International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(18s) 2026 pp. 1003-1018
https://doi.org/10.70917/ijcisim-2026-4938
Article

FedStoneNet-Hybrid: A Privacy-Preserving Federated CNN–Transformer Framework for Kidney Stone Detection from CT Imaging

Guda Madhu[1], Nirmalajyothi Narisetty[2]

[1]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad -500043, Telangana, India.
[1]Email id: madhuguda86@gmail.com ORCID 0009-0004-7864-9675.
[2]Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad -500043, Telangana, India.
[2]Email id: nirmala.narisetty@gmail.com, ORCID ID: 0000-0002-4810-9676.

Abstract: One of the most frequent urological issues that calls for accurate, timely diagnosis and appropriate treatment is renal stone disease, and CT is very helpful in diagnosing renal stones. Methods of deep learning possess effective automated detection potential. However, most prevailing methods involve centralized training, raising privacy and multi-institutional collaboration concerns. In response to these challenges, this research presents FedStoneNet-Hybrid, a federated learning framework designed to protect the privacy of kidney stone detection on distributed CT data using a hybrid CNN and transformer. Under the proposed scheme, each institution will locally train its hybrid model, where CNN layers will capture a fine-grained spatial representation and transformer modules will capture the global contextual relationship of pollution data. The Federated Averaging algorithm aggregates only the model parameters at the central server rather than the raw medical data. Thus, safe and scalable collaborative learning can be enabled. As observed from the experimental results, the proposed model achieves good performance with an accuracy of 97.2%, sensitivity of 96.8%, specificity of 96.9%, F1 score of 0.96.85, and AUC of 0.991. The model shows solid calibration with a low Expected Calibration Error and Brier Score. Based on these findings, the suggested framework serves as an accurate, scalable, and privacy-preserving approach for distributed medical imaging.
Keywords: Kidney stone detection; Federated learning; Hybrid CNN–Transformer; Computed tomography (CT); Medical image analysis; Privacy-preserving AI; Distributed learning; Deep learning; Calibration; Clinical decision support

1. Introduction

Kidney stones are a common urological condition affecting a significant number of people worldwide. It often requires a timely diagnosis to assess the harms associated with disease progression [1]. The gold standard for detecting kidney stones is computed tomographic (CT) imaging due to its good resolution and diagnostic accuracy [2]. Inter-observer variability can affect diagnostic consistency in clinical practice, particularly with manual interpretation of CT scans, which is time-consuming [3].

In the recent past, deep learning (DL) techniques have been used in the automated analysis of medical images with efficiency and improved accuracy [4]. Convolutional neural networks (CNNs) perform well in learning local spatial features from CT images, while attention-based transformer architectures can learn global contextual relationships [5]. Despite these advancements, most existing approaches rely on a centralized training process. Patients' privacy, data protection, and regulatory compliance are serious issues with such strategies, especially in health care [6].

To mitigate these concerns, federated learning (FL) was proposed to train models in a distributed manner without sharing raw data. In this framework, organizations train models locally, and only the model parameters are exchanged and aggregated at a central server. Federated learning is invoked for data privacy and collaborative learning, but the present FL-based approaches rely on conventional deep architectures [8]. They do not take into account the complementary strengths of various models.

Also, issues concerning the heterogeneous distribution of data (non-IID data), lack of generalization across institutions, and the absence of reliable confidence estimation remain unresolved. Due to the limitations of current methods, there is a need for an integrated framework which can offer high diagnostic accuracy, good generalization and privacy protection at the same time.

To overcome these challenges, researchers propose FedStoneNet-Hybrid, a new federated learning framework that employs a hybrid CNN-Transformer architecture to detect kidney stones from distributed CT imaging data. Specifically, the CNN modules extract fine-grained local features while the transformer modules capture global contextual dependencies in the proposed method. Using this hybrid architecture for federated learning, contributions from a number of sites are accomplished without compromising privacy.

2. Related Works

Kidney stone disease is a common urological disease whose timely diagnosis is made by a CT scan. In recent years, several deep learning (DL) techniques have shown promise in automating kidney stone detection, which can enhance diagnostic accuracy and ease the burden of diagnosis in clinical practice [9] [10]. Numerous strategies have been proposed for addressing the challenges of feature representation, data heterogeneity, and privacy preservation. Various types of AI models are available for image segmentation tasks. Recently, there have been varied developments in image processing. Nonetheless, such schemes are increasingly facing scalability, computational efficiency, and safe multi-organization deployment challenges [11]. This section undertakes a literature review and identifies gaps in the research framework proposed in the present study.

Ye, Yuguang et al. (2026) [12] designed a multimodal deep learning framework by leveraging CT images and clinical data for gallbladder stone diagnosis. Integrating imaging features with patient-related clinical information enhances diagnostic accuracy and robustness. Although it functions in multi-institutional healthcare environments, it does have privacy challenges. This highlights the need for decentralized methods such as Federated Learning.

Bharathi, V. et al. (2026) [13] examined the use of federated learning (FL) for kidney stone detection with privacy. The FL study verified that the collaborative model can train on distributed datasets without sharing raw medical data while keeping patient data private. Their framework successfully tackled privacy problems, but like other works, it was also based on classical deep learning architectures and did not exploit the full potential of hybrid architectures.

The aim of the research presented by Lilhore, Umesh Kumar et al. (2026) was to develop a hybrid quantum-inspired transformer network to predict kidney stones from CT images and urine data. By dynamically summing and selecting routes, our attention method achieved good generalization and linear complexity. Despite its demonstrative performance, the methodology is highly computationally complex, which, lacking a distributed learning framework, is not scalable in real-world clinical settings.

Diab, Nabil et al. (2026) [15] developed a kidney disease diagnosis strategy using federated optimization. A federated framework can use a Mayfly optimization algorithm to train the model while preserving privacy. Despite this, the study focuses on structured clinical data, not imaging modalities, and is therefore less suitable for CT-based kidney stone detection tasks.

Pimpalkar, Amit et al. (2025) [16] have proposed fine-tuned deep learning models for the classification of kidney disease using CT images, and used an optimized CNN architecture to achieve high accuracy. In privacy-sensitive contexts, an application for the data training process was not possible, though it was centralized. Further, it did not consider the data heterogeneity in different institutions. Panchal, Neha et al. (2025) [17] proposed a deep learning-based kidney stone detection model using CT scan images. The study showed CNN-based approaches can learn discriminative features for the identification of stones. The model runs on a centralized dataset, which does not accommodate variability across multi-institutional data sources, restricting generalisation.

Rao, Neha J. et al. (2025) [18] proposed a deep-learning framework for determining and classifying the clinical severity of kidney stones in two stages. First, it detects the presence of stones and then classifies the severity level appropriately. Even though the framework works well, it is computationally intensive and relies on centralized training.

