


## Article

# Evaluation of Vision Transformers for Multi-Organ Tumor Classification Using MRI and CT Imaging

Óscar A. Martín  and Javier Sánchez <sup>\*,†</sup> 

Centro de Tecnologías de la Imagen (CTIM), Instituto Universitario de Cibernética, Empresas y Sociedad (IUCES), 35017 Las Palmas de Gran Canaria, Spain; oscar.martin104@alu.ulpgc.es

\* Correspondence: jsanchez@ulpgc.es; Tel.: +34-928-458710

† Current address: Department of Computer Science, University of Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain.

## Abstract

Using neural networks has become the standard technique for medical diagnostics, especially in cancer detection and classification. This work evaluates the performance of Vision Transformer architectures, including Swin Transformer and MaxViT, for several datasets of magnetic resonance imaging (MRI) and computed tomography (CT) scans. We used three training sets of images with brain, lung, and kidney tumors. Each dataset included different classification labels, from brain gliomas and meningiomas to benign and malignant lung conditions and kidney anomalies such as cysts and cancers. This work aims to analyze the behavior of the neural networks in each dataset and the benefits of combining different image modalities and tumor classes. We designed several experiments by fine-tuning the models on combined and individual datasets. The results revealed that the Swin Transformer achieved the highest accuracy, with an average of 99.0% on single datasets and reaching 99.43% on the combined dataset. This research highlights the adaptability of Transformer-based models to various human organs and image modalities. The main contribution lies in evaluating multiple ViT architectures across multi-organ tumor datasets, demonstrating their generalization to multi-organ classification. Integrating these models across diverse datasets could mark a significant advance in precision medicine, paving the way for more efficient healthcare solutions.

**Keywords:** brain tumor; lung tumor; kidney tumor; neural networks; Vision Transformer; Swin Transformer; MaxViT



Academic Editors: D. J. Lee and Dong Zhang

Received: 18 June 2025

Revised: 13 July 2025

Accepted: 23 July 2025

Published: 25 July 2025

**Citation:** Martín, Ó.A.; Sánchez, J. Evaluation of Vision Transformers for Multi-Organ Tumor Classification Using MRI and CT Imaging. *Electronics* **2025**, *14*, 2976. <https://doi.org/10.3390/electronics14152976>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Cancer is a term that encompasses a broad group of diseases. The statistics show that it is the leading cause of death affecting society, resulting in millions of deaths each year. In 2020, there were an estimated 19.3 million new cases of cancer and 9.9 million cancer-related deaths [1].

Over the last decade, medical technology has undergone rapid improvement. Neural networks play a pivotal role in early and accurate cancer diagnoses, providing healthcare professionals with automated tools to detect diseases efficiently and minimizing human error and workload. They have yielded significant improvements in disease research and diagnosis. In particular, the use of deep neural networks for detection and classification has been widely applied in several types of tumors, such as brain [2], lung [3], and kidney tumors [4]. Nevertheless, only a few studies have assessed the performance of these neural

networks in detecting and classifying different types of diseases using a single network and the benefits this can bring.

This work specifically investigates the use of recent models for image-based diagnosis, particularly using MRI and CT scans for detecting tumors. We explore architectures based on Vision Transformer, which excels at processing images and identifying anomalous patterns.

In the last two years, these architectures have evolved into more efficient and domain-specific architectures. Models like BiFormer [5] and EfficientViT [6] introduced sparse and cascaded attention mechanisms to reduce computation while maintaining accuracy. Lightweight designs such as SeaFormer++ [7] and MetaSwin [8] focused on mobile and medical segmentation tasks by incorporating squeeze-enhanced or pooled token mechanisms. In medical imaging, architectures like SwinUNETR [9] and MedFormer [10] emerged, combining Transformer attention with encoder–decoder structures to handle multimodal and 3D data.

Our work aimed to classify different types of tumors related to various human organs and distinct image formats into a single model. This allowed us to assess the capabilities of these architectures and their generalizability to multiple diseases and image formats. The objective was to integrate different data into a single Transformer model for detecting cancerous and benign diseases. The Transformer architecture has been successfully applied in computer vision, superseding convolutional neural networks (CNNs) in many tasks. The differences between CNNs and Transformers are notable from an architectural perspective, and the choice between them depends on the specific task. We analyzed several architectures, such as Vision Transformer (ViT) [11], Swin Transformer [12,13], and MaxVit [14].

For the training of these networks, we selected three datasets: one of brain images [15], containing three types of tumors (gliomas, meningiomas, and pituitary tumors); another of lung images [16,17], containing CT scans of lung cancer and healthy lungs, with three distinct classes (normal, benign, and malignant); and a dataset of kidney images [18], containing four different categories (non-tumor, stone, cyst, and cancer). The first dataset included 15,000 MRIs, with 5000 images in each subclass; the second dataset contained 1190 CT scans of lung cancer and healthy lungs; and the third one had 12,446 CT scans distributed in four classes. Each dataset was based on 2D slices of a specific modality and organ, thus limiting intra-organ multimodal generalization.

The experimental results showed the performance of each neural network using transfer learning. In many cases, the models' accuracy was above 99%, with the Swin Transformer providing outstanding results, followed by MaxViT. These promising results demonstrate that Vision Transformers can play a key role in medical image analysis.

The main contribution of this work lies in its systematic evaluation of multiple ViT variants across heterogeneous medical imaging datasets and different tumor types. Unlike prior studies that typically focus on a single organ or imaging modality, our approach explores the potential of ViTs to generalize within a unified framework, highlighting their adaptability in multi-organ classification tasks. By comparing model performance on individual datasets versus a combined dataset, we offer novel insights into how the models leverage shared patterns and modality-specific features through their self-attention mechanisms.

Section 2 summarizes state-of-the-art works on brain, lung, and kidney tumor detection methods. Section 3 details the datasets used in this work and the neural networks employed in the classification task. The results in Section 4 assess the performance of the neural networks for each dataset and the accuracy by combining the three datasets in a single set. Finally, the conclusions in Section 5 summarize the main ideas and contributions of this work and propose some ideas for future work.

## 2. Related Work

Artificial intelligence is making significant advances in tumor detection, with deep learning models, like EfficientNet [19], achieving high accuracy rates using MRI images. Other standard models, such as VGG and MobileNet, have also shown outstanding results [20]. Furthermore, the use of Transformer models, such as Vision and Swin Transformers, is being explored to overcome the limitations of CNNs in medical image classification and segmentation. These models can capture global relationships in images, which is crucial for accuracy in medical diagnosis.

A recent work [21] highlights the importance of segmenting medical images based on Vision Transformers instead of convolutional networks. The latter are effective in capturing local correlations, although they are limited in capturing global relationships.

### 2.1. Brain Tumors

Brain tumor classification has received significant attention in the last few years, particularly through the analysis of MR images. Various studies have been conducted to enhance the performance of brain tumor classification using different methodologies. The work presented in [22] focused on brain tumor classification based on T1-weighted contrast-enhanced MRI. They proposed a dataset that has been used in many subsequent works.

Traditional techniques typically classify images based on two main steps: feature extraction and classification. The features proposed in [23] were based on PCA and GIST techniques, and the classification was carried out through a regularized extreme learning machine. The last step can also involve the use of neural networks, like in [24], which relies on the 2D Discrete Wavelet Transform (DWT) and 2D Gabor filters to extract statistical features from MRI. The approach presented in [25] tackled the problem of MRI segmentation and classification using genetic algorithms. The authors of [26] explored brain tumor classification using Capsule Networks. These types of networks have several benefits over CNNs, as they are robust to rotation and affine transformations and require less training data.

The system proposed in [27] consisted of three main steps: tumor segmentation, data augmentation, and deep feature extraction and classification. It relied on extensive data augmentation techniques and the fine-tuning of a VGG-19 network. Many works [28] have extensively utilized pure CNN models for brain tumor classification, obtaining high accuracy in various datasets.

Several works [29] have conducted a performance analysis of transfer learning with VGG-16, ResNet-50, or Inception-v3 for automatic prediction of tumor cells in the brain. A recent work [20] reported a detailed performance assessment analysis of several convolutional architectures from different perspectives, drawing important conclusions about these techniques. Several models obtained high accuracy, even for networks with a relatively low number of parameters, such as EfficientNet [19].

Nevertheless, convolutional networks have several limitations, such as their local receptive field and difficulty in capturing global spatial dependencies, which can hinder performance with CT and MRI data.

### 2.2. Lung Tumors

Lung cancer remains one of the leading causes of cancer-related deaths worldwide. Early and accurate detection through imaging techniques, such as CT scans, is crucial for improving patient outcomes. Classification of lung tumors using machine learning techniques has been an active area of research.

Early work in lung tumor classification primarily utilized traditional machine learning algorithms such as Support Vector Machines (SVMs), k-Nearest Neighbors (k-NNs), and

Decision Trees. For instance, the authors of [30] employed an SVM for classifying lung nodules in CT images, achieving notable accuracy by optimizing hyperparameters and using feature selection techniques to reduce dimensionality. Similarly, the method in [31] used k-NN combined with an SVM, demonstrating improved classification performance on small datasets.

With the advent of deep learning, CNNs have become the preferred choice. For example, the work presented in [32] developed a model that outperformed traditional methods by automatically learning hierarchical features from CT images. Their approach significantly reduced the need for manual feature extraction, leading to higher accuracy and robustness. On the other hand, the work in [33] further advanced this field by introducing a multi-view network that integrated information from multiple CT slices. This method demonstrated superior performance in distinguishing between benign and malignant nodules compared to single-view models.

