EVALUATION OF VISION TRANSFORMERS FOR MULTIMODAL IMAGE CLASSIFICATION: A CASE STUDY ON BRAIN, LUNG, AND KIDNEY TUMORS

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

Javier Sánchez
Centro de Tecnologías de la Imagen (CTIM)
Instituto Universitario de Cibernética, Empresas y Sociedad (IUCES)
University of Las Palmas de Gran Canaria
35017 Las Palmas de Gran Canaria, Spain
jsanchez@ulpgc.es

ABSTRACT

Neural networks have become the standard technique for medical diagnostics, especially in cancer detection and classification. This work evaluates the performance of Vision Transformers architectures, including Swin Transformer and MaxViT, in 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 includes 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 image modalities. The results revealed that the Swin Transformer provided high accuracy, achieving up to 99.9% for kidney tumor classification and 99.3% accuracy in a combined dataset. MaxViT also provided excellent results in individual datasets but performed poorly when data is combined. This research highlights the adaptability of Transformer-based models to various image modalities and features. However, challenges persist, including limited annotated data and interpretability issues. Future works will expand this study by incorporating other image modalities and enhancing diagnostic capabilities. Integrating these models across diverse datasets could mark a pivotal advance in precision medicine, paving the way for more efficient and comprehensive healthcare solutions.

Keywords Brain tumor · Lung tumor · Kidney tumor · Neural Networks · Vision Transformer · Swin Transformer · MaxViT

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. According to [1], there were an estimated 19.3 million new cases of cancer and 9.9 million cancer-related deaths in 2020.

In the last decade, medical technology has rapidly improved, with neural networks playing a key role. The intersection of computer science and medicine has yielded significant improvements in disease research and diagnosis. Artificial intelligence (AI) has lightened the burden on medical professionals, enhancing diagnostic efficiency. AI finds applications in diverse areas, from health-monitoring wearables to DNA analysis algorithms.

Technology plays a crucial role in early and accurate cancer diagnoses, providing healthcare professionals with automated tools to detect diseases efficiently and minimizing human error and workload. In particular, the use of deep neural networks for the detection and classification has become widely used in several types of tumors, such as brain [2], lung [3], or kidney tumors [4]. Nevertheless, not many 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 magnetic resonance imaging (MRI) and computed tomography (CT) scans to detect tumors. We explore architectures based on the Vision Transformer and its variants, which excel at processing images and identifying anomalous patterns. Our work aims to classify different types of tumors with distinct image formats into a single model. This will allow us to assess the capabilities of these architectures in generalizing to multiple diseases and image formats. The objective is to integrate different data and image formats into a single Transformer model, adapting to cancerous and benign disease detection. 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 will depend on the specific task. We analyze several architectures, such as the Vision Transformer (ViT) [5], the Swin Transformer [6, 7], and MaxVit [8].

For the training of these networks, we selected three datasets: one of brain images [9], containing three types of tumors (*gliomas*, *meningiomas*, and *pituitary tumors*); another of lung images [10, 11], containing CT scans of lung cancer and healthy lungs, with three distinct classes (*normal*, *benign*, and *malignant*); and another dataset of kidney images [12], containing four different categories (*non-tumor*, *stone*, *cyst*, and *cancer*). The first dataset includes 15,000 MRIs, with 5,000 images in each subclass; the second dataset contains 1,190 CT scans of lung cancer and healthy lungs; and the third one has 12,446 CT scans distributed in four classes.

The experimental results analyze the performance of each neural network when trained from scratch and using transfer learning. In many cases, the models' accuracy is 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.

Section 2 summarizes the 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 the work and propose some ideas for future works.

2   Related Work

Artificial intelligence is making significant advances in tumor detection, with deep learning models, like EfficentNet [13], achieving high accuracy rates with MRI images. Other standard models, such as VGG and MobileNet, have also shown outstanding results [14]. Furthermore, the use of Transformer models, such as the Vision and the Swin Transformers, are 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 [15] 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 [16] focused on the classification of brain tumors 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 then classification. The features proposed in [17] 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 the work proposed

in [18], which relies on the 2D Discrete Wavelet Transform (DWT) and 2D Gabor filters to extract statistical features from MRI and then feed into a neural network. The approach presented in [19] tackled the problem of segmentation and classification of MRI using genetic algorithms. The authors of [20] explored brain tumor classification using Capsule Networks. This type of network has several benefits over CNNs since they are robust to rotation and affine transformations and require less training data.

The system proposed in [21] consists 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 [22] have extensively used pure CNN models for brain tumor classification, obtaining high accuracy in different datasets.

