Proceedings of the 5th International Conference on Advanced Science & Engineering ICOASE 2025, University of Zakho and Duhok Polytechnic University, 24th – 25th September 2025

(Special issue for Passer Journal of Basic and Applied Sciences)
Paper No. 12

A Deep Learning Approach for Automated Kidney Tumor Classification

Hivi Kamal Ismael [1,2], Wafaa Mustafa Abduallah [3]

[1] Department of Information Technology, Akre University for Applied Science, Akre 42004, Kurdistan Region, Iraq.
[2] Department of Information Technology, Duhok Polytechnic University, Duhok 42001, Kurdistan Region, Iraq.
[3] Department of Cybersecurity Engineering, Duhok Polytechnic University, Duhok 42001, Kurdistan Region, Iraq.
Email: hevi.kamal@auas.edu.krd , Wafaa.abduallah@dpu.edu.krd

Abstract

Correct classification of kidney tumors in computed tomography (CT) images is crucial for early diagnosis and treatment planning. In this study, we propose a hybrid architecture in which ResNet50 is employed for localized spatial feature extraction, while Vision Transformer (ViT) enables global contextual learning to automatically classify kidney tumors into multiple classes. A single-stage paradigm was then performed to classify CT images into one of the three clinically relevant categories: normal tissue, benign tumor and malignant tumor. All classifications and evaluations were performed at the slice level, where each axial CT slice was treated as an independent input sample. The model was independently trained and tested using two publicly available datasets of CT images, KAUH-Kidney, and CT-Kidney. It so adapts advanced preprocessing methods like class-aware data augmentation, normalization, and focal loss to deal with class imbalance. To thoroughly evaluate performance, we used a 5-fold cross-validation. On the KAUH-Kidney dataset, the hybrid model reached an accuracy of 99.53% and on the CT-Kidney dataset it detected 99.73% of the samples with the perfect macro and micro AUC scores on both datasets. Where the combined approach outperformed both standalone CNN or transformer-based architectures in accuracy, F1-score, and generalization. This study demonstrates the potential of hybrid deep learning frameworks to assist in the more accurate, efficient and automated classification of kidney tumors in clinical practice settings.

Keywords: Kidney Tumor Classification, Multi-class Classification, Deep Learning, ResNet50, Vision Transformer (ViT), CT Imaging.

1 Introduction

Image classification plays a crucial role in computer vision and medical imaging, enabling systems to analyze and categorize visual data effectively [1]. In healthcare, these systems accelerate diagnosis, reduce manual workload, and support clinical decisions. Deep learning (DL), particularly Convolutional Neural Networks (CNNs), has proven highly effective in extracting patterns from medical images without manual feature engineering [2].

Kidney cancer is a serious global health issue, ranking among the top ten most common cancers in both men and women. Worldwide, more than 430,000 new cases and 180,000 deaths were reported in 2020 alone [3]. In the U.S., over 76,000 new cases and nearly 14,000 deaths were reported in 2021[4]. Early classification of kidney tumors is essential for selecting appropriate treatment strategies, as tumor subtypes differ significantly in behavior [5]. However, these tumors often lack early symptoms and are frequently discovered incidentally during imaging [6]. Misclassification or delayed diagnosis can lead to disease progression, metastasis, or unnecessary surgical interventions, impacting both survival rates and quality of life. With the advent of imaging techniques as CT and MRI, detection percentage has increased [7], and DL models have shown potential for automatization of the classification process [8].

In comparison with conventional machine learning, DL techniques achieve higher accuracy, and in some cases experience better performance than professional radiologists [9]. CNN-based techniques have been widely applied for benign and malignant tumor discrimination [10], and recent advances in this technology have generalized to multi-class tumor classification [11]. Attention mechanisms and transformer modules have been applied to CNN based architectures including VGG16, ResNet50 and EfficientNet [12,13].

Hybrid models using the combination of CNNs and Vision Transformers (ViTs) have shown better classification accuracy because they are able to extract both spatial and contextual features. These methods also contribute to

improved comprehension of tumor heterogeneity and predictive behaviours [14]. The application of the publicly available datasets, such as TCGA, KiTS19 and C4K-KiTS19, has also made it possible for these models to be evaluated and tested in the standardized way [15].

In this study, we propose a hybrid model that combines ResNet50 capturing spatial features and ViT learning global context to enhance kidney tumor classification. Furthermore, the complexity of kidney tumors, as a result of their subtle early appearance, diverse size and texture, and overlapping image characteristics, make it difficult for conventional classification methods. To address these issues, our model uses both local and global image features: ResNet50 learns the fine-grained spatial details to distinguish between tumor subtypes, while vision transformer (ViT) captures that contextual information spread across the entire image. This integrated approach was primarily used to enhance classification accuracy in a setting characterized by tumor heterogeneity and imaging uncertainty.

Unlike existing studies that rely only on conventional CNNs, transformers, or binary classification pipelines, this work introduces a unified hybrid ResNet50–Vision Transformer (ViT) framework for end-to-end multi-class kidney tumor classification from CT images. The proposed model explicitly combines the complementary strengths of ResNet50 for local spatial and structural feature extraction with the ViT's ability to model long-range dependencies and global contextual information. In addition, this study addresses the severe class imbalance commonly present in kidney tumor datasets through the joint use of class-specific data augmentation and focal loss, improving robustness across Normal, Benign, and Malignant categories. Furthermore, the framework is systematically evaluated on two independent public kidney CT datasets using strict patient-wise 5-fold cross-validation, and is enhanced with Grad-CAM-based visual explanations to improve model interpretability and support its potential role as a clinical decision-support system.The rest of this paper is organized as follows: Section 2 reviews the related work; Section 3 presents the methodology; Section 4 discusses the experimental results; and Section 5 concludes the study.

2 Related Studies

Deep learning methods have recently made significant improvements in classifying and detecting kidney tumors such as CT scan images. For example, [6] generated a new dataset with 8,400 CT images of 120 patients. They performed multiple transformations on the images including normalization, format conversion, and augmentation, to enhance the data quality. They also applied edge detection techniques from OpenCV to extract important features. Using custom CNN models, specifically CNN-6 and CNN-4, they achieved high accuracies of 97% and 92%, respectively, outperforming standard pre-trained models like VGG16. However, their study had limitations because it used data from only one institution, making it less general. It also did not include 3D imaging or external validation methods.

Following this, [16] introduced a method called SSLSD-KTD, combining masked autoencoders and self-distillation with a Vision Transformer model. Using two datasets, KAUH-kidney and CT-kidney, they improved image features through advanced preprocessing. Their approach resulted in high accuracies of 98.04% on the KAUH-kidney dataset and 82.14% on the CT-kidney dataset. When incorporating transfer learning techniques, accuracy increased further, setting a new benchmark in automatic tumor detection from CT scans. However, this method relies on labeled data, and it is unclear if it can handle larger datasets, which remain challenges.

In another relevant study, [17] tested several CNN models, including VGG16, ResNet50, U-Net, and Crossbar-Net, for detecting kidney tumors. They demonstrated that incorporating attention mechanisms and squeeze-and-excitation blocks improved model performance, achieving up to 96.56% accuracy with ResNet50 and approximately 95% with their own CNN model. However, computational costs and large dataset requirements limited the feasibility of broader implementation, underscoring a trade-off between accuracy and computational demands.

[18] highlighted the efficacy of transfer learning by developing a computer-assisted diagnosis system that combined traditional methods of machine learning with models of deep learning such as DenseNet-121 and ResNet-101. Through meticulous image preprocessing and feature extraction, they achieved very high accuracy (98.22%), precision (95.77%), recall (93.23%), and an F1-score (98.43%) with DenseNet-121. Their hybrid approach substantially enhanced the efficiency and accuracy of kidney tumor detection. Although highly effective, this approach requires validation on broader and more diverse datasets to confirm its generalizability.

