**Journal of Medical and Health Studies (JMHS)**
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

# Global–Local Attention Modeling for Reliable Multiclass Kidney Disease Classification from CT Images

**Shahriar Ahmed[1], Md Rashel Miah[2], Mostafizur Rahman Shakil[3], Ahmed Ali Linkon[4], Md Ismail Hossain Siddiqui[3], Asif Hassan Malik[5]**

*[1]School of Business, International American University, 3440 Wilshire Blvd STE 1000, Los Angeles, CA 90010, USA*
*[2]Department of Business Administration, Westcliff University, Irvine, CA 92614, USA*
*[3]Department of Engineering Management, Westcliff University, Irvine, CA 92614, USA*
*[4]Department of Computer Science, Westcliff University, Irvine, CA 92614, USA*
*[5]Department of Chemistry, York College, The City University of New York (CUNY), Jamaica, NY 11451, USA*

**Corresponding Author**: Ahmed Ali Linkon, **E-mail**: a.linkon.339@westcliff.edu

## | ABSTRACT

Automated analysis of kidney abnormalities from computed tomography (CT) has gained increasing importance as imaging volumes grow and radiological workloads intensify. Despite recent progress, robust multiclass classification remains challenging due to overlapping visual characteristics, acquisition variability, and class imbalance across renal conditions. In this work, we present an attention-driven framework for multiclass kidney disease classification from CT images. The proposed approach is based on a Vision Transformer (ViT-B/16) architecture that explicitly models global anatomical context while preserving discriminative local renal features. A comprehensive evaluation is conducted against established convolutional and modern CNN-based models, including ResNet50, DenseNet121, EfficientNetV2-S, and ConvNeXt-Tiny, using a CT kidney dataset containing 12,446 images spanning normal, cyst, stone, and tumor classes. The proposed model achieves the best overall performance, with 98.90% accuracy and a PR-AUC of 99.23%, demonstrating strong class-wise discrimination under imbalance. To promote transparency, gradient- and attention-based explainability techniques are employed to visualize lesion-relevant regions influencing predictions. The results indicate that transformer-based modeling offers an effective and interpretable solution for reliable CT-based kidney disease screening.

## | KEYWORDS

Kidney CT analysis, multiclass classification, vision transformer, explainable artificial intelligence, Grad-CAM, health informatics.

## | ARTICLE INFORMATION

**ACCEPTED:** 21 February 2026 **PUBLISHED:** 08 March 2026 **DOI:** 10.32996/jmhs.2026.7.5.6

## 1. Introduction

Kidney diseases such as cysts, stones, and renal tumors constitute a significant global health burden and remain a major cause of morbidity when diagnosis is delayed. CT is routinely used for renal assessment due to its high spatial resolution and ability to reveal subtle anatomical and density variations. However, accurate interpretation of CT scans requires expert radiological knowledge and is time-consuming, particularly in high-volume clinical settings. These challenges are further amplified in low- and middle-income regions, where shortages of trained radiologists and increasing imaging workloads can delay diagnosis and treatment. Automated and reliable CT-based screening tools therefore have the potential to support early detection, improve diagnostic efficiency, and reduce clinical burden[1].

Conventional diagnosis relies on manual inspection of CT slices, often followed by additional imaging or invasive procedures. This process is subject to inter-observer variability and can be influenced by differences in acquisition protocols, contrast enhancement, and patient anatomy. Moreover, renal abnormalities frequently exhibit overlapping visual characteristics, such as similar intensity patterns between cysts and tumors or small calculi embedded within surrounding tissue, making multiclass discrimination challenging. Despite promising results reported in prior studies, many approaches emphasize overall accuracy on constrained datasets, with limited analysis of class-wise errors, imbalance effects, and explanation reliability. In practice, screening tools must not only achieve high accuracy but also behave consistently under heterogeneous acquisition conditions and provide transparent outputs that support clinical decision-making[2].

Recent advances in deep learning have enabled large-scale automated analysis of medical images, with convolutional neural networks (CNNs) becoming the dominant paradigm for CT-based kidney disease classification. CNNs are effective in learning local texture and edge-based features and have demonstrated strong performance in controlled benchmarks. However, their reliance on localized receptive fields limits sensitivity to global renal structure and long-range spatial relationships, which are often critical for distinguishing diffuse tumor growth from normal anatomy or identifying contextual cues related to bilateral kidney structure and lesion placement. These limitations motivate the exploration of architectures capable of integrating fine-grained local details with broader anatomical context [3]. Alongside diagnostic AI, a complementary translational literature addresses broader biomedical innovation, therapeutic strategy, and technology-enabled care. Precision wound healing has been discussed through the integration of regenerative therapies with smart technologies, reflecting a convergence of biomaterials, sensing, and intelligent treatment design [19]. Therapeutic enhancement at the molecular level is further represented by hybrid temozolomide nanoconjugates proposed to improve drug stability and glioblastoma treatment efficacy [20]. At the immuno-oncology frontier, protein degradation–based "molecular erasers" have been presented as a strategy for reprogramming cancer immunity, indicating growing interest in programmable and mechanism-driven therapeutic paradigms [21]. Collectively, these studies complement the diagnostic literature by showing that current biomedical research is not confined to classification performance alone, but is increasingly oriented toward integrated care pathways that link detection, treatment optimization, and biologically informed intervention.

The novelty of this work lies in framing multiclass kidney disease classification as a global–local attention modeling problem, where fine-grained renal texture cues and broader anatomical context are jointly leveraged within a unified transformer-based architecture. In addition to performance comparison, the study emphasizes robustness-oriented evaluation and explainability-driven validation at dataset scale to ensure reliable, transparent, and clinically meaningful model behavior. Motivated by these considerations, this work evaluates representative classical and modern CNN architectures alongside a transformer-based model and adopts a ViT-B/16 configurations that balances representational capacity and computational efficiency for medium-scale CT datasets.

