Journal of Computer Science

Research Article

Explainable MedAttn-ResNet50 for Automated Kidney Tumor Classification from CT Images With Grad-CAM Visualization

D. Vetrithangam[1], Pramod Kumar[2], Shilpasree S[3], Akanksha Kulkarni[4], Gara Jaya Raju[5], Aswathy S Nair[6], Subramanian Selvakumar[7] and Puneet Kumar[8]

[1]*Department of Computer Science & Engineering, University Institute of Engineering, Chandigarh University, Mohali-140413, Punjab, India*
[2]*Department of Computer Science and Engineering, Ganga Institute of Technology and Management, Jhajjar, Haryana -124001, India*
[3]*Department of Computer Science and Engineering, National Institute of Engineering, Mysuru-570018, Karnataka, India*
[4]*Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed) University, Pune, India*
[5]*Department of Computer Science and Engineering, Aditya University, Surampalem, - 533437, India*
[6]*Department of Computer Applications, Marian College, Kuttikkanam Autonomous, Kerala-685531, India*
[7]*Department of Computer Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia*
[8]*Department of Computer Science & Engineering, University Institute of Engineering, Chandigarh University, Mohali-140413, Punjab, India*

Article history
Received: 04-03-2026
Revised: 23-06-2026
Accepted: 15-07-2026

Corresponding Author:
Subramanian Selvakumar
Department of Computer Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia
Email: sscseau9@gmail.com

Abstract: Kidney tumor classification from CT imaging remains challenging due to CT-specific intensity variations, tumor heterogeneity, and limited clinically meaningful interpretability in existing deep learning models. To address these limitations, we propose an Explainable MedAttn-ResNet50 framework that integrates radiology-driven preprocessing, multi-scale feature extraction, and medical attention mechanisms within a unified architecture. The preprocessing pipeline applies radiology-inspired intensity normalization, contrast enhancement, and data augmentation to preserve diagnostically relevant tissue contrast. A Multi-Scale Feature Extraction Module (MSFEM) captures tumors of varying shapes and sizes, improving robustness to heterogeneous cases, while spatial, channel, and medical knowledge–gated attention guides the network to focus on anatomically relevant regions. Grad-CAM is applied at deeper layers to generate high-resolution visual explanations and was assessed qualitatively via representative overlays. The model was trained and evaluated on 10,000 CT images and achieved 99.40% accuracy, with 99.80% sensitivity, 99.00% specificity, and 0.993 AUC, outperforming several baseline architectures. These results demonstrate that the proposed method improves predictive performance while providing clinically aligned interpretability for kidney tumor classification.

Keywords: Explainable AI, Kidney Tumor Classification, MedAttn-ResNet50, Grad-CAM, CT Imaging, Deep Learning, Attention Mechanisms, Medical Image Analysis

Introduction

Kidney cancer is one of the most common malignancies of the urinary system, accounting for nearly 2–3% of all adult cancers worldwide (Zi et al., 2024). Early detection and accurate classification of kidney tumors are crucial for making timely clinical decisions and initiating treatments such as surgery, chemotherapy, or targeted therapy (Mahmud et al., 2023). CT remains the go-to diagnostic tool for renal neoplasms, providing detailed anatomical views, tissue differentiation using Hounsfield Units, and cross-sectional renal imaging. Interpreting CT scans manually by radiologists takes time and results in

significant variability between different radiologists The challenges that make it difficult to diagnose tumors are based on the appearance of subtle differences in the appearance of the tumor; the overlap of intensities as seen on the CT scan (Gofrit and Orevi, 2016); and the variations in the size of the tumor Therefore, this may cause delays and inconsistencies in medical decisions Computer-aided diagnosis systems powered with artificial intelligence have a dual benefit they increase the accuracy of the diagnosis while reducing the workload of radiologists (Gharaibeh et al., 2022). Medical image classification is an area where Convolutional Neural Networks (CNNs) excel due to their Deep Learning architecture. The same architectures that have proven effective in medical image classification, such as ResNet, DenseNet, Transfer Learning, and Inception, have been shown to be successful in many areas, including identifying lung nodules and classifying brain and kidney tumors (Abdelrahman and Viriri, 2023; Vetrithangam et al., 2023; Chakraborty and Vetrithangam, 2023).

Transfer learning using CNNs trained on Natural Images enhanced classification performance over traditional Machine Learning techniques in prior works in Kidney Tumor Analysis. Some studies also explored attention mechanisms to identify salient tumor regions, while other studies used Hybrid Feature Extraction techniques to combine Hand-Crafted Features and Outputs from CNNs (Alonge et al., 2025). Although some progress has been made, there are still major deficiencies. Standard data processing techniques (for example, z-score) cannot account for the inherent differences in intensity of a CT scan. The second is tumor heterogeneity; due to the fact that the receptive field size of the CNN is constant, it can be difficult for the CNN to find characteristics that have been learned from one type of tumor morphology to apply to another type. The third is explainability: Although models can predict patient outcomes, they do so in a "black box" and therefore lack interpretability, which limits their clinical acceptance. Although recent attempts at creating visualizations using heat maps can aid clinicians in understanding the model's decision-making process, these visualizations are generally too broad and do not reflect how clinicians would normally diagnose patients. The integration of clinically relevant explainability, robust multi-scale feature extraction, and domain-specific data pre-processing into an architecture remains largely an open problem (Alonge et al., 2025). Although significant progress has been made in applying deep learning to CT-based tumor classification, there is still no consensus on how to integrate domain-specific adaptations with interpretability frameworks systematically. Existing approaches focus only on the accuracy rather than addressing the transparency and incorporating interpretability mechanisms. So, they lack CT-specific intensity normalization. Existing literature did not explore multi-scale tumor representations, which limits the robustness to tumors that have different sizes and whose contrast levels are different (Yang et al., 2022). So the unresolved question is how to build a deep learning model for the tasks of tumor classification that ensures robustness, interpretability, and domain adaptation. This research aims to address the aforementioned gaps by proposing MedAttn-ResNet50, an explainable deep learning architecture tailored for kidney tumor classification from CT images. The core objective is to combine the strengths of residual learning, radiology-inspired preprocessing, attention mechanisms, and multi-scale feature extraction within a unified pipeline. Specifically, radiology-inspired intensity normalization and CT-domain feature enhancement are employed to emphasize diagnostically relevant tissues. Residual learning facilitates effective and efficient training of large architectures while ensuring that the gradients can flow. The spatial, channel, and medical knowledge-gated attention mechanisms are used to improve the interpretability of the network by focusing the model's attention on the most relevant tumor regions. A Multi-Scale Feature Extraction Module (MSFEM) is developed to extract features at multiple scales to better support the modeling of tumor heterogeneity in diverse clinical cases. The significant contributions of this research are summarized as follows:

1. Domain-Specific Preprocessing: We have created a radiology-specific pre-processing pipeline for CT images, which includes: Radiology-inspired intensity normalization and contrast enhancement, min-max normalization, and data augmentation (rotations, shears, translations, zooms); so the model can be trained using the unique CT image tissue properties
2. Explainable MedAttn-ResNet50 Architecture: We have modified the ResNet-50 base model by adding batch normalization specific to CT images and medical attention module types (channel, spatial, and medical knowledge gate) to provide radiologists with an understanding of where the model is focusing in the images; therefore increasing both model transparency and the interpretability of the model's outputs
3. Multi-Scale Feature Extraction Module (MSFEM): We have proposed a new multi-branched convolutional network (1x1, 3x3, 5x5, and dilated 3x3) to extract various-sized tumor characteristics; this provides adaptability and robustness to tumor heterogeneity

4. Integration of Grad-CAM Explainability: We apply Gradient-weighted Class Activation Mapping (Grad-CAM) at deeper layers (Stage 5) to produce high-resolution heatmaps

This allows visualization of tumor-relevant areas, aligning model outputs with clinical thinking and enhancing radiologists' trust. To solve the essential issues of renal tumor CT classification, each design component of our proposed model addresses specific problems. For preprocessing specifically for CT, we apply Radiology-inspired intensity enhancement and normalization to improve contrast consistency and feature stability. The Multi-Scale Feature Extraction Module (MSFEM) captures the tumor heterogeneity from different receptive-field scales. Here, tumor heterogeneity refers to the size, texture, and boundary appearance of each tumor. Ultimately, interpretability is enhanced through the MedAttn (medical knowledge-gated attention) block, which directs the network's focus to clinically relevant areas, and Grad-CAM, which highlights the image areas that affect the model's decision.

To the best of our knowledge, few previous studies have jointly integrated radiology-inspired preprocessing, medical attention, multi-scale feature extraction, and Grad-CAM-based explainability within a unified framework for kidney tumor classification.

Literature Review

This section reviews existing literature on machine learning and deep learning models used for kidney tumor classification. Examining their strengths and limitations lays the groundwork for the proposed explainable MedAttn-ResNet50 model.

Alzu'bi et al. (2022) developed deep learning models to detect and classify kidney tumors in CT scans. CNN-6 & ResNet50 both performed extremely well in terms of their accuracy, which were both at 97 and 96%, respectively. The four-layer CNN-4 model had an accuracy of 92%. VGG16 was the least accurate of all the models tested, with a low accuracy of 60%. Custom CNNs are able to classify images that have moderate complexity, using the images to determine the diagnostic accuracy of the images; this can potentially help reduce the workload of radiologists and also potentially increase the speed at which radiologists can intervene in the patient care process. Majid et al. (2023) conducted research utilizing machine learning to create a kidney tumor detection model in CT scans using fine-tuning and transfer learning techniques. This research work used many types of classifiers, such as Random Forest and SVM, as well as deep neural networks, ResNet-101, and DenseNet-121. We utilized three different resolutions in our images (32x32, 64x64, 128x128) to evaluate the use of GLCM feature extraction and found that the fine-tuned ResNet-101 and DenseNet-121 achieved a high accuracy of 94.09% when the images were evaluated at 32x32 resolution. This is an improvement over previous studies that suggest that transfer learning is effective for rapid and cost-effective kidney tumor classifications. However, we are concerned regarding the lack of information regarding the dataset as well as the decrease in performance of the models when the image resolution increased; therefore, it is necessary to continue evaluating the performance and real-world application of these models. Mehedi et al. (2022) suggested a two-stage deep learning system for kidney tumor classification. In the initial phase, UNet and SegNet were employed for the manual segmentation of kidneys within CT scans. MobileNetV2 and InceptionV3 models were then trained on the CT kidney dataset, which contained normal, cyst, tumor, and stone classes. MobileNetV2 and InceptionV3 achieved peak test accuracies of 95.29 and 97.38%, respectively, exhibiting robust sensitivity and specificity, while Grad-CAM was employed to enhance model interpretability. High classification accuracy and improved interpretability through explainable AI are the main advantages. Manual segmentation has limitations in scalability. Moreover, using one publicly available dataset with no external validation limits its generalizability and clinical applicability.

Li et al. (2023) developed radiomics to build a binary logistic regression model that automatically differentiates kidney tumors from normal tissue on arterial phase CT, which they validated on 210 KiTS19 challenge cases. This study defined three radiomic signatures from gross tumor and contralateral renal parenchyma volumes with performance exceeding 85% in both training and validation cohorts using an automatic streamlined pipeline. The findings here may not apply elsewhere, as the dataset was only from one institution and may include selected features without disclosing external validation and no stratification by tumor histology. Martín and Sánchez (2025) conducted a multi-center study leveraging MRI and CT image datasets. They evaluated the performance of the Swin and MaxViT Transformer architectures for classifying brain, lung, and kidney tumors as malignant or benign. The Swin Transformer demonstrated incredible performance, as reported in a study. It achieved 99% or better accuracy on each dataset alone as well as 99.4% on the entire set of data. Due to the system's ability to process both imaging information, diagnostic reliability was increased, and it shows prospects for multimodal, multi-class tumor analysis. The results suggest that the technology might find application in clinical practice in the near future. Nonetheless, the technology still relies on a small amount of annotated data, and interpreting transformer models is difficult. That's why further validation and explanatory activity will need to take place before clinical use. Kaur et al. (2024) created a CNN model to classify kidney images as normal

(non-tumor) and tumor kidney images.

The optimizer used for the CNN model was Adam, and Data Augmentation was also used to make the model generalize better on unseen data. The CNN showed good performance on the Training Set and Validation Set with an Accuracy of 97.69% and 95.31%, respectively. The model's precision and recall both scored 0.97. This methodology seems like it has a lot of potential, but there are many limitations. First, the small size of the dataset makes it hard to consider. Second, this methodology requires external validation as well as more sophisticated model designs. In other words, the constraints suggest that this methodology will need additional development using larger datasets as well as improved architectures before it can be trusted in practice. Sudharson and Kokil (2020) employed an ensemble classifier leveraging transfer learning. This work employed neural networks such as ResNet-101, ShuffleNet, and MobileNet-v2. The networks were tolerant to any distortion in the nature of the images and combined the features found by classifiers such as SVM and Majority voting for the classification of the ultrasound images of the kidneys, normal and cysts, stone or malignancy, and improved still further by the use of a perception evaluator. The quality of the image thus produced was very good even in the presence of noise. This method claimed to have produced better results than published so far, with deep learning and classical techniques having been used. The model showed an accuracy of 96.354% with clean data and 95.58% in the presence of noise. These points point to the reliability of the technique and its independence from noise, which is so necessary in a clinical setting. The model showed good reliability with a wide range of the qualities of the images and number of classes, but an extreme deficiency shows a difficulty where the classes of the sets also remain to be tested, and the clinical need for acceptance by further testing and assessment, and practical use is still required. A multi-layer CNN for the automation of CKD detection and classification using ultrasound is presented by Hannan and Pal (2023). Features are selected, and predictions are given in batch form, where an accuracy of 87·4% is achieved. This kind of system would reduce the radiologists' workload and help them deliver faster diagnoses. Our results show moderate accuracy and less than conventional forms. However, the generalizability of this technique was not possible owing to the lack of availability of the dataset, its requirements and details, as well as criteria such as sensitivity and specificity, and more validation and reporting needs to be carried out before any such batch processing technique could be employed in real clinical practice.