Finally, [2] implemented a hybrid deep learning model combining two most famous architectures VGG16 and VGG19, using transfer learning. They utilized careful image scaling and augmentation techniques along with dropout layers to prevent overfitting. The model classified the kidney pathology to normal, cystic, stones and tumor class with the outstanding accuracy of (100%) against minimal loss and is easy to train with minimum time between epochs.

This study represented a strong starting point for subsequent work focused on increasing the speed and scalability of kidney tumor diagnosis. While their work has produced excellent results there are limitations to their research. It's a short training time, there is no external validation, and restrictions to the dataset. This highlights the necessity to conduct more studies to validate the results to be widely applicable to clinical practice.

Overall, these studies demonstrate the many effective deep learning strategies, including dataset generation and effective preprocessing methods, as well as learned model and transfer learning based strategies that improved accuracy, robustness of kidney tumor classification and detection from CT scans.

3 Materials and Methods

The study follows a structured methodology to classify kidney tumors using a hybrid deep learning model that integrates ResNet50 and Vision Transformer (ViT). The proposed pipeline consists of several key stages: data preprocessing, model training, evaluation, and deployment. Figure 1 illustrates the complete workflow, including dataset preparation, preprocessing techniques such as image resizing, normalization, and data augmentation, as well as model training and optimization. The hybrid model utilizes ResNet50 for spatial feature extraction and ViT for global contextual learning. The final trained model is validated and tested to ensure robust classification performance.

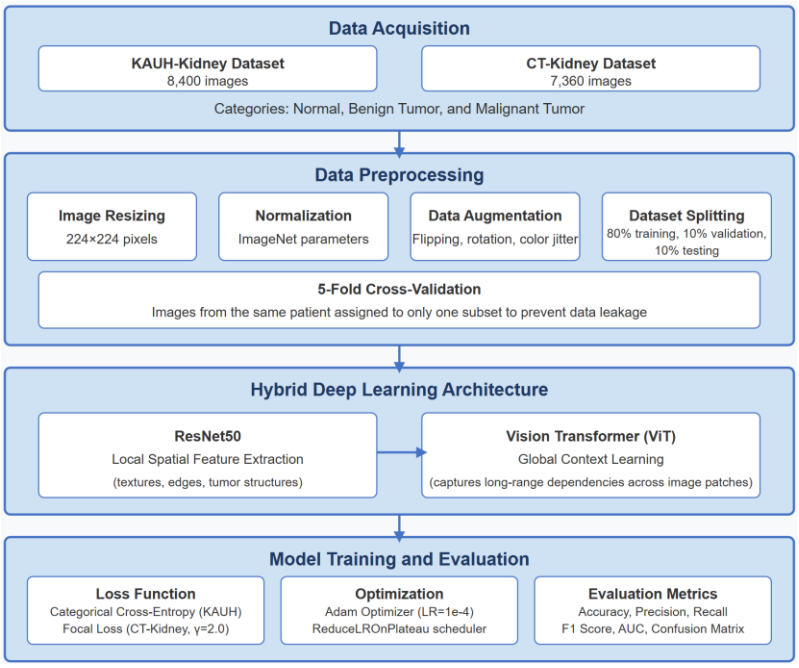


Figure 1: Overview of the proposed deep learning methodology for kidney tumor classification.

The workflow includes data acquisition, preprocessing, hybrid deep learning architecture (resnet50 and vision transformer), model training, and evaluation, emphasizing the use of 5-fold cross-validation to prevent data leakage.

3. 1 Data Acquisition

This study utilized two publicly available kidney CT scan datasets, the KAUH-Kidney and CT-Kidney datasets, both of which were independently used to train and evaluate a model of deep learning for classifying kidney tumors into three categories: normal, benign tumor, and malignant tumor.

3. 1. 1 KAUH-Kidney Dataset

A dataset utilized in this study is the KAUH-Kidney dataset [6]. It comprises 8,400 images acquired from 120 adult patients at the CT scanning department of Jordan KAUH Hospital. The images are supplied in DICOM format as axial slices, commonly used for the exchange and transmission of medical images. The dataset comprises CT scans, both with and without contrast, obtained from adult patients aged 30 to 80 years, namely 55 females and 65 males, who had belly and pelvis CT imaging. Of the 120 patients, 60 diagnosed with benign or malignant tumors, while the other 60 were classified as normal cases. **In this study, a subset of 7,770 images from 111 patients was selected and used**

for training and evaluation. While obtaining experimental results, the dataset was split into 80% for training, 10% for validation, and 10% for testing, ensuring that images of the same patient were present in only one subset. The detailed dataset distribution is presented in Table 1.

Table 1: Distribution of the KAUH-Kidney Dataset

Category	Training (80%)	Validation (10%)	Testing (10%)	Total
Normal	2,859	342	369	3,570
Tumor	3,357	435	408	4,200
Total	6,216	777	777	7,770

3. 1. 2 CT-Kidney Dataset

The second dataset utilized in this study, the CT-Kidney dataset, originally comprises 12,446 CT images obtained from various hospitals in Dhaka, Bangladesh, using urogram-based scanning protocols [19]. The dataset includes images of the entire abdomen as well as various kidney conditions, including cysts, kidney stones, normal kidneys, and kidney tumors. The images were collected from PACS (Picture Archiving and Communication System) in DICOM format, with both contrast-enhanced and non-contrast axial and coronal scans selected for analysis. For this study, a subset of 7,360 images was selected, focusing exclusively on normal and tumor kidney cases, while images containing cysts and kidney stones were excluded. This subset consists of 5,077 Normal kidney images and 2,283 Tumor images. The dataset was split into 80% training, 10% validation, and 10% testing, ensuring that contrast-enhanced and non-contrast variations were not assigned across different subsets. The detailed distribution of this dataset is provided in Table 2.

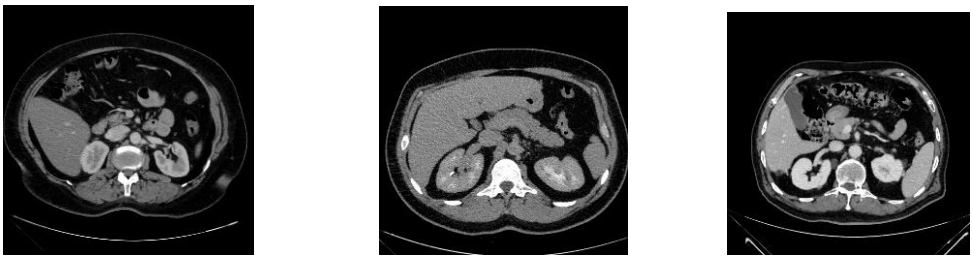
Table 2: Distribution of the CT-Kidney Dataset

Category	Training (80%)	Validation (10%)	Testing (10%)	Total
Normal	4,059	508	510	5,077
Tumor	1,829	228	226	2,283
Total	5,888	736	736	7,360

Both datasets were **used separately**, meaning that the **trained model was evaluated independently on each dataset** rather than on a combined dataset. In both datasets, CT scans were provided as axial slices in DICOM format. Each slice was treated as an independent sample during training, validation, and testing. Consequently, all labels, predictions, and performance metrics reported in this study are defined at the slice level rather than at the patient level.

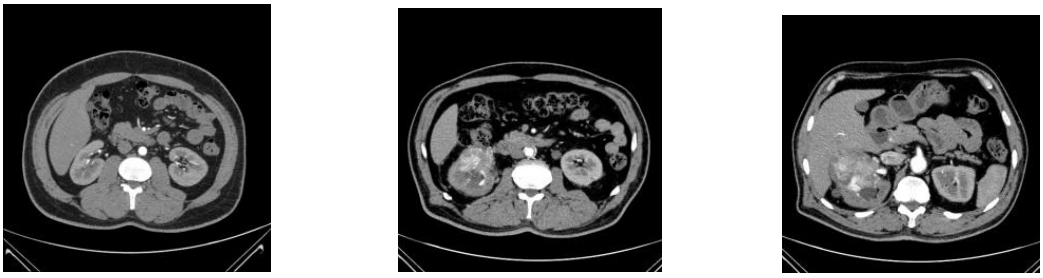
3. 1. 3 Representative Sample Images

To provide visual context for the classification task, sample CT images from both datasets are shown. Figure 2A includes images from the KAUH-Kidney dataset: (a) normal kidney, (b) benign tumor, and (c) malignant tumor. These images, captured in axial view, vary in contrast and tumor appearance. Figure 2B shows similar examples from the CT-Kidney dataset, highlighting differences in image quality and tumor presentation due to varied sources and scanning protocols. Including these examples helps illustrate the visual challenges the model addresses and the diversity in the data.



