INSPIRE INTELLIGENCE

Inspire Intelligence Journal
ISSN: 3110-942X, Vol. 1, No. 2, 2026
doi.org/10.65718/inspireAI.2026.1008

InspireQuill.org
Inspire Academic Press (Canada)

Research Paper OPEN ACCESS

Multi-Phase CNN-ViT-Wavelet Fusion with Attention for Robust Kidney Stone Detection from Ultrasound Images

Faizan Ahmad [1], Uzair Ishtiaq [2, *], Malik M. Ali Shahid [3]

[1] Department of Computer Science, Superior University, Lahore, 547700, Pakistan
[2] Department of Artificial Intelligence, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[3] Department of Computer Science, Namal University, Mianwali, Pakistan
[*] Correspondence: uzair@um.edu.my

Abstract

Accurate detection of kidney stones in ultrasound images is hindered by low contrast, speckle noise, and operator-dependent variability. The objective of this study is to develop a robust multi-phase model that integrates the local, global, and frequency-domain features to enable a reliable and generalizable kidney stone classification. Three convolutional neural networks (ResNet50, DenseNet121, and EfficientNet-B0) and vision transformer variants (ViT-Base, Swin-Transformer, and DeiT-Small) were independently trained on a primary renal ultrasound dataset. The dataset was split patient-wise in a 70-15-15 manner; extensive data augmentation, normalization, and five-fold cross-validation were implemented to avoid data leakage and ensure the model's robustness. The best-performing CNN and ViT features were fused at the feature level with an attention mechanism and classified using ML classifiers, among which XGBoost demonstrated the optimal performance. Next, a discrete wavelet transform (DWT) branch was incorporated to acquire complementary frequency information for further enhancing the discriminative capability. The multi-phase framework achieved 97.9%, 97.8%, and 0.997 for accuracy, F1-score, and AUC, respectively, on the internal dataset. Similarly, it obtained 94.3%, 94.0%, and 0.970 for accuracy, F1-score, and AUC on the external renal ultrasound dataset. These results demonstrate a robust generalization backed up by Grad-CAM and attention maps. The Multi-Phase Framework (MPF) offers a consistent, generalizable, and fully automated method for kidney stone detection in ultrasound images, supporting improved diagnostic performance.

Keywords: Kidney Stone Detection; Ultrasound Imaging; Convolutional Neural Network (CNN); Vision Transformer (ViT); Wavelet Transform; Attention Mechanism; Feature Fusion; External Validation

1. Introduction

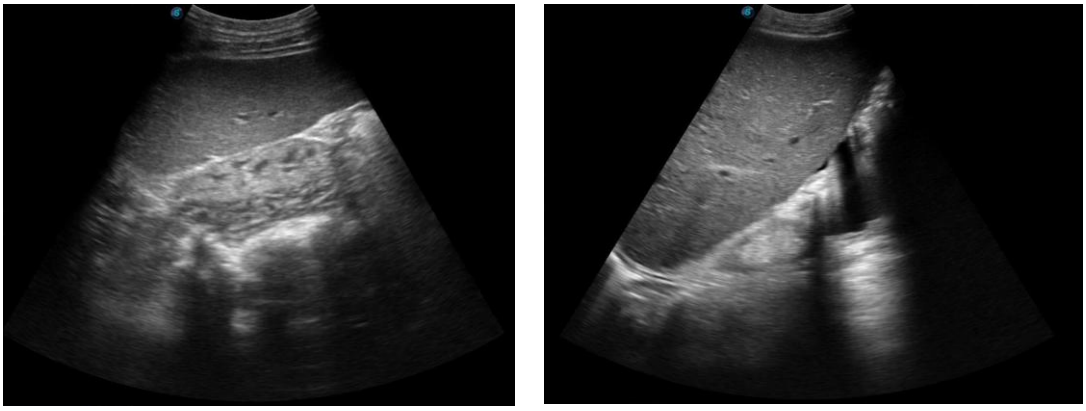
Kidney stones are one of the most common urological disorders all over the world, impacting millions of individuals and bringing serious health issues because of painful hardness, urinary tract blockage, persistent infection, and permanent kidney damage[1, 2]. Kidney stones are a medical issue of great concern, not only to

Academic Editor: Ghazanfar Latif

Received: 19/12/2025
Revised: 24/02/2026
Accepted: 10/03/2026
Published: 19/03/2026

Citation
Ahmad, F., Ishtiaq, U., & Shahid, M. M. A. (2026). Multi-phase CNN–ViT–wavelet fusion with attention for robust kidney stone detection from ultrasound images. *Inspire Intelligence Journal*, 1(2), 85–105.

disrupt the life of patients but also to incur great healthcare expenses. To avert severe complications, there is a strong need to identify the kidney stones early and accurately, with proper guidance on treatment strategies, and reduce hospitalization rates[2-4]. Ultrasound imaging is widely used as the primary diagnostic modality for detecting kidney stones because it is a noninvasive, accessible, economical, and radiation-free procedure[5-7]. Ultrasound can continuously monitor the urinary tract in real time and is applicable in such situations where the procedure needs to be repeated, especially in pediatric and pregnant patients. Although these are the benefits, but there are number of sensitivities of ultrasound imaging, such as speckle noise, low image contrast, operator dependence, and overlapping anatomical structures, which may complicate the precise detection of kidney stones [8, 9]. Such limitations drive the development of automated computer-aided diagnosis (CAD) systems that leverage the recent advances in artificial intelligence and deep learning to enhance the detection performance and reliability [10-13]. Historically, traditional medical image processing methods, including thresholding, edge detection, and morphological intervention, have been used to detect kidney stones[14-16]. These methods are aimed at per-pixel analysis and handcrafted feature extraction to localize the stone areas. Although working well in controlled conditions, traditional methods tend to fail in clinical practice, where stones can be tiny, partially obscured, or even located in low-contrast areas. With the introduction of deep learning (especially convolutional neural networks (CNNs)), medical image analysis has undergone a revolution, enabling automated extraction of hierarchical features directly from raw imaging data [6, 7, 17]. CNNs can also extract local spatial patterns, textures, and shapes, which are important for distinguishing kidney stones from adjacent tissue. They can form complex feature representations that enable them to classify well even under noisy, heterogeneous ultrasound conditions [18, 19].



(a) **(b)**

Figure 1. Representative kidney ultrasound images; **(a)** Kidney presenting with nephrolithiasis, characterized by a hyperechoic focus accompanied by posterior acoustic shadowing; **(b)** The normal kidney exhibits a regular structure with distinct differentiation between the cortex and medulla.

The incorporation of CNNs with complementary architecture has been studied recently to further improve their performance. Integrated CNN-LSTM networks utilized the chronological information in imaging data, obtaining patterns that could not be observable in a single frame [13]. Correspondingly, CNN-SVM frameworks combined the deep feature extraction with traditional machine learning algorithms, utilizing the support vector machine generalization to elevate the robustness[20]. The Discrete Wavelet Transform (DWT) has also been combined into CNNs to obtain the frequency-domain features of kidney stones, offering the additional discriminatory information and enhancing the feature richness [21, 22]. Even more recently, Vision Transformers (ViTs) have become formidable global feature extractors, complementing the local feature-capture abilities of CNNs[23, 24]. Multi-

branch hybrid architectures that combine CNNs and ViTs have achieved better performance by leveraging the advantages of local and global feature representations [25]. Attention-based approaches further enhanced such hybrid frameworks by rapidly emphasizing clinically important structures by reducing the noise and irrelevant structures, particularly in low-quality ultrasound images[26, 27]. Ensemble learning with multiple CNNs or CNN-ViT models has been demonstrated to be more robust, less prone to variance, and less prone to overfitting[28-31]. Recent work has improved the results by using these types of methods and kidney stone datasets with impressive performances between 95%-98%, with external validation studies indicating that they can generalize well around 94-95% [32-34]. Ablation studies show that a combination of frequency-domain features, attention modules, and multi-branch architectures not only enhances performance but also stabilizes learning across cross-validation folds [21, 35, 36]. Even with these achievements, several issues remain, like many previous studies using small, homogeneous, or institution-based datasets, which limit the generalizability of the trained models [32, 36]. Although CNNs are good at capturing local spatial features, but they can miss global contextual details that ViTs handle. Such multi-modal or hybrid structures, including CNNs, ViTs, wavelet transforms, attention modules, and classical machine learning classifiers, are currently being explored to realize state-of-the-art performance[36-38]. Development of an efficient kidney stone detection pipeline with emphasis on careful preprocessing, patient-wise data division, augmentation approaches, and external validation to guarantee clinical usability [21, 32, 39].

This paper presents a three-phase framework of automated kidney stone detection from ultrasound images. The first phase involved the systematic analysis of various CNN and ViT architectures to have a reference of their initial performance and to select the top-performing CNN and ViT for further development [26, 36, 40-42]. The second phase involves selecting and extracting features from top-performing CNN and ViT classifiers and enhancing them with attention mechanisms to highlight the most important features and maximize feature fusion to enhance the classification performance[37, 43]. The third phase adds another wavelet branch that captures the frequency-domain properties of kidney stones. This branch is connected to CNN and ViT feature representation and processed using classical machine learning classifiers to achieve the final classification results[21, 36, 41]. The proposed framework is evaluated on internal and external datasets, with extensive ablation studies conducted to determine the importance of each component [21, 32, 36, 39]. To achieve the highest accuracy and minimize overfitting by systematically combining spatial, frequency, and global contextual information, showing that hybrid deep learning architectures can offer high performance in kidney stone detection in clinical settings[44-47].

