Journal of Artificial Intelligence in Medical Applications (JAIMA)
Vol. 1, No. 1, 2025, **pp**. 89–105
ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.740

Integrating Vision Transformers with Transfer Learning for Enhanced Kidney Cancer Classification in CT Imaging

Ali Mahmoud Ali [1,*], Mahmood Khalsan[1,2,3], Muntadher Idrees Ali[1,4], Mabrouka Ali Jelban[5], Teresa Abuya[6]

[1]Department of Information Technology, College of Science, University of Warith Al- Anbiyaa, Karbala 56001, Iraq; ali.mahmoud@uowa.edu.iq, mahmood.jasim@uowa.edu.iq, muntather.adriess@uowa.edu.iq
[2]Scientific Department, Warith International Cancer Institute, Karbala, 56001, Iraq; mahmoud@mizan.edu.iq
[3]Southern Italy Tailored Medicine Innovation Center, STaiMIC (Startup Innovativa S.R.L.), Bari, Italy.
[4]Department of Artificial intelligence and Robotics, Computer engineering, ShahidBahonar University of Kerman, Kerman, Iran; mintidriss@gmail.com
[5]College of Nursing, University of Tripoli, Tripoli, Libya; M.jelpan@uot.edu.ly
[6]Department of Computing Sciences, School of Information Science and Technology, Kisii University, Kisii P.O. Box 408-40200, Kenya; tkwamboka@kisiiuniversity.ac.ke

Received: 27/08/2025, Revised: 19/10/2025, Accepted: 07/11/2025, Published: 30/12/2025

ABSTRACT: Accurate classification of kidney tumors from CT images remains a significant challenge due to the complexity of anatomical structures and the visual similarity between normal and pathological tissues. While conventional CNN-based methods have achieved promising results, they often suffer from overfitting and limited generalization, especially when trained on fixed 80-20 splits. This study aims to design a lightweight yet accurate framework for kidney tumor classification using a Vision Transformer Transfer Learning approach. The proposed model leverages a ViT-Tiny architecture pre-trained on ImageNet and fine-tuned on a carefully curated subset of the CT Kidney Dataset, distinguishing between normal and tumor cases. Preprocessing steps and optimization strategies were employed to enhance both model performance and computational efficiency. To improve robustness and avoid the overfitting, the model was evaluated using both 5-fold and 10-fold cross-validation, along with a more challenging 69%–31% train–test split. Experimental results show that the VTTL framework achieved a classification accuracy of 99.32%, an F1-score of 97.97%, a precision of 98.28%, and a recall of 97.97%, outperforming or matching several state-of-the-art CNN, hybrid, and transformer-based methods. The VTTL model offers a powerful and efficient solution for automated kidney tumor detection, making it highly suitable for deployment in low-resource clinical environments and online diagnostic platforms.

Keywords: Kidney Cancer, Vision Transformer, ViT, Transfer Learning, ViT-tiny, CT images.

1. INTRODUCTION

Kidney cancer, particularly renal cell carcinoma, is one of the most prevalent malignancies of the urinary tract and accounts for a significant number of cancer-related deaths worldwide [1]. Early and accurate diagnosis is critical for effective treatment planning and long-term patient survival. However, many kidney tumors remain asymptomatic in their early stages and are often discovered incidentally during imaging for unrelated conditions. As the incidence of kidney cancer continues to rise, especially in aging populations, there is an urgent need for reliable diagnostic tools that can support timely and precise identification of abnormal kidney growths [2].

Medical imaging plays a central role in the detection, diagnosis, and monitoring of kidney cancer. Among various imaging modalities[3], Computed Tomography (CT) remains the gold standard for evaluating kidney tumors due to its ability to provide high-resolution, cross-sectional images of renal structures. CT scans enable clinicians to assess tumor size, shape, location, and vascular involvement with considerable detail. The widespread availability and rapid acquisition of CT images have made them indispensable in clinical workflows[4], particularly for staging and surgical planning. However, the manual interpretation of CT scans is time-consuming and subject to inter-observer variability, highlighting the need for automated and intelligent diagnostic systems. Despite significant advances in computer-aided diagnosis, traditional methods—particularly those based on classical machine learning or early deep learning models like CNNs—face notable limitations [5]. CNNs often require large amounts of labeled data to perform well, which is a challenge in the medical domain where data annotation is time-consuming and requires expert input. Additionally, these models are typically resource-intensive, relying on deep, complex architectures that may not be feasible for deployment in low-resource clinical settings. Furthermore, they tend to focus on local features due to their convolutional nature, which may lead to the loss of important global context in complex

medical images[6]. These limitations underscore the need for more efficient, scalable, and context-aware models for accurate kidney tumor classification.

Although Convolutional Neural Networks (CNNs) continue to dominate medical image classification[7], they often fall short in capturing long-range dependencies and spatial relationships within complex medical imaging data [1]. Vision Transformers (ViTs) offer a promising alternative by leveraging global attention mechanisms to model relationships across image patches. However, their adoption in medical applications remains limited due to three key challenges. First, data scarcity is a major obstacle. ViTs typically require large-scale datasets for effective training, yet most medical datasets are relatively small and imbalanced. As a result, many existing studies struggle to adapt ViTs to these constrained conditions without encountering overfitting issues [8]. Second, computational demands present a barrier to practical use. Many ViT architectures are computationally intensive, making them unsuitable for deployment in low-resource settings, such as hospitals and clinics in developing regions. Third, contextual and representational bias remains an underexplored concern. There is a lack of comprehensive evaluation of ViT-based models trained exclusively on medical data—such as CT scans—especially in diverse and non-Western healthcare contexts. This study addresses these limitations by (i) employing transfer learning with a lightweight variant of the Vision Transformer to reduce model complexity, and (ii) developing tailored preprocessing and augmentation strategies to convert grayscale CT images into a format compatible with ViT input requirements. These innovations aim to make transformer-based solutions more practical and accessible for medical image classification, without demanding high-end computational resources.

The aim of this study is to develop an efficient and lightweight deep learning framework for the accurate classification of kidney tumors using CT images. By leveraging a Vision Transformer Transfer Learning (VTTL) approach, the study seeks to address the limitations of traditional CNN-based models in handling complex spatial relationships and global image context, while also overcoming common challenges associated with data scarcity and high computational demands. The proposed framework is designed to balance diagnostic performance with efficiency, making it suitable for deployment in real-world clinical environments, including those with limited computational resources. The main contributions of this study can be summarized as follows:

- Developed a resource-efficient Vision Transformer Transfer Learning (VTTL) framework using a pre-trained ViT-Tiny model, adapted for binary classification of kidney CT images (normal vs. tumor).
- Designed a preprocessing pipeline to convert grayscale CT images to RGB format, normalize them according to ImageNet standards, and apply data augmentation to enhance generalization.
- Demonstrated the practical effectiveness of transfer learning on a real-world medical dataset collected from hospitals in Dhaka, Bangladesh, despite the limited sample size.
- Optimized the model for low-resource environments using minimal input resolution (96×96), small batch sizes, and early stopping, enabling deployment without high-end GPUs.
- Conducted comprehensive performance evaluation by comparing the proposed model with MLP and CNN-LSTM architectures, showing superior accuracy and strong generalization.
- Ensured transparency and reproducibility by using a publicly available dataset and clearly reporting training configurations to support replicability in the medical AI research community.

The remainder of this paper is organized as follows. Section 2 reviews the related work on CT image-based detection. Section 3 describes the proposed approach, including data preparation, the Vision Transformer with transfer learning, and dataset collection. Section 4 presents the experimental results and discussion. Finally, Section 5 concludes the paper.

2. RELATED WORKS

Recent advancements in machine learning (ML) and deep learning (DL) have significantly contributed to the improvement of classification accuracy in various cancer detection tasks, including kidney cancer. Traditionally, earlier methods relied on hand-crafted features combined with statistical classifiers to detect abnormalities in CT images. However, these techniques often struggled to capture the complex visual patterns required for reliable medical diagnosis. In contrast, Convolutional Neural Networks (CNNs) have emerged as a dominant paradigm, integrating feature extraction and classification into a unified framework, resulting in notable improvements in diagnostic performance [1]. The introduction of the Vision Transformer (ViT) by Dosovitskiy et al. (2021) [9] marked a breakthrough in computer vision, replacing convolutional operations with attention-based mechanisms. Unlike CNNs, ViTs divide images into fixed-size patches and use transformer encoder blocks to model long-range dependencies and capture global spatial context. While early ViT models

were considered data-hungry and computationally demanding, recent variants such as ViT-Tiny and DistilViT have demonstrated reduced complexity and improved usability for smaller datasets.