(a) Normal Kidney (b) Benign Tumor (c) Malignant Tumor

Figure 2A: Representative axial CT images from the KAUH-Kidney dataset.



(a) Normal Kidney (b) Benign Tumor (c) Malignant Tumor

Figure 2B: Representative axial CT images from the CT-Kidney dataset.

3. 2 Data Preprocessing

Preprocessing is a key step in medical image classification using deep learning techniques, ensuring data consistency, enhancing tumor feature visibility, and improving model performance. A single application of a standardized preprocessing pipeline was used on the KAUH-Kidney and CT-Kidney datasets to provide consistency in terms of input sizes, intensity normalization, and augmentation strategies. Standardization of image property along with maintaining variations which helps the model to generalize efficiently under different imagery conditions is introduced in the form of image preprocessing like resizing, normalization, data augmentation, and class-balancing techniques. During data loading, for minority classes, class specific augmentation techniques were applied, and a weighted sampling was performed. These pre-processing techniques improve classification performance and robustness, especially for cases with lower sample availability.

3. 2. 1 Image Resizing

Due to variations in scanning equipment and resolution, all images were resized to 224×224 pixels to ensure consistency in input dimensions.

- For training images, random resized cropping was applied to introduce minor spatial variations and enhance generalization.
- For validation and testing images, bilinear interpolation was used to maintain smooth scaling and preserve anatomical details.

Standardizing image dimensions ensures compatibility with the deep learning model while reducing the complexity of computation and memory overhead.

3. 2. 2 Normalization

To standardize pixel intensity distributions and enhance model performance, ImageNet-based normalization was applied across all dataset splits. The normalization equation is:

$$X' = \frac{X - \mu}{\sigma} \qquad (1)$$

where **X'** is the normalized pixel intensity, **X** is the original intensity, **μ** is the mean, and **σ** is the standard deviation. The ImageNet normalization parameters were applied as follows. As the CT scan images were originally single-channel grayscale, they were replicated into three channels using NumPy before normalization to match the input dimensionality expected by pretrained models (e.g., ResNet50 and ViT). This allowed the use of standard ImageNet normalization parameters for compatibility.

- **Mean:** [0.485, 0.456, 0.406]
- **Standard Deviation:** [0.229, 0.224, 0.225]

This ensures that the model focuses on tumor-related features rather than brightness variations due to different scanning protocols.

3. 2. 3 Data Augmentation

Several data augmentation techniques have been used during training to increase generalization and reduce overfitting. These included random resized cropping (60%–100%) to simulate variations in scanning conditions while retaining anatomical structures, horizontal flipping to enhance orientation invariance (as tumors can arise on either kidney) and rotation (±15° for majority class images and ±20° for minority classes) due to patient positioning and scanner direction. We used the color jittering (brightness and contrast adjustments) to replicate variations in imaging conditions, and then transformed all the images into PyTorch tensor (ToTensor) for efficient feeding to the model. For all training images, random resized cropping (scale range: 0.8–1.0), horizontal flipping, rotation (±15°), and color jittering (brightness and contrast: ±0.2) were applied. For minority classes in the CT-Kidney dataset, stronger augmentations were used with larger rotation angles (±20°), and cropping scales between 0.6–1.0.

A minority-aware augmentation strategy was used to handle the class imbalance especially for the under-represented tumor classes i.e., malignant tumors. This already comprised much more powerful transformations such as larger rotation angles, more extreme color jittering and more aggressive cropping. These techniques were intended to increase the diversity of training samples and strengthen the capability of the model to learn discriminative features from minority categories.

Data augmentation has been well-demonstrated in previous work as an effective way to improving generalization and reducing overfitting in deep learning models, especially in medical imaging tasks [20,21].

3. 2. 4 Dataset Splitting and 5-Fold Cross-Validation

For a strong performance evaluation and a better generalization of the model, we used a 5-fold cross-validation strategy separately for KAUH-Kidney and CT-Kidney datasets. Each dataset was further partitioned into five subsets, in these subsets, images from the same patient were only in one subsets for preventing data leakage. For each fold, four subsets (80%) were utilized for model training, and one subset (20%) was used for validation. The described process was repeated five times with the restriction that the subset was selected for validation only once, which offers a more comprehensive and reliable performance evaluation. After completing the cross-validation phase, the final model was evaluated on an independent test set (10%), which remained untouched during training. It is important to note that although patient-wise splitting was strictly enforced to avoid data leakage, model predictions and performance evaluation were conducted at the slice level, without aggregating predictions across multiple slices belonging to the same patient. Additionally, for the CT-Kidney dataset, preprocessing included class-balancing strategies during training, such as weighted random sampling combined with minority-specific augmentation techniques. These methods ensured equitable representation of all tumor classes across cross-validation folds and reduced the risk of model bias toward majority classes. No class balancing strategies were applied to the KAUH-Kidney dataset, as its tumor classes were relatively balanced in distribution.

3. 3 The proposed deep learning architecture

To classify kidney tumors, a hybrid deep learning model was developed by integrating ResNet50 and ViT (ViT-Base-ResNet50-224-in21k from the TIMM library). The proposed architecture combines the strengths of convolutional neural networks (CNNs) for local spatial feature extraction and Vision Transformers (ViT) for capturing contextual global relationships. for global contextual relationships. The overall architecture and workflow of the model are illustrated in Figure 3, which provides a visual representation of the complete pipeline, starting from image preprocessing to final classification. The ResNet50 module is utilized to extract spatial features through convolutional operations, which are then passed to the Vision Transformer (ViT) for global feature learning using self-attention mechanisms. The classification head outputs predictions into three categories: Normal Kidney, Benign Tumor, and Malignant Tumor.

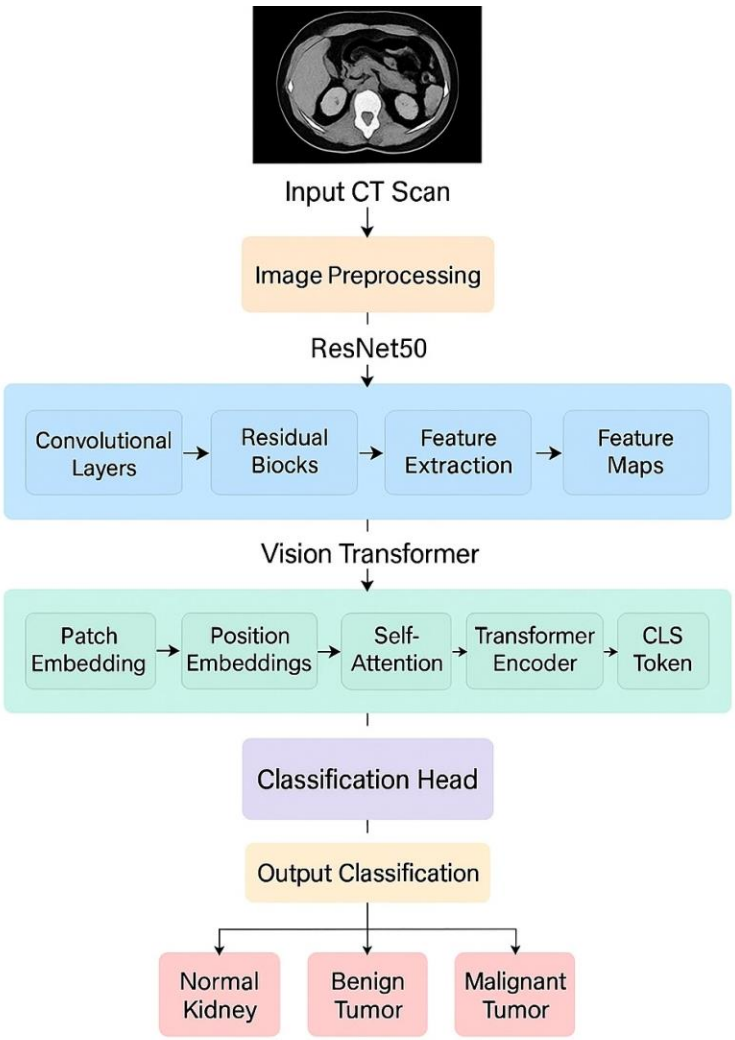


Figure 3: Overview of the proposed deep learning model architecture for kidney tumor classification.