2. Materials and Methods

2.1. Dataset Description

Two publicly accessible kidney ultrasound datasets were employed to train and evaluate the Multi-Phase Framework (MPF). The internal kidney ultrasound images (stone/No Stone) are comprised of 9,416 grayscale images: 5,002 describing the kidney stones and 4,414 depicting the normal kidney samples. Images were acquired from multiple clinical ultrasound machines during the regular diagnostic process, unveiling a natural diversity in depth, gain, probe type, and field of view. Patient demographic data, including age, sex, and comorbidities, were not available. To prevent data leakage and promote generalizability, the dataset was split patient-wise in a 70-15-15 manner, and extensive data augmentation on the training set, normalization, and five-fold cross-validation were implemented.

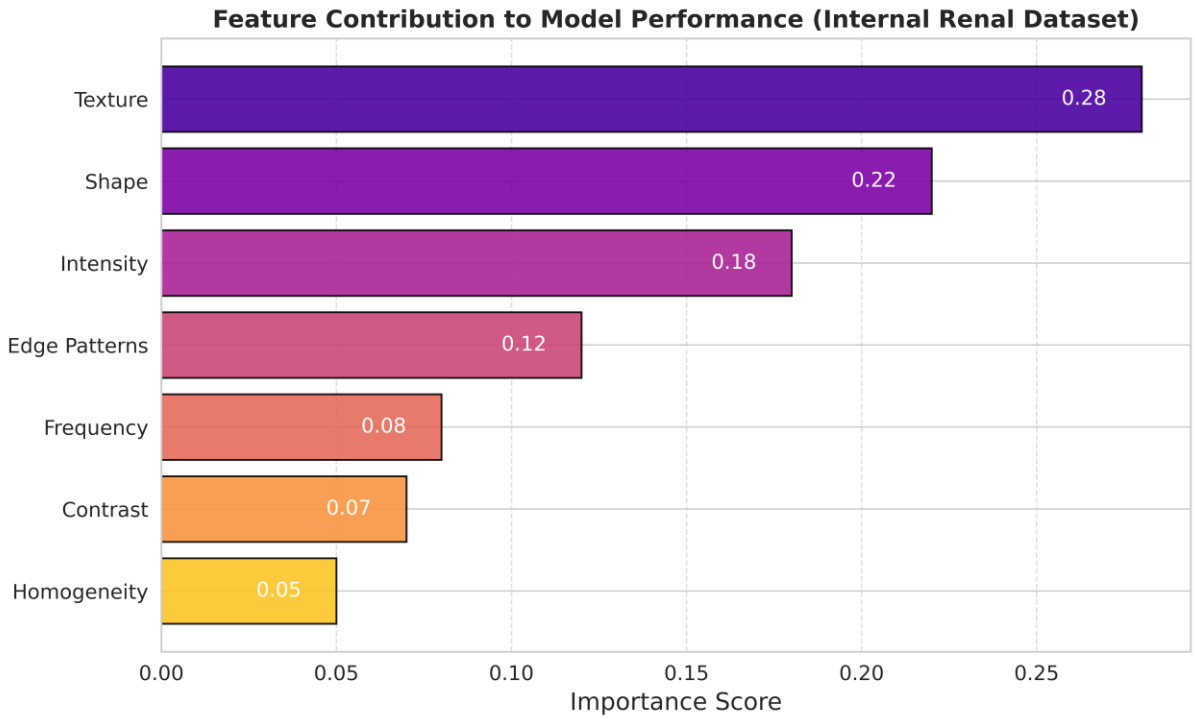


Figure 2. Contribution of internal Renal Ultrasound dataset features to model performance.

The external dataset comprised 134 grayscale images categorized into stones and normal classes and was used for testing to assess the generalizability of MPF. Both datasets contained images in .jpg and .png formats with varying resolutions, which were normalized during preprocessing to ensure uniform input across all models. A summary of both the internal and external renal datasets is provided in Table 1.

Table 1. Internal and External Renal Ultrasound Dataset Summary

Dataset	Class	Total Images	Selected	Excluded
Internal	Normal	4414	4414	0
	Stone	5002	5002	0
External	Normal	67	67	0
	Stone	67	67	0

2.1.1. Preprocessing and Augmentation

All images were resized to 224×224×3 pixels, and median filtering was applied to reduce speckle noise while pixel volumes were normalized to the range [0,1] using the following method:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

The internal dataset depicts a little class imbalance, with the Stone class comprising more samples than the Normal class. Data augmentation was applied only on the training set through different augmentation techniques like rotation, horizontal and vertical flipping, zooming, shifting, brightness, etc., as shown in Table 2 below. The

external dataset endured only normalization, and its original labels (0 and 1) were mapped to 0 as Normal and 1 as Stone to keep uniformity with the internal dataset.

Table 2. Number of images per class before and after augmentation for the internal Renal ultrasound dataset.

Class	Original Samples	Training (70%)	Validation (15%)	Testing (15%)	Augmentation ×3
Normal (No Stone)	4414	3089	662	663	9267
Stone	5002	3501	750	751	10503
Total	9416	6590	1412	1414	19770

2.1.2. Baseline Models Evolution

To establish benchmark performance, three convolutional neural network (CNN) architectures and three Vision Transformer (ViT) models were trained independently. Convolutional Neural Networks (CNNs) were primarily developed to capture the local spatial features, including edges, textures, and the boundaries of objects. In contrast, Vision Transformers (ViTs) aimed to derive global contextual information by processing images in segments. The effectiveness of the models were evaluated using key metrics such as accuracy, precision, recall, F1-score, and standard deviation (SD) as defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision+Recall)} \quad (5)$$

$$SD = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(M_k - M')^2} \quad (6)$$

Here,M_k denotes the metric (such as accuracy) in fold k, M' represents the mean metric across K = 5 folds, and TP, TN, FP, and FN indicate true positives, true negatives, false positives, and false negatives, respectively.

Best-performing CNN and ViT models were selected from baseline performance for feature extraction and merged using an attention mechanism to obtain supplementary information, therefore integrating the local and global representations while fused vector was classified using XGBoost and Random Forest (RF) algorithms to generate the final results.

2.1.3. Multi-Phase Feature Extraction

A multi-branch architecture was established to extract comprehensive information from ultrasound images, building on the best-performing CNN-ViT fusion method. Three parallel branches: a CNN branch for local texture and structural features, a ViT branch for global contextual features, and a wavelet transformation branch to collect frequency-domain information that aids in the identification of kidney structures that are not easily visible in the spatial domain.

$$F_{fused} = \sum_{i=1}^{B} a_i F_i, \quad a_i = \frac{e^{\wedge}(w_i)^T F_i}{\sum_{j=1}^{B} e^{\wedge}(w_i)^T F_i} \tag{7}$$

In this context, F_i denotes the feature vector from branch i, w_i represents the learnable attention weights, and a_i indicates the importance of each branch. The fused feature vector is then passed to dense layer followed by dropout to dimensionality reduction and then input fed into classical machine learning classifiers, specifically XGBoost and Random Forest, for final classification. Attention maps are generated and visualized to confirm that the model focuses on clinically relevant kidney-stone regions, thereby enhancing interpretability.

2.1.4. Training and Implementation

TensorFlow with additional support from NumPy, OpenCV, Albumentations, and scikit-learn and NVIDIA GPUs were used to implement the framework in Python. The Adam optimizer, 50–200 training epochs, batch sizes of 16–32, and a learning rate of 1×10^{-4} were utilized, and early stopping with a patience of 10 was used to reduce the overfitting, while binary cross-entropy served as the loss function. Details of the training summary of all baseline, ensemble and MPF models are shown in Table 3.

Table 3. Experimental Configuration and Training Parameters of all Models.

Parameter	**Configuration**
Optimizer	Adam
Learning Rate	1e-4(reduced by factor 0.1)
Batch Size	32
Epochs	200
Loss Function	Categorical Cross-Entropy
Data Augmentation	Random rotation (horizontal/vertical flip, zoom), color jitter
Early Stopping	Enabled (based on validation loss)
Hardware	NVIDIA RTX 3090
Framework	Python, TensorfFlow
RAM	128GB
Initialization	Pre-trained ImageNet weights

2.1.5. Cross-Validation and Ablation Study

A 5-fold cross-validation was implemented on the internal dataset to evaluate robustness and stability. Ablation studies were conducted to understand the contribution of each component, namely, CNN combined with ViT, wavelet, attention, dense layer, and classifiers while the external dataset was useful for testing the generalization performance on unseen data.

2.1.6. Evaluation Metrics

Performance was evaluated using the metrics defined in Equations (2) to (6), with standard deviation (SD) included to indicate variability across folds. These metrics were calculated for both the fused CNN and ViT features prior to the multi-branch stage, as well as for the final MPF.

2.1.7. Interpretability and Visualization

Grad-CAM and global attention maps were used to visualize the most contributing regions of the model's predictions. These visualizations demonstrated that the model concentrated on clinically relevant kidney-stone regions, thereby enhancing interpretability and supporting confidence in its potential clinical application.

