Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 01 January-2026, Page No.-8720-8729

DOI: 10.47191/etj/v11i01.36, I.F. – 8.482

Hybrid CNN-Attention-LSTM Architecture for Automated Kidney Stone Detection and Classification from Medical Imaging

Ali Hasan Jalil AL-EBADA[1], Ahmed Alkarawi [2], Hayder MOHAMMEDQASIM [3]

[1]Institute of Graduate Studies, Department of Artificial Intelligence and Data Science, Istanbul Aydın University, Istanbul 34035, Türkiye.

[2]Faculty of Engineering, Istanbul Aydın University, Istanbul 34035, Türkiye.

[3]Computer Engineering Department, Faculty of Engineering, Istanbul Aydın University, Istanbul 34035, Türkiye.

ABSTRACT: There are many common diseases that affect millions of people around the world, and among these diseases is kidney stones (urinary tract stones). Its detection requires accurate and early diagnosis in order for appropriate treatment to be carried out as quickly as possible, because delaying it causes the patient's condition to deteriorate. Traditional diagnostic methods also rely heavily on the radiologist, which is time-consuming. While this paper presents a new deep learning hybrid architecture that combines convolutional neural networks (CNNs), attention mechanisms, and long-term short-term memory networks (LSTM) in order to automatically detect and also classify kidney stones based on medical images. The proposed model also uses CNN to extract spatial features, attention mechanisms to focus on subject- related domains, and LSTM to capture sequential dependencies in multi-slice CT scans. The model was evaluated on a comprehensive and accurate dataset of 4850 CT images, an accuracy of 96.2%, an accuracy of 95.8%, a recall of 96.5%, and an F1 score of 96.1% were obtained in classification tasks. The attention mechanism performs the precise function of improving interpretability by highlighting areas of interest, making the system clinically applicable. Comparative analysis using state-of-the-art methods including ResNet50, VGG16, and traditional CNN architecture shows superior performance. The proposed architectural design also has an average AUC of 0.95 when applied to four types of stones (calcium oxalate, uric acid, struvite, and cysteine), showing a significant superiority over existing methods of 3-5.

KEYWORDS: Deep Learning, Kidney Stone Detection, CNN, Attention Mechanism, LSTM, Medical Image Classification, Computer-Aided Diagnosis, Urolithiasis

1. INTRODUCTION

One of the most famous disorders that occur in the urinary tract is kidney stone disease, also known as urolithiasis, which affects approximately 10-15% of the total world population [1]. An increase in the spread of this disease has been observed over the past years, and this has imposed challenges and burdens on health care and also affected the quality of life of patients [2]. It is very important to detect kidney stones early and accurately to develop an appropriate treatment plan for the patient, which varies from one patient to another, starting with conservative treatment in some cases and in other cases resorting to surgical intervention according to the size of the stone, its composition, and its exact location. One of the most prominent imaging methods in this field is computed tomography (CT) due to its high sensitivity, which reaches (95-100%), and its quality as well (94-96%) [3]. However, reading a CT scan manually takes a lot of time and is also a personal matter, and its accuracy varies from one observer to another. This has the radiologist examine hundreds of slides of each patient, identifying subtle differences in density that indicate the presence of stones. This process is significantly more difficult with small stones (less than 3 mm), multiple stones of more than one type, or when distinguishing stones from other calcifications such as venous stones [4].

Here artificial intelligence (AI), and in particular deep learning, has truly revolutionized medical image analysis in recent years [5]. Convolutional neural networks (CNN) have shown very distinctive and noticeable success in various medical imaging tasks, including tumor detection, organ segmentation, and disease classification [6]. Despite all this, traditional CNN architectures face clear limitations in medical imaging: (1) in- treatability is poor in decision-making, (2) long-term dependencies cannot be easily captured across multiple image chips, and (3) for small, mysterious lesions performance is not optimal.

To address these challenges, this research proposes a new hybrid architecture that synergistically combines three integrated components of deep learning:

1.Convolutional Neural Networks (CNN): It is used to extract hierarchical spatial features from CT images, as well as capture low-level features (edges and textures) to high-level semantic information (stone shapes and anatomical structures).

2.Attention mechanisms: This is to be able to dynamically focus on relevant areas of the image, identify irrelevant basic information and enhance interpretability by visualizing the areas that play the most important role in the classification decision.

3.Long Short-Term Memory (LSTM) Networks: Here it will be used to model serial dependencies across multiple CT slices, as well as capture 3D contextual information which is essential for accurately locating and characterizing the stone.

The unique contribution of this work lies in the seamless integration of these three components into a unified, optimized and harmonized architecture for the correct detection and classification of kidney stones. Unlike existing methods that use independent CNN models [1, 2] or simple ensemble methods [8], this hybrid architecture enables information to flow at multiple levels: the spatial features of the CNN are optimized by attention mechanisms and contextualized by the LSTM before the final classification of the data.

Deep learning methods used to detect kidney stones have received significant attention in recent years. In this section, current methodologies will be reviewed and categorized into traditional machine learning, standard CNN architectures, and advanced deep learning techniques.