The model combines ResNet50 for local spatial feature extraction and Vision Transformer (ViT) for global contextual learning. The architecture predicts the classification result to its three categories: Normal Kidney, Benign Tumor, Malignant Tumor.

3. 3. 1 ResNet50 for Feature Extraction

The ResNet50 architecture is a 50-layer deep CNN architecture that aims to extract the spatial features from the medical images while solving the vanishing gradient problem through residual learning techniques. In this study, the pre-trained ResNet50 was applied as feature extractor. It performs its intelligent hierarchical architecture based on convolution, residual connections, and max-pooling layers to first extract features intelligently and then the resulting features are fed into the ViT classifier.

3. 3. 2 Vision Transformer (ViT) for Global Context Learning

Vision Transformer (ViT) is an architecture of deep learning that applies transformer-based self-attention mechanisms to both image classification and analysis, and serves as a strong alternative to traditional CNNs. Unlike CNNs, which depend on localized feature extraction through convolutional filters, ViTs considers images as sequences of patches, handling them similarly to how words are managed in natural language processing (NLP) for image processing. The patch-based method allows ViT to learn global contextual relationships, which is extremely useful for complex tasks of medical imaging, like tumor classification.

3. 3. 3 Hybrid Architecture (ResNet50 + ViT)

The proposed hybrid model integrates Convolutional Neural Networks (CNNs) and transformers to enhance the classification of kidney tumors in CT images. Specifically, we adopt a pre-trained ViT-Base-ResNet50-224-in21k model from the TIMM library, where the ResNet50 backbone acts as the feature extraction module within the transformer architecture, rather than functioning as an independent CNN. ResNet50 is known for its ability to capture low-level spatial features such as edges, textures, and fine anatomical details through convolutional operations and residual connections, which facilitate deep network training and stability. ResNet50 is widely famous for its capability to learn low-level spatial features such as edges, textures and fine anatomical details through the application of Convolutional operation followed by residual connections, which in turn leads to deep network training and stability.

The Vision Transformer (ViT), on the other hand, employs a self-attention which learns global relationships between spatial regions of the image. This is particularly relevant in the context of medical imaging, where the large scale cues surrounding a lesion or organ are essential for its classification. Incorporating the local feature extraction of ResNet50 and ability to model long-range relationships of ViT, the hybrid enjoys a more discriminative and richer feature representation. This integration enhances the robustness of the classification, which is especially important when dealing with complex multi-class problems such as normal, benign, and malignant classes of tissue to separate between at different spatial resolutions.

The hybrid architecture integrates a ResNet50 backbone with a Vision Transformer (ViT) head. Specifically, the ResNet50 extracts 7×7×2048 feature maps, which are reshaped into 49 patches of 2048 dimensions each and treated as token embeddings for the ViT module. These tokens are then fed into a Transformer encoder with 12 layers, 12 attention heads, and learnable positional embeddings. This design preserves both spatial and semantic context for final classification.

3. 4 Model Training and Optimization

The proposed hybrid ResNet50 and Vision Transformer (ViT) model was implemented using PyTorch and trained on an NVIDIA CUDA-enabled GPU to accelerate computations. The training process involved separate training and evaluation phases for each dataset (KAUH-Kidney and CT-Kidney), comprising training, validation, and independent testing stages, to rigorously assess model performance.

3. 4. 1 Loss Function

The model was trained with categorical cross-entropy loss, which is a commonly used loss function for multi-class prediction of the difference between predicted class probabilities and the true class label. It is formulated as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \qquad (2)$$

where y_i represents the ground truth label, and $\hat{y}_i$ denotes the predicted probability for class i.

Specifically, the CT-Kidney dataset, contained a target class ratio imbalance, thus we used the Focal Loss function (with $\gamma = 2.0$) and class-specific weights during model training to lessen the effect of class imbalance and to concentrate learning more on those samples that were hard to classify. In contrast, the KAUH-Kidney dataset did not have heavily imbalanced classes, so we were able to use standard CrossEntropyLoss without additional weighting mechanisms.

3. 4. 2 Optimization and Training Strategy

The model was trained using Adam optimizer with initial learning rate of 1×10^{-4}, this allows the model to converge in a stable adaptable way. In order to improve generalization, we used a ReduceLROnPlateau scheduler which decreased the learning rate when validation accuracy remained constant for three consecutive epochs. Model checkpointing was employed to preserve the best-performing model during the training process. The general training strategy employed batch size of 16 and training for maximum of 20 epochs, effectively balancing efficiency and performance. Moreover, data augmentation and ImageNet-based normalization was used to ensure robustness and

consistency in performance with respect to different imaging conditions. The ReduceLROnPlateau scheduler was used with a patience of 3 epochs and a reduction factor of 0.1. The model was trained using Adam optimizer with a learning rate of 1e-4 and a batch size of 16. Focal loss was implemented using $\gamma = 2.0$ and class-wise α values based on the inverse frequency of each class.

4 Results and Discussion

In this section, the experiments performed for kidney tumor classification using the proposed hybrid deep learning model (ResNet50 + Vision Transformer (ViT)). The model was created in Python with PyTorch and related libraries and trained and tested using 5-fold cross validation on the KAUH-Kidney and CT-Kidney datasets. For optimal performance, the parameters were optimized based on the preliminary experiments conducted in a GPU-enabled Google Colab environment. Subsequent sections discuss the detailed results and including precision, F1-score, accuracy, recall, AUC, and others.

4. 1 Experiments Setup

The proposed model was implemented on **Google Colab** using **Python 3** and essential deep learning libraries, including **PyTorch, timm, torchvision, h5py, and scikit-learn**. Additional Python packages such **as numpy, matplotlib, and tqdm** were utilized for **data processing, visualization, and progress monitoring**. The experiments were performed on a **cloud workstation equipped with an NVIDIA Tesla T4 GPU** (~16GB RAM, ~12GB GPU memory, and ~50GB of disk space). The parameters were chosen based on experiments that produced the most optimal results for those listed in Table 3. The detailed descriptions are as follows:

(i) Steps per Epoch:

Although not explicitly specified in the code, the number of training steps per epoch is implicitly determined by the size of the training dataset and the batch size (16), with a WeightedRandomSampler ensuring balanced class distribution.

(ii) Epochs:
Each fold of the 5-fold cross-validation was trained for 20 epochs, ensuring multiple complete passes through the training data.

(iii) Validation Steps:

Similarly, the number of validation steps per epoch is determined by dividing the total number of validation samples by the batch size.

(iv) Loss Function:

The model supports two loss functions:

- **Focal Loss** (gamma=2.0) for handling class imbalance.
- **Weighted Cross-Entropy Loss**, with class weights computed dynamically from dataset distributions.

(v) Optimizer:
The Adam optimizer was chosen for its adaptive learning rate and effective handling of sparse gradients, crucial for our image classification task.

(vi) Activation Function:

A softmax activation function was used in the final classification layer to generate distributions of probabilities over the three classes.

(vii) Learning Rate:

A starting learning rate of 1×10^{-4} was set, with a ReduceLROnPlateau scheduler applied to decrease the learning rate by a factor of 0.1 if the validation loss showed no improvement over 3 consecutive epochs.

(viii) Batch Size:

A batch size of 16 was selected to balance training efficiency with the available GPU memory.