Several works [23] have conducted a performance analysis of transfer learning in CNN models (such as VGG-16, ResNet-50, or Inception-v3) for automatic prediction of tumor cells in the brain. A recent work [14] reported a detailed performance assessment analysis of many convolutional architectures from different perspectives, drawing important conclusions about these techniques. Several models can attain high accuracy, even for networks with a relatively low number of parameters like EffientNet [13].

2.2 Lung tumors

Lung cancer remains one of the leading causes of cancer-related deaths worldwide. Early and accurate detection through imaging techniques like CT scans is crucial for improving patient outcomes. Deep learning models have shown significant promise in automating the classification of lung tumors from CT scans.

Classification of lung tumors using machine learning techniques has been an active area of research. Numerous studies have explored various algorithms and methodologies to improve diagnostic accuracy and efficiency.

Early work in lung tumor classification primarily utilized traditional machine learning algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Decision Trees. For instance, the authors of [24] employed SVM for classifying lung nodules in CT images, achieving notable accuracy by optimizing hyperparameters and using feature selection techniques to reduce dimensionality. Similarly, [25] used k-NN combined with SVM, demonstrating improved classification performance on small datasets.

With the advent of deep learning, CNNs have become the preferred choice for image-based lung tumor classification. The work presented in [26] developed a deep CNN model that outperformed traditional methods by automatically learning hierarchical features from raw CT images. Their approach significantly reduced the need for manual feature extraction, leading to higher accuracy and robustness. On the other hand, [27] further advanced this field by introducing a multi-view CNN that integrates information from multiple CT slices, enhancing the ability of the model to capture spatial dependencies and improve classification accuracy. This method demonstrated superior performance in distinguishing between benign and malignant nodules compared to single-view CNNs.

Ensemble methods, which combine the predictions of multiple models, have been explored to enhance classification performance. For example, the authors of [28] utilized an ensemble of CNNs and several traditional machine-learning techniques to classify lung tumors, achieving 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 in scenarios with limited labeled data. The model presented in [29] employed transfer learning using pre-trained CNNs, achieving high classification accuracy with reduced training time. Their approach demonstrated that leveraging pre-trained models can enhance performance, especially when dealing with small medical datasets.

Recent studies have focused on improving the interpretability and explainability of machine learning models in lung tumor classification. The work in [30] introduced an explainable AI framework that combines a 3D U-Net with attention mechanisms to highlight the most relevant regions in CT images.

2.3 Kidney tumors

The application of machine learning to the classification of kidney tumors has 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 [31] 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, [32] 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 (e.g., clear cell renal carcinoma, papillary renal cell carcinoma) 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, [33] explored handcrafted features combined with SVMs for detecting renal masses, achieving promising results in sensitivity and specificity. Similarly, [34] utilized radiomics features in combination with eight machine-learning techniques, like logistic regression, decision trees, and SVMs, to classify renal tumors.

Advancements in deep learning techniques have also played a crucial role in kidney tumor detection. For example, [35] proposed a CNN-based U-Net architecture with an attention mechanism for kidney tumor detection. On the other hand, [36] developed a 3D U-Net-based deep convolutional neural network for automatically segmenting kidney and renal masses, achieving high accuracy in tumor segmentation. Furthermore, [37] and [38] focused on deep learning approaches for kidney tumor detection and classification, utilizing models such as 2D-CNN, ResNet, and ResNeXt to improve the detection accuracy. The work in [39] explored self-supervised learning for kidney tumor classification on CT images. In addition, [40] studied the development of an LED-based near-infrared sensor for human kidney tumor diagnostics, providing 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, differentiating tissues, although it is more expensive than tomography.

3.1 Datasets

We selected three datasets related to brain, lung, and kidney tumors. In the next sections, we explain each dataset 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 the complete tumor removal, making early detection crucial for timely and appropriate treatment before the tumor grows.

In this work, we will test 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.