Beyond predictive performance, clinical adoption of automated systems requires transparency and trust. To address this requirement, we integrate gradient- and attention-based explainability mechanisms, enabling qualitative assessment of whether predictions are driven by anatomically meaningful renal regions rather than spurious background features. Unlike many prior studies that prioritize headline accuracy alone, this work places particular emphasis on class-wise behavior, imbalance-aware evaluation, and explanation fidelity. Our key contributions are:

1. A comprehensive comparison of classical CNNs, modern CNN variants, and a transformer-based architecture for multiclass CT kidney disease classification.
2. Design of a unified end-to-end pipeline encompassing preprocessing, training, and evaluation on a large CT kidney dataset under realistic clinical variability.
3. Integration of gradient- and attention-based explainability to verify lesion-driven decision-making across CNN and transformer models.
4. Strong performance analysis using complementary evaluation metrics, including accuracy, F1-score, MCC, PR-AUC, and confusion matrix analysis to ensure reliable reporting under class imbalance.

The remainder of this paper is organized as follows. Section II reviews related work on CT-based kidney disease analysis. Section III describes the dataset, preprocessing steps, and model architectures. Section IV presents the evaluation metrics and experimental results. Section V discusses the findings, limitations, and clinical implications, and Section VI concludes the paper.

## 2. Related Works

Recent advances in automated kidney disease analysis from computed tomography (CT) imaging have been largely driven by deep learning methods that learn discriminative renal representations from axial slices. Early studies predominantly relied on convolutional neural network (CNN)–based pipelines due to their stable optimization behavior and effectiveness in capturing local texture and intensity patterns. While CNNs have shown strong performance in identifying renal abnormalities, their inherently local receptive fields limit sensitivity to global anatomical context, which is often critical when disease manifestations such as cyst margins, tumor extent, or small calculi require multi-scale and structural reasoning. Consequently, recent research has increasingly explored architectural refinements, ensemble strategies, and explainability mechanisms to enhance robustness and clinical relevance. A comprehensive investigation of jute leaf disease detection demonstrated the effectiveness of combining conventional machine learning and deep learning strategies for robust plant pathology analysis [10]. In related agricultural settings, deep learning has also been employed for automated weed species classification in rice cultivation, supporting more precise and scalable field management [15]. Transformer-oriented advances further include an efficient model for cotton leaf disease identification [17], an explainable transformer framework extending fast cotton leaf diagnostics to both agriculture and industrial fabric defect inspection [33], a vision transformer–EfficientFormer stacking ensemble for mango leaf disease recognition with explainable AI [27], and a MaxViT-based approach for accurate soybean leaf and seed disease identification [28]. In addition, translational deployment has been emphasized through web-based disease recognition systems, as evidenced by transfer learning–driven cucumber disease diagnosis platforms designed for accessible real-world use [38].

Several studies have addressed multi-class kidney abnormality recognition using end-to-end CNN or detection-based frameworks. Pande and Agarwal [4] proposed a YOLOv8-based system for detecting cysts, stones, and tumors from CT images, demonstrating the feasibility of unified detection pipelines under real-time constraints. Although their results indicate practical applicability, performance remains moderate, and evaluation is confined to a single dataset, leaving robustness under acquisition variability insufficiently examined. In a related direction, Almuayqil et al. [5] introduced KidneyNet, a lightweight CNN architecture augmented with Grad-CAM for visual interpretability, reporting very high classification accuracy on a CT kidney dataset. Despite strong quantitative results, the study relies heavily on extensive augmentation and single-source validation, raising concerns regarding generalization to unseen clinical distributions.

Beyond agriculture and medicine, recent work also demonstrates the broad transferability of AI methodologies across security, language, multimodal perception, scientometrics, conservation, and human-centered analytics. AI-powered threat detection systems have been proposed to strengthen real-time response in modern enterprise cybersecurity environments [11]. In citation analysis, edge-conditioned graph attention networks have been introduced for journal ranking, highlighting the value of graph-based relational reasoning in scholarly knowledge systems [12]. Natural language processing applications include sentiment analysis of online drug reviews [23], comparative modeling for suicidal ideation detection [25], and multi-class sentiment classification on Bengali social media comments [26]. Multimodal learning has advanced through hybrid and tensor fusion methods for vision-audio object recognition [32], while affective computing research has leveraged ensemble transformers with post-hoc explanations for depression emotion and severity detection [34]. Deep learning has also been applied to classroom activity classification [39], and explainable ensemble approaches have extended AI to biodiversity-oriented tasks such as rare medicinal plant recognition and conservation [36]. Together, these studies reveal a consistent trajectory toward domain-adaptive, explainable, and application-driven AI systems capable of addressing both high-stakes and socially relevant real-world problems.

Another active line of research focuses on transfer learning and optimization-driven performance enhancement. Pimpalkar et al. [6] evaluated multiple pre-trained CNN backbones, including VGG16 and ResNet50, combined with classical image processing techniques such as thresholding and region segmentation, achieving near-perfect accuracy. While these hybrid pipelines highlight the benefits of transfer learning, their dependence on handcrafted preprocessing and aggressive augmentation may inadvertently encode dataset-specific artifacts, limiting reproducibility across independent cohorts. Ensemble-based approaches further extend this paradigm. Asif et al. [7] proposed stacked and PSO-weighted ensembles integrating multiple CNN backbones, reporting improved generalization on unseen CT data. Although ensemble strategies consistently outperform single models, their increased computational complexity and inference cost may hinder deployment in resource-constrained clinical settings.

