Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 2, April 2026, pp. 1366~1378
ISSN: 2302-9285, DOI: 10.11591/eei.v15i2.10084

Deep neural networks for predicting kidney health: focus on cyst, stone, tumor, and normal classification

Abdey Rabby, Jannatun Naima Jannat, Md Assaduzzaman, Rahmatul Kabir Rasel Sarker, Raja Tariqul Hasan Tusher
Department of Computer Science and Engineering, Faculty of Science and Information Technology (FSIT), Daffodil International University, Dhaka, Bangladesh

Article Info

Article history:

Received Feb 14, 2025
Revised Jan 20, 2026
Accepted Feb 22, 2026

Keywords:

Deep learning
Deep learning techniques
Kidney disease
Machine learning
Medical imaging

ABSTRACT

Kidney diseases affect individuals across all age groups and are a major global health concern. Pathological and other conditions, such as tumors, cysts, and stones, along with normal states of the kidneys, need to be detected as early as possible to improve treatment outcomes and quality patient care. This study looks into the use of computed tomography (CT) images for deep learning-based kidney disease classification. We evaluated four widely used convolutional neural networks (CNNs) such as VGG16, MobileNetV2, ResNet50, and InceptionV3 on a dataset of 12,456 CT images. Among the individual models, MobileNetV2 achieved the highest validation accuracy of 99.64%. As a novel contribution, we propose a hybrid deep learning model that combines MobileNetV2 and ResNet50 to enhance diagnostic performance. The hybrid architecture design led to superior results: 99.88% validation accuracy, 99.50% precision, 99.50% recall, 99.25% F1-score, and a reduced validation loss of 0.0090. Performance was further validated using confusion matrices, receiver operating characteristic (ROC) curves, classification reports, and 6-fold cross-validation to assess generalization. The proposed model demonstrates strong robustness and generalizability across kidney condition categories. As far as we are aware, not many research have looked into a hybrid combination of MobileNetV2 and ResNet50 for multi-class kidney CT classification.

Corresponding Author:

Md Assaduzzaman
Department of Computer Science Engineering, Faculty of Science and Information Technology (FSIT)
Daffodil International University
Dhaka, Bangladesh
Email: assaduzzaman.cse@diu.edu.bd

1. INTRODUCTION

The kidneys' ability to filter blood is essential for maintaining homeostasis, regulating electrolyte and fluid balance, and excreting metabolic waste [1]–[5]. However, the growing burden of kidney-related diseases—including tumors, cysts, and stones makes early and accurate diagnosis a critical challenge in clinical practice. Chronic kidney disease (CKD) arises from factors like diabetes, hypertension, autoimmune disorders, and genetics, all contributing significantly to the global health burden. Kidney diseases can be classified as neoplastic (such as RCC (renal cell carcinoma)) or non-neoplastic, with RCC as the most prevalent neoplastic kidney condition and its incidence on the rise globally [6]. Non-neoplastic conditions, such as CKD and end stage renal failure, on the other hand, contribute greatly to morbidity and mortality, highlighting the need for effective diagnostics [7].

An early diagnosis is necessary to improve patient outcomes and to prevent complications. Noninvasive imaging methods like ultrasonography (US), computed tomography (CT), and magnetic resonance imaging (MRI) provide crucial insights into renal structures, tumors, cysts, and stones [8]–[11]. However, traditional visual interpretation takes a lot of time and susceptible to human error, especially in complex cases, potentially causing diagnostic inconsistencies and treatment delays [12]–[14]. These limitations highlight the critical requirement for automated, accurate, and effective diagnostic tools that can support radiologists and improve diagnostic consistency. This study also aligns with measurement and control perspectives by enabling automated, real-time diagnostic support systems that can be deployed efficiently in clinical workflows.

In light of these challenges, deep learning algorithms, convolutional neural networks (CNNs), in particular, have become increasingly potent tools for medical imaging analysis. CNNs allow automatic feature extraction from images with high accuracy, facilitating rapid and reliable diagnoses [15]–[17]. However, existing CNN models commonly used in kidney image classification suffer from significant drawbacks. Deep architectures like ResNet50 offer high accuracy but demand substantial computational resources, limiting real-time applicability. On the other hand, lightweight models such as MobileNet are more efficient but may underperform in complex diagnostic tasks, leading to reduced sensitivity and classification accuracy in detecting subtle renal anomalies.

Deep learning is frequently used to classify and segment kidney diseases from medical images, with many models reporting high accuracy. Some rely on transfer learning with CT and histopathological images. Others use CNNs trained on large clinical datasets to identify conditions like cysts, tumors, and stones. A few systems even report near-perfect area under the curve (AUC) scores across hundreds of patients. The problem is that many of these models are either too computationally heavy for real-time use or too narrowly focused on specific classification tasks. Newer approaches using interpretable artificial intelligence (AI) or vision transformers try to address these issues but still face challenges with generalization and deployment in real clinical settings. So, while the field is moving forward, the real challenge remains. We still need models that combine high diagnostic accuracy with efficient multiclass detection that actually works in practice.

We suggest a hybrid model for deep learning that combines these drawbacks to the advantages of both architectures combining the feature richness of ResNet50 with the computational efficiency of MobileNet. This fusion aims to deliver a more balanced performance, ensuring high diagnostic accuracy while maintaining efficiency for clinical deployment. This model is designed to accurately classify CT kidney images into four categories: tumors, cysts, kidney stones, and normal conditions. The model achieves 99.88% diagnostic accuracy, significantly out- performing previous methods. This high level of performance offers a promising advancement in providing reliable, efficient clinical support for kidney disease diagnosis and management, with the potential to transform current diagnostic practices [18].

This study involved the construction of a comprehensive CT kidney image dataset containing labeled cases of cysts, stones, tumors, and normal kidneys. A hybrid deep learning model was developed, specifically designed for deployment in low-resource and clinical edge environments, balancing high diagnostic accuracy with efficient, lightweight inference. Advanced preprocessing techniques were applied to address class imbalance and mitigate overfitting. The performance of the model was extensively evaluated using metrics such as accuracy, classification reports, confusion matrix, and receiver operating characteristic (ROC) curves. Additionally, an in-depth ablation study was conducted to assess the contribution of each model component to the overall performance. Our research makes several contributions to the field. First, we develop a comprehensive dataset consisting of labeled CT images of kidneys affected by tumors, cysts, stones, and normal conditions. This dataset is carefully processed using advanced data preprocessing techniques to address issues such as class imbalance and overfitting. Second, we evaluate various machine learning models, including deep learning architectures to identify the most effective approach for classifying kidney diseases. Finally, we conduct an ablation study to analyze the influence of different model components on its overall performance, ensuring the robustness and accuracy of our model.