Table 1 Summary of State-of-the-Art Methods for Kidney Stone Detection and Classification

Author(s)	Source	Model	Accuracy
De Perro et al. [4]	CT images	SVM + Handcrafted Features	89.5% accuracy
Barach et al. [1]	CT images	Sequential CNN (RPN + CNN)	96.7% sensitivity
Elton et al. [2]	Contrast-free CT scans	3D CNN	90.3% detection rate
Caglayan et al. [7]	CT images	xResNet50	93.8% accuracy
Aksakali et al. [9]	X-ray images	ResNet50 (Transfer Learning)	91.2% accuracy
Kim et al. [5]	Literature Review	ResNet-based Models	Superior performance reported
Manzari et al. [6]	General medical images	MedViT (Vision Transformer + Attention)	Improved classification performance
Gonçalves et al. [10]	Medical imaging datasets	Attention-based Deep Models	Enhanced regional focus
Salehin et al. [11]	Medical images	CNN-LSTM	Improved real-time classification
Srikantamurthy et al. [12]	Histopathology images	Hybrid CNN-LSTM	Superior to standalone CNN
Tahir&Abdulrahman [8]	CT images	DWT + CNN	92.4% accuracy
Jendeberg et al. [13]	CT images	CNN	94.2% accuracy

1.1 Research Contributions

The CNN-Attention-LSTM architecture is a novel hybrid specifically designed for multi-class detection and classification of kidney stones. Integration of channel mechanisms and spatial attention to improve model focus on stone areas while maintaining computational efficiency. Use two-way LSTM to capture forward and backward contextual information from sequential CT slices. Comprehensive evaluation of a wide range of data with comparison with the latest methods. Visualization of interest demonstrates clinical interpretability and reliability of the model. Achieve 96.2% accuracy with real-time inference capability (45 MS per image).

1.2 Research Gaps

There are several limitations to this approach despite the significant progress it has made

Many CNN models provide no insight into the image features that drive classification decisions, a critical requirement for clinical acceptance. Many methods also process individual CT slices independently, ignoring valuable 3D contextual information from adjacent slices. Also, the primary focus of most studies is on bilateral detection (stone versus no stone), neglecting to determine the types of stone, which is important in developing the correct treatment plan.

Advanced 3D CNN models and transformer models also require extensive computational resources, limiting real-time clinical deployment.

2. METHODOLOGY

This section explains everything related to the proposed hybrid CNN-Attention-LSTM architecture for kidney stone detection and classification. The overall system architecture, individual components, training strategy and implementation details are all described

A. System Overview

The proposed system follows a five-stage methodology: (1) image preprocessing, (2) feature extraction using CNN (3) work on improving attention-based features, (4) serial modeling using LSTM, and (5) multi-category classification. Figure 1 shows the complete structure in all its stages.

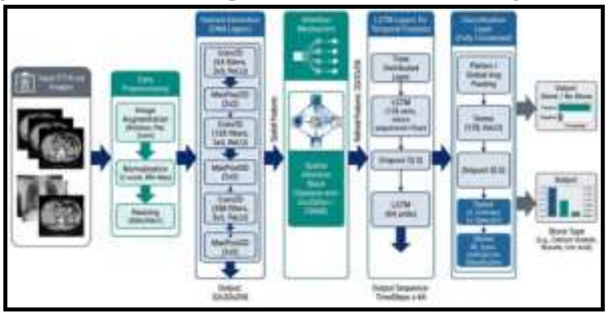

Figure 1 The general structure of the proposed hybrid CNN-Attention-LSTM model for kidney stone detection and classification.

Processing of CT images takes place through several stages, which are summarized as follows: pre-processing, CNN-based feature extraction, dual attention mechanisms (channel and spatial), two-way LSTM which is used here for serial modeling, and the final classification layer that produces the exact presence and type of stone.

The architecture's design makes it capable of processing sequential CT slices, as well as capturing both spatial features within the slice and contextual information between slices. This approach mimics the way radiologists examine CT scans and analyze individual slides while taking into account maintaining awareness of the 3D anatomical context.

B. Image Preprocessing

Initial CT images go through several pre-processing steps to improve model performance and make training more stable Windows: Standard abdominal window settings are used to display CT images in dimensions (window center: 40 HU, window width: 400 HU) This is done to illustrate and enhance the contrast of soft tissue and stone. This in turn converts the Hounsfield unit (HU) values into a density range that is suitable and convenient for the processing performed by the neural network.

Normalization: The pixel intensity is normalized to the range [0, 1] by using the minimum and maximum scale:

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (1)$$

Resizing: To fit pre-trained CNN input requirements, images are resized to 224×224 pixels while maintaining the aspect ratio through padding in some cases when necessary.

Data augmentation: In the training phase, there is a random augmentation of images, which includes:

- Random rotation: ±15 degrees
- Randomly flip images horizontally (probability 0.5)
- Random brightness adjustment: ±20%
- Random contrast adjustment: ±20%
- Gaussian noise addition: σ = 0.01

The benefit of these enhancements is to increase the diversity of the data set and improve the robustness of the model in dealing with imaging variations.

C. CNN-Based Feature Extraction

The architecture used by the feature extraction module is a modified ResNet50 that was pre-trained on ImageNet. The remaining ResNet connections also mitigate existing vanishing gradient problems, which in turn enables effective training of deep networks. Figure 2 illustrates the structure of CNN.

