OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 03 - Issue 06 (2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

DEEP LEARNING-BASED CLINICAL DECISION SUPPORT SYSTEM FOR AUTOMATED KIDNEY STONE DETECTION IN CT IMAGES USING CUSTOM CNN AND TRANSFER LEARNING

Amjad khan[1], Muhammad Javed[2], Nasir Gul[3], Saif Ullah Noor[4], Reyan[5], Ashraf ullah[6]

[1, 2, 3, 4, 5, 6]Department of Computer Science, University of Science and Technology Bannu, Khyber Pakhtunkhwa, Pakistan.

[*]Corresponding Author: (amjadkhanbscs@gmail.com)
DOI: (*https://doi.org/10.71146/kjmr952*)

Article Info

Abstract

Kidney stone disease (nephrolithiasis) is a prevalent urological condition that is arguably the most prevalent one and also has a high burden of disease for the healthcare system. Even today, non-contrast computed tomography (NCCT) remains the gold standard imaging technique for stone detection, although the manual reading of computed tomography (CT) scans may be time-consuming and observer-dependent, especially for high-volume clinical practice. In this work, we develop an automated deep learning system for CT image detection of kidney stones from axial images. Two CNN-based systems were designed and tested: one with a specifically designed CNN architecture optimized for computation efficiency (cnnD) and another (cnnT) with a transfer learning approach using an architecture pre-trained on ImageNet. There were a total of 3,154 annotated CT images, which were used for training (2,522 images), validation (316 images), and testing (316 images). Pre-processing and normalization of standard images and data augmentation were used to enhance the model's robustness and generalisation. The custom CNN model's accuracy of 94.3%, precision of 0.95, recall of 0.94, and F1-score of 0.94 on the independent test set were validated in the experiment. Therefore, transfer learning for medical image analysis was demonstrated to be effective when the fine-tuned VGG16 model classified the images with a higher accuracy (96.5%). The custom CNN had less computation needed and also had quicker inference time as compared to VGG16, even though the latter achieved better predictive performance and thus could be applied in real clinical settings. The results suggest that the deep learning methods are useful for detecting kidney stones with high accuracy and efficiency from CT scans, and potentially for assisting radiologists in their clinical diagnosis and decisions. Overall, the proposed framework emphasizes the potential of using AI throughout the diagnostic process, and its goal is to improve the efficiency, uniformity, and accessibility of the diagnosis.

Keywords: *Kidney Stone Detection, Nephrolithiasis, Deep Learning, Convolutional Neural Networks, Computed Tomography, VGG16, Transfer Learning, Clinical Decision Support.*

Introduction

In its clinical aspect, kidney stone disorder is called nephrolithiasis, which is still one of the most common urological diseases in the world and remains a major health problem. It occurs in about 10–15% of the world's population, and most of the population that is affected has a recurrence rate, 45% of people suffering from another attack in 10 years. Severe clinical complications of kidney stones may include progressive renal function impairment, recurrent infections, urinary tract stones, and/or acute renal colic. Nephrolithiasis is a post-modern phenomenon, with a number of lifestyle and metabolic risk factors being associated (obesity, diabetes mellitus, dietary intake, and reduced fluid consumption) [2, 3].

Correct and prompt diagnosis is crucial to successful management and prevention of complications. Out of the available imaging modalities, non-contrast computed tomography (NCCT) has a high sensitivity and specificity for detecting urinary calculi, and it is considered the preferred imaging modality [4]. Although computer tomography is highly diagnostic, its interpretation is hard work, which still requires a lot of expertise in the radiology field. Reporting delays, inter-observer disagreement, and potential for missed subtle pathological diagnosis, especially when dealing with a large stone or multiple stones, may be a problem with the increasing volume of cases in a busy clinical setting [5], [7]. Computer-aided diagnostic systems have been studied as a support system for medical image analysis; in fact, these systems have many of the above-mentioned internal limitations. Initial methods were mainly based on manually constructed image features and old and standard machine learning algorithms. These techniques had been demonstrated to have some potential, but often had limitations due to image quality, anatomical complexity, and differences in acquisition protocols, which prevented their use in clinical practice in real-world scenarios [8, 9]. Over the past few years, great improvements have been made in the domain of medical image analysis, using deep learning techniques. Recently, convolutional neural networks (CNNs) have shown great promise in automatically learning hierarchical image representation and achieved success in many clinical applications of disease detection, segmentation, and classification problems in different medical fields. Khan et al. [10] presented a study achieving good results with deep learning models for medical image analysis, while Maqsood and Khan [11] and Hekmat et al. [12] obtained excellent results in the field of disease detection. Islam et al. [15] also showed positive results in medical image analysis using deep learning models. CNN-based models can directly learn implicitly complex visual patterns from image data and demonstrate better diagnostic performance when compared to the process of manual feature engineering. Nevertheless, studies in particular on the automatic detection of kidney stones are still quite limited. Prior to this, Shetty et al. [13] and Pree Danan et al. [14] have shown that deep learning could be used for kidney stone detection, Elton et al. [16] have provided proof that the technology is also effective for analyzing kidney stones, and Cui et al. [18] have attempted to interpret the models learned by the deep learning systems using detailed explanations. However, there have been a number of investigations that have considered predictive accuracy, but not much emphasis has been put on the efficiency of the computations, their ability to be deployed in a clinic, or the interpretability of the learned models. Keeping in mind such challenges, this study proposed a deep learning system for automated kidney stone detection in axial CT images. Two complementary classification approaches are explored, namely a tailored CNN that provides efficient inferences with a lower number of calculations, and a transfer learning CNN

that builds a model on the basis of pre-trained feature representations to boost the classification performance. The database, made up of 3,154 labelled CT scans, served as the training and testing material for the two models to conduct an extensive analysis of their accuracy, effectiveness, and applicability.

The major contributions.

1. . Automated detection algorithm for a computationally efficient convolutional neural network. The classification of kidney stones based on the CT images.
2. Comprehensive evaluation of a custom CNN architecture and a fine-tuned VGG16 transfer Evaluation of learning model in the same experimental conditions
3. Analysis of the trade-off between diagnostic performance and computational efficiency to support practical clinical deployment.
4. Integration of Grad-CAM-based visual explanations to improve model transparency and interpretability.
5. Development of a web-based clinical decision-support framework for automated kidney stone detection and assessment.

Literature Review

The integration of artificial intelligence (AI) and deep learning methods in medical imaging has revolutionized the medical diagnosis and identification of many illnesses, such as nephrolithiasis. Machine learning, convolutional neural networks (CNNs), transfer learning and explainable artificial intelligence (XAI) have been used to develop efficient automated systems to support the use of clinicians to locate kidney stones with an accuracy of up to 95%. Various imaging techniques have been studied such as computed tomography (CT), ultrasound imaging and non-imaging clinical parameters for the better diagnostic efficiency and to minimize human interaction with manual interpretation. To improve such CNNs limitations Asif et al. [6] introduced two ensemble deep learning architectures namely Stacked Ensemble Net (SEN) and PSO Weighted Avg Net (PWAN). They used various pre-trained networks like InceptionV3, ResNetV2, Inc option, MobileNet and Xception and leveraged only a small portion of the features to build the framework so that its classification performance is more robust. On the one hand, the extension of Particle Swarm Optimization (PSO) was employed to improve the predictive performance, where the ensemble weights were optimized. Experimental results showed the models had higher accuracy than the standalone models and visualizations provided by Grad-CAM. However, ensemble architectures can be computationally demanding and may not be widely used in clinical settings due to these demands. In order to lower the cost and radiation dose of the computed tomography (CT) imaging, Asaye et al. [7] discussed an ultrasonic diagnostic system. They used their approach, including image preprocessing, segmentation, and feature extraction based on Gray Level Co-occurrence Matrix (GLCM), as well as the classification of classification algorithms such as support vector machine, artificial neural network, and Naive Bayes classifiers. The highest accuracy (98.4%) and area under the curve (AUC, 0.98) were achieved by the K-nearest neighbors (KNN) classifier. Another advantage of the study is the accurate estimation of size. While these are promising results, current systems which use ultrasound suffer from limited image quality and anatomic variability, and are highly dependent on the skill of the operator. Various alternative non-