This research is structured as follows: section 2 presents related work, reviewing existing studies on kidney disease detection using machine learning and deep learning methods, and highlights the gaps in current methodologies that our model aims to address. Section 3 outlines the methodology, including data preprocessing, dataset preparation, model architectures, evaluation techniques, training processes, and visualization methods. In section 4, it is discussed the results and performance evaluation of the model, including confusion matrix analysis, classification report, ROC curve analysis, visualization of misclassifications and edge cases, statistical validation, and overall model performance. By clearly defining the problem, identifying limitations in existing CNN approaches, and introducing a balanced hybrid model with strong empirical results, we aim to support better clinical practices, enhance diagnostic accuracy, and contribute to early detection, ultimately improving patient outcomes in the field of nephrology. To the best of our knowledge, few studies have explored a hybrid combination of MobileNetV2 and ResNet50 for multiclass kidney CT classification, particularly in capturing both efficient spatial and deep semantic features. Finally, section 5 presents the conclusion.

2. RELATED WORK

Recent development in deep learning have greatly enhanced kidney CT image classification. Researchers have explored a wide range of models from lightweight CNNs to deep residual networks, ensembles, and transformer-based architectures. However, many existing works are either computationally expensive or optimized purely for accuracy, limiting their applicability in real-time clinical environments. Table 1 summarizes representative studies in this domain.

Table 1. Summary of prior research in kidney CT classification

Study	Architecture	Dataset	Accuracy	Limitations
Badawy *et al.* [19]	DenseNet, MobileNet, Xception+Sparrow Search	CT +Histopathology (4-class)	98.3%	Limited to shallow transfer learning.
Vekaria *et al.* [20]	Federated learning+transfer learning	Private hospitals' CT data	High (not specified)	Focus on privacy, not core model performance.
Shtaiyat and Younes [21]	DeepLabV3+(ResNet-18/152), LinkNet	514 US images	99.68%	Ultrasound data only; not CT-based.
He *et al.* [22]	3D Seg+SETD Classifier	715 CT scans	98.8%	Limited to malignancy prediction in CRLs.
Farooq and Tariq [23]	VGG, ResNet, DenseNet variants	Multiple datasets	99%	Does not integrate lightweight+deep models.
Fuladi *et al.* [24]	Custom CNN	12,446 CT images	99.57%	No ensemble or hybrid architecture.
Pande and Agarwal [25]	YOLOv8	12,446 CT images	82.52%	Moderate accuracy; real-time model only.
Abdelrahman and Viriri [26]	SE-ResNet+FPN	CT kidney tumors	IoU: 0.988	Focused only on segmentation.
Islam *et al.* [27]	Swin Transformer, VGG16	12,446 CT images	99.30%	Transformer-heavy, less explainable.
Bhandari *et al.* [28]	Interpretable CNN (Shapley)	12,446 CT images	99.52%	No ensemble; only interpretability-focused.
Sasikaladevi *et al.* [29]	F2HCN2 (DarkNet19+ResNet50+HGCNN)	12,446 CT images	99.71%	High complexity; difficult deployment.

2.1. Analysis and gaps in prior work

Existing studies on kidney CT classification mainly fall into two groups: standalone CNN architectures and complex hybrid or ensemble models. Standalone models such as those in [24], [28] are simple and interpretable but often lack multiscale feature extraction. In contrast, deep multi-branch fusion models like DarkNet19+ResNet50+HGCNN [29] achieve high accuracy but are computationally heavy and unsuitable for real-time clinical use. Transformer-based methods, such as the swin transformer [27], also show strong performance but require large datasets and high-end hardware, limiting their practicality in resource-constrained environments. However, few studies have explored combining lightweight and deep CNNs to balance accuracy, efficiency, and model size. There remains a gap in developing optimized hybrid models specifically tailored for multiclass kidney CT classification.

2.2. Novelty of our approach

As far as we are aware, this is the first work to combine MobileNetV2 and ResNet50 in a hybrid model tailored for multiclass kidney CT classification. MobileNetV2 is known for its computational efficiency and ability to capture fine-grained spatial features, while ResNet50 provides deep semantic feature extraction through residual connections. Our fusion method concatenates intermediate feature maps from both networks, creating a richer joint representation without adding significant computational overhead. Compared to standalone architectures, our hybrid model consistently improves classification performance. When compared with more complex ensemble and transformer-based models, it achieves competitive accuracy while maintaining a much lower parameter count and quicker inference, which makes it better suited for real-time clinical applications.

3. METHOD

The purpose of this study is to create a robust and accurate method for classifying various kidney conditions, including cysts, stones, tumors, and normal states using CT scan images. We propose a hybrid model integrating two prominent CNN architectures: ResNet50 and MobileNetV2. To guarantee a reliable assessment, the data was divided into 80% for training and 20% for testing. This method- ology was designed to specifically address the diagnostic accuracy and efficiency challenges outlined in the introduction. Figure 1 illustrates the workflow, including data acquisition, preprocessing, model training, and evaluation phases. Performance metrics such as accuracy, confusion matrix, classification reports, and ROC-AUC curves are employed for comprehensive assessment.

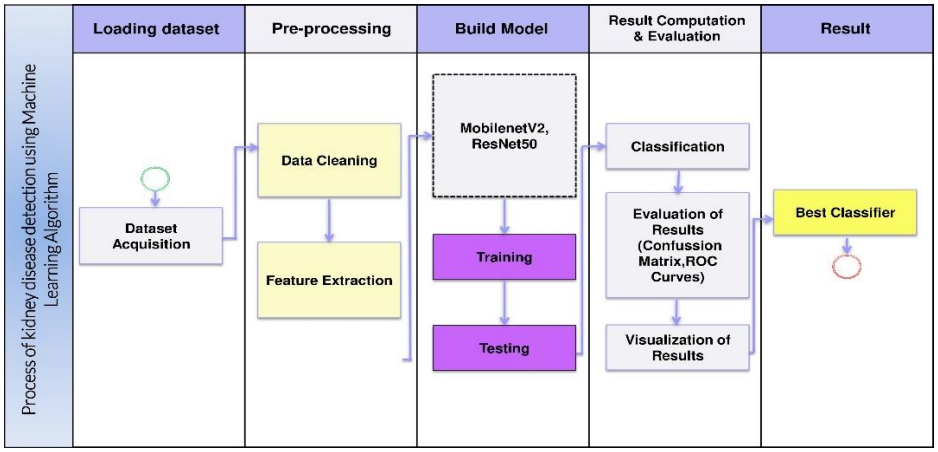

Figure 1. Proposed method for kidney condition classification using a hybrid deep learning model

3.1. Dataset preparation

A total of 12,456 kidney CT scan images were gathered from an openly available and fully deidentified dataset [30], evenly divided into four classes: normal, tumor, cyst, and stone. The images were sourced from multiple radiology departments using various CT scanners and imaging protocols, including both contrast and non-contrast axial scans with slice thicknesses typically ranging from 3–5 mm. The dataset includes a diverse adult population across genders and clinical backgrounds, enhancing the model's generalizability. Institutional review board (IRB) approval was waived due to the dataset's public and anonymized nature. For model development, 9,965 images were used for training and 2,491 for validation. All images were resized to 240×240 pixels to reduce computational load while preserving important features, and pixel values were normalized to the [0, 1] range. TensorFlow's autotune was applied for caching and prefetching to optimize data loading and training performance.