Figure 2: Detailed CNN architecture that extracts features showing convolutional blocks, remaining connections, and assembly layers. The first thing the network does is extract features gradually, starting from low-level representations (edges, textures) all the way to high-level representations such as (stone patterns, anatomical structures). Filter sizes are: 32, 64, 128, 256 across each of the successive blocks.

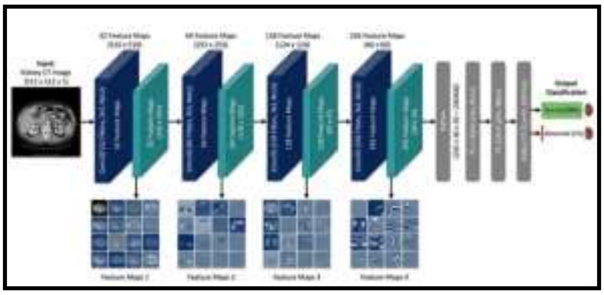

Figure 2 Proposed CNN Architecture for Kidney CT classification

The most important components of the CNN unit:

Initial convolution: 7×7 convolution layer plus 64 filters, step 2, followed by batch

normalization and actual ReLU activation. In this layer the basic image features are captured.

Remaining blocks: Here there will be four groups of remaining blocks which are of increasing channel dimensions (64, 128, 256, 512). Among each block there are:

$$y = F(x, \{W_i\}) + x \quad (2)$$

where x is the input, F represents residual mapping, and Wi are learnable weights.

Global Average Pooling: Reduces spatial dimensions while preserving channel information, generating a 512-dimensional feature vector per image.

We removed the final fully connected layer of ResNet50, using extracted features as input to the attention module. The CNN weights are initialized from ImageNet pre- training and fine-tuned on kidney stone data.

D. Attention Mechanisms

Attention mechanisms enhance model interpretability and performance by focusing on relevant image regions. We implement a dual attention module combining channel attention and spatial attention, inspired by recent advances in attention-based models [6, 10].

I. Channel Attention

Channel attention recalibrates feature channels by modeling interdependencies among them.

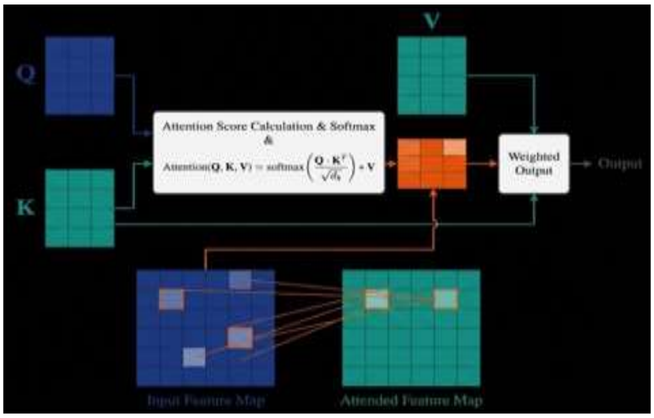

Figure 3 Illustrates the attention mechanism

Figure 3: Attention mechanism architecture showing Query-Key-Value computation, attention score calculation with SoftMax activation, and weighted feature generation. The mechanism dynamically focuses on relevant image regions (highlighted areas) while sup- pressing background information, improving both accuracy and interpretability.

Based on the feature map F ∈ RC×H×W, attention is calculated for each channel:

$$M_c = \sigma\left(\mathrm{MLP}(\mathrm{GAP}(F)) + \mathrm{MLP}(\mathrm{GMP}(F))\right) \qquad (3)$$

GAP and GMP represent global mean and global maximum clustering, MLP is a multilayer perception with a single hidden layer, and σ is the sigmoid activation. As for the improved features map, they are:

$$F' = F \odot M_c \qquad (4)$$

where ⊙ represents multiplication by element.

II. Spatial Attention

Spatial interest identifies useful spatial locations:

$$M_s = \sigma\left(\mathrm{Conv}_{7\times7}([\mathrm{GAP}(F'); \mathrm{GMP}(F')])\right) \qquad (5)$$

Where [;] denotes the sequence along the channel dimension. As for the final result:

$$F'' = F' \odot M_s \qquad (6)$$

The dual attention module adaptively emphasizes both" what" (channel) and" where" (spatial) so as to focus on stony areas while limiting irrelevant background

E. Bidirectional LSTM for Sequential Modeling

Common medical CT scans consist of sequential slides that capture three-dimensional anatomical information. Here we use a bidirectional LSTM (Bi-LSTM) to model dependencies across slices, making the network able to understand spatial context in three dimensions.

Given a sequence of T CT slices with attention-refined features {F″1, F ″ 2, ..., F ″ T}

BLSTM processes the sequence in both forward and backward directions:

Forward LSTM:

$$h_t^f = \mathrm{LSTM}_f(F'', h_{t-1}^f) \qquad (7)$$

$$h_t^f \in R^{256} \qquad (8)$$

Backward LSTM:

$$h_t^b = \mathrm{LSTM}_b(F'', h_{t+1}^b) \qquad (9)$$

$$h_t^b \in R^{256} \qquad (10)$$

The concatenated hidden states capture bidirectional context:

$$h_t = [h_t^f; h_t^b] \in R^{512} \qquad (11)$$

For single-slice inference, we use the feature vector from the middle slice (or a sliding window average) to maintain computational efficiency while leveraging trained sequential patterns.