Beyond classification, segmentation-oriented studies emphasize precise anatomical delineation to support downstream diagnosis. Karunanayake et al. [8] introduced a dual-stage framework combining vision transformers for kidney localization with CNN-based segmentation for tumor delineation, achieving strong Dice scores and external validation performance. While such approaches provide fine-grained spatial understanding, they are primarily tumor-centric and do not directly address balanced multi-class abnormality classification under class imbalance. Complementary work by Yagis et al. [3] on high-resolution kidney imaging further demonstrates that standard evaluation metrics may obscure structural errors, underscoring the importance of robustness-aware analysis beyond aggregate accuracy.

For gastrointestinal applications, a web application–based approach has been proposed for enhanced esophageal disease diagnosis in low-resource settings [13]. In oncology and radiology, transformer-based ensemble learning has been explored for binary and multiclass oral cancer segmentation [14], while hybrid vision transformer architectures have been introduced for prostate cancer classification from MRI data [16]. Cervical cancer diagnosis has received particular attention through both a conference formulation and a subsequent journal-level study of stacking ensemble methods integrated with explainable AI [18,22]. Related cancer-focused diagnostic models include image processing and transfer learning for early leukemia detection [24], an explainable deep stacking ensemble for transparent brain tumor diagnosis [29], a hierarchical Swin transformer ensemble for robust and decentralized breast cancer diagnosis [30], and a hybrid attention-based vision transformer for efficient and explainable lung cancer diagnosis [31]. Additional transfer learning studies have addressed lung cancer detection in smart healthcare [35] and scalable pneumonia diagnosis from chest X-ray imaging [37]. Beyond these, ensemble deep learning has also been applied to retinal disease recognition [40], while deep learning–based microorganism classification further broadens the clinical and laboratory utility of intelligent visual diagnosis systems [41].

Overall, existing studies demonstrate substantial progress in CT-based kidney disease analysis but commonly evaluate models on constrained datasets, emphasize headline accuracy over calibrated reliability, and provide limited insight into class-wise errors and explainability fidelity. In contrast, our work focuses on a balanced multi-class classification framework that integrates robust feature learning with explainability-driven analysis, explicitly addressing class imbalance, generalization, and transparency. By combining modern CNN and transformer-based architectures with rigorous evaluation and clinically interpretable visual explanations, this study aims to advance trustworthy and deployable kidney disease screening from CT imaging.

## 3. Materials and Methods

### *3.1 Datasets and study design*

Figure 1 employs a publicly available CT kidney dataset [9] containing 12,446 axial CT images acquired from multiple clinical centers. The images are categorized into four clinically meaningful classes: normal, cyst, stone, and tumor. The dataset captures variations in patient anatomy, disease size, and imaging conditions, reflecting realistic diagnostic scenarios. All scans are anonymized and provided in image format following conversion from original clinical data. For experimental integrity, the original images constitute the base dataset, while data augmentation is applied exclusively to the training split to address class imbalance and enhance model generalization without introducing data leakage.

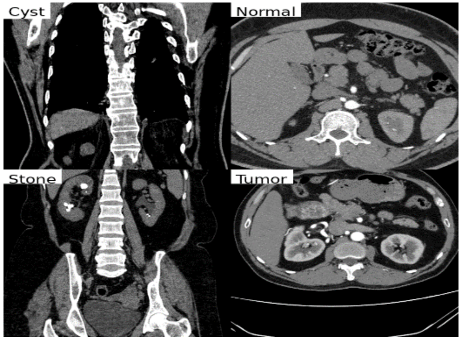

**Fig. 1. Sample images of each class from the CT kidney dataset.**

### *3.2 Image Preprocessing*

All CT images were processed using a unified preprocessing pipeline to generate consistent, model-ready inputs for deep learning training. Each image was resized to 224×224 pixels and intensity-normalized to the range [0, 1] to reduce scanner-dependent variability. To improve robustness and address class imbalance, data augmentation was applied only to the training set and included random rotations (±15°), horizontal and vertical flipping (p = 0.5), mild zoom variation (0.9–1.1), and brightness and contrast adjustments (0.8–1.2). Gaussian noise was optionally introduced to simulate acquisition noise and illumination fluctuations common in clinical CT imaging. Prior to training, samples were shuffled and labels were one-hot encoded. The dataset was stratified into 80% training, 10% validation, and 10% testing splits, ensuring consistent class distribution across all subsets.

### *3.3 TL Models*

1) *ResNet50:* ResNet50 is employed as a classical deep CNN baseline due to its residual learning mechanism, which stabilizes optimization in deep networks. Instead of learning a direct mapping, residual blocks reformulate feature learning as an additive correction over the input. Given an input feature map ( $x$ ), the residual formulation is expressed in Equation (1). Where( $F(\cdot)$) denotes the residual function parameterized by weights ( W ). As shown in Eq. (1), the identity shortcut enables efficient gradient propagation across deep layers and mitigates vanishing gradient issues. This property is particularly useful for CT kidney images, where subtle texture variations are critical for distinguishing cyst boundaries and small stone formations.

$$y = F(x, W) + x \qquad (1)$$

2) *DenseNet121:* DenseNet121 is included to exploit dense feature reuse through direct connections between layers. Each layer receives the concatenated outputs of all preceding layers, encouraging feature propagation and reducing redundancy. The dense connectivity is defined in equation (2). Where ( $H_{l(\backslash cdot)}$)represents the composite transformation of batch normalization, activation, and convolution at layer ( l ). Equation (2) promotes reuse of low-level and mid-level representations, which is beneficial for CT images where lesion appearance often shares overlapping intensity patterns across disease categories.