3. Results

3.1.1. Baseline Models Evolution

During this phase, individual convolutional neural network (CNN) and Vision Transformer (ViT) models were systematically evaluated using the internal renal ultrasound dataset. The dataset was partitioned by patient-wise into 70% for training, 15% for validation, and 15% for testing to prevent data leakage and ensure reliable performance assessment as shown in Table 4 below. Among all the base CNN models, EfficientNet-B0 recorded the best performance of 91.8% accuracy, a 91.0% F1 score, and an AUC score of 0.978, narrowly outperforming than ResNet50 and DenseNet121, indicating its strong ability to learn the local textures, edges, and structural features of kidney tissue, which are critical for kidney stone detection. On the other hand, Swin-Transformer outperformed among all ViTs with 93.7% accuracy, 93.2% F1-score, and the highest AUC of 0.985. This demonstrates its ability to effectively manage the long-range dependencies found in the kidney region. The findings validate that Convolutional Neural Networks (CNNs) are highly proficient at identifying local features while Vision Transformers (ViTs) are highly proficient in utilizing global contextual information. The results highlighted that CNNs are strong at extracting local features and ViTs are good at utilizing global contextual information.

Table 4. Accuracy, F1-score, and AUC of baseline CNN and ViT models on the renal ultrasound dataset.

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
ResNet50 (Baseline)	91.2	90.6	90.1	90.3	0.952
DenseNet121 (Baseline)	91.5	90.5	90.3	90.4	0.960
EfficientNet-B0 (Baseline)	91.8	91.5	90.5	91.0	0.978
ViT-Base (Baseline)	93.5	93.0	92.8	92.9	0.972
Swin-Transformer	93.7	93.3	93.1	93.2	0.985
DeiT-Small	92.7	92.1	91.9	92.0	0.965

3.1.2. Confusion Matrix Analysis of Baseline Models

The performance of three baseline Convolutional Neural Networks (ResNet50, DenseNet121, and EfficientNet-B0) and three Vision Transformer (ViT-Base, Swin-Transformer, and DeiT-Small) models on the internal renal ultrasound dataset was evaluated using confusion matrices as shown in Figure 3.

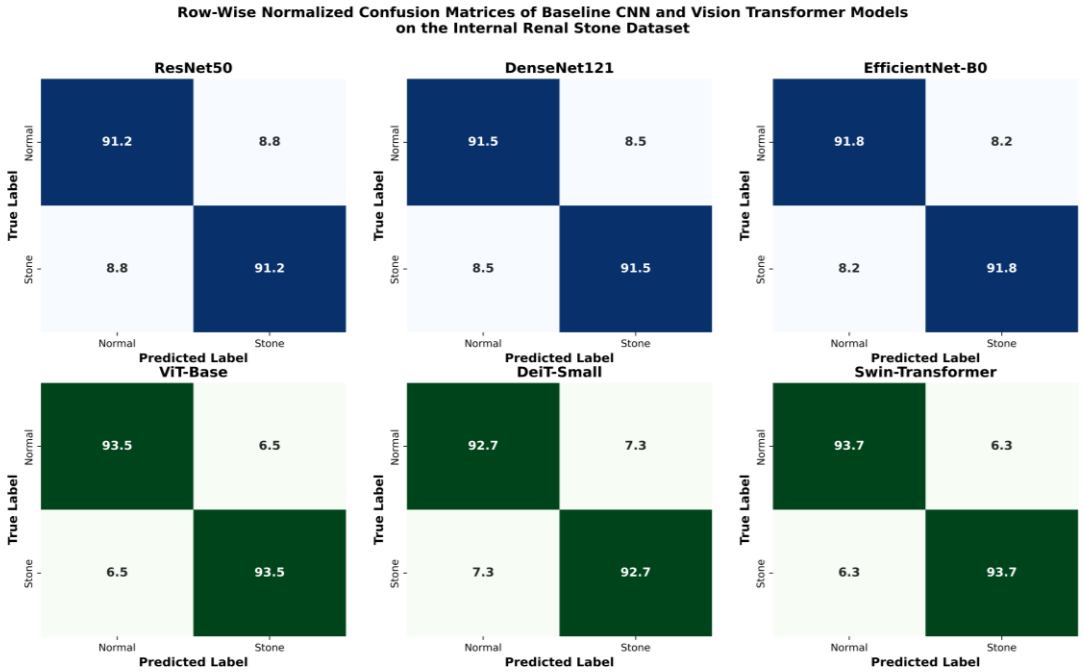


Figure 3. Confusion matrices of baseline CNN and ViT models for Stone vs. Normal classification, showing class-wise correct and incorrect predictions.

The dataset comprised labeled ultrasound images categorized into two classes: Stone and Normal, enabling a binary classification task. All models were trained using standardized deep learning pipelines with distinct training, validation, and test sets, and their performance was measured on the held-out test set, while confusion matrices visually summarize the model performance. Diagonal elements indicate the proportion of correctly classified instances for each class, while off-diagonal elements represent misclassifications. For instance, ViT-based models highlight the superior predictive performance compared to CNN variations and these matrices alleviate the elaborate analysis of class-specific performance. All models demonstrated strong performance, similar to that of CNNs, which achieved accuracies between 91.2% and 92.0%, while Vision Transformer (ViT) variants generally outperformed CNNs, with accuracy from 92.7% to 93.8%. Notably, the Swin-Transformer model achieved the highest performance, reaching 93.8%, which indicates that the Swin Transformer offers the most effective feature representation for ultrasound images. On average, ViT models increased performance by about 2 to 3% as compared to CNN baselines, which highlights the strong ability of transformer-based architecture to capture global context and structural patterns in medical imaging. In medical imaging, confusion matrices are important because they quantify the overall predictive accuracy and further identify potential sources of error, such as false positives and false negatives, which are critical for clinical decision-making. Comparisons of various deep learning architectures on the renal ultrasound dataset are done in both numerical and graphical confusion matrices. The present analysis helps in model selection and further optimization to support the development of reliable automated kidney stone detection systems.

3.1.3. ROC Curve Analysis of Baseline Models

Figure 4 below shows Receiver Operating Characteristic curves for six baseline deep learning models that were evaluated on the internal renal ultrasound dataset. The left subplot depicts the ROC curves of CNN variants (ResNet50, DenseNet121, and EfficientNet-B0) to exhibit their discriminative capacity between Stone and Normal

classes, while the central subplot presents the Receiver Operating Characteristic (ROC) curves for Vision Transformer models, including ViT-Base, Swin-Transformer, and DeiT-Small.

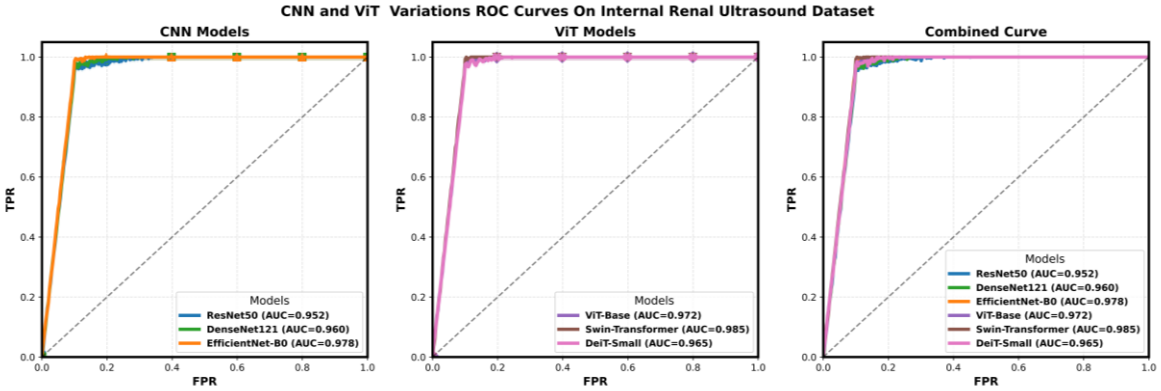


Figure 4.ROC curves of baseline CNN and ViT models on the internal renal ultrasound dataset, showing classification performance and AUC values for Stone vs. Normal cases.

These curves illustrate the correlation between the true positive rate and the false positive rate across a range of classification thresholds. Point markers depict the thresholds, while slight prediction variations are implicit by semi-transparent bands. The right subplot combined all CNNs and ViTs together, providing a way to directly compare local- and global-feature-based architectures. The Area Under the Curve (AUC) measures systematically assess the discriminative abilities of each model, and elevated AUC values indicate enhanced classification performance. The findings indicate that ViT-based models are more efficient at identifying global contextual features, whereas CNNs exhibit superior effectiveness at identifying local structural features.

3.2. Ensemble Model (CNN–ViT) Fusion Classification

In Phase 2, a comparative performance analysis was conducted among the top-performing convolutional neural network (CNN) model, EfficientNet-B0, the leading Vision Transformer (ViT), Swin-Transformer, and their ensemble. The aim was to assess the utility of feature-level fusion for kidney stone detection from ultrasound images. As a strong CNN baseline, EfficientNet-B0 could achieve 91.8% accuracy and a 91.0% F1-score, while its major advantage is the ability to capture local texture patterns and fine spatial details that are primary for distinguishing the tiny and faint stone structures in ultrasound images. Conversely, the convolutional architectures are naturally limited in modeling long-range contextual dependencies that constrain their overall discriminative power. The Swin Transformer, however, demonstrated impressive performance, achieving a 93.7% accuracy and a 93.2% F1-score, thereby emphasizing the advantages of hierarchical self-attention mechanisms in capturing global anatomical context and long-range relationships. The Swin-Transformer's superior capacity to differentiate between stone and normal cases is further evidenced by the AUC value of 98.5, particularly within the ensemble model that integrates features from EfficientNet-B0 and the Swin-Transformer, which resulted in the most significant performance improvements.