imaging methods have been investigated too. Gulhane et al. [8] devised a machine-learning model that likely can reduce the need for radiological imaging techniques by using parameters from biochemical urine analysis. The authors assessed multiple machine learning and deep learning models to use a dataset of urinary biomarkers. Among the three with the best performance was their improved deep neural network, which suggests that there are plenty of potential avenues for non-invasive screening. However, the model has limited generalizability due to the limited size of the dataset. For deep learning approach related to CT analysis, Sharma et al. [9] proposed a Hybrid ResNet101-Custom CNN with Feature Fusion approach for discriminative learning. The framework was trained using a large scale CT dataset and gave very high classification performance. While the achieved accuracy is near perfect and the results shows potential of hybrid deep learning models, there is need for further testing with these models on independent external data sets to validate robustness and generalizability.

In this case, Kabir and Lee [5] have illustrated that the deep learning approach can be used for recognizing words such as “stone” through RGB-D images using the Mask R-CNN, thus broadening the applications of the algorithm outside the scope of medicine. The 3D algorithm can be used in identifying and locating any obstructing stone inside the underground pipeline. Even though the application area is different from kidney stone diagnosis, they emphasized the benefits of combining object detection with depth perception for accurate localization. Nowadays, there has been much emphasis on making a trade-off between performance and efficiency in diagnosing kidney stones. According to [10], the work done by Mallikarjun et al., they combined CNN with SVM approaches to recognize CT scans of kidney stones, which gave enhanced results through deep learning combined with classical machine learning algorithms. Likewise, Kumar et al. [11] designed a light and efficient CNN architecture for deployment in resource constrained environments to get high accuracy with less inference time. But, due to lack of explain ability tools, the clinical transparency and user trust is limited. The field of medical image analysis (MIA) is a very successful example of transfer learning, particularly in cases where annotated data sets are scarce. Zhang et al. performed an investigation and classification study on the use of ResNet50 and Efficient Net for Kidney Stone classification and obtained desirable diagnostic performance [12]. Transfer learning produces better feature representation and faster convergence, but in most cases, these models need larger computing power and limited usage in healthcare in low-power environments. With the rise of explainable AI, the specialized research in this field remains. The attention-based network and the Grad-CAM technique were proposed by Patel et al. (2013) in order to identify the presence of kidney stones. The approach provides greater interpretability of the Deep Learning model due to the ability to demonstrate areas of the image that have the greatest effect on the decision made. However, the increased computational cost remains an issue for real-world use. From this analysis, it can be concluded that significant advances have been made in automating kidney stone detection techniques. Yet, there are still some key aspects to take into consideration since the problem of computational costs and feasibility in terms of interpretability is often ignored in the papers discussed. In addition, there are relatively few direct comparisons made between lightweight custom CNN architectures and transfer learning architectures like VGG16. Less work has been done on the integration of explainability techniques in deployable systems for clinical decision support. The present study seeks to create and test a new deep learning-based framework for automated kidney stone detection in computed tomography (CT) images, which is efficient, interpretable and clinically applicable, motivating such research by the gaps identified.

Table 1: Synthesis and Research Gap

Study	Modality	Approach	Accuracy	Key Strength	Limitation
Asif et al. [1]	CT	Ensemble CNNs	High	Robust feature learning	Computationally expensive
Asaye et al. [2]	Ultrasound	ML + GLCM	98.4%	Radiation-free diagnosis	Operator dependency
Gulhane et al. [3]	Urine Analysis	Deep Neural Network	90.0%	Non-invasive screening	Limited dataset size
Sharma et al. [4]	CT	Hybrid CNN-ResNet	99%	Multi-class classification	Potential overfitting
Kabir and Lee [5]	RGB-D Images	Mask R-CNN	92.0%	Three-dimensional localization	Non-medical application
Mallikarjuna et al. [6]	CT	CNN-SVM Hybrid	96.1%	Improved robustness	Limited dataset
Kumar et al. [7]	CT	Lightweight CNN	~95.0%	Fast inference	Limited explainability
Zhang et al. [8]	CT	Transfer Learning	98.0%	Strong feature extraction	High computational cost
Patel et al. [9]	CT	XAI-Based CNN	96.8%	Improved interpretability	Computational overhead

Methodology

This study presents an automated deep learning framework for kidney stone detection from axial computed tomography (CT) images. The proposed framework consists of five stages: (A) data acquisition and preprocessing, (B) model development, (C) training and optimization, (D) performance evaluation and explain ability, and A workflow diagram of the proposed system is illustrated in Fig. 1.

A. Dataset Acquisition and Preparation

In this study, the dataset was downloaded from the public dataset, Axial CT Imaging Dataset: Kidney Stone Detection, available in the Kaggle repository. The data consists of labelled axial computed tomography (CT) images of two classes: Stone and Non-Stone. Stratified sampling was

employed to separate the data for model construction in an impartial method. validation, and test sets, with a ratio of 80:10:10, making sure there was no change in class distribution among the sets. All images were scaled to 224 × 224 and then normalized to [0,1] before training. The training set was augmented with data rotation, zooming, translation, horizontal flipping, shearing to improve the generalization of the model and prevent overfitting.