(ix) Input Dimensions:

All input images were resized to 224×224 pixels with three channels (RGB), conforming to the requirements of the pre-trained networks.

(x) Feature Extraction:

The ResNet50 module, initialized with ImageNet-21K pretrained weights, serves as the spatial feature extractor by capturing textures, edges, and tumor structures.

(xi) Global Context Learning:

The Vision Transformer (ViT) module processes the feature embeddings to capture long-range dependencies and global contextual information, enhancing the classification accuracy.

(xii) Output Neurons:

The final classification layer consists of 3 neurons, corresponding to the classes: Normal, Tumor-Benign, and Tumor-Malignant.

Table 3: Model parameters.

Parameter	Value
Step_Per_Epoch	Training images/batch size
Epochs	20 (for each fold in 5-fold cross-validation)
Validation Steps	Validation images/batch size
Loss	Focal Loss (γ=2.0) / Weighted Cross-Entropy
Optimizer	Adam
Activation Function	Softmax
Learning Rate	0.0001 (Reduced by factor of 0.1 if no improvement)
Scheduler	ReduceLROnPlateau (patience=3, factor=0.1)
Batch Size	16
Input Dimensions	224 × 224 shape with 3 RGB channels
Feature Extraction	ResNet50 (Pretrained on ImageNet-21K)
Global Context Learning	Vision Transformer (ViT): 12 layers, 12 heads, 2048-dim tokens
Output Neuron	3 (Normal, Tumor-Benign, Tumor-Malignant)

4. 2 Model Evaluation

To thoroughly evaluate classification performance, precision, F1-score, accuracy, recall and area under curve (AUC) test set was used to evaluate the trained model. Accuracy (AC) is a metric for evaluating the performance of a classification model [22]. It represents the overall proportion of correctly classified samples, The formula for accuracy calculation is provided below [23]:

$$Accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)} \qquad (3)$$

Where **TP** (True Positives) and **TN** (True Negatives) represent correctly predicted cases, while **FP** (False Positives) and **FN** (False Negatives) denote misclassified instances. Softmax activation is used in the final classification layer to produce class probability scores. The predicted class label is the one with the highest probability score, which leads to reliable classification performance.

To evaluate classification reliability, precision was used to measure the ratio of correctly predicted positive cases to the total predicted positives [24]:

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

Additionally, recall measured the model's capability to accurately identify all relevant instances within a class, calculated as [24]:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

Since precision and recall often have an inverse relationship, the F1-score was computed to provide a balanced metric, ensuring that both aspects were equally considered [24]:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

AUC evaluated the discriminative ability of the model, calculated for macro-averaged and micro-averaged scenarios to address class imbalance. The area under the curve represented the models' behaviors in different situations. The AUC can be calculated as follows [25]:

$$AUC = \frac{\sum R_i(I_p) - I_p((I_p + 1)/2)}{\mathrm{I_p} + I_n} \quad (7)$$

where I_p and I_n represent the number of positive and negative instances, respectively, and R_i denotes the rank of positive instances.

These metrics were computed using sklearn.metrics during final testing, ensuring an objective and standardized performance assessment. All evaluation metrics, including accuracy, precision, recall, F1-score, and macro/micro AUC, were computed based on slice-level predictions by comparing each predicted CT slice label with its corresponding ground-truth annotation. By using multiple evaluation metrics, the model's effectiveness in classifying **Normal, Tumor-Benign, and Tumor-Malignant** cases was rigorously validated. A confusion matrix was jointly generated for each test fold and then aggregated across all folds to examine misclassifications by category. This was a useful indication to see the strength and deficiencies of the model giving it the power to classify between Normal, Benign Tumor, and Malignant Tumor cases.

4. 3 Experimental Results and Analysis

This section reports the quantitative performance of the proposed hybrid ResNet50 + Vision Transformer (ViT) model on the KAUH-Kidney and CT-Kidney datasets. Results are compared with individual ResNet50 and ViT baselines using the standard metrics of classification: precision, F1-score, accuracy, recall, macro/micro AUC. The hybrid model consistently outperformed standalone architectures across all metrics and datasets.

4. 3. 1 Results on KAUH-Kidney Dataset

The hybrid model achieved excellent performance on the KAUH-Kidney dataset, with an accuracy of 99.53%, macro F1-score of 99.47%, and macro/micro AUC of 100.00%. These results were higher than those of the individual ResNet50 (99.23% accuracy) and ViT (96.27% accuracy) models. Figure 4 presents the final confusion matrix, which is summed across all five cross-validation folds, showing minimal misclassifications across the three classes: Normal, Tumor-Benign, and Tumor-Malignant. To clarify, the confusion matrices in Figures 4 and 8 represent the cumulative results across the test sets from all five cross-validation folds. These totals align precisely with the distribution described in Tables 1 and 2. To prevent any data leakage, patient-wise splitting was strictly enforced, images from a single patient were included in only one subset (train, validation, or test). The training behavior of the model is illustrated in Figures 5 to 7. Figure 5 displays the average AUC curves (macro and micro) across all five folds, Figure 6 shows the average training and validation loss curves, and Figure 7 presents the average training and validation accuracy curves. These figures confirm that the model converged efficiently during training and maintained stable generalization throughout all epochs and validation folds.

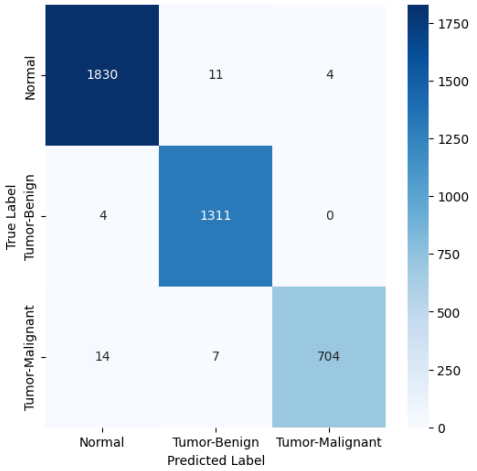


Figure 4: Confusion matrix of the hybrid model on the KAUH-Kidney dataset, summed across all five folds.

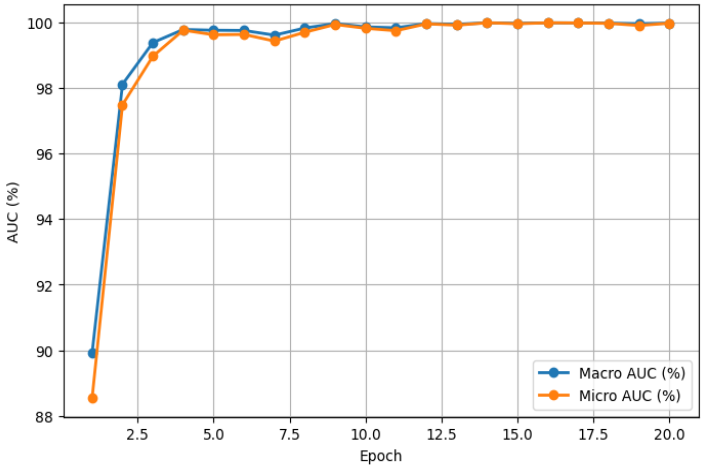


Figure 5: Average macro and micro AUC curves of the hybrid model over 5 folds (KAUH-Kidney dataset).

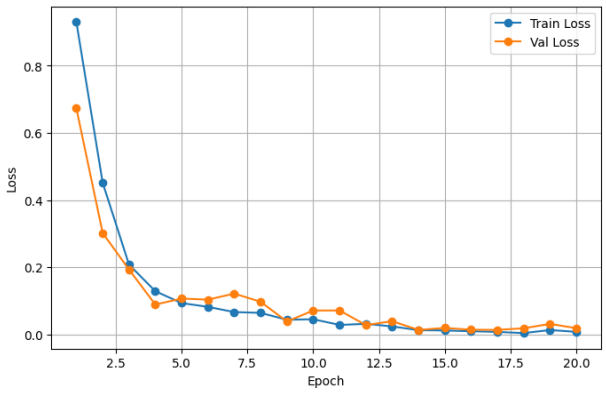


Figure 6: Average training and validation loss curves of the hybrid model over 5 folds (KAUH-Kidney dataset).