Consequently, the ensemble model, which combined local texture features with global contextual representations, attained an accuracy of 95.3%, an F1-score of 94.8%, and an AUC of 99.1, as illustrated in the bar chart presented in Figure 5. These findings indicate that the integration of complementary features from CNN and ViT architectures enhance a more resilient and discerning feature space.

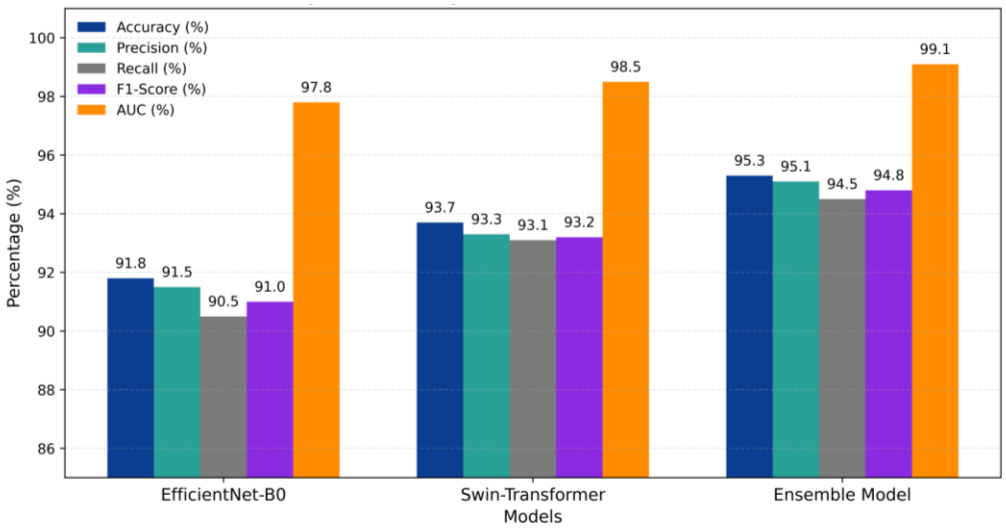


Figure 5. Comparison of Accuracy, Precision, Recall, F1-Score, and AUC for top-performing CNN (EfficientNet-B0), ViT (Swin-Transformer), and their Ensemble model on the internal renal ultrasound dataset.

3.2.1. Confusion Matrix of Ensemble Model

The confusion matrices in Figure 6 below present the row-wise normalized classification performance of the ensemble model, which integrates EfficientNet-B0 and Swin Transformer with feature fusion and attention, using XGBoost and Random Forest classifiers on the internal renal ultrasound dataset. The model correctly classifies 95.0% and 96.3% of normal and stone cases, respectively, having negligible misclassification rates of 5.0% and 3.7% false positives and false negatives, respectively. This stable performance reflects the robust sensitivity and specificity, enabling the performance of XGBoost in utilizing fused CNN–ViT features. In contrast, the Random Forest ensemble also obtained strong results, with 95.4% normal and 95.1% stone cases, which shows its strong performance.

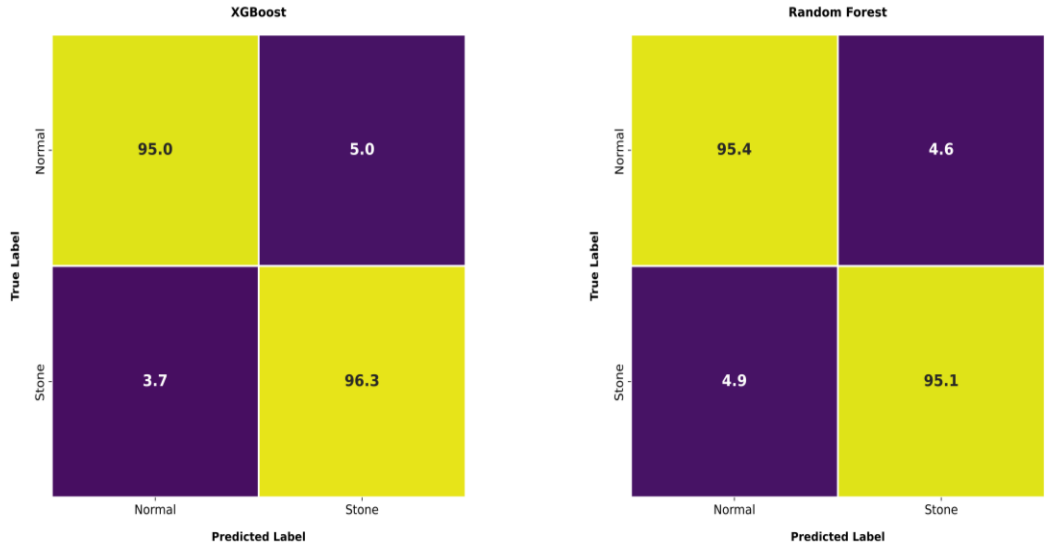


Figure 6. Confusion matrix of the Ensemble Model (CNN+ViT) on the internal renal ultrasound dataset.

3.2.2. Ensemble Model Ablation Analysis

Ablation study in Table 5 below highlights the impacts of feature fusion and attention mechanisms in the ensemble model employing XGBoost and Random Forest algorithms. Without fusion and attention, the classifiers show decreased performance, proposing narrow discriminative capacity when depending solely on isolated deep features. Adding the feature fusion produces significant improvements across all evaluation metrics, highlighting that combining the CNN and Vision Transformer representations enables complementary feature learning. Optimal outcomes are achieved when both fusion and attention mechanisms are utilized while Random Forest exhibits an accuracy of 94.9 ± 0.14%, an F1-score of 94.2 ± 0.16%, and an AUC of 0.983; in contrast, XGBoost attains an accuracy of 95.3 ± 0.11%, an F1-score of 94.8 ± 0.13%, and an AUC of 0.991. Moreover, the attention mechanism augments the integrated features by highlighting diagnostically relevant regions, thus enhancing the precision-recall balance and diminishing misclassification rates. These findings highlight the significance of amalgamating fusion and attention to enhance group performance. XGBoost demonstrates marginally enhanced balance and generalization in kidney stone identification using ultrasound images.

Table 5. Ablation study of XGBoost and RF ensembles showing Accuracy ±STD, F1 ±STD, and AUC.

Classifier	**Fusion**	**Attention**	**Accuracy(%) ±Std(%)**	**F1-Score(%) ±Std(%)**	**AUC**
XGBoost	YES	NO	95.1 ±0.13	94.2 ±0.11	0.989
	YES	YES	95.3 ±0.11	94.8 ±0.13	0.991
RF	YES	NO	94.6±0.16	93.3 ±0.18	0.980
	YES	YES	94.9 ±0.14	94.2 ±0.16	0.983

3.3. *Proposed Multi-Phase Framework*

Figure 6 below compares the performance of the MPF, which integrates EfficientNet-B0, Swin Transformer, and the Discrete Wavelet Transform (DWT), with an ensemble model comprising EfficientNet-B0 and Swin Transformer, both evaluated on an internal kidney stone ultrasound dataset. The training and validation accuracy over 200 epochs in the left panel, with the MPF, achieved a better and stable training validation accuracy of 0.98, while the ensemble model had a stable training validation accuracy of 0.93. Results confirmed that adding a DWT branch with CNN-ViT provides better feature extraction and overall classification performance. The MPF reflects faster convergence and keeps lower loss over the training as against the ensemble model, which indicates the better learning stability and generalization, while checkpoints indicate the specific epochs at which model weights were saved, further supporting the reliable superiority of the MPF. Overall results demonstrate that adding the DWT branch with EfficientNet-B0 and the Swin Transformer provides much better performance than the baseline and ensemble model.

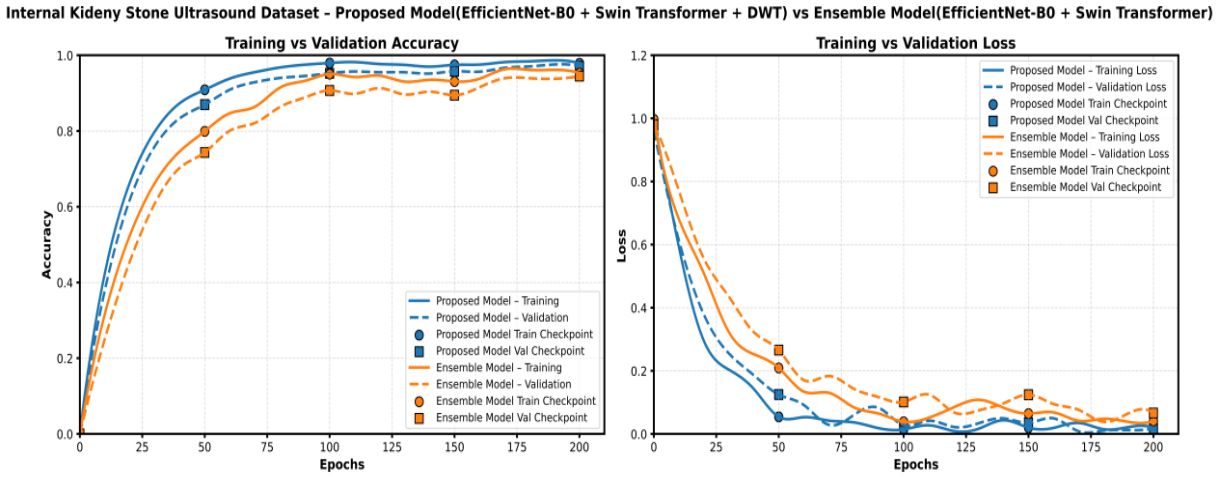


Figure 7. Training and validation loss and accuracy of the proposed and baseline ensemble models on internal Renal Ultrasound dataset.