Figure 1 depicts several examples of brain tumors in the dataset [9]. This dataset forms part of a larger collection that encompasses eight distinct cancer types from different organs: acute lymphoblastic leukemia, brain, breast, cervical,
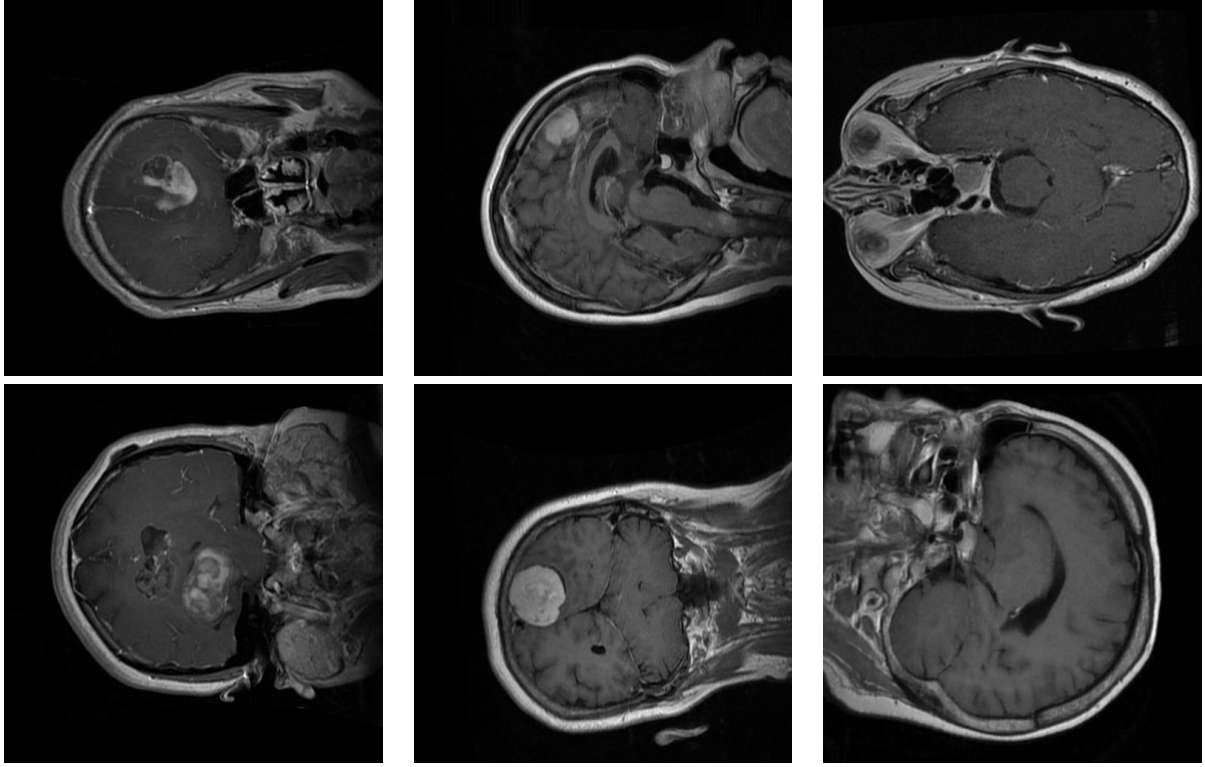
4

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.

kidney, lung, colon, and oral cancer. The images are in JPEG format with dimensions of 512×512 pixels. We selected brain cancer, which contains the three previous tumor subclasses. Each subclass includes the same number of images (5,000 MRI).

3.1.2 Lung Tumor Dataset

Lung tumors can arise from various lung cells, including 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 size. Lung tumors often do not cause symptoms and are usually detected during imaging tests conducted for reasons unrelated to lung cancer.

Since these frequently go unnoticed 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 [10, 11], which 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 comprises 110 cases, varying in patient age, sex, residential area, and socioeconomic status, among other variables. The total number of images is 1,190, distributed as shown in Table 1.

Type	#Cases	Images
Healthy Lung	55	416
Benign Tumor Lung	15	213
Cancerous Tumor Lung	40	561