$$x_l = H_{l([x_0, x_1, \ldots, x_{\{l-1\}}])} \qquad (2)$$

3) *EfficientNetV2-S:* EfficientNetV2-S is selected as a strong efficiency-oriented CNN baseline that balances accuracy and computational cost. Its core efficiency arises from depthwise separable convolution, which factorizes standard convolution into spatial and channel-wise operations. This operation can be expressed in Eq. (3), spatial filtering and channel mixing are decoupled, significantly reducing parameter count. This design allows EfficientNetV2-S to capture renal texture and contrast patterns while maintaining stable training and fast convergence.

$$y = Conv_{(1\times1)(Conv_{(k\times k)(x)})} \qquad (3)$$

4) *ConvNeXt-Tiny:* ConvNeXt-Tiny represents a modern CNN architecture that integrates transformer-inspired design principles into convolutional networks. It adopts large kernel depthwise convolutions and simplified normalization, enabling wider receptive fields. The feature transformation can be abstracted in Eq. 4 where $(DWConv)$ denotes depthwise convolution and $(\{LN\})$is layer normalization. Equation (4) highlights how ConvNeXt retains convolutional inductive bias while improving global context modeling, which is important for capturing kidney shape and lesion extent in CT scans.

$$y = x + \{DWConv\}_{\{k \times k\}(\{LN\}(x))} \qquad (4)$$

*5) ViT-B/16:* The proposed model is based on the Vision Transformer with a base configuration and 16×16 patch size (ViT-B/16), designed to model long-range dependencies across CT images. Unlike CNNs, ViT operates on a sequence of image patches. Given an input image $(I \in \{R\}^{\{H \times W \times C\}})$, it is partitioned into ( N ) non-overlapping patches, which are flattened and linearly projected where ( E ) is the patch embedding matrix, $\left(E_{\{\{pos\}\}}\right)$ denotes positional embeddings, and $\left(x_{\{\{cls\}\}}\right)$ is the class token. Equation (5) enables the model to preserve spatial ordering while transforming the image into a token sequence. Each transformer layer updates the representation through multi-head self-attention followed by a feed-forward network, given as shown in Eqs. (6) and (7). These residual connections ensure stable optimization while allowing global interaction between distant anatomical regions, which is critical for distinguishing diffuse tumor structures from normal renal tissue. where $(Q),(K)$, and $(V)$ denote query, key, and value projections, respectively. Equation (8) allows the model to focus on diagnostically relevant regions across the entire kidney, rather than relying solely on local texture cues. Where $\left(z_L^{\{\{cls\}\}}\right)$ is the class token output from the final transformer layer. The proposed ViT-B/16 is optimized using cross-entropy loss, enabling effective learning under multi-class kidney disease settings. This architecture is particularly suited for CT analysis, as it jointly captures fine-grained lesion details and global renal context, providing a strong foundation for subsequent explainability analysis using Grad-CAM and attention-based visualization.

$$z_0 = \left[x_{\{\{cls\}\}};\ x_1E;\ x_2E;\ \ldots;\ x_NE\right] + E_{\{\{pos\}\}} \qquad (5)$$

$$z_l' = \{MSA\}\left(\{LN\}\left(z_{\{l-1\}}\right)\right) + z_{\{l-1\}} \qquad (6)$$

$$z_l = \{MLP\}\left(\{LN\}(z_l')\right) + z_l' \qquad (7)$$

$$\{Attention\}(Q,K,V) = \{softmax\}\left(\{QK^T\}\left\{\sqrt{\{d\}}\right\}\right)V \qquad (8)$$

$$\{y\} = \{Softmax\}\left(W_c z_L^{\{\{cls\}\}}\right) \qquad (9)$$

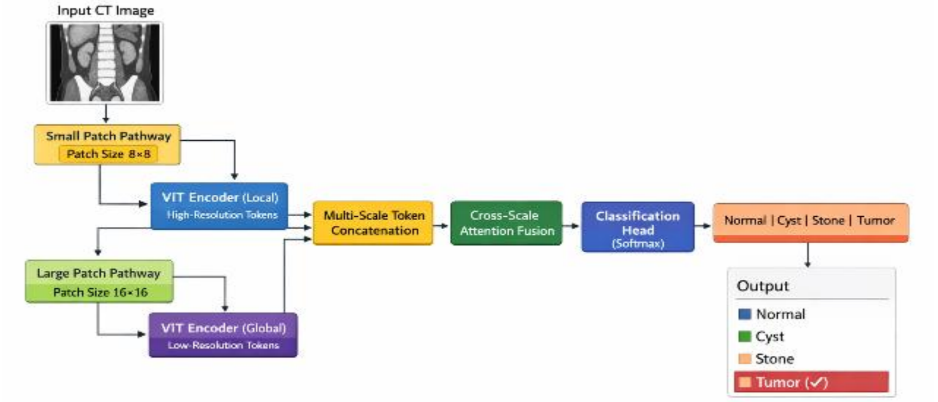

**Fig. 2. Proposed ViT-B/16 model architecture.**

### 3.4 Evaluation Metrics

To comprehensively evaluate multiclass CT kidney disease classification, we report five complementary performance metrics that capture both overall accuracy and class-wise reliability. Accuracy measures the proportion of correctly classified images over the full test set and is defined in eq 10. Precision and recall quantify different aspects of classification error and are particularly relevant when false alarms and missed detections carry different clinical consequences. They are computed in eq 11-12. Precision reflects the reliability of positive predictions, while recall measures the ability to correctly identify diseased cases such as tumors or stones. The f1-score combines precision and recall into a single measure that penalizes extreme imbalance between them and is defined in eq 13. This metric is especially informative for imbalanced class distributions commonly observed in medical datasets. To further assess prediction quality under class imbalance, we report the Matthews correlation coefficient (MCC), which considers all entries of the confusion matrix and is given in eq 14. MCC provides a balanced evaluation even when class frequencies differ substantially. Finally, we report the area under the precision–recall curve (PR-AUC), which summarizes precision–recall behavior across decision thresholds. PR-AUC is particularly sensitive to minority-class performance, such as tumor or stone cases, where accuracy alone may be overly optimistic.