Ensemble methods have also been explored to enhance classification performance. For instance, an ensemble of CNNs and several traditional machine-learning techniques [34] achieved improved accuracy and robustness by mitigating the weaknesses of individual models.

Transfer learning, which involves fine-tuning pre-trained models on specific datasets, has gained popularity due to its effectiveness with reduced, labeled data. The model presented in [35] employed transfer learning using pre-trained models, achieving high accuracy with reduced training times. Their approach demonstrated that leveraging pre-trained models can enhance performance when dealing with small datasets.

Recent studies have shown that transfer learning improves classification with limited data using ViT architectures [36,37]. Our work also confirms this trend, as shown in Section 4.

### 2.3. Kidney Tumors

The application of machine learning to the classification of kidney tumors has also seen significant advancements in recent years. Various methods and technologies have been developed to aid in the early detection and accurate segmentation of kidney tumors. For instance, the work in [38] utilized a computer-aided detection system for kidney tumors on abdominal CT scans, employing a gray-level threshold method for segmentation and texture analysis for tumor detection. Similarly, the work in [39] presented a semi-automatic kidney tumor detection and segmentation method using atlas-based segmentation.

The classification of kidney tumors into subtypes such as benign, malignant, and histological categories is crucial for guiding treatment strategies. Early approaches often utilized traditional techniques, such as SVMs and Random Forests, relying on manually extracted features such as texture, shape, and intensity descriptors. For example, the method explained in [40] explored handcrafted features with SVMs for detecting renal masses, achieving promising results in sensitivity and specificity. Similarly, the use of radiomics features in combination with machine learning techniques [41] helped improve the classification of renal tumors.

Advancements in deep learning techniques have also played a crucial role. For example, a CNN-based U-Net architecture with an attention mechanism [42] and a 3D U-Net-based deep convolutional neural network [43] achieved high accuracy in tumor segmentation. Furthermore, recent deep learning approaches, utilizing models such as 2D-CNN, ResNet, and ResNeXt [44,45], have improved the detection accuracy. The work in [46] explored self-supervised learning for kidney tumor classification on CT images. In addition, the work in [47] developed an LED-based near-infrared sensor for human kidney tumor diagnostics to provide a simplified alternative to conventional NIR spectroscopic methods.

### 3. Materials and Methods

This section details the datasets and methods employed in our study. Our work is based on two types of medical imaging: computed tomography, which uses X-rays to obtain detailed images, useful in oncology and other medical areas, and magnetic resonance imaging, which produces detailed, three-dimensional images using magnetic fields and electromagnetic waves, thereby differentiating tissues, although it is more expensive than tomography.

#### 3.1. Datasets

We selected three datasets related to brain, lung, and kidney tumors. Next, we will explain each one in detail.

##### 3.1.1. Brain Tumor Dataset

Brain tumors originate from the abnormal growth of cancerous cells within the brain. They can be benign or malignant and may affect various areas, leading to symptoms depending on their location, size, and severity. Research indicates that patient survival largely depends on complete tumor removal, making early detection crucial for timely and appropriate treatment before the tumor grows.

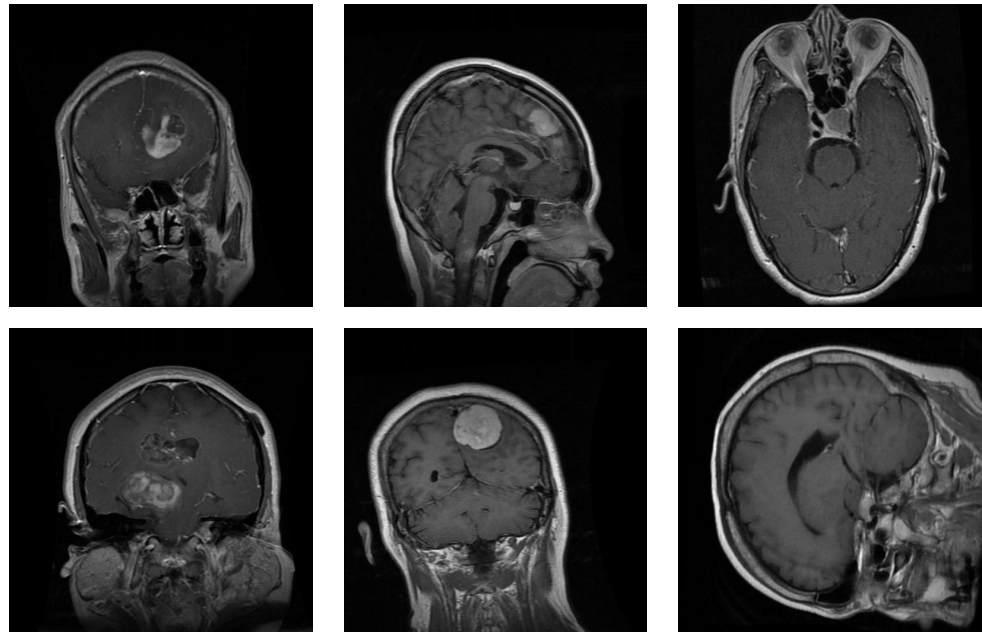
In this work, we tested the classification of the following brain tumors:

- Glioma: This common tumor type has several variants. Some are considered non-dangerous, but others are quite aggressive, spreading to healthy brain tissues and causing pressure on the brain or spinal cord. Gliomas are cancerous cells resembling glial cells surrounding brain and spinal cord nerve cells. They grow within the white matter and dangerously spread through healthy brain tissues.
- Meningioma: This is the most common primary brain tumor, classified into three grades based on their characteristics, with the high grade being rare.
  - Grade I: The most common is a low-grade tumor with slow growth.
  - Grade II: An intermediate grade, known as atypical meningioma, characterized by a higher likelihood of recurrence after removal.
  - Grade III: The highest grade, less common, characterized as malignant and called anaplastic meningioma. It differs in appearance from normal cells and expands rapidly. Atypical and anaplastic meningiomas can spread throughout the brain and other body organs. A mass on the outer tissue layer of the brain can identify these meningiomas. With this data, the importance of including this cancer type in the dataset for early detection training is evident.
- Pituitary Tumor: Generally, this type of brain tumor is non-invasive. It is likely to be benign with low growth potential. However, there is a risk of growth and spread to other parts, such as the optic nerve or carotid arteries, presenting more aggressive characteristics. These often go unnoticed and are detected during tests for reasons unrelated to cancer.

The brain tumor dataset used in this study is part of the Multi-Cancer Imaging Dataset [15]. This dataset includes MRI images categorized into the three tumor types. The brain cancer images were obtained from other sources [22,48].

Figure 1 depicts several examples of brain tumors in the dataset. This dataset forms part of a larger collection that encompasses eight distinct cancer types from different organs: acute lymphoblastic leukemia, brain, breast, cervical, kidney, lung, colon, and oral cancer. The images are in JPEG format with dimensions of  $512 \times 512$  pixels. We selected brain cancer, which contained the three previous tumor subclasses. Each subclass included the same number of images (5000 MRIs).





**Figure 1.** Samples from the brain dataset: the first column shows two examples of glioma tumors; the middle column shows meningioma tumors; and the last column depicts images of pituitary tumors.

### 3.1.2. Lung Tumor Dataset

Lung tumors can arise from various cells in the bronchi, bronchioles, and pulmonary alveoli. Lung cancer is a harmful disease that poses a lethal threat to patients' health. This causes more deaths than breast, colon, and prostate cancers combined. However, the prognosis depends on the type of cancer, its spread, and its size. Lung tumors often do not cause symptoms and are usually detected during imaging tests conducted for reasons unrelated to lung cancer.

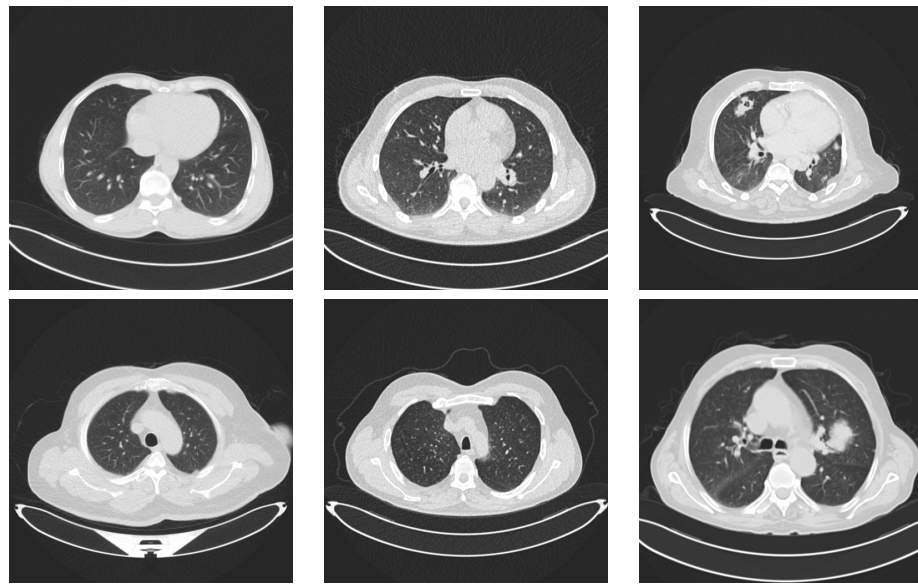
Since these are not frequently detected in clinical histories, developing a predictive model that can provide information on identifying a potential lung tumor could be beneficial. An early diagnosis of lung cancer, when the tumor has not spread to other tissues, can lead to a favorable prognosis in up to 90% of patients.