Transfer learning has become an essential technique in medical image analysis, especially where annotated datasets are limited [10]. By leveraging representations learned from large-scale natural image datasets, models can be fine-tuned to perform specific medical tasks. Within the ViT domain, transfer learning has proven valuable in adapting large transformer models to smaller-scale medical imaging problems. For example, Chen et al. (2021) [11] and Z. Chen et al. (2023) [12] successfully applied ViT-based transfer learning to detect retinal diseases, classify lung abnormalities, and analyze histopathological data—achieving performance comparable to or better than CNN-based models. In kidney imaging, CNNs have traditionally been used to identify conditions such as stones, tumors, and cysts from CT scans, using architectures like ResNet, DenseNet, and Inception. Bingo et al. (2023) [13], for instance, used a CNN-based model to classify kidney diseases as a multiclass problem, achieving high accuracy but also encountering challenges such as data imbalance and overfitting.

Despite these advancements, few studies have explored the application of transformer-based models for kidney CT analysis. Given the intricate textural and spatial characteristics of CT scans, the self-attention mechanism of ViTs may offer superior performance by capturing clinically relevant patterns without relying on the inductive biases inherent to CNNs. Recent research has started to evaluate hybrid and transformer-based models in this context. Islam et al. (2022) [14], in a comprehensive study using 12,446 kidney CT images, compared three CNN models (VGG16, InceptionV3, ResNet50) with three transformer-based models (EANet, CCT, Swin Transformer). Their results showed that the Swin Transformer achieved an impressive 99.3% accuracy, outperforming all CNN counterparts. However, their study did not assess computational efficiency in real-time settings, limiting its practical implications. Similarly, Asif et al. (2022) [15] proposed a VGG19-based model with an augmented naive Inception module, achieving 99.25% precision, but at a high computational cost. Bhandari et al. (2023) [16] proposed a lightweight CNN model with interpretability tools (LIME, SHAP), attaining 99.52% accuracy. More recently, Maqsood et al. (2024) [17] introduced SpinalZFNet, a fusion of SpinalNet and ZFNet, which achieved 99.8% accuracy and demonstrated strong sensitivity and specificity. Nevertheless, these CNN-based models lack the global attention mechanism that is central to ViTs, which is crucial for identifying subtle diagnostic patterns in complex CT data.

In low-resource clinical environments, deep learning faces several constraints, including limited computational hardware, small and imbalanced datasets, and a shortage of annotated medical images. For real-world applicability, there is a pressing need for lightweight, low-latency, and efficient models [31]. Techniques such as model pruning, quantization, and architecture simplification (e.g., ViT-Tiny) have been introduced to reduce computational demands while preserving performance [18]. Research in this direction focuses on achieving high accuracy and generalization with minimal hardware requirements, particularly in clinical decision-support systems. This study contributes to that goal by demonstrating how a compact ViT model can achieve competitive classification performance on a medical imaging task, despite limited data and computational resources. The practical implementation of deep learning in under-resourced clinical settings requires models that are not only accurate but also lightweight and easily deployable on standard hardware. Challenges such as restricted GPU access, minimal training data, and the need for real-time inference demand carefully optimized solutions. Although recent work by Bhandari et al. [16] and Maqsood et al. [17] has highlighted the potential of lightweight CNN models, there remains a lack of research on compact Vision Transformer architectures under similar constraints. This study addresses that gap by introducing a ViT-based transfer learning framework capable of performing accurate binary classification of kidney tumors from CT scans in low-resource environments. A summary of the related works discussed in this section is presented in Table 1.

Table 1: Summary of Related Work in Kidney CT Classification and Vision Transformer Applications

Author / Year	Method/Model	Key Findings	Limitations
Dosovitskiy et al. (2021) [9]	Vision Transformer (ViT)	Replaced CNNs with self-attention for global feature learning	Initially data-hungry, improved in later versions (ViT-tiny, DistilViT)
C.-F. Chen et al. (2021) [11]	Cross-Attention Multi-Scale Vision Transformer	Handle with small and large patch tokens efficiently	Outperformed DeiT by ~2% on ImageNet1K with only moderate increase in complexity. Efficient with linear attention cost.
Z. Chen et al. (2023) [12]	ViT-Adapter	Introduced a lightweight, pretraining-free adapter for plain ViT, enabling it to perform on par with vision-specific transformers in dense vision tasks.	Achieved 60.9 box AP and 53.0 mask AP on COCO test-dev, without extra detection data. Suitable for downstream adaptation.
Bingo et al. (2023) [13]	CNN (ResNet/DenseNet)	High accuracy in detecting cysts, stones, tumors	Struggled with imbalance and overfitting

Islam et al. (2022) [14]	Swin Transformer, CCT, EANet vs CNNs	Swin Transformer achieved 99.3% accuracy—best overall	The lack of comparison with other existing models or datasets for kidney disease diagnosis.
Asif et al. (2022) [15]	VGG19 + Naïve Inception (TL)	99.25% accuracy, effective TL method	Computationally heavy
Bhandari et al. (2023) [16]	Lightweight CNN + XAI (LIME, SHAP)	99.52% accuracy with interpretability, efficient design	No external dataset validation
Maqsood et al. (2024) [17]	SpinalZFNet (SpinalNet + ZFNet)	99.8% accuracy, excellent sensitivity/specificity	Performance depends on high-quality pre-processing
Maniyar et al. (2023) [31]	CNN	Achieved 92% accuracy	Lower than other DL methods, lacks ViT comparison

As evident from the reviewed literature, existing studies have demonstrated the promise of both CNN-based and transformer-based models for kidney CT image classification. However, several gaps remain unaddressed. Many transformer-based models, such as those proposed by Dosovitskiy et al. and Islam et al., show impressive accuracy but suffer from high computational requirements and limited real-world validation in low-resource settings. Others, like Bhandari et al. and Maqsood et al., introduce lightweight CNN architectures but lack the global context modeling capabilities inherent to Vision Transformers. Furthermore, most works rely on large-scale or externally preprocessed datasets, with little focus on efficiency or deployment feasibility in practical clinical environments. To fill these gaps, our proposed approach integrates a compact Vision Transformer (ViT-Tiny) with transfer learning and a customized preprocessing pipeline, specifically optimized for binary classification of kidney tumors in grayscale CT images. By achieving high accuracy on a modestly sized, real-world dataset while maintaining low computational cost, our framework offers a balanced, deployable solution suitable for resource-constrained clinical settings.

3. The Proposed Methodology

The proposed methodology follows a structured Vision Transformer Transfer Learning (VTTL) pipeline tailored for efficient kidney CT image classification. Initially, the model parameters are defined, including image size, batch size, learning rate, number of epochs, weight decay, and early stopping criteria. The training components—such as the loss function and optimizer—are then set up to guide model convergence. The dataset is prepared by applying appropriate image transformations, constructing a custom dataset, and creating DataLoaders for efficient training. The ViT model is configured by selecting a suitable architecture (ViT-Tiny), loading a pre-trained model, and attaching a classification head specific to the binary task. Once configured, the training loop is executed, consisting of training and validation phases with early stopping to prevent overfitting. Throughout training, the learning curve is monitored by tracking average losses and identifying signs of overfitting [19]. Finally, the trained model is evaluated on the validation set, and key performance metrics are generated to assess accuracy, precision, recall, and overall effectiveness. The details of the proposed methodology are presented in Figure 1.