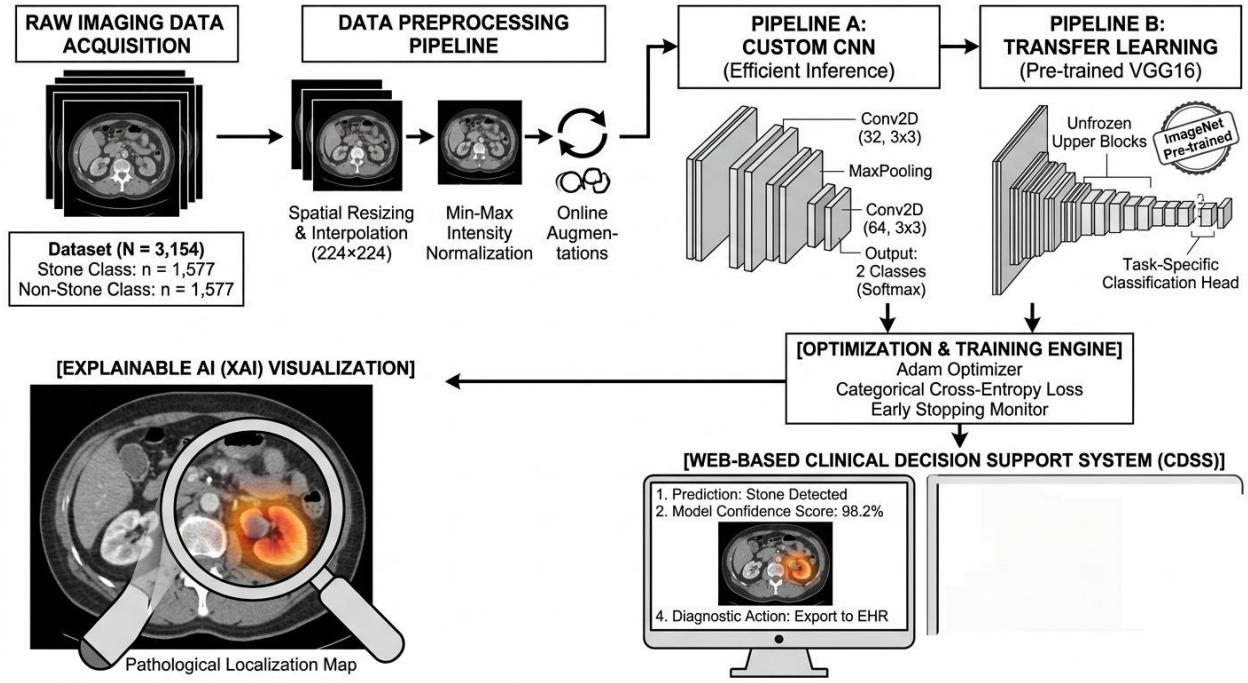

Figure 1 System Diagram

B. Custom CNN Architecture

A custom convolutional neural network (CNN) was developed as a baseline model for kidney stone classification. The architecture consists of multiple convolutional layers followed by Rectified Linear Unit (ReLU) activation functions and max-pooling operations for hierarchical feature extraction. The extracted features were flattened and passed through fully connected layers before final classification using a SoftMax output layer.

The forward propagation process is expressed as

$$[Z_l = W_l * X_l + b_l]$$

$$[A_l = ReLU(Z_l)]$$

$$[P_l = MaxPool(A_l)]$$

where (W_l) and (b_l) represent trainable weights and biases, respectively.

The final classification output is obtained using

$$\left[\hat{Y} = Softmax\left(W_f F + b_f\right)\right]$$

where (F) denotes the flattened feature representation.

C. Transfer Learning Using VGG16

In order to explore the effectiveness of transfer learning, a second model was created with a pre-trained VGG16 model trained on the ImageNet dataset. In the first step, the convolutional base was frozen and used as a feature extractor. The network was followed by a custom classification head that comprises Global Average Pooling, Batch Normalization, Dense layers, and Dropout regularization.

The model prediction is represented as

$$\hat{Y} = g\left(f_{VGG16}(X)\right)$$

where (f_{VGG16}) denotes the pre-trained feature extraction component and $\left(g(\cdot)\right)$ represents the task-specific classification head.

Following the initial convergence, a part of the upper layers of the VGG16 backbone was unfrozen and fine-tuned with a reduced learning rate to make use of learned representations for CT imaging characteristics.

D. Training and Optimization

We trained both models using the Adam optimizer. The loss function was categorical cross-entropy:

$$\left[L = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\widehat{y_{i,c}}\right)\right]$$

where $\left(y_{i,c}\right)$ and $\left(\widehat{y_{i,c}}\right)$ represent the ground-truth and predicted probabilities, respectively.
A batch size of 32 was employed for all experiments. The custom CNN was trained for a maximum of 100 epochs, whereas the VGG16 model was trained for up to 50 epochs. Early stopping with a patience of seven epochs was implemented to prevent overfitting. In addition, the ReduceLROnPlateau strategy was applied to dynamically adjust the learning rate according to validation performance:

$$\eta_{new} = 0.3 \times \eta_{old}$$

where (\eta) denotes the learning rate.

E. Performance Evaluation

The trained models were evaluated on an independent test set using standard classification metrics, including Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Accuracy is defined as

$$[Accuracy = \{TP + TN\}\{TP + TN + FP + FN]$$

Precision is calculated as

$$\left[Precision = \frac{TP}{TP + FP}\right]$$

Recall (Sensitivity) is calculated as

$$\left[Recall = \frac{TP}{TP + FN}\right]$$

The F1-score is computed as

$$\left[F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}\right]$$

These metrics provide a comprehensive assessment of model performance and diagnostic reliability.

F. Model Explainability

In order to increase transparency, Grad-CAM is used to determine which areas in an image have the highest influence on the prediction process in the model. This is done by computing gradient-weighted significance for neurons based on target class c.

$$\left[\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}\right]$$

The resulting Grad-CAM heatmap was generated as

$$\left[L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)\right]$$

The resulting Grad-CAM heatmaps were placed onto the original CT images to provide visual explanations for the model predictions, showing the precise image regions that influenced the decision

Results and Discussion

This section presents a detailed examination of the proposed automatic kidney stone detection method employing Custom Convolutional Neural Network (CNN) using axial computed tomography (CT) images. For validation of effectiveness of the model, it was evaluated by analyzing the experimental results through parameters such as accuracy, loss convergence, precision, recall, F1 score and performance through a confusion matrix.

4.1 Custom CNN Model Summary

The Custom CNN model was designed with two convolutional and pooling layers followed by a dense architecture. The architecture is optimized for a compromise between complexity and generalization, with a total number of trainable parameters of 23,907,650.

Table 2 Architecture of the Custom CNN Model

Layer (Type)	Output Shape	Parameters
Conv2D (32 filters, 3×3)	(222 × 222 × 32)	896
MaxPooling2D (2×2)	(111 × 111 × 32)	0
Conv2D (64 filters, 3×3)	(109 × 109 × 64)	18,496
MaxPooling2D (2×2)	(54 × 54 × 64)	0
Flatten	(186,624)	0
Dense (128 neurons, ReLU)	(128)	23,888,000
Dense (2 neurons, SoftMax)	(2)	258
Total Parameters		**23,907,650**

The model was trained on 2,522 images, validated on 316 images and tested on 316 images, all balanced equally between Stone and non-Stone classes.

Training and Validation Performance

The model was trained for a maximum of 100 epochs with Adam optimizer and categorical cross-entropy loss with an initial learning rate of 0.001. Overfitting was avoided using early stopping and learning rate reduction mechanisms.

Table 3 Epoch-wise Training and Validation Results

Epoch	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
1	0.7133	0.8766	0.6794	0.3252
2	0.9163	0.9241	0.2166	0.1610
3	0.9715	0.9304	0.0873	0.1693
4	0.9830	0.9304	0.0602	0.1414
5	0.9893	0.9430	0.0359	0.1470

6	0.9917	0.9399	0.0303	0.1373
7	0.9921	0.9399	0.0265	0.1233
8	0.9905	0.9494	0.0323	0.1142
9	0.9897	0.9430	0.0315	0.1551
10	0.9929	0.9430	0.0213	0.1351
11	0.9937	0.9494	0.0201	0.1130
12	0.9933	0.9399	0.0166	0.1312
13	0.9937	0.9399	0.0159	0.1388
14	0.9952	0.9462	0.0153	0.1132
15	0.9964	0.9494	0.0096	0.1253
16	0.9968	0.9494	0.0057	0.1196

Training Behavior Analysis

The standard a model displayed a quick convergence the accuracy increased from 71.33% to 99.68% in the first 16 epochs. Validation metric accuracy was stable at 93-95%, demonstrating strong generalization and no overfitting. The respective loss values declined in a consistent way, verifying efficient optimization techniques.