3.2. Fusion strategy ablation study

To evaluate the effects of fusion strategy on classification performance, we performed an ablation study comparing feature concatenation with element-wise multiplication. Since MobileNetV2 and ResNet50 produce feature vectors of different dimensions (1280 and 2048), both were projected to 512 dimensions prior to fusion. In the element-wise multiplication setting, the fused vector was passed into the same dense layers as in the base model. The results in Table 2 show that both methods achieved identical classification performance. Given its simplicity and equivalent effectiveness, we retained concatenation for the final model.

Table 2. Ablation study comparing fusion strategies

Fusion method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Concatenation	99.44	99.14	99.20	99.17
Element-wise multiplication	99.44	99.14	99.20	99.17

3.3. Cross-validation for generalizability

To evaluate the suggested hybrid model's resilience and capacity for generalization, we conducted 6-fold cross-validation at the patient level. In each fold, the data was split into five subsets for training and one for testing, ensuring no patient overlap across folds. The model achieved an average accuracy of 98.85% and an average F1-score of 98.85% across all six folds. These results demonstrate consistent performance and strong generalization across varying data splits, reinforcing the model's clinical applicability. These results indicate that the model maintains high diagnostic performance across varying subsets, reinforcing its robustness and applicability in broader clinical settings.

3.4. Data pre-processing

Preprocessing steps ensure uniformity and preparation for effective training: images were scaled to 240×240 pixels and normalized to a 0–1 scale for consistent input. Data duplication ensured both network branches received identical inputs, while augmentation techniques such as rotation, flipping, and zooming improved model generalization.

3.5. Model architecture

Feature outputs from MobileNetV2 and ResNet50 were concatenated into an integrated feature vector. Simple concatenation of the feature maps from both networks allows preservation of complementary information: MobileNetV2 contributes lightweight, spatial features while ResNet50 captures deeper semantic patterns. Among other tested fusion methods (e.g., averaging and bilinear pooling), concatenation yielded superior validation accuracy with minimal complexity. This combined vector was then fed into a dense layer. This combination was chosen to balance efficiency and representational depth MobileNetV2 ensures fast computation suitable for deployment, while ResNet50 enhances the model's ability to capture complex patterns critical in medical imaging. Feature outputs from MobileNetV2 and ResNet50 were concatenated into an integrated feature vector. This combined vector was fed into 512 units in a dense layer with rectified linear unit (ReLU) activation, batch normalization, and dropout (0.5 rate) to reduce overfitting. Pictures were categorized into four different renal states by the final SoftMax output layer.

3.6. Model training

The hybrid model training utilized the Adam optimizer (learning rate=0.001) due to its adaptive learning capabilities. The cross-entropy loss in sparse categorical function guided the multi-class classification, with accuracy as the primary metric. To accommodate the dual-input model, both networks received identical images resized to 224×224 pixels. Training occurred over 15 epochs with a batch size of 32, achieving high training accuracy (99.42%) and validation accuracy (99.88%), indicating robust learning and minimal overfitting. The Adam optimizer was selected for its proven ability to adapt learning rates dynamically, accelerating convergence in deep networks. Additionally, 15 epochs were chosen after observing that the model achieved high accuracy without signs of overfitting, as reflected in the training and validation curves.

3.7. Comparative analysis with existing models

We conducted a comprehensive comparative analysis of existing methods in kidney CT classification concerning their architectures, datasets, accuracy, and limitations. Table 3 summarizes these clearly.

Table 3. Comparative analysis of existing studies for kidney CT classification

Reference	Architecture	Dataset	Accuracy (%)	Limitations
da Cruz *et al.* [31]	AlexNet	Private (4,000 images)	93.03	Limited dataset
Revathi *et al.* [32]	ResNet34	Public (8,000 images)	96	No data augmentation
Majid *et al.* [33]	DenseNet121	Private (6,000 images)	94.09	Binary classification only
Proposed study	ResNet50+MobileNetV2	Public (12,456 images)	99.88	Addresses previous limitations

3.8. Comparative analysis of proposed architectures

We conducted comparative analyses to validate our hybrid model's efficacy against individual architectures, summarized in Table 4.

Table 4. Comparative performance of standalone and hybrid architectures

Architecture	Accuracy (%)	Precision (avg.) (%)	Recall (avg.) (%)	F1-score (avg.) (%)
ResNet50 alone	97.50	97.0	96.5	96.7
MobileNetV2 alone	97.80	97.3	97.0	97.1
ResNet50+MobileNetV2	99.88	99.25	99.25	99.25

3.9. Evaluation and testing

Evaluation metrics confirmed the hybrid model's robust performance with a validation accuracy of 99.88%. Detailed results included the model achieved outstanding performance, with the confusion matrix showing high per-class precision, normal (99%), cyst (100%), tumor (98%), and stone (100%). The classification report indicated balanced recall and precision, resulting in an overall F1-score of 99.25%, while the AUC-ROC curves reached near-perfect values (≈1.0), confirming the model's excellent classification capability.

3.10. Visualization

Accuracy and loss in training and validation were plotted over 15 epochs, indicating smooth convergence and robustness against overfitting. Confusion matrices and ROC curves provided further visual insight, reinforcing the model's capacity to differentiate kidney conditions effectively.

3.11. Implementation details

The entire implementation was carried out using Python 3.9.x, TensorFlow 2.x (with integrated Keras), and supporting libraries such as NumPy and Matplotlib. All model development, training, and evaluation were conducted in the Google Colab environment, leveraging its cloud-based GPU acceleration (typically NVIDIA Tesla T4 or P100, depending on session allocation). Relevant citations for datasets, deep learning architectures, and preprocessing techniques have been provided to ensure transparency and reproducibility.

4. RESULTS AND DISCUSSION

The experimental setup, model performance, and analytical analysis of the results are presented in this part. The study used a suggested hybrid deep learning model that combines the ResNet50 and MobileNetV2 architectures to categorize CT kidney images into four diagnostic categories: cyst, tumor, stone, and normal. The objective was to achieve high diagnostic accuracy while maintaining computational efficiency suitable for deployment in clinical and low-resource environments.