Bala, Saroj, et al. (2025) [19] utilized YOLOv10 and channel attention for better detection of kidney stones. Their model improved localization and detection accuracy by focusing on salient CT regions. Such detection models, though, may fail with small or low-contrast stones and require large annotated datasets.

Model training largely depends on the availability of data. Abdalla, Peshraw Ahmed, et al. (2025) [20] created a freely accessible CT picture database to enable kidney stone diagnosis. These databases enable easy benchmarking and reproduction of results in AI-related research. Nevertheless, the dataset itself doesn't solve privacy issues; sharing norms prevent the exchange of data between institutions.

The study by Vasudeva, Nisha, et al. (2025) [21] reviewed artificial intelligence-based radiographic imaging techniques for diagnosing kidney stones, providing a broader perspective on the field. According to their analysis, deep learning models are becoming increasingly important for enhancing diagnostic reliability, although they also mention challenges such as data heterogeneity, model interpretability, and clinical implementation.

An ensemble deep learning technique that enhances classification performance with different late fusion models has been proposed by Rathva, Varsha, et al. (2025) [22]. Ensemble approaches are computationally more expensive and can make the model more complex. Hence, due to limiting variance and enhancing robustness, they are less suited for real-time or distributed deployment.

Researchers also examined mixed framework structures. Amolkumar N. Jadhav Developing a CNN-LSTM model, et al. (2025) [23] developed a CNN-LSTM model to anticipate kidney stones. CNNs gather spatial features, while LSTMs learn temporal or sequential patterns in data. In the case of medical images, transformer-based models are more effective than LSTM-based models, which are good at capturing long-range spatial dependencies.

Degadwala, Sheshang, and Rathva, Varsha (2024) [24] provided a comprehensive overview of the ML and DL techniques for kidney stone detection. Their paper reveals that CNNs perform better than feature-based methods, which leads to the introduction of the transition phrase between the two. The paper mentions challenges such as the unavailability of datasets, unbalanced data, and the need for reliable feature representation.

Building on this, Ravi, Vinayakumar, et al. (2024) [25] propose an automatic kidney stone detection system employing a guided bilateral feature detector. The method mainly uses handcrafted features for kidney stone identification in CT images. Feature-based methods tend to be less scalable and adaptable than deep learning, despite being effective in controlled situations for complex heterogeneous datasets.

To overcome these limitations, Fuladi, Siddhesh, et al. (2024) [26] present a CNN-based framework for the detection and treatment of kidney stones. Their model improved performance by automatically learning hierarchical features from CT images. Nevertheless, the method requires centralized training and fails to employ techniques for distributed data and privacy protection.

Asif, Sohaib et al. (2024) [27] suggested a deep learning model based on optimized fusion combining several architectures for detection. Their fusion technique enhances the variety and strength of characteristics, thereby strengthening classification capabilities. Nonetheless, these fusion models may lead to a higher computational cost and limited scalability due to their complexity.

Research Gap

Kidney stone detection has made remarkable advances using deep learning but continues to face some challenges. Many existing approaches use a centralized training scheme, which can lead to privacy leakage. Moreover, these methods may not apply to a multi-institutional healthcare setting [28]. Although CNN-based models can capture local spatial representation, they are unable to model global contextual relations. On the other hand, transformer-based ones are sophisticated, but their complexity is very high [29]. Besides, ensemble and hybrid models increase accuracy but also increase model complexity. They are not suitable for scalable deployment. Moreover, there have been few works on non-identical (non-IID) data distributions in distributed clinical datasets.

To overcome these limitations, propose a FedStoneNet-Hybrid framework that fuses CNN-based local feature extraction with transformer-based global context modelling in a federated learning approach. Through extensive experimentation. The multi-federation framework can enable privacy-preserving, scalable, and robust kidney stone detection across heterogeneous dataset types while maintaining high accuracy and reliable prediction performance. Thus, our work is ready to be deployed in real-world clinical settings.

3. Materials and Methodology

The proposed hybrid model, FedStoneNet, is for detecting kidney stones from CT scan data. It normalizes, resizes, and augments the dataset using the KiTS19 database. A hybrid CNN and Transformer architecture is used to fuse local spatial features with global contexts. The FedAvg model combines the local models' updated parameters for training across multiple nodes. The encoding predicts a joint distribution with disentangled representations. This leads to more understandable outcomes and better calibration of independent and joint prediction. The framework guides scalable and robust privacy-preserving analysis of medical images.

3.1.Design of the dataset and experiments.

The publicly available KiTS19 (Kidney Tumor Segmentation Challenge 2019) dataset was adapted for this study [30]. The dataset has high-resolution CT scans of the abdomen and a kidney segmentation mask dataset to extract ROIs and generate labels. According to the NITF format, the dataset includes approximately 300 patient scans, typically at 512 × 512 resolution. To conduct this research, CT volumes were transformed into 2D slices and annotated for binary classification (stone vs non-stone). The data was split into 70% training, 10% validation, and 20% testing without any patient overlap. To replicate a realistic distributed healthcare scenario, the dataset was divided between 5-10 federation nodes, which represented independent institutions with dissimilar (non-IID) data distributions.

3.2.Data preprocessing

The proposed framework uses the publicly available KiTS19 dataset that contains high-res abdominal CT scans. The dataset was designed for kidney tumor segmentation. This dataset is rich in renal imaging data. It can also be used to create a kidney stone detection task by simply preprocessing the images.

Due to the unavailability of explicit kidney stone annotations, a reliable and reproducible approach has been adopted to generate binary labels (stone vs. non-stone) using a Hounsfield Unit (HU)-based labelling strategy.

Let the input CT slice be represented as in Eq. (1)

$$X \in R^{H \times W} \quad (1)$$

where each pixel $x_{i,j}$ corresponds to an intensity value in Hounsfield Units.

Preprocessing was carried out to standardize the input. The CT intensities were clipped to [-1000, 1000] Hounsfield Units and normalized using min–max scaling. All slices were resized to 224 by 224 pixels.

Gaussian filtering was used to reduce noise and minimize imaging artifacts. Kidney zones were extracted whenever available to focus on relevant structures. To improve generalization and prevent overfitting, techniques such as rotation, horizontal flipping, and contrast were used.

3.3.Feature extraction and model architecture

The FedStoneNet-Hybrid model integrates attention networks into CNNs to capture global and local features, which is beneficial for identifying the 10 most important features, such as color, height, and others, through a series of experiments conducted during training and testing. The CNN backbone retrieves local features such as edges, textures, and structural patterns linked with kidney stones. The features are converted into feature maps and then into token representations, which the transformer module will process. By utilizing multi-head self-attention, the transformer can effectively process long-range dependencies and relationships between different regions.