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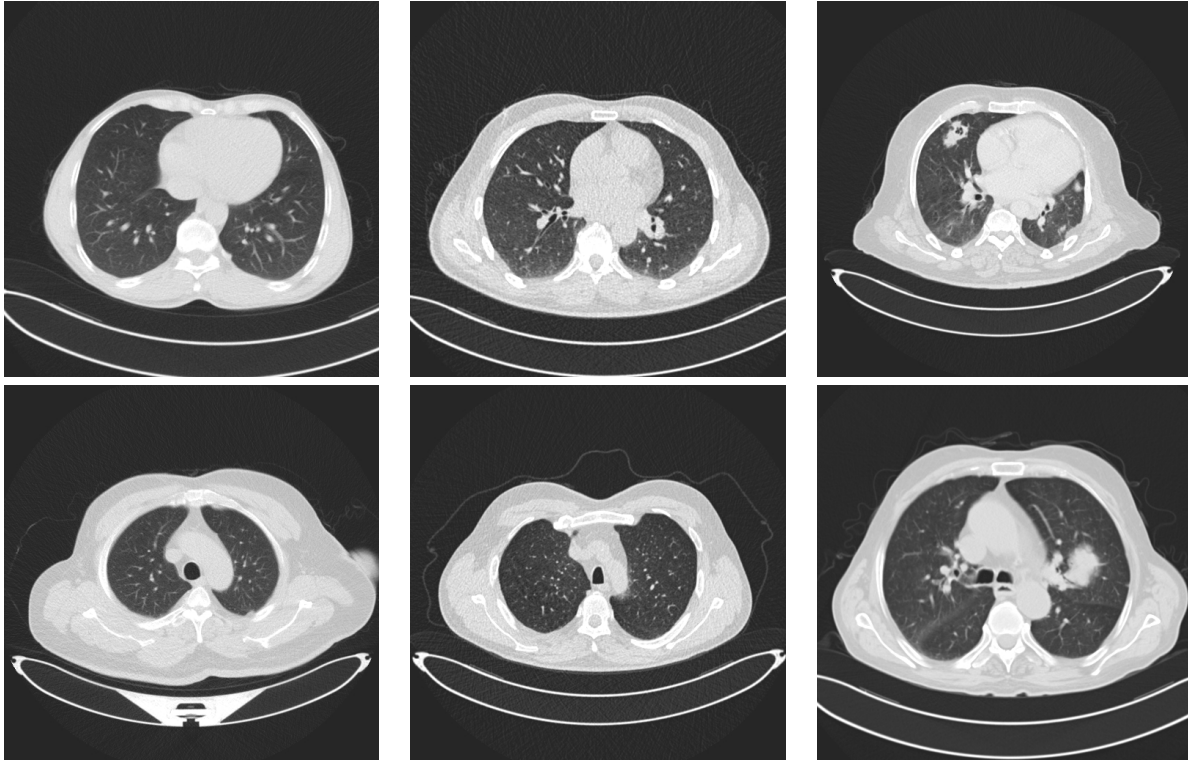
Figure 2: Examples of lung images. The images in the first column represent lungs with no tumors. The images in the middle column represent images of benign tumors and the last column contains two images with malignant tumors.

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 during the stipulated time after the diagnosis of the disease or the start of treatment.

We selected a dataset of kidney images from Kaggle [12], 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 the tumor, achieving a cure rate of over 70%.

Figure 3 shows several samples from the dataset, with two scans of each dataset class.
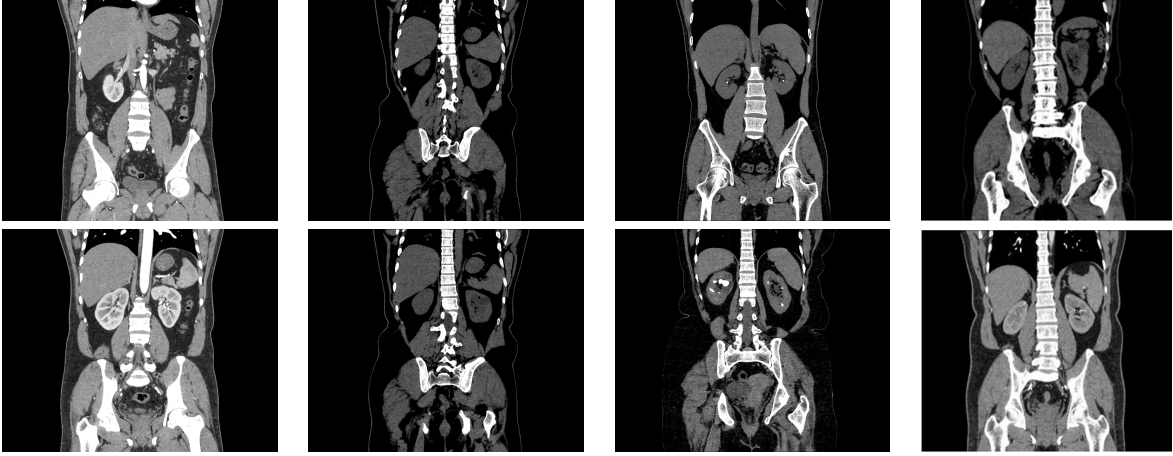
Figure 3: Examples from the kidney dataset. The two images in the first column represent CT images of *normal* kidneys, the two images in the second column depict kidneys with *cysts*, The images on the third column represent CT images of kidneys with *stones*, and the images in the last column depict 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 contains a total of 12,446 images. Table 2 shows the distribution of images into the four classes.

Class	Number of images
Normal	5,077
Cyst	3,709
Stone	1,377
Cancer	2,283

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

3.2   Neural Networks

In this work, we compare the performance of three vision Transformers, i.e., the Vision Transformer (ViT) [5], the Swin Transformer [6], and MaxViT [8]. These architectures have surpassed the performance of previous CNNs on many computer vision tasks.

3.2.1   Vision Transformer

The Vision Transformer (ViT) [5] 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.