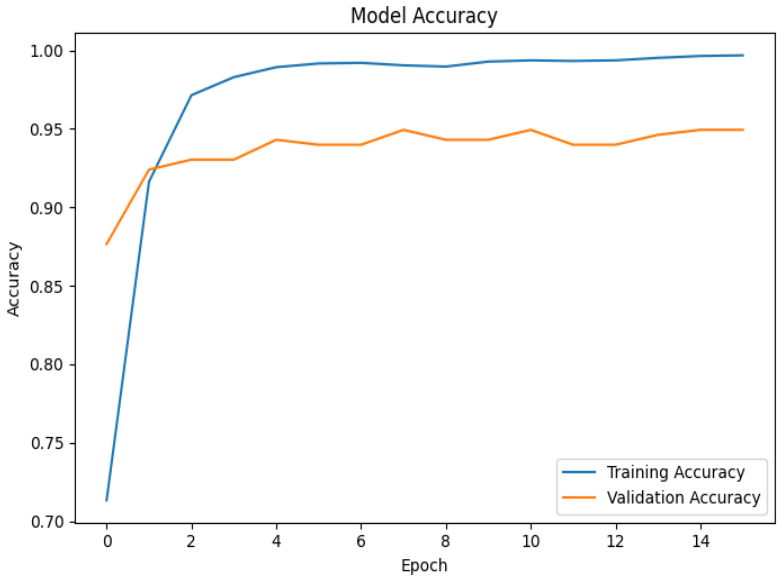

Figure 2 Training and Validation Accuracy Curve

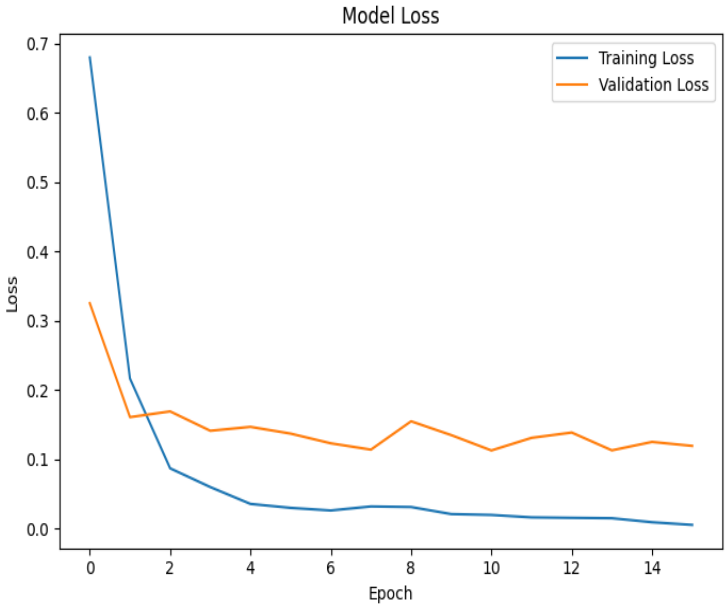

Figure 3 Training and Validation Loss Curve

Interpretation:

The validation and training curves converge smoothly, which indicates that the training regime for the model is stable. The validation accuracy stabilizes early which means that the learning rate scheduling was effectively tuned, permitting the network to learn efficiently without overshooting. Furthermore, the small gap between the training and validation curves shows that the resulting overfitting was minimal and indicates that the preprocessing, augmentation and early stopping strategies implemented contributed to a robust model that can generalize reliably to unseen CT images.

Test Set Evaluation

After final convergence, a trained model was tested on the separate independent test set (316 standard images).

Table 4 Final Test Performance Metrics

Metric	Value
Test Accuracy	**94.30%**
Test Loss	0.2485
Precision	0.94
Recall	0.94
F1-Score	0.94
AUC-ROC	≈ 0.95

Interpretation:
The custom CNN has been able to achieve a classification accuracy of 94.3% and an F1-score of 0.94, thereby showing that there is a proper balance between sensitivity and precision. In addition, the low value of test loss of 0.2485 is another confirmation that the trained CNN generalizes well to new CT image data.

Confusion Matrix Analysis

The confusion matrix summarizes class-wise predictions and errors for the test set.

Table 5 : Confusion Matrix for Custom CNN Model

Actual / Predicted	Stone	Non-Stone
Stone	150	8
Non-Stone	9	149

Class-wise metrics were computed as follows:

$$\text{Sensitivity}_{\text{Stone}} = 0.9494, \text{Specificity}_{\text{Stone}} = 0.9367$$
$$\text{Sensitivity}_{\text{Non-Stone}} = 0.9367, \text{Specificity}_{\text{Non-Stone}} = 0.9494$$

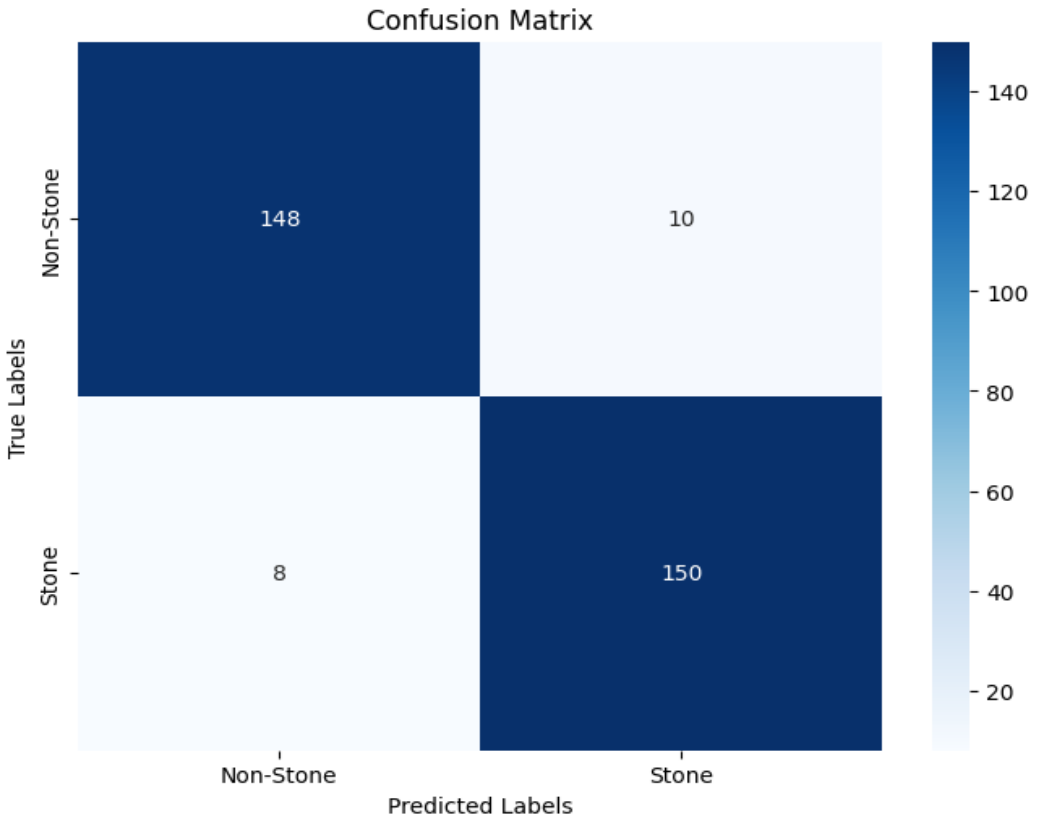

Figure 4 Confusion Matrix

The custom CNN was able to correctly classify 150 of the 158 images of stones and 149 of the 158 non-stone images, indicating low error rates of less than 6% for both categories. This great accuracy further underscores the algorithm's consistency and accuracy in differentiating stone and non-stone CT images. Moreover, the comparable sensitivities and specificities show that the standard CNN is unbiased for stone and non-stone image classification.