The hybrid feature representation is defined in Eq. (2)

$$F = Transformer(CNN(X)) \quad (2)$$

Where X is the input CT image, and F is the ultimate feature embedding after classification. The resulting feature vector (about 256–512 dimensions) thus helps with classification. The proposed architecture is shown in Fig. 1.

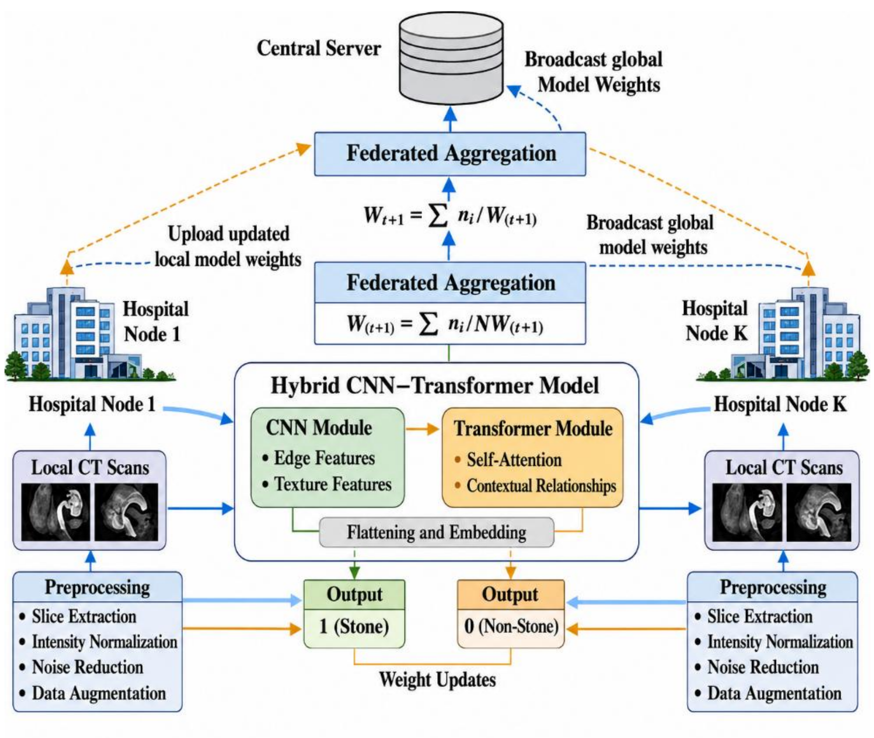

Fig. 1 Architecture of the proposed FedStoneNet-Hybrid framework.

The federated hybrid CNN–Transformer framework for kidney stone detection from distributed CT imaging data. It has three main components. The first one is the local hospital nodes, then a hybrid deep learning model, and finally the central federated server. The preprocessing stage of the local CT scans at each hospital node is responsible for slice extraction, intensity normalization, noise reduction, and data augmentation. This step will make sure that the input is 'standardized' and 'high quality' for model training.

The hybrid CNN-Transformer model is fed with pre-processed images. The CNN module extracts low-level and mid-level features, such as edges and textures related to kidney stones. After that, these features are transformed through self-attention to capture global relationships within the image through the transformer module. The final classification output indicates the presence (stone) or absence (non-stone) of kidney stones based on the combined feature representation.

Each hospital trains the model separately and generates updated model weights using its local data. Key patient information, rather than raw data, is all that is transmitted to the central server to ensure privacy. A global model is built by the central server using the federated aggregation (FedAvg) process. It combines the local model updates of all nodes. The improved global model is then sent back to all hospitals for the following round of training.

This kind of training enables institutions to collaborate with each other to share knowledge. This ensures that the model generalizes better on different sets of data and does not violate privacy constraints. In the federated framework, CNN and Transformer together achieve accurate feature extraction and modelling of the global context. Consequently, this enhances the diagnostic performance.

3.4.Federated learning framework

The model is trained using a federated learning paradigm to protect patient privacy. The participating nodes train a local model with their own private dataset without sharing the raw data. Throughout the various rounds of communication, the locally updated model parameters are transmitted to a central server, which aggregates those parameters using the Federated Averaging (FedAvg) algorithm given in Eq. (3).

$$W^{(1)} = \sum_{i=1}^{K} \frac{n_i}{N} W_i^{(1)} \tag{3}$$

where $W_i^{(1)}$represents the local model weights from node i, $n_i$is the number of samples at node i, and $N$is the total number of samples across all nodes.

The global model is sent back to all nodes to enable collaborative learning at institutions while keeping data private.

3.5.Performance Analysis

The training procedure of the proposed framework is summarized as follows.

Input

- Distributed datasets $D_1, D_2, \dots, D_K$
- Initial global weights $W^{(0)}$
- Communication rounds T

Step 1: **Initialize** global hybrid CNN–Transformer model. $W^{(0)}$

Step 2: **For each round $t = 1$to T:**

a. **Broadcast** $W^{(t)}$to all nodes

b. **For each node i(in parallel):**

- Preprocess local CT images
- Train hybrid model (CNN → Transformer → Classifier)
- Update local weights $W_i^{(1)}$
- Send weights to the server

C. **Aggregate (FedAvg):**

$$W^{(1)} = \sum_{i=1}^{K} \frac{n_i}{N} \cdot W_i^{(1)}$$

Step 3: **Repeat** until convergence

Output

- Final global model $W^{(T)}$

The suggested FedStoneNet-Hybrid Algorithm helps in privacy-preserving collaborative learning among various hospitals using Federated Learning. First, the central server initiates a hybrid CNN-Transformer model to include all nodes. Before merging with other hospitals, each hospital preprocesses its own local CT images. The previously processed data is sent to the hybrid model, where the CNN collects local spatial features while the Transformer forms global contextual relationships. Classification to detect kidney stones (stone vs non-stone) using the model.

Every hospital modifies the weights for its local model utilizing its own dataset. Only the modified weights are sent to the central server instead of raw data to maintain raw data privacy. The server combines these weights to produce a better version of the global model using FedAvg. After that, this revised model is sent back to all hospitals for the next training. The above process is repeated iteratively for several rounds of communication. The model is trained on many different data distributions (non-IID data) while ensuring privacy through collaboration. The last global model is suitable for real-world application scenarios of multi-institutional medical institutions because of its high accuracy, robustness, and generalization.

3.6.Training configuration

The model was developed with standard deep learning frameworks (TensorFlow/PyTorch). Training was performed with the Adam optimizer using a learning rate of 0.0001 and a batch size of 16–32. Each local model completed 30–50 epochs of training and was federated for ~75 communication rounds until convergence. All experiments were performed in the GPU environment.