We selected a dataset from Kaggle [16,17] that includes CT scans of lung cancer and healthy lungs. The dataset has three classes: normal, benign, and malignant. Figure 2 depicts two examples of each class in the dataset.

The images were in DICOM format and were preprocessed to anonymize patient identifiers. The dataset comprised 110 cases, varying in patient age, sex, residential area, and socioeconomic status, among other variables. It contained 1190 images, although we detected and removed duplicate files, with a total of 1054 unique images. These were distributed as shown in Table 1.

**Table 1.** Distribution of samples in the lung dataset. For each type of class, this table shows the number of cases in the second column and the number of images in the third column.

Type	Images
Healthy Lung	405
Benign Tumor Lung	102
Cancerous Tumor Lung	547



**Figure 2.** Examples of lung images. Images in the first column represent lungs with no tumors. The images in the middle column represent benign tumors and the last column contains two images with malignant tumors.

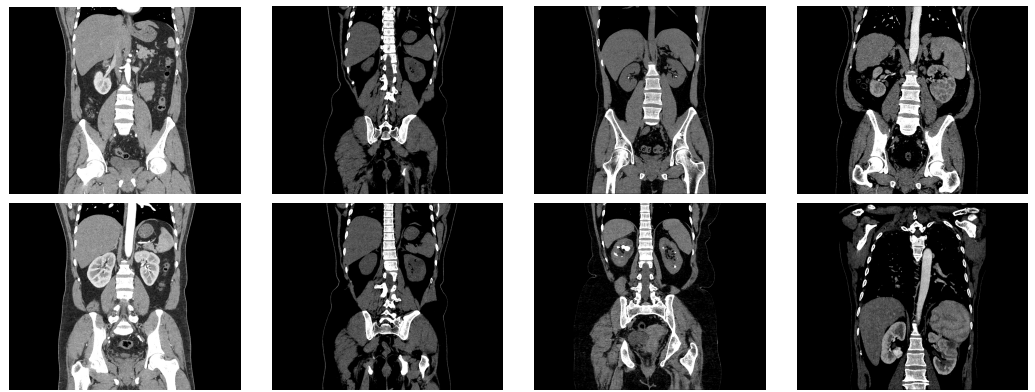
### 3.1.3. Kidney Dataset

In the United States, approximately 430,000 individuals were diagnosed with kidney cancer in the year 2020, and the number of cases has been increasing over the decades. It ranks as the sixth most common type of cancer among men and ninth among women. The relative survival rate is 77% after five years, depending on various factors such as treatment, cancer spread, and patient age. This statistic refers to the patient's life expectancy after the diagnosis of the disease or the start of treatment.

We selected a dataset of kidney images from Kaggle [18], classified into four different categories:

- **Non-tumor:** This category represents images of healthy kidneys.
- **Stone:** This category represents images where an abnormal solid material is located in the kidney. This type of material occurs in the kidneys when levels of certain minerals are high. It usually presents with sharp pain in the lower back or groin, or even blood in the urine. Diagnosis depends on physical health, laboratory tests, and the size of the stone visible in the images. Generally, treatment involves removing or breaking the stone into pieces.
- **Cyst:** This class represents images where small aqueous or liquid structures are identified and are not normally harmful to health. It is necessary to distinguish between a simple cyst, which has no risk of turning into kidney cancer, and a complex cyst, which does have a risk of becoming kidney cancer, even though this risk is low and would require careful diagnosis by a urological surgeon.
- **Cancer:** This represents images with a cancerous mass. Renal cell carcinoma is the most common type of tumor, accounting for up to 90% of kidney cancer. Many cases are detected in the early stages of the tumor. Generally, it does not present symptoms, and surgery is the standard treatment to remove it, achieving a cure rate of over 70%.

Figure 3 shows several samples from the dataset, with two scans of each dataset class.



**Figure 3.** Examples from the kidney dataset. The two images in the first column represent CT images of normal kidneys, the two images in the second column depict kidneys with cysts, the images in the third column represent CT images of kidneys with stones, and the images in the last column depict kidneys with cancer.

Images of the dataset were selected from axial and coronal anatomical planes with and without contrast for the entire abdomen. The data format was DICOM, and the images were converted to JPG format. The dataset originally contained 12,446 samples and, after removing duplicate samples, this was reduced to 11,929 images. Table 2 shows the distribution into the four classes.

**Table 2.** Distribution of samples in the classes of the kidney dataset.

Class	Number of Images
Normal	5002
Cyst	3284
Stone	1360
Cancer	2283

Samples from the datasets were preprocessed by resizing all images to  $224 \times 224$  pixels, converting DICOM files (for lung and kidney datasets) to JPEG format, and ensuring 3-channel RGB format for compatibility with Vision Transformer. Note that using 2D RGB slices prevents models from leveraging 3D data or time-series information, limiting volumetric understanding.

### 3.2. Neural Networks

In this work, we compared the performance of three Vision Transformers, i.e., Vision Transformer (ViT) [11], Swin Transformer [12], and MaxViT [14], which have provided competitive results in various computer vision tasks [21,49]. We selected Swin for its hierarchical, efficient design and MaxViT for its state-of-the-art hybrid attention capabilities.

#### 3.2.1. Vision Transformer

The Vision Transformer [11] is a model designed for image classification tasks, employing a Transformer-like architecture adapted for visual data. ViT begins by dividing an input image into fixed-size patches. This resembles how the original Transformer model breaks down text into tokens. Each image patch is flattened and transformed into a vector through linear embedding. This process converts the 2D patch into a 1D vector that the Transformer can process. Since the Transformer architecture does not inherently process sequential data, positional embeddings are added to the patch embeddings to retain the order of the patches, which is crucial for understanding the spatial relationships within the image. The sequence of patch embeddings, now with positional information, is fed into a



standard Transformer encoder. The encoder consists of layers of multi-head self-attention and feed-forward neural networks.

For classification tasks, an extra learnable token, often referred to as the classification token (CLS), is added to the sequence. The state of this token at the output of the Transformer encoder captures the global information about the image, which is used for the final classification.

ViT has demonstrated remarkable abilities, achieving comparable or better performance than traditional CNNs on various computer vision tasks [49]. It represents a significant shift in how models process visual information, leveraging the power of self-attention mechanisms to capture global dependencies within the image.

Table 3 shows different configurations of the ViT model. Due to computational limitations, only the base model ViT-Base was tested, which has two alternatives, ViT-b-16 and ViT-b-32, for input patches of  $16 \times 16$  pixels and  $32 \times 32$  pixels, respectively. Smaller patches have a higher computational cost since the size of the Transformer sequence is inversely proportional to the square of the patch size. Therefore, we selected  $32 \times 32$  patches to reduce computational complexity, as smaller patches (e.g.,  $16 \times 16$ ) increase FLOPs and GPU memory usage.

**Table 3.** Details of different variants of the ViT model.

Model	Layers	Hidden Layers	MLP Size	Heads	Parameters (in Millions)
ViT-Base	12	768	3072	12	86
ViT-Large	24	1024	4096	16	307
ViT-Huge	32	1280	5120	16	632

### 3.2.2. Swin Transformer

The Swin Transformer [12,13] is a type of Vision Transformer adapted for computer vision tasks, including image classification, object detection, and semantic segmentation. Unlike other Vision Transformers that compute self-attention globally across the entire image, the Swin Transformer divides the image into smaller windows and applies self-attention within these local windows. This innovation significantly reduces computational complexity while maintaining its effectiveness. This architecture builds upon the ViT framework and introduces a hierarchical approach for processing images. It starts with smaller patches in the initial layers and progressively merges them into larger patches in deeper layers. This enables detailed image processing, capturing both local and global contexts.

It achieves linear complexity depending on the input image size. This is in contrast to other architectures that have quadratic complexity due to global self-attention. The efficiency of the Swin Transformer makes it a general-purpose backbone for various vision tasks. Due to its hierarchical feature maps and efficient computation, it is a versatile backbone for image classification tasks. It combines the power of Transformers with localized self-attention, making it an effective choice for processing medical images and detecting anomalies. Its benefits lie in improved efficiency, hierarchical feature extraction, and suitability for various computer vision tasks.

### 3.2.3. MaxViT Transformer

The MaxViT Transformer [14] is a hybrid architecture that combines the strengths of CNNs and Vision Transformers to create a powerful image classification model. It integrates the inductive biases of CNNs with the global receptive field of ViTs. This combination allows it to achieve high performance across various parameters and metrics. The architecture introduces a multi-axis attention mechanism, incorporating blocked local and dilated global attention. This design enables the model to capture local and global

spatial interactions at linear complexity, regardless of input resolution. Similar to traditional CNNs, MaxViT follows a hierarchical design. It builds upon a new type of basic building block that unifies MBConv [50] blocks and grid attention layers, allowing the model to reach a global context throughout the entire network. This model scales well with large datasets and maintains linear complexity with respect to the grid attention used, making it suitable for high-resolution inputs.