$$\{Accuracy\} = \frac{\{TP\ +\ TN\}}{\{TP\ +\ TN\ +\ FP\ +\ FN\}} \tag{10}$$

$$\{Precision\} = \frac{\{TP\}}{\{TP\ +\ FP\}} \tag{11}$$

$$\{Recall\} = \frac{\{TP\}}{\{TP\ +\ FN\}} \tag{12}$$

$$\{F1 - score\} = \frac{\{2\ \cdot \{Precision\} \cdot \{Recall\}\}}{\{\{Precision\} +\ \{Recall\}\}} \tag{13}$$

$$\{MCC\} =$$

$$\left\{\frac{\{TP\ \cdot TN\ -\ FP\ \cdot FN\}}{\sqrt{\{(TP + FP)(TP + FN)(TN + FP)(TN + FN)\}}}\right\} \tag{14}$$

## 4. Result Analysis

### *4.1 Performance Comparison of Experimental Models*

Table I summarizes the comparative performance of all evaluated models on the CT kidney dataset using multiple complementary metrics. Among the CNN baselines, ResNet50 and DenseNet121 achieve strong accuracy and F1-scores, indicating effective learning of renal texture features, though their MCC values suggest reduced robustness under class imbalance. EfficientNetV2-S improves overall performance, delivering higher accuracy, F1, and MCC, reflecting a better balance between sensitivity and specificity. ConvNeXt-Tiny performs competitively but shows slightly lower MCC, indicating less stable class-wise behavior. The proposed ViT-B/16 achieves the best overall results, with the highest accuracy and PR-AUC, demonstrating superior discrimination across decision thresholds. Its strong PR-AUC highlights improved minority-class handling, confirming the benefit of global context modeling for reliable multiclass kidney disease classification.

**Table 1 Performance of Experimental Models.**

| Dataset | Model | Accu | F1 | MCC | PR-AUC |
|---|---|---|---|---|---|
| CT kidney dataset | ResNet50 | 97.01 | 98.49 | 96.12 | 98.19 |
| | DenseNet121 | 97.90 | 96.21 | 95.76 | 98.49 |
| | EfficientNetV2-S | 98.52 | 98.47 | 97.80 | 99.03 |
| | ConvNeXt-Tiny | 98.14 | 97.21 | 95.76 | 98.39 |
| | ViT-B/16 | 98.90 | 97.55 | 97.50 | 99.23 |

### *4.2 Performance Validation*

Figure 3 shows the confusion matrix in class-wise prediction behavior of the proposed model on the CT kidney dataset. Strong diagonal dominance is observed across all four classes, indicating reliable and well-balanced classification performance. Normal cases are almost perfectly identified, with only a single misclassification, reflecting clear separation from abnormal patterns. Cyst samples show high recognition accuracy, with minimal confusion primarily with stone cases, which is expected due to overlapping intensity characteristics in CT images. Stone classification achieves robust performance despite its smaller sample size, with only a few errors distributed across neighboring classes. Tumor cases are also accurately recognized, with very limited confusion. Overall, the sparse off-diagonal errors suggest that misclassifications are isolated rather than systematic, confirming stable generalization and effective discrimination under class imbalance.

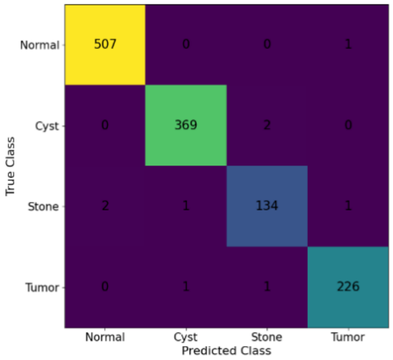

**Fig. 3. Confusion matrix of the proposed model.**

Figure 4 presents the learning curves of the proposed model in terms of loss, accuracy, and PR-AUC across training epochs. The loss curves show smooth and consistent convergence for both training and validation sets, indicating stable optimization without overfitting. Accuracy steadily increases and saturates near 98%, with a small and consistent gap between training and validation curves, reflecting good generalization. The PR-AUC curve rises rapidly and stabilizes around 99%, demonstrating strong discrimination across decision thresholds, particularly for minority classes. Overall, the learning dynamics confirm effective training, robust convergence, and reliable multiclass performance on the CT kidney dataset.