3.7.Evaluation metrics

To assess the model, several classification and calibration metrics were analysed for the proposed model. Measures of classification performance included accuracy, sensitivity, specificity, F1-score, and Area Under the Curve (AUC) to evaluate class separability. To measure prediction reliability, calibration metrics such as Expected Calibration Error (ECE) and Brier Score. These parameters are key to evaluating confidence estimation for clinical applications.

This methodology proposes hybrid deep learning, along with the federated learning process, for effective and accurate kidney stone detection. The federated scheme allows multiple nodes to extract features using a combination of CNN and Transformer securely. Models learned using non-IID data can generalize better to the real world. The preprocessing and training methods promote stability and convergence. The calibration metrics allow for reliable and clinically meaningful predictions. Overall, it provides an efficient and scalable framework for multi-institutional medical image analysis.

4. Result and Discussion

This section provides a complete analysis of the suggested FedStoneNet-Hybrid with circulated CT plots. The effectiveness of a classifier can be measured using classification metrics, namely accuracy, sensitivity, specificity, F1-score, AUC, and calibration metrics such as Expected Calibration Error (ECE) and Brier Score. The results show that the proposed model performs better compared to the baseline models, namely the CNN, Transformer, and federated CNN models. Brilliant results emerge when combining these strategies. This section also contains ablation studies to verify the contribution of individual components, which consistently improve with the introduction of the CNN, Transformer, and federated optimization. In addition, the federated learning effect is analyzed, which increases scalability, privacy protection, and generalization on non-IID data distributions.

4.1.Experimental setup

The assessment of the FedStoneNet-Hybrid model was performed on CT images of the KiTS19 dataset and TCIA samples. The dataset was partitioned into 70% for training, 10% for validation, and 20% for testing. To mimic a real-world distributed healthcare system, the data were split across 5–10 federated nodes with non-IID distributions.

A hybrid CNN–Transformer model that performs federated averaging to aggregate local parameters was used. There were close to 75 communication rounds of training done.

4.2.Quantitative Performance Evaluation

The performance of the proposed FedStoneNet-Hybrid model through a quantitative assessment using standard classification and calibration metrics. The performance is measured using accuracy, sensitivity, specificity, F1-score, and AUC. Furthermore, evaluate reliability using Expected Calibration Error (ECE) and Brier Score. Results were summarised in Table 1, and Fig. 2 shows the model has high predictive capability and robustness. Metrics for accuracy and confidence in the predictions are to be measured to understand them better.

Table 1. Overall Performance of Proposed Model

Metric	Value (%)
Accuracy	**97.2**
Sensitivity	96.8
Specificity	96.9
F1-Score	96.85
AUC	**0.991**
ECE	0.021
Brier Score	0.034

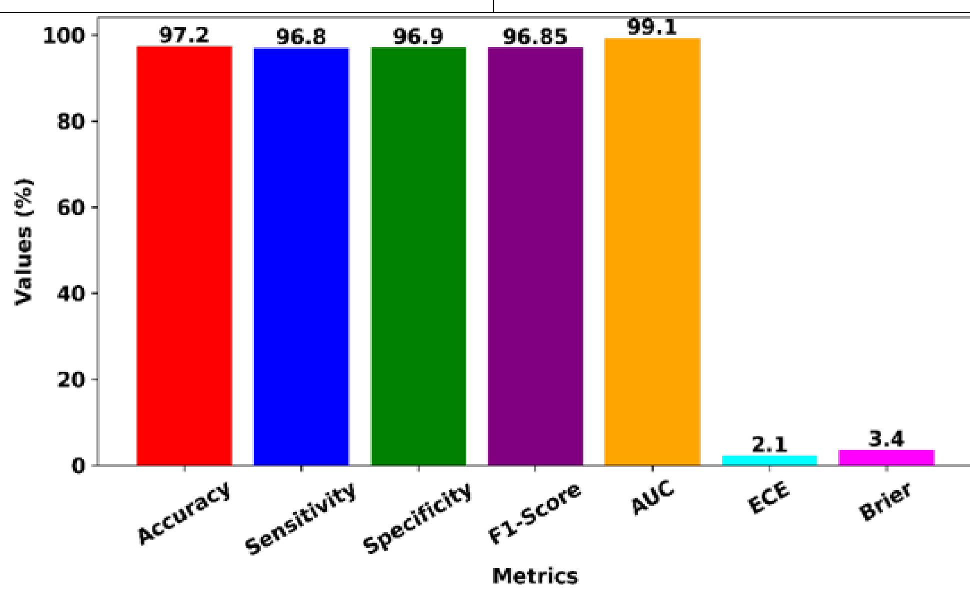

Fig. 2 Performance metrics of the proposed model.

Fig. 2 shows the performance of the proposed FedStoneNet-Hybrid model using classification and calibration metrics. The model achieves high classification metrics, including Accuracy (97.2%), Sensitivity (96.8%), Specificity (96.9%), and F1-score (96.85%). It indicates a strong ability to differentiate between sensitivity and non-stone classes. The model clearly differentiates between classes, as the AUC is 99.1%, and this can be seen across all decision thresholds. In addition to performance, the graph shows the calibration metric. A low ECE (2.1%) and Brier Score (3.4%) indicate that the probabilities have learned to predict the outcome rather accurately and reliably.

4.3.Evaluating Performance with Baseline Models

The proposed model is compared with the baseline CNN, vision transformer, and federated CNN models. Ensuring a fair comparison between the methods is key, which will be evaluated based on accuracy, F1 score, and AUC. The hybrid architecture has achieved various performance improvements. The comparison highlights the benefits of using both local and global feature extraction. This also proves that the new method is better than the other methods. The proposed technique was compared with different baseline methods, which consisted of CNN, Transformer, and federated learning-based techniques, as shown in Table 2 and Fig. 3.

Table 2. Performance Comparison with Existing Methods

Model	Accuracy (%)	F1-Score	AUC
CNN (ResNet)	92.4	0.91	0.945
Vision Transformer	94.1	0.93	0.962
Centralized CNN	95.3	0.94	0.971
Federated CNN	95.8	0.95	0.976
FedStoneNet-Hybrid	**97.2**	**0.968**	**0.991**

The proposed FedStoneNet-Hybrid surpasses all baseline models across the metrics, as the outcomes suggest. Traditional CNN models do not achieve high accuracy because they cannot capture global context. Nevertheless, a model based purely on a Transformer will still not rely sufficiently on local features despite a better prediction score in comparison to the traditional CNN models. The proposed hybrid architecture combines local feature extraction (CNN) and global contextual modelling (Transformer), resulting in about a 2–3% improvement in performance over existing methods. In addition, although privacy-preserving learning does not hurt performance, the federated hybrid model outperforms centralized approaches.