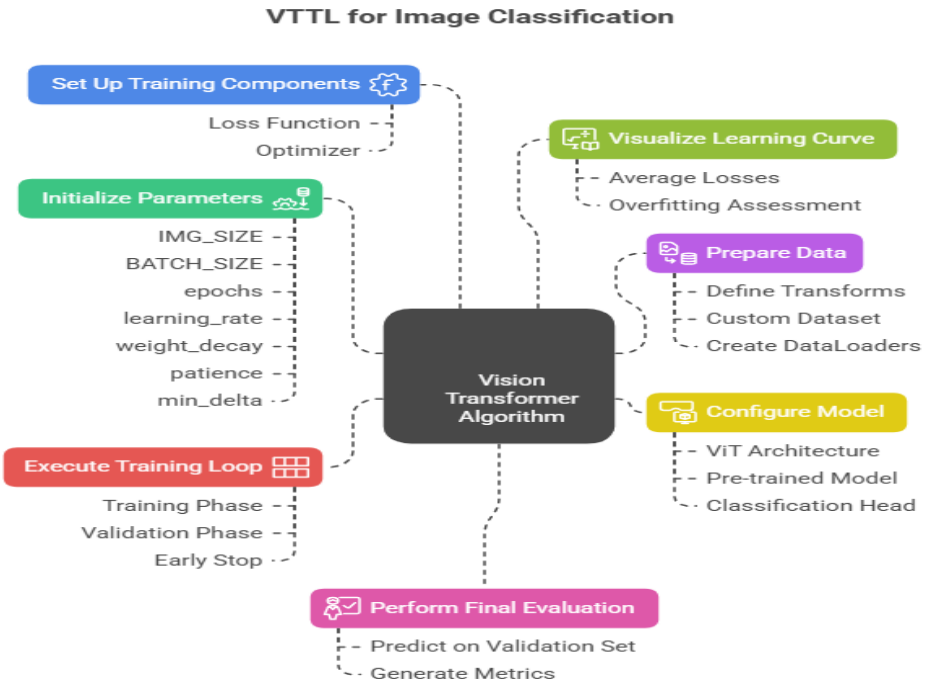

Figure 1. The proposed Methodology

3.1 The Dataset Acquisition and Preparation

The study employs the CT Kidney Dataset, which is a retrospective collection of labelled CT images within the Picture Archiving and Communication System (PACS) of various hospitals in the city of Dhaka in Bangladesh. Patients were categorized or diagnosed according to four criteria: normal, which included the number of samples (5077); cystic, which included the number of samples (3709); and tumour-filled. Last but not least, there are 2283 samples in the fourth category, where the patient has a cyst, and there are 1377 samples overall. This means that there are 12446 gland samples in all categories. Figure (2) shows representative examples of the dataset. It should be mentioned that the DICOM-type pictures utilized were converted to JPEG format after each patient's information was removed. The dataset's classes are shown in Figure (3). You can download the dataset from https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone. The dataset was used in two sections, training and validation:

- Training Set: 5,161 (3,553 Normal) + (1,598 Tumour) images.
- Validation Set: 2,209 images (1,524 normal + 685 tumour).

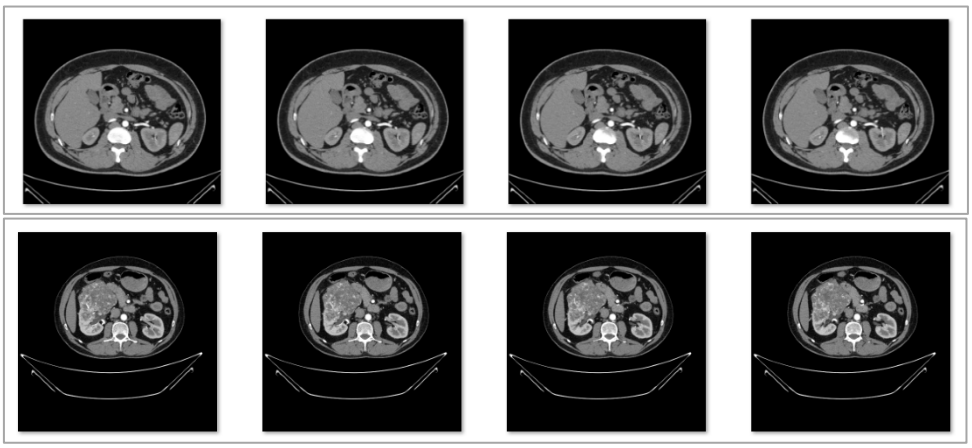

Figure 1. Examples of the utilized dataset

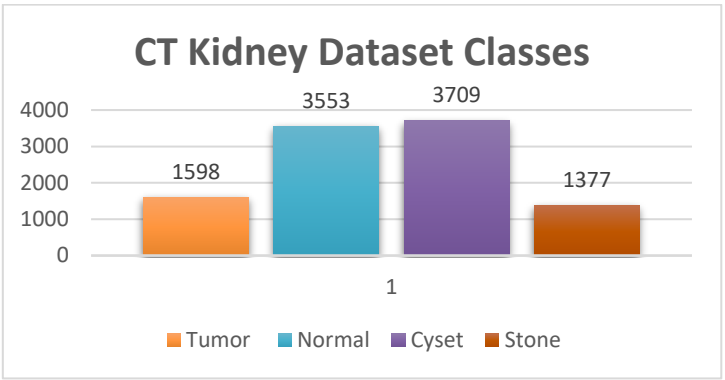

Figure 2. CT Kidney Dataset Classes.

3.2 Image Preprocessing and Data Augmentation

Images are then processed with a selected combination of the following torchvision. transforms preprocessing and augmentations to each image loaded by Kidney CT Dataset: Summary: Kidney CT Image Preprocessing and Augmentation.

1. Image Resize

All images were resized to 224×224pixels to reduce memory and computational cost.

2. Grayscale to RGB Conversion

Greyscale images are converted to 3-channel RGB by stacking the same channel three times:
np.stack([img, img, img], axis=-1) Necessary for compatibility with pre-trained ViT models.

3. Tensor Conversion & Normalization

Images converted to PyTorch tensors using transforms. ToTensor().
Normalized using ImageNet mean and std values:

Mean: [0.485, 0.456, 0.406]
Std: [0.229, 0.224, 0.225]
Ensures compatibility with transfer learning from ImageNet.
4. Data Augmentation (Training Set Only)
Random Horizontal Flip: with 50% probability.
Random Rotation: by ±15 degrees.
Purpose: increase diversity of training data and reduce overfitting.
5. Data Loader Settings
BATCH_SIZE = 16 for memory efficiency.
num_workers = 2 for parallel loading to avoid I/O bottlenecks.

This preprocessing pipeline improves efficiency and compatibility with ViT models and helps achieve better generalization during training.

3.3 Model Architecture Vision Transformer with Transfer Learning (VTTL)

The proposed model employs a Vision Transformer (ViT) architecture enhanced through transfer learning for the classification of kidney CT images. Unlike Convolutional Neural Networks (CNNs), which rely on local receptive fields and hierarchical feature extraction, the ViT processes images as sequences of fixed-size patches. Each input image is divided into non-overlapping patches (e.g., 16×16), and these patches are then linearly embedded and combined with positional encodings to retain spatial information. A special classification token ([CLS]) is also appended to the sequence, which ultimately gathers the output representation used for classification. This entire sequence is then fed into the transformer encoder blocks, originally designed for natural language processing tasks, to learn contextual relationships across the entire image.

To mitigate the data and computational demands typically associated with training ViTs from scratch, a transfer learning strategy is employed. A pre-trained ViT model (e.g., ViT-Tiny) trained on large-scale datasets like ImageNet is fine-tuned on the specific CT kidney dataset. This allows the model to benefit from general visual feature representations learned during pretraining, making it more suitable for the relatively small and domain-specific medical dataset. The final classification is performed by a lightweight classification head consisting of dropout and linear layers, distinguishing between normal and tumor classes. This design not only ensures high accuracy but also makes the model efficient enough for real-world medical applications, including low-resource clinical settings. The details of the proposed model are presented in Figure 4 and Table 2.