The ViT architecture has shown remarkable abilities, achieving comparable or even better performance than traditional CNNs on many computer vision tasks. 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 will be tested. Starting from this variant, we choose two alternatives, ViT-b-16 and ViT-b-32, for input patches

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

of 16x16 pixels and 32x32 pixels, respectively. The model with smaller patches has a higher computational cost since the size of the Transformer sequence is inversely proportional to the square of the patch size.

3.2.2 Swin Transformer

The Swin Transformer [6, 7] 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 but 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 computation complexity for 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 suitable as a general-purpose backbone for various vision tasks. Due to its hierarchical feature maps and efficient computation, the Swin Transformer 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 like tumors. Its benefits lie in improved efficiency, hierarchical feature extraction, and suitability for various computer vision tasks.

3.2.3 MaxViT Transformer

The MaxViT Transformer [8] is a hybrid architecture that combines the strengths of CNNs and Vision Transformers to create a powerful image classification model. MaxViT 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 [41] blocks and grid attention layers, allowing the model to reach a global context throughout the entire network. MaxViT scales well with large datasets, and maintains linear complexity concerning the grid attention used, making it suitable for high-resolution inputs.

In the experiments, we will test different configurations of these architectures, like the ViT-b-16 and ViT-b-32 models. Regarding the Swin Transformer, there are three variants: *tiny*, *small*, and *base*. In the case of MaxViT, we will test the base model. When considering performance, the focus is on the relationship between training time and final accuracy. Models that require extensive computation, both in terms of time and GPU memory load, were discarded.

Table 4: Complexity of the models and its variants.

Models	Parameters	FLOPS (in Millions)	Size (in Megabytes)
ViT-b-16	86,415,592	456.96	360.24
ViT-b-32	88,185,064	456.96	344.59
Swin-t	28,288,354	123.43	141.22
Swin-s	49,606,258	208.40	236.33
Swin-b	87,768,224	363.69	397.41
MaxViT-b	119,880,192	138.50	545.68

The variants of the models that we used in the experiments are presented in Table 4. For each model variant, we summarize the number of parameters, the floating point operations per second (FLOPS), and their size in megabytes.

3.3   Experimental setup

We use the Adam optimizer for learning the parameters of the neural networks, which is frequently used in image classification tasks. For the loss function, we choose Categorical Cross-Entropy, given by the following expression:

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

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 yields a value between 0 and 1, representing a probability for each label or class trained in the model. When the error is very high, cross-entropy significantly penalizes those values that deviate from the expected predictions.

The metric used to compare the models in the experiment is *accuracy*. It is an intuitive performance measure and is calculated as a ratio of correctly predicted observations to the total observations. Its formula is given by:

$$\text{accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}. \tag{2}$$

The source code was implemented in Python and the deep learning models were implemented using the PyTorch framework. The training was carried out in an NVIDIA GPU GeForce RTX 2060 SUPER. 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: Ranges from 4 to 32 images per batch.
- Image Size: Images were scaled to a fixed dimension of 224x224 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 are deactivated at a rate of 30% and 40%. A dropout value within this range has always been active.
- Epochs: the number of epochs ranges from 5 to 200, depending on the dataset and mini-batch size applied.

4   Results

Each dataset was split into a training, validation, and test set. The training set was about 60-80% of the dataset and was used to learn the parameters of the networks. The size of the validation set was around 5-15% of the datasets. This set was used to validate the training process and to find the best hyperparameters. The size of the test set was about 15%. This was used to evaluate the performance of models. This set was never used during training, serving as a measure of the effectiveness of models.

Table 5 shows the results when we train the models from scratch. In this case, we selected the most accurate version for each architecture. In the last row, we also show the results when we combine all the datasets and increase the output layer to include all the labels. We obtain the best accuracy for the Kidney and Lung datasets with the Swin-t model, and the best results for the Brain and the combination of all datasets with the MaxViT-b model.

Table 5: Accuracy of the models training from scratch. This table shows the accuracy of the ViT-b-32, Swin-t, and MaxViT-b for each dataset. The last row shows 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-b
Brain	81.5%	92.4%	**96.4%**
Kidney	91.4%	**93.0%**	92.2%
Lung	98.2%	**98.9%**	96.0%
All datasets	89.9%	91.6%	**95.7%**

These results show that no model stands out above the others, and the ranking depends on the chosen dataset. In general, combining the datasets does not yield better results, providing poorer results for the three models. This might mean that images with different characteristics do not contribute to the generalization of the neural networks.

Table 6 shows similar results when we use transfer learning. We load the parameters of the models that were learned with ImageNet. In this case, we free all the parameters and train the models with a low learning rate.