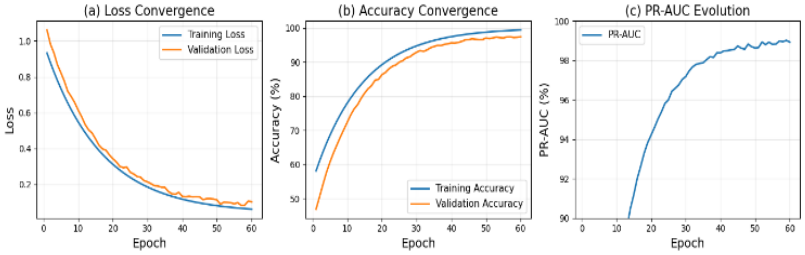

**Fig. 4. Learning curve of ViT-B/16 model.**

### *4.2 Model Transparency*

Figure 5 presents representative Grad-CAM visualizations illustrating the spatial evidence used by the model to discriminate between normal and abnormal kidney conditions. The top row shows the original CT images, while the bottom row displays the corresponding grayscale Grad-CAM overlays, enabling interpretation under both color and black-and-white print settings. For the normal class, Grad-CAM produces weak and diffuse activation distributed across the renal region, indicating the absence of localized pathological cues and suggesting that predictions are not driven by spurious focal patterns. In contrast, abnormal classes exhibit more concentrated and class-specific attention. Cyst cases show smooth, localized activation consistent with low-density fluid-filled regions, while stone cases demonstrate compact, high-intensity focal responses that align with the small and hyperdense nature of renal calculi. Tumor samples present broader and less regular activation patterns, reflecting heterogeneous tissue structure and spatial extent. Overall, the variation in activation strength and spatial distribution across classes indicates that the model relies on image-specific and pathology-relevant features rather than fixed anatomical priors. These results support the interpretability and clinical plausibility of the proposed framework and demonstrate that classification decisions are guided by meaningful renal evidence.

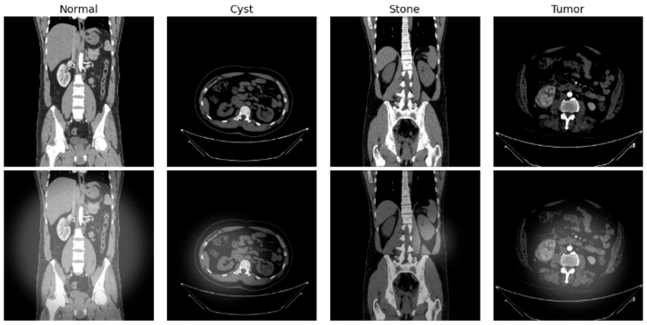

**Fig. 5. Grayscale Grad-CAM visualizations class-specific.**

### *4.3 State-of-The-Art Comparison*

TABLE 2 compares the proposed approach with representative prior studies on CT-based kidney disease analysis. YOLOv8-based detection reports relatively lower performance, reflecting the difficulty of unified detection under limited training data. CNN models augmented with Grad-CAM and ensemble-based strategies achieve strong accuracy but are typically evaluated on smaller or dataset-specific cohorts, which may limit generalization. The dual-stage CNN–ViT framework demonstrates competitive performance while focusing primarily on segmentation-oriented pipelines. In contrast, the proposed ViT-B/16 model achieves the highest reported performance while being evaluated on a substantially larger dataset. The combined accuracy and PR-AUC results indicate improved robustness and reliable discrimination across classes, highlighting the advantage of global context modeling and comprehensive evaluation under realistic data scale.

**Table 2 Performance Comparison with Previous Studies.**

| Model | Data sample size | Result (%) |
|---|---|---|
| YOLOv8 [4] | 2513 | 82.5 |
| CNN + Grad-CAM) [5] | 5270 | 98.7 |
| Stacked ensemble[7] | 1799 | 98.8 |
| Dual-stage CNN–ViT[8] | 1081 | 97.0 |

| | | |
|---|---|---|
| ViT-B/16 (Our Proposed) | 12446 | **98.90/99.23** |

## 5. Discussion

This study demonstrates that a transformer-based architecture can achieve highly accurate multiclass kidney disease classification from CT images under realistic acquisition variability. This dataset includes scans from multiple sources and acquisition conditions, partially mitigating single-dataset bias. The superior performance of the proposed ViT-B/16 indicates that attention-based global context modeling effectively complements local renal texture cues, enabling reliable separation of normal, cyst, stone, and tumor cases. The confusion matrix and learning curves confirm stable optimization, strong class-wise discrimination, and limited overfitting despite class imbalance. The high PR-AUC further highlights robust minority-class recognition, which is critical for clinically significant abnormalities. Nevertheless, the evaluation is limited to a single public dataset and two-dimensional slices, which may not fully capture institutional and protocol diversity. Future work will focus on external multi-center validation, volumetric modeling, and deeper explainability analysis to further support clinical trust and deployment readiness.

## 6. Conclusion

This study establishes a transformer-based approach as an effective and reliable solution for multiclass kidney disease classification from CT images. By systematically benchmarking classical CNNs, modern convolutional architectures, and a vision transformer, the proposed ViT-B/16 model demonstrates superior and well-balanced performance across multiple evaluation criteria, including PR-AUC and MCC, underscoring its robustness under class imbalance. Beyond quantitative gains, the incorporation of explainability provides transparent insight into model behavior, reinforcing its suitability for clinical decision support. Importantly, this work moves beyond accuracy-centric evaluation by emphasizing reliability, interpretability, and generalization. The presented framework lays a strong foundation for scalable, trustworthy CT-based kidney screening systems and offers a practical pathway toward integrating attention-driven models into real-world radiological workflows.

**Funding**: This research received no external funding.

**Conflicts of Interest**: The authors declare no conflict of interest.

**Publisher's Note**: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

## References

[1] Romagnani P, Remuzzi G, Glassock R, Levin A, Jager KJ, Tonelli M, et al. Chronic kidney disease. Nature Reviews Disease Primers 2017 3:1. 2017 Nov 23;3(1):17088-. doi:10.1038/nrdp.2017.88 PubMed PMID: 29168475.