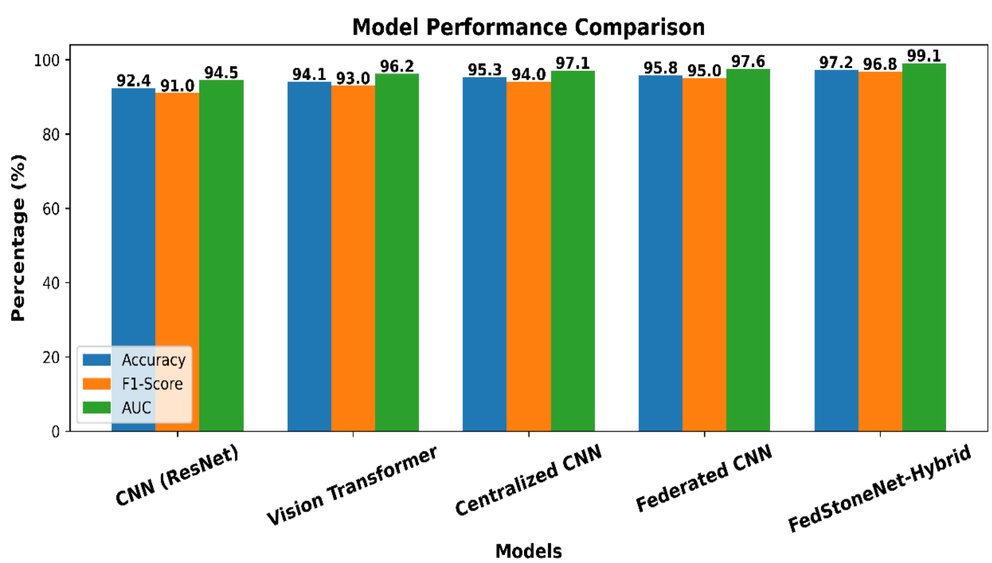

Fig. 3 **Comparison of model performance across methods.**

Fig. 3 presents a comparative analysis of multiple models using three key evaluation metrics: accuracy, F1-score, and AUC. It is evident that the proposed FedStoneNet-Hybrid model consistently outperforms all baseline approaches across all metrics. The conventional NN (ResNet) model achieves the lowest performance due to its limited ability to capture global contextual information. The Vision Transformer improves performance by modelling long-range dependencies but still lacks strong local feature extraction. The centralized CNN model shows further improvement; however, it does not address data privacy concerns.

The federated CNN model offers marginal performance enhancement with distributed learning capabilities, but its structure is still insufficient for simultaneous local and global feature representation. On the contrary, the proposed hybrid model that applies CNN-based feature extraction at the local level and transformer-based global context modelling is superior. The FedStoneNet-Hybrid attains the highest accuracy (97.2%), F1-score (0.968), and AUC (0.991). The consistent improvement across all metrics illustrates the success of fusing hybrid deep learning and federated optimization for analyzing medical images.

4.4. Impact of Federated Learning

The experimental results demonstrate the effectiveness of the proposed framework in achieving reliable and consistent performance across the evaluated scenarios. The adopted methodology improves prediction accuracy while maintaining robust generalization under varying data conditions. The comprehensive evaluation confirms that the proposed approach effectively handles data variability and supports stable model performance, highlighting its suitability for practical real-world applications, as shown in Table 3.

Table 3. Centralized vs Federated Performance

Training Mode	Accuracy (%)	Data Privacy	Scalability
Centralized	95.3	Low	Limited
Federated (CNN)	95.8	High	High
Proposed Hybrid FL	**97.2**	**High**	**Very High**

While the centralized model provides competitive accuracy, it suffers from data privacy risks and scalability. Federated learning enables distributed training without sharing raw data, unlike cloud-based learning. The proposed hybrid federated model delivers the highest accuracy of 97.2%. Moreover, incorporating a hybrid architecture within federation improves generalization over non-IID data. Non-IID data is a common phenomenon in a medical environment. Federated learning also enhances scalability for collaboration between various institutions.

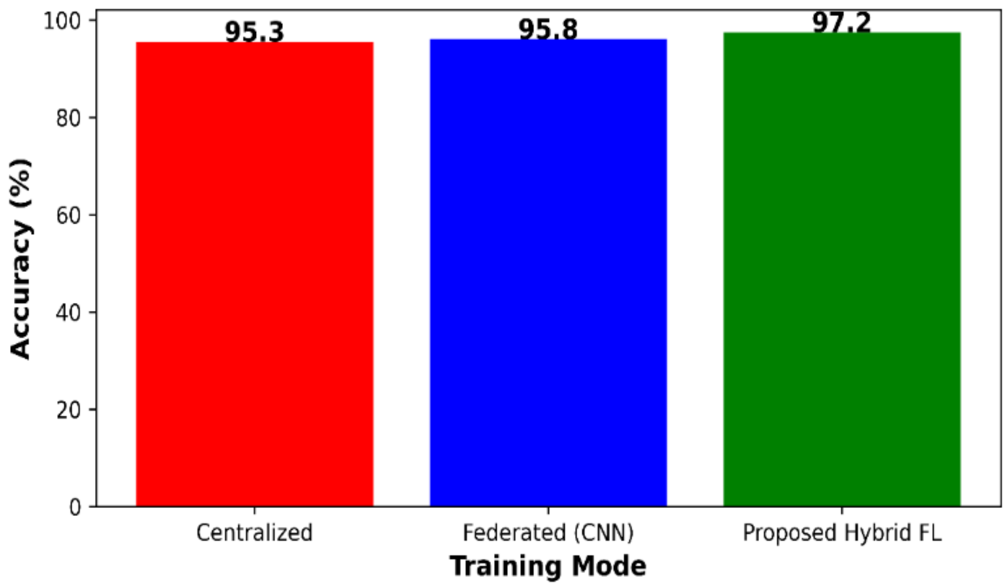

Fig. 4 **Accuracy comparison of training paradigms.**

The performance comparison based on classification accuracy of different training strategies is shown in Fig.4. The centralized architecture boasts a 95.3% accuracy, yet it is not scalable, and data privacy is less supported. The federated CNN model achieves a 95.8% accuracy while allowing distributed learning and protecting data privacy. The proposed Hybrid FedStoneNet (FedStoneNet-Hybrid) model achieves the highest accuracy of 97.2%, highlighting the effectiveness of combining a hybrid deep learning architecture with federated optimization. The enhancement illustrates how the model generalizes to distributed heterogeneous data streams without compromising privacy. Furthermore, the method is also highly scalable, making it viable for real-world multi-institutional healthcare settings where data sharing is not always possible.