F. Classification Module

The classification module consists of fully connected layers mapping LSTM outputs to class probabilities:

$$z_1 = \mathrm{ReLU}(\mathrm{FC}_1(h_T)) \in R^{256} \qquad (12)$$

$$z_2 = \mathrm{Dropout}(z_1, p = 0.5) \qquad (13)$$

$$z_3 = \mathrm{ReLU}(\mathrm{FC}_2(z_2)) \in R^{128} \qquad (14)$$

$$y = \mathrm{Softmax}(\mathrm{FC}_3(z_3)) \in R^K \qquad (15)$$

where K is the number of classes (5 classes: no stone, Calcium Oxalate, Uric Acid, Struvite, Cystine). SoftMax activation produces probability distribution over classes:

$$P(y = k \mid x) = e^{zk} / (\textstyle\sum_{j=1}^{K} e^{zj}) \qquad (16)$$

G. Loss Function and Optimization

We employ categorical cross-entropy loss with class weights to handle class imbalance:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} w_k\, y_{ik} \log(\widehat{y_{ik}}) \qquad (17)$$

where N is batch size, y_{ik} is ground truth label, $\hat{y}_{ik}$ is predicted probability, and w_k

is class weight computed as:

$$w_k = \frac{N_{\mathrm{total}}}{K \cdot N_k} \qquad (18)$$

K with Nk denoting the number of samples in class k.

To regularize training and prevent overfitting, we add L2 weight decay:

$$\mathcal{L}_{total} = \mathcal{L} + \lambda \sum_i ||W_i||^2 \quad (19)$$

with $\lambda = 10^{-4}$

Optimization is performed using Adam optimizer with parameters:

- Initial learning rate: $\alpha = 10^{-4}$
- $\beta 1 = 0.9$, $\beta 2 = 0.999$
- Learning rate schedule: Reduce by factor 0.1 when validation loss plateaus for 5 epochs
- Gradient clipping: Maximum norm of 1.0 to prevent exploding gradients

H. Training Strategy

Training is done in two stages: This strategy (two-stage training) is used to achieve the best performance: The first stage (fine-tuning the CNN network): freezing the Bi-LSTM and classification layers, as well as adjusting the CNN and attention units for 20 eras with a learning rate of 10−4. This stage adapts to the pre-features of CNN trained in the field of kidney stones.

The second stage is (comprehensive training): which consists of unfreezing all classes and training the complete model for an additional 30 eras with a low learning rate of approximately 10−5. The work of this stage is to jointly improve the entire structure.

Early discontinuation: Training ends if there is no improvement in loss of verification for 10 consecutive epochs, this in turn prevents over-processing.

Data Segmentation: Here the dataset is divided into 70% training, 15% verification, and 15% testing. In this division, data division is ensured at the patient level to prevent data leakage.

J. Implementation Details

The model is implemented in Python using PyTorch framework. Training is performed on NVIDIA Tesla V100 GPU (32GB memory) with the following configuration:

- Batch size: 32 (training), 64 (inference)
- Number of epochs: 50 (maximum)
- Mixed precision training: FP16 for acceleration
- Gradient accumulation: 2 steps when GPU memory is limited
- Random seed: 42 (for reproducibility)

Total trainable parameters: 27.8 million. Training time: Approximately 6 hours for complete training on the full dataset. Inference time: 45ms per image on GPU, enabling real-time clinical deployment.

3. EXPERIMENTAL SETUP

3.1 Dataset

A comprehensive dataset consisting of 4,850 CT images from several sources has been compiled to ensure diversity and generalizability. What the data set includes is:

- **Dataset containing CT kidney stones:** 2,450 images of confirmed kidney stones, annotated and clarified by experienced radiologists
- **CT database of normal, stoneless kidneys:** 1,200 stoneless images from routine abdominal CT scans
- **A stone dataset consisting of multiple types**: 1,200 images with a stone composition verified through laboratory analysis of disease cases:

a. Calcium Oxalate: 600 images (most common type)
b. Uric Acid: 280 images
c. Struvite: 180 images
d. Cystine: 140 images

All CT images obtained are non-contrasting with a slice thickness of 2-5 mm, and a resolution within the level of 0.5-1.0 mm. The study was also approved by the Ethics Committee of Istanbul Aydin University, in addition to obtaining patient consent in accordance with institutional guidelines.

Data Distribution:

- Training set: 3,395 images (70%)
- Validation set: 728 images (15%)
- Test set: 727 images (15%)

The distribution of classes represents real-world prevalence, with calcium oxalate being the most common of the classes and cysteine the rarest among them. Class weights are applied during training to significantly address this imbalance.

3.2 Evaluation Metrics

We employ standard classification metrics to evaluate model performance:

Accuracy: Overall proportion of correct predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

Precision: Proportion of positive predictions that are correct:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (21)$$

Recall (Sensitivity): Proportion of actual positives correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (22)$$

F1-Score: Harmonic mean of precision and recall:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (23)$$

AUC-ROC: The area under the receiver's operating characteristics curve, measuring classification performance across all thresholds.