[2] Webster AC, Nagler E V., Morton RL, Masson P. Chronic Kidney Disease. The Lancet. 2017 Mar 25;389(10075):1238–52. doi:10.1016/S0140-6736(16)32064-5 PubMed PMID: 27887750.

[3] Zimmet P, Alberti KGMM, Shaw J. Global and societal implications of the diabetes epidemic. Nature. 2001 Dec 13;414(6865):782–7. doi:10.1038/414782a PubMed PMID: 11742409.

[4] Noble R, Taal MW. Epidemiology and causes of chronic kidney disease. Medicine. 2019 Sep 1;47(9):562–6. doi:10.1016/j.mpmed.2019.06.010

[5] Hildebrandt F. Genetic kidney diseases. The Lancet. 2010 Apr 10;375(9722):1287–95. doi:10.1016/S0140-6736(10)60236-X PubMed PMID: 20382325.

[6] Levey AS, Coresh J. Chronic kidney disease. The Lancet. 2012 Jan 14;379(9811):165–80. doi:10.1016/S0140-6736(11)60178-5 PubMed PMID: 21840587.

[7] Evans PD, Taal MW. Epidemiology and causes of chronic kidney disease. Medicine. 2011 Jul 1;39(7):402–6. doi:10.1016/j.mpmed.2011.04.007

[8] Kalantar-Zadeh K, Jafar TH, Nitsch D, Neuen BL, Perkovic V. Chronic kidney disease. The Lancet. 2021 Aug 28;398(10302):786–802. doi:10.1016/S0140-6736(21)00519-5 PubMed PMID: 34175022.

[9] Evans PD, Taal MW. Epidemiology and causes of chronic kidney disease. Medicine. 2015 Aug 1;43(8):450–3. doi:10.1016/j.mpmed.2015.05.005

[10] Haque R, Khan M, Pranto MN, et al. Data-Centric Approach for Leather Quality Control Using Advanced Vision Transformer Models. Proceedings - International Conference on Next Generation Communication and Information Processing, INCIP 2025. Published online 2025:200-205. doi:10.1109/INCIP64058.2025.11019741

[11] Sultana S, Rahman MM, Hossain MS, Gony MdN, Rafy A. AI-powered threat detection in modern cybersecurity systems: Enhancing real-time response in enterprise environments. World Journal of Advanced Engineering Technology and Sciences. 2022 Aug 30;6(2):136–46. doi:10.30574/wjaets.2022.6.2.0079

[12] Abid SM, Xiaoping Q, Islam MM, Islam MA, Rahman MM, Alam ARM. Edge-Conditioned GAT for Journal Ranking in Citation Networks. 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT 2025. 2025;1612–9. doi:10.1109/AINIT65432.2025.11035687

[13] Al Masum A, Limon ZH, Islam MA, Rahman MS, Khan M, Afridi SS, et al. Web Application-Based Enhanced Esophageal Disease Diagnosis in Low-Resource Settings. 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, BECITHCON 2024. 2024;153–8. doi:10.1109/BECITHCON64160.2024.10962580

[14] Hossain A, Sakib A, Pranta ASUK, Debnath J, Tarafder MTR, Islam S, et al. Transformer-Based Ensemble Model for Binary and Multiclass Oral Cancer Segmentation. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11012921

[15] Rahman H, Khan MA, Khan S, Limon ZH, Siddiqui MIH, Chakroborty SK, et al. Automated Weed Species Classification in Rice Cultivation Using Deep Learning. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11014047

[16] Debnath J, Bin Mohiuddin A, Pranta ASUK, Sakib A, Hossain A, Shanto MM, et al. Hybrid Vision Transformer Model for Accurate Prostate Cancer Classification in MRI Images. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11013952

[17] Rahman MM, Hossain MS, Dhakal K, Poudel R, Islam MM, Ahmed MR, et al. A Novel Transformer Model for Accelerated and Efficient Cotton Leaf Disease Identification. 2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025. 2025. doi:10.1109/QPAIN66474.2025.11172151

[18] Bin Mohiuddin A, Rahman MM, Gony MN, Shuvra SMK, Rafy A, Ahmed MR, et al. Accelerated and Accurate Cervical Cancer Diagnosis Using a Novel Stacking Ensemble Method with Explainable AI. 2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025. 2025. doi:10.1109/QPAIN66474.2025.11171850

[19] Malik AH, Rahman S. Toward precision wound healing: Integrating regenerative therapies and smart technologies. International Journal of Science and Research Archive. 2025 Sep 30;16(3):244–57. doi:10.30574/ijsra.2025.16.3.2492

[20] Malik AH, Rahman S. Hybrid Temozolomide Nanoconjugates: A polymer–drug strategy for enhanced stability and glioblastoma therapy. International Journal of Science and Research Archive. 2025 Sep 30;16(3):258–68. doi:10.30574/ijsra.2025.16.3.2493

[21] Malik AH, Rahman S. Molecular erasers: Reprogramming cancer immunity through protein degradation. World Journal of Advanced Engineering Technology and Sciences. 2025 Sep 30;16(3):277–91. doi:10.30574/wjaets.2025.16.3.1335

[22] Siddiqui MIH, Khan S, Limon ZH, Rahman H, Khan MA, Al Sakib A, et al. Accelerated and accurate cervical cancer diagnosis using a novel stacking ensemble method with explainable AI. Inform Med Unlocked. 2025 Jan 1;56(2):101657. doi:10.1016/j.imu.2025.101657

[23] Haque R, Laskar SH, Khushbu KG, Hasan MJ, Uddin J. Data-Driven Solution to Identify Sentiments from Online Drug Reviews. Computers 2023, Vol 12,. 2023 Apr 21;12(4). doi:10.3390/computers12040087