Recent advances in CT image denoising have demonstrated the effectiveness of deep learning approaches for preserving diagnostically important structures. Zhang et al. (2025) reviewed CNN-, Transformer-, and diffusion-based denoising methods and highlighted the importance of diagnosis-oriented image restoration. Unlike these approaches, the proposed Explainable MedAttn-ResNet50 framework focuses on robust feature learning without incorporating an explicit denoising stage. Residual learning architectures have shown superior capability in preserving anatomical structures while reducing noise. Eulig et al. (2025) investigated the behavior and robustness of low-dose CT denoising networks and highlighted the importance of interpretability in clinical applications. Unlike these approaches, the proposed Explainable MedAttn-ResNet50 framework focuses on robust feature learning and attention-guided classification without employing an explicit denoising stage.

A method based on a two-stage decision-theoretic approach was proposed by Kumar et al. (2024) for the diagnosis of kidney tumors. Initially, MobileNetV3 was used for the classification of the available CT images as either normal or cancerous, and this achieved accuracy and precision values of 99.1% and 99%, respectively. U-Net (custom)-based segmentation was then used for tumor areas, followed by training to obtain an average value of the Dice coefficient of 0.9445. This two-stage modular approach will improve the efficiency and accuracy of the diagnostic procedure, as tumor-positive cases will benefit from segmentation. While prior attention and Grad-CAM studies offer visual explanations, they are often not clinically aligned with CT tumor analysis because they use generic preprocessing and non–domain–aware attention. In contrast, our method integrates radiology-inspired preprocessing, medical knowledge–gated attention, MSFEM for multi-scale tumor heterogeneity, and deeper-layer Grad-CAM placement to provide more clinically meaningful explanations.

Materials and Methods

This section outlines the preprocessing steps for CT kidney images and introduces the MedAttn-ResNet50 methodology. The methods allow for consistent input, reliable feature extraction, and understandable classification results.

Dataset

The dataset contains 10,000 CT scan images, where each of the two classes (normal and diseased) has 5,000 images of the kidney. All of the images presented were acquired in the coronal plane, providing a complete cross-section of the abdomen. The images are high-contrast grayscale pictures, much like one would find in a clinical computed tomography scan. Bilateral kidneys, surrounding soft tissue, and adjacent abdominal organs, along with the vertebral column, are easily visualized within the images. The imaging

parameters are consistent and provide excellent tissue contrast, making it possible to distinguish between normal kidney tissue and pathological areas easily. The images encompass the entire kidney, apex to base, allowing for easy identification and location of pathology. The grayscale intensity values of the images represent the full dynamic range of the images; bone structures are depicted as hyperintense (white), soft tissues appear in intermediate gray levels, and air/gas appears hypointense (black). The images are captured at the same spatial resolution and using the same window/level setting for viewing soft tissue within the abdomen. The dataset contains a wide variety of presentations of pathological conditions in the tumor-positive group, whereas the normal group depicts normal kidney anatomy without masses, cysts, or any other type of abnormality. As such, this well-balanced and high-quality dataset is an excellent choice for training and testing deep learning algorithms designed to detect and classify kidney tumors. Scanner vendor/protocol details (kVp, slice thickness, reconstruction kernel) were unavailable in this dataset, which contained JPEG files. The complete dataset containing 10,000 CT images was partitioned into training, validation, and test sets consisting of 7,000, 2,000, and 1,000 images, respectively, corresponding to a 70:20:10 ratio. The test set was held out during training and used exclusively for final performance evaluation. Consequently, the confusion matrix and all reported performance metrics were computed using the independent test set of 1,000 images.

The dataset is available at the source link: https://www.kaggle.com/datasets/d243prasannaasole/brain-kidney-and-oral-cancer-dataset.

Preprocessing

All CT slices were preprocessed before training to ensure uniformity and to emphasize diagnostically relevant information:

Resizing

Each grayscale slice was resampled to 224×224 pixels via bilinear interpolation. This fixed input size matches standard CNN backbones (e.g., ResNet-50, VGG) and preserves sufficient anatomical detail of the kidney and any tumors while removing excess background.

Intensity Contrast Enhancement

Since the dataset consists of JPEG images rather than raw DICOM scans, absolute Hounsfield Unit values are unavailable. Therefore, conventional HU windowing cannot be performed directly. Instead, radiology-inspired intensity enhancement based on grayscale intensity distributions and tissue appearance characteristics is employed to emphasize diagnostically relevant structures and suppress irrelevant background variations.

Intensity Normalization

After clipping, pixel values x were linearly rescaled using min–max normalization as shown in Equation (1). This method scales input values to the range [0,1] (or sometimes [–1,1]). It is well-suited for CNNs with ReLU activations, which perform better with positively bounded inputs, and works efficiently on imaging datasets that are relatively homogeneous in acquisition settings and windowing. The main advantages of this approach are its simplicity, stability, and ability to preserve relative contrast; however, its limitation is that new test images with values outside the training range may be clipped, potentially leading to loss of intensity information:

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

Channel Handling

Since the model accepts single-channel input, we directly fed the normalized grayscale image. (Alternatively, one could duplicate the single channel to three channels for pretrained RGB backbones.

Data Augmentation

To take into account possible variability of acquisitions and to improve the generalization of results, we used random transformations with a low risk of distorting anatomy, applied randomly at each epoch during training (zoom scale of 10–20%, flip (horizontal/vertical), ±15° rotation, shear, and translation). Thus, it was possible to simulate different patient positioning and tumor sizes.

Robustness to Image Variability and Artifacts

Because the dataset consists of diagnostic-quality CT images, no explicit denoising stage was applied. Nevertheless, in lower-dose or noisy clinical scenarios, conventional filters or deep learning–based denoising methods may be incorporated prior to classification. Uniform preprocessing of images leads to uniform size, uniform intensity, and reduced image variability; this improves training stability and classification accuracy in MedAttn-ResNet50.