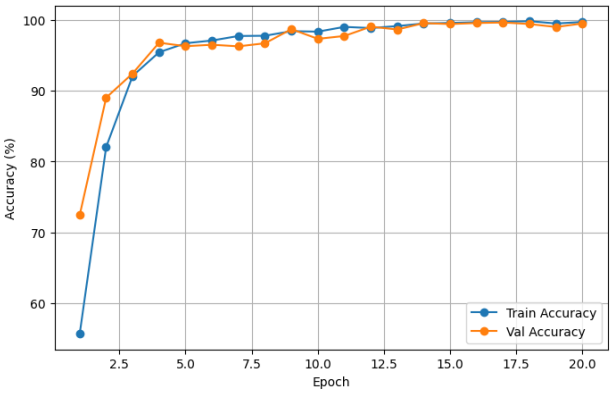


Figure 7: Average training and validation accuracy curves of the hybrid model over 5 folds (KAUH-Kidney dataset).

4. 3. 2 Results on CT-Kidney Dataset

The hybrid model also performed strongly on the CT-Kidney dataset, achieving 99.73% accuracy, a macro F1-score of 99.39%, and perfect macro/micro AUC scores of 100.00%. These results were higher than those of the individual ResNet50 (98.64% accuracy) and ViT (99.05% accuracy) models. Figure 8 presents the final confusion matrix, which is summed across all five cross-validation folds, showing minimal misclassifications across the three classes: Normal, Tumor-Benign, and Tumor-Malignant. The training behavior of the model is illustrated in Figures 9 to 11. Figure 9 displays the average AUC curves (macro and micro) across all five folds, Figure 10 shows the average training and validation loss curves, and Figure 11 presents the average training and validation accuracy curves. These figures indicate consistent learning and strong generalization throughout the training process.

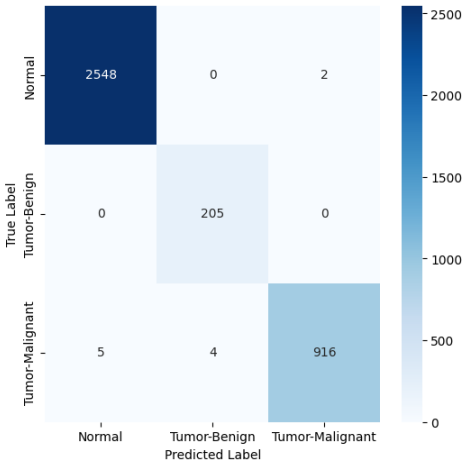


Figure 8: Confusion matrix of the hybrid model on the CT-Kidney dataset, summed across all five folds.

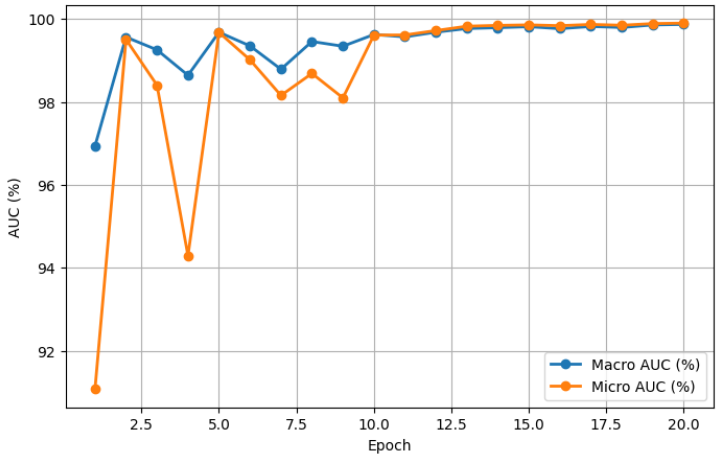


Figure 9: Average macro and micro AUC curves of the hybrid model over 5 folds (CT-Kidney dataset).

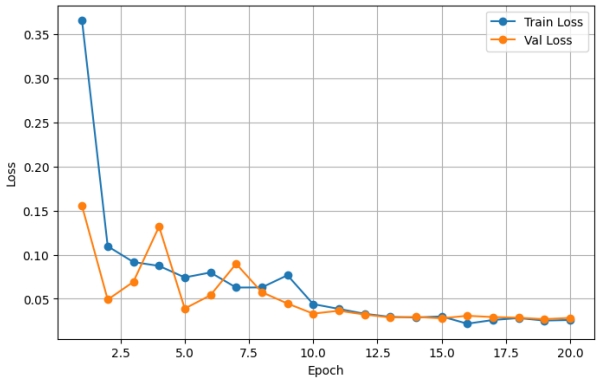


Figure 10: Average training and validation loss curves Curves of the hybrid model over 5 folds (CT-Kidney dataset).

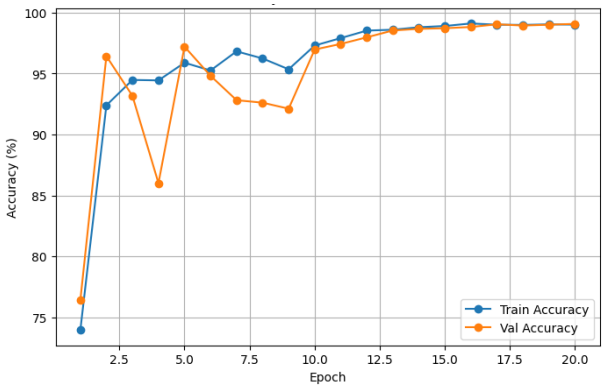


Figure 11: Average training and validation accuracy of the hybrid model over 5 folds (CT-Kidney dataset).

4. 3. 3 Experimental Results Discussion

The analysis of the experimental results demonstrates that the proposed hybrid deep learning model, which integrates ResNet50 and Vision Transformer (ViT), can accurately classify CT scans of kidney tumors into three clinically significant categories: Normal, Benign Tumor, and Malignant Tumor. We apply our model to two datasets (KAUH-Kidney and CT-Kidney) and achieve superior performance compared with other CNN and transformer-based models in the listed primary evaluation metrics, such as accuracy, F1-score, and AUC. The good performance is due to the fact that the model is able to extract representations from images in two different ways: the ResNet50 extracts low-level spatial and structural features (edges and textures) and the Vision Transformer (ViT) models long-range dependencies and global information over the image. The model integrates these two representations to analyze tumor patterns, ensuring reliable multi-class classification when complexity or ambiguity exists. One of the major advantages of the model is its good performance over both datasets, despite the variety of the data collection methods and source regions. Although cross-dataset validation is important for measuring real-world generalization, differences in acquisition protocols, image characteristics, and annotation strategies between public kidney CT datasets introduce significant domain shift. Therefore, in this study, **each dataset is evaluated independently using intra-dataset 5-fold cross-validation**, while cross-dataset evaluation combined with harmonization techniques is left as future work. The robustness of the model was also confirmed by 5-fold cross-validation. Furthermore, the class

imbalance in the CT-Kidney dataset was successfully solved by the class-specific data augmentation and focal loss technique. Compared with the binary or multi-stage methods in the past, the proposed method with end-to-end multi-class classification offers a more streamlined and computationally efficient multi-class classification framework that may support clinical decision-making. Previous models often could not discriminate between benign and malignant tumors or relied on limited datasets that did not generalize well. In contrast, our model performs well on multiple datasets using a one-step prediction framework. To make the model more interpretable, an important condition for clinical applicability, we integrated Gradient-weighted Class Activation Mapping (Grad-CAM) into the model to visualize which parts of the CT scan did contribute the most to the prediction of the model. As shown in Figure 12, Grad-CAM emphasizes tumor-related regions, which helps clinicians better understand why the model makes a particular decision and increases the confidence of the model for the clinical application.