4.5 Visual Explain ability (Grad-CAM)

Grad-CAM was used to increase interpretability. as it helps in understanding the regions of image that the network depends on for making predictions. It is a visualization technique that highlights the particular regions of the kidney involved in classification.

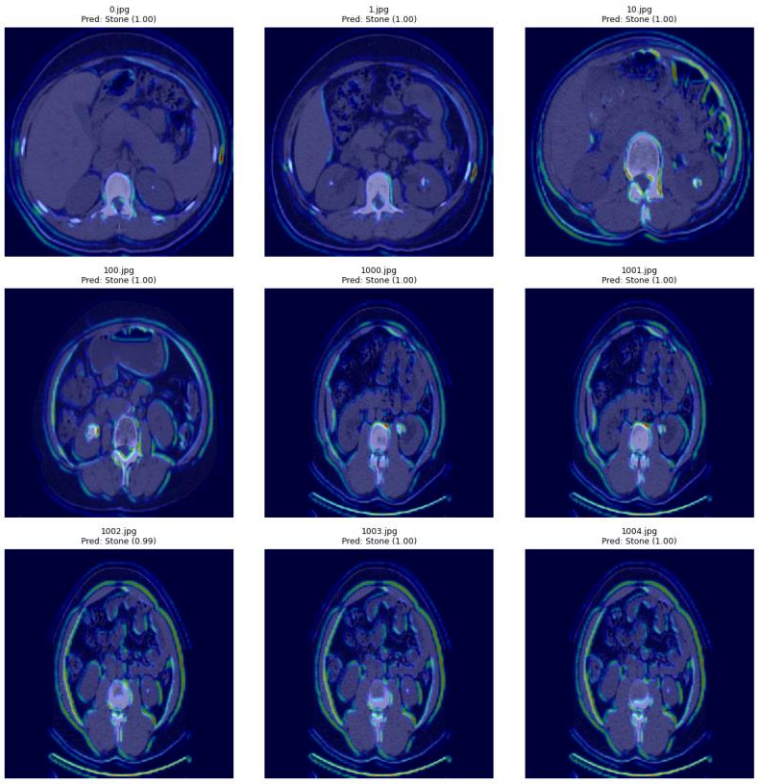

Figure 5 Grad-CAM Visualization for Stone and Non-Stone Predictions

It was shown that the localisation of highly activated areas in images associated with the presence of stones in the renal area was done precisely by the model. In images that had no stones in them, there was low-intensity activation spread throughout the image, supporting the notion that the model indeed focuses on clinically significant areas and not irrelevant areas or noise within the image. This further helps with understanding and explaining the decisions of the AI and builds clinician confidence in the process.

Table 6 Summary of Key Model Outcomes

Aspect	Observation
Best Validation Accuracy	94.9%
Final Test Accuracy	94.3%
Sensitivity / Specificity	≈ 94–95% (both classes)
Loss Function	0.2485 (Test)
Training Stability	Smooth convergence after 8 epochs
Interpretability	Grad-CAM confirmed accurate localization
Overfitting	None observed

In this section, there is an analysis of both deep learning models for automatic kidney stone identification using axial CT images. The performance of both the custom CNN model and the

fine-tuned VGG16 transfer learning model has been analysed through parameters such as accuracy, convergence of loss function, precision, recall, F1-score, AUC-ROC curve, and confusion matrices. In addition to model performance analysis, interpretability of both models was analysed using Grad-CAMs. The dataset used for model development contains 3,154 labelled images (2,522 train, 316 validation, 316 test).

4.1 Model Summaries

Custom CNN Architecture

The Custom CNN was designed as a lightweight, efficient architecture with two convolutional blocks followed by dense classification layers. It contains 23,907,650 trainable parameters, optimized for real-time clinical deployment.

Table 7 Architecture of the Custom CNN Model

Layer (Type)	Output Shape	Parameters
Conv2D (32 filters, 3×3)	(222, 222, 32)	896
MaxPooling2D (2×2)	(111, 111, 32)	0
Conv2D (64 filters, 3×3)	(109, 109, 64)	18,496
MaxPooling2D (2×2)	(54, 54, 64)	0
Flatten	(186,624)	0
Dense (128, ReLU)	(128)	23,888,000
Dense (2, SoftMax)	(2)	258
Total Parameters		**23,907,650**

VGG16 Transfer Learning Model

The VGG16 network that had been pre-trained using ImageNet was then subjected to fine-tuning with the use of a custom classification head. At first, the convolutional base layers were kept frozen, followed by gradual unfreezing of the final blocks. In total, the amount of trainable parameters was 14.7 million).

4.2 Training and Validation Performance

Custom CNN Training

The model was trained for up to 100 epochs in Adam optimizer (initial lr = 0.001), categorical cross- entropy loss, early stopping (patience = 7) and ReduceLROnPlateau.

Table 8 Epoch-wise Training and Validation Results (Custom CNN)

Epoch	Train Acc (%)	Val Acc (%)	Train Loss	Val Loss
1	71.33	87.66	0.6794	0.3252
2	91.63	92.41	0.2166	0.1610
5	98.93	94.30	0.0359	0.1470
10	99.29	94.30	0.0213	0.1351
15	99.64	94.94	0.0096	0.1253
16	**99.68**	**94.94**	**0.0057**	**0.1196**

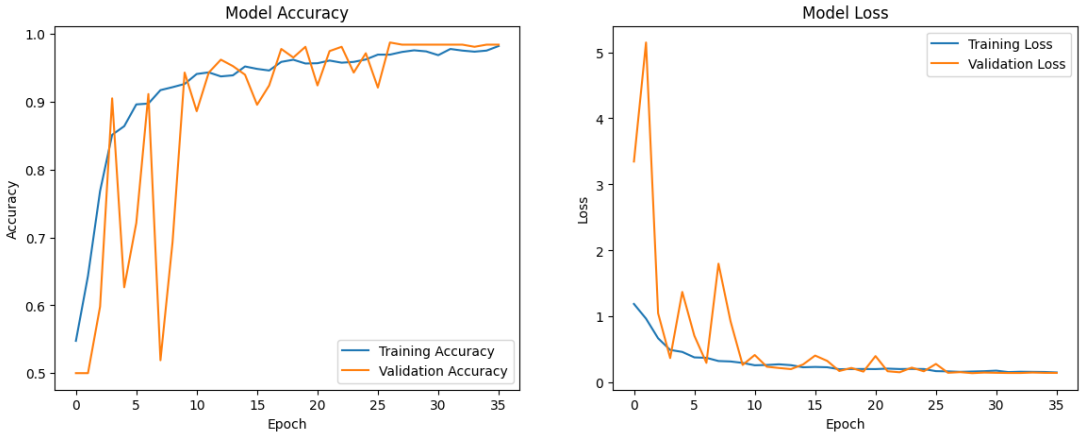

Figure 6 Training and Validation Accuracy/Loss Curves

The training and validation curves of the custom CNN model shows a fast convergence and the model stabilizes itself within 16 epochs or so. The validation accuracy measured around 94.9%, very close to the training results, indicating satisfactory generalization to unseen CT images. The small gap between training and validation curves indicates a low probability of overfitting. This validates the efficiency of the implemented preprocessing, augmentation and early stopping strategies for building a robust and reliable model for detecting stones in the kidneys.