The bar chart in Figure 8 presents the performance metrics of MPF on both internal and external renal ultrasound datasets. Four key metrics, accuracy, precision, recall, and F1-score, were evaluated for both internal and external datasets, where the internal dataset uniformly yields moderately better results, which indicates strong learning, whereas the external dataset results exhibit the model's robust generalization capability.

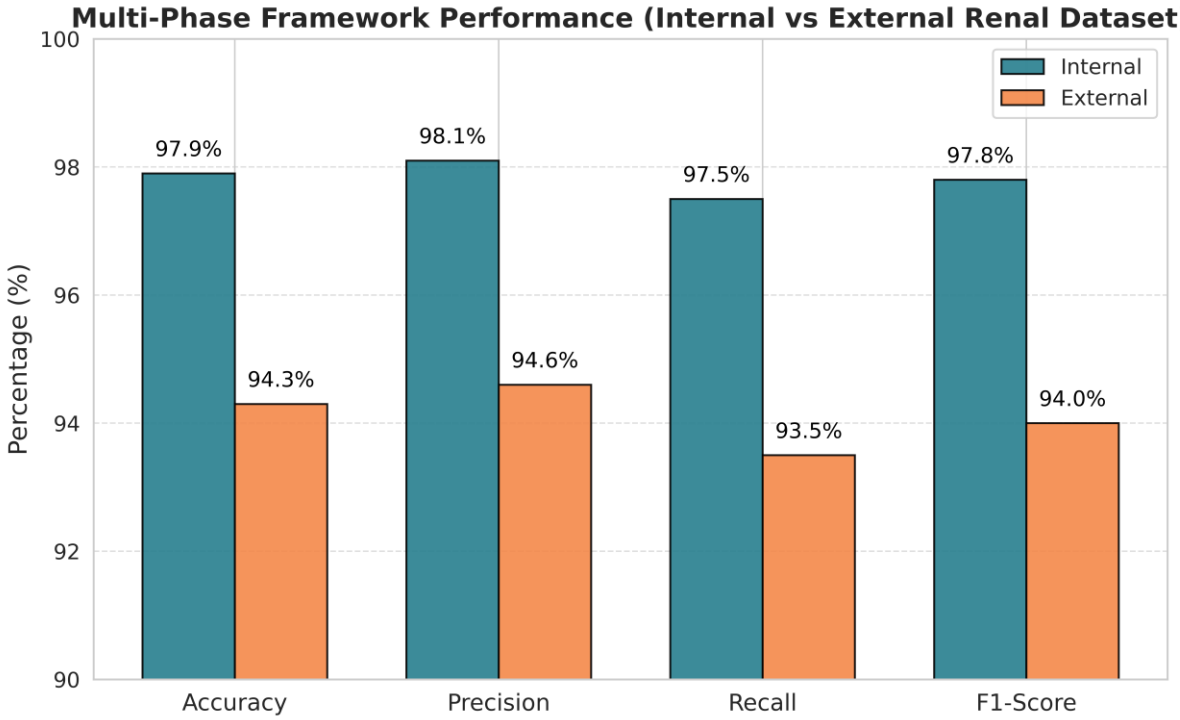


Figure 8. Model performance on internal vs. external renal ultrasound datasets across Accuracy, Precision, Recall, and F1-Score.

3.3.1. Cross-Dataset Ablation Analysis

To evaluate the constancy and robustness of the Multi-Phase Framework (MPF), a 5-fold cross-validation was analyzed on the internal renal ultrasound dataset. An ablation study shows that by adding Fusion, Attention, and Dense layers, the performance of the proposed model was significantly improved, as summarized in Table 7 below. Starting with fusion only achieved 97.0%, 96.8%, and 0.991 accuracy, F1-score, and AUC, while adding the attention achieved 97.5% and 97.3% accuracy and F1-score, respectively. Adding dense layers further enhanced the performance, as evidenced by a 97.9%, 97.8%, and 0.997 accuracy, F1-score, and AUC, respectively, in the XGBoost model. Likewise, Random Forest exhibited a comparable pattern; Fusion alone attained a 97.2% accuracy, while the complete configuration achieved 97.6%. These findings indicate that attention-guided feature fusion and dense representations substantially enhance predictive capabilities.

Table 7. Ablation study of XGBoost and RF ensembles on Internal Renal Ultrasound Dataset (Accuracy ±STD, F1 ±STD, AUC).

Classifier	Fusion	Attention	Dense Layer	Accuracy(%) ±Std(%)	Precision(%) ±Std(%)	Recall(%) ±Std(%)	F1-Score(%) Std(%)	AUC
XGBoost	YES	NO	NO	97.0 ±0.3	96.8 ±0.3	96.7±0.3	96.8 ±0.3	0.991
	YES	YES	NO	97.5 ±0.3	97.3±0.3	97.2 ±0.3	97.3 ±0.3	0.994
	YES	YES	YES	97.9 ±0.3	98.1 ±0.3	97.5±0.3	97.8 ±0.3	0.997
RF	YES	NO	NO	97.2±0.3	97.0 ±0.3	96.9 ±0.3	97.0 ±0.3	0.981
	YES	YES	NO	97.4 ±0.3	97.2±0.3	97.1 ±0.3	97.2 ±0.3	0.985
	YES	YES	YES	97.6 ±0.3	97.4 ±0.3	97.3±0.3	97.4 ±0.3	0.987

To evaluate the generalization capability of the proposed model, the trained XGBoost and Random Forest ensembles were assessed on an independent external renal ultrasound dataset without additional retraining. The results on the external dataset are shown in Table 8, where Random Forest achieved 94.1% accuracy, a 94.0% F1-score, and an AUC of 0.968, while XGBoost achieved 94.3% accuracy, a 94.1% F1-score, and an AUC of 0.970. These metrics indicate that the models' predictive performance is only slightly reduced when applied to different domains, compared to internal validation. External evaluation supports the framework's suitability for use in various imaging conditions, confirming its robustness and potential for therapeutic applications.

Table 8. External Renal Ultrasound Dataset evaluation without retraining, showing generalization capability (Accuracy, F1, AUC).

Classifier	Fusion	Attention	Dense Layer	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
XGBoost	YES	NO	NO	93.9	93.8	93.5	93.7	0.963
	YES	YES	NO	94.1	94.0	93.8	93.9	0.967
	YES	YES	YES	94.3	94.6	93.5	94.1	0.970
RF	YES	NO	NO	93.6	93.6	93.4	93.3	0.961
	YES	YES	NO	93.8	94.0	93.7	93.8	0.965
	YES	YES	YES	94.1	94.2	94.0	94.0	0.968

Table 9 below illustrates the aggregated performance of XGBoost on both internal and external datasets, highlighting the cumulative effect of all architectural improvements. The model demonstrated its excellent performance on the internal dataset, achieving an accuracy of 97.9%, a 97.8% F1-score, and an AUC of 0.997,

thereby confirming the effectiveness of the fusion, attention, and dense layers. Furthermore, the model displayed strong generalization capabilities on the external dataset, attaining an accuracy of 94.3%, an F1-score of 94.1%, and an AUC of 0.968. These findings provide a novel benchmark for automated kidney stone detection utilizing ultrasound images, highlighting the significance of multi-branch feature integration, attention-guided refinement, and dense representation learning in producing reliable, realistic predictions across different datasets. These measures indicate that the model's predictive performance is barely reduced when applied to different domains, compared to internal validation. Consequently, the external evaluation confirms the framework's effectiveness across multiple diagnostic contexts, affirming its dependability and potential across different domains.

Table 9. Performance of XGBoost on Internal and External Renal Ultrasound datasets, demonstrating optimal model configuration (Accuracy, F1, AUC)

Dataset	Classifier	Fusion	Attention	Dense Layer	Accuracy(%)	F1-score(%)	AUC
Internal	XGBoost	YES	YES	YES	97.9 ±0.03	97.8 ±0.03	0.997
External	XGBoost	YES	YES	YES	94.3	94.1	0.968

3.3.2. Row-Wise Confusion Matrix Analysis of Proposed Model

The confusion matrices depicted in Figure 9 illustrate the efficacy of the MPF on the internal renal dataset, assessed using 5k-fold cross-validation, and the external dataset with XGBoost. The model attained classification accuracy of 99.5% for kidney stone cases and 98.7% for normal cases on the internal dataset, with misclassification rates between 0.5% and 1.3%, indicating precise differentiation between stone and normal occurrences. XGBoost achieved an accuracy of 96.7% for stone instances and 97.1% for normal cases on the external dataset, demonstrating robust generalization across varied datasets. These findings imply that the model's integration of feature fusion and attention mechanisms facilitates elevated sensitivity and specificity within both internal and external cohorts, consequently underscoring its reliability and dependability in predicting kidney stones.