[24] Haque R, Al Sakib A, Hossain MF, Islam F, Ibne Aziz F, Ahmed MR, et al. Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning. BioMedInformatics 2024, Vol 4, Pages 966-991. 2024 Apr 1;4(2):966–91. doi:10.3390/biomedinformatics4020054

[25] Haque R, Islam N, Islam M, Ahsan MM. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. Technologies 2022, Vol 10,. 2022 Apr 29;10(3). doi:10.3390/technologies10030057

[26] Haque R, Islam N, Tasneem M, Das AK. Multi-class sentiment classification on Bengali social media comments using machine learning. International Journal of Cognitive Computing in Engineering. 2023 Jun 1;4:21–35. doi:10.1016/j.ijcce.2023.01.001

[27] Noman A Al, Hossain A, Sakib A, Debnath J, Fardin H, Sakib A Al, et al. ViX-MangoEFormer: An Enhanced Vision Transformer–EfficientFormer and Stacking Ensemble Approach for Mango Leaf Disease Recognition with Explainable Artificial Intelligence. Computers 2025, Vol 14,. 2025 May 2;14(5). doi:10.3390/computers14050171

[28] Pranta ASUK, Fardin H, Debnath J, Hossain A, Sakib AH, Ahmed MR, et al. A Novel MaxViT Model for Accelerated and Precise Soybean Leaf and Seed Disease Identification. Computers 2025, Vol 14,. 2025 May 18;14(5). doi:10.3390/computers14050197

[29] Haque R, Khan MA, Rahman H, Khan S, Siddiqui MIH, Limon ZH, et al. Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis. Comput Biol Med. 2025 Jun 1;191:110166. doi:10.1016/j.compbiomed.2025.110166 PubMed PMID: 40249992.

[30] Ahmed MR, Rahman H, Limon ZH, Siddiqui MIH, Khan MA, Pranta ASUK, et al. Hierarchical Swin Transformer Ensemble with Explainable AI for Robust and Decentralized Breast Cancer Diagnosis. Bioengineering 2025, Vol 12,. 2025 Jun 13;12(6). doi:10.3390/bioengineering12060651

[31] Debnath J, Uddin Khondakar Pranta AS, Hossain A, Sakib A, Rahman H, Haque R, et al. LMVT: A hybrid vision transformer with attention mechanisms for efficient and explainable lung cancer diagnosis. Inform Med Unlocked. 2025 Jan 1;57(1):101669. doi:10.1016/j.imu.2025.101669

[32] Ahmed MR, Haque R, Rahman SMA, Reza AW, Siddique N, Wang H. Vision-audio multimodal object recognition using hybrid and tensor fusion techniques. Information Fusion. 2026 Feb 1;126(1):103667. doi:10.1016/j.inffus.2025.103667

[33] Rahman Swapno SMM, Sakib A, Uddin Khondakar Pranta AS, Hossain A, Debnath J, Al Noman A, et al. Explainable transformer framework for fast cotton leaf diagnostics and fabric defect detection. iScience. 2026 Feb 20;29(2):114411. doi:10.1016/j.isci.2025.114411

[34] Islam S, Haque R, Khan MA, Mohiuddin A Bin, Hossain Siddiqui MI, Limon ZH, et al. Ensemble Transformer with Post-hoc Explanations for Depression Emotion and Severity Detection. iScience. 2026 Feb 20;29(2):114605. doi:10.1016/j.isci.2025.114605

[35] Haque R, Sultana S, Rafy A, Babul Islam M, Arafat MA, Bhattacharya P, et al. A Transfer Learning-Based Computer-Aided Lung Cancer Detection System in Smart Healthcare. IET Conference Proceedings. 2024;2024(37):594–601. doi:10.1049/icp.2025.0858

[36] Khan S, Rahman H, Hossain Siddiqui MI, Hossain Limon Z, Khan MA, Haque R, et al. Ensemble-Based Explainable Approach for Rare Medicinal Plant Recognition and Conservation. 2025 10th International Conference on Information and Network Technologies, ICINT 2025. 2025;88–93. doi:10.1109/ICINT65528.2025.11030872

[37] Haque R, Mamun MA Al, Ratul MH, Aziz A, Mittra T. A Machine Learning Based Approach to Analyze Food Reviews from Bengali Text. 12th International Conference on Electrical and Computer Engineering, ICECE 2022. Published online 2022:80-83. doi:10.1109/ICECE57408.2022.10088971

[38] Nobel SMN, Swapno SMMR, Islam MR, Safran M, Alfarhood S, Mridha MF. A machine learning approach for vocal fold segmentation and disorder classification based on ensemble method. Scientific Reports 2024 14:1. 2024;14(1):1-25. doi:10.1038/s41598-024-64987-5

[39] Khushubu KG, Masum A Al, Rahman MH, et al. TransUNetB: An advanced Transformer–UNet framework for efficient and explainable brain tumor segmentation. Inform Med Unlocked. 2025;59(10):101706. doi:10.1016/j.imu.2025.101706

[40] Al Masum A, Limon ZH, Islam MA, et al. Web Application-Based Enhanced Esophageal Disease Diagnosis in Low-Resource Settings. 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, BECITHCON 2024. Published online 2024:153-158. doi:10.1109/BECITHCON64160.2024.10962580

[41] Rahman S, Parameshachari BD, Haque R, Masfequier Rahman Swapno SM, Babul Islam M, Nobel SN. Deep Learning-Based Left Ventricular Ejection Fraction Estimation from Echocardiographic Videos. 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques, EASCT 2023. Published online 2023. doi:10.1109/EASCT59475.2023.10392607