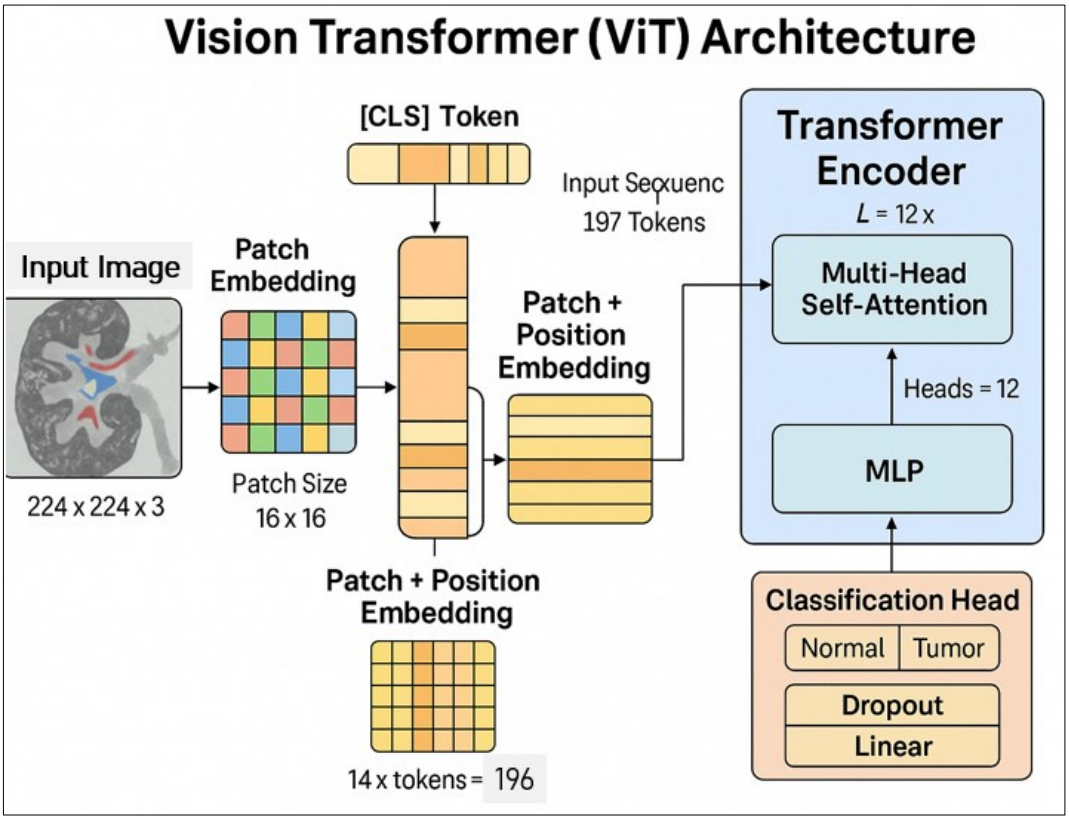

Figure 3. VIT Architecture

Table 2: Information of the Vision Transformer (ViT) image processing

Element	Value	Explanation
Image Size	224×224	Automatically resized by timm to match ViT input
Patch Size	16×16	Fixed value defined in the model architecture
Number of Patches	14 * 14 = 196	Number of patches along the width: 224/16=14 Number of patches along the height: 224/16=14
Total Number of Tokens	197	196 patches + 1 classification token [CLS]

3.3.1 VTTL Pre-trained Model

The timm library is loading and loading the vit_tiny_patch16_224 pre-trained ViT model. This specific variant was selected because its size and the amount of resources needed are significantly reduced in contrast to other larger ViT models (e.g., vit_base_patch16_224), which is why this variant can be successfully applied in the environment that has limited computational resources. Its pretrained=True indicator makes sure that the model will be initialized on the weights trained on the ImageNet dataset. The size of input images is passed to timm. create_model with the parameter img_size=224, which tells the model input image dimensions so that it can strategically adjust the size of its patch embeddings.

3.3.2 Outline of ViT Architecture (Conceptual)

The ViT architecture works the following way:

Patch Embedding: The given image is then broken into a grid of small patches of a set size (e.g., 16x16 pixels in patch16). Every patch is then squashed into a one-dimensional vector and linearly embedded into higher-dimension space.

- Positional Embeddings: To ensure spatial information, a positional embedding is learnt in alignment with the patch embeddings, which are annotations of the original position of each patch within the image.
- [CLS] Token: A special classifiable token ([CLS]) is added before the patch embedding sequence. The last state of this token through the Transformer encoder is utilised in classifying.
- Transformer Encoder blocks: the heart of the ViT. It is made of several layers that are the same (e.g., depth=6 in ViT-tiny). Within every block there are:
- Multi-Head Self-Attention (MHA): This module is used to give individual significance (or patches/tokens) compared to each other when processing a given patch. It uses query, key, and value vectors to calculate what it terms attention scores using input embeddings. The attention mechanism is Attention(Q,K,V)=softmax(dk QKT)V when there are n heads and dk = the dimension of the key vectors. The multi-head attention mechanism enables the model to attend to any information simultaneously across several subspaces of representations.
- Layer Normalization (nn.LayerNorm): Rescales the input to each sublayer before using MHA and MLP to normalize the inputs.
- Feed-Forward Network (MLP): A two-layer neural network that is used separately on every entry in the sequence.
- Classifier Head: The result of the Transformer encoder of the [CLS] token is sent through a closing nn.Linear layer (self.vit.head) to create the classification logits to the num_classes (2 in this paper). A new linear layer is added (to replace the original one in the pre-trained model), which fits the task.

The whole VTTL model was subsequently transferred to the suitable computing device, with preference being given to a CUDA-enabled GPU where possible; otherwise, it fell back to the CPU.

3.3.3 Model VTTL Training

In the training process, epochs could be arbitrarily limited to a maximum of 50, with early stopping applied to avoid overfitting and to find the best state of the model.

- Loss Function: Cross-Entropy Loss was chosen as an optimization target [20], which is recommended to use in the case of binary classification. The loss of a batch of N samples is a:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} * log(\hat{P}ij) \quad (1)$$

in which C represents classes, and yij = 1 in case sample i belongs to the jth class and 0 otherwise, where p^ij are the estimated probabilities.

- Optimizer: Adam optimizer (torch.optim.Adam) was used as the parameter update. As it is a trend in fine-tuning a pre-trained model, a learning_rate of 0.0001 was chosen.
- L2 Regularization (Weight decay): The weight_decay of 1e-4 was utilized in the Adam optimizer. This L2 regularization term increases the penalty on large weights, which in turn promotes simpler patterns of learning by the model and makes it less likely to overfit [21]:

$$L_{total} = L_{original} + \frac{1}{2} * \lambda \sum_{\omega \in W} \omega^2 \quad (2)$$

- Early stopping logic: It is an essential mechanism that may be applied to dynamically terminate training when generalization performance in the validation set stabilizes, or shows a worse functioning so that computational resources may be saved and selecting the best-performing model may be achieved.
- patience = 7: The training loop will delay 7 successive epochs and not show important progress in Val Loss.
- min_delta = 0.0001: Val Loss should increase beyond 0.0001 in order to say that a difference is an improvement.The best_val_loss is monitored, and in case a smaller Val Loss is found (more than min_delta improvement), the model state_dict() is stored into best_vit_model_low_resource.pth.
- In case that Val Loss fails to improve within patient epochs, the training loop is halted.
- The weights are then loaded in the model after the loop to make sure that the evaluation will be done using the version of the model with the best performance. At every epoch, the model switches between training (model.train()) and evaluation (model.eval()) mode, (1) computing and accumulating train_loss and val_loss. Training progress was monitored with the help of progress bars (tqdm). Table 3 presents a detailed overview of the and neurons (parameters) applied in the proposed model VTTL.

Table 3. Layers and Neurons (Parameters) Overview

Component	Details
Patch Size	16×16 pixels
Input Image Size	96×96 (customized in your code via img_size=96)
Number of Patches	(96 / 16) ² = 36 patches
Embedding Dimension	192 (each patch gets a 192-dimensional vector.)
Transformer Encoder Layers	12 layers (blocks)
Multi-Head Self-Attention	Each block has 3 attention heads
MLP Hidden Layer Size	Each MLP in a block has a hidden size of 4×192=768 neurones.
Total Parameters	≈ 5.7 million (pretrained ViT-Tiny)
Final Classification Head	You override it with nn.Linear(192,2) for binary classification

3.4 Model Evaluation

After training each of the four models, including the Vision Transformer (ViT), Multi-Layer Perceptron (MLP), and the Hybrid CNN-LSTM, a thorough assessment of the model is carried out based on a standard set of performance measures and visualizations that are applied to the validation set of each[22]. Model evaluation is essential in healthcare and algorithm development since it gives a measure of the model for decision-making in different areas. The performance of the model is compre hensively assessed using a classification report, which provides a detailed breakdown of key metrics for each class. These metrics are calculated from the gathered predictions and true labels. A typical classification report, often generated using a library like sklearn.metrics.classification_report, includes the following:

1. Accuracy: Percentage ratio of the number of cases that have been classified correctly out of all the cases [23]. It becomes very useful when the distribution differences of the class are not considerable.

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

2. Precision: Precisely the ratio between actual true positive records and the records that have been categorized as positive in the classification. It measures the extent of the reality of the positive prediction[24].

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

3. (Sensitivity): The true positive instances divided by the total positive instances; this gives the percentage of true positive instances among the actual positive instances [25]. To be specific, it defines the measure of true positive instances of the data that has been modelled.

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

4. F1 Score: The F-measure, which denotes the harmonic mean of both precision and recall, the two measures thus resulting in a single figure. The class distribution is often an imbalanced one in classification problems, and this is where it is particularly helpful [23].

$$F1\_score = \frac{2rp}{(r + p)} = \frac{(2 * TP)}{(2 * TP + FP + FN)} \quad (6)$$

3. Confusion Matrix: As shown in Figure 5, the Seaborn plot was used to plot a confusion matrix so as to visualize and quantify classification errors. This was useful to determine misclassification tendencies and possible bias towards certain classes.