In the experiments, we tested different configurations of these architectures. Regarding the Swin Transformer and MaxViT, we tested their tiny variants. When considering performance, the focus was on the relationship between training time and accuracy. Models that required extensive computation, both in terms of time and GPU memory load, were discarded.

Table 4 shows the variants used in the experiments. For each one, we summarize the number of parameters, the floating point operations per second (FLOPS), and their size in megabytes. In this case, we analyzed models based not only on accuracy but also on FLOPs, or inference time, to study practical deployment in clinical settings.

**Table 4.** Complexity of the models.

Models	Number of Parameters	FLOPS (Millions)	Size (Megabytes)
ViT-b-32	88,185,064	456.96	344.59
Swin-t	28,288,354	123.43	141.22
MaxViT-t	30,919,624	5600.0	118.80

### 3.3. Experimental Setup

We used the Adam optimizer for learning the parameters of the neural networks, which is frequently used in image classification tasks, with parameters  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . Since the problem was a multiclass classification problem, we chose the Categorical Cross-Entropy loss function, given by the following expression:

$$\mathcal{L}(\hat{y}_i, y_i) = - \sum_{i=1}^N y_i \cdot \log \hat{y}_i, \quad (1)$$

where  $y_i$  is the true value of sample  $i$ ,  $\hat{y}_i$  is the prediction given by the neural network, and  $N$  is the number of classes. This function yielded a value between 0 and 1, representing a probability for each label or class trained in the model. When the error was high, cross-entropy significantly penalized values that deviated from the expected predictions.

Each dataset was split into a training, validation, and test set. The size of the training set was 80% of the samples and was used to learn the parameters of the networks. The size of the validation set was 10% of the samples and was used to validate the training process and find appropriate hyperparameters. The test comprised 10% of the samples and was used to evaluate the performance of the models.

The number of epochs in the first fine-tuning phase was 50, with batch size 16, image size 224, learning rate  $10^{-3}$ , and dropout rate 30%. In the second phase, the number of epochs was reduced to 15 and the learning rate was set to  $10^{-4}$ .

The metrics used to compare the models were accuracy, precision, recall, F1-score, and AUC-ROC. These metrics relied on the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values of the classifications. *Accuracy* was an intuitive performance measure and was calculated as a ratio of correctly predicted observations to the total observations, given by

$$accuracy = \frac{TP + TN}{Total}. \quad (2)$$

*Precision* measured the number of positive predictions that were correctly estimated and was calculated as

$$precision = \frac{TP}{TP + FP}. \quad (3)$$

*Recall*, or sensitivity, measured the number of actual positives that were correctly identified. High recall meant few false negatives and was calculated as

$$recall = \frac{TP}{TP + FN}. \quad (4)$$

F1-score was the harmonic mean between the precision and recall. It balanced the two metrics and was especially useful for imbalanced classes. It was calculated as

$$F1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \quad (5)$$

For dealing with imbalanced classes, we also tested weighting the loss function and oversampling minority classes, although they did not offer noticeable gains in our results.

AUC-ROC was the Area Under the Receiver Operating Characteristic Curve. It measured the area under the ROC curve, which was the relation between the True Positive Rate and the False Positive Rate. In the experiments, we also used confusion matrices to represent the misclassifications between the predicted and actual values of each class.

The source code was implemented in Python 3.10 and the deep learning models were implemented using the PyTorch 2.1 framework. The training was carried out on an NVIDIA GeForce RTX 2060 SUPER (NVIDIA Corporation, Santa Clara, CA, USA). The hyperparameters were tuned depending on the neural network architecture, dataset complexity, and available computational resources. We explored the following range of values:

- Mini-batch Size: From 4 to 32 images per batch.
- Image Size: Images were scaled to a fixed dimension of  $224 \times 224$  pixels.
- Image Channels: We used three channels (RGB).
- Learning Rate: The learning rate was set between  $10^{-3}$  and  $10^{-5}$ , depending on the learning technique and the trained model.
- Dropout: In the final layer, neuron connections were deactivated at a rate of 30% and 40%. A dropout value within this range was always active.
- Epochs: The number of epochs in the training process ranged from 5 to 200, depending on the dataset and mini-batch size applied.

To mitigate overfitting during training, we implemented several strategies: dropout (30–40%) on fully connected layers, early stopping based on validation loss, transfer learning with frozen layers, and reduced learning rates during fine-tuning phases.

## 4. Results

### 4.1. Results with Transfer Learning

In this section, we analyze the global performance of the models for each dataset alone and in combination. We trained the models using a two-step transfer learning approach: in the first step, we optimized the head of the models learning the new classification problem while preserving the backbone of the networks; in the second step, once the models were adapted to the new problem, we optimized all the parameters with a small learning rate to further improve the accuracy.

Table 5 shows the results at the end of the first step. This table compares the accuracy of the three neural networks across the individual and combined datasets. The results indicate that ViT-b-32 consistently outperformed the other two models across all scenarios. It achieved the highest accuracy on the brain (93.27%), kidney (97.15%), and lung (93.33%)

datasets individually. Moreover, its performance remained strong even when trained on the combined dataset, achieving 95.43% accuracy. This suggests that ViT-b-32 is particularly robust and generalizes well across different types of tumor data.

**Table 5.** Accuracy of the models after the first transfer learning step. This table shows the accuracy of the models for each dataset. The last row depicts the results of each model by combining the three datasets. Bold letters highlight the best results in each row.

Dataset	ViT-b-32	Swin-t	MaxViT-t
Brain	<b>93.27%</b>	93.07%	91.47%
Kidney	<b>97.15%</b>	95.55%	88.51%
Lung	<b>93.33%</b>	90.48%	87.62%
All datasets	<b>95.43%</b>	93.82%	83.60%

The Swin-t model performed similarly to ViT-b-32 in most cases. On the brain dataset, its accuracy (93.07%) was nearly equal to that of ViT-b-32. However, there was a slightly larger performance gap in the kidney (95.55%) and lung (90.48%) datasets. When evaluated on the combined dataset, Swin-t achieved an accuracy of 93.82%, which, although lower than ViT-b-32, still indicates relatively strong generalization capabilities.

In contrast, the MaxViT-t model showed significantly lower performance across all datasets. It presented particularly weak results on the kidney tumor dataset (88.51%). Its performance dropped when trained on the combined dataset, achieving an accuracy of 83.60%. These results suggest that MaxViT-t may struggle to capture relevant features effectively in this domain or may require more task-specific fine-tuning to achieve competitive results.

The results of the second step showed a notable improvement across all models and datasets when compared to the initial training results, indicating the need for further optimization; see Table 6. This comparison provides key insights into how each model benefited from additional fine-tuning.

**Table 6.** Accuracy of the models using transfer learning. This table shows the accuracy of ViT-b-32, Swin-t, and MaxViT-t for each dataset. The last row depicts the results of each model by combining the three datasets. Bold letters highlight the best results in each row.

Dataset	ViT-b-32	Swin-t	MaxViT-t
Brain	97.07%	<b>99.53%</b>	99.27%
Kidney	97.73%	<b>99.75%</b>	99.75%
Lung	95.24%	<b>97.14%</b>	94.29%
All datasets	97.03%	<b>99.43%</b>	98.68%

The accuracy of ViT-b-32 increased on the brain dataset from 93.27% to 97.07% and on the kidney dataset from 97.15% to 97.73%. On the lung dataset, the accuracy improved from 93.33% to 95.24%, and on the combined dataset, it rose from 95.43% to 97.03%. Its relative ranking dropped as the other two models showed even greater gains.

The Swin-t model showed the most significant improvements. On the brain dataset, its accuracy increased from 93.07% to 99.53%, and on the kidney tumor dataset, it increased from 95.55% to 99.75%. For lung cancer, its accuracy increased from 90.48% to 97.14%, and on the combined dataset, it increased from 93.82% to 99.43%. These substantial gains positioned Swin-t as the top performer across all categories after fine-tuning. The model overtook ViT-b-32 with significant margins, demonstrating high adaptability to the tumor datasets.

MaxViT-t showed a noticeable improvement after the second fine-tuning step. It showed greater gains after full fine-tuning, likely due to its deeper, high-capacity architecture and global–local attention. Its brain tumor classification accuracy jumped from 91.47% to 99.27% and its kidney dataset performance leapt from 88.51% to 99.75%, matching Swin-t. Lung tumor classification improved from 87.62% to 94.29% and the combined dataset result increased significantly from 83.60% to 98.68%. This transformation suggests that MaxViT-t may require more extensive training to fully unlock its potential, possibly due to its complex architecture or greater capacity.

The performance of the three neural networks on the brain dataset revealed important distinctions not only in terms of accuracy but also when considering training and inference times, parameter count, and computational cost (FLOPS), as depicted in Table 7. The training times were calculated as the average seconds per epoch in the two fine-tuning phases. The inference times were calculated as the average milliseconds per image of the training set.

**Table 7.** Performance assessment for the brain dataset: the first column shows the model used in each experiment; the second column depicts the accuracy of the model for the test set; the third column shows the training time in seconds per epoch; the fourth column shows the inference time in milliseconds per image; the fifth column shows the number (#) of parameters of each network; and the last column shows the number of FLOPS. The last two columns are copied from Table 4.