Although the proposed hybrid ResNet50–Vision Transformer framework demonstrates excellent classification performance, this study does not claim direct clinical deployment or replacement of expert radiologists. All experiments were conducted at the slice level, without aggregating predictions across multiple slices for patient-level diagnosis, and no direct comparison with radiologist assessments was performed. Therefore, the reported results should be interpreted as a proof-of-concept demonstrating the potential of hybrid deep learning models for kidney tumor classification rather than a fully validated clinical diagnostic system. The integration of Grad-CAM-based visual explanations enhances model transparency by highlighting anatomically relevant regions associated with tumor predictions, thereby supporting interpretability and trust. In a clinical workflow, such a system could serve as a decision-support tool by assisting radiologists in prioritizing suspicious slices, reducing reading time, and providing complementary insights rather than replacing expert judgment. Future work will focus on patient-level aggregation strategies, comparison with radiologist performance, and prospective multi-institutional validation to establish clinical reliability and real-world applicability.

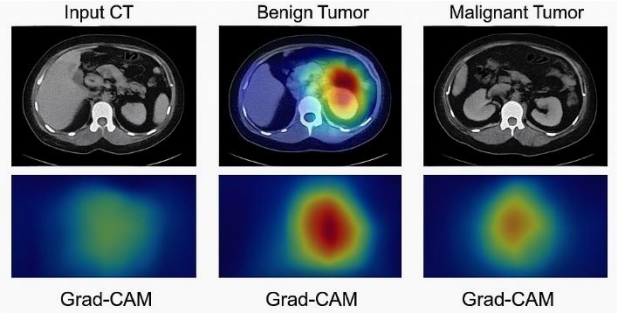


Figure 12: Grad-CAM visualization

Although kidney tumors are inherently three-dimensional structures, a 2D slice-based classification strategy was deliberately adopted in this study. The public CT datasets used exhibit substantial variability in slice thickness, number of slices per patient, and acquisition protocols, which complicates direct volumetric modeling without extensive harmonization. Under these conditions, the proposed 2D hybrid ResNet50–Vision Transformer framework provides a computationally efficient and robust solution, enabling consistent evaluation across heterogeneous datasets while effectively capturing both local and global image characteristics.

Despite the promising results, several limitations must be acknowledged. First, the model was trained and evaluated solely on CT images, which may limit its generalizability to other imaging modalities such as MRI or ultrasound. Different modalities provide varying levels of contrast and spatial resolution, which can influence feature representation and, consequently, classification accuracy. Second, although we used two independent datasets, both were based on CT scans and may not capture the full spectrum of imaging variability present in real-world clinical environments. Factors such as scanner type, acquisition protocol, patient demographics, and institutional bias can significantly impact model performance. Therefore, future work should focus on expanding the dataset diversity to include multi-modal imaging and more varied clinical sources in order to assess robustness and enhance the model's applicability in broader clinical settings.

In clinical applications, the implications of false positives and false negatives must be carefully considered. Misclassifying a malignant tumor as benign or normal could delay urgent treatment and increase the risk of metastasis or poor prognosis. Conversely, false positives may lead to unnecessary interventions or patient anxiety. Our model's high recall (sensitivity) across folds suggests a reduced likelihood of false negatives, which is particularly critical in

cancer detection. The confusion matrices in Figures 4 and 8 further demonstrate the model's ability to correctly classify malignancies with minimal confusion among classes. Moreover, although this study used two independent datasets, both were based on CT imaging modalities. To assess generalization across diverse populations and scanning protocols, we enforced patient-wise data splits and included images with different contrasts and scanner types. Nevertheless, to further validate the model's robustness, future work will explore external multi-institutional datasets and other modalities such as MRI and ultrasound, ensuring broader applicability across clinical settings.

4. 3. 4 Comparative Performance Summary

The full comparison of classification performances of ResNet50, Vision Transformer (ViT), and the proposed hybrid model ResNet50+ViT is presented in Table 4 for both datasets. The hybrid model outperformed the baseline approaches in all the main evaluation metrics. More specifically, it achieved a 3.16% improvement in F1-score over ViT on the KAUH-Kidney dataset, and a 3.49% improvement in recall on the CT-Kidney dataset, indicating a better generalization among the tumor classes, even in terms of data imbalance.

Table 4: Comparative Performance of All Models

Model	Dataset	Accuracy (%)	Precision (Macro)	Recall (Macro)	F1-score (Macro)	Macro AUC (%)	Micro AUC (%)
ResNet50	KAUH-Kidney	99.23	99.39	99.31	99.35	99.98	99.99
ViT		96.27	96.49	96.19	96.31	99.6	99.61
Hybrid (ResNet+ViT)		**99.53**	**99.6**	**99.35**	**99.47**	**100**	**100**
ResNet50	CT-Kidney	98.64	96.4	95.03	95.7	99.67	99.84
ViT		99.05	97.51	96.21	96.84	99.85	99.95
Hybrid (ResNet+ViT)		**99.73**	**99.08**	**99.7**	**99.39**	**100**	**100**

4. 3. 5 Comparison with Other Related Studies

The hybrid deep learning models that we proposed (ResNet50 + ViT) achieved better performance than the previous published studies in the classification of the kidney CT images into three clinical categories: Normal, Benign Tumors and Malignant Tumors. Previous studies have largely focused on binary classification or categorization of conditions; in contrast, this study proposes a reliable and accurate multi-class classification leading to clinically acceptable performance, while further facilitating feature learning through integrated convolutional and transformer-based modeling. The study by Alzu'bi et al. (2022) employed a two-stage approach. They first classified whether a tumor was present (Normal vs Tumor), then if the Tumor was Benign or Malignant. They used models like CNN-6, ResNet50, and VGG16 for detection, with CNN-6 achieving **97% accuracy**. For tumor type classification, they used a simpler CNN-4 model, which reached **92% accuracy**. However, their system did not combine Normal, Benign, and Malignant into one classification task, and they only used data from one hospital. **Özbay et al. (2024)** proposed a self-supervised learning model called SSLSD-KTD, using Masked Autoencoders and Self-Distillation. Their model classified CT images as either **Normal** or **Tumor**, without separating tumor types. It achieved **98.04% accuracy** on the KAUH dataset and **82.14%** on the CT-Kidney dataset. After applying transfer learning, accuracy increased to **99.82%** and **95.24%,** respectively. While the results were strong, the model did not classify Benign and Malignant tumors. In another study, Majid et al. (2023) implemented both traditional machine learning (e.g., Random Forest, SVM) and deep learning models (DenseNet-121, ResNet-101) for Normal versus Tumor classification using a dataset of 12,446 CT images. DenseNet-121 achieved an accuracy of 98.22%, but the model did not include separate categories for Benign and Malignant tumors. The focus was mainly on detecting whether a tumor was present. **Prommakhot and Srinonchat (2024)** proposed a hybrid model using **VGG16 and VGG19**, targeting multi-class classification of four categories: **Normal**, **Cyst**, **Stone**, and **Tumor**. While the model achieved **100% accuracy** across all classes. However, it did not divide tumors into **Benign** and **Malignant** categories, and it was not tested on different datasets, which raises concerns about possible overfitting and how well the model would perform on new, unseen data. In contrast, this study presents a single multi-class classification model that was trained and tested separately on two datasets: KAUH-Kidney and CT-Kidney. Unlike previous works that used multiple stages, this approach works in one step. The proposed hybrid model, combining ResNet50 and Vision Transformer (ViT), achieved 99.53%

accuracy on the KAUH-Kidney dataset and 99.73% accuracy on the CT-Kidney dataset. It also achieved AUC scores of 100% (perfect) as macro and micro metrics and high F1-scores of 99.47% and 99.39%, respectively. The results demonstrate that the model outperforms other previous methods and provides a reliable and practical approach for classifying kidney tumors for clinical utility in routine cytology practice. Table 5 shows a detailed comparison of the classification tasks, methods, datasets, and accuracies of all the studies referenced in this comparison.