Table 6: Accuracy of the models using transfer learning. This table shows the accuracy of the ViT-b-32, Swin-t, and MaxViT-b for each dataset. The last row shows 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-b
Brain	94.8%	95.2%	**96.9**%
Kidney	97.9%	**99.4**%	98.1%
Lung	96.3%	**98.8**%	97.2%
All datasets	96.5%	**97.2**%	96.6%

The accuracy of the models increases with respect to the previous experiment, meaning that transfer learning must be the choice for training the models. The experiment with all the datasets provided much better results, although still insufficient to beat the training with standalone datasets. Swin-t obtained the best results for the Kidney, Lung, and all the datasets combined, whereas MaxViT-b obtained the best results for the Brain dataset.

In the following experiment in Table 7, we show the results of training the models in two steps: first, we free the last classification layer and train with a large learning rate; then, we free all the network parameters and train with a small learning rate.

Table 7: Accuracy of the models using transfer learning and fine-tuning. This table shows the accuracy of the ViT-b-32, Swin-t, and MaxViT-b for each dataset. The last row shows the results of each model by combining the three datasets. Bold letters highlight the best results in each row.

Tumor Type	Vit-b-32	Swin-t	MaxVit-b
Brain	93.5%	**99.7**%	98.7%
Kidney	95.6%	**99.9**%	99.8%
Lung	97.8%	95.1%	**98.8**%
All datasets	97.4%	**99.3**%	97.5%

The results in this case are still more accurate than in the previous two experiments. We observe that the Swin-t model outperforms the other models in two datasets, and provides competitive results for the combined dataset.

Table 8 compares the results of the methods using the kidney dataset. This table shows the accuracy of the models and other information, such as the number of parameters, FLOPS, and its size in megabytes, that allows us to understand its overall performance.

Table 8: Performance assessment of different configurations of vision Transformers for the kidney dataset: The first column shows the version of the model used in each experiment; the second column depicts the accuracy of the model for the test set; the third column shows the number of parameters of each network; the fourth column, the number of FLOPS; and the last column, the size of the networks in megabytes. The last two columns are copied from Table 4.

Model	Accuracy	Number of Parameters	FLOPS (in Millions)	Size (in Megabytes)
ViT-b-16	92.4	86,415,592	456.96	360.24
ViT-b-32	95.6	88,185,064	456.96	344.59
Swin-t	99.9	28,288,354	123.43	141.22
Swin-s	99.3	49,606,258	208.40	236.33
Swin-b	98.8	87,768,224	363.69	397.41
MaxViT-b	**99.8**	119,880,192	138.50	545.68

These results show that Vit_b_32 achieves better accuracy than Vit_b_16 and is faster. Despite having more parameters, Vit_b_32 occupies less memory space, which could explain the shorter training time. This efficiency may be due to differences in the model's structure, data types, and precision. On the other hand, Swin_t achieves better accuracy than Swin_s and Swin_b, and is 1.76 times faster. The smaller model that achieves better results in less training time likely reflects the complexity of the datasets. It is possible that, for these datasets, a model with a larger number of parameters is not necessary. More parameters may lead to overfitting, as the validation accuracy tends to reach a lower relative maximum with longer training times compared to the smaller architecture.

MaxViT_b also achieves a very high accuracy (99.8%) compared with other architectures within a reasonable training time. Looking at Table 7, we can raise the following conclusions for each data set: ViT underperforms the other models in all datasets; MaxVit, on the other hand, obtains the best performance in the Brain dataset and the combined dataset; and the Swin model obtains higher accuracy in the Kidney and Lung datasets.

5   Conclusion

In this study, we assessed the performance of several neural networks focusing specifically on the Vision Transformer and its key variants. Our models were trained on MRI and CT scans, individually and in combination. The models were designed to make accurate predictions without being influenced by image modality, ensuring robustness across different image types. Additionally, the number of classes and cancer typology varied with the datasets, with brain images containing more specific cancer types than the kidney and lung datasets.

Two architectures —-MaxViT and Swin Transformer—- emerged as the top performers. Swin Transformer demonstrated remarkable performance, achieving an accuracy of 99.3%, while MaxViT followed closely with 97.5%. These results were obtained on a comprehensive dataset that included unrelated diseases and a variety of medical image formats, highlighting the adaptability of these architectures. Although MaxViT and the original Vision Transformer model showed strong predictive capabilities, Swin Transformer consistently delivered superior results in the experiments. Deep learning models have revolutionized brain, lung, and kidney tumor classification from MRI and CT, with improved accuracy and efficiency. However, challenges such as computational complexity, generalization, and the need for large annotated datasets remain.