Model	Accuracy	Training Time	Inference Time	#Parameters	FLOPS (Mill)
ViT-b-32	97.07%	87 s/epoch	5.8 ms/image	88,185,064	456.96
Swin-t	99.53%	142 s/epoch	16.7 ms/image	28,288,354	123.43
MaxViT-t	99.27%	155 s/epoch	41.3 ms/image	30,919,624	5600.0

Swin-t achieved the highest accuracy (99.53%) while maintaining the lowest FLOP count (123.43 M) and the fewest parameters (28.3 M), striking an excellent balance between performance and efficiency. It also exhibited moderate training (142 s/epoch) and inference (16.7 ms/image) times, making it well-suited for applications requiring rapid, resource-efficient predictions. Its windowed self-attention and hierarchical design enable effective feature extraction with lower complexity, making it suitable for real-time and resource-constrained applications.

MaxViT-t achieved a similar accuracy (99.27%) but incurred a substantial computational burden. Despite a comparable parameter count to Swin-t (30.9 M), it required the highest number of FLOPs (5600 M), resulting in the longest training (155 s/epoch) and inference (41.3 ms/image) times. This suggests that MaxViT-t’s hybrid architecture demands substantial resources due to its multi-axis attention and convolutional fusion, highlighting the resource-intensive nature of its architecture.

ViT-b-32, while having the fewest FLOPs among the three models (456.96 M), lagged in accuracy (97.07%) and had the highest number of parameters (88.2 M). It maintained the fastest inference time (5.8 ms/image) and the shortest training time (87 s/epoch), but its lower accuracy and larger model size suggest limited efficiency gains.

Therefore, we may conclude that Swin-t offers the most favorable trade-off between accuracy and computational efficiency, while MaxViT-t, though accurate, demands significantly more resources. ViT-b-32 is less competitive. Swin-t is optimized for lightweight and fast inference, making it ideal for applications with limited computational resources, such as edge devices or clinical settings where rapid predictions are essential.

#### 4.2. Results for Individual Datasets

In this section, we provide more details about the performance of each model and dataset using all the metrics.

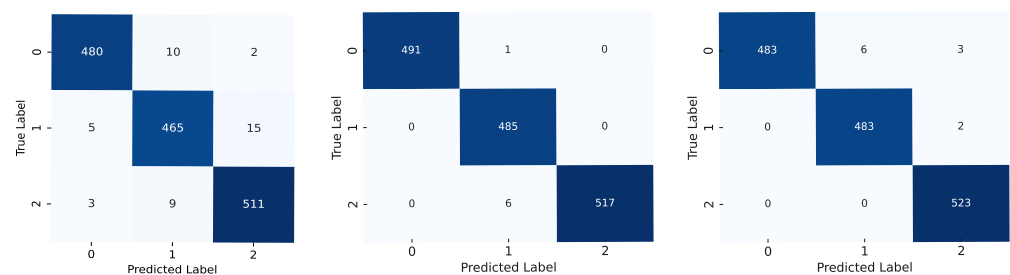


The results for the brain tumor dataset reveal excellent performance across all three models, with Swin-t and MaxViT-t outperforming ViT-b-32, especially in terms of classification precision and general robustness; see Table 8.

**Table 8.** Performance of the models for the brain dataset. Bold letters indicate the best result in each metric.

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
ViT-b-32	97.07	0.971	0.971	0.971	0.998
Swin-t	<b>99.53</b>	<b>0.995</b>	<b>0.996</b>	<b>0.995</b>	<b>1.0</b>
MaxViT-t	99.27	0.993	0.993	0.993	<b>1.0</b>

The precision, recall, and F1-score (0.971) of ViT-b-32 were aligned, showing that the model detects brain tumors without favoring one type of error. Swin-t stood out as the top performer on this dataset, with a high AUC-ROC, indicating flawless discrimination between tumor classes. Its precision (0.995) and recall (0.996) were both exceptionally high, with an F1-score of 0.995, reflecting high reliability and minimal error rate. MaxViT-t also performed exceptionally well, with a high F1-score of 0.993 and an AUC-ROC showing excellent class separability. Although marginally behind Swin-t in terms of overall metrics, it was still highly effective for this classification task. These results reflect a few misclassifications in the confusion matrices of Figure 4.



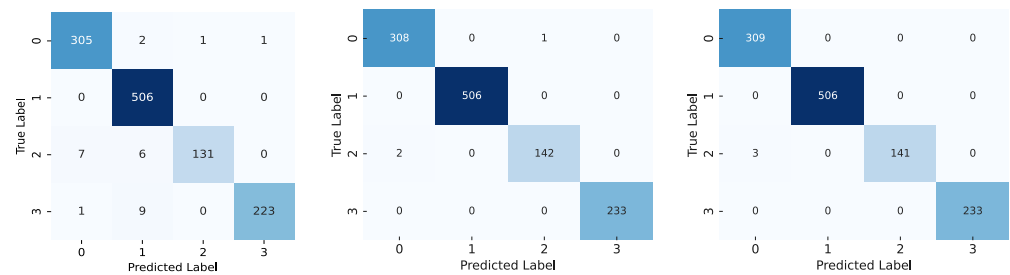
**Figure 4.** Confusion matrices for the models using the brain dataset: results for the ViT-b-32 model are on the left, those for the Swin-t model are in the middle, and those for the MaxViT-t model are on the right. The label codes are glioma (0), meningioma (1), and pituitary (2).

The performance results for the kidney dataset in Table 9 show that ViT-b-32 also stood behind the other two models. Its precision (0.983) and recall (0.964) indicated that it missed more true cases. The AUC-ROC of 0.997 was high, showing a good overall discrimination ability. Swin-t was the top performer, with a precision of 0.997 and a recall of 0.996. This performance, especially given its lightweight architecture, emphasizes Swin-t's remarkable efficiency and robustness for kidney tumor classification.

MaxViT-t matched Swin-t in accuracy and slightly surpassed it in precision (0.998), indicating even fewer false positives. However, its recall (0.995) was marginally lower than Swin-t's, though this difference is negligible in practice. Its AUC-ROC confirmed that it achieved highly effective class separation. Their confusion matrices in Figure 5 reflect the low misclassification rates of these two models.

**Table 9.** Performance of the models for the kidney dataset. Bold letters indicate the best result in each metric.

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
ViT-b-32	97.73	0.983	0.964	0.972	0.997
Swin-t	<b>99.75</b>	0.997	<b>0.996</b>	<b>0.996</b>	<b>1.0</b>
MaxViT-t	<b>99.75</b>	<b>0.998</b>	0.995	<b>0.996</b>	<b>1.0</b>



**Figure 5.** Confusion matrices for the models using the kidney dataset: results for the ViT-b-32 model are on the left, those for the Swin-t model are in the middle, and those for the MaxViT-t model are on the right. The label codes are cyst (0), normal (1), stone (2), and tumor (3).

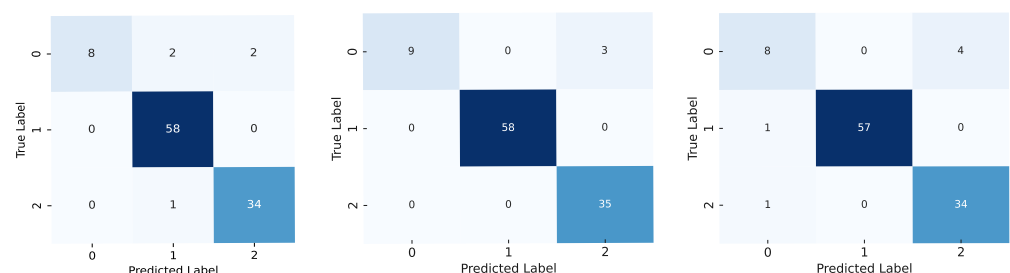
Table 10 shows the results for the lung dataset, which had fewer samples than the other datasets and presented imbalanced classes. This experiment also demonstrated the outstanding performance of Swin-t in terms of overall accuracy and balance across evaluation metrics. ViT-b-32's recall (0.879) was noticeably lower, suggesting that the model missed more true values. Swin-t had the highest accuracy of 97.14%, with good precision (0.974) and recall (0.917). These values indicate that the Swin-t predictions captured many actual tumor cases. Its AUC-ROC of 0.993 further reinforced its excellent class separation ability.

MaxViT-t, on the other hand, showed the lowest performance on this dataset, with an accuracy of 94.29%, and the lowest scores in all other metrics. Its precision (0.898) and recall (0.877) indicated more false positives and negatives compared to other models. The F1-score of 0.883 suggested a weaker balance in classification, and the AUC-ROC of 0.979 ranked below that of both ViT-b-32 and Swin-t.

The confusion matrices in Figure 6 show that Swin-t had fewer misclassifications, only in the third class.

**Table 10.** Performance of the models for the lung dataset. Bold letters indicate the best result in each metric.

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
ViT-b-32	95.24	0.965	0.879	0.911	0.982
Swin-t	<b>97.14</b>	<b>0.974</b>	<b>0.917</b>	<b>0.939</b>	<b>0.993</b>
MaxViT-t	94.29	0.898	0.874	0.883	0.979



**Figure 6.** Confusion matrices for the models using the lung dataset: results for the ViT-b-32 model are on the left, those for the Swin-t model are in the middle, and those for the MaxViT-t model are on the right. The label codes are benign (0), malignant (1), and normal (2).

We may conclude that Swin-t was the best overall performer when we used separate datasets as it combined high accuracy, robust generalization, and balanced metrics across all three tumor datasets. It also adapted better to smaller datasets with imbalanced classes.

