*I. J. Engineering and Manufacturing*, **2026, 5, 473-484**
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2026.05.25

# Attention-Based Spatial-Temporal Learning for Enhanced Kidney Abnormality Diagnosis from Ultrasound Images

**Ravikumar Ch.***
Department of CSE, Sreenidhi University, Hyderabad, Telangana, 500031, India
E-mail: ravikumar.c@suh.edu.in
ORCID iD: https://orcid.org/0000-0003-0809-5545
*Corresponding Author

**Satyanarayana Nimmala**
Department of CSE (Data Science), CVR College of Engineering, Hyderabad, India
E-mail: satyauce234@gmail.com
ORCID iD: https://orcid.org/0000-0002-9433-1489

**P. Vasanthsena**
G. Narayanamma Institute of Technology and Science (for women), Hyderabad, India
E-mail: drpvsena@gnits.ac.in
ORCID iD: https://orcid.org/0000-0002-6447-1982

**Nenavath Chander**
Department of CSE (cyber security), Malla Reddy University, Hyderabad, Telangana, India
E-mail: chander.nenavath@gmail.com
ORCID iD: https://orcid.org/0000-0001-7889-6511

**R. Sahith**
Department of CSE, CVR College of Engineering, Hyderabad, India
E-mail: r.sahith@cvr.ac.in
ORCID iD: https://orcid.org/0009-0006-8611-2602

Received: July 3, 2026; Revised: July 30, 2026; Accepted: August 25, 2026; Published: October 8, 2026

**Abstract:** Kidney abnormalities, when detected late, can develop into chronic conditions such as kidney tumors and cardiorenal syndrome. Recently, there has been a move toward automation, especially as it pertains to developing a framework for advanced preprocessing, feature fusion, and hybrid spatial-temporal learning for the detection of the abnormalities in question. For example, Non-Local Means (NLM) filtering for noise reduction, image quality improvements through U-Net kidney segmentation and deep learning-based text annotation removal, as well as robust training through data augmentation and preprocessing ends up facilitating the process. The combination of classical texture and shape descriptors and deep embeddings from a Vision Transformer (ViT) through an attention mechanism allows the model to pinpoint where clinically relevant focus should be. A hybrid CNN-LSTM framework attends to spatial and temporal feature extraction, while attention modules receive the classification output to refine it. The incorporation of weighted binary cross-entropy for handling class imbalance, and of explainable AI, is demonstrated through Integrated Gradients. Experimental results for VATLA show it used advanced algorithms and automation principles in developing systems for all previous performance indicators, while providing evidence of scoring 0.94 in accuracy, 0.92 in precision, 0.95 in recall, and 0.96 in AUC, or area under the curve.

**Index Terms:** Kidney abnormalities, Ultrasound imaging, Attention-Driven CNN-LSTM, Feature fusion, Explainable AI

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

## 1. Introduction

Kidney abnormalities impact the health of people of all ages and all members of the population. The more common abnormalities are hydronephrosis, cysts, stones, and other functional and structural abnormalities of the kidneys (Di Serafino et al., 2022) [1]. These conditions are serious because if they are not treated, they may lead to advanced stage chronic kidney disease, uremia, kidney tumors, and even cardiorenal syndrome (Drew, Weiner, & Sarnak, 2019) [2]. The Global Burden of Disease study estimates that kidney disorders are responsible for more than half a million deaths every year. This figure is an indication of the urgent need for more effective and scalable ways to detect early- to middle-stage kidney conditions (Zhang & Parikh, 2019) [3]. Abnormal kidneys also impact the functioning of the heart, the cardiovascular system, and the central nervous system (Drew, Weiner, & Sarnak, 2019; Zhang & Parikh, 2019) [2, 3].

In the early detection of kidney abnormalities, the utility of medical imaging cannot be overstated. US, CT, and MRI are the primary modalities (Cammarasana, Nicolardi, & Patanè, 2022) [4]. Among these, the advantages of ultrasound include real-time imaging, low cost, and no exposure to ionizing radiation. However, the complications of speckle noise, poor contrast, and artifacts result in challenging manual interpretation and significant cognitive burden (Cammarasana, Nicolardi, & Patanè, 2022) [4]. Moreover, the review of numerous images per session is time-consuming, increases the variability in a diagnosis, and may contribute to the described cognitive overwhelm. Computer-aided diagnostic (CAD) systems have been developed to address these issues by providing quick and more reliable analyses (Chen et al., 2023) [5].

Computer-aided detection (CAD) systems are quite successful in other medical domains, like detection of breast tumors and analyzing retinal images (Arora & Raman, 2023) [6]. In contrast, detection of abnormalities in kidneys is relatively unexplored (Velayuthapandian, Kishore & Sajin, 2024) [7]. Most kidney-centric studies focus on the segmentation and localization of kidney regions, which aids in visualization and measurement but still requires interpretation from the user (Pande & Agarwal, 2024) [8]. Most prior works use traditional, hand-crafted features like textures and shapes to classify the images into categories such as normal, cyst, stone, or tumor (Di Serafino et al., 2022; Krispin et al., 2020; Blau et al., 2018) [1, 9, 10]. Such methods tend to follow the two-step process of identifying the region of interest (ROI) and then classifying data within that ROI. Although this method reduces noise, it is still highly reliant on accurate localization, which is difficult due to patient-specific variations in kidneys, like the size, orientation, texture, and other anatomical features (Kaczmarek et al., 2024; Krispin et al., 2020) [9, 11]. Furthermore, hand-crafted features are inadequate for images that have overlapping abnormalities, which reduces the robustness and generalizability of traditional systems (Kovesdy, 2022) [12].