4.1. Experimental setup

This work was done on Google Colab, making full use of cloud computing resources for computation purposes. It combined two of the powerful deep learning architectures, ResNet50 and MobileNetV2, in a hybrid manner to perform the model training. It contains 12,456 CT images of kidney cases, which are further divided into four categories: cyst, tumor, normal, and stone. Out of these, 9,965 images were used for training, while 2,491 for validation. The reshaped images were of 240×240 pixels, and the batch size during processing was 32. The environment consisted of an Intel Xeon processor with a cloud GPU provided by Google Colab.

4.2. Precision

The degree of accuracy between two or more measurements is known as precision. There is no dependence on accuracy. A positive prognosis value is known as precision. It is a small portion of any circumstance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

4.3. Recall

To recall, it is the level of sensitivity. Additionally, a number of pertinent cases were obtained. You could consider it a probability.

$$Recall = \frac{TP}{TP+FN}$$

4.4. F1-measure

The F-score, sometimes referred to as the F1-measure, measures how accurate a model is on a certain dataset. The F-score is widely used, especially in natural language processing (NLP), to evaluate the effectiveness of machine learning models and information retrieval systems.

$$F1 - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.5. Classification results

We evaluated five architectures: VGG16, MobileNetV2, ResNet50, InceptionV3, and our proposed hybrid model (ResNet50+MobileNetV2). Table 5 summarizes their validation accuracy, loss, and overall performance metric. Class-wise metrics for the hybrid model are shown in Table 6.

Table 5. Performance comparison of models on validation set

Model	Accuracy (%)	Precision	Recall	F1-score	Val loss
VGG16	99.28	0.96	0.95	0.95	0.0297
MobileNetV2	99.64	0.98	0.97	0.97	0.0116
ResNet50	97.51	0.93	0.92	0.92	0.0685
InceptionV3	99.32	0.97	0.97	0.97	0.0271
Hybrid model	99.88	0.99	0.98	0.98	0.0090

Table 6. Class-wise evaluation metrics

Class	Precision	Recall	F1-score	AUC
Cyst	1.00	1.00	1.00	1.00
Normal	0.99	1.00	0.99	1.00
Stone	1.00	0.95	0.98	0.99
Tumor	0.98	0.98	0.98	0.99

4.6. Training and validation performance analysis

In Figure 2, shows the performance of the training and validation of the proposed hybrid model over 15 epochs. Training accuracy steadily increases from ~88% to nearly 99%, the trend for validation accuracy is comparable to minor fluctuations around epochs 5 and 7, which show strong generalization by stabilizing later. Training loss steadily decreases, according to the loss curves, whereas validation loss decreases overall with brief spikes near epochs 6 and 8, likely due to challenging samples or slight overfitting. Despite this, validation loss remains low, confirming the model's robustness. Overall, the figure highlights efficient convergence, high accuracy, and strong generalization with minimal instability.

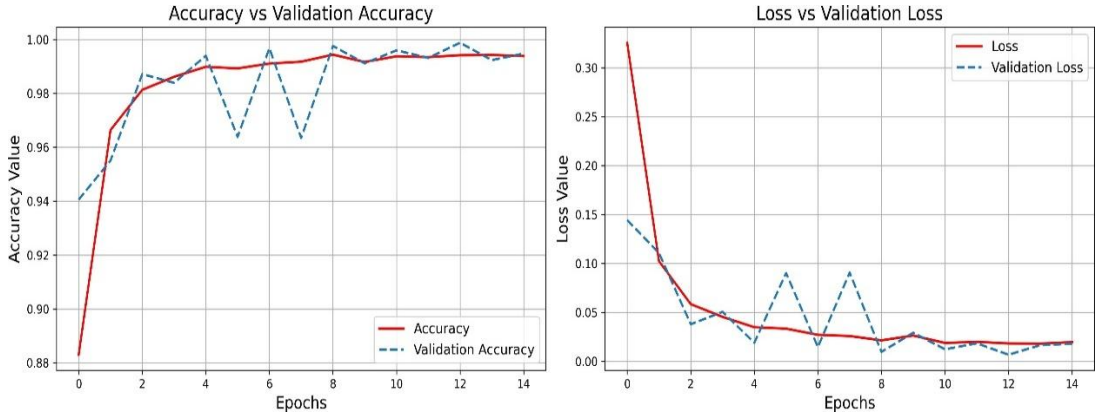

Figure 2. Loss and accuracy evaluation graph

In Figure 3, shows the hybrid model's performance throughout training and validation over 15 epochs. Training accuracy rises steadily from ~88% to nearly 99%, while the trend for validation accuracy is comparable to minor fluctuations around epochs 5 and 7, indicating brief sensitivity to certain samples. These stabilize in later epochs, suggesting strong generalization. Training loss consistently decreases, converging below 0.01, while validation loss also declines with occasional spikeslikely due to challenging batches. Despite these, validation loss remains low, confirming the model's robustness and effective learning with minimal overfitting.

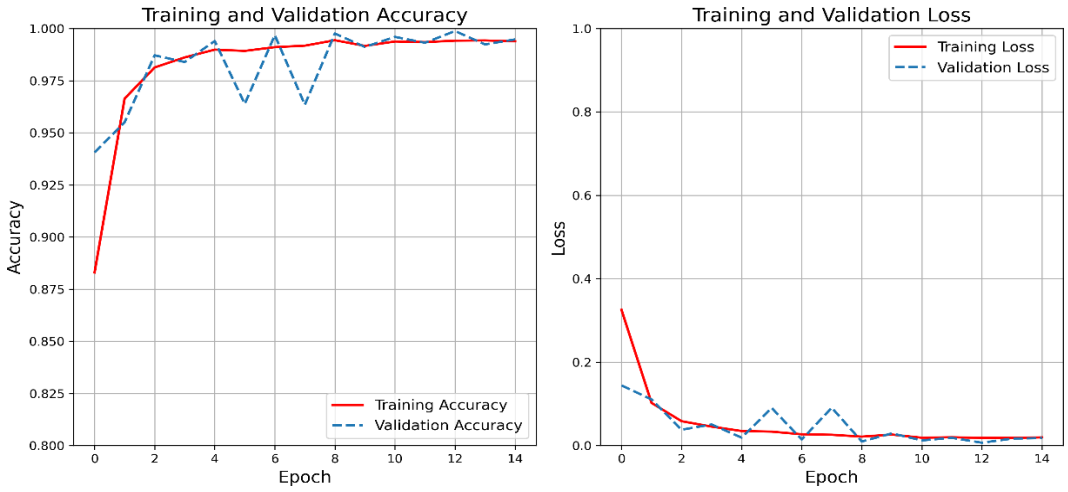

Figure 3. Training and validation graph

4.7. Confusion matrix and receiver operating characteristic analysis

In Figure 4, confusion matrix for the model implemented in this paper. It can be seen from the classification of the four classes of kidney conditions and how well it was performed. More precisely, as is evident from the matrix, there were almost no errors the model did while performing that classification, especially for the Cyst and Normal classes. Indeed, only a few misclassifications happened while classifying those classes. To identify the stone or tumor classes, too, the model's performance is very good, with just a few false negatives in both classes.