Residual Learning Backbone

MedAttn-ResNet50 employs the ResNet-50 backbone for residual learning to resolve vanishing gradients that occur in deep neural networks. The network is trained on residual learning, rather than learning a direct mapping between input data x and output y of the form *H(x):*

$$F(x) = H(x) - x, H(x) = F(x) + x \quad (2)$$

The above representation signifies where x is the input,

H(x) is the intended mapping, and *F(x)* is the residual mapping. In addition, the "identity" skip connection in the network preserves information to enable better gradient flow through each iteration, resulting in faster convergence. Four main stages of ResNet-50. Residual bottleneck blocks are formed in each stage:

- Stage 2: 3 blocks producing 256 channels
- Stage 3: 4 blocks producing 512 channels
- Stage 4: 6 blocks producing 1024 channels
- Stage 5: 3 blocks producing 2048 channels

From basic textural elements to intermediate organ outlines to the complex semantic features that are required to identify tumors, the learning process is developed.

Attention Mechanisms (MedAttn Block)

After Stage 5, a dedicated attention module (MedAttn block) is introduced to enhance feature discrimination and enable the network to focus on clinically relevant structures. The module consists of three complementary components: Spatial attention, channel attention, and a Medical Knowledge Gate.

Channel attention adaptively reweights feature channels according to their importance, using global average pooling followed by fully connected layers. Spatial attention highlights informative regions while suppressing irrelevant background information. In this work, the medical prior information *M* is derived from CT-domain intensity patterns and tissue appearance characteristics preserved in the grayscale JPEG images. Although Hounsfield Unit values are not directly available in JPEG data, the intensity distributions and texture information retained in the images provide clinically relevant cues for distinguishing tumor and normal tissues. Consequently, the Medical Knowledge Gate utilizes these radiological characteristics to guide feature refinement toward diagnostically important regions. The first channel attention. Global pooling and fully connected layers are integrated with the domain knowledge using the Medical Knowledge Gate, and Equation (3) summarizes the unified feature representation that comes out of this step. The final feature representation is obtained by combining the outputs of spatial attention, channel attention, and the Medical Knowledge Gate as:

$$F_{\{out\}} = F \odot A_s(F) \odot A_c(F) \odot A_m(F) \quad (3)$$

Where A_s, A_c, and A_m denote spatial, channel, and medical gate attention, respectively:

$$A_m(F, M) = \text{Sigmoid}(W_m[GAP(F) \parallel M] + b_m) \quad (4)$$

- F denotes the input feature map
- $A_s(F)$ and $A_c(F)$ represent spatial and channel attention, respectively
- M denotes the medical prior information (e.g., grayscale intensity distributions and tissue appearance characteristics preserved in JPEG images)
- $GAP(\cdot)$is global average pooling
- $\parallel$denotes feature concatenation
- $W_m$and $b_m$are learnable parameters
- $\sigma(\cdot)$is the sigmoid activation function

Thus, the complete attention formulation becomes:

$$F_{out} = F \odot A_s(F) \odot A_c(F) \odot \sigma(W_m[GAP(F) \parallel M]+b_m) \quad (5)$$

Which explicitly shows that the Medical Knowledge Gate incorporates CT-domain prior information M into the attention mechanism, distinguishing it from conventional SE or CBAM attention. The Medical Knowledge Gate is designed to incorporate CT-domain priors into the attention process. Let *M* denote prior information derived from grayscale intensity distributions and tissue appearance characteristics preserved in the JPEG images. After global average pooling of the feature map *F*, the pooled features are concatenated with *M* and passed through a learnable linear transformation followed by a sigmoid activation to generate the medical attention weights. Through the incorporation of these radiological cues into the gating process, the network is directed towards focusing on diagnostically relevant regions while suppressing irrelevant background structures. In contrast to standard SE or CBAM attention mechanisms that simply derive attention weights from learned features, the Medical Knowledge Gate proposed here utilizes meaningful CT image characteristics from the domain to modify the gating signal. This, in turn, improves feature enhancement and model clinical interpretability.

Multi-Scale Feature Extraction (MSFEM)

The Multi-Scale Feature Extraction Module (MSFEM) is used to capture complementary features at different receptive-field scales since the kidney tumors significantly differ from each other in size, texture, and appearance of boundaries. The part includes four parallel convolution branches having kernel configurations of 1×1, 3×3, 5×5, and dilated 3×3. The 1×1 branch helps with preserving fine local details, while the 3×3 and 5×5 convolutions gather medium- and large-scale information, respectively. The enlargement of the 3×3 branch without an increase in parameters is proposed to allow the obtaining of a larger anatomical context.

Let Fi denote the feature maps generated by the i^{th} branch, and w_i represent the corresponding learnable fusion weights. Equation (4) presents the computation of the multi-scale feature representation.

$$F_{MSFEM} = \sum_{i=1}^{4} w_i F_i, \sum_{i=1}^{4} w_i = 1 \quad (6)$$

The outputs of 1×1, 3×3, 5×5, and dilated 3×3 branches are denoted by F1, F2, F3 and F4, respectively. The maps of features are then combined through a 1×1 convolution to achieve a compact 512-channel map. The adaptive fusion mechanism allows the network to model tumor heterogeneity efficiently and enhances robustness against tumor shape and size variations. We apply a weighting scheme on the features per convolutional branch, F_i, which uses learned weights, w_i, that sum to one. This mechanism enables the adaptability of tumor feature recognition across scales.

Classification Head

After performing multi-scale fusion, Global Average Pooling (GAP) reduces the spatial feature maps into a small vector as defined in (5):

$$v = GAP(F) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} F(i,j) \quad (7)$$

The feature vector is then passed through ReLU-activated fully connected layers with dropout for regularization by the classification head. As can be seen from Equation (6), the output of the final Softmax layer represents probabilities for binary classification (normal vs tumor):

$$P(y = c) = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}} \quad (8)$$

Training Strategy and Hyperparameters

As depicted in Figure 1, the proposed MedAttn-ResNet50 model was trained with well-tuned hyperparameters to attain stability and clinical reliability. Each CT slice was resized to 224×224 pixels and subjected to radiology-inspired intensity enhancement followed by min–max normalization to [0,1]. In order to enhance the generalization of the studied model, data augmentations such as flips, small rotations (±15°), zooming, shearing, and translations were carried out on the original data. We trained with a batch size of 32 for 50 epochs with AdamW for the initial learning rate of 1e-4, weight decay of 1e-4, and cosine annealing with warmup to gradually increase the learning rate. The loss function employed was binary cross-entropy with logits and label smoothing (0.05) for better calibration. To avoid overfitting and unstable updates, early stopping with a patience of eight epochs and gradient clipping (norm 1.0) are applied. Moreover, mixed-precision training (FP16/AMP) is used for speeding up the computation.