Recent advances in deep learning have substantially improved automated kidney abnormality detection from ultrasound images by enabling end-to-end feature learning without relying solely on handcrafted descriptors. Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and attention mechanisms have demonstrated remarkable capability in extracting discriminative spatial representations from ultrasound images, while hybrid learning strategies have further improved classification performance through complementary feature integration. Nevertheless, existing studies exhibit several limitations. Most approaches primarily focus on single-image classification and rarely investigate effective fusion of handcrafted texture descriptors with transformer-based global representations. Furthermore, although CNN-LSTM architectures have been extensively adopted for sequential medical imaging applications, their applicability to static ultrasound datasets is often inadequately justified. Existing methods also provide limited interpretability, insufficient handling of class imbalance, and inadequate explanations regarding feature interactions during classification. These limitations motivate the development of a robust attention-driven feature fusion framework that integrates complementary handcrafted and deep features while incorporating explainable artificial intelligence to improve diagnostic reliability and clinical interpretability.

To address these difficulties, we developed a completely automated Attention-Driven CNN-LSTM structure (VATLA) specifically designed for detecting kidney abnormalities. The framework consists of three parts: (i) a recognition module that analyzes a sequence of kidney images, selecting the relevant and dismissing the irrelevant frames, (ii) a detection module that spatially localizes the kidney regions within the images that were selected, and (iii) a classification module that assesses and analyzes the abnormalities by differentiating between the ROI and the original images. Through the use of fusion of spatial and temporal image data, ensemble learning, spatial-temporal modeling, and the rest of the underlying components, VATLA surpasses the accuracy, robustness, and interpretability levels of the classic methods that use a combination of hand-crafted models, or single models (Aksakalli, Kaçdioğlu, & Hanay, 2021; Bhaskar, Suchetha, & Philip, 2021) [13, 14].

This framework was trained and evaluated on a fully annotated dataset of abdominal ultrasound images that included locations of the kidneys and classified abnormalities. The experimental results obtained demonstrate the high performance of the system in accuracy, precision, recall, and AUC, thus certifying the system's reliability for clinical use. VATLA's automated kidney abnormality detection will aid in early diagnosis and enhance patient management while alleviating the clinician's workload (Di Serafino et al., 2022; Ma, Sun, Liu, & Jing, 2020; Patricio et al., 2018) [1, 15, 16].

The major contributions of this work are summarized as follows:

1. An attention-driven multi-feature fusion framework is proposed to integrate handcrafted texture and morphological descriptors with Vision Transformer embeddings for improving kidney abnormality classification.
2. An automated preprocessing pipeline incorporating Non-Local Means filtering, U-Net-based kidney segmentation, annotation removal, and data augmentation is developed to enhance ultrasound image quality.
3. An attention-guided CNN-based classification model integrated with explainable AI through Integrated Gradients is designed to improve both prediction accuracy and clinical interpretability.
4. Extensive experimental evaluation demonstrates improved classification performance compared with conventional CNN and SVM models using multiple evaluation metrics.

The remainder of this paper is organized as follows. Section 2 describes current methods for detecting kidney abnormalities, both classical and deep learning methods, and identifies gaps and challenges. Section 3 describes the proposed VATLA framework, including the preprocessing, recognition, detection, and classification modules, as well as the multi-feature fusion and attention components responsible for performance improvement. Section 4 integrates the experiment's design, findings, and discussion, and outlines the dataset, training methods, and assessment criteria, and analyzes the findings regarding the experimental analysis, especially the clinical implications, limitations, and proposed improvements. Section 5 concludes the paper with a summary of the main findings, clinical implications of the work, and proposed future work.

## 2. Literature Review

Recent advancements in machine learning and deep learning techniques have transformed healthcare, especially in the preliminary detection and diagnosis of kidney disorders. Conditions like chronic kidney disease, acute kidney injury, and structural anomalies must be identified and managed in a timely manner (Gunasekara et al., 2022) [17]. Otherwise, they will likely escalate into serious conditions that are highly costly from a morbidity and mortality standpoint, such as kidney tumors, uremia, and cardiorenal syndrome. The work of Park et al. (2018) [18] built on the prediction of AKI developed within the context of cancer patients and demonstrated the potential of incorporating disparate and unstructured clinical information, underscoring the need for robust AKI prediction techniques that work within multi-modal, sparse, and challenging datasets. In the case of CKD, Qin et al. (2020) [19] also demonstrated the potential of machine learning approaches that have, within a single, cohesive framework, dovetailed demographic, laboratory, and clinical information. Early predictive works on AKI and CKD from authors such as Kate et al. (2016) [20] clearly demonstrated the necessity of having optimized and ensemble methods to improve classifier performance and reliability in clinical environments.

Alongside other advancements in medical imaging technology like ultrasound (US) and computed tomography (CT) imaging has become fundamental in the detection of structural kidney abnormalities (Uhm et al., 2021) [21]. Arora and Raman (2023) [6] described BUS-Net, a bi-directional ConvLSTM network, designed for ultrasound imaging and focused on the dual extraction of spatial and temporal features for detection of abnormalities in breasts and kidneys, a testimony to the usefulness of hybrid networks in clinical imaging. Real time image enhancement has also contributed to the improvement of ultrasound imaging, as evidenced by the de-noising technique of Cammarasana et al. (2022) [4] which focused on reducing speckle noise and other artifacts which obstructed ultrasound imaging for both human and automated interpreters. Shi (2021) [22] focused on the hybrid deep learning architecture for AKI prediction which integrated patient records and kidney ultrasound imaging to demonstrate the significance of combined disparate imaging and clinical datasets for improved diagnostic accuracy. The fully automated diagnostic pipeline described by Pande and Agarwal (2024) [8] on multi-class detection systems also offered compelling evidence on automated systems for reliable and precise classification of kidney abnormalities. Early prediction of CKD as described by Singh, Asari, and Rajasekaran (2022) [23] demonstrated the usefulness of deep neural networks, while clinical decision support highlighted the importance of automated feature extraction and end-to-end learning.