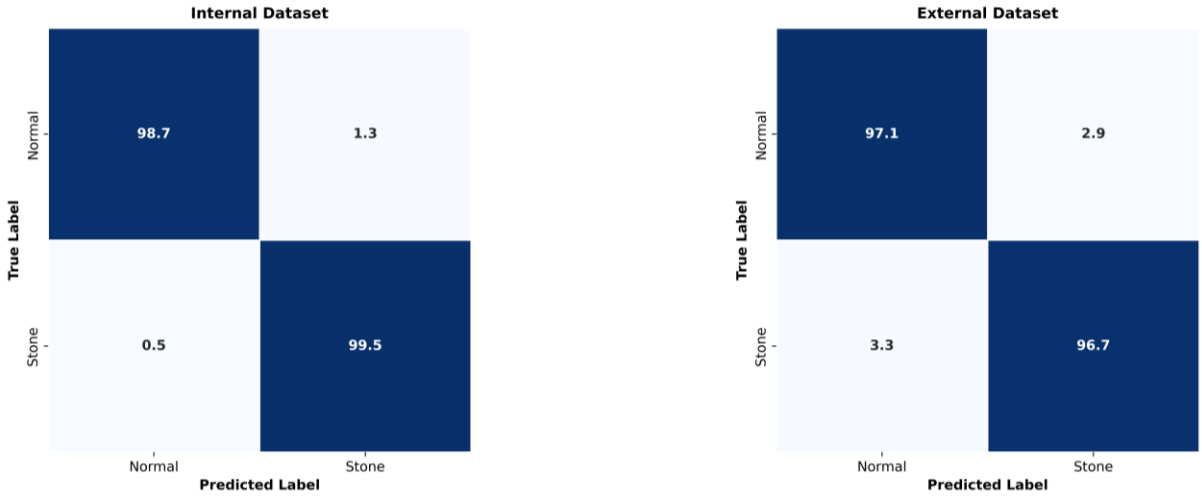


Figure 9. Confusion matrix of the proposed model on internal and external renal ultrasound datasets showing class-wise prediction accuracy for Stone and Normal cases.

3.3.3. Receiver Operating Curve Analysis of Proposed framework

Figure 10 compares XGBoost and Random Forest (RF) classifiers assessed on internal and external renal ultrasound datasets. Receiver operating characteristic (ROC) curves are shown in a single 1×3 panel layout, allowing for direct visual comparison across datasets.

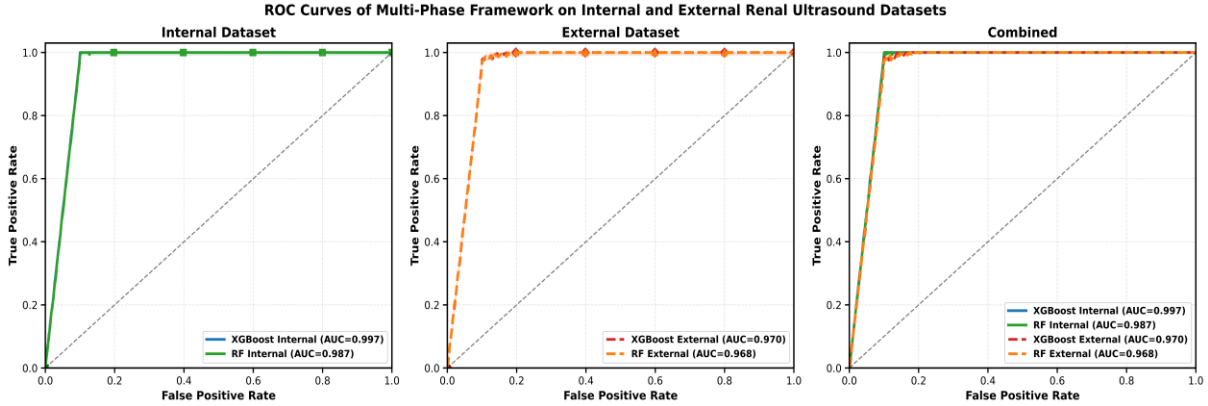


Figure 10.ROC curves of Proposed Framework on the internal and external renal ultrasound dataset, showing classification performance and AUC values for Stone vs. Normal cases.

The leftmost panel illustrates the performance on the internal dataset, with XGBoost achieving an AUC of 0.997 and RF attaining an AUC of 0.987. The central panel presents the outcomes derived from the external dataset; here, the AUC values are only slightly diminished (XGBoost: 0.970, RF: 0.968), thereby suggesting robust generalization capabilities when applied to real-world data. The rightmost panel offers a holistic perspective, integrating both internal and external datasets by aggregating all Receiver Operating Characteristic (ROC) curves into a unified graphical format. Within this chart, solid lines denote the performance metrics for the internal dataset, whereas dashed lines signify the results obtained from the external dataset.

3.3.4. Predicted Probability Curve Analysis

To visualize the predictive performance of the proposed model on both internal and external renal ultrasound datasets, a smooth predicted probability curve with overlaid bubble markers was developed as shown in Figure 11 below. The normalized sample index is highlighted on the horizontal axis, while the predicted probability on the vertical axis shows the positive class (Stone), and the predicted probability curve provides a visual representation of the model's predictive behavior across the entire dataset. Individual predictions are represented by bubbles; those colored green signify predictions that closely align with the actual class, whereas red bubbles indicate substantial divergence from the true label. Within the internal dataset, which includes the model's training and validation samples, the prevalence of green bubbles suggests a high accuracy rate of 97.9%, with only a small proportion of red bubbles (2.0%) representing isolated misclassifications. The model's predictive accuracy on the external dataset is robust, with a 94.8% rate of correct predictions. However, the presence of a small proportion of red bubbles, specifically 5.0%, indicates a degree of forecast uncertainty when applied to novel data. Despite such limitations, the model demonstrates robust predictive performance on the external dataset, with an accuracy of 94.3%. The existence of a marginally elevated percentage of red bubbles, namely 5.0%, indicates a degree of predictive uncertainty when the model is utilized on novel data. The summary, each subplot clearly highlights the ratios of both correct and incorrect predictions, providing quick quantitative context for the depicted projections. This visualization clearly illustrates the progression of predicted probability and the distribution of strong versus uncertain predictions throughout the samples.

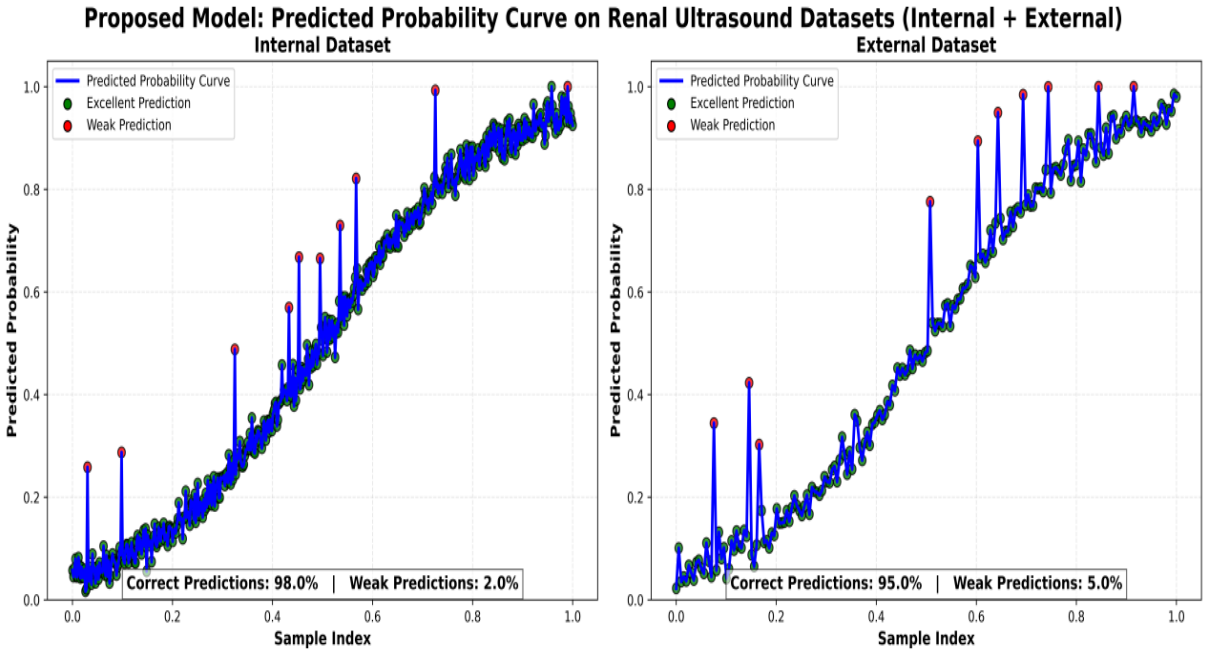


Figure 11. Proposed Framework Predicted Probability Analysis on the internal and external renal ultrasound dataset.