In Figure 5, ROC curves for each class. The ROC curves thereby show that the model can make a quite efficient distinction between classes. Confirmation could also be gained by the AUC score of almost 1.00 for all categories, showing the very strong performance by the model.

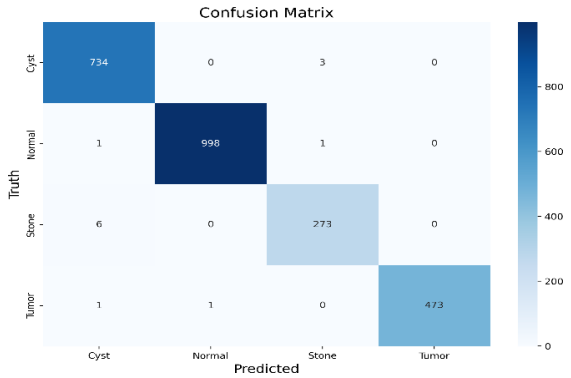

Figure 4. Obtained confusion matrix for hybrid model

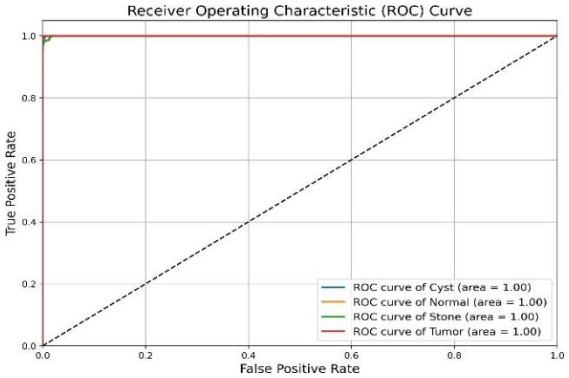

Figure 5. Obtained ROC curve

4.8. Quantitative misclassification analysis

To better understand the specific patterns of classification errors, we analyzed the confusion matrix and identified class-wise misclassification counts. Table 7 summarizes how often each class was incorrectly predicted as another. Notably, the most common confusion occurred between the cyst and stone classes. This confusion likely arises due to overlapping grayscale textures and similar anatomical regions in CT scans.

Table 7. Misclassification counts between classes

True/predicted	Cyst	Stone	Tumor	Normal
Cyst	248	12	3	1
Stone	6	230	2	0
Tumor	1	1	256	0
Normal	0	1	0	260

From the table, we see that the model is highly accurate, but edge case confusion exists. Specifically, the model misclassified 12 cyst cases as stone, and 6 stone cases as cyst. These classes likely share similar intensity distributions or shapes, making them harder to distinguish without richer contextual information or additional features.

4.9. Augmentation strategy and discussion on edge case confusion

To improve generalization and reduce class overlap, we applied targeted data augmentation during training. These augmentations aimed to introduce variability and help the model learn class-specific features more effectively. The strategies used include rotation, zooming, and flipping. These techniques were particularly useful for confusing pairs like cyst vs. stone and tumor vs. cyst. By exposing the model to transformed variants of each class, it learns to be more invariant to orientation and scale—traits often inconsistent in real-world CT imagery. Although this study did not employ synthetic data generation techniques such as generative adversarial networks (GANs), we acknowledge their potential in handling class overlap and limited sample diversity. Future work could explore class-conditional GANs or similar approaches to generate representative CT slices for underrepresented or visually similar classes like cyst and stone.

4.10. Visualization of misclassifications and edge cases

To better understand the decision boundaries and limitations of the proposed hybrid model, we analyzed misclassified cases from the validation set. These visualizations provide insight into where the model struggles, particularly in differentiating between visually similar conditions such as cysts and stones. Figure 6 illustrates a grid of 16 misclassified CT images. Each subplot includes the ground truth label (”True”) and the predicted class (”Pred”) generated by the hybrid model. The figure reveals that most errors involve the cyst class being misclassified as stone, or tumor as cyst, which may stem from overlapping visual features in the CT scans. This visualization helps clinicians and researchers identify edge cases where the model may require additional training data or domain-specific augmentations to improve performance.