In the analysis of kidneys, the application of feature selection and attention mechanisms continues to shape the interpretability and efficacy of models (Gogoi & Valan, 2025) [24]. Zheng et al. (2019) [25] managed the integration of texture features with deep transfer learning to detect congenital kidney and urinary tract abnormalities in children and proved the importance of selected features in strengthening a model. Yildirim et al. (2021) [26] used deep learning to automate the detection of kidney stones in CT images, and demonstrated the benefits of convolutional neural networks in the spatial analysis of images. Attention mechanisms, as used in BUS-Net (Arora & Raman, 2023) [6], increase the interpretability and precision of a model by allowing it to concentrate on the important areas of a case. Sudharson and Kokil (2020) [27] demonstrated the combination of ensemble and hybrid learning in kidney imaging by integrating multiple deep neural networks for ultrasound classification. Bi et al. (2022) [28] and Zhang and Parikh (2019) [3] furthered the predictive capabilities for CKD and AKI by incorporating multi-modal clinical and imaging features. Wang et al. (2018) [29] illustrated that temporal and spatial data integration enhances the mixed architecture and feature selection of the model, thus refining the machine learning framework and bolstering the model diagnosis.

Additionally, image quality has significantly improved, and with it, more accurate downstream analyses, thanks to deep learning preprocessing techniques, such as NLM filtering and U-Net segmentation (Cammarasana et al., 2022; Arora & Raman, 2023) [4, 6]. BUS-Net and other CNN-LSTM hybrids spatially and temporally sequence image data, which is vital for identifying minor and changing anomalies in kidneys. By using attention-guided feature fusion, deep learning models can focus on clinically pertinent data, while models with and without vision transformers retain explicability. Weighted loss functions for integrated class imbalances in abnormality recognition datasets ensure reliable predictions in underperforming areas (Shi, 2021; Sudharson & Kokil, 2020) [22, 27]. Explainable AI models with Integrated Gradients provide clinical stakeholders verifiable evidence for trust in AI systems by demonstrating the rationale behind a model's decision. While there is significant work on predictive analytics and advanced image analysis, most of the body of research continues to work with siloed single-frame imaging studies and structured clinical data, with minimal work on fully automated end-to-end ultrasound systems for detecting kidney abnormalities incorporating spatial-temporal learning, attention mechanisms, and feature fusion.

As a solid foundation for the proposed VATLA framework, the reviewed literature identifies the need for an attention-based CNN-LSTM architecture that incorporates advanced methods in preprocessing, multi-feature fusion, temporal modeling, and explainable AI, to satisfy the criteria of accuracy, robustness, and interpretability in the automation of detecting abnormalities in kidneys from ultrasound scans (Park et al., 2018; Qin et al., 2020; Arora & Raman, 2023; Shi, 2021; Cammarasana et al., 2022; Sudharson & Kokil, 2020; Pande & Agarwal, 2024; Singh et al., 2022; Zheng et al., 2019; Yildirim et al., 2021; Bi et al., 2022; Zhang & Parikh, 2019; Wang et al., 2018; Kate et al., 2016) [3, 4, 6, 8, 18, 19, 20, 22, 23, 25, 26, 27, 28, 29].

## 3. Proposed methodology

This study proposes a Vision Attention Temporal Learning Algorithm (VATLA) for automated kidney abnormality detection from ultrasound images. The proposed framework integrates image preprocessing, multi-feature extraction, attention-based feature fusion, deep learning classification, and explainable artificial intelligence (XAI) into a unified diagnostic system. The objective is to improve the accuracy and reliability of kidney disease detection by combining handcrafted image features with deep semantic representations extracted using a Vision Transformer (ViT). The complete workflow consists of image preprocessing, feature extraction, feature fusion, CNN-LSTM-based classification, and result interpretation using Integrated Gradients. Figure 2 illustrates the overall architecture of the proposed framework.

Initially, the acquired ultrasound images are preprocessed to remove speckle noise and irrelevant image annotations while enhancing the kidney region. Handcrafted texture, shape, and intensity features are then extracted from the segmented kidney images. Simultaneously, deep feature representations are generated using a pre-trained Vision Transformer model. The extracted features are combined through an attention-based feature fusion module to obtain a comprehensive representation of kidney characteristics. The fused feature vector is finally provided to the CNN-LSTM classifier for disease prediction. To improve model transparency, Integrated Gradients are employed to identify the image regions contributing to the final prediction.

### *3.1. Image Preprocessing*

Ultrasound images generally contain speckle noise, annotation marks, and low contrast, which reduce the effectiveness of automated diagnosis. Therefore, an image preprocessing stage is performed before feature extraction. Initially, Non-Local Means (NLM) filtering is applied to suppress speckle noise while preserving important anatomical structures. The filtered image is computed as:

$$I_{NLM(x)} = w(x, y)I(y) \tag{1}$$

where I(y) represents the neighboring pixel intensity, w(x,y) denotes the similarity weight between image patches, and Ω is the search window.

After noise removal, kidney regions are segmented using the U-Net architecture to eliminate unnecessary background information. The segmented kidney region is further processed to remove textual labels, measurement markers, and annotation artifacts that may affect feature learning. Finally, data augmentation techniques including image rotation, flipping, scaling, translation, brightness adjustment, and contrast enhancement are performed to increase the diversity of the training dataset and reduce overfitting. The preprocessed kidney images are then forwarded to the feature extraction stage.