#### 4.3. Results for the Combined Dataset

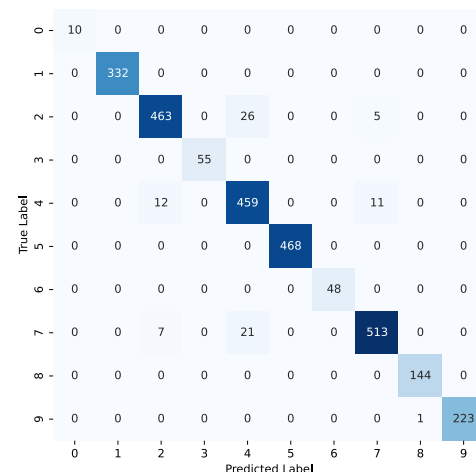
When we combined all the datasets and trained the models, we also obtained high performance in general. Looking at Table 11, the results show that Swin-t provided the best overall performance across all metrics.

ViT-b-32 yielded good results in precision and recall and a high AUC-ROC. It presented slightly more misclassifications in several classes (e.g., glioma and meningioma), possibly due to class similarity or overlap; see Figure 7.

The confusion matrix of the Swin-t model, in Figure 8, reveals a few off-diagonal misclassifications. It had the highest recall and F1-score, indicating a strong ability to capture true positives while maintaining precision.

**Table 11.** Performance of the models when we combined all the datasets. Bold letters indicate the best result in each metric.

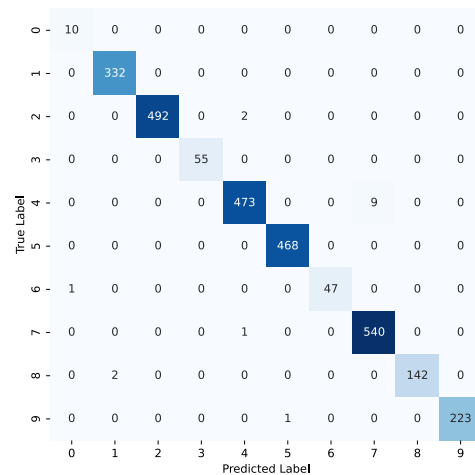
Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
ViT-b-32	97.03	0.983	0.983	0.983	0.999
Swin-t	<b>99.43</b>	<b>0.988</b>	<b>0.994</b>	<b>0.991</b>	<b>1.00</b>
MaxViT-t	98.68	0.974	0.984	0.978	<b>1.00</b>



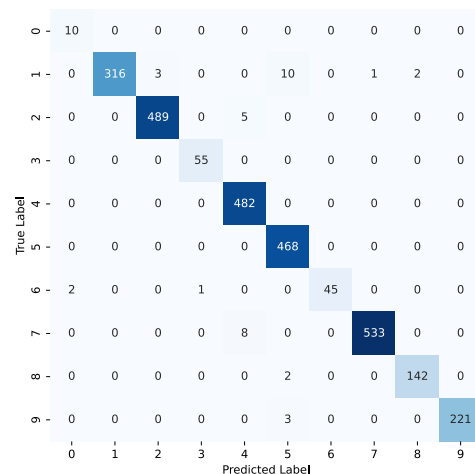
**Figure 7.** Confusion matrix for the ViT-32-b model using the combined datasets. The label codes are as follows: benign (0—lung), cyst (1—kidney), glioma (2—brain), malignant (3—lung), meningioma (4—brain), normal (5—kidney), normal (6—lung), pituitary (7—brain), stone (8—kidney), tumor (9—kidney).

MaxViT-T also provided good performance, although behind the Swin-t models in all metrics and ViT-b-32 in overall scores. Its confusion matrix in Figure 9 shows misclassifications, particularly between similar or adjacent classes. It had a high AUC-ROC, which suggests good separability in probability estimates.

It is interesting to note that the results for the less representative classes, with few samples, were satisfactory, obtaining a high rate of correct classifications. This behavior was similar for the three architectures. This highlights the capacity of Vision Transformers to deal with imbalanced data. We observed fewer misclassifications in the lung labels compared to the confusion matrices in Figure 6, which indicates that the models took advantage of the contents of the other datasets.



**Figure 8.** Confusion matrix for the Swin-t model using the combined datasets. The label codes are as follows: benign (0—lung), cyst (1—kidney), glioma (2—brain), malignant (3—lung), meningioma (4—brain), normal (5—kidney), normal (6—lung), pituitary (7—brain), stone (8—kidney), tumor (9—kidney).



**Figure 9.** Confusion matrix for the MaxViT-t model using the combined datasets. The label codes are as follows: benign (0—lung), cyst (1—kidney), glioma (2—brain), malignant (3—lung), meningioma (4—brain), normal (5—kidney), normal (6—lung), pituitary (7—brain), stone (8—kidney), tumor (9—kidney).

Table 12 compares the results when the models were trained with independent datasets and when trained with the combined dataset. We averaged the results of the independent datasets. In this case, ViT-b-32 significantly improved when trained on the combined dataset, suggesting that it benefited the most from more diverse data.

**Table 12.** Comparison of the performance of the models when trained with individual and combined datasets. The results of the individual datasets from the previous tables were averaged (avg). Bold letters indicate the best result in each metric.

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
ViT-b-32 Individual (avg)	96.77	0.975	0.949	0.959	0.994
ViT-b-32 Combined	97.03	0.983	0.983	0.983	0.999
Swin-t Individual (avg)	98.96	<b>0.988</b>	0.975	0.980	0.998
Swin-t Combined	<b>99.43</b>	<b>0.988</b>	<b>0.994</b>	<b>0.991</b>	<b>1.00</b>
MaxViT-t Individual (avg)	97.99	0.966	0.961	0.962	0.995
MaxViT-t Combined	98.68	0.974	0.984	0.978	<b>1.00</b>

The Swin-t model showed the best performance in both scenarios, but slightly better when trained on the combined dataset. MaxViT-t already performed strongly with individual datasets, but it had a small gain from training on the combined data. Training on the combined dataset generally yielded better, or at least equivalent, performance for the three models, especially for ViT-b-32.

Given these results, performance on the combined dataset suggested some modality transfer capacity, although each organ used a fixed modality.

Table 13 details the per-class evaluation of the Swin-t model on the combined dataset. It demonstrated a strong and balanced performance across all categories, regardless of class size or complexity. Notably, the model handled rare classes—lung tumors—with high accuracy, achieving perfect recall and F1-scores above 0.95. This indicates that the model is sensitive enough to detect underrepresented classes without sacrificing precision.

**Table 13.** Per-class evaluation metrics for the Swin-t model on the combined dataset, including precision, recall, and F1-score for each of the 10 classes. The results highlight the model’s robustness across both majority and minority classes, with strong performance even in low-support categories such as benign and malignant lung tumors.

Dataset	Class	Precision	Recall	F1-Score
Brain	Glioma	1.000	0.996	0.998
	Meningioma	0.995	0.981	0.988
	Pituitary	0.984	0.998	0.991
Kidney	Cyst	0.991	1.000	0.996
	Normal	0.998	1.000	0.999
	Stone	0.993	0.986	0.989
	Tumor	1.000	0.995	0.997
Lung	Benign	0.909	1.000	0.952
	Malignant	1.000	1.000	1.000
	Normal	1.000	0.979	0.989

For the kidney and brain tumor classes, which represent a broader range of conditions and higher support, the model consistently yielded F1-scores near 0.99. Classes such as kidney tumor, cyst, and pituitary tumor showed minimal misclassification, and confusion between classes remained low. This suggests that the Swin-t model can effectively distinguish subtle variations in medical imaging patterns across modalities (MRI and CT).

The model achieved a macro-average F1-score of 0.990, indicating reliable generalization across heterogeneous data. These results validate its capacity to perform accurate multi-organ classification in the presence of label imbalance.

#### 4.4. Comparison with State-of-the-Art Methods

Next, we compared our results for the Swin-t model with state-of-the-art methods. We selected recent methods with top-tier accuracy in each dataset. Table 14 details, for each dataset, the methods, year of publication, techniques employed in each work, and the reported accuracy.

Our fine-tuned Swin-t model exceeded the best previously reported accuracies on the brain dataset, achieving 99.53%, which is higher than the 98.70% reported in [51], using a Vision Transformer, and also higher than the 98.6% obtained in [52], with an improved version of the EfficientNet model.

For the kidney dataset, the accuracy of the Swin-t model was on par with previous works. A recent hybrid CNN+Relief method [53] reported 99.37%, which is lower than our result, and a fine-tuned InceptionV3 model [54] attained 99.96%, which is slightly higher than the 99.75% accuracy obtained with the Swin-t model. In this case, transformer-based



models reached the same state-of-the-art performance on kidney CT as the specialized CNN approaches.

**Table 14.** Comparison with state-of-the-art methods using similar datasets. The table compares the accuracy of prior works with the results of the Swin-t model. These are represented in bold letters.