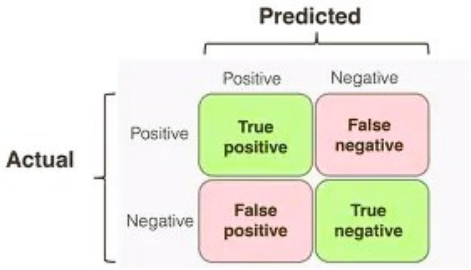

Figure 4. Confusion Matrix

4. AUC and ROC Curve: We calculated Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) of each model. These are measures used to determine how well the models classify the two categories of observations at different thresholds, and they are especially useful when solving medical diagnosis problems.

True Positive Rate (TPR):

$$TPR \frac{(TP)}{(TP + FN)} \quad (7)$$

False Positive Rate (FPR):

$$FPR \frac{(FP)}{(FP + TN)} \quad (8)$$

5. Learning Curves: Loss and accuracy of training and validation curves were then evaluated in epochs to observe the learning dynamics and to identify the association of overfitting or underfitting during training [19].

6. Visualization of Class Distribution: A visual breakdown of the sample count of each resulting class (normal vs. tumour) was also provided to indicate the presence of a possible existing imbalance between classes, which may have a strong impact on model behavior.

7. Cross-Validation: The k-fold cross-validation was carried out to avoid single-data-split overfitting and have a robust model. This gave averaged performance measures over several folds to increase the accuracy of the comparison between models [26]. This predetermined and unifying system of evaluation gives a transparent and rigorous benchmarking of the three suggested models, which are VTTL, MLP, and CNN-LSTM, and shows their unique strengths and weaknesses in the binary identification process of kidney CT images. The Vision Transformer (VTTL) model, specifically the vit_tiny_patch16_224 architecture, was employed for image classification. Input images were resized to 96×96 pixels, processed in batches of 8, and trained for a maximum of 50 epochs. The training utilized the Adam optimizer with a learning rate of 0.0001 and a weight decay (L2 regularization) of 1×10 −4. CrossEntropyLoss served as the loss function, and early stopping was implemented with a patience of 7 epochs and a minimum delta of 0.0001 to prevent overfitting.

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.740

Algorithm1: Vision Transformer Transfer Learning for Kidney CT Classification

```
1. Initialization and Imports
Initialize required libraries:
   - OS, NumPy, Pandas, Matplotlib, Torch, Sklearn, Seaborn, Timm, etc.
2. Set Image Size and Transformations
Set IMG_SIZE = 224
Define 'train_transform':
   - Convert to PIL Image
   - Resize to IMG_SIZE
   - Apply random flip and rotation
   - Normalize image
Define 'val_transform':
   - Convert to PIL Image
   - Resize to IMG_SIZE
   - Normalize image
3. Define Custom Dataset Class: KidneyCTDataset
Class: KidneyCTDataset(root_dir, transform)
   - For each subfolder (class label):
      - For each image file:
         - Save image path and label
   Function __len__():
      Return number of images
   Function __getitem__(index):
      - Load image at given index
      - Convert grayscale to RGB (3 channels)
      - Apply transformation
      - Return image_tensor, label_tensor
4. Load Training and Validation Data
Set 'train_dir' and 'val_dir' paths
Create 'train_dataset' using KidneyCTDataset(train_dir, train_transform)
Create 'val_dataset' using KidneyCTDataset(val_dir, val_transform)
Set BATCH_SIZE = 8
Create DataLoaders:
   train_loader ← DataLoader(train_dataset, batch_size, shuffle=True)
   val_loader   ← DataLoader(val_dataset, batch_size, shuffle=False)
Extract class names from dataset
5. Define Vision Transformer Model: VTTL
Class: ViTTransferLearning(num_classes, model_name, pretrained, img_size)
   - Load pretrained ViT using timm.create_model
   - Replace classification head with new Linear(num_features, num_classes)
   Function forward(x):
      Return output of ViT model
6. Training Configuration
Set device = GPU if available else CPU
Initialize model using ViTTransferLearning (vit_tiny_patch16_224)
Set loss function: CrossEntropyLoss
Set optimizer: Adam with lr=0.0001 and weight_decay=1e-4
Set early stopping parameters:
   epochs = 50, patience = 7, min_delta = 0.0001
Initialize best_val_loss = infinity, early_stop = False
Initialize empty lists for train_losses and val_losses
7. Training Loop
For each epoch in epochs:
   - Train Phase:
      For each batch in train_loader:
         - Move inputs and labels to device
         - Zero gradients
         - Forward pass
         - Compute loss
         - Backward pass and update weights
         - Accumulate training loss
   - Validation Phase:
      For each batch in val_loader:
         - Move inputs and labels to device
         - Forward pass
         - Compute loss
         - Accumulate validation loss
   - Calculate average train and val losses
   - Append to loss history
   - Early Stopping Check:
      If val_loss improved:
         - Save model
         - Reset epochs_no_improve
      Else:
```

```
        - Increment epochs_no_improve
    If epochs_no_improve >= patience:
        - Trigger early stopping
        - Load best model weights
        - Break loop
8. Plot Learning Curve
Plot train_losses and val_losses over epochs
9. Model Evaluation
Set model to eval mode
Initialize lists for all_preds and all_labels
For each batch in val_loader:
   - Predict class labels
   - Append predictions and ground truth labels
Print classification report using sklearn
Plot confusion matrix with seaborn
10. Cross-validation
- 5 k-fold cross-validation
- 10-fold cross-validation
-model evaluation after cross-validation
```

4. Results and Discussion

This section provides the experimental findings of the three deep learning models, namely a Multi-Layer Perceptron (MLP), a hybrid Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM), and a Vision Transformer (ViT). All models were tested using the identical binary classification phenomenon (normal vs. tumour), and the performance of the models was evaluated by using confusion matrices, learning curves, classification reports, and cross-validation. The use of the Vision Transformer Transfer Learning (VTTL) solution in binary classification of kidney CT scans, separating Normal and Tumour classes, produced results that are very strong and clinically interesting. Although this model was trained on a small sample size, the VTTL model had remarkable generalization abilities, solid performance, and discriminative strength, and therefore, transformer-based methods can be considered efficient in resource-limited medical imaging settings.

4.1. Training Dynamics and Convergence Behavior

Training was done in an environment that was CUDA enabled, and this made the process of computation quite efficient on various epochs. The number of training epochs completed was 16, and the early stopping functionality was set to stop training in case the validation loss did not decrease during seven consecutive epochs. The strategy played a crucial part in reducing overfitting, especially since the size of the dataset was comparatively small. The learning curves as shown in the Figure 6 give an insight into the optimisation path of the model. Remarkably, the training started with a greater initial loss (0.2619) than the validation loss (0.0679) at Epoch 1, which is common in initial stages of a fine-tuning transfer learning process, where pre-trained weights quickly adapt to the new features with regard to the target domain. A short decrease and increase in validation loss occur after Epoch 2, which indicates a temporary instability or the changes of a learning rate scheduler. This volatility, however, was only observed momentarily, and further on, validation loss as well as training loss also fell, accepting each other.

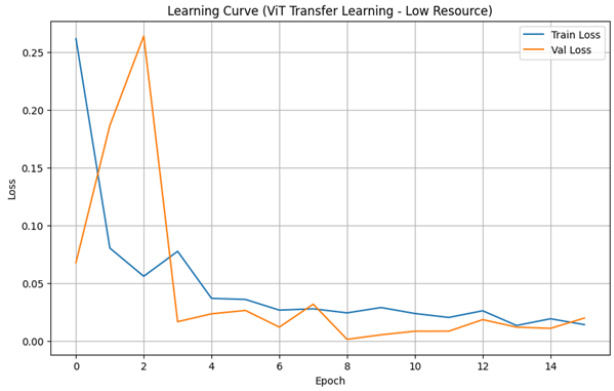

Figure 5. learning curve

A validation loss attained a minimum value of 0.0015 at Epoch 9 and signified the optimal generalization condition of the model. The subsequent gradual flattening of the improvements caused the automatic termination at Epoch 16. The gap between training and validation losses is small at every step of

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.740

training, indicating well-managed model fitting and the lack of overfitting. Table 4 provides a details comparative analysis of the evaluated models MLP, CNN-LSTM, and ViT (VTTL).

Table 4. Comparative Summary and Performance Benchmarking

Metric	MLP	CNN-LSTM	ViT (VTTL)
Accuracy	0.97	0.95	0.994
F1-Score (Macro Avg)	0.97	0.94	0.99
Recall (Tumour)	0.98	1.00	0.99
AUC	0.98	0.99	0.999

4.1.1. Confusion Matrix Interpretation

Interpretation of Confusion Matrix: A detailed picture of the modeled predictions is provided in the confusion matrix (Figure 7):

•True Positives (Normal): The model returned a positive prediction on 1524 normal images and no false positives, thus a good specificity.