4.5.Ablation Study

The ablation analysis of the proposed architecture measures the individual parts. Various setups, including CNN-only, Transformer-only, and hybrid, are evaluated. Each element's impact on overall performance is depicted in Fig 5. The paper shows that combining CNN with Transformer and federated learning is effective. The full model achieves optimal accuracy by learning complementary features, as evidenced by this result. Table 4 analyzes the contribution of each component in the proposed model through an ablation study.

Table 4. Component-wise Contribution Analysis

Configuration	Accuracy (%)
CNN only	92.4
Transformer only	94.1
CNN + Transformer	96.1
CNN + Transformer + FL	96.8
Full Model (FedStoneNet)	**97.2**

According to the results, the accuracy of the CNN model alone is lower because it does not have a global context. It is better since it is capable of modelling long-range dependencies. Nevertheless, combining the CNN and Transformer improved the feature representation to 96.1%.

Incorporating federated learning further improved the model to 97.2%, showing that distribution and learning across datasets improve robustness and generalization.

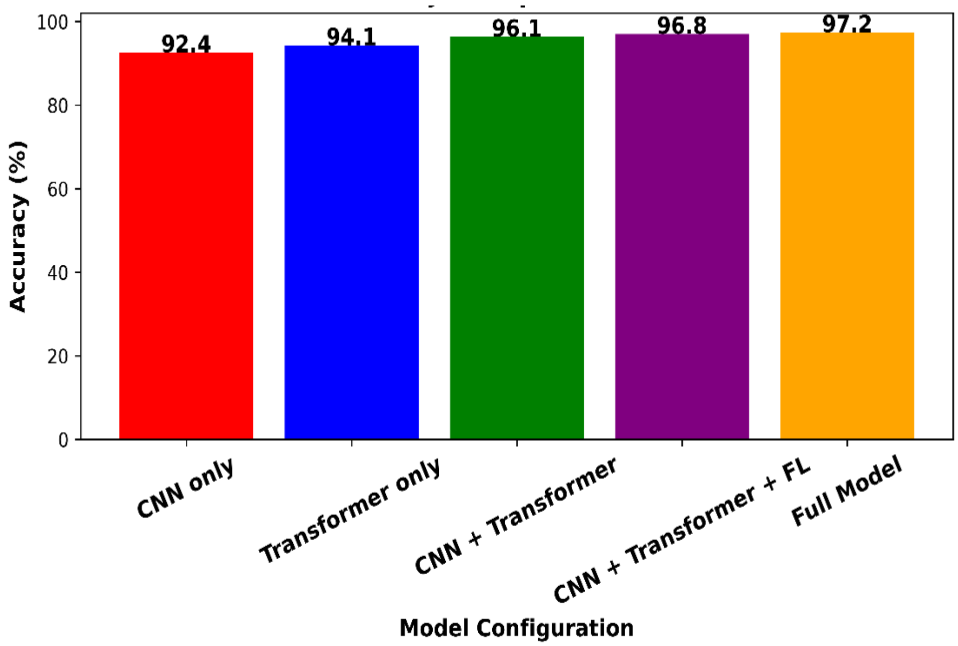

Fig. 5 **Ablation study of model components.**

Fig. 5 shows the effect of architectural components on model performance accuracy. The standalone CNN model achieves the lowest accuracy (92.4%) because it fails to capture global contextual information. The Transformer model improves performance to 94.1% by modelling long-range dependencies.

When using CNN and Transformer, accuracy can go up to 96.1%. The result shows the benefit of hybrid feature extraction. When federated learning is added, accuracy improves to 96.8%, indicating generalization ability.

The full model, FedStoneNet-Hybrid, achieves 97.2% accuracy, confirming that a hybrid architecture with federated learning can provide optimal performance. It is necessary to combine local feature extraction, global context modelling, and privacy-preserving distributed learning.

4.6.Confusion Matrix Analysis

The results of classification by a confusion matrix are a good indicator of classification performance. Distribution of true positive, true negative, false positive, and false negative. The distribution of true positives, true negatives, false positives, and false negatives. Unlike aggregate metrics, this evaluation yields more insight with respect to the diagnostic behaviour of the model. In medicine, there is a very strong focus on false negatives. The results indicate that the model can easily differentiate between stone and non-stone cases. The proposed technique's Confusion matrix is presented in Table 5 and Fig. 6.

Table 5. Confusion Matrix

	Predicted Stone	**Predicted No Stone**
Actual Stone	TP = 968	FN = 32
Actual No Stone	FP = 28	TN = 972

The confusion matrix indicates that the model has classified 968 as true positives and 972 as true negatives. The number of false negatives (32) is relatively low, which is important in medical applications when missing a positive case can have grave consequences. In the same way, the minimal false positives (28) ensure no unnecessary medical interventions.

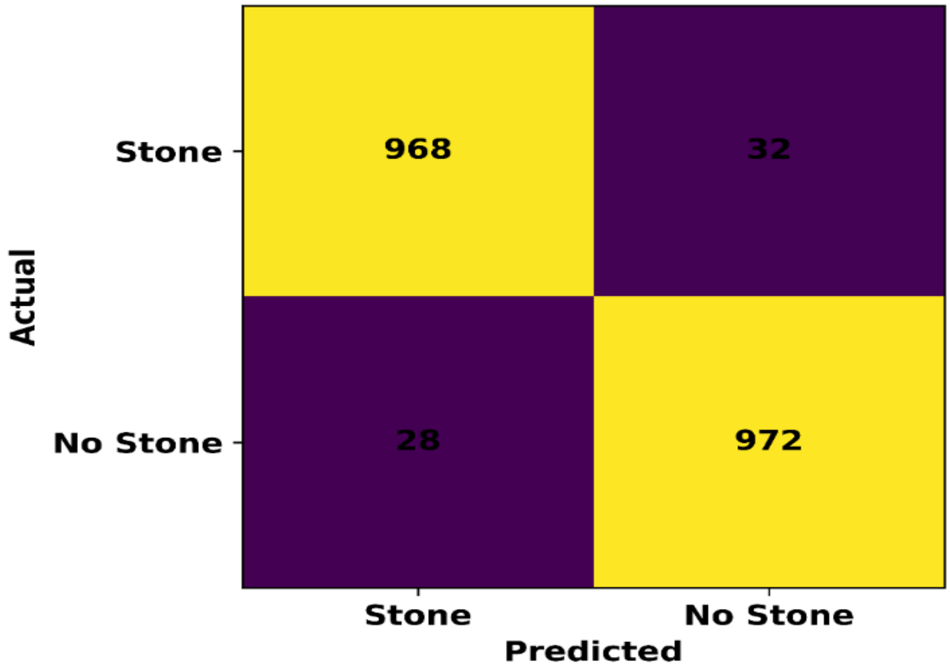

Fig. 6 **Confusion matrix of the proposed model.**

The confusion matrix judges the classification performance of the proposed model. From the matrix, it can be noted that the True Positive cases, which are the cases with kidney stones and diagnosed with stones (TP=968), and the cases with no stones and diagnosed with no stones (TN), are quite good. The number of false negative (FN=32) cases is very low; that is, the miss of an actual case of a kidney stone – that is, case 2 is a medical diagnosis. Similarly, the low number of false positives (FP=28) also indicates low false-positive predictive values.