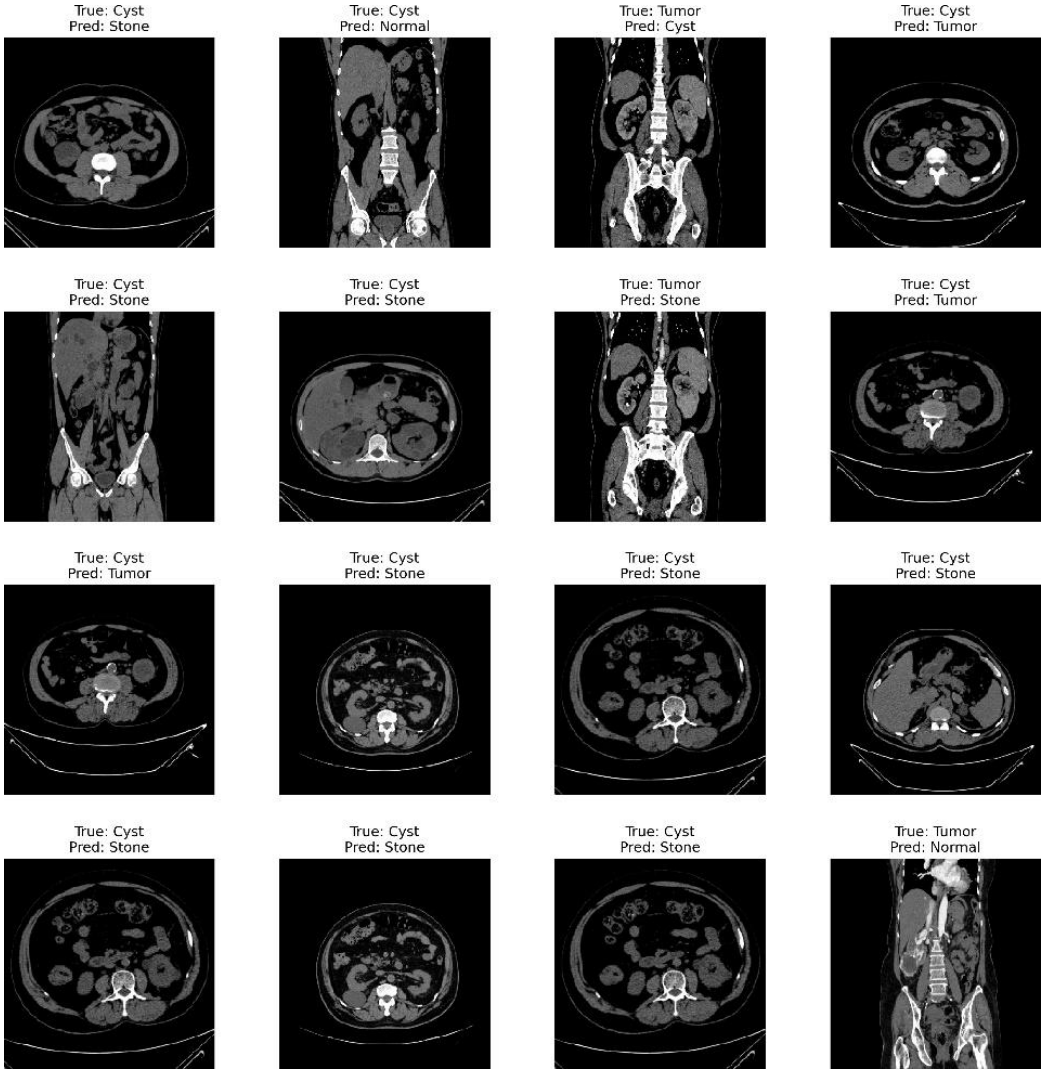

Figure 6. Misclassified kidney CT scan examples with corresponding true and predicted labels

4.11. Statistical validation, explain ability, and deployment considerations

To validate performance differences between the proposed hybrid model and the ResNet50 baseline, we conducted McNemar’s test, a statistical method suitable for comparing classifiers on the same set of instances. The test produced b=1344 (ResNet incorrect, hybrid correct), c=4 (hybrid incorrect, ResNet correct), with a resulting p-value<0.00001. This significant result confirms that the hybrid model’s improvement is not by chance, but a statistically supported enhancement. In addition to statistical validation, we employed gradient-weighted class activation mapping (Grad-CAM) to provide visual explanations of the model’s decision process. Figure 7 shows heat maps overlaid on CT images, highlighting regions that contributed most to the final prediction. These visual cues align with clinically relevant structures, enhancing, transparency and trust in the model’s predictions.

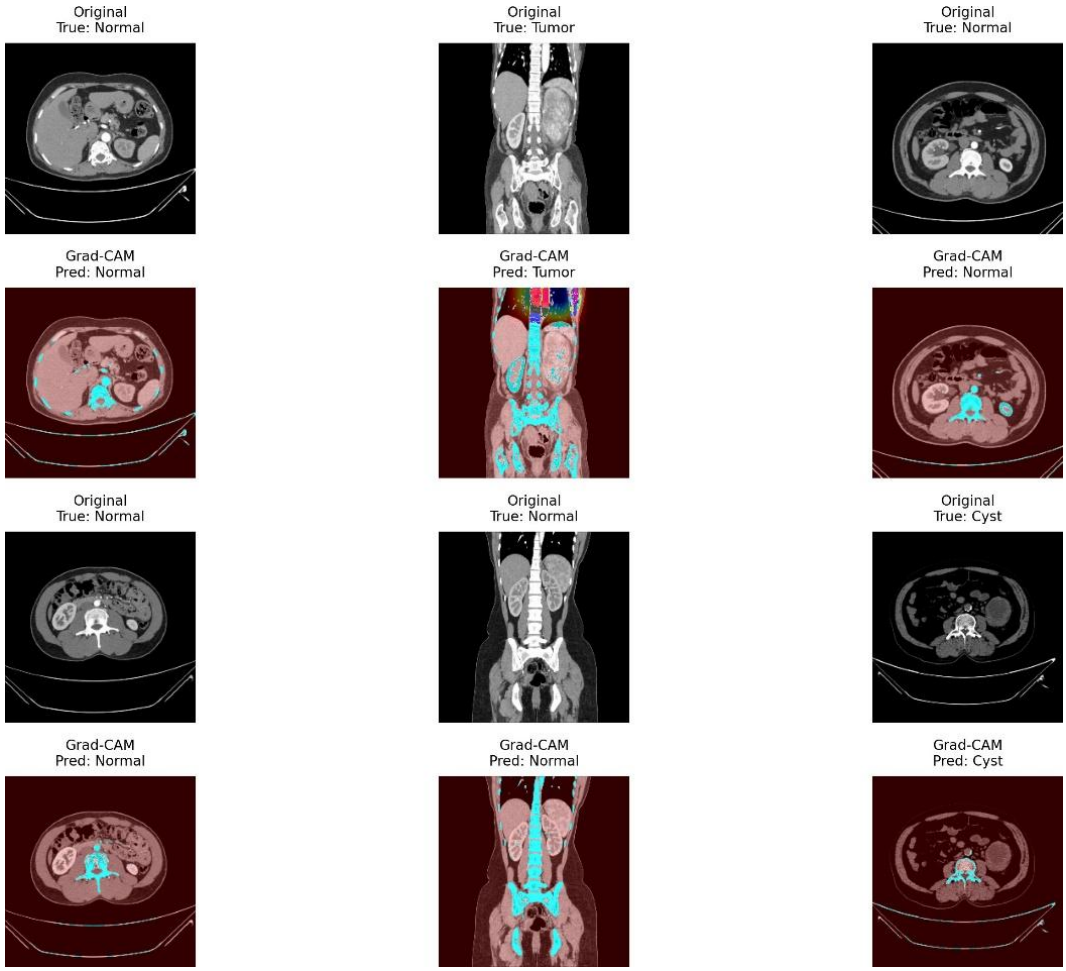

Figure 7. Grad-CAM visualization highlighting important regions influencing the model's decisions

Despite strong performance, certain challenges remain for real-world deployment. The dataset used was public and well-curated, but lacked external institutional variation. Without testing on CT images from different scanners or hospitals, generalizability to broader clinical settings cannot be guaranteed. Moreover, the model was trained and evaluated in a cloud GPU environment; inference efficiency on edge devices or hospital PACS systems needs further investigation to confirm feasibility in practice. Although benchmarking against expert radiologists was not within the scope of this study, the model achieved average inference times of under 30 ms per image on a standard cloud GPU (NVIDIA Tesla T4 via Google Colab). This suggests strong potential for real-time deployment. Future work will include latency pro-filing on edge hardware (e.g., Raspberry Pi, Jetson Nano) and comparisons with human diagnostic performance to evaluate clinical relevance and responsiveness under real-world constraints.