•True Positives (Tumour): 684 tumour cases were diagnosed correctly, which shows here a high sensitivity.

•False Positives: Impressively, there were no false positive results of the tumour class, where zero false alarms were dropped by the model that determined whether the patient had a tumour or not.

•False Negatives: A single tumour image was classified as normal in error; however, this was the only error in what was otherwise a perfect performance.

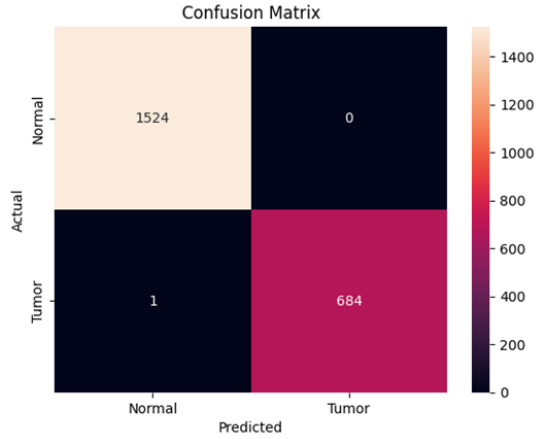

Figure 6. Confusion Matrix

The fact that there are almost no off-diagonal elements shows how well the model draws a discriminative boundary between the classes, which is a crucial need in diagnostic tools where the false negative and the false positive would mean a great deal.

4.1.2. Classification Metrics and Clinical Relevance

Its classification report and the ROC curve are as given below:

Classification Report:

	precision	recall	f1-score	support
Normal	1.00	1.00	1.00	1524
Tumor	1.00	1.00	1.00	685
accuracy			1.00	2209
macro avg	1.00	1.00	1.00	2209
weighted avg	1.00	1.00	1.00	2209

Figure 7: Classification Metrics

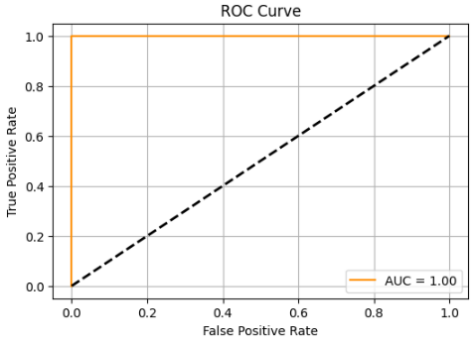

Figure 8: ROC for VTTL

Precision values of 1.00 of the two classes denote that the positive predictions were accurate (both of Normal and Tumour). This can be especially helpful in case of tumour detection because of false positives, where unnecessary biopsies or treatment are to be avoided.

A recall of 1.00 in both classes means that the model was able to capture all the cases. The misclassification in the tumour category does not affect the recall more significantly, because it occupies a small percentage in the total test set.

The precision (F1-score, which is the harmonic mean of precision and recall) was a perfect score of 1.00 across the two classes, which is a balanced model that is not crucially biased towards any of the metrics: precision or recall.The overall accuracy was 100 percent; it reflects perfect classification and 2209 validation samples. Support figures reinstate a slight class imbalance (1524 normal vs. 685 tumour), against which the model functioned just fine without skewing results and tampering with minority-class performance. The performance makes it clear that Vision Transformer outperforms other architecture approaches like MLPs and CNN-based hybrids remarkably in such a task. With only one wrongly classified tumour image, the error rate is only 0.045%, an outstanding rate even as far as medical applications of imaging are concerned, where accuracy in diagnosis is everything. The classification metrics and the ROC curve supporting these results are shown in Figures 8 and 9, respectively. These findings are interesting in clinical view. False positive results are avoided, so there are minimal unnecessary follow-ups to take, and false negatives are only around one percent, so a high diagnostic sensitivity is also achieved, which is important in early tumour detection. Such balance is rarely present at the same time in the traditional models. Further, the success of the model on a small dataset already supports the potential of transfer learning, particularly in combination with the attention capabilities as enabled by the ViT architecture. This enables efficient feature engineering and contextual reasoning, especially critical to those areas of lesser volume with high variation, such as medical imaging.

4.2 VTTL performance on Vision Transformer based on K-Fold cross-validation

In order to rigorously assess the stability, generalizability, and discriminative effectiveness of the Vision Transformer (ViT) framework, a complete K-fold cross validation is utilized through 5-fold and 10-fold settings. It is a common technique in the machine learning literature to be a resilient method to prevent overfitting and to determine model performance on data that the model has never seen. The 5-fold cross-validation results show that the performance of the model is quite high over all folds, where the classification accuracies are 1.00, 1.00, 0.98, 0.93, and 0.99, respectively. The calculated mean accuracy is 0.9798, which indicates good predictive accuracy. A significant decrease in Fold 4 was, however, noticed, with the accuracy falling to 0.93. This degradation can reasonably be explained by an imbalance of the classes in the validation set or an increased number of more heterogeneous/difficult samples that may have introduced distributional variability and hence affected the accuracy of the model. Although this decreased the classification fidelity, the model has maintained satisfactory classification fidelity in the fold. The F1-score of the class of the label Tumour was 0.89, and the Normal one was 0.94, both of which indicate desirable harmony between recall and precision even in more challenging validation groups. As shown in Figure 10.

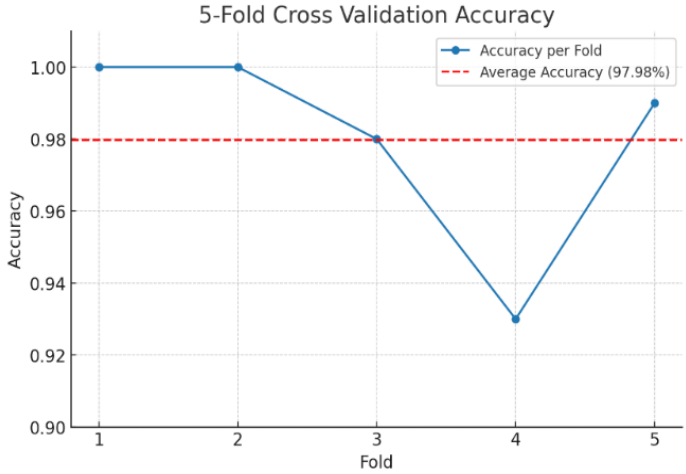

Figure 9. 5-fold Cross-Validation

Folds 1, 2, and 5 achieved near-perfect performance. Fold 4 showed slightly reduced recall for Normal (0.90) and lower precision for Tumour (0.81), resulting in an accuracy dip to 0.93. Despite this, the macro and weighted average F1-scores across folds remained consistently high. To an even greater extent, evidence of the consistency and generalization capabilities of the model emerged in the 10-fold cross-validation. The

overall scores in precision of all the folds were drastically above 1.00, and only a slight 0.98 was recorded in Fold 8. The average accuracy of the ten folds was 0.9944 and highlighted a low loss of performance. As illustrated in Figure 11.

.

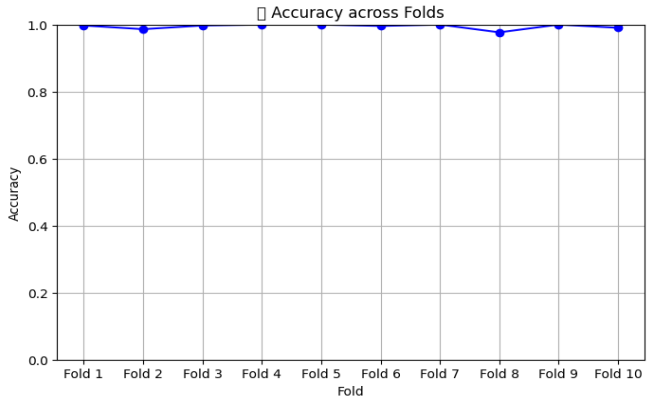

Figure 10. 10-Fold Cross-Validation.

Early stopping was observed every time in cross-validation schemes around epochs 9-12, meaning that the ViT model was able to rapidly find the optimal parameters. Such behavior indicates the effective organization of the learning process and implies a good use of representational potential without an overload of training. In addition, the loss on validations between folds was very close to zero, which means that there is little overfitting and similar training and validation distributions. Most importantly, F1-scores of all classes and folds had an upper prevalence of above 0.98, whereas the precision-recall trade-offs were especially elevated in clinically significant classes (e.g., "Tumour"). This underlines the strength of the Vision Transformer and its flexibility, particularly with respect to minor textural-based variations associated with CT imaging.