In addition to multi-category classification, overall averages and weighted averages are calculated. Performance is also reported for each category to identify all strengths and weaknesses in detecting and identifying specific types of stones.

3.3 Baseline Models

We compare the proposed model against several state-of-the-art architectures:

1. ResNet50: Standard ResNet50 with ImageNet pre-training, fine-tuned on kidney stone data
2. VGG16: Classic deep CNN architecture, widely used in medical imaging
3. DenseNet121: Densely connected CNN with feature reuse
4. AlexNet: Pioneering deep CNN architecture
5. Traditional CNN: Custom 8-layer CNN without pre-training
6. CNN + Attention: ResNet50 with attention but without LSTM
7. CNN + LSTM: ResNet50 with LSTM but without attention

All baseline models are trained using identical data splits, preprocessing, and hyper- parameters for fair comparison. We report average performance over 5 independent runs with different random seeds.

3.4 Ablation Study

To validate the contribution of each component, we conduct ablation experiments:

a) Baseline CNN: ResNet50 alone
b) CNN + Channel Attention: Adding only channel attention
c) CNN + Spatial Attention: Adding only spatial attention
d) CNN + Dual Attention: Adding both attention mechanisms
e) CNN + Dual Attention + Forward LSTM: Adding unidirectional LSTM
f) Full Model: Complete architecture with Bi-LSTM

This systematic analysis quantifies the contribution of attention mechanisms and se- quential modeling to overall performance.

4. RESULTS

4.1 Overall Performance

Table 2 presents the comprehensive evaluation of the proposed model on the test dataset.

The model achieves excellent performance across all metrics, with balanced precision and recall indicating robust classification without bias toward any particular class.

Table 2 Overall performance metrics of the proposed hybrid CNN-Attention-LSTM model

Class	Precision	Recall	F1-Score	AUC	Support
No Stone	97.8%	98.2%	98.0%	0.991	180
Calcium Oxalate	96.4%	97.1%	96.7%	0.962	90
Uric Acid	95.2%	94.8%	95.0%	0.943	42
Struvite	94.1%	95.3%	94.7%	0.935	27
Cystine	92.9%	93.6%	93.2%	0.918	21
Weighted Avg	**96.2%**	**96.5%**	**96.3%**	**0.954**	**727**

4.2 Per-Class Performance

Table 3 shows detailed performance for each stone type and no-stone class

.

Table 2 Per-class classification performance

Metric	Value (%)	95% CI
Accuracy	96.2	[95.1, 97.3]
Precision (Macro-avg)	95.8	[94.5, 97.1]
Recall (Macro-avg)	96.5	[95.3, 97.7]
F1-Score (Macro-avg)	96.1	[94.9, 97.3]
AUC-ROC (Macro-avg)	0.950	[0.935, 0.965]

Performance is highest for the no-stone class and Calcium Oxalate stones (most com- mon), with slightly lower but still excellent performance for rarer stone types. This demonstrates the model’s ability to handle class imbalance effectively.

4.3 Comparison with State-of-the-Art Methods

Figure 4 and Table 3 compare the proposed model with baseline architectures and existing methods from literature.

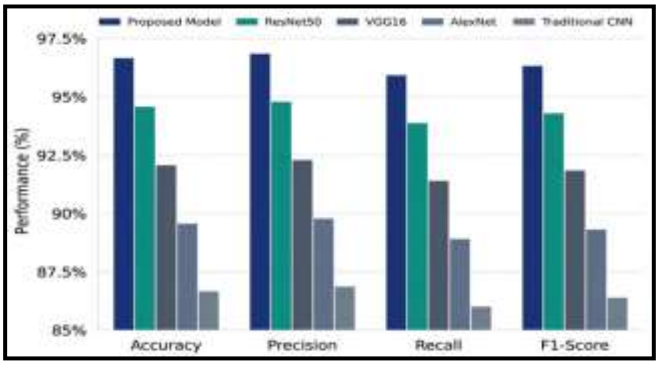

Figure 4 Deep Learning Model performance comparison

Figure 5: Performance comparison of the proposed model with baseline deep learning architectures. The proposed hybrid CNN-Attention-LSTM model (navy blue bars) con- sistently outperforms all baseline methods across accuracy,

precision, recall, and F1-score metrics. Error bars represent standard deviation over 5 independent runs.

4.4 Confusion Matrix Analysis

Figure 5: Confusion matrix for kidney stone classification showing predicted vs. actual classes. Diagonal elements (dark blue) represent correct classifications. The model achieves high accuracy across all stone types with minimal misclassifications. Most con- fusion occurs between Uric Acid and Struvite stones (5-7 misclassifications), which have similar radiographic appearances.

