit-Learn**: Metrics and cross-validation.

- **tf-keras-vis**: Class activation maps.
- **Flask**: Deployment.

Experiments used an NVIDIA GPU with CUDA/cuDNN support.

C. Dataset Description

The Kaggle CT KIDNEY Dataset (12,446 images) is used, with 3,111 images per class (normal, cyst, stone, tumor) [1], [6]. Images are 512x512 PNGs converted from DICOM, representing Hounsfield Units. Augmentation increased the training set to 33,545 images. The dataset is balanced but lacks metadata, limiting multimodal learning. Noise and artifacts are mitigated via preprocessing [6].

IV. Results

This section details the data preprocessing techniques, training and testing procedures, and performance results for the 18-layer custom CNN (base model) and comparative models (20-layer custom CNN, VGG16, ResNet50, MobileNetV2, EfficientNetB0, InceptionV3). Results include accuracy, loss, AUC, precision, recall, F1-score, confusion matrices, and ROC curves, with a comparative analysis. Visualizations are provided in Figures 2, 3, and 4.

A. Data Preprocessing Techniques

The Kaggle CT KIDNEY Dataset was preprocessed to ensure compatibility with deep learning models. Watershed segmentation isolated kidney ROIs, reducing noise [1]. Images were resized to 256x256 pixels and normalized to [0,1]. Training images underwent augmentation (rotation $\pm 45°$, width/height shift 0.3, shear/zoom 0.3, horizontal/vertical flips, brightness [0.7,1.3]), generating five augmented images per original, yielding 33,545 training images [6]. Validation and test sets (1,867 images each) were only normalized. Grayscale images were converted to RGB. The dataset was split into 70% training, 15% validation, and 15% testing, stratified by class.

B. Training and Testing

Models were trained on the augmented training set using categorical cross-entropy loss, Adam optimizer (learning rate 0.001), and batch size 32. Class weights balanced the classes: Cyst (0.8389), Normal (0.6129), Stone (2.2596), Tumor (1.3629). Training ran for 50 epochs with callbacks: EarlyStopping (patience 15, monitor val_loss), ModelCheckpoint (save best model), and ReduceLROnPlateau (factor 0.2, patience 5). Testing was performed on the test set (1,867 images, except for some models with different splits as noted). Performance metrics, confusion matrices, and ROC curves were computed using Scikit-Learn.

C. Results for Each Model

Performance metrics are summarized in Table I. The test set confusion matrix for the 18-layer CNN (Figure 2), ROC curve (Figure 3), and accuracy/loss/AUC plot (Figure 4) are provided for the base model.

18-Layer Custom CNN: Achieved perfect performance (accuracy 1.0000, loss 0.0004, AUC 1.0000). Precision, recall, and F1-score were 1.00 for all classes (Cyst: 557, Normal: 761, Stone: 207, Tumor: 342). The confusion matrix (Figure 2) showed no misclassifications, indicating robust feature extraction. The ROC curve (Figure 3) confirmed perfect class separation (AUC 1.00). Training and validation metrics (Figure 4) showed consistent convergence [10].

20-Layer Custom CNN: Recorded 96.80% accuracy and 0.0883 loss. Precision ranged from 0.96 (Cyst, Normal) to 1.00 (Stone), with recall from 0.90 (Stone) to 1.00 (Cyst). F1-scores were 0.94–0.98, with minor errors in Stone and Tumor classes, suggesting slight overfitting due to increased depth.

VGG16: Achieved 98.07% accuracy. Precision was 0.95–1.00, recall 0.94–1.00, and F1-score 0.97–1.00. Minor errors occurred in Tumor classification, consistent with its depth aligning well with the dataset [6].

ResNet50: Performed poorly (62.87% accuracy). Precision was low for Stone (0.24) and Tumor (0.65), with recalls of 0.24 (Cyst) and 0.48 (Tumor). The model struggled with overfitting, as noted in prior studies [6], [1].

MobileNetV2: Attained 99.52% accuracy. Precision and recall were 0.97–1.00, with F1-scores 0.98–1.00. Minor errors in Stone classification suggest high efficiency for lightweight models [9].