3.3.5. Explainability Analysis Using Grad-CAM and Attention Maps

An explainability analysis was performed to qualitatively assess the decision-making behavior of the MPF and to confirm that the models concentrate on clinically significant regions (Figure 12). Class-discriminative heatmaps were produced utilizing Grad-CAM for CNN architectures, namely ResNet50, DenseNet121, and EfficientNet-B0. This was accomplished by calculating the gradients of the projected class score concerning the final convolutional feature maps. These visualizations were then compared to the ground-truth labels, which were either Normal or Stone, and incorporated confidence scores derived from softmax probabilities.. For Vision Transformer models (ViT-Base, Swin-Transformer, DeiT-Small), explainability was derived from attention map visualization, showing how attention is distributed across image patches. Since attention maps are not class-specific, ground-truth labels were not enforced at this level, and confidence scores represent final prediction probabilities. The study also examined improved results by combining the strengths of EfficientNet-B0 and Swin-Transformer architectures. This model uses Grad-CAM and attention mechanisms to improve its interpretability. The combined results consistently show increased confidence, which indicates greater stability and less uncertainty.

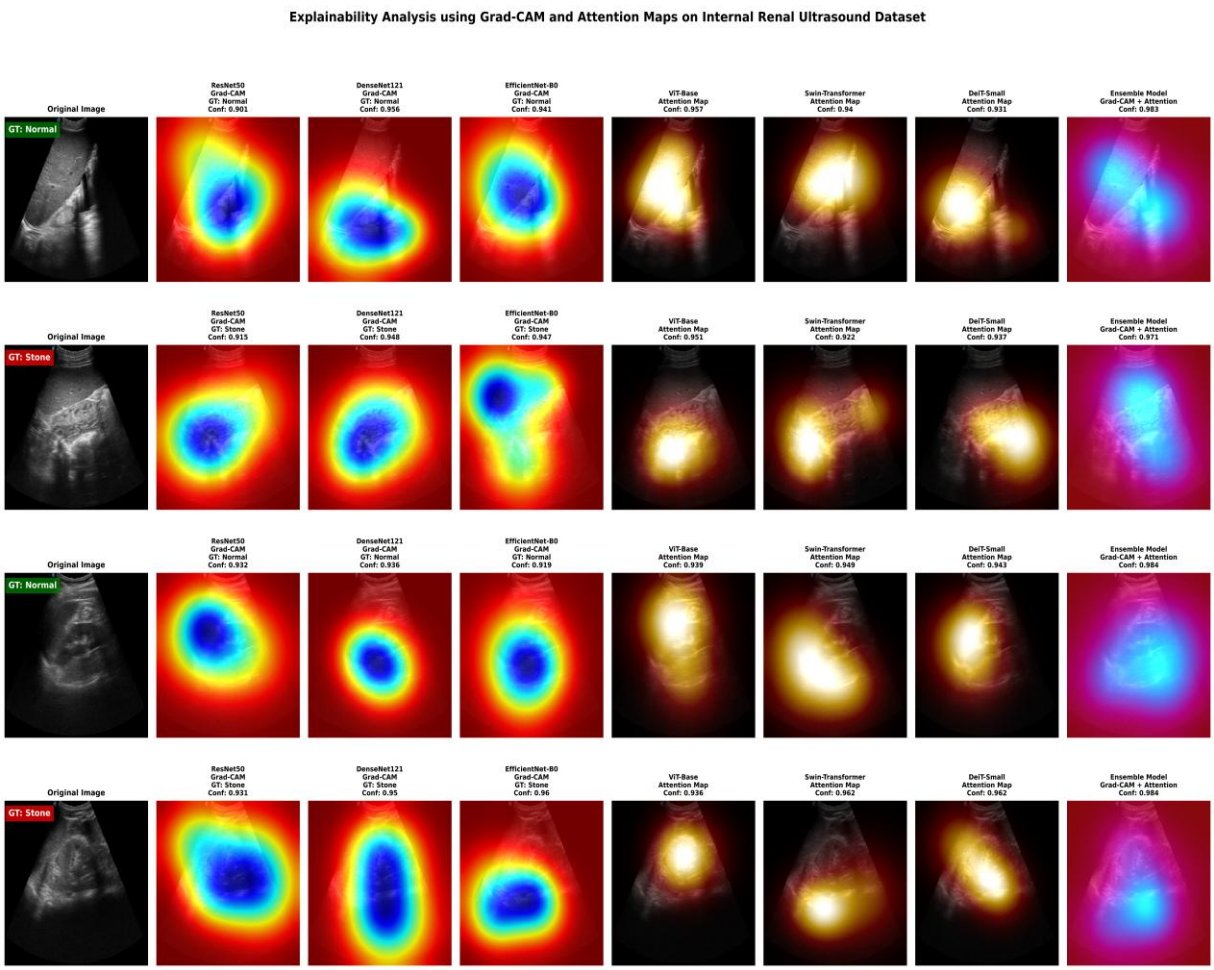


Figure 12. Grad-CAM and attention visualizations of CNN and ViT variations on internal renal ultrasound images.

4. Discussion

The current study demonstrates the advantages of integrating local, global, and frequency-domain features to improve the kidney stone detection from ultrasound images. Dense and attention-based feature fusion balances the contributions of all branches, reducing the dominance of CNN features that may result from their high-dimensional vector size. The Multi-Phase Framework consistently outperforms the baseline CNN,ViT and ensemble models, indicating its ability to capture complementary information that single-branch architectures cannot effectively utilize. Several limitations should still exist in the current study: first, the lack of demographic information, such as patient age, sex, and comorbidities, limited subgroup analysis and hindered the evaluation of model performance across diverse populations. Second, the model only processes 2D static ultrasound images and does not use temporal information from video sequences. Thirdly, the large dimensionality of CNN features may limit the adoption of real-time applications on portable ultrasound devices, necessitating substantial memory and processing power. Furthermore, the wavelet transform complicates preprocessing and requires meticulous parameter tuning to ensure equitable contribution from each feature branch. Techniques such as Score-CAM or Smooth Grad-CAM, which are sophisticated explainable AI methodologies, can enhance the transparency of clinical decisions, consequently fostering greater acceptance and trust. Unlabeled ultrasound images present an opportunity for semi-supervised or self-supervised learning, potentially enhancing feature representations, especially when dataset sizes

are constrained. The model's real-time performance on portable devices could facilitate point-of-care diagnostics in resource-limited environments, and the framework's expansion to include multi-class renal disorders is also proposed. This study presents a robust, adaptable, and broadly applicable framework for identifying kidney stones using ultrasound imaging. By incorporating attention-based fusion with multi-branch feature extraction, the model strikes a balance between clinical applicability and technological precision. These findings pave the way for future investigations and broader implementation in practical ultrasound diagnostics, while also underscoring the promise of multi-branch deep learning architectures in medical image processing.

Table 10. Automated Kidney Stone Detection and classification, Datasets, Results, Methods and Limitations Analysis.

Year	Dataset	Method	Accuracy	Limitation1	Limitation2	Handled
2025[18]	CT Dataset	Ensemble CNN	100	Limited Generalization	Overfit	Yes
2025[13]	Ultrasound (images)	Hybrid CNN-LSTM	97	No External Validation	No Explainable AI	Yes
2024[27]	CT kidney Dataset	CNNs	96.0	No Explainable AI	NO External Validation	Yes
2023[17]	Coronal CT	Lightweight CNNs	96%	No External Validation	Limited Dataset	Yes
2022[24]	2,959 CT images	CNN Variations	63-93	No Explainable AI	Generalization Challenges	Yes

5. Conclusions

A Multi-Phase Framework is introduced for kidney stone detection in ultrasound images that integrates the convolutional neural networks (CNNs) to extract local features, Vision Transformers (ViTs) to capture global contextual information, and a wavelet-based branch to obtain the frequency-domain features and combine them using a dense attention-based fusion module. The CNN branch identifies local textures, including edges and stone boundaries; the ViT branch provides anatomical context at a global scale; and the wavelet branch extracts high- and low-frequency patterns that spatial feature extractors may overlook. Interpretability of the framework was enhanced through the use of saliency maps and layer-wise relevance propagation (LRP). These methods confirmed that model predictions focus on clinically relevant regions on an internal dataset of 9,416 ultrasound images, demonstrating 97.9% accuracy, with precision, recall, and F1-scores all exceeding up to 97%, and an area under the curve (AUC) of 0.997, indicating strong discriminative capability. External validation on an independent dataset yielded 94.3% accuracy, confirming the solid generalization without any risk of overfitting, while the ablation studies verified that each branch, including the wavelet module, significantly enhances overall performance. The proposed MPF offers a strong, non-invasive, and easily understood method for identifying kidney stones. By combining local, global, and frequency-domain data with dense attention refinement and explainable AI techniques, the framework is well-suited for clinical settings and practical diagnostic use. These results highlight the promise of multi-branch, attention-guided deep learning architectures to improve accuracy and reliability in ultrasound-based medical imaging.