In summary, the detection and classification of tumors have progressed significantly, with deep learning driving the most recent breakthroughs. Future research should focus on developing more efficient models, improving generalization across diverse datasets, and integrating multimodal data. Explainability also remains a critical concern when deploying these models in clinical settings. We will explore input formats beyond images, such as genomic and DNA sequences, to enhance the model's applicability across multi-modal medical data. This expansion would require adapting the current model architectures to accommodate non-image data effectively. Additionally, developing lightweight models for deployment in clinical settings remains an interesting area of research.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] S Deepak and PM Ameer. Brain tumor classification using deep cnn features via transfer learning. *Computers in biology and medicine*, 111:103345, 2019.

[3] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

[4] Meyke Hermsen, Thomas de Bel, Marjolijn Den Boer, Eric J Steenbergen, Jesper Kers, Sandrine Florquin, Joris JTH Roelofs, Mark D Stegall, Mariam P Alexander, Byron H Smith, et al. Deep learning–based histopathologic assessment of kidney tissue. *Journal of the American Society of Nephrology*, 30(10):1968–1979, 2019.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021.

[7] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, June 2022.

[8] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 459–479, Cham, 2022. Springer Nature Switzerland.

[9] Kaggle. Multi cancer dataset. `https://www.kaggle.com/datasets/obulisainaren/multi-cancer/data`, 2022. Accessed: 2018-06-26.

[10] Hamdalla Alyasriy and Muayed AL-Huseiny. The IQ-OTH/NCCD lung cancer dataset. *Mendeley Data, V4*, 2023.

[11] Kaggle. The IQ-OTH/NCCD lung cancer dataset. `https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-dataset`, 2022. Accessed: 2018-06-26.

[12] Kaggle. CT KIDNEY DATASET: Normal-cyst-tumor and stone. `https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone`, 2022. Accessed: 2018-06-26.

[13] Juan Manuel Medina and Javier Sánchez. High accuracy brain tumor classification with EfficientNet and magnetic resonance images. In *5th International Conference on Advances in Signal Processing and Artificial Intelligence*. International Frequency Sensor Association (IFSA) Publishing, SL, 2023.

[14] Daniel Reyes and Javier Sánchez. Performance of convolutional neural networks for the classification of brain tumors using magnetic resonance imaging. *Heliyon*, 10(3):e25468, 2024.

[15] Asifullah Khan, Zunaira Rauf, Abdul Rehman Khan, Saima Rathore, Saddam Hussain Khan, Najmus Saher Shah, Umair Farooq, Hifsa Asif, Aqsa Asif, Umme Zahoora, Rafi Ullah Khalil, Suleman Qamar, Umme Hani Asif, Faiza Babar Khan, Abdul Majid, and Jeonghwan Gwak. A recent survey of vision transformers for medical image segmentation. Technical report, 2023.

[16] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one*, 10(10):e0140381, 2015.

[17] Abdu Gumaei, Mohammad Mehedi Hassan, Md Rafiul Hassan, Abdulhameed Alelaiwi, and Giancarlo Fortino. A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification. *IEEE Access*, 7:36266–36273, 2019.

[18] Mustafa R Ismael and Ikhlas Abdel-Qader. Brain tumor classification via statistical features and back-propagation neural network. In *2018 IEEE international conference on electro/information technology (EIT)*, pages 0252–0257. IEEE, 2018.

[19] Nilesh Bhaskarrao Bahadure, Arun Kumar Ray, and Har Pal Thethi. Comparative approach of mri-based brain tumor segmentation and classification using genetic algorithm. *Journal of digital imaging*, 31:477–489, 2018.

[20] Parnian Afshar, Arash Mohammadi, and Konstantinos N Plataniotis. Brain tumor type classification via capsule networks. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 3129–3133. IEEE, 2018.

[21] Muhammad Sajjad, Salman Khan, Khan Muhammad, Wanqing Wu, Amin Ullah, and Sung Wook Baik. Multi-grade brain tumor classification using deep cnn with extensive data augmentation. *Journal of computational science*, 30:174–182, 2019.

[22] Wadhah Ayadi, Wajdi Elhamzi, Imen Charfi, and Mohamed Atri. Deep cnn for brain tumor classification. *Neural processing letters*, 53:671–700, 2021.

[23] Chetana Srinivas, Nandini Prasad KS, Mohammed Zakariah, Yousef Ajmi Alothaibi, Kamran Shaukat, B Partibane, and Halifa Awal. Deep transfer learning approaches in performance analysis of brain tumor classification using mri images. *Journal of Healthcare Engineering*, 2022(1):3264367, 2022.

[24] Devinder Kumar, Alexander Wong, and David A. Clausi. Lung nodule classification using deep features in ct images. In *2015 12th Conference on Computer and Robot Vision*, pages 133–138, 2015.