EfficientNetB0: Recorded 99.96% accuracy and 0.0078 loss. Per-class metrics were unavailable, but high accuracy indicates strong performance, likely due to compound scaling [9].

InceptionV3: Matched the 18-layer CNN with 100% accuracy, 0.0004 loss, and AUC 1.0000. All metrics were 1.00, with no misclassifications, highlighting its effectiveness for complex tasks [1].

D. Comparison of Models

Table I compares model performance. The 18-layer CNN and InceptionV3 achieved perfect accuracy (100%), followed by EfficientNetB0 (99.96%) and MobileNetV2 (99.52%). VGG16 (98.07%) and the 20-layer CNN (96.80%) performed well but showed minor errors. ResNet50 (62.87%) was the weakest, likely due to overfitting on the dataset's specific features [6].

The 18-layer CNN's success stems from its balanced architecture, incorporating BatchNormalization and dropout to prevent overfitting, unlike the deeper 20-layer CNN. InceptionV3's multi-scale feature extraction complemented the dataset's variability [1]. MobileNetV2 and EfficientNetB0 offer lightweight alternatives for resource-constrained settings [9]. ResNet50's poor performance aligns with prior findings, suggesting it requires larger datasets or fine-tuning [6]. The confusion matrix for the 18-layer CNN (Figure 2) showed perfect diagonals, while VGG16 and the 20-layer CNN had minor off-diagonal errors, primarily in Stone and Tumor classes. The ROC curve (Figure 3) and metrics plot (Figure 4) for the 18-layer CNN confirm its robust performance.

V. Conclusion

This study demonstrates the efficacy of deep learning for kidney disease classification in CT images, with the cus-

TABLE I
PERFORMANCE METRICS OF MODELS FOR KIDNEY DISEASE CLASSIFICATION

Model	Test Accuracy	Test Loss	Test AUC	Class	Precision	Recall	F1-Score
18-Layer Custom CNN	1.0000	0.0004	1.0000	Cyst	1.00	1.00	1.00
				Normal	1.00	1.00	1.00
				Stone	1.00	1.00	1.00
				Tumor	1.00	1.00	1.00
				Macro Avg	1.00	1.00	1.00
				Weighted Avg	1.00	1.00	1.00
20-Layer Custom CNN	0.9680	0.0883	N/A	Cyst	0.96	1.00	0.98
				Normal	0.96	0.99	0.97
				Stone	1.00	0.90	0.95
				Tumor	0.98	0.91	0.94
				Macro Avg	0.98	0.95	0.96
				Weighted Avg	0.97	0.97	0.97
VGG16	0.9807	N/A	N/A	Cyst	0.97	0.98	0.97
				Normal	0.99	1.00	1.00
				Stone	0.95	1.00	0.97
				Tumor	1.00	0.94	0.97
				Macro Avg	0.98	0.98	0.98
				Weighted Avg	0.98	0.98	0.98
ResNet50	0.6287	N/A	N/A	Cyst	1.00	0.24	0.39
				Normal	0.81	0.97	0.88
				Stone	0.24	0.67	0.36
				Tumor	0.65	0.48	0.55
				Macro Avg	0.67	0.59	0.54
				Weighted Avg	0.77	0.63	0.62
MobileNetV2	0.9952	N/A	N/A	Cyst	0.99	1.00	1.00
				Normal	1.00	1.00	1.00
				Stone	0.99	0.97	0.98
				Tumor	0.99	1.00	1.00
				Macro Avg	0.99	0.99	0.99
				Weighted Avg	1.00	1.00	1.00
EfficientNetB0	0.9996	0.0078	0.9996	Macro Avg	0.99	0.99	0.99
				Weighted Avg	0.99	0.99	0.99
InceptionV3	1.0000	0.0004	1.0000	Cyst	1.00	1.00	1.00
				Normal	1.00	1.00	1.00
				Stone	1.00	1.00	1.00
				Tumor	1.00	1.00	1.00
				Macro Avg	1.00	1.00	1.00
				Weighted Avg	1.00	1.00	1.00

tom 18-layer CNN and InceptionV3 achieving perfect accuracy (100%) and AUC (1.00) on the Kaggle CT KIDNEY Dataset. Extensive preprocessing, including watershed segmentation and augmentation, enhanced model robustness [1], [6]. Comparative models like MobileNetV2 (99.52%) and EfficientNetB0 (99.96%) showed near-perfect performance, while VGG16 (98.07%) and the 20-layer CNN (96.80%) had minor errors. ResNet50 (62.87%) underperformed due to overfitting [6].