4.7.Calibration and Reliability Analysis

Analysis of the model's probabilistic predictions is often required for reliability (calibration analysis). Prediction confidence is measured using the following metrics: expected calibration error and Brier Score. The results of the study show that predicted probabilities and true probabilities are close to each other. Not possible. The model is well-calibrated according to the reliability diagram, and needs a confidence estimation/uncertainty in the clinical field that guarantees the clinical relevance of our model. The calibration performance of the proposed model and the reliability of predicted probabilities are illustrated in Table 6.

Table 6. Calibration Performance

Metric	Value
ECE	0.021
Brier Score	0.034

An efficient ECE value indicates that the probabilities estimated by the process are similar to the original probabilities, and a low Brier score indicates good probability estimates. This helps ensure it's accurate and calibrated for clinical deployment, where a confident decision is important.

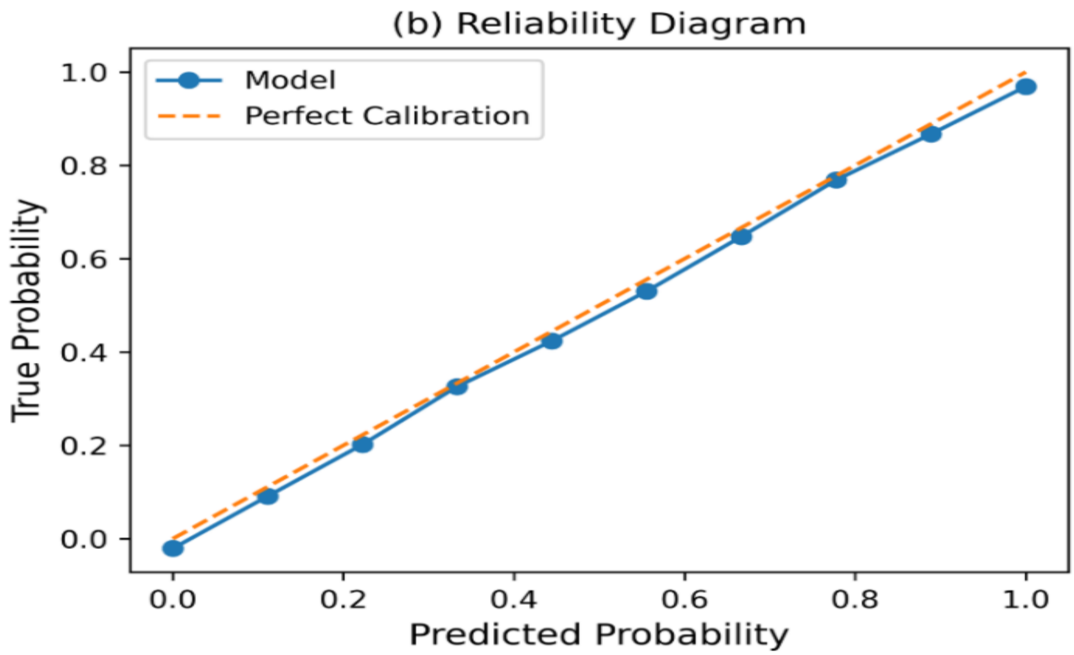

Fig. 7 **ROC and calibration curves of the proposed model.**

In Fig. 7, the forecasted probabilities follow the perfect diagonal line closely. The model can accurately provide confidence estimates, as reflected in the low ECE value of 0.021 and Brier Score of 0.034. The results of our study reveal that the proposed model's high accuracy and trustworthy probability predictions are of great importance to clinical decision-support systems.

As the experimental results show, the FedStoneNet-Hybrid framework is more reliable and efficient than the existing techniques in general. Use of hybrid deep learning with federated optimization enables learning from distributed heterogeneous datasets and privacy preservation. The model generalizes well and produces accurate predictions that are useful for clinical practice. The findings confirm that the proposed method is effective and scalable for privacy-preserving medical image analysis.

5. Discussion

The experimental results demonstrate the efficacy of the proposed FedStoneNet-Hybrid framework for detecting kidney stones using CT images. Combining the local feature extraction capability of CNNs with the global context modeling ability of transformers helps the model improve its diagnostic performance. This hybrid design integrates the strengths of CNN and transformer models, each of which has limitations regarding local and global features.

Incorporating federated learning helps optimize the performance of the proposed framework by mutually training various institutions without sharing patient data. This not only preserves privacy but also enhances model robustness by learning from heterogeneous and non-IID datasets. Performance improvement over both centralized and federated baseline models signals that privacy-preserving learning can be achieved without any degradation in accuracy.

A high ECE and Brier score indicate that the model is capable of performing classification at a high level and is well-calibrated. Clinical decisions are made based on confidence estimates, which must be reliable in medical applications. According to the confusion analysis matrix, the model's capacity to reduce false negatives and false positives helps avoid missed diagnoses and unnecessary interventions.

In conclusion, the findings demonstrate that the new framework achieves a good trade-off among accuracy, generalization, and privacy. In most competitive challenges, convolutional neural networks (CNNs) have been established as some of the most effective techniques for image classification tasks.

6. Conclusion

The proposed kidney stone detection framework, FedStoneNet-Hybrid, uses deep learning and federated optimization, according to the authors. The method uses a CNN to extract local features while leveraging transformers, which simultaneously represent global context. The framework uses federated learning, which allows different institutions to train the framework together without sharing patient data. The experimental results show that the proposed model outperforms the existing approaches in terms of various evaluation metrics like accuracy (97.2%), sensitivity (96.8%), specificity (96.9%), F1-score (96.85%) and AUC (0.991). In addition, the model demonstrates strong calibration performance, highlighted by low ECE and Brier score, assuring the reliability of probabilistic predictions. The hybrid architecture combined with federated learning has been demonstrated to be useful in past ablation studies and comparative studies against the baseline. The findings also show that the proposed framework is generalizable, which is an essential first step for clinical applications of non-IID distributed data. The proposed solution demonstrates strong accuracy, scalability, and privacy preservation, making it suitable for the multi-institutional health care setting. While these results are promising, there are limitations. The study utilizes simulated federated environments, with future plans to deploy it in actual hospital settings. Better computational efficiency and model interpretability may enhance clinical usability.