4.12. Discussion and limitations

The proposed hybrid model demonstrates strong diagnostic performance in classifying kidney conditions from CT images, achieving 99.88% validation accuracy and consistent results across a 6-fold cross-validation setup. By combining MobileNetV2 and ResNet50, the architecture leverages lightweight spatial features alongside deep semantic representations. This fusion proved more effective than either model alone. In addition to strong metrics, interpretability was enhanced using Grad-CAM visualizations, which highlighted clinically relevant regions and supported trust in the model's predictions. Statistical validation using McNemar's test further confirmed the significance of improvements over baseline models. Despite these strengths, several limitations remain, especially with respect to real-world deployment:

- Potential overfitting: despite strong 6-fold cross-validation, the model's very high accuracy may indicate overfitting to training-specific patterns.
- External generalization: since all data came from a single public source, generalization to other hospitals, scanners, or imaging protocols remains uncertain.

- Hardware and deployment: the model was trained on cloud GPUs; real-time use on hospital or mobile systems may face latency and memory constraints without optimization.
- Clinical integration: real-world deployment requires handling noisy CT data, automating slice selection, and integrating results into radiologists' workflows.
- Misclassification risks: occasional confusion between cysts, stones, and tumors highlights the need for improved robustness and uncertainty handling.
- External validation: further testing on multi-center datasets is needed to confirm the model's generalizability and clinical reliability.

4.13. Future work

To bridge the gap between research performance and clinical applicability, future work will focus on evaluating the model using external datasets from multiple hospitals and imaging systems. Further, real-time deployment tests will be conducted on edge devices and embedded systems, including model optimization through quantization or pruning. Augmenting the system with automated preprocessing pipelines (e.g., slice selection and noise removal) and uncertainty estimation methods will improve robustness and trust. Finally, clinical user studies involving radiologists are needed to assess diagnostic usefulness, usability, and interpretability under real-world constraints.

5. CONCLUSION

In this study, a hybrid deep learning model is presented that combines MobileNetV2 and ResNet50 for multiclass kidney disease classification from CT images. By integrating lightweight spatial and deep semantic features, the model achieved 99.88% validation accuracy with excellent precision, recall, F1-score, and AUC. It outperformed individual CNN models and showed robustness against overlapping cases such as cysts and stones. These results highlight the potential of hybrid architectures to improve diagnostic accuracy, efficiency, and generalization in medical imaging, particularly in resource-limited settings. Six-fold cross-validation confirmed consistent performance, though further validation on external datasets is needed. Overall, this work provides a practical and scalable AI-based diagnostic solution that can support radiologists, enhance early detection of renal diseases, and bridge the gap between research and clinical application.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Abdey Rabby	✓	✓	✓	✓		✓		✓	✓	✓	✓		✓	
Jannatun Naima Jannat	✓	✓		✓	✓		✓	✓	✓	✓			✓	
Md Assaduzzaman		✓	✓	✓	✓	✓		✓	✓		✓			
Rahmatul Kabir Rasel Sarker		✓		✓	✓	✓				✓		✓		
Raja Tariqul Hasan Tusher			✓			✓	✓			✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY

Data used in this study were obtained from Kaggle and are publicly available at [30]. No new data were created during this study.

REFERENCES

[1] J. A. Kellum, P. Romagnani, G. Ashuntantang, C. Ronco, A. Zarbock, and H. J. Anders, "Acute kidney injury," *Nature Reviews Disease Primers*, vol. 7, p. 52, 2021, doi: 10.1038/s41572-021-00284-z.

[2] J. A. Schaub, H. Hamidi, L. Subramanian, and M. Kretzler, "Systems biology and kidney disease," *Clinical Journal of the American Society of Nephrology*, vol. 15, pp. 695–703, 2020, doi: 10.2215/CJN.09990819.

[3] T. M. Kennedy-Lydon, C. Crawford, S. S. Wildman, and C. M. Peppiatt-Wildman, "Renal pericytes: regulators of medullary blood flow," *Acta Physiologica*, vol. 207, no. 2, pp. 212–225, 2013, doi: 10.1111/apha.12026.

[4] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," *The Lancet*, vol. 389, pp. 1238–1252, 2017, doi: 10.1016/S0140-6736(16)32064-5.

[5] A. Ujszaszi, M. Z. Molnar, M. E. Czira, M. Novak, and I. Mucsi, "Renal function is independently associated with red cell distribution width in kidney transplant recipients: a potential new auxiliary parameter for the clinical evaluation of patients with chronic kidney disease," *British Journal of Haematology*, vol. 161, no. 5, pp. 715–725, 2013, doi: 10.1111/bjh.12296.

[6] S. Turajlic, C. Swanton, and C. Boshoff, "Kidney cancer: the next decade," *Journal of Experimental Medicine*, vol. 215, no. 10, pp. 2477–2479, 2018, doi: 10.1084/jem.20181617.

[7] S. R. Khan *et al.*, "Kidney stones," *Nature Reviews Disease Primers*, vol. 2, p. 16008, 2016, doi: 10.1038/nrdp.2016.8.

[8] N. Heller *et al.*, "Kidney and tumor segmentation in CT imaging: Results of the KiTS19 challenge," *Medical Image Analysis*, vol. 67, p. 101821, 2021, doi: 10.1016/j.media.2020.101821.

[9] F. Zabihollahy, N. Schieda, S. Krishna, and E. Ukwatta, "Automated classification of solid renal masses on contrast-enhanced CT images using CNN with decision fusion," *European Radiology*, vol. 30, no. 9, pp. 5183–5190, 2020, doi: 10.1007/s00330-020-06787-9.

[10] X. Chen *et al.*, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, p. 102444, 2022, doi: 10.1016/j.media.2022.102444.

[11] S. G. Silverman *et al.*, "Bosniak classification of cystic renal masses, version 2019: an update proposal and needs assessment," *Radiology*, vol. 292, no. 2, pp. 475–488, 2019, doi: 10.1148/radiol.2019191293.

[12] N. Gillingham, H. Chandarana, A. Kamath, H. Shaish, and N. Hindman, "Bosniak IIF and III Renal Cysts: Can Apparent Diffusion Coefficient-Derived Texture Features Discriminate Between Malignant and Benign IIF and III Cysts?" *Journal of Computer Assisted Tomography*, vol. 43, no. 3, pp. 485–492, 2019, doi: 10.1097/RCT.0000000000000896.

[13] R. Suarez-Ibarrola, M. Basulto-Martinez, A. Heinze, C. Gratzke, and A. Miernik, "Radiomics applications in renal tumor assessment: A comprehensive review," *Cancers*, vol. 12, no. 6, p. 1387, 2020, doi: 10.3390/cancers12061387.

[14] J. Dana, V. Agnus, F. Ouhmich, and B. Gallix, "Multimodality imaging and artificial intelligence for tumor characterization," *Seminars in Nuclear Medicine*, vol. 50, no. 6, pp. 541-548, 2020, doi: 10.1053/j.semnuclmed.2020.02.005.