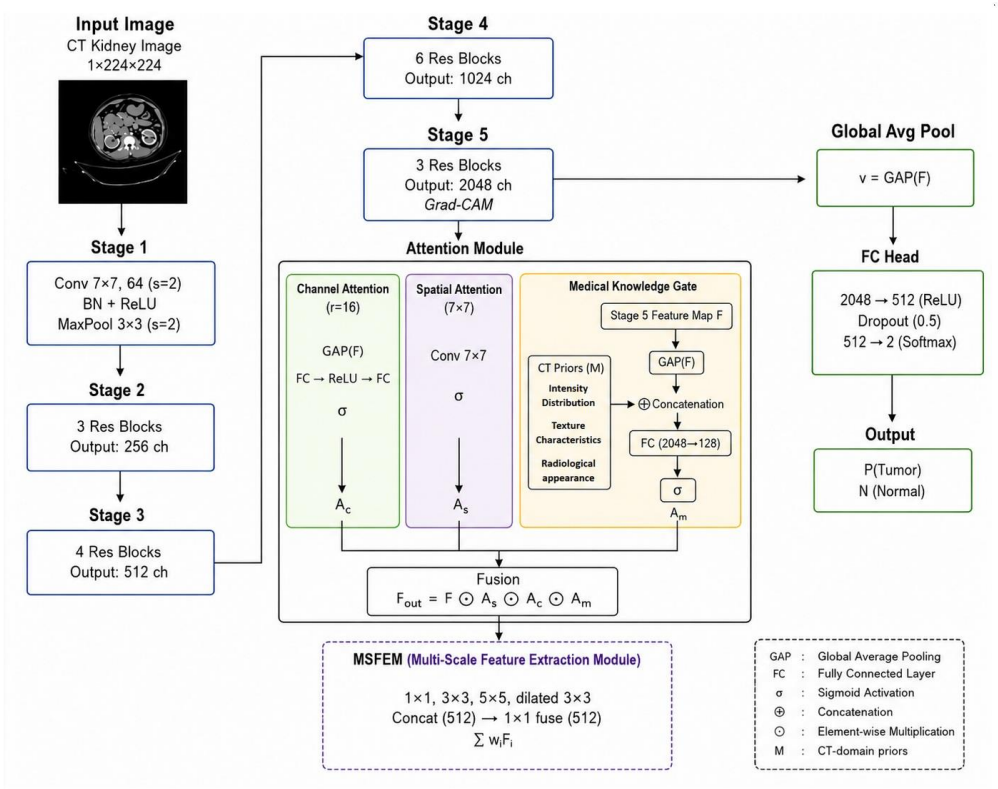

Fig. 1: Proposed Explainable MedAttn-ResNet50 architecture for automated kidney tumor classification

Within the architecture, the attention modules were defined to have a channel reduction ratio of 16, a spatial attention kernel size of 7×7, and the medical gate driven by global pooling with a dropout of 0.1. Our Multi-Scale Feature Extraction Module (MSFEM) implements a distinct approach. This module incorporates 1 x 1, 3 x 3, and 5 x 5 convolutions, along with a dilated 3 x 3 convolution, in parallel within a single CNN block. It then fuses these diverse features using softmax-normalized learnable weights. The resulting output is a distinguished 512-channel feature representation. In contrast to standard multi-scale CNN blocks that concatenate features uniformly, our approach allows for adaptive scale selection at each input by the network. Global average pooling was utilized before the two fully connected layers (2048→512→2) with ReLU activation and 0.5 dropout just before the softmax prediction presented in the Classification Head. The evaluation of the model was done using performance metrics AUC, F1-Score, Sensitivity, Specificity, Precision, Accuracy, etc., on the Validation Set by tuning the decision threshold for maximizing the F1-Score. All these hyperparameters, when used together, ensured the model was robust, interpretable, and performed well on the dataset for an Automated Kidney Tumor Detection System.

Explainability With Grad-CAM

To make the model explainable, Grad-CAM was applied at the output of Stage 5 (2048-channel feature maps). Grad-CAM computes gradients of the target class with respect to convolutional feature maps, producing class-discriminative heatmaps. These heatmaps are superimposed on the input CT slices, highlighting tumor regions that contributed to classification. This step ensures transparency and clinical trust in the model's predictions.

Evaluation Metrics

Classification tasks are almost always evaluated by means of accuracy. It measures the percentage of correctly classified samples with respect to the total samples. In medical imaging classification, accuracy indicates how well the model differentiates between tumor and normal kidney CT slices, with accuracy defined as in (7):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

Where:

- TP (True Positives): Tumor images correctly classified as tumor
- TN (True Negatives): Normal images correctly classified as normal
- FP (False Positives): Normal images incorrectly classified as tumors
- FN (False Negatives): Tumor images incorrectly classified as normal

Sensitivity is a metric that was established specifically to measure a classifier. As shown in Equation (8), recall is the number of actual positives we correctly identified through our model:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (10)$$

As stated in Equation (9), specificity (true negative rate) represents the ability of a model to identify normal cases correctly. Specifically, it is defined as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (11)$$

The AUC (Area under Curve) summarizes the ROC curve, giving a general overview of the performance of a binary classifier. This measures the probability that a random positive sample will obtain a higher score than a randomly selected negative sample according to model predictions. Using the trapezoidal rule, the formula for AUC is shown in equation (10):

$$\text{AUC} = \sum_{i=1}^{n-1}(\text{FPR}_{i+1} - \text{FPR}_i).\frac{(\text{FPR}_{i+1}+\text{FPR}_i)}{2} \quad (12)$$

Where TPR[i] and FPR[i] represent the TPR and FPR values at the i^{th} point on the ROC curve, and n is the total number of points on the curve.

Results and Discussion

The method has been implemented and validated on a 4 GB GPU with 12 GB RAM and 1 TB storage using Python's TensorFlow, PyTorch, NumPy, and scikit-learn. In total, the Explainable MedAttn-ResNet50 training and validation accuracy is about 50 epochs. As shown in Figure 2, the training/validation accuracy curves of our proposed Explainable MedAttn-ResNet50 model are shown. 50 epochs. Both curves show a smooth increase during training and converge after almost thirty epochs, which indicates a stable learning process. The close match of training accuracy and validation accuracy suggests that the model does not overfit and generalizes well to data it has not seen. The mini-batch stochasticity and data augmentation cause minor oscillations in the curves. The proposed framework for automated kidney tumor classification demonstrates high robustness with a validation accuracy of around 99.40%.

The MedAttn-ResNet50 model achieved 99.40 % accuracy, as per the confusion matrix shown in Fig. 3. Further, it is an explainable model, which makes it easy to understand the decisions it takes. Also, the validation was done with 1000 CT kidney images.