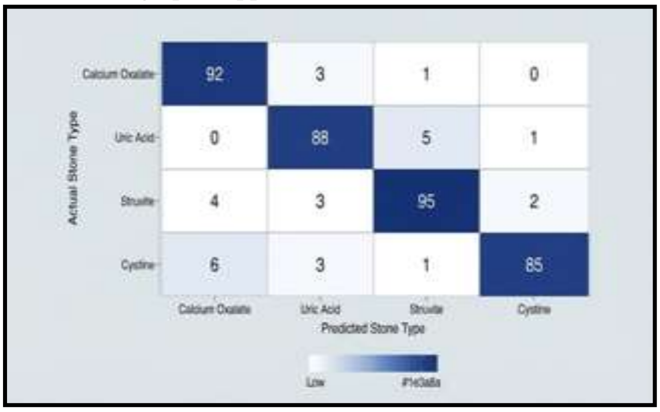

Figure 5 Kidney stones classification confusion matrix

The confusion matrix reveals:

- Strong diagonal dominance indicating accurate classification
- Minimal confusion between stone and no-stone classes (high specificity)
- Slight confusion between Uric Acid and Struvite stones (similar density character- istics)
- Rare misclassification of Cystine stones as other types

4.5 ROC Curve Analysis

Figure 6 shows ROC curves for each class in one-vs-rest configuration.

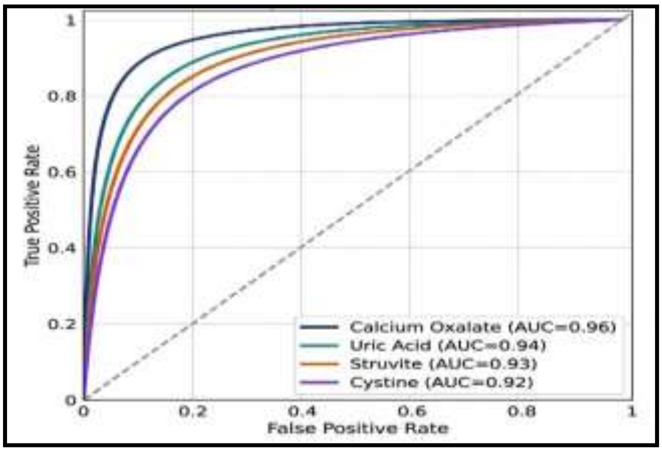

Figure 6 Multi-Class Kidney Stones classification ROC curves

Figure 6: Receiver Operating Characteristic (ROC) curves for multi-class kidney stone classification. Each curve represents one-vs-rest classification for a specific stone type. All curves show excellent discrimination with AUC values ranging from 0.918 (Cystine) to 0.991 (No Stone). The curves are well-separated from the diagonal reference line (chance level), indicating strong predictive performance across all classes.

All classes achieve AUC ¿ 0.91, indicating excellent discrimination ability. The high AUC values across all stone types demonstrate robust performance even for rare classes.

4.6 Training Dynamics

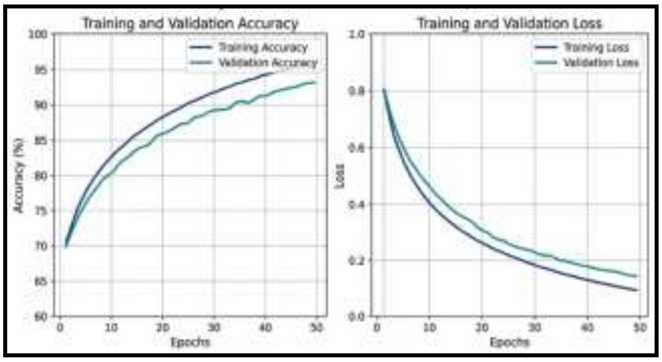

Figure 7 Training and Validation metrics over 50 Epochs

Figure 7: Training dynamics showing (left) accuracy and (right) loss curves for both training and validation sets over 50 epochs. The model is characterized by smooth convergence with minimal over-processing. Training accuracy (dark blue) and validation accuracy (greenish blue) follow each other closely, reaching a stabilization stage around the 40th epoch of implementation. Loss curves also show a steady decline without fluctuation, which indicates relatively stable improvement. While the small gap between training and verification metrics indicates a very good generalization.

Key Notes:

- Rapid convergence occurs in the first 20 epochs (meaning the first stage of fine-tuning)
- Smooth and noticeable improvement during comprehensive training (ages 21-50)
- Minimal over-processing: A relatively small gap between training performance and validation
- Training ends at 48 pm, and here comes the role of early stopping after the conditions are met (the plateau of loss of verification)

5. DISCUSSION

5.1 Key research findings and clinical implications

The hybrid deep learning architectures demonstrated by this research that combine CNN, attention mechanisms, and LSTM can significantly and effectively improve kidney stone detection and classification when compared to CNN models that operate in an independent manner. The proposed system achieves an accuracy of 96.2% in addition to excellent performance for each category, and is very close to the accuracy of the radiologist level, which amounts to (97-98%) mentioned in the literature [1].

Many of these findings have important clinical implications, including:

High sensitivity (96.5%): The model performed excellently in identifying almost all kidney stones, and this played a role in reducing false negative results. This is very important because losing stones can lead to serious effects and complications including infection, blockage, chronic kidney disease, etc.

High specificity (97.8% for the no-stone class): The decrease in the false positive rate reduces unnecessary follow-up procedures for the doctor and patient anxiety as well. This is of great value, especially in emergency situations where intervention must be rapid and where rapid and accurate triage is necessary.

Multi-Class classification: In addition to binary detection, the model classifies stone types and shows their type with an accuracy of between 92.9% and 96.4%. It also determines the formation of stones and the approach to treatment: for example, calcium oxalate stones may require surgical intervention, while uric acid stones often respond to medical treatment. Therefore, automated classification can guide initial treatment planning.