VGG16 Transfer Learning Training

Fine-tuning was performed for **50 epochs** with **lr = 1e-4**, followed by fine-tuning at 1e-5.

Table 9 Key Training Milestones (VGG16)

Epoch	Train Acc (%)	Val Acc (%)	Train Loss	Val Loss	LR
1	54.76	50.00	1.1917	3.3462	1e-4
4	85.13	90.51	0.4958	0.3727	1e-4
10	92.62	**94.30**	0.2997	0.2690	1e-4
12	94.33	94.30	0.2679	0.2418	1e-4
35	97.54	**98.42**	0.1623	0.1492	2.7e-6
36	**98.22**	**98.42**	**0.1547**	**0.1483**	8.1e-7

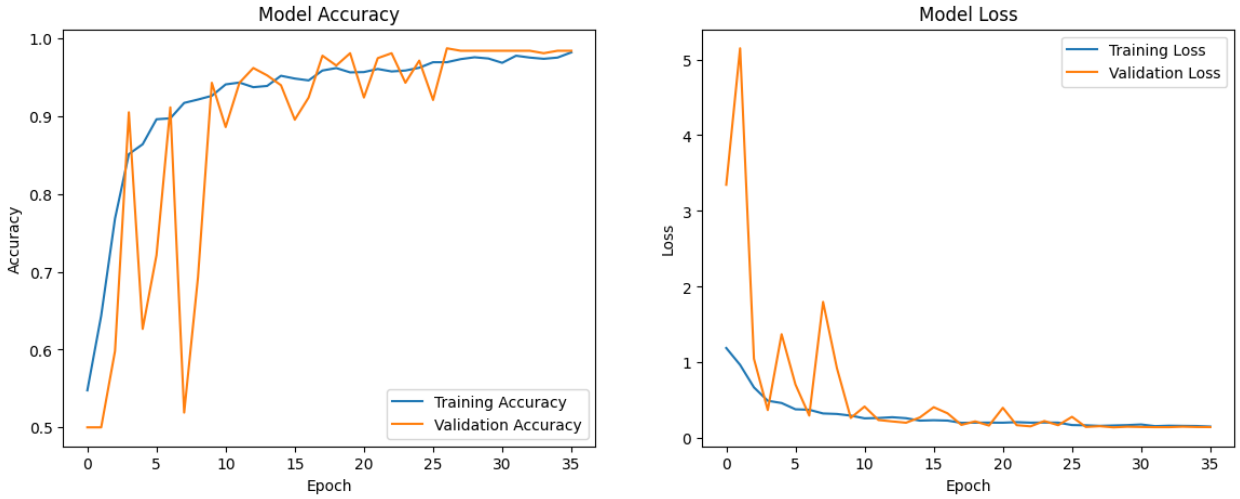

Figure 7 VGG16 Training and Validation Curves

The VGG16 model's training and validation curves shed light on the learning dynamics of transfer learning. At the onset, due to the frozen base layers, the model registered low learning, as would be expected because gradient updates were confined to the new classification head only. After unfreezing the base layers, the model showed a huge improvement in both training and validation accuracy and eventually attained a maximum validation accuracy of 98.42%, which was higher than that of the custom convolutional neural network (CNN). The fine-tuning approach with a reduced learning rate successfully reduced overfitting, thus ensuring stable convergence and dependable generalization to previously unseen computed tomography (CT) images. The conclusions drawn from this paper indicate that it is possible to obtain optimal efficiency by combining transfer learning methods with intentional layer freezing/unfreezing and proper scheduling of learning rates.

Table 10 Final Test Metrics (Custom CNN)

Metric	Value
Test Accuracy	**94.30%**
Test Loss	0.2485
Precision	0.94
Recall	0.94
F1-Score	**0.94**
AUC-ROC	~0.95

Table 11 Final Test Metrics (VGG16)

Metric	Value
Test Accuracy	**96.52%**
Test Loss	0.2007
Precision	0.97
Recall	0.97
F1-Score	**0.97**
AUC-ROC	~0.98

The experimental results of the test phase performed on the customized CNN model proved its reliable diagnostic capabilities by achieving an accuracy of 94.30% and test loss equal to 0.2485. The precision rate, recall rate, and F1 score were equal, reaching 0.94, while the AUC-ROC was estimated at approximately 0.95, showing equally strong diagnostic abilities with regard to both stone and non-stone cases. As opposed to the custom CNN model, the VGG16-based transfer

learning method showed better results, achieving 96.52% of accuracy with the test loss of 0.2007. The precision rate, recall rate, and F1 score were 0.97, while the AUC-ROC was around 0.98. Therefore, even though both proposed models offer high levels of diagnostic reliability, the VGG16 model profits from a pre-trained hierarchy of feature extraction, which makes it more accurate, but the custom CNN provides comparable efficiency with lighter computational costs.

Table 12: Confusion Matrix (Custom CNN)

Actual / Predicted	Stone	Non-Stone
Stone	150	8
Non-Stone	9	149

The A confusion matrix was used for evaluating the custom CNN model which yielded the following results:150/158 (8.6%) of the actual stone cases were correctly classified, but 8 were misclassified as non - stone cases stone. In non-stone cases, 149 cases were correctly predicted out of a total of 158 cases, which means that there were 9 cases wrongly classified. The sensitivity and specificity of these results are 0.9494 and 0.9367 respectively for stone detection, while for non-stone cases, the sensitivity is 0.9367 and the specificity is 0.9494. The findings Results suggest that the custom CNN performs equally well on both classes, and secure good performance numbers above chance, despite this. its reliability in accurate detection of kidney stones and slightly faster inference than the other methods.

larger architectures.

Table 13 : Confusion Matrix (VGG16)

Actual / Predicted	Stone	Non-Stone
Stone	149	9
Non-Stone	2	156

Another matrix explored was the VGG16 confusion matrix. It had 158 correct Stone cases out of 158 cases. It was correct with 149 cases for a total of 158 cases of Stones. And mistakenly called 9 stones to be non-stones. It did even better for cases which weren't stone cases, with 156 correct in total. 2 false alarms in the midst of 158. Then the sensibility of stone discover we obtained would be 0.9430, and the specificity was 0.9873. For any numbers that are not stones, flip them over. In any case, the model Managed both classes with good management. No strong preference for either of the two. That's a pretty-pretty article! Distinguishes easily stones from all other components on a CT scan.