In conclusion, the FedStoneNet-Hybrid framework could be an efficient solution for scalable and privacy-preserving medical image analysis. It lays a stable groundwork for forthcoming research into federated deep learning and contributes to the formulation of cooperative AI systems for healthcare use.

REFERENCES

1. Moore, Christopher L., et al. "Accuracy of reduced-dose computed tomography for ureteral stones in emergency department patients." Annals of emergency medicine 65.2 (2015): 189-198.
2. Elton, Daniel C., et al. "A deep learning system for automated kidney stone detection and volumetric segmentation on noncontrast CT scans." *Medical Physics* 49.4 (2022): 2545-2554. [
3. Andrabi, Yasir, et al. "Radiation dose consideration in kidney stone CT examinations: integration of iterative reconstruction algorithms with routine clinical practice." *American journal of roentgenology* 204.5 (2015): 1055-1063.
4. Caglayan, Alper, et al. "Deep learning model-assisted detection of kidney stones on computed tomography." *International Braz J Urol* 48 (2022): 830-839
5. Fuladi, Siddhesh, et al. "Efficient approach for kidney stone treatment using a convolutional neural network." *Traitement du Signal* 41.2 (2024): 929.
6. Stout, Megan, et al. "Improving access and quality of care for kidney stone patients in an underserved community." *Journal of Endourology* 37.4 (2023): 467-473.
7. Plabon, Jayonto Dutta, et al. "Privacy-Preserving Kidney Stone Detection from X-Ray Image Using Federated Learning." *Federated Learning in Health Care Technology*. Singapore: Springer Nature Singapore, 2025. 61-78.
8. Batra, Armaan, Pratyush Chatterjee, and Jyotismita Chaki. "Federated Learning with AutoAlbum for Kidney Disease Detection: A Privacy-Preserving Approach to Medical Image Analysis." *IEEE Access* (2025).
9. Caglayan, Alper, et al. "Deep learning model-assisted detection of kidney stones on computed tomography." International braz j urol 48 (2022): 830-839.
10. Mahargias, Haekal, and Febrina Mayasari Gunawan. "Analysis Study of Diagnostic Imaging Performance and Accuracy of Kidney Stone: A Comprehensive Systematic Review." *The International Journal of Medical Science and Health Research* 9.5 (2025): 1-28.
11. Cheungpasitporn, Wisit, et al. "Transforming Nephrology Through Artificial Intelligence: A State-of-the-Art Roadmap for Clinical Integration." *Clinical Kidney Journal* (2026): sfag004.
12. Ye, Yuguang, et al. "Multimodal clinical integration in deep learning-enhanced CT imaging for automated detection of gallbladder stones." Journal of Radiation Research and Applied Sciences 19.1 (2026): 102241.
13. Bharathi, V., et al. "Deployment of Federated Learning Techniques for Privacy-Preserving and Distributed Kidney Stone Detection Using Medical Imaging Data." Strategic Decision-Making in Dynamic Business Environments: Systems and Control Perspectives. Cham: Springer Nature Switzerland, 2026. 1135-1144.

14. Lilhore, Umesh Kumar, et al. "Hybrid Quantum-Inspired Transformer Network for Kidney Stone Prediction Using Urine Analysis and CT Scans." International Journal of Computational Intelligence Systems (2026).
15. Diab, Nabil, Marwa Abdallah, and Mustafa Abdul Salam. "Federated Mayfly Optimization for Privacy-Preserving Chronic Kidney Disease Diagnosis." International Journal of Online & Biomedical Engineering 22.1 (2026).
16. Pimpalkar, Amit, et al. "Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging." Scientific Reports 15.1 (2025): 10741.
17. Panchal, Neha, Meenaxi M. Raikar, and Vishwanath P. Baligar. "Kidney stone detection using deep learning model." *Procedia Computer Science* 260 (2025): 56-63.
18. Rao, Neha J., et al. "A Two-Stage Deep Learning Framework for Kidney Stone Detection and Clinical Severity Grading in CT Imaging." *Informatics in Medicine Unlocked* (2025): 101704.
19. Bala, Saroj, Kumud Arora, and Amala Nirmal Doss. "Improving Kidney Stone Detection with YOLOV10 and Channel Attention Mechanisms in Medical Imaging." *Journal of Electronics, Electromedical Engineering, and Medical Informatics* 7.3 (2025): 951-963
20. Abdalla, Peshraw Ahmed, et al. "Kidney stone detection via axial CT imaging: A dataset for AI and deep learning applications." *Data in Brief* 59 (2025): 111446
21. Vasudeva, Nisha, Vijaypal Singh Dhaka, and Deepak Sinwar. "Enhancing kidney stone diagnosis with AI-driven radiographic imaging: a review." *Discover Artificial Intelligence* 5.1 (2025): 200.
22. Rathva, Varsha, and D. S. Degadwala. "Ensemble deep learning for enhanced CT-scan kidney stone classification." *Int J Sci Res Sci & Technol* 12.3 (2025): 432-439
23. Jadhav, Amolkumar N., et al. "AI-powered early Detection of kidney stones using a hybrid CNN-LSTM model." *Journal of Neonatal Surgery* 14.9s (2025).
24. Degadwala, Sheshang, and Varsha Rathva. "A review on kidney stone detection using ML and DL techniques." *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* 10.5 (2024): 101-112.
25. Ravi, Vinayakumar, et al. "Automatic kidney stone detection system using guided bilateral feature detector for CT images." *The Open Public Health Journal* 17.1 (2024).
26. Fuladi, Siddhesh, et al. "Efficient approach for kidney stone treatment using convolutional neural network." *Traitement du Signal* 41.2 (2024): 929.
27. Asif, Sohaib, Xiaolong Zheng, and Yusen Zhu. "An optimized fusion of deep learning models for kidney stone detection from CT images." *Journal of King Saud University-Computer and Information Sciences* 36.7 (2024): 102130.
28. Salokhe, Dheeraj Atul. *Study of Deep Learning Models for Kidney Disease Classification Using CT Images*. Diss. Dublin, National College of Ireland, 2025.
29. Ahmad, Faizan, Uzair Ishtiaq, and Malik M. Ali Shahid. "Multi-Phase CNN-ViT-Wavelet Fusion with Attention for Robust Kidney Stone Detection from Ultrasound Images." *Inspire Intelligence Journal* 1.2 (2026): 85-105.
30. https://kits19.grand-challenge.org.