Dataset	Method	Year	Techniques	Accuracy
Brain	Reddy et al. [51]	2024	ViT (FTVT-L16)	98.70%
	Ishaq et al. [52]	2025	Improved EfficientNet	98.60%
	<b>Our result</b>	<b>2025</b>	<b>Swin Transformer</b>	<b>99.53%</b>
Kidney	Bingol et al. [53]	2023	Hybrid CNN (Relief + WNN)	99.37%
	Pimpalkar et al. [54]	2025	InceptionV3 CNN	99.96%
	<b>Our result</b>	<b>2025</b>	<b>Swin Transformer</b>	<b>99.75%</b>
Lung	Pathan et al. [55]	2024	Optimized CNN with SCA	99.00%
	Jian et al. [56]	2025	CNN + GRU	99.77%
	<b>Our result</b>	<b>2025</b>	<b>Swin Transformer</b>	<b>97.14%</b>

For lung CT classification, the work in [56] obtained 99.77% accuracy using a CNN combined with a GRU model, and a recent optimized CNN model [55] reported 99.0%. Although the Swin-t model achieved lower results (97.14%), it is still a high performance for a Transformer on a reduced dataset.

Therefore, our results are on par with or better than those in the literature. We note that training the model with the combined dataset yielded even better results, especially for the lung dataset, which had fewer samples. Vision Transformers benefit from increasing amounts of data and leverage the variability of multi-organ modalities.

## 5. Conclusions

This work presented a comprehensive evaluation of three state-of-the-art Vision Transformer architectures—ViT-b-32, Swin-t, and MaxViT-t—applied to the classification of brain, lung, and kidney tumors using MRI and CT scans. We analyzed the performance in terms of accuracy, precision, recall, F1-score, and AUC-ROC using individual and combined datasets.

The results of the first fine-tuning step indicated that ViT-b-32 initially offered the most consistent performance across all datasets, particularly benefiting from diverse training data. However, after the second fine-tuning phase, Swin-t emerged as the most effective model, outperforming the others in every metric and dataset configuration. It achieved the highest accuracy (99.75%) and maintained a low computational footprint, making it suitable for deployment in clinical environments with limited resources.

MaxViT-t, while initially underperforming, demonstrated significant gains after full fine-tuning, especially for the kidney and brain datasets. However, its high computational cost makes it less practical for real-time applications, despite its high accuracy. ViT-b-32, although competitive in early stages and particularly strong in recall and generalization, was outperformed by the other models in accuracy and efficiency.

Training with a combined dataset provided improved generalization and robustness, particularly for ViT-b-32 and MaxViT-t, which benefited from the added data diversity. This is more important for less-represented and imbalanced classes. The results highlight the potential of Vision Transformers to handle multi-organ and multiclass medical imaging tasks within a unified framework.

This study confirms that Transformer-based architectures are highly effective for medical image classification, especially when fine-tuned appropriately. Swin-t stands out

with the best trade-off between performance and efficiency, making it a strong candidate for future deployment in computer-aided diagnosis systems.

A limitation of this study is the lack of intra-organ multimodal analysis. While our experiments demonstrated strong performance across different organs and imaging modalities, each organ in this study was represented by a single modality—MRI for the brain and CT for the lung and kidney—without evaluating how the models performed when presented with different modalities for the same anatomical region. This limited our ability to fully assess the robustness and adaptability of Vision Transformers to modality variation within a single pathology.

In future work, we will explore the use of datasets containing multiple imaging modalities for the same condition (e.g., MRI and PET for brain tumors). Additionally, attention map analysis could be used to investigate whether the models focus on consistent anatomical features across modalities. This would not only improve interpretability but also provide valuable insights into how different imaging techniques contribute to models' decision-making processes. Future work may also explore ensemble approaches to combine the strengths of each model. We will also study federated learning for privacy-preserving training and real-time inference deployment in clinical settings.

**Author Contributions:** Conceptualization, Ó.A.M. and J.S.; methodology, J.S.; software, Ó.A.M. and J.S.; validation, Ó.A.M. and J.S.; investigation, Ó.A.M. and J.S.; writing—original draft preparation, Ó.A.M. and J.S.; writing—review and editing, J.S.; supervision, J.S.; project administration, J.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** This study did not involve any direct interaction with human participants or the collection of personally identifiable information. It was conducted using anonymized, publicly available datasets of tumor images from the brain, lung, and kidney, sourced from the Kaggle platform. As these datasets are fully de-identified and openly accessible, Institutional Review Board (IRB) approval was not necessary.

**Data Availability Statement:** The data used in this study are openly available in Kaggle at the following URLs: the brain tumor dataset is available at <https://www.kaggle.com/datasets/obulisainaren/multi-cancer/versions/1> (accessed on 2 November 2024); the lung cancer dataset is available at <https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-dataset> (accessed on 2 November 2024); and the kidney dataset is available at <https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone> (accessed on 2 November 2024).

**Acknowledgments:** During the preparation of this manuscript, the authors used ChatGPT o4, Gemini 2.5 Pro, and DeepSeek-V3 to translate, generate, and improve the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

MRI	Magnetic Resonance Imaging
CT	Computed Tomography
ViT	Vision Transformer
CNN	Convolutional Neural Network
DWT	Discrete Wavelet Transform
SVM	Support Vector Machine

## References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [\[CrossRef\]](#)
2. Deepak, S.; Ameer, P. Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* **2019**, *111*, 103345. [\[CrossRef\]](#)
3. Coudray, N.; Ocampo, P.S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.L.; Razavian, N.; Tsirigos, A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **2018**, *24*, 1559–1567. [\[CrossRef\]](#)
4. Hermsen, M.; de Bel, T.; Den Boer, M.; Steenbergen, E.J.; Kers, J.; Florquin, S.; Roelofs, J.J.; Stegall, M.D.; Alexander, M.P.; Smith, B.H.; et al. Deep learning-based histopathologic assessment of kidney tissue. *J. Am. Soc. Nephrol.* **2019**, *30*, 1968–1979. [\[CrossRef\]](#)
5. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R. BiFormer: Vision Transformer with Bi-Level Routing Attention. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333. [\[CrossRef\]](#)
6. Cai, H.; Li, J.; Hu, M.; Gan, C.; Han, S. EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 17302–17313.
7. Wan, Q.; Huang, Z.; Lu, J.; Yu, G.; Zhang, L. SeaFormer++: Squeeze-enhanced axial transformer for mobile visual recognition. *Int. J. Comput. Vis.* **2025**, *133*, 3645–3666. [\[CrossRef\]](#)
8. Lee, S.; Lee, M. MetaSwin: A unified meta vision transformer model for medical image segmentation. *PeerJ Comput. Sci.* **2024**, *10*, e1762. [\[CrossRef\]](#) [\[PubMed\]](#)
9. He, Y.; Nath, V.; Yang, D.; Tang, Y.; Myronenko, A.; Xu, D. SwinUNETR-V2: Stronger Swin Transformers with Stagewise Convolutions for 3D Medical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2023*; Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R., Eds.; Springer: Cham, Switzerland, 2023; pp. 416–426. [\[CrossRef\]](#)
10. Wang, Y.; Huang, N.; Li, T.; Yan, Y.; Zhang, X. Medformer: A Multi-Granularity Patching Transformer for Medical Time-Series Classification. In *Proceedings of the Advances in Neural Information Processing Systems*; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 36314–36341.
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [\[CrossRef\]](#)
12. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [\[CrossRef\]](#)
13. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12009–12019. [\[CrossRef\]](#)
14. Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. MaxViT: Multi-axis Vision Transformer. In *Proceedings of the European Conference on Computer Vision (ECCV 2022)*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; pp. 459–479. [\[CrossRef\]](#)
15. Naren, O.S. Multi Cancer Dataset [Dataset]. 2022. Available online: <https://www.kaggle.com/datasets/obulisainaren/multi-cancer/versions/1> (accessed on 2 November 2024). [\[CrossRef\]](#)
16. Alyasriy, H.; AL-Huseiny, M. The IQ-OTH/NCCD Lung Cancer Dataset. Mendeley Data, V4. 2023. Available online: <https://data.mendeley.com/datasets/bhmdr45bh2/4> (accessed on 2 November 2024). [\[CrossRef\]](#)
17. Kaggle. The IQ-OTH/NCCD Lung Cancer Dataset. 2022. Available online: <https://www.kaggle.com/datasets/hamdallak/the-igothnccd-lung-cancer-dataset> (accessed on 2 November 2024).
18. Kaggle. CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone. 2022. Available online: <https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone> (accessed on 26 June 2018).
19. Medina, J.M.; Sánchez, J. High Accuracy Brain Tumor Classification with EfficientNet and Magnetic Resonance Images. In Proceedings of the 5th International Conference on Advances in Signal Processing and Artificial Intelligence, Tenerife, Spain, 7–9 June 2023; International Frequency Sensor Association (IFSA) Publishing: Barcelona, Spain, 2023.
20. Reyes, D.; Sánchez, J. Performance of convolutional neural networks for the classification of brain tumors using magnetic resonance imaging. *Heliyon* **2024**, *10*, e25468. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Khan, A.; Rauf, Z.; Khan, A.R.; Rathore, S.; Khan, S.H.; Shah, N.S.; Farooq, U.; Asif, H.; Asif, A.; Zahoor, U.; et al. *A Recent Survey of Vision Transformers for Medical Image Segmentation*; Technical Report; Cornell University: Ithaca, NY, USA, 2023. [\[CrossRef\]](#)