All the key metrics of the scoring were high in the case of the ViT-based architecture compared to the traditional ones, namely, the Multi-Layer Perceptron (MLP) and the Hybrid CNN-LSTM. The Area Under the Receiver Operating Characteristic Curve (AUC) of the model was also very high and was close to 1.00 in 5-fold and 10-fold cross-validation techniques. Such an almost perfect AUC indicates high discriminating capacity between the normal and pathological kidney tissues. Conversely, the MLP model took a relatively shorter time to converge and was stable in reliability, though it is not robust when it comes to sensitivity. In its turn, the high recall value of the tumour category obtained by the hybrid CNN-LSTM model is a highly desirable feature in a real-world scenario where false negative results are undesired. Its performance, however, as a whole, especially in precision and generalization, was marginally lower than the performance of the ViT. Figure 12 shows a comparative table of all the models tested with their relative performance evaluation metrics. The results indicate that the ViT model has greater accuracy, F1-score, recall, and AUC compared to the other models.

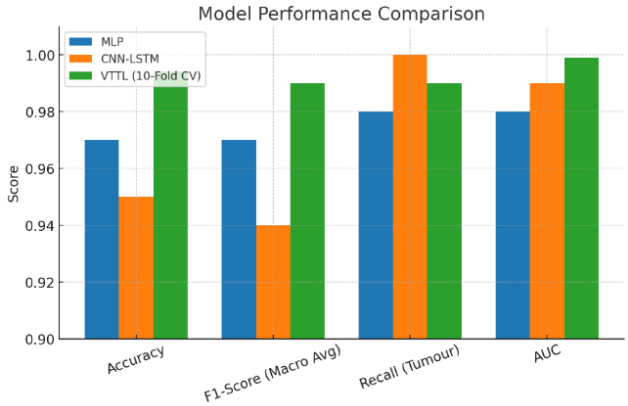

Figure 11: Comparative Performance Summary of Models

In order to support that the consistency and generalizability of our proposed Vision Transformer Transfer Learning (VTTL) model hold, we used 5-fold and 10-fold cross-validation schemes. This practice makes it so that the measured performance estimates do not depend on a certain data partition and rather the actual ability of the model to generalize over the unseen samples. The outcome of these assessments is summarized and contrasted with current state-of-the-art techniques in Table 5, which proves that our model

can display competitive accuracy. The proposed VTTL model demonstrates superior generalization compared to previous studies, effectively mitigating overfitting. This improvement is attributed to our use of 10-fold cross-validation, which ensures performance consistency across multiple data partitions, unlike other works relying solely on a fixed 80-20 split. Additionally, our evaluation employed a more challenging 69% training and 31% testing split, yet the model maintained near-perfect accuracy, confirming its robustness and generalization capability.

Table 5. Comparison Between the Proposed System and Existing Literature

Author / Year [Ref]	Method Used	Accuracy	F1-Score	Precision	Recall
Islam et al. (2022) [14]	Swin Transformers	99.30%	99.6%	99.3%	99.3%
Asif et al. (2022) [15]	VGG19 + Naïve Inception	99.25%	N/A	100%	97%
Narmada et al. (2022) [27]	CNN	99.36%	N/A	100%	96.5%
Alzu'bi et al. (2022) [28]	CNN (ResNet50, CNN-4, VGG16)	97%	N/A	100%	96.5%
Qadir & Abd (2023) [29]	DenseNet201 + Random Forest	99.44%	99.425%	99.45%	99.475%
Bhandari et al. (2023) [16]	Customized Lightweight CNN	99.52%	99.7%	99.6%	99.9%
Maqsood et al. (2024) [17]	SpinalZFNet (SpinalNet + ZFNet)	99.8%	99.7%	99.75%	99.9%
Aronson et al. (2024) [30]	MIScnn (3D U-Net for segmentation)	69%	77%	77%	77%
Kadhim & Mohammed (2025)[18]	Lightweight CNN + Hybrid Crow Swarm Optimization	100%	97.49%	97.97%	98.28%
Kadhim & Mohammed (2025) [8]	MLP (with GLCM + Gabor filters)	99.64%	99.5%	99.5%	99.5%
Maniyar et al. (2024)[31]	CNN	92%	90.5%	90.02%	91%
Proposed Model with cross-validation (CV=10)	**ViT-tiny + Transfer Learning**	**99.44 %**	**97.97%**	**98.28%**	**97.97%**

4.3 Discussion

It makes a difference that the Vision Transformer has the potential to address long-range dependency with the help of the self-attention mechanisms, which are challenging to be performed in high-resolution medical images with the help of the traditional convolutional or recurrent architectures. This can be very beneficial in CT scans, where the spatial correlations and minor anatomical differences are paramount. Furthermore, the patch-based tokenization used by the ViT, as well as position-aware embeddings, can enable meaningful abstraction of the structural features, which means that it will be able to classify a wider range of images with accurate results using even a relatively small amount of data available, a consistent issue with medical imaging categories.

The VTTL consistently outperformed traditional models such as CNN, CNN-LSTM, and MLP due to its ability to capture global contextual relationships across the entire image through self-attention mechanisms. This global modelling is particularly advantageous in medical imaging in diagnosis using medical images where very fine and spatially separated pathological patterns have to be identified. Use of pre-trained weights allowed VTTL to achieve rapid convergence and strong generalization in a data-scarce environment. The quantitative measures of accuracy, F1-scores, and AUC were exceedingly high and noticed a nearly equal level of classification accuracy in 5-fold and 10-fold cross-validation analysis with the classification results having classification metrics very close to perfect and a mere 1 instance of misclassification which indicates high sensitivity and specificity, which is vital in the domain of clinical safety. Moreover, VTTL was robustness against dataset class imbalance without additional augmentation strategies and eliminated the need for handcrafted features, learning discriminative and task-relevant representations, independently. The combination of these features enables VTTL to be seen as a better, generalizable and stable solution to high-stakes medical images classification problems like that of kidney tumour classification of CT scans.

Conclusively, it can be said that the empirical results obtained in this paper showcase the effectiveness of transformer-based architectures when processing medical images, especially on the task of kidney CT classification. Although MLP and hybrid CNN-LSTM representations continue to have their usefulness in some restrictions (relevance, hardware, or class-sensitive sensitivity), the ViT model proves to outperform all the others in terms of the most thorough and clinically relevant performance under all test resolutions. The experimental evidence reveals the immeasurable potential of Vision Transformers in the medical field to perform difficult classification. The VTTL model attained state-of-the-art performance, generalized across training data, and showed clinical suitability in terms of high specificity and sensitivity. These findings promote a wider application of transformer-based architectures in medical AI workflows, and especially in cases that have limited data boxes available and where the objective values are access to accuracy estimates in therapy.

5. Conclusion

In this study, we proposed a Vision Transformer Transfer Learning (VTTL) framework for the classification of kidney CT images into normal and tumor categories. By leveraging the ViT-Tiny architecture pre-trained on ImageNet, along with tailored preprocessing and augmentation techniques, the model was optimized for low-resource environments typical of many clinical settings. Despite working with a relatively small and slightly imbalanced dataset, the VTTL model outperformed baseline architectures such as MLP and CNN-LSTM across all major evaluation metrics, achieving 99.4% accuracy, an F1-score of 0.99, and an AUC of 0.999. Its consistent performance under both 5-fold and 10-fold cross-validation further highlights its generalizability and robustness. With low false positive and false negative rates, the model shows strong potential for clinical application, especially in high-stakes scenarios where diagnostic errors carry significant consequences. This research reinforces the viability of Vision Transformers in medical image analysis, demonstrating their strength not only in capturing complex spatial dependencies via self-attention but also in adapting effectively to domain-specific tasks through transfer learning. To further advance this work, we propose three directions for future research: (1) expand the use of ViT-based models across diverse medical imaging modalities such as MRI and X-ray, particularly in settings with limited labeled data; (2) incorporate more diverse and multi-class datasets representing a broader range of kidney pathologies to evaluate the scalability and clinical robustness of the model; and (3) integrate visual explanation tools like attention maps or Grad-CAM to enhance interpretability and support clinical trust in automated decision-making systems.

ACKNOWLEDGEMENTS

The authors are grateful to the authors of the CT Kidney Dataset (Normal, Cyst, Tumor, and Stone) available from Kaggle. Appreciation is also extended to the University of Warith Al-Anbiyaa and Warith International Cancer Institute which had to offer the computational facilities and technical support without which this work would have not been possible.