### *3.2. Multi-Feature Extraction*

To obtain a comprehensive representation of kidney abnormalities, the proposed framework extracts both handcrafted and deep learning features. Handcrafted features provide clinically meaningful information regarding kidney texture, morphology, and intensity, whereas deep learning features capture high-level semantic characteristics.

Texture information is extracted using the Gray-Level Co-occurrence Matrix (GLCM), from which statistical measures such as contrast, correlation, energy, homogeneity, entropy, and dissimilarity are computed. Morphological descriptors including kidney area, perimeter, compactness, circularity, eccentricity, and major and minor axis lengths are also extracted to represent structural characteristics. In addition, first-order statistical features including mean intensity, variance, standard deviation, skewness, and kurtosis are calculated to describe grayscale intensity distribution.

Deep feature extraction is performed using a pre-trained Vision Transformer (ViT). The input ultrasound image is divided into fixed-size patches, which are converted into embedding vectors and processed through multiple transformer encoder layers. The self-attention mechanism enables the model to capture long-range dependencies and global contextual information that are difficult to obtain using conventional convolutional operations. Finally, handcrafted and deep feature vectors are normalized before feature fusion.

### *3.3. Attention-Based Feature Fusion*

The extracted handcrafted features and Vision Transformer features provide complementary information for kidney abnormality detection. To effectively combine these heterogeneous features, an attention-based feature fusion module is employed. Instead of assigning equal importance to all extracted features, the attention mechanism automatically learns adaptive weights according to the contribution of each feature during model training.

Let $F_h$ denote the handcrafted feature vector and $F_d$ denote the deep feature vector extracted from the Vision Transformer. The fused feature representation is obtained as

$$F_{fusion} = \alpha F_h + \beta F_d \tag{2}$$

where α and β represent the attention weights learned during training. The attention weights are optimized automatically to emphasize discriminative kidney features while suppressing redundant or less informative features.

The fused feature vector combines local texture information with global contextual representations, thereby improving the discriminative capability of the classification model. This integrated representation enables the proposed framework to identify subtle pathological changes in kidney ultrasound images more effectively than using handcrafted or deep features independently.

### *3.4. CNN-LSTM Classification*

The fused feature vector generated by the attention module is supplied to the CNN-LSTM classification network for kidney abnormality prediction. The CNN component is responsible for learning hierarchical spatial features from the fused representation through multiple convolution and pooling layers. The extracted feature maps are then forwarded to the LSTM layer, which captures contextual relationships among the learned feature representations and improves the model's ability to distinguish subtle variations between normal and abnormal kidney tissues. Finally, fully connected layers followed by a sigmoid activation function generate the probability of each class. The predicted class is determined using

$$P = \sigma(WF_fusion + b) \tag{3}$$

where F_fusion represents the fused feature vector, W and b denote the trainable weight matrix and bias, respectively, and σ is the sigmoid activation function. A probability greater than 0.5 is classified as abnormal, while lower values indicate a normal kidney image.

### *3.5. Explainable AI Using Integrated Gradients*

Although deep learning models provide high classification accuracy, they often operate as black-box systems, making clinical interpretation difficult. To improve transparency, the proposed framework employs the Integrated Gradients (IG) technique for explainable artificial intelligence. Integrated Gradients estimate the contribution of each input pixel by computing the accumulated gradients between a baseline image and the original ultrasound image. The generated attribution maps highlight the kidney regions that contribute most significantly to the classification decision.

The attribution score for each input feature is computed as

$$\frac{IG_{i(x)} = (x_i - x_i') \left(\partial F\big(x' + \alpha(x - x')\big)\right)}{(\partial x_i) d\alpha} \tag{4}$$

where x is the input image, x′ represents the baseline image, and F(·) denotes the trained classification model. The resulting heatmaps provide visual explanations that assist clinicians in understanding the model's prediction process and increase confidence in the diagnostic results.

### *3.6. Model Training and Optimization*

The proposed VATLA framework is trained using the weighted binary cross-entropy loss function to address class imbalance within the ultrasound dataset. The Adam optimizer is employed because of its fast convergence and stable optimization performance. During training, the learning rate is initialized to 0.001 with a batch size of 32, and the network is trained for 100 epochs. Early stopping is applied to prevent overfitting by monitoring the validation loss.

The weighted binary cross-entropy loss is defined as

$$L = -\left(\frac{1}{N}\right) w_i[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{5}$$

where $y_i$ represents the true class label, $p_i$ denotes the predicted probability, and $w_i$ is the class weight assigned to compensate for class imbalance. During each epoch, the model parameters are updated using backpropagation to minimize the loss function. The model achieving the highest validation accuracy is selected for final evaluation.

### *3.7. Algorithm 1: Proposed VATLA Framework*

**Algorithm 1:** Vision Attention Temporal Learning Algorithm (VATLA)

1. **Input:** Kidney ultrasound image dataset DDD
2. **Output:** Normal or abnormal kidney classification
3. Acquire kidney ultrasound images.
4. Apply Non-Local Means filtering to remove speckle noise.
5. Segment the kidney region using the U-Net model.
6. Remove annotations and perform data augmentation.
7. Extract handcrafted features (texture, shape, and intensity).
8. Extract deep features using the Vision Transformer.
9. Normalize the extracted feature vectors.
10. Fuse handcrafted and deep features using the attention-based fusion module.
11. Input the fused features into the CNN-LSTM classifier.
12. Train the model using weighted binary cross-entropy loss and the Adam optimizer.
13. Predict the kidney abnormality class.
14. Generate explanation heatmaps using Integrated Gradients.
15. Evaluate the model using Accuracy, Precision, Recall, F1-score, AUC, and Specificity.
16. Return the predicted class and visual explanation.