Informed Consent Statement: Not applicable

Data Availability Statement: This study has been conducted using two publicly available datasets and Links for downloading these datasets are following:

Internal: https://www.kaggle.com/datasets/gurjeetkaurmangat/kidney-ultrasound-images-stone-and-no-stone

External: https://www.kaggle.com/datasets/zaynebnouiri/renal-data

Acknowledgments: In the authors would like to acknowledge the support received during the course of this study. No additional administrative or technical assistance beyond the listed author contributions was involved. During the preparation of this manuscript, the authors used Grammarly for the purposes of language refinement, grammatical correction, and improving the clarity and readability of the text. The authors have reviewed and edited all generated suggestions and take full responsibility for the accuracy, integrity, and originality of the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
ViT	Vision Transformer
DWT	Discrete Wavelet Transform
RF	Random Forest
AUC	Area Under the Curve
F1-Score	F1-Measure / Harmonic Mean of Precision and Recall
SD	Standard Deviation
Grad-CAM	Gradient-weighted Class Activation Mapping
LRP	Layer-wise Relevance Propagation
MPF	Multi-Phase Framework

References

[1]. Xu, W., et al. (2026). Position-aware and decision-guided domain incremental learning framework for multi-view kidney stone detection. *Biomedical Signal Processing and Control, 112*, 108366.
[2]. Lu, D., et al. (2026). A computer vision model for automated kidney stone segmentation and evaluation of its performance vs surgeons. *BJU International, 137*(1), 87–94.
[3]. Sh, A. S., & Sherdorbek O‘lmasbek o‘g, M. (2026). Diagnosis of kidney stone disease in children. *Pedagogs International Research Journal, 98*(1), 48–50.
[4]. Cao, J., et al. (2026). Uric acid levels mediate the association between four dietary indices and kidney stones in US adults: A cross-sectional study of NHANES 2007–2018. *PLOS ONE, 21*(1), e0339839.
[5]. Indurani, M., et al. (2026). Modern urological approaches to renal stone disease: Diagnostic advances and therapeutic innovations. *Kidneys, 15*(1), 155–166.

[6]. Vasudeva, N., Dhaka, V. S., & Sinwar, D. (2025). Enhancing kidney stone diagnosis with AI-driven radiographic imaging: A review. *Discover Artificial Intelligence, 5*(1), 200.
[7]. Panchal, N., Raikar, M. M., & Baligar, V. P. (2025). Kidney stone detection using deep learning model. *Procedia Computer Science, 260*, 56–63.
[8]. Munir, S., & Imran, M. (2025). Detection of kidney stones using CT scan imaging.
[9]. Kulandaivelu, G., et al. (2025). An implementation of adaptive multi-CNN feature fusion model with attention mechanism with improved heuristic algorithm for kidney stone detection. *Computational Intelligence, 41*(1), e70028.
[10]. Achararit, P., et al. (2025). A novel approach for automated renal stone detection from KUB radiographs in Thai population. In *Proceedings of the IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE.
[11]. Tang, Z., et al. (2025). An optimized bidirectional recurrent neural network for kidney stone detection based on developed bald eagle search method in CT scan images. *Scientific Reports, 15*(1), 37109.
[12]. Yalçın, S. (2025). Kidney stone detection using an EfficientNet-based method. *Computer Science, 10*(1), 1–10.
[13]. Jadhav, A. N., et al. (2025). AI-powered early detection of kidney stones using a hybrid CNN-LSTM model. *Journal of Neonatal Surgery, 14*(9s).
[14]. Hossain, M. N., et al. (2025). Detection and classification of kidney disease from CT images: An automated deep learning approach. *Technologies, 13*(11), 508.
[15]. Rapelang, S. L., & Obagbuwa, I. C. (2025). Hybrid support vector machine–convolutional neural networks multi-classification models for detection of kidney stones. *International Journal of Imaging Systems and Technology, 35*(4), e70128.
[16]. Srinivas, A. C. M. V., et al. (2025). Secure and intelligent framework for kidney stone detection using a CNN-XGBoost hybrid model with blockchain integration. In *Proceedings of the World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS)*. IEEE.
[17]. Wang, F., Silvestre, G., & Curran, K. M. (2023). Lightweight framework for automated kidney stone detection using coronal CT images. *arXiv preprint*.
[18]. Sharma, K., et al. (2025). Hybrid deep learning framework for classification of kidney CT images: Diagnosis of stones, cysts, and tumors. *arXiv preprint*.
[19]. Salem, A., & Mondal, A. (2024). A CNN approach to polygenic risk prediction of kidney stone formation. *arXiv preprint*.
[20]. Srivastava, R., & Kumar, P. (2022). A CNN-SVM hybrid model for the classification of thyroid nodules in medical ultrasound images. *International Journal of Grid and Utility Computing, 13*(6), 624–639.
[21]. Tahir, F. S., & Abdulrahman, A. A. (2023). Kidney stones detection based on deep learning and discrete wavelet transform. *Indonesian Journal of Electrical Engineering and Computer Science, 31*(3), 1829.
[22]. Li, X., et al. (2023). Deep learning attention mechanism in medical image analysis: Basics and beyonds. *International Journal of Network Dynamics and Intelligence*, 93–116.
[23]. Göker, H. (2024). Transfer learning-based classification of kidney stone from computed tomography images. In *Proceedings of the International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE.
[24]. Islam, M. N., et al. (2022). Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Scientific Reports, 12*(1), 11440.
[25]. Choi, H.-S., et al. (2023). Transfer learning for effective urolithiasis detection. *International Neurourology Journal, 27*(Suppl 1), S21.
[26]. Vishmitha, D., et al. (2022). Kidney stone detection using deep learning and transfer learning. In *Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE.
[27]. Naresh, P., et al. (2024). Transfer learning based kidney stone detection using ResNet50 with medical images. In *Disruptive Technologies in Computing and Communication Systems* (pp. 286–291). CRC Press.
[28]. Obaid, W., et al. (2025). Noisy ultrasound kidney image classifications using deep learning ensembles and Grad-CAM analysis. *AI, 6*(8), 172.
[29]. Koo, J. C., et al. (2023). Non-annotated renal histopathological image analysis with deep ensemble learning. *Quantitative Imaging in Medicine and Surgery, 13*(9), 5902.
[30]. Tsai, C. M., & Lee, J.-D. (2024). Grad-CAM visualization and ensemble learning for improved gastrointestinal disease classification using CNNs. In *Proceedings of the IEEE Global Conference on Consumer Electronics (GCCE)*. IEEE.

[31]. Villegas-Jimenez, A., Flores-Araiza, D., & Lopez-Tiro, F. (2023). Causal scoring medical image explanations: A case study on ex-vivo kidney stone images. *arXiv preprint*.

[32]. Balanageshwara, S., & Aashitha, S. (2024). Automated kidney stone detection and prediction using enhanced deep learning models. *Cuestiones de Fisioterapia, 53*(03), 1153–1162.

[33]. Sitaraman, S. R. (n.d.). Adaptive CNN-LSTM and neuro-fuzzy integration for edge AI and IoMT-enabled chronic kidney disease prediction.

[34]. Ogundokun, R. O., et al. (2022). Hybrid InceptionV3-SVM-based approach for human posture detection in health monitoring systems. *Algorithms, 15*(11), 410.

[35]. Luo, X., Wang, Y., & Ou-Yang, L. (2025). LGFFM: A localized and globalized frequency fusion model for ultrasound image segmentation. *IEEE Transactions on Medical Imaging*.

[36]. Ramesh, P., et al. (2023). Automatic kidney stone detection using deep learning method. *Journal of Advanced Zoology, 44*, 100–109.

[37]. Billah, M., et al. (2024). Real-time object detection in medical imaging using YOLO models for kidney stone detection. *European Journal of Computer Science and Information Technology, 12*(7), 54–65.

[38]. Rahman, T., & Uddin, M. S. (2024). Speckle noise reduction and segmentation of kidney regions from ultrasound image. In *Proceedings of the International Conference on Informatics, Electronics and Vision (ICIEV)*. IEEE.

[39]. Asaye, Y. A., Annamalai, P., & Ayalew, L. G. (2025). Detection of kidney stone from ultrasound images using machine learning algorithms. *Scientific African, 28*, e02618.

[40]. Karthikeyan, M., et al. (2024). Kidney stone detection using deep learning techniques. In *Proceedings of the International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE.

[41]. Yildirim, K., et al. (2021). Deep learning model for automated kidney stone detection using coronal CT images. *Computers in Biology and Medicine, 135*, 104569.

[42]. Huang, Z.-H., et al. (2023). Design and validation of a deep learning model for renal stone detection and segmentation on kidney–ureter–bladder images. *Bioengineering, 10*(8), 970.

[43]. Ramadhani, A., & Salam, A. (2024). Deployment of web-based YOLO for CT scan kidney stone detection. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 8*(3), 1357–1368.

[44]. Borges, T., et al. (2022). Kidney stone detection using ultrasound images. *Proceedings of the Advancement in Electronics & Communication Engineering*.

[45]. Kanagamalliga, S., et al. (2024). Cutting-edge techniques for enhanced kidney stone detection in ultrasound imaging. In *Proceedings of the International Conference on Computing, Sciences and Communications (ICCSC)*. IEEE.

[46]. Venkatrao, K., & Kareemulla, S. (2023). HDLNET: A hybrid deep learning network model with intelligent IoT for detection and classification of chronic kidney disease. *IEEE Access, 11*, 99638–99652.

[47]. Pimpalkar, A., et al. (2025). Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. *Scientific Reports, 15*(1), 10741.