FUNDING INFORMATION

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHORS CONTRIBUATIONS

Ali Mahmoud Ali – Conceptualization, Methodology, Data Curation, Writing – Original Draft Preparation.
Mahmood Khalsan – Model Development, Experimental Design, Writing – Review & Editing.
Muntadher Idrees Ali – Implementation, Software Development, Data Preprocessing.
Mabrouka Ali Jelban – Literature Review, Data Interpretation, Writing – Review & Editing.
Teresa Abuya – Validation, Visualization, and Proofreading.

COFLICTS OF INTERESTS

The authors declare no conflicts of interest regarding the publication of this paper.

EITHICAL APPROVAL

Not applicable. This study used publicly available, anonymized data and did not involve any experiments with human participants or animals.

DATA AVAILABILITY STATEMENTS

The dataset used in this study is publicly available at: https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone.

REFERENCES

[1] D. Abdalredha Kadhim and M. Abed Mohammed, "A Comprehensive Review of Artificial Intelligence Approaches in Kidney Cancer Medical Images Diagnosis, Datasets, Challenges and Issues and Future Directions," International Journal of Mathematics, Statistics, and Computer Science, vol. 2, pp. 199–243, May 2024, doi: 10.59543/ijmscs.v2i.9747.

[2] A. M. Ali and M. A. Mohammed, "A Comprehensive Review of Artificial Intelligence Approaches in Omics Data Processing: Evaluating Progress and Challenges," International Journal of Mathematics, Statistics, and Computer Science, vol. 2, pp. 114–167, Dec. 2023, doi: 10.59543/ijmscs.v2i.8703.

[3] M. A. Mohammed, K. H. Abdulkareem, A. M. Dinar, and B. G. Zapirain, "Rise of Deep Learning Clinical Applications and Challenges in Omics Data: A Systematic Review," Feb. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/diagnostics13040664.

[4] M. A. Mohammed et al., "Novel Crow Swarm Optimization Algorithm and Selection Approach for Optimal Deep Learning COVID-19 Diagnostic Model," Comput Intell Neurosci, vol. 2022, 2022, doi: 10.1155/2022/1307944.

[5] L. Yuan et al., "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet." [Online]. Available: https://github.com/yitu-opensource/T2T-ViT

[6] Mithunraj et al., "Optimizing CNN performance in medical imaging with cross-modality pre-training: A study using MobileNetV3," Intell Based Med, vol. 12, Jan. 2025, doi: 10.1016/j.ibmed.2025.100281.

[7] S. H. Nowfal, V. Eswaramoorthy, V. P. Arivanantham, B. Marapelli, K. Swaroopa, and E. M. V. Dyana, "Diabetic Retinopathy Image Lesion Segmentation with Feature Fusion Relation Transformer Network," Journal of Machine and Computing, vol. 4, no. 4, pp. 1032–1043, Oct. 2024, doi: 10.53759/7669/jmc202404096.
[8] D. A. Kadhim and M. A. Mohammed, "Advanced Machine Learning Models for Accurate Kidney Cancer Classification Using CT Images," Mesopotamian Journal of Big Data, vol. 2025, pp. 1–25, Jan. 2025, doi: 10.58496/MJBD/2025/001.
[9] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929
[10] A. A. Mukhlif, B. Al-Khateeb, and M. A. Mohammed, "Breast cancer images Classification using a new transfer learning technique," Iraqi Journal for Computer Science and Mathematics, vol. 4, no. 1, pp. 167–180, 2023, doi: 10.52866/ijcsm.2023.01.01.0014.
[11] C.-F. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2103.14899
[12] Z. Chen et al., "Vision Transformer Adapter for Dense Predictions," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2205.08534
[13] H. Bingol, M. Yildirim, K. Yildirim, and B. Alatas, "Automatic classification of kidney CT images with relief based novel hybrid deep model," PeerJ Comput Sci, vol. 9, 2023, doi: 10.7717/peerj-cs.1717.
[14] 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, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-15634-4.
[15] S. Asif, Y. Wenhui, S. Jinhai, Q. U. Ain, Y. Yueyang, and H. Jin, "Modeling a Fine-Tuned Deep Convolutional Neural Network for Diagnosis of Kidney Diseases from CT Images," in Proceedings - 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 2571–2576. doi: 10.1109/BIBM55620.2022.9995615.
[16] M. Bhandari, P. Yogarajah, M. S. Kavitha, and J. Condell, "Exploring the Capabilities of a Lightweight CNN Model in Accurately Identifying Renal Abnormalities: Cysts, Stones, and Tumors, Using LIME and SHAP," Applied Sciences (Switzerland), vol. 13, no. 5, Mar. 2023, doi: 10.3390/app13053125.
[17] F. Maqsood et al., "Artificial Intelligence-Based Classification of CT Images Using a Hybrid SpinalZFNet," Interdiscip Sci, vol. 16, no. 4, pp. 907–925, Dec. 2024, doi: 10.1007/s12539-024-00649-4.
[18] D. A. Kadhim and M. A. Mohammed, "Efficient Kidney Cancer Classification from CT Images Using a Lightweight Convolutional Neural Network Optimized with an Enhanced Crow Swarm Optimization Algorithm," Journal of Soft Computing and Data Mining, vol. 6, no. 1, pp. 200–229, Jun. 2025, doi: 10.30880/jscdm.2025.06.01.014.
[19] A. M. Ali and M. A. Mohammed, "Enhanced Cancer Subclassification Using Multi-Omics Clustering and Quantum Cat Swarm Optimization," Iraqi Journal for Computer Science and Mathematics, vol. 5, no. 3, pp. 552–582, 2024, doi: 10.52866/ijcsm.2024.05.03.035.
[20] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2304.07288
[21] I. Loshchilov and F. Hutter, "DECOUPLED WEIGHT DECAY REGULARIZATION." [Online]. Available: https://github.com/loshchil/AdamW-and-SGDW
[22] F. Abedi et al., "Chimp Optimization Algorithm Based Feature Selection with Machine Learning for Medical Data Classification," Computer Systems Science and Engineering, vol. 47, no. 3, pp. 2791–2814, 2023, doi: 10.32604/csse.2023.038762.
[23] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," Nov. 01, 2018, Springer New York LLC. doi: 10.1007/s10916-018-1088-1.
[24] M. Khalsan, M. Mu, E. S. Al-Shamery, S. Ajit, L. R. Machado, and M. Opoku Agyeman, "A Novel Fuzzy Classifier Model for Cancer Classification Using Gene Expression Data," IEEE Access, vol. 11, pp. 115161–115178, 2023, doi: 10.1109/ACCESS.2023.3325381.
[25] M. Khalsan, M. Mu, E. S. Al-Shamery, L. MacHado, M. O. Agyeman, and S. Ajit, "Intersection Three Feature Selection and Machine Learning Approaches for Cancer Classification," in Proceedings of 2023 International Conference on System Science and Engineering, ICSSE 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 427–433. doi: 10.1109/ICSSE58758.2023.10227163.
[26] M. Khalsan, M. Mu, E. S. Al-Shamery, L. MacHado, M. O. Agyeman, and S. Ajit, "Enhancing Cancer Classification Through the Development of a Fuzzy Gene Selection-Wrapper Plus Method," in 5th IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 39–44. doi: 10.1109/IICAIET59451.2023.10291820.
[27] N. Narmada, V. Shekhar, and T. Singh, "Classification of Kidney Ailments using CNN in CT Images," in 2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICCCNT54827.2022.9984256.
[28] D. Alzu'Bi et al., "Kidney Tumor Detection and Classification Based on Deep Learning Approaches: A New Dataset in CT Scans," J Healthc Eng, vol. 2022, 2022, doi: 10.1155/2022/3861161.
[29] A. M. Qadir and D. F. Abd, "Kidney Diseases Classification using Hybrid Transfer-Learning DenseNet201-Based and Random Forest Classifier," Kurdistan Journal of Applied Research, pp. 131–144, Jan. 2023, doi: 10.24017/science.2022.2.11.
[30] L. Aronson, R. Ngnitewe Massa'a, Md, S. Jamal, S. Gardezi, and A. L. Wentland, "Automatic Segmentation of the Kidneys and Cystic Renal Lesions on Non-Contrast CT Using a Convolutional Neural Network."
[31] V. Kumashi, M. Kudari, and S. N. Anupama, "Kidney Disease Classification," in 15th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2024, Grenze Scientific Society, 2024, pp. 180–187. doi: 10.48175/ijarsct-19981.