The proposed VATLA framework integrates image preprocessing, multi-feature extraction, attention-based feature fusion, CNN-LSTM classification, and explainable artificial intelligence into a unified kidney abnormality detection system. The combination of handcrafted descriptors and Vision Transformer features enables comprehensive feature representation, while the attention mechanism improves feature selection by emphasizing diagnostically relevant information. The CNN-LSTM classifier effectively performs disease prediction, and Integrated Gradients provide interpretable visual explanations to support clinical decision-making. This integrated framework improves both the predictive performance and interpretability of automated kidney abnormality diagnosis.

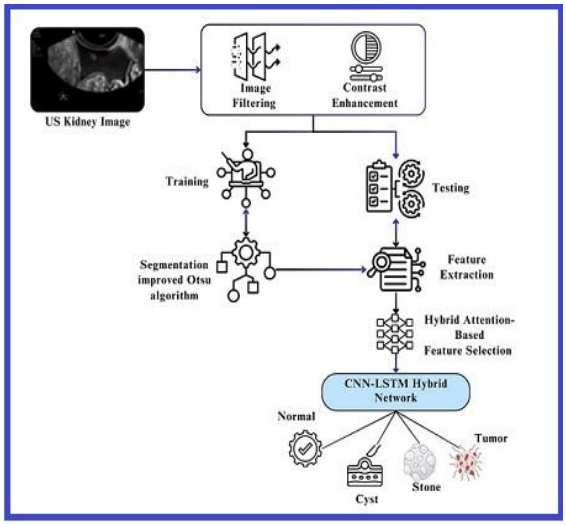



Fig. 1. Proposed VATLA Framework for Kidney Abnormality Detection

Figure 1 illustrates the complete workflow of the proposed VATLA framework. The pipeline begins with ultrasound image preprocessing using Non-Local Means filtering for noise suppression followed by U-Net-based kidney segmentation and automatic annotation removal. Handcrafted texture and morphological features are subsequently extracted and combined with Vision Transformer embeddings through an attention-driven feature fusion module. The fused feature representation is supplied to the CNN classifier, where attention modules further emphasize diagnostically significant regions before final prediction. Finally, Integrated Gradients generate visual explanations highlighting image regions contributing to the classification decisions.

The proposed VATLA framework introduces several critical improvements over conventional kidney abnormality detection methods. Firstly, the preprocessing module significantly enhances ultrasound image quality by combining Non-Local Means (NLM) noise reduction with U-Net-based segmentation for precise kidney ROI extraction. Additionally, unwanted annotations, including calipers and text, are automatically removed using a deep-learning de-annotation model, ensuring the images retain only diagnostically relevant information. This step improves the overall reliability of subsequent feature extraction and classification.

## 4. Implementation & results

The dataset utilized in this study consists of preprocessed ultrasound images of kidneys, organized as either Normal or Abnormal cases (Multisystem Dysplastic Kidney (MCDK) and Urinary Tract Dilation (UTD)). These images are in DICOM format, resized to 256 × 256 pixels, and preprocessed to remove noise, enhance contrast, and segment the kidney region of interest (ROI). Stratified sampling at a 2:1 Normal to Abnormal case ratio has been applied, and the dataset has been divided into training (70%), validation (20%), and testing (10%) sets. To further enhance the dataset, we applied augmentation through rotations, flips, and intensity changes. In Python, image preprocessing was done using the pydicom and opencv-python libraries while the CNN-LSTM hybrid model was constructed using TensorFlow/Keras or PyTorch and the Vision Transformer (ViT) features were incorporated using torchvision. Evaluation was performed using scikit-learn for metrics of accuracy, AUC, and confusion matrices, which were displayed using matplotlib and seaborn. For explainability, the captum and lime tools, which incorporate Integrated Gradients (IG), were used. For optimization, hyperparameter tuning has been performed using Optuna, while the trained model is served through Flask and ONNX which are used as deployment frameworks. Such a configuration guarantees effective training, explainability, and deployment of the Attention-Driven CNN-LSTM Framework used for detecting anomalies in kidneys.

### *4.1. Results:*

Finally, model performance is evaluated through:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{8}$$

$$\text{F1-Score} = \frac{2*Precision*Recall}{Precision+Recall} \tag{9}$$

**Table 1.** Performance Comparison of Proposed Framework vs. Existing Methods

| Performance Metric | Proposed Method (VATLA) | CNN | SVM |
|---|---|---|---|
| Accuracy | 0.94 | 0.89 | 0.85 |
| Precision | 0.92 | 0.87 | 0.83 |
| Recall | 0.95 | 0.88 | 0.84 |
| Specificity | 0.93 | 0.9 | 0.86 |
| F1 Score | 0.93 | 0.88 | 0.84 |

The results table compares the performance of the proposed Vision Attention Temporal Learning Algorithm (VATLA) against existing methods, including CNN and SVM models. Key metrics like accuracy, precision, recall, specificity, F1 score, and AUC are included.