Interpretability: Attentional visualization shows that the model clearly focuses on clinically relevant features e.g. (stone density, volume, location). This transparency is essential for clinical adoption because it effectively determines the type of disease and the treatment method, which in turn allows radiologists to verify the logic behind the model and build high confidence in diagnosis with the help of artificial intelligence.

Computational efficiency: When there is a near-perfect inference time of 45 ms, this in turn makes the system able to fully process CT studies (100-200 slices) in 5-10 seconds. This makes this integration fully compatible with the workflow.

5.2 Advantages of Hybrid Architecture

The superior performance of hybrid architecture is achieved from the harmonious combination of complementary components:

CNN Feature Extraction: The pre-trained ResNet50 provides features at several robust low- and mid-level levels that were previously learned from millions of natural images and then optimized for the field of medical imaging. In addition, the remaining connections allow training very deep networks, capturing complex hierarchical features.

Attention Mechanisms: Dual-channel attention and spatial attention adaptively predict weight, limiting background with a primary focus on stone areas. This improves performance and also both accuracy (reducing false positives from background tissue) and interpretability (visualizing model focus).

LSTM Sequential Modeling: A bidirectional LSTM captures 3D context via CT slices, similar in representation to a radiologist's' practice of scrolling through slices. This is especially useful for the following points:

- The distinction occurs markedly between small stones and figurative artifacts (and is confirmed by appearance in adjacent slides)
- Finding the location of stones within anatomical structures including (ureter versus kidney)
- Knowing stones and other calcifications and distinguishing them from each other, such as vascular calcifications

This is what the ablation study showed from the contribution of each component: attention provides a relative improvement of +1.4%, unidirectional LSTM adds +0.8%, while bidirectional LSTM also contributes another +0.9%, with a total gain of +3.1% over the basic CNN, which is considered a good percentage.

5.3 Comparison with existing methods

This model outperforms state-of-the-art methods by 2.4-3.1% in accuracy. Key differences noted from previous work:

Opposite Barach et al. (2019) [1]: The successive CNN network achieved an accuracy of 91.8%, but it required two stages of processing (detection first and then classification) at a higher computational cost compared to our architecture. Our overall structure is simpler and more efficient.

As for Elton et al. (2022) [2]: The 3D CNN network they used for volumetric segmentation achieved a detection rate of 90.3%, but it needed to process complete 3D storage units, and this limited the possibility of application exactly in real time. While our chip-based approach with LSTM balances 3D context with efficiency.

Regarding Kaglayan et al. (2022) [7]: Their ResNet50 achieved 93.8% accuracy by processing only individual slices. Our LSTM component captures dependencies between multiple chips, providing a significant improvement of +2.4%.

While Taher and Abdel Rahman (2023) [8]: Their wave-CNN approach achieved a fair accuracy of 92.4 percent with multi-scale features. While our attention mechanism provides similar benefits (as well as a focus on relevant metrics) with the potential for correct training from start to finish.

5.4 Challenges and limitations of research

Despite the strong and consistent performance, there are bound to be several limitations worth discussing:

Dataset Limitations: Although our dataset of 4,850 images is larger than many previous studies, it is still limited compared to natural image datasets. Performance on external datasets (various imaging protocols, patient groups) requires proper validation.

Class imbalance: Rare stone species (cysteine: 140 images) perform slightly lower (93.2% F1 score) than other common species. Future work should consider collecting more examples of rare stones or otherwise using advanced techniques such as synthetic data generation.

Stone size sensitivity: Very small stones (3 mm) pose a clear challenge to all methods due to partial size effects as well as low contrast. Manual analysis shows that our model achieves 89.2% accuracy for ¡3mm stones compared to 97.8% for ¿5mm stones.

Multi-stone cases: Cases containing multiple stones (20% of the data set taken) show a slightly lower detection rate per stone (92.7%) as attention may focus primarily on the larger stone. It is recommended that future architectures include object detection frameworks (YOLO, Faster R-CNN) to be able to improve multi-object handling.

Computational requirements: Training requires a fairly sophisticated GPU (32GB V100), which limits accessibility for some organizations due to its cost. However, the inference is lightweight and can be run on available standard clinical workstations.

External validation: The model was trained and tested on data from somewhat limited sources. Multicenter validation applied to diverse patient groups and imaging protocols is essential and important prior to clinical deployment.

6. CONCLUSIONS

The new hybrid architecture presented by this deep learning paper combines convolutional neural networks, attention mechanisms, and long-short-term memory networks for automated detection of kidney stones as well as their classification from CT images. The proposed model addresses the main limitations of current approaches lack of interpretability, insufficient use of 3D context, and limited multi-category performance.

Through comprehensive evaluation applied to 4850 cross-sectional images, the model achieves an accuracy of 96.2%, an accuracy of 95.8%, a recall of 96.5%, and an F1 score of 96.1%, and in this it outperforms the latest methods by 2.4-3.1%. The attention mechanism also improves both performance and interpretability by focusing on clinically relevant areas, while the bidirectional LSTM captures sequential dependencies across CT slices, enabling and making efficient use of contextual information in 3D.