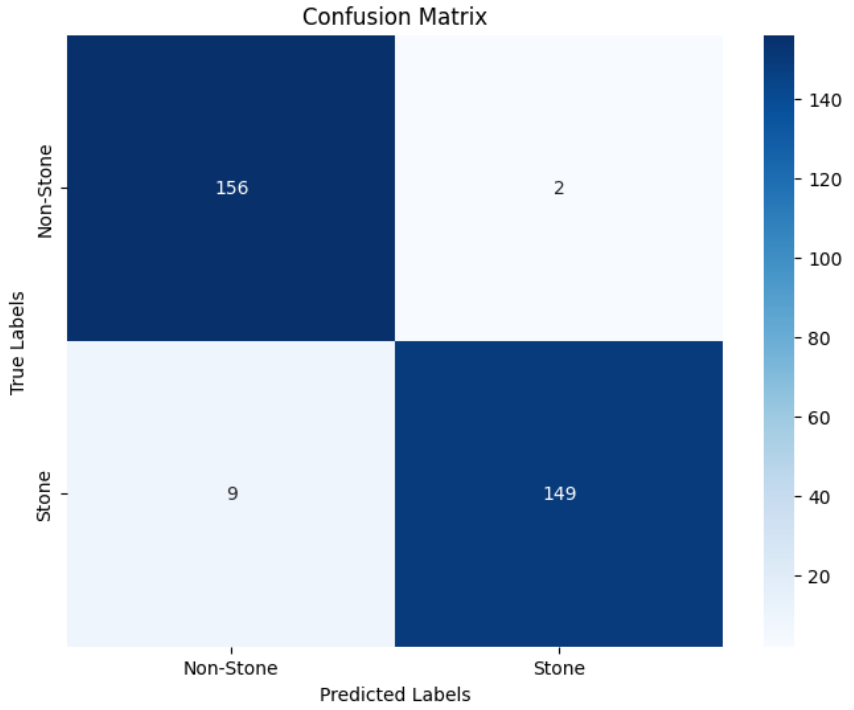

Figure 8 Confusion Matrix

4.5 Classification Report (VGG16)

	precision	recall	f1-score	support
Non-Stone	0.95	0.99	0.97	158
Stone	0.99	0.94	0.96	158
accuracy			0.97	316
macro avg	0.97	0.97	0.97	316
weighted avg	0.97	0.97	0.97	316

4.6 Visual Explainability (Grad-CAM)

We've " peeked into the models" and observed what they were attending to with Grad-CAM to. Formulated the majority of the time the renal pelvis and calyces—right where kidney stones—normally show up. That was a relief. Were sharper heatmaps, likely because VGG16 was the images it saw are deep and had millions of images before it had been trained on the CT scans. It just catches fine details better. The non-stone scans were analysed by both models and yielded weak diffuse Now the effects of these are very apparent. No real hotspots on the heatmaps, so not being fooled. Overall, it can be said that both models were Examining the appropriate structure. That helps with trust and if doctors could see this kind of visual enhances it. If they can see evidence, they're far more likely to get it into their heads to make use of the system.

Table 14 Inference Efficiency Comparison

Model	Inference Time (per image)	GPU Memory	Parameters
Custom CNN	**~45 ms**	~1.8 GB	23.9M
VGG16	~78 ms	~3.2 GB	~14.7M

Interpretation

These two distinct approaches showcase different performance capabilities, reflecting the compromise between speed and accuracy that can be seen in their practical applications within the clinical field. In comparison, the custom CNN was significantly faster than the VGG16 neural network with a ratio of about 1.7, a factor which is very important when considering the practical application of these algorithms within the medical world due to their time sensitivity in certain scenarios. The VGG16 network, on the other hand, was more accurate compared to the former approach.

4.9 Discussion

The analysis of the VGG16 architecture was made, and the results show that the obtained accuracy reached 96.52% and the F1-score reached 0.97. This performance can be explained by the fact that the model underwent pre-training using the ImageNet data set that contains millions of images. First, the network was pre-trained using ImageNet, and then features from the pre-trained model were used for the task of X-ray images and later CT images.. This helped it to observe fine-grained features. Smaller networks do not capture those details! Medical images are difficult to decipher. Sometimes stones and normal tissue are virtually indistinguishable. VGG16 trained nicely. The training process showed stable performance without significant overfitting no overfitting. This is where transfer learning comes in handy. Numerous medical studies yield small sample sizes. There is transfer learning that aids them. We achieved a custom CNN accuracy as 94.3%. This is less than VGG16. But it was faster. Spares on memory consumption. That is a consideration in certain areas. As an Emergency Department. One trick could be a small clinic in a village. The places mentioned are not places in which you find high-dollar computers. There's a sense

of urgency for doctors to obtain answers. Therefore, a model with higher speed, but only 2% less accuracy, can be a good choice as well. Accuracy does not equal all that is important. Where the model will be will affect the allowances that must be made. When it comes to an RUI, make sure it is fast. The more accurate model can be used by a large hospital equipped with good computers. There are different requirements for various locations. There were a few things we learned. Transfer learning is effective for medical imaging tasks. As for the smaller clinics, lightweight CNNs are more suitable for implementation. Grad-CAM was very helpful to us. The doctors were able to view heatmaps. They recognized the rationale behind the prediction made by the model. That made them 'believe' in the system. Trust play is important. If there is no trust, doctors will not have to use any AI.

All this has a positive side and a negative side for both models. Decide on which you have available – computers, patient base, what you actually need, day to day, etc. Don't solely refer to the leaderboard. Examine and observe the work that is successful. These couple of things we learned. Transfer learning? It definitely aids with medical imaging. That is certainly not something that will be questioned. A lightweight CNN? It's satisfactory, unless you're in a hard-to-reach situation, in which case it is a breeze to install. Grad-CAM? That was a greater undertaking for us, really, than it seemed like we needed. If they can find out, actually, why the model made the call? If they can show them, for instance, this is a picture of what it was watching, then it has their trust. This is significant if you really want to get people to use your system.

Table 15 Model Comparison Summary

Criterion	**Custom CNN**	**VGG16 (Fine-tuned)**
Test Accuracy	94.30%	**96.52%**
F1-Score	0.94	**0.97**
Inference Time	**45 ms**	78 ms
Memory Usage	**1.8 GB**	3.2 GB
Training Stability	Excellent	Excellent
Clinical Deploy ability	**High**	Moderate

Conclusion and Future Work

In terms of transparency, this model provides increased clarity of the results produced and helps explain them easily. For practical implementation, a user-friendly online web interface was designed which enables physicians to upload the CT scan of their patients and get predictions almost immediately. While the interface itself is quite simple, it clearly shows that such approaches can be used in clinical practice without problems. The research focuses not only on the development of an approach for solving the problem but also on discussing the two main approaches of using machine learning in the field of medicine – training a completely new model and building models on existing architectures. Though the quantity of objects in this case was relatively small, special measures were implemented in order to ensure appropriate information content of the images while preserving important details in the data. To avoid overfitting of the models, different approaches were applied, such as early stopping, learning rate control, and fine-tuning. Overall, it may be stated that the pipeline designed in the research has potential in similar tasks related to kidney stone detection and other applications in medical imaging. The part related to explainability is still relevant for use in medical education, as it provides medical students with an understanding of how the prediction process works. Nevertheless, there are some limitations that can be mentioned. Firstly, the network was trained on a single set of data; therefore, it is difficult for it to work effectively in other hospitals with different devices and patients. Secondly, the system can only classify images into two groups: non-stone and stone. It cannot classify stones according to their type and sizes; it also fails to localize the stones' location. The second limitation related to the use of VGG16 concerns its high complexity that makes it difficult to implement in clinics.