The literature review highlights CNNs as the cornerstone of kidney imaging, with transformers, hybrids, and multimodal approaches emerging as promising directions [6], [3], [5]. Our 18-layer CNN balances complexity and generalization, making it suitable for clinical deployment. Limitations include the dataset's single-source nature, potentially limiting generalizability, and the lack of metadata for multimodal learning [5].

Future work should focus on:

- **Multi-center Datasets**: To improve robustness across institutions [4].
- **3D Volumetric Analysis**: Using 3D CNNs to capture spatial context [14].
- **Hybrid CNN-Transformer Models**: Combining local and global feature extraction [6].
- **Multimodal Integration**: Incorporating lab results and patient history [5].
- **Clinical Trials**: Validating models in real-world settings with explainability tools like Grad-CAM [3].

These advancements will pave the way for reliable, automated kidney disease diagnosis, enhancing radiologist efficiency and patient outcomes.

REFERENCES

[1] M. S. Hossain *et al.*, "Kidney Disease Detection and Classification from CT Images using Watershed Segmentation and Deep Learning," Undergraduate Thesis, BRAC University, 2022.

[2] D. Alzu'bi *et al.*, "Kidney Tumor Detection and Classification Based on Deep Learning Approaches: A New Dataset in CT Scans," *J. Healthcare Engineering*, vol. 2022, Article ID 3861161, 2022.