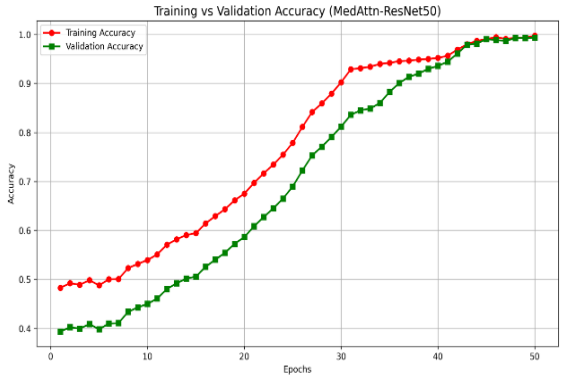

Fig. 2: Training and Validation Accuracy of the Proposed Model

This is taken from a bigger database of 10000 images, which are split into 7000 training images, 2000 validation images, and 1000 test images. The confusion matrix obtained from the independent test set of one thousand CT images is shown in Figure 3. Although the dataset contains 10,000 images in total, only the held-out test samples were used for final evaluation to ensure unbiased evaluation of performance. The model predicted 495 of 496 tumor images and 499 of 504 normal images correctly. Thus, one false negative and five false positives were noted. The presented results exhibit high sensitivity along with a low false alarm rate. Thus, the proposed framework is potentially applicable for the automated classification of kidney tumors.

Figure 4 presents the Receiver Operating Characteristic (ROC) curve of the proposed Explainable MedAttn-ResNet50 model for tumor vs. normal classification. ROC curve plots TPR (TPR / sensitivity) against FPR (FPR = 1 − specificity) through decision thresholds. The analysis of the ROC curve yields a threshold-independent measure of discrimination performance.

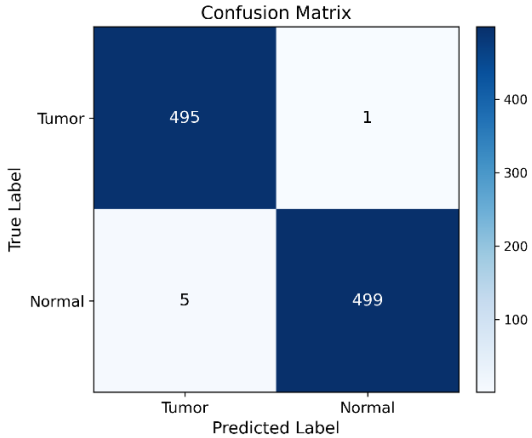

Fig. 3: Confusion matrix of the proposed Explainable MedAttn-ResNet50 model

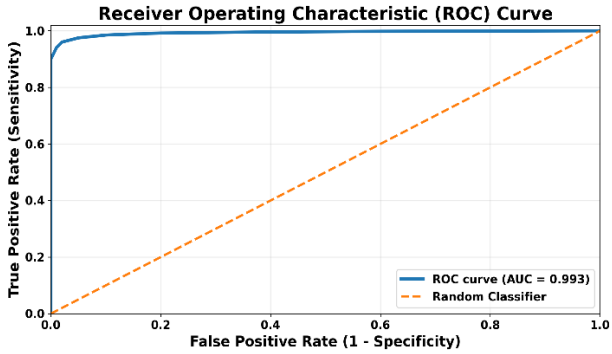

Fig. 4: ROC curve of the proposed Explainable MedAttn-ResNet50 model for tumor classification

The ROC curve remains close to the upper-left corner across a wide range of thresholds, indicating that the proposed model achieves a high true positive rate while maintaining a low false positive rate. The high AUC value of 0.993 demonstrates excellent discrimination capability and suggests that the model is robust to threshold variations. This behavior reflects the effectiveness of the MedAttn block and MSFEM in extracting discriminative tumor representations while minimizing false alarms. As shown in the confusion matrix, the proposed model correctly identified 495 of 496 tumor images and 499 of 504 normal images, resulting in one false negative and five false positives. Accordingly, the model achieved an accuracy of 99.40%, a sensitivity of 99.80%, and a specificity of approximately 99.0%, as presented in Table 1. Likewise, the ROC analysis shows AUC is approximately 0.993, which means a global diagnosis is excellent. In practical terms, the high AUC indicates that the Explainable MedAttn-ResNet50 model performs well over a large range of thresholds, making it a reliable option for tumor identification that minimizes false positives and false negatives. Figure 5 presents representative kidney CT images together with their corresponding Grad-CAM overlays. The side-by-side comparison demonstrates that the proposed Explainable MedAttn-ResNet50 model effectively focuses on anatomically significant regions, particularly tumor-affected areas and lesion boundaries, while suppressing irrelevant background information. Analysis of the Grad-CAM visualizations indicates that most correctly classified cases exhibit strong activations over clinically relevant regions.

Table 1: Performance of the proposed Explainable MedAttn-ResNet50 on the test set

Metrics	Value (%)
Accuracy	99.4
Sensitivity (Recall / TPR)	99.8
Specificity (TNR)	99
F1-Score	99.4
AUC	0.993

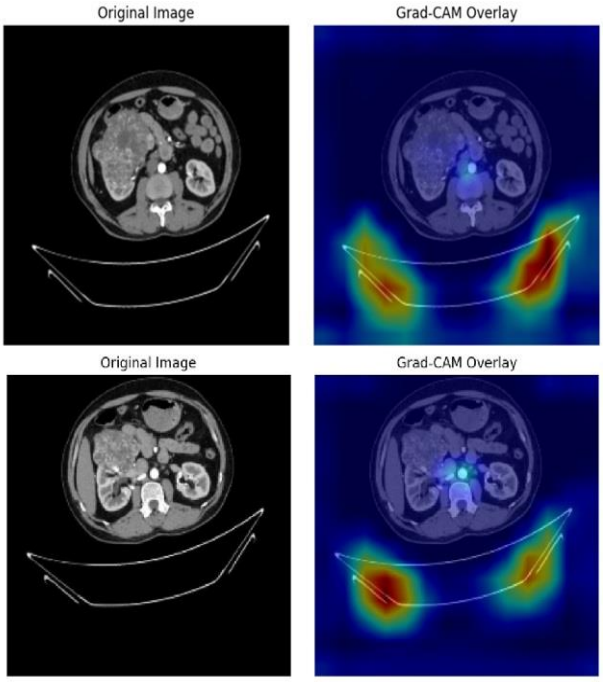

Fig. 5: Grad-CAM visualizations of kidney CT images using the proposed Explainable MedAttn-ResNet50 model

The few misclassified cases are mainly associated with low-contrast lesions, ambiguous boundaries, or image artifacts, which reduce the distinction between tumor and normal tissues. Overall, the Grad-CAM overlays confirm that the model consistently attends to diagnostically meaningful regions, thereby enhancing the interpretability and potential clinical utility of the proposed framework.