Table 5: Comparative analysis of recent kidney tumor classification methods using CT images

Ref.	Classification Task	Method	Dataset	Accuracy
[6]	Stage 1: Normal vs Tumor Stage 2: Benign vs Malignant	CNN-6, ResNet50, VGG16 (detection); CNN-4 (classification)	KAUH-Kidney (8,400 images)	97% (Detection), 92% (Classification)
[16]	Normal vs Tumor	SSLSD-KTD (Masked Autoencoder + Self-Distillation)	KAUH-Kidney, CT-Kidney	99.82% (KAUH), 95.24% (CT)
[18]	Normal vs Tumor	DenseNet-121, ResNet-101, Random Forest (RF), Support Vector Machine (SVM), LightGBM (LGBM)	CT-Kidney (12,446 images)	98.22%
[2]	Normal, Cyst, Stone, Tumor	VGG16, VGG19 (combined model)	CT-Kidney (12,446 images)	100%
Proposed method	**Normal, Benign, Malignant**	Hybrid Model: ResNet50 + Vision Transformer (ViT)	KAUH-Kidney, CT-Kidney	**99.53%** (KAUH), **99.73%** (CT)

5 Conclusion

In this study, a hybrid deep learning framework combining ResNet50 and Vision Transformer (ViT) was proposed for multi-class automated classification of kidney tumors from CT images. The fully automated model was developed to classify kidney conditions directly into three clinically relevant categories: normal tissue, benign tumor, and malignant tumor, thus overcoming the limitations of previous research, which mainly focused on binary classification or multi-stage processes. The proposed model was trained and tested separately on two different CT datasets derived from independent sources, namely KAUH-Kidney and CT-Kidney. It showed superior performance, obtaining strong accuracy and AUC scores on both datasets. The model adopted CNN-based spatial feature extraction to learn local features and transformer-based global context learning to capture complex patterns in tumors. These results demonstrate that the hybrid model developed through the proposed method has strong potential to enhance the accuracy and automation of kidney tumor diagnosis, thereby supporting radiologists. Future work may include model robustness improvement, use of the model on other imaging modalities, and integration of more clinical data to improve diagnostic performance and generalization. In addition, we plan to validate the proposed model using real-world clinical datasets from collaborating healthcare institutions. This external validation will allow us to evaluate the model's generalizability and robustness across diverse imaging conditions, patient populations, and clinical settings

Authors contribution

Hivi Kamal Ismael contributed to the conceptualization, methodology design, implementation, data preprocessing, model development, experimentation, analysis of results, and manuscript writing. Wafaa Mustafa Abduallah contributed to the supervision of the research, validation of the methodology, interpretation of the results, critical revision of the manuscript, and overall guidance of the study. Both authors reviewed and approved the final version of the manuscript.

Conflict of interests

The authors declare no conflicts of interest related to this study. The research was conducted independently without any financial or commercial relationships that could influence the results or interpretation. The authors also confirm

that no external funding was received for this research. The funders had no role in the design of the study, data collection, analysis, manuscript preparation, or decision to publish.

References

[1] S. Solanki, U. P. Singh, S. S. Chouhan, and S. Jain, "Brain tumor detection and classification using intelligence techniques: an overview," *IEEE Access*, **11**, 12870–12886, (2023).

[2] A. Prommakhot and J. Srinonchat, "VGGNet Integration for Kidney Tumor Classification," *Proceeding - 12th Int. Electr. Eng. Congr. Smart Fact. Intell. Technol. Tomorrow, iEECON* **2024**(iEECON), 1–6, doi: 10.1109/iEECON60677.2024.10537904 (2024).

[3] S. Wang *et al.*, "Comprehensive Analysis of Ferroptosis Regulators With Regard to PD-L1 and Immune Infiltration in Clear Cell Renal Cell Carcinoma," *Front. Cell Dev. Biol.*, **9**(July), doi: 10.3389/fcell.2021.676142 (2021).

[4] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021.," *CA. Cancer J. Clin.*, **71**(1), 7–33, (2021).

[5] B. Ljungberg *et al.*, "European association of urology guidelines on renal cell carcinoma: the 2019 update," *Eur. Urol.*, **75**(5), 799–810, (2019).

[6] D. Alzu'Bi *et al.*, "Kidney Tumor Detection and Classification Based on Deep Learning Approaches: A New Dataset in CT Scans," *J. Healthc. Eng.*, **2022**, , doi: 10.1155/2022/3861161(2022).

[7] U. Capitanio and F. Montorsi, "Renal cancer," *Lancet*, **387**(10021), 894–906, (2016).

[8] I. AlShourbaji, P. Kachare, W. Zogaan, L. J. Muhammad, and L. Abualigah, "Learning features using an optimized artificial neural network for breast cancer diagnosis," *SN Comput. Sci.*, **3**(3), 229, (2022).

[9] Y. Pu *et al.*, "Variational autoencoder for deep learning of images, labels and captions," *Adv. Neural Inf. Process. Syst.*, **29**, (2016).

[10] H. Qassim, A. Verma, and D. Feinzimer, "Compressed residual-VGG16 CNN model for big data places image recognition," in *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, IEEE, 169–175 (2018).

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).

[12] L. Zhou, Z. Zhang, Y.-C. Chen, Z.-Y. Zhao, X.-D. Yin, and H.-B. Jiang, "A deep learning-based radiomics model for differentiating benign and malignant renal tumors," *Transl. Oncol.*, **12**(2), 292–300, (2019).

[13] K.-H. Uhm, S.-W. Jung, S.-H. Hong, and S.-J. Ko, "Lesion-Aware Cross-Phase Attention Network for Renal Tumor Subtype Classification on Multi-Phase CT Scans," 1–14, doi: 10.1016/j.compbiomed.2024.108746 (2024).

[14] V. E. Reuter and S. K. Tickoo, "Differential diagnosis of renal tumours with clear cell histology," *Pathology*, **42**(4), 374–383, (2010).

[15] Z. Lin *et al.*, "Recognizing pathology of renal tumor from macroscopic cross-section image by deep learning," *Biomed. Eng. Online*, **22**(1), 1–20, doi: 10.1186/s12938-023-01064-4 (2023).

[16] E. Özbay, F. A. Özbay, and F. S. Gharehchopogh, "Kidney Tumor Classification on CT images using Self-supervised Learning," *Comput. Biol. Med.*, **176**(May), doi: 10.1016/j.compbiomed.2024.108554 (2024).

[17] P. T. Akkasaligar, S. Pattar, A. Uppin, K. Kori, and A. Kulakarni, "Kidney Tumor Detection Using Computed Tomography Scan Images Through CNN," *Int. Conf. Distrib. Comput. Optim. Tech. ICDCOT 2024*, pp. 1–6, (2024), doi: 10.1109/ICDCOT61034.2024.10516254.

[18] M. Majid *et al.*, "Enhanced Transfer Learning Strategies for Effective Kidney Tumor Classification with CT Imaging," *Int. J. Adv. Comput. Sci. Appl.*, **14**(8), 421–432, doi: 10.14569/IJACSA.2023.0140847 (2023).

[19] M. N. Islam, M. Hasan, M. K. Hossain, M. G. R. Alam, M. Z. Uddin, and A. Soylu, "Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography," *Sci. Rep.*, **12**(1), 1–14, (2022).

[20] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. big data*, **6**(1), 1–48, (2019).

[21] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv Prepr. arXiv1712.04621*, (2017).

[22] D. Al-obidi and S. Kacmaz, "Facial Features Recognition Based on Their Shape and Color Using YOLOv8," in *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, IEEE, 1–6 (2023).

[23] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, and G. Srivastava, "Deep neural networks to predict diabetic retinopathy," *J. Ambient Intell. Humaniz. Comput.*, **14**(5), 5407–5420, doi: 10.1007/s12652-020-01963-7 (2023).

[24] S. Babichev, I. Liakh, and I. Kalinina, "Applying a Recurrent Neural Network-Based Deep Learning Model for Gene Expression Data Classification," *Appl. Sci.*, **13**(21), 11823, (2023).

[25] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," in *Healthcare*, MDPI, 541 (2022).