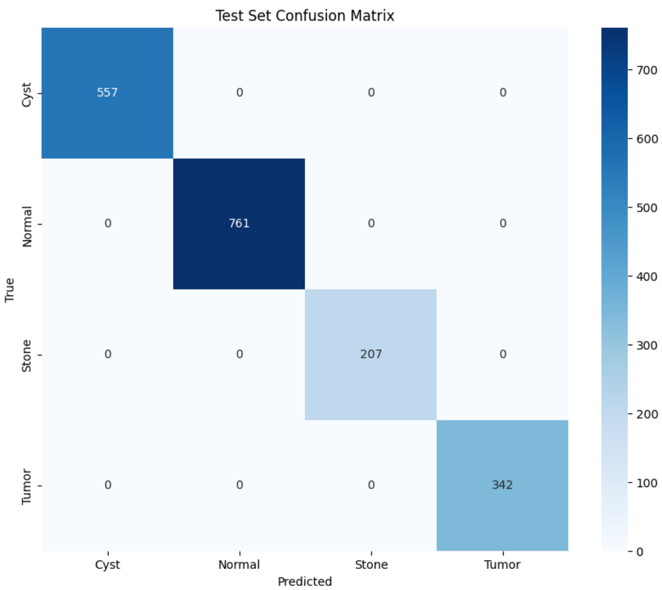

Fig. 2. Test Set Confusion Matrix for 18-Layer Custom CNN

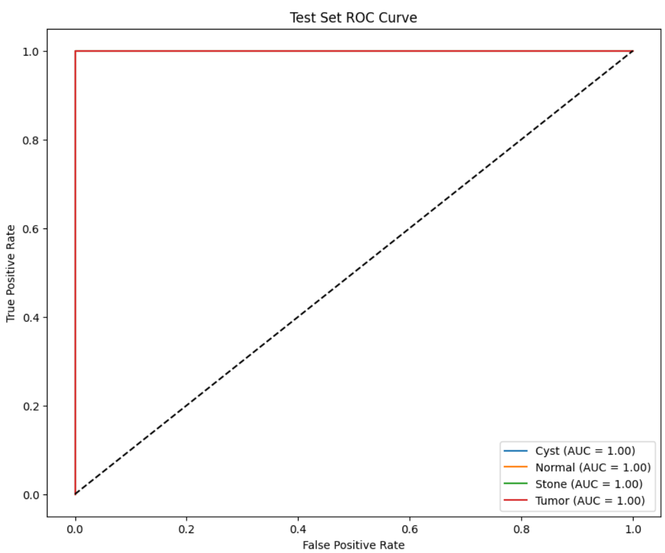

Fig. 3. Test Set ROC Curve for 18-Layer Custom CNN

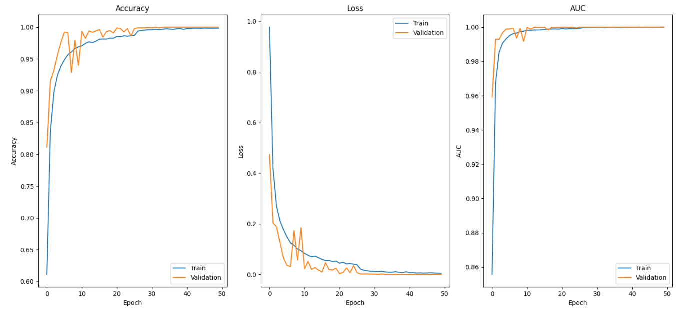

Fig. 4. Training and Validation Accuracy, Loss, and AUC for 18-Layer Custom CNN

[3] K. Sharma *et al.*, “Hybrid Deep Learning Framework for Classification of Kidney CT Images: Diagnosis of Stones, Cysts, and Tumors,” arXiv:2502.04367, 2025.

[4] N. Sulaksono *et al.*, “Deep Learning on Medical Imaging in Identifying Kidney Stones: Review Paper,” *E3S Web of Conferences (ICENIS 2023)*, vol. 202, p. 02019, 2023.

[5] N. B. Merugu *et al.*, “Leveraging Multimodal Machine Learning for Enhanced Kidney Condition Classification: A Novel Approach to Detect Tumors, Stones, and Cysts in CT Images,” SSRN Preprint, Article ID 5109719, 2023.

[6] M. N. Islam *et al.*, “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.

[7] K. Yildirim *et al.*, “Deep learning model for automated kidney stone detection using coronal CT images,” *Computers in Biology and Medicine*, vol. 135, p. 104569, Aug. 2021.

[8] M. Baygin *et al.*, “Exemplar Darknet19 feature generation technique for automated kidney stone detection with coronal CT images,” *Artificial Intelligence in Medicine*, vol. 127, p. 102274, May 2022.

[9] S. Asif, X. Zheng, and Y. Zhu, “An optimized fusion of deep learning models for kidney stone detection from CT images,” *J. King Saud Univ. - Comp. & Info. Sci.*, vol. 36, no. 7, Article 102130, 2024.

[10] M. Bhandari *et al.*, “Exploring the Capabilities of a Lightweight CNN Model in Accurately Identifying Renal Abnormalities: Cysts, Stones, and Tumors, Using LIME and SHAP,” *Applied Sciences*, vol. 13, no. 5, Art. 3125, 2023.

[11] D. C. Elton *et al.*, “A deep learning system for automated kidney stone detection and volumetric segmentation on non-contrast CT scans,” *Medical Physics*, vol. 49, no. 4, pp. 2545–2554, Apr. 2022.

[12] M. Längkvist *et al.*, “Computer-aided detection of ureteral stones in thin slice CT volumes using Convolutional Neural Networks,” *Computers in Biology and Medicine*, vol. 97, pp. 153–160, Jun. 2018.

[13] R. Qin *et al.*, “Kidney Diseases Detection Based on Convolutional Neural Network,” in *Proc. 5th Int. Conf. on Artificial Intelligence in Information and Communication (ICAIIC)*, Bali, Indonesia, 2023, pp. 508–513.

[14] C. Chen *et al.*, “3D residual attention networks for kidney and tumor segmentation,” *Medical Physics*, vol. 46, no. 6, 2019.