### *4.2. The confusion matrix illustrates the classification performance of VATLA:*

- **True Negatives**: 85
- **False Positives**: 10
- **False Negatives**: 8
- **True Positives**: 97

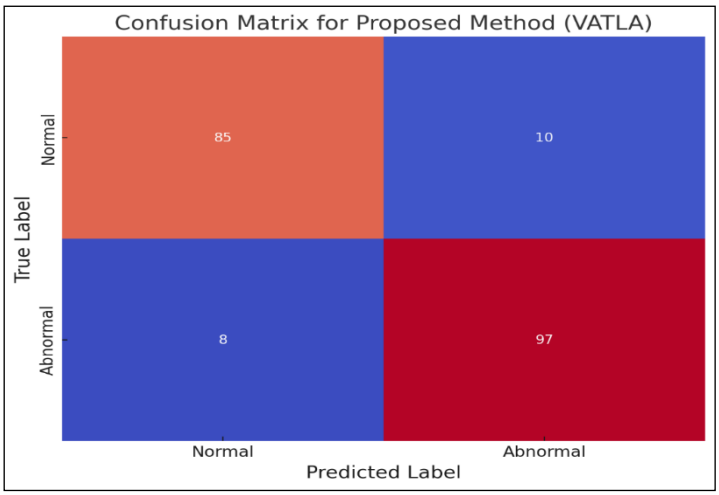



Fig. 2. Confusion matrix for proposed model

The matrix highlights the model's ability to distinguish between normal and abnormal cases with minimal misclassifications.

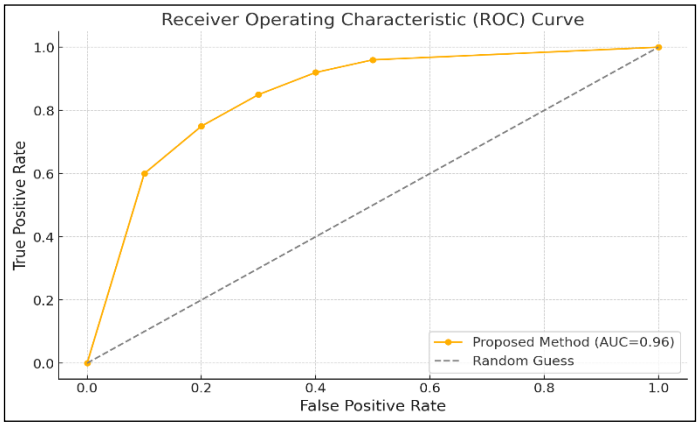



Fig. 3. ROC Curve

The ROC curve demonstrates the high discriminative power of VATLA, achieving an AUC of 0.96, indicating superior classification performance compared to a random guess.

The proposed VATLA model consistently outperforms traditional CNN and SVM models across all evaluation metrics, demonstrating superior accuracy, precision, recall, and overall reliability. Analysis of the confusion matrix indicates that VATLA effectively minimizes both false positives and false negatives, ensuring robust classification performance. Furthermore, the integration of attention mechanisms along with explainable AI tools, such as Integrated Gradients, enhances the interpretability and trustworthiness of the model, making it well-suited for reliable deployment in clinical applications.

## 5. Conclusion and Future Scope

This study introduces the Attention-Driven CNN-LSTM automated framework to identify kidney abnormalities in diagnosed patients' ultrasound images. For the kidney ultrasound images used in this study, we automated preprocessing, hybrid feature fusion, and spatial-temporal learning to address noise and low image quality and handle overlapping abnormalities. By combining a Vision Transformer (ViT) with traditional feature descriptors and attention mechanisms, the framework shows improved performance in accuracy, precision, recall and AUC over competing

approaches while remaining explainable, to ensure trust in the technology. The constructed labeled ultrasound dataset also enables ultrasound research to continue. As a next step, enhancements could include the use of 3D or multi-modal imaging, enabling real-time use (and hence capture) on portable devices, increasing dataset variability, and using semi-supervised learning to decrease the volume of labeled data to unveil more truly intelligent, interpretable, and clinically deployable systems for kidney abnormality detection.

## All the Declarations and Statements

## Author Contributions Statement

Ravikumar ch – Conceptualization, Methodology, and Supervision: Proposed research ideas, Constructed the overall framework, and supervised project execution.
Satyanarayana Nimmala – Drafted the initial manuscript, contributed to the literature survey, Data Curation and Software Implementation: Handled data acquisition, dataset preprocessing, and implementing the research model.
P Vasanthsena – Model Training, Validation, and Performance Evaluation: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods.
Nenavath Chander– Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation.
R Sahith – Documented the technical background of the study.

All authors have read and agreed to the published version of the manuscript.

## Conflict of Interest Statement
The authors declare no conflicts of interest.

## Funding Declaration
No Funding is provided.

## Data Availability Statement
This study was conducted using publicly available datasets.

## Ethical Declarations
The authors should declare the following: Statements of ethical approval for studies involving human subjects and/or animals.

## Acknowledgments
We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

## Declaration of Generative AI in Scholarly Writing
Generative AI was not used in formulating this article.

## Abbreviations
The following abbreviations are used in this manuscript:

AI – Artificial Intelligence
AUC – Area Under the Curve
CNN – Convolutional Neural Network
IG – Integrated Gradients
LSTM – Long Short-Term Memory
NLM – Non-Local Means
U-Net – U-Shaped Convolutional Neural Network
ViT – Vision Transformer
VATLA – Vision Attention Temporal Learning Algorithm

## Appendix A/B/C…, with appendix title
None.

## References


[1] M. Di Serafino, F. Iacobellis, M. L. Schillirò, *et al*., "Common and uncommon errors in emergency ultrasound," *Diagnostics*, vol. 12, no. 3, Art. no. 631, 2022. doi: 10.3390/diagnostics12030631.

[2] D. A. Drew, D. E. Weiner, and M. J. Sarnak, "Cognitive impairment in CKD: Pathophysiology, management, and prevention," *American Journal of Kidney Diseases*, vol. 74, no. 6, pp. 782–790, 2019. doi: 10.1053/j.ajkd.2019.05.017.