As shown in Table 2, the ablation results show that the proposed MSFEM can significantly boost performance over a single-scale 3×3 convolution block. This gain is qualitatively expected because the kidney tumors are highly heterogeneous in their size, texture, and boundary appearance. MSFEM, which integrates multi-scale receptive fields (RSs) and adaptively fuses them, is able to detect fine edges of lesions while capturing a broader anatomical context to avoid confusion with normal structures, which improves its sensitivity and specificity. Unlike the single-scale ablation model, the proposed model represents tumors of varying scales, reducing misclassification. Compared with the ablation model without MSFEM, the proposed Explainable MedAttn-ResNet50 achieves clear gains of +3.8% accuracy, +3.4% sensitivity, +4.2% specificity, and +0.043 AUC, demonstrating the strong contribution of MSFEM to overall classification performance. Kidney tumors are one of the most common cancers in adults. The model developed has high sensitivity, which will reduce the number of missing tumor cases, and its high specificity will limit false alarms that result in unwanted follow-up examinations. The last few tasks remaining are problems that occur in rare cases, such as low-contrast lesions, tumor boundaries, partial volume effect, or image artifacts. These are also relatively difficult to interpret for experts. The radiology-inspired intensity preprocessing and normalization help mitigate intensity variations across scans. Also, the use of MSFEM helps represent the heterogeneous appearance of the tumor by learning features at multiple receptive-field scales. Finally, the MedAttn helps to attend to clinically relevant regions as per Grad-CAM overlays. However, Grad-CAM validation in this study is qualitative; future work will include expert ROI-based localization evaluation (e.g., IoU/Dice) and multi-institutional testing to confirm generalizability across scanners, acquisition protocols, and tumor subtypes. Our Explainable MedAttn-ResNet50 shows great performance when comparing against the best methods of kidney tumor classification, as seen in Figure 6. Although CNN-6 and ResNet50 are capable of high-level predictive abilities, the lack of transparency and their variability dependent on medical image data limit them from being used clinically.

Similarly, other complex networks such as DenseNet-121 (95.10%), MobileNetV2 (95.29%), and sequential CNN models (95.31%), while providing a great level of accuracy, were also deemed unsuitable for clinical use. As we can see from Figure 6, InceptionV3 had a slightly better accuracy than all the ensemble models, with an accuracy of 97.38% vs 96.54%, respectively. However, Swin Transformer and MobileNetV3 provide high predictive performance but offer limited interpretability compared with the proposed framework. Our proposed Explainable MedAttn-ResNet50 was able to perform very well on both validation and test sets, with a validation accuracy of 99.40%. The integration of multiple modules that focus on medical attention, multi-scale feature extraction, and Grad-CAM explainability improved both predictive accuracy and clinical interpretation, making it applicable for use in real-world diagnostic workflows.

Table 2: Ablation results of Explainable MedAttn-ResNet50 with and without MSFEM

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Proposed model (Explainable MedAttn-ResNet50)	99.4	99.8	99	0.993
Ablation model:				
Without MSFEM (replace MSFEM with a single standard 3×3 conv block)	95.6	96.4	94.8	0.95
	3.8	3.4	4.2	0.043

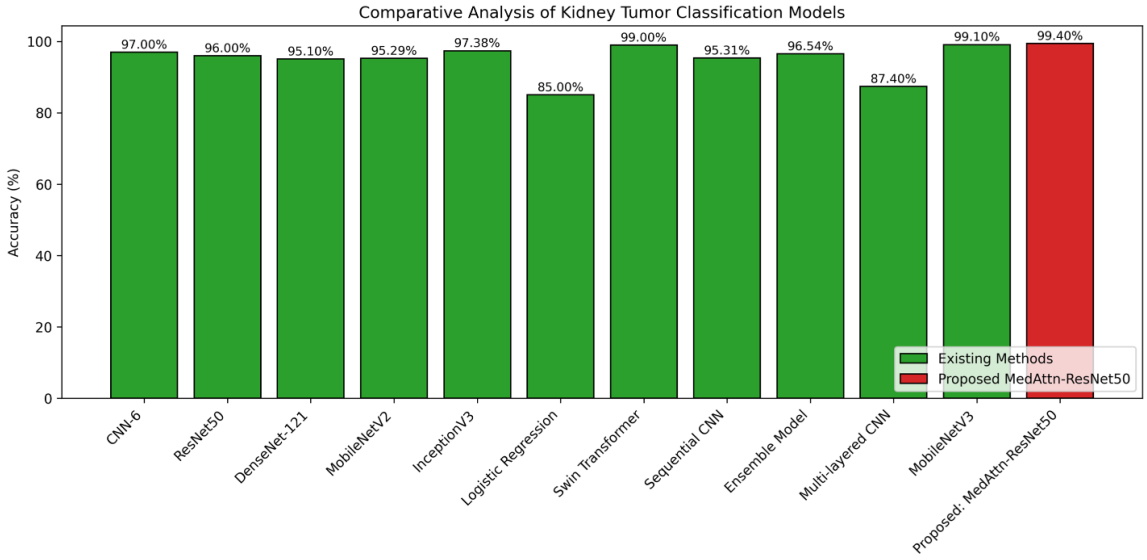

Fig. 6: Comparative accuracy of existing deep learning models and the proposed Explainable MedAttn-ResNet50 model for kidney tumor classification

Conclusion

In this study, we proposed an Explainable MedAttn-ResNet50 framework for automated kidney tumor classification from CT images by integrating CT-domain intensity-based preprocessing, a Medical Attention (MedAttn) block, a Multi-Scale Feature Extraction Module (MSFEM), and Grad-CAM-based visual explanations. The proposed framework effectively captures tumor heterogeneity while providing clinically meaningful interpretability. Experimental results on a dataset of 10,000 CT images demonstrated superior performance, achieving 99.40% accuracy, 99.80% sensitivity, 99.00% specificity, and an AUC of 0.993. The major contributions of this work include:

i) The introduction of a domain-aware medical attention mechanism for enhanced feature discrimination
ii) The development of an MSFEM module for capturing multi-scale tumor characteristics
iii) The integration of Grad-CAM visualizations to improve model transparency and clinical interpretability

Despite these promising results, the current framework is limited to binary classification and qualitative interpretability assessment. Future work will focus on subtype-wise kidney tumor classification, quantitative evaluation of Grad-CAM localization using expert-defined ROIs and IoU/Dice metrics, and multi-center validation to further assess the robustness and generalizability of the proposed approach.

Acknowledgment

All authors have mutually agreed to participate.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

D. Vetrithangam: Conceptualization, Methodology, Software, Write Original manuscript.

Pramod Kumar: Data curation, formal data analysis, Write Original manuscript.

Shilpasree S: Methodology, Formal data analysis, Validation, Visualization, Review and Revise Original manuscript.

Akanksha Kulkarni: Investigation, Visualization.

Gara Jaya Raju: Supervision, Resources, Validation.

Aswathy S Nair: Software, Investigation, Visualization.