22. Cheng, J.; Huang, W.; Cao, S.; Yang, R.; Yang, W.; Yun, Z.; Wang, Z.; Feng, Q. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS ONE* **2015**, *10*, e0140381. [\[CrossRef\]](#)
23. Gumaei, A.; Hassan, M.M.; Hassan, M.R.; Alelaiwi, A.; Fortino, G. A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification. *IEEE Access* **2019**, *7*, 36266–36273. [\[CrossRef\]](#)
24. Ismael, M.R.; Abdel-Qader, I. Brain tumor classification via statistical features and back-propagation neural network. In Proceedings of the IEEE International Conference on Electro/Information Technology (EIT), Rochester, MI, USA, 3–5 May 2018; pp. 0252–0257. [\[CrossRef\]](#)
25. Bahadure, N.B.; Ray, A.K.; Thethi, H.P. Comparative approach of MRI-based brain tumor segmentation and classification using genetic algorithm. *J. Digit. Imaging* **2018**, *31*, 477–489. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Afshar, P.; Mohammadi, A.; Plataniotis, K.N. Brain tumor type classification via capsule networks. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 3129–3133. [\[CrossRef\]](#)
27. Sajjad, M.; Khan, S.; Muhammad, K.; Wu, W.; Ullah, A.; Baik, S.W. Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *J. Comput. Sci.* **2019**, *30*, 174–182. [\[CrossRef\]](#)
28. Ayadi, W.; Elhamzi, W.; Charfi, I.; Atri, M. Deep CNN for brain tumor classification. *Neural Process. Lett.* **2021**, *53*, 671–700. [\[CrossRef\]](#)
29. Srinivas, C.; KS, N.P.; Zakariah, M.; Alothaibi, Y.A.; Shaukat, K.; Partibane, B.; Awal, H. Deep transfer learning approaches in performance analysis of brain tumor classification using MRI images. *J. Healthc. Eng.* **2022**, *2022*, 3264367. [\[CrossRef\]](#)
30. Kumar, D.; Wong, A.; Clausi, D.A. Lung Nodule Classification Using Deep Features in CT Images. In Proceedings of the 12th Conference on Computer and Robot Vision, Halifax, NS, Canada, 3–5 June 2015; pp. 133–138. [\[CrossRef\]](#)
31. Sathishkumar, R.; Kalaivasan, K.; Prabhakaran, A.; Aravind, M. Detection of lung cancer using SVM classifier and KNN algorithm. In Proceedings of the International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 29–30 March 2019; pp. 1–7. [\[CrossRef\]](#)
32. Shen, W.; Zhou, M.; Yang, F.; Yang, C.; Tian, J. Multi-scale Convolutional Neural Networks for Lung Nodule Classification. In *Proceedings of the Information Processing in Medical Imaging*; Ourselin, S., Alexander, D.C., Westin, C.F., Cardoso, M.J., Eds.; Springer: Cham, Switzerland, 2015; pp. 588–599. [\[CrossRef\]](#)
33. Liu, X.; Hou, F.; Qin, H.; Hao, A. Multi-view multi-scale CNNs for lung nodule type classification from CT images. *Pattern Recognit.* **2018**, *77*, 262–275. [\[CrossRef\]](#)
34. Zhang, B.; Qi, S.; Monkam, P.; Li, C.; Yang, F.; Yao, Y.D.; Qian, W. Ensemble learners of multiple deep CNNs for pulmonary nodules classification using CT images. *IEEE Access* **2019**, *7*, 110358–110371. [\[CrossRef\]](#)
35. Hussein, S.; Kandel, P.; Bolan, C.W.; Wallace, M.B.; Bagci, U. Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches. *IEEE Trans. Med. Imaging* **2019**, *38*, 1777–1787. [\[CrossRef\]](#)
36. Nejad, R.R.; Hooshmand, S. HViT4Lung: Hybrid Vision Transformers Augmented by Transfer Learning to Enhance Lung Cancer Diagnosis. In Proceedings of the 2023 5th International Conference on Bio-Engineering for Smart Technologies (BioSMART), Paris, France, 7–9 June 2023; pp. 1–7. [\[CrossRef\]](#)
37. Gai, L.; Xing, M.; Chen, W.; Zhang, Y.; Qiao, X. Comparing CNN-based and transformer-based models for identifying lung cancer: Which is more effective? *Multimed. Tools Appl.* **2024**, *83*, 59253–59269. [\[CrossRef\]](#)
38. Kim, D.Y.; Park, J.W. Computer-Aided Detection of Kidney Tumor on Abdominal Computed Tomography Scans. *Acta Radiol.* **2004**, *45*, 791–795. [\[CrossRef\]](#)
39. Zhou, B.; Chen, L. Atlas-based semi-automatic kidney tumor detection and segmentation in CT images. In Proceedings of the 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, China, 15–17 October 2016; pp. 1397–1401. [\[CrossRef\]](#)
40. Feng, Z.; Rong, P.; Cao, P.; Zhou, Q.; Zhu, W.; Yan, Z.; Liu, Q.; Wang, W. Machine learning-based quantitative texture analysis of CT images of small renal masses: Differentiation of angiomyolipoma without visible fat from renal cell carcinoma. *Eur. Radiol.* **2018**, *28*, 1625–1633. [\[CrossRef\]](#)
41. Erdim, C.; Yardimci, A.H.; Bektas, C.T.; Kocak, B.; Koca, S.B.; Demir, H.; Kilickesmez, O. Prediction of benign and malignant solid renal masses: Machine learning-based CT texture analysis. *Acad. Radiol.* **2020**, *27*, 1422–1429. [\[CrossRef\]](#)
42. Rathnayaka, P.; Jayasundara, V.; Nawaratne, R.; De Silva, D.; Ranasinghe, W.; Alahakoon, D. *Kidney Tumor Detection Using Attention Based U-Net*; Technical Report; University of Minnesota: Minneapolis, MN, USA, 2019.
43. Lin, Z.; Cui, Y.; Liu, J.; Sun, Z.; Ma, S.; Zhang, X.; Wang, X. Automated segmentation of kidney and renal mass and automated detection of renal mass in CT urography using 3D U-Net-based deep convolutional neural network. *Eur. Radiol.* **2021**, *31*, 5021–5031. [\[CrossRef\]](#)
44. Alzu'bi, D.; Abdullah, M.; Hmeidi, I.; AlAzab, R.; Gharaibeh, M.; El-Heis, M.; Almotairi, K.H.; Forestiero, A.; Hussein, A.M.; Abualgah, L. Kidney tumor detection and classification based on deep learning approaches: A new dataset in CT scans. *J. Healthc. Eng.* **2022**, *2022*, 3861161. [\[CrossRef\]](#)

45. Praveen, S.P.; Sidharth, S.R.; Priya, T.K.; Kavuri, Y.S.; Sindhura, S.M.; Donepudi, S. ResNet and ResNeXt-Powered Kidney Tumor Detection: A Robust Approach on a Subset of the KAUH Dataset. In Proceedings of the 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 11–13 December 2023; pp. 749–757. [\[CrossRef\]](#)
46. Özbay, E.; Özbay, F.A.; Gharehchopogh, F.S. Kidney Tumor Classification on CT images using Self-supervised Learning. *Comput. Biol. Med.* **2024**, *176*, 108554. [\[CrossRef\]](#)
47. Bogomolov, A.; Zabarylo, U.; Kirsanov, D.; Belikova, V.; Ageev, V.; Usenov, I.; Galyanin, V.; Minet, O.; Sakharova, T.; Danielyan, G.; et al. Development and testing of an LED-based near-infrared sensor for human kidney tumor diagnostics. *Sensors* **2017**, *17*, 1914. [\[CrossRef\]](#)
48. Cheng, J. Brain Tumor Dataset. 2017. Available online: [https://figshare.com/articles/dataset/brain\\_tumor\\_dataset/1512427/8](https://figshare.com/articles/dataset/brain_tumor_dataset/1512427/8) (accessed on 2 November 2024). [\[CrossRef\]](#)
49. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision Transformers for Dense Prediction. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 12159–12168. [\[CrossRef\]](#)
50. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 10096–10106. [\[CrossRef\]](#)
51. Reddy, C.K.K.; Reddy, P.A.; Janapati, H.; Assiri, B.; Shuaib, M.; Alam, S.; Sheneamer, A. A fine-tuned vision transformer based enhanced multi-class brain tumor classification using MRI scan imagery. *Front. Oncol.* **2024**, *14*, 1400341. [\[CrossRef\]](#)
52. Ishaq, A.; Ullah, F.U.M.; Hamandawana, P.; Cho, D.J.; Chung, T.S. Improved EfficientNet Architecture for Multi-Grade Brain Tumor Detection. *Electronics* **2025**, *14*, 710. [\[CrossRef\]](#)
53. Bingol, H.; Yildirim, M.; Yildirim, K.; Alatas, B. Automatic classification of kidney CT images with relief based novel hybrid deep model. *PeerJ Comput. Sci.* **2023**, *9*, e1717. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Pimpalkar, A.; Saini, D.K.J.B.; Shelke, N.; Balodi, A.; Rapate, G.; Tolani, M. Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. *Sci. Rep.* **2025**, *15*, 10741. [\[CrossRef\]](#)
55. Pathan, S.; Ali, T.; P G, S.; P, V.K.; Rao, D. An optimized convolutional neural network architecture for lung cancer detection. *APL Bioeng.* **2024**, *8*, 026121. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Jian, W.; Haq, A.U.; Afzal, N.; Khan, S.; Alsolai, H.; Alanazi, S.M.; Zamani, A.T. Developing an innovative lung cancer detection model for accurate diagnosis in AI healthcare systems. *Sci. Rep.* **2025**, *15*, 22945. [\[CrossRef\]](#) [\[PubMed\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.