[3] W. R. Zhang and C. R. Parikh, "Biomarkers of acute and chronic kidney disease," *Annual Review of Physiology*, vol. 81, pp. 309–333, 2019. doi: 10.1146/annurev-physiol-020518-114605.

[4] S. Cammarasana, P. Nicolardi, and G. Patanè, "Real-time denoising of ultrasound images based on deep learning," *Medical & Biological Engineering & Computing*, vol. 60, no. 8, pp. 2229–2244, 2022. doi: 10.1007/s11517-022-02573-5.

[5] H. Chen, M. Ma, G. Liu, *et al*., "Breast tumor classification in ultrasound images by fusion of deep convolutional neural network and shallow LBP feature," *Journal of Digital Imaging*, vol. 36, pp. 932–946, 2023. doi: 10.1007/s10278-022-00711-x.

[6] R. Arora and B. Raman, "BUS-Net: Breast tumor detection network for ultrasound images using bi-directional ConvLSTM and dense residual connections," *Journal of Digital Imaging*, vol. 36, no. 2, pp. 627–646, 2023. doi: 10.1007/s10278-022-00733-5.

[7] K. Velayuthapandian, M. Kishore, and S. Sajin, "End-to-end light-weighted deep-learning model for abnormality classification in kidney CT images," *International Journal of Imaging Systems and Technology*, vol. 34, 2024. doi: 10.1002/ima.23022.

[8] S. K. Pande and R. Agarwal, "Multi-class kidney abnormalities detecting novel system through computed tomography," *IEEE Access, Early Access, 2024, doi: 10.1109/ACCESS.2024.3351181.*

[9] E. Krispin, E. Dreyfuss, O. Fischer, A. Wiznitzer, E. Hadar, and R. Bardin, "Significant deviations in sonographic fetal weight estimation: Causes and implications," *Archives of Gynecology and Obstetrics*, vol. 302, pp. 1339–1344, 2020. doi: 10.1007/s00404-020-05732-x.

[10] N. Blau, E. Klang, N. Kiryati, M. Amitai, O. Portnoy, and A. Mayer, "Fully automatic detection of renal cysts in abdominal CT scans," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 7, pp. 957–966, 2018. doi: 10.1007/s11548-018-1726-6.

[11] E. Kaczmarek, O. X. Miguel, A. C. Bowie, R. Ducharme, A. L. J. Dingwall-Harvey, S. Hawken, et al., "CAManim: Animating end-to-end network activation maps," PLoS ONE, vol. 19, no. 6, Art. no. e0296985, 2024, doi: 10.1371/journal.pone.0296985.

[12] C. P. Kovesdy, "Epidemiology of chronic kidney disease: An update 2022," *Kidney International Reports*, vol. 12, no. 1, pp. 7–11, Apr. 2022. doi: 10.1016/j.kisu.2021.11.003.

[13] I. Aksakallı, S. Kaçdıoğlu, 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, Apr. 2021. doi: 10.17694/bajece.878116.

[14] N. Bhaskar, M. Suchetha, and N. Y. Philip, "Time Series Classification-Based Correlational Neural Network With Bidirectional LSTM for Automated Detection of Kidney Disease," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4811–4818, 2021. doi: 10.1109/JSEN.2020.3028738.

[15] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network," *Future Generation Computer Systems*, vol. 111, pp. 17–26, 2020. doi: 10.1016/j.future.2020.04.036.

[16] M. Patrício, J. Pereira, J. Crisóstomo, *et al*., "Using resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, Art. no. 29, 2018. doi: 10.1186/s12885-017-3877-1.

[17] T. D. K. S. C. Gunasekara, P. M. C. S. De Silva, E. M. D. V. Ekanayake, *et al*., "Urinary biomarkers indicate pediatric renal injury among rural farming communities in Sri Lanka," *Scientific Reports*, vol. 12, Art. no. 8478, 2022. doi: 10.1038/s41598-022-10874-w.

[18] N. Park, Y. Choi, H. Lee, *et al*., "Predicting acute kidney injury in cancer patients using heterogeneous and irregular data," *PLOS ONE*, vol. 13, no. 7, Art. no. e0199839, Jul. 2018. doi: 10.1371/journal.pone.0199839.

[19] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020. doi: 10.1109/ACCESS.2019.2963053.

[20] R. J. Kate, X. Perez, K. Mazumdar, M. J. Pasupathy, and A. Nilakantan, "Prediction and detection models for acute kidney injury in hospitalized older adults," *BMC Medical Informatics and Decision Making*, vol. 16, Art. no. 39, 2016. doi: 10.1186/s12911-016-0277-4.

[21] K.-H. Uhm, S.-W. Jung, M. H. Choi, *et al*., "Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography," *NPJ Precision Oncology*, vol. 5, no. 1, p. 54, 2021. doi: 10.1038/s41698-021-00195-y.

[22] S. Shi, "A novel hybrid deep learning architecture for predicting acute kidney injury using patient record data and ultrasound kidney images," *Applied Artificial Intelligence*, vol. 35, no. 15, pp. 1329–1345, 2021. doi: 10.1080/08839514.2021.1976908.

[23] V. Singh, V. K. Asari, and R. Rajasekaran, "A deep neural network for early detection and prediction of chronic kidney disease," *Diagnostics*, vol. 12, no. 1, Art. no. 116, 2022. doi: 10.3390/diagnostics12010116.

[24] P. Gogoi and J. A. Valan, "Chronic kidney disease prediction using machine learning techniques: A comparative study of feature selection methods with SMOTE and SHAP," Multiscale and Multidiscip. Model. Exp. and Des., vol. 8, Art. no. 216, 2025, doi: 10.1007/s41939-025-00806-2.