Subramanian Selvakumar: Literature review, Supervision, Review and Revise Original manuscript.

Puneet Kumar: Revise Original manuscript.

Ethics

This study does not involve human participants, patients, clinical trials, or animal subjects. Therefore, ethical approval and informed consent were not required.

References

Abdelrahman, A., & Viriri, S. (2023). FPN-SE-ResNet Model for Accurate Diagnosis of Kidney Tumors Using CT Images. *Applied Sciences*, *13*(17), 9802. https://doi.org/10.3390/app13179802

Alonge, M., Isreal, O., & Olorunniwo, O. E. (2025). Integrating multi-scale attention mechanisms into CNN architectures for enhanced kidney lesion classification in CT imaging. *ResearchGate*.

Alzu'bi, D., Abdullah, M., Hmeidi, I., AlAzab, R., Gharaibeh, M., El-Heis, M., Almotairi, K. H., Forestiero, A., Hussein, A. M., & Abualigah, L. (2022). Kidney Tumor Detection and Classification Based on Deep Learning Approaches: A New Dataset in CT Scans. *Journal of Healthcare Engineering*, *2022*, 1–22. https://doi.org/10.1155/2022/3861161

Chakraborty, A., & Vetrithangam, D. (2023). ExRAN: Deep Ensemble Majority Voting using Transfer Learning for Brain tumor Identification from Magnetic Resonance Imaging. *2023 International Conference on Inventive Computation Technologies (ICICT)*, 130–134. https://doi.org/10.1109/icict57646.2023.10134332

Eulig, E., Jäger, F., Maier, J., Ommer, B., & Kachelrieß, M. (2025). Reconstructing and analyzing the invariances of low-dose CT image denoising networks. *Medical Physics*, *52*(1), 188–200. https://doi.org/10.1002/mp.17413

Gharaibeh, M., Alzu'bi, D., Abdullah, M., Hmeidi, I., Al Nasar, M. R., Abualigah, L., & Gandomi, A. H. (2022). Radiology Imaging Scans for Early Diagnosis of Kidney Tumors: A Review of Data Analytics-Based Machine Learning and Deep Learning Approaches. *Big Data and Cognitive Computing*, *6*(1), 29. https://doi.org/10.3390/bdcc6010029

Gofrit, O. N., & Orevi, M. (2016). Diagnostic Challenges of Kidney Cancer: A Systematic Review of the Role of Positron Emission Tomography-Computerized Tomography. *Journal of Urology*, *196*(3), 648–657. https://doi.org/10.1016/j.juro.2016.02.2992

Hannan, S. A., & Pal, P. (2023). Detection and classification of kidney disease using convolutional neural networks. *Journal of Neurology and Neurorehabilitation Research*, *8*(2), 1–7.

Kaur, G., Sharma, N., Malhotra, S., Devlival, S., & Gupta, R. (2024). Kidney Tumor Detection and Classification Using Convolutional Neural Network Architecture. *2024 2nd World Conference on Communication & Computing (WCONF)*, 1–5. https://doi.org/10.1109/wconf61366.2024.10691954

Kumar, A., Nelson, L., & Venu, V. S. (2024). Enhancing Kidney Disease Classification through Transfer Learning with VGG16. *2024 2nd International Conference on Computer, Communication and Control (IC4)*, 1–6. https://doi.org/10.1109/ic457434.2024.10486470

Li, Y., Gao, X., Tang, X., Lin, S., & Pang, H. (2023). Research on automatic classification technology of kidney tumor and normal kidney tissue based on computed tomography radiomics. *Frontiers in Oncology*, *13*, 1013085. https://doi.org/10.3389/fonc.2023.1013085

Mahmud, S., Abbas, T. O., Mushtak, A., Prithula, J., & Chowdhury, M. E. H. (2023). Kidney Cancer Diagnosis and Surgery Selection by Machine Learning from CT Scans Combined with Clinical Metadata. *Cancers*, *15*(12), 3189. https://doi.org/10.3390/cancers15123189

Majid, M., Gulzar, Y., Ayoub, S., Khan, F., Reegu, F. A., Mir, M. S., Jaziri, W., & Soomro, A. B. (2023). Enhanced Transfer Learning Strategies for Effective Kidney Tumor Classification with CT Imaging. *International Journal of Advanced Computer Science and Applications*, *14*(8), 0140847. https://doi.org/10.14569/ijacsa.2023.0140847

Martín, Ó. A., & Sánchez, J. (2025). Evaluation of vision transformers for multimodal image classification: A case study on brain, lung, and kidney tumors. *ArXiv Preprint (ArXiv:2502.05517*. https://doi.org/10.48550/arXiv.2502.05517

Mehedi, M. H. K., Haque, E., Radin, S. Y., Ur Rahman, Md. A., Reza, Md. T., & Alam, Md. G. R. (2022). Kidney Tumor Segmentation and Classification using Deep Neural Network on CT Images. *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–7. https://doi.org/10.1109/dicta56598.2022.10034638

Sudharson, S., & Kokil, P. (2020). An ensemble of deep neural networks for kidney ultrasound image classification. *Computer Methods and Programs in Biomedicine*, *197*, 105709. https://doi.org/10.1016/j.cmpb.2020.105709

Vetrithangam, D., Ebin, PM., Nareshkumar, P., Arunadevi, B., Fathima, A., & Ramesh Kumar, A. (2023). Enhanced Vgg-19 Model for Accurate Brain Tumor Prediction. *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, 1–6. https://doi.org/10.1109/nmitcon58196.2023.10276235

Yang, E., Kim, C. K., Guan, Y., Koo, B.-B., & Kim, J.-H. (2022). 3D multi-scale residual fully convolutional neural network for segmentation of extremely large-sized kidney tumor. *Computer Methods and Programs in Biomedicine*, *215*, 106616. https://doi.org/10.1016/j.cmpb.2022.106616

Zhang, W., On, C. K., Alfred, R., Chai, S. S., & Chiou Sheng, C. (2025). Deep Learning Models for Low-Dose CT Denoising: A Forward Survey (2022–2025). *2025 International Conference on Advances in Machine Intelligence, and Cybersecurity Technologies (AMICT)*, 284–289. https://doi.org/10.1109/amict65811.2025.11402821

Zi, H., Liu, M.-Y., Luo, L.-S., Huang, Q., Luo, P.-C., Luan, H.-H., Huang, J., Wang, D.-Q., Wang, Y.-B., Zhang, Y.-Y., Yu, R.-P., Li, Y.-T., Zheng, H., Liu, T.-Z., Fan, Y., & Zeng, X.-T. (2024). Global burden of benign prostatic hyperplasia, urinary tract infections, urolithiasis, bladder cancer, kidney cancer, and prostate cancer from 1990 to 2021. *Military Medical Research*, *11*(1), 64. https://doi.org/10.1186/s40779-024-00569-w