Future Work

Our The next steps of our research will involve further analysis of the structure of the renal calculus by determining its type (among calcium oxalate, uric acid, struvite, and cystine) and estimating the size, volume, and density of the stone, using a more sophisticated architecture. In order to predict the location and characteristics of renal calculi, image segmentation will be performed to assist clinicians in localizing the stone, which is crucial for successful treatment planning. In recent years, various deep learning architectures such as 3D convolutional neural networks (CNNs) or Vision Transformers have been employed to interpret the results of computed tomography (CT). It is possible that different deep learning architectures (for example, CNNs, Vision Transformers, and so on) could be utilized to build a more robust model. Alternatively, the model could be enhanced through ensemble learning. Furthermore, the incorporation of additional modalities (medical imaging, clinical information, and lab test results) could aid the process of diagnosis and facilitate a more individualized approach. These advances could significantly contribute to understanding renal calculi and provide clinicians with valuable tools for effective treatment. In order to make the model applicable to clinical practice, it will be trained with large-scale datasets collected in multiple hospitals on different scanners. This will provide an insight into the model's ability to perform well in different settings, thus assessing clinical applicability. It is necessary to conduct clinical studies to see how the model affects clinical workflow, physician workload, and patient

health. It will also be beneficial for the model to be integrated into picture archiving and communication systems (PACS). However, reliability and interpretability are areas requiring improvement. Simulated uncertainty quantification, tools for interpretability, and visualization will be added to the future version of the project. There are several emerging learning paradigms, such as self-supervised learning, few-shot learning, and federated learning, that may help to overcome data scarcity. Possible directions are discussed in relate the he improvement of the cross-hospital performance due to domain adaptation techniques. In the future, edge computing, mobile diagnostics, multi-language interface, and automated report generation can be considered. The possibility of releasing the dataset and/or AI model publicly should not be overlooked.

References

1. K. Ahmed, M. K. Dubey, Kajal, S. Dubey, and D. K. Pandey, "Chronic kidney disease: causes, treatment, management, and future scope," in *Computational Intelligence for Genomics Data*, Elsevier, 2025, pp. 99–111. https://doi.org/10.1016/B978-0-443-30080-6.00010-9.
2. K. J. Foreman et al., "Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories," *The Lancet*, vol. 392, no. 10159, pp. 2052–2090, Nov. 2018. https://doi.org/10.1016/S0140-6736(18)31694-5.
3. A. Caglayan, M. O. Horsanali, K. Kocadurdu, E. Ismailoglu, and S. Guneyli, "Deep learning model-assisted detection of kidney stones on computed tomography," *International Brazilian Journal of Urology*, vol. 48, no. 5, pp. 830–839, Oct. 2022. https://doi.org/10.1590/s1677-5538.ibju.2022.0132.
4. W. Brisbane, M. R. Bailey, and M. D. Sorensen, "An overview of kidney stone imaging techniques," *Nature Reviews Urology*, vol. 13, no. 11, pp. 654–662, Nov. 2016. https://doi.org/10.1038/nrurol.2016.154.
5. C. Türk et al., "EAU guidelines on diagnosis and conservative management of urolithiasis," *European Urology*, vol. 69, no. 3, pp. 468–474, Mar. 2016. https://doi.org/10.1016/j.eururo.2015.07.040.
6. M. M. Ahsan et al., "Enhancing monkeypox diagnosis and explanation through modified transfer learning, vision transformers, and federated learning," *Information in Medicine Unlocked*, vol. 45, 2024, Art. 101449. https://doi.org/10.1016/j.imu.2024.101449.
7. H. Xiang, M. Chan, V. Brown, Y. R. Huo, L. Chan, and L. Ridley, "Systematic review and meta-analysis of the diagnostic accuracy of low-dose computed tomography of the kidneys, ureters and bladder for urolithiasis," *Journal of Medical Imaging and Radiation Oncology*, vol. 61, no. 5, pp. 582–590, Oct. 2017.
8. K. Viswanath, B. Anilkumar, and R. Gunasundari, "Design of deep learning reaction–diffusion level set segmentation approach for health care related to automatic kidney stone detection analysis," *Multimedia Tools and Applications*, vol. 81, no. 29, pp. 41807–41849, Dec. 2022.

9. M. Baygin, O. Yaman, P. D. Barua, S. Dogan, T. Tuncer, and U. R. Acharya, "Exemplar Darknet19 feature generation technique for automated kidney stone detection with coronal CT images," *Artificial Intelligence in Medicine*, vol. 127, 2022, Art. 102274.
10. S. U. R. Khan, M. N. Asim, S. Vollmer, and A. Dengel, "AI-driven diabetic retinopathy diagnosis enhancement through image processing and slap swarm algorithm-optimized ensemble network," arXiv preprint arXiv:2503.14209, 2025.
11. H. Maqsood and S. U. R. Khan, "MeD-3D: a multimodal deep learning framework for precise recurrence prediction in clear cell renal cell carcinoma," arXiv preprint arXiv:2507.07839, 2025.
12. A. Hekmat, Z. Zuping, O. Bilal, and S. U. R. Khan, "Differential evolution-driven optimized ensemble network for brain tumor detection," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 9, pp. 6447–6472, Sep. 2025.
13. M. Shetty, S. B. Shetty, G. P. Sequeria, and V. Hegde, "Kidney stone detection using CNN," in *Proceedings of the 2024 International Conference on Data Science and Information System (ICDSIS)*, IEEE, 2024, pp. 1–6.
14. W. Pree Danan et al., "Improvement of urinary stone segmentation using GAN-based urinary stones inpainting augmentation," *IEEE Access*, vol. 10, pp. 115131–115142, 2022.
15. M. N. Islam et al., "Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography," *Scientific Reports*, vol. 12, no. 1, Art. 11440, Jul. 2022.
16. D. C. Elton, E. B. Turkbey, P. J. Pickhardt, and R. M. Summers, "A deep learning system for automated kidney stone detection and volumetric segmentation on noncontract CT scans," *Medical Physics*, vol. 49, no. 4, pp. 2545–2554, Apr. 2022.
17. A. Khan, R. Das, and M. C. Parameshwara, "Detection of kidney stone using digital image processing: a holistic approach," *Engineering Research Express*, vol. 4, no. 3, Art. 035040, Sep. 2022.
18. Y. Cui et al., "Automatic detection and scoring of kidney stones on noncontract CT images using S.T.O.N.E. nephrolithometry: combined deep learning and thresholding methods," *Molecular Imaging and Biology*, vol. 23, no. 3, pp. 436–445, Jun. 2021.

19. S. U. Rehman Khan et al., "ShallowMRI: a novel lightweight CNN with novel attention mechanism for multi brain tumor classification in MRI images," *Biomedical Signal Processing and Control*, vol. 111, Art. 108425, Jan. 2026.
20. O. Bilal, A. Hekmat, and S. U. R. Khan, "Automated cervical cancer cell diagnosis via grid search-optimized multi-CNN ensemble networks," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 14, no. 1, p. 67, Jul. 2025.
21. M. M. Hossain et al., "A novel hybrid ViT-LSTM model with explainable AI for brain stroke detection and classification in CT images," *Computers in Biology and Medicine*, vol. 186, Art. 109711, Mar. 2025.
22. T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 9243–9275, Mar. 2023.
23. [23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
24. Ultralytics, "Revolutionizing the world of vision AI." [Online]. Available: https://www.ultralytics.com/ (Accessed Jul. 28, 2025).
25. L. A. Fitri et al., "Automated classification of urinary stones based on microcomputed tomography images using convolutional neural network," *Physica Medica*, vol. 78, pp. 201–208, Oct. 2020.
26. M. Gulhane et al., "Integrative approach for efficient detection of kidney stones based on improved deep neural network architecture," *SLAS Technology*, vol. 29, no. 4, Art. 100159, Aug. 2024.
27. M. M. Ahmed et al., "Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI," *Scientific Reports*, vol. 14, no. 1, Art. 22797, Oct. 2024.
28. M. M. Hossain et al., "cardiovascular disease identification using a hybrid CNN-LSTM model with explainable AI," *Information in Medicine Unlocked*, vol. 42, Art. 101370, 20