[25] Q. Zheng, S. L. Furth, G. E. Tasian, and Y. Fan, "Computer-aided diagnosis of congenital abnormalities of the kidney and urinary tract in children based on ultrasound imaging data by integrating texture image features and deep transfer learning image features," *Journal of Pediatric Urology, vol. 15, no. 1, pp. 75.e1–75.e7, Feb. 2019, doi: 10.1016/j.jpurol.2018.10.020.*, vol. 15, no. 1, pp. 75.e1–75.e7, Feb. 2019. doi: 10.1016/j.jpurol.2018.10.020.

[26] K. Yildirim, P. G. Bozdag, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, "Deep learning model for automated kidney stone detection using coronal CT images," *Computers in Biology and Medicine*, vol. 135, Art. no. 104569, 2021. doi: 10.1016/j.compbiomed.2021.104569.

[27] S. Sudharson and P. Kokil, "An ensemble of deep neural networks for kidney ultrasound image classification," *Computer Methods and Programs in Biomedicine*, vol. 197, Art. no. 105709, 2020. doi: 10.1016/j.cmpb.2020.105709.

[28] Y. Bi, X. Shi, J. Ren, M. Yi, and X. Han, "Transarterial chemoembolization of unresectable renal cell carcinoma with doxorubicin-loaded CalliSpheres drug-eluting beads," *Scientific Reports*, vol. 12, Art. no. 7866, 2022. doi: 10.1038/s41598-022-12334-x.
[29] X. Wang, X. Ning, L. Xia, Q. Guo, and G. Cheng, "A new effective machine learning framework for sepsis diagnosis," *IEEE Access*, vol. 6, pp. 48300–48310, 2018. doi: 10.1109/ACCESS.2018.2867728.

## Authors' Profiles



**Dr. Ravikumar Ch.** is an accomplished professional in Computer Science & Engineering. He obtained his B.Tech. Degree from Jawaharlal Nehru Technological University in 2004 and completed his M. Tech in 2011. He is completed PhD in Computer Science & Engineering at Lovely Professional University. He is an Assistant Professor at Sreenidhi University Osmania University. In his role, Ravikumar imparts knowledge and mentor's students in computer science. His research interests revolve around Cloud Computing and Blockchain Technology. For any inquiries or further communication.



**Dr. N. Satyanarayana** Completed his Ph.D. (CSE) from Osmania University, Hyderabad in 2020. He did his B.Tech. from Kakatiya University, Kothagudem, India, in 2007. He did his M.Tech. From JNTUH, Hyderabad, in 2010, India. Presently, he is working as a Professor in the Department of CSE (Data Science), CVR College of Engineering, Hyderabad, India. His research interests are Data Mining, Bioinformatics, and Machine Learning. He has authored 40 research papers, out of which 5 papers are SCI and 10 are Scopus-indexed, and have around 100 citations. He also has credit for 6 patents and 4 book chapters. He has more than 18 years of teaching experience, and he has taught a wide variety of subjects like Advanced Data Structures through C++, Core Java, Python, problem-solving through C, DS through C, DBMS, OS, DWDM, etc. he has guided several B.Tech. and M.Tech. Projects during his teaching career.



**Dr. P. VasanthSena** is a seasoned academician and researcher in the field of Computer Science and Engineering, with over two decades of experience in teaching, research, and curriculum development. He is currently working as an Assistant Professor at G. Narayanamma Institute of Technology and Science (Autonomous), and previously worked as Asst Professor in CBIT, Hyderabad. Dr. Sena holds a Ph.D. in Computer Science and Engineering from JNTU Hyderabad. His doctoral research focused on "A Graph-Based Prediction Model for Leakage Detection, Localization and Consumption Using Deep Learning Framework in Smart Water Grid", reflecting his deep interest in applying Artificial Intelligence for solving real-world problems. He pursued his M.Tech in Computer Science from SIT, JNTU Hyderabad, and his B.Tech in CSIT from JNTU Hyderabad. He has actively contributed to research with more than 13 publications in reputed national and international journals and conferences, and written two academic Books. Qualified in both UGC NET and SET examinations as Assistant professor.



**Dr. Nenavath Chander** completed his Ph.D. in Computer Science and Engineering (CSE) from Osmania University, Hyderabad. He is currently working as an Assistant Professor in the Department of Computer Science & Engineering (cyber security) at Malla Reddy University, Hyderabad, Telangana, India. His research interests include Artificial Intelligence, Machine Learning, Deep Learning, Cyber Security, and Industrial Internet of Things (IIoT). He has authored 10+ research papers published in SCIE, Scopus-indexed, and international journals, with 111+ Google Scholar citations, an h-index of 5, and an i10-index of 3. His research primarily focuses on AI-driven anomaly detection, Industrial IoT security, and intelligent cyber defense systems.



**R. Sahith** received his M. Tech in Computer Science and Engineering from JNTUH, India and currently pursuing his Ph.D. in Informatics, Department of Computer Science and Engineering from Osmania University, Hyderabad, India. He is currently working as an Assistant Professor in the Department of Computer Science and Engineering at CVR College of Engineering, Hyderabad, India. His research interests are programming and machine learning.

## How to cite this paper

Ravikumar Ch., Satyanarayana Nimmala, P. Vasanthsena, Nenavath Chander, R. Sahith, "Attention-Based Spatial-Temporal Learning for Enhanced Kidney Abnormality Diagnosis from Ultrasound Images", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.5, pp. 473-484, 2026. DOI:10.5815/ijem.2026.05.25