[25] R Sathishkumar, K Kalaiarasan, A Prabhakaran, and M Aravind. Detection of lung cancer using svm classifier and knn algorithm. In *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, pages 1–7. IEEE, 2019.

[26] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale convolutional neural networks for lung nodule classification. In Sebastien Ourselin, Daniel C. Alexander, Carl-Fredrik Westin, and M. Jorge Cardoso, editors, *Information Processing in Medical Imaging*, pages 588–599, Cham, 2015. Springer International Publishing.

[27] Xinglong Liu, Fei Hou, Hong Qin, and Aimin Hao. Multi-view multi-scale cnns for lung nodule type classification from ct images. *Pattern Recognition*, 77:262–275, 2018.

[28] Baihua Zhang, Shouliang Qi, Patrice Monkam, Chen Li, Fan Yang, Yu-Dong Yao, and Wei Qian. Ensemble learners of multiple deep cnns for pulmonary nodules classification using ct images. *IEEE Access*, 7:110358–110371, 2019.

[29] Sarfaraz Hussein, Pujan Kandel, Candice W. Bolan, Michael B. Wallace, and Ulas Bagci. Lung and pancreatic tumor characterization in the deep learning era: Novel supervised and unsupervised learning approaches. *IEEE Transactions on Medical Imaging*, 38(8):1777–1787, 2019.

[30] Chenglong Wang, Yun Liu, Fen Wang, Chengxiu Zhang, Yida Wang, Mei Yuan, and Guang Yang. Towards reliable and explainable ai model for pulmonary nodule diagnosis. *Biomedical Signal Processing and Control*, 88:105646, 2024.

[31] D.-Y. Kim and J.-W. Park. Computer-aided detection of kidney tumor on abdominal computed tomography scans. *Acta Radiologica*, 45(7):791–795, 2004.

[32] Bowei Zhou and Li Chen. Atlas-based semi-automatic kidney tumor detection and segmentation in ct images. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1397–1401. IEEE, 2016.

[33] Zhichao Feng, Pengfei Rong, Peng Cao, Qingyu Zhou, Wenwei Zhu, Zhimin Yan, Qianyun Liu, and Wei Wang. Machine learning-based quantitative texture analysis of ct images of small renal masses: differentiation of angiomyolipoma without visible fat from renal cell carcinoma. *European radiology*, 28:1625–1633, 2018.

[34] Cagri Erdim, Aytul Hande Yardimci, Ceyda Turan Bektas, Burak Kocak, Sevim Baykal Koca, Hale Demir, and Ozgur Kilickesmez. Prediction of benign and malignant solid renal masses: machine learning-based ct texture analysis. *Academic radiology*, 27(10):1422–1429, 2020.

[35] Prabod Rathnayaka, Vinoj Jayasundara, Rashmika Nawaratne, Daswin De Silva, Weranja Ranasinghe, and Damminda Alahakoon. Kidney tumor detection using attention based u-net. 2019.

[36] Zhiyong Lin, Yingpu Cui, Jia Liu, Zhaonan Sun, Shuai Ma, Xiaodong Zhang, and Xiaoying Wang. 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. *European Radiology*, 31:5021–5031, 2021.

[37] Dalia Alzu'bi, Malak Abdullah, Ismail Hmeidi, Rami AlAzab, Maha Gharaibeh, Mwaffaq El-Heis, Khaled H Almotairi, Agostino Forestiero, Ahmad MohdAziz Hussein, and Laith Abualigah. Kidney tumor detection and classification based on deep learning approaches: a new dataset in ct scans. *Journal of Healthcare Engineering*, 2022(1):3861161, 2022.

[38] S Phani Praveen, Sampathi Rao Sidharth, Teella Komali Priya, Yaswant Sai Kavuri, Salopantula Mani Sindhura, and Swapna Donepudi. Resnet and resnext-powered kidney tumor detection: A robust approach on a subset of the kauh dataset. In *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pages 749–757. IEEE, 2023.

[39] Erdal Özbay, Feyza Altunbey Özbay, and Farhad Soleimanian Gharehchopogh. Kidney tumor classification on ct images using self-supervised learning. *Computers in Biology and Medicine*, 176:108554, 2024.

[40] Andrey Bogomolov, Urszula Zabarylo, Dmitry Kirsanov, Valeria Belikova, Vladimir Ageev, Iskander Usenov, Vladislav Galyanin, Olaf Minet, Tatiana Sakharova, Georgy Danielyan, et al. Development and testing of an led-based near-infrared sensor for human kidney tumor diagnostics. *Sensors*, 17(8):1914, 2017.

[41] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.