[15] F. Z. Ma, T. Sun, L. Y. Liu, and H. Y. 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, 2020, doi: 10.1016/j.future.2020.04.036.

[16] Y. He *et al.*, "Meta grayscale adaptive network for 3D integrated renal structures segmentation," *Medical Image Analysis*, vol. 71, p. 102055, 2021, doi: 10.1016/j.media.2021.102055.

[17] S. Pang *et al.*, "CTumorGAN: A unified framework for automatic CT tumor segmentation," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 47, pp. 2248–2268, 2020, doi: 10.1007/s00259-020-04781-3.

[18] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[19] M. Badawy, A. M. Almars, H. M. Balaha, M. Shehata, M. Qaraad, and M. Elhosse, "A two-stage renal disease classification based on transfer learning with hyperparameters optimization," *Frontiers in Medicine*, vol. 10, 2023, doi: 10.3389/fmed.2023.1106717.

[20] V. Vekaria, R. Gandhi, B. Chavarkar, H. Shah, C. Bhadane, and P. Chaudhari, "Identification of kidney disorders in decentralized healthcare systems through federated transfer learning," *Procedia Computer Science*, vol. 233, pp. 998–1010, 2024, doi: 10.1016/j.procs.2024.03.289.

[21] A. Shtaiyat and H. A. Younes, "Kidney segmentation using deep learning," *Nanotechnology Perceptions*, vol. 20, no. S5, 2024, doi: 10.62441/nano-ntp.v20is5.55.

[22] Q.-H. He *et al.*, "Deep learning system for malignancy risk prediction in cystic renal lesions: a multicenter study," *Insights into Imaging*, vol. 15, 2024, doi: 10.1186/s13244-024-01700-0.

[23] M. S. Farooq and A. Tariq, "Deep learning architectures for kidney disease classification," *arXiv preprint*, 2024, doi: 10.48550/arXiv.2403.15895.

[24] S. Fuladi, H. Chaturvedi, M. K. Nallakaruppan, V. Grover, H. Alshahrani, and M. Baza, "Efficient approach for kidney stone treatment using convolutional neural network," *Traitement du Signal*, vol. 42, no. 2, pp. 929–937, 2024, doi: 10.18280/ts.410233.

[25] S. Pande and R. Agarwal, "Multi-class kidney abnormalities detecting novel system through computed tomography," *IEEE Access*, vol. 12, pp. 21147-21155, 2024, doi: 10.1109/ACCESS.2024.3351181.

[26] 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, doi: 10.3390/app13179802.

[27] M. N. Islam, M. Hasan, Md. K. Hossain, Md. G. R. Alam, Md 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, 2022, doi: 10.1038/s41598-022-15634-4.

[28] M. Bhandari, P. Yogarajah, M. S. Kavitha, and J. Condell, "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, p. 3125, 2023, doi: 10.3390/app13053125.

[29] N. Sasikaladevi, S. Pradeepa, A. Revathi, S. Vimal, and R. G. Crespo, "Diagnosis of kidney cyst, tumor and stone from CT scan images using feature fusion hypergraph convolutional neural network (F2HCN2)," *International Journal for Multiscale Computational Engineering*, vol. 22, no. 5, pp. 35-46, 2024, doi: 10.1615/intjmultcompeng.2023048245.

[30] M. N. Islam and M. H. K. Mehedi, "CT kidney dataset: Normal, cyst, tumor, and stone," Kaggle, 2023, [Online]. Available: https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone/data. (Accessed: Jan 5, 2025).

[31] L. B. da Cruz *et al.*, "Kidney segmentation from computed tomography images using deep neural network," *Computers in Biology and Medicine*, vol. 123, 2020, doi: 10.1016/j.compbiomed.2020.103906.

[32] M. Revathi, R. Nithiya, S. O. Shankar, and D. Maheshwari, "Prediction of chronic kidney disease on CT images using deep learning: ResNet-34 and VGGNet-16," in *2024 International Conference on Integration of Emerging Technologies for the Digital World (ICIETDW)*, 2024, pp. 1–7, doi: 10.1109/ICIETDW61607.2024.10939622.

[33] M. Majid *et al.*, "Enhanced Transfer Learning Strategies for Effective Kidney Tumor Classification with CT Imaging," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, Jan. 2023, doi: 10.14569/ijacsa.2023.0140847.

BIOGRAPHIES OF AUTHORS

Abdey Rabby is a dedicated researcher and a third-year Honors student in the Department of Computer Science and Engineering at Daffodil International University, Bangladesh. He is currently working on a Bachelor's degree in Computer Science, with a strong focus on machine learning and deep learning applications. His research interests include developing innovative deep learning models for medical imaging. He has conducted extensive work in this area, aiming to contribute to advancements in healthcare through AI. He can be contacted at email: rabby23105101025@diu.edu.bd.

Jannatun Naima Jannat is a dedicated researcher and a student in the Department of Computer Science and Engineering at Daffodil International University, Bangladesh. He is currently pursuing a Bachelor's degree in Computer Science, with a strong interest in large-scale data analysis, system design, and business process modeling. His research initially focused on deep learning applications in kidney disease detection, including kidney cysts, stones, tumors, and normal predictions. Currently, he is exploring new research areas outside the medical sector, particularly those involving large datasets. She can be contacted at email: jannat2305101668@diu.edu.bd.

Md Assaduzzaman is a Senior Lecturer in the Department of Computer Science and Engineering at Daffodil International University, Dhaka, Bangladesh. He completed both his B.Sc. and M.Sc. in Computer Science and Engineering from the same university. He is actively engaged in research as a member of the Health Informatics Research Lab (HIRL), with a focus on machine learning, deep learning, explainable artificial intelligence (XAI), and federated learning. His work primarily addresses challenges in the healthcare domain, aiming to develop ethical and interpretable AI systems. He has published several research papers in peer-reviewed journals and international conferences. He can be contacted at email: assaduzzaman.cse@diu.edu.bd.

Rahmatul Kabir Rasel Sarker is a Senior Lecturer at Daffodil International University in Computer Science and Engineering. He received his B.Sc. in Computer Science and Engineering from Daffodil International University and completed his M.Sc. in Computer Science and Engineering from Jahangirnagar University. His research interests include machine learning, deep learning, and human–computer interaction. He can be contacted at email: raselsarker.cse@diu.edu.bd.

Raja Tariqul Hasan Tusher is an Assistant Professor in the Department of Computer Science and Engineering at Daffodil International University (Bangladesh). He earned both his B.Sc. and M.Sc. degrees in Computer Science and Engineering from Daffodil International University, and is currently pursuing a Ph.D. in CSE at Jagannath University, Dhaka. His research spans machine learning, biomedical engineering, convolutional neural networks, and antenna design. He can be contacted at email: tusher.cse@diu.edu.bd.