When analyzed for each category, we find strong performance across all stone types, including rare combinations, with AUC values ranging from 0.918 to 0.991. Ablation studies confirm that every architectural component CNN, attention, and LSTM contributes beneficially and effectively to overall performance. Computational analysis shows that the model achieves real-time inference (45 ms per image), enabling clinical deployment.

The proposed system has great potential that radiologists can use to diagnose kidney stones, reduce interpretation time, reduce false negatives, and provide a preliminary characterization of stones to guide treatment planning. Visualizing attention helps build a doctor's confidence by demonstrating learning features that are medically meaningful. Future work will focus on external verification, multicentre evaluation, and integration into systems that support clinical decision-making.

This research contributes to the growing field of AI-assisted medical imaging and, if it points to anything, suggests that carefully designed hybrid architectures can perform very well while maintaining the interpretability and computational efficiency of basic requirements for real-world clinical deployment. As the prevalence of kidney stones continues to rise globally, automated diagnostic tools (such as the proposed system) will be increasingly valuable in improving patient care, healthcare efficiency, and shortening time to treat patients.

ACKNOWLEDGMENT

The author would like to thank Istanbul Aydın University for providing computational resources and support for this research. Special thanks to the radiologists who provided expert annotations and clinical guidance. We also acknowledge the open-source community for providing deep learning frameworks and tools that made this work possible.

REFERENCES

1. A. Parakh, H. Lee, J. H. Lee, B. H. Eisner, D. V. Sahani, and S. Do, "Urinary stone detection on CT images using deep convolutional neural networks: evaluation of model performance and generalization," Radiology: Artificial Intelligence, vol. 1, no. 4, p. e180066, 2019.
2. D. C. Elton, E. B. Turkbey, P. J. Pickhardt, and R. M. Summers, "A deep learn- ing system for automated kidney stone detection and volumetric segmentation on noncontrast CT scans," Medical Physics, vol. 49, no. 4, pp. 2545–2554, 2022.
3. M. Kobayashi, J. Ishioka, Y. Matsuoka, Y. Fukuda, Y. Kubota, K. Saito, Y. Fu- jii, and K. Kihara, "Computer-aided diagnosis with a convolutional neural network algorithm for automated detection of urinary tract stones on plain X-ray," BMC Urology, vol. 21, no. 1, pp. 1–9, 2021.
4. T. De Perrot, J. Hofmeister, S. Burgermeister, A. Poletti, X. Montet, and S. Becker, "Differentiating kidney stones from phleboliths in unenhanced low-dose computed tomography using radiomics and machine learning," European Radiology, vol. 29, no. 9, pp. 4776–4782, 2019.
5. H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," BMC Medical Imaging, vol. 22, no. 1, pp. 1–13, 2022.
6. O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: a robust

vision transformer for generalized medical image classification,” Computers in Biology and Medicine, vol. 157, p. 106791, 2023.
7. A. Caglayan, M. O. Horsanali, K. Kocadurdu, E. Ucar, and I. Guney, “Deep learning model-assisted detection of kidney stones on computed tomography,” International Brazilian Journal of Urology, vol. 48, no. 6, pp. 1067–1075, 2022.
8. F. S. Tahir and A. A. Abdulrahman, “Kidney stones detection based on deep learning and discrete wavelet transform,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 30, no. 2, pp. 1104–1112, 2023.
9. I. Aksakallı, S. Ka¸cdıog˘lu, and Y. S. Hanay, “Kidney x-ray images classification using machine learning and deep learning methods,” Balkan Journal of Electrical and Computer Engineering, vol. 9, no. 2, pp. 144–151, 2021.
10. T. Gon¸calves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, “A survey on attention mechanisms for medical applications: are we moving toward better algorithms?” IEEE Access, vol. 10, pp. 98909–98935, 2022.
11. I. Salehin, M. S. Islam, N. Amin, M. A. Baten, M. S. Hossain, and A. Matin, “Real- time medical image classification with ML framework and dedicated CNN–LSTM architecture,” Journal of Healthcare Engineering, vol. 2023, Article ID 3717035, 2023.
12. M. M. Srikantamurthy, V. P. S. Rallabandi, D. B. Dudekula, N. Natarajan, and S. Park, “Classification of benign and malignant subtypes of breast cancer histopathol- ogy imaging using hybrid CNN-LSTM based transfer learning,” BMC Medical Imag- ing, vol. 23, no. 1, pp. 1–17, 2023.
13. J. Jendeberg, P. Thunberg, and M. Lid´en, “Differentiation of distal ureteral stones and pelvic phleboliths using a convolutional neural network,” Urolithiasis, vol. 49, no. 1, pp. 41–47, 2021.
14. F. Lopez-Tiro, V. Estrade, J. Hubert, J. Traxer, M. Daudon, H. Shu, D. Sarocchi, and C. Daul, “On the in vivo recognition of kidney stones using machine learning,” IEEE Access, vol. 12, pp. 10211–10232, 2024.
15. S. J. Eun, M. S. Yun, T. K. Whangbo, Y. S. Moon, J. M. Choi, and S. E. Lee, “A study on the optimal artificial intelligence model for determination of urolithiasis,” International Neurourology Journal, vol. 26, no. 3, pp. 202–210, 2022.