Vision Transformers for Lung Cancer Detection

Dylan Govender
University of KwaZulu-Natal
Durban, South Africa
221040222@stu.ukzn.ac.za

Yuvika Singh
University of KwaZulu-Natal
Durban, South Africa
SinghY1@ukzn.ac.za

Abstract—*Lung cancer is a significant contributor to cancer-related mortality. With recent advancements in Computer Vision, Vision Transformers have gained traction and shown remarkable success in medical image analysis. This study explored the potential of Vision Transformer models (ViT, CVT, CCT ViT, Parallel ViT, Efficient ViT) compared to established state-of-the-art architectures (CNN) for lung cancer detection via medical imaging modalities, including CT and Histopathological scans. This work evaluated the impact of data availability and different training approaches on model performance. The training approaches included but were not limited to, Supervised Learning and Transfer Learning. Established evaluation metrics such as accuracy, recall, precision, F1-score, and area under the ROC curve (AUC-ROC) assessed model performance in terms of detection efficacy, data validity, and computational efficiency. ViT achieved an accuracy of 94% on a balanced dataset and an accuracy of 87% on an imbalanced dataset trained from the ground up. Cost-sensitive evaluation metrics, such as cost matrix and weighted loss, analysed model performance by considering the real-world implications of different types of errors, especially in cases where misdiagnosing a cancer case is far more critical.*

Index Terms—Accuracy, Area Under the ROC Curve (AUC-ROC), Balanced Dataset, Cancer-Related Mortality, CNN, Computational Efficiency, Cost Matrix, Cost-Sensitive Evaluation Metrics, CT Scans, CVT, Data Availability, Data Validity, Detection Efficacy, Efficient ViT, Evaluation Metrics, F1-Score, Histopathological Scans, Imbalanced Dataset, Lung Cancer Detection, Medical Image Analysis, Misdiagnosis, Parallel ViT, Precision, Real-World Implications, Recall, Supervised Learning, Transfer Learning, Vision Transformer (ViT), Weighted Loss

1. INTRODUCTION

Cancer is the leading cause of death worldwide, with lung cancer being one of the most severe causes of cancer-related mortality, according to the World Health Organisation (WHO) [1]. Detection of cancer at an early stage can significantly reduce the mortality rate. Early diagnosis and screening are two commonly used methods for the detection of cancer [2]. The conventional use of early diagnosis required the expertise of diagnostic radiologists, which were time-consuming, susceptible to human error, and affected by subjective interpretation, leading to variability between different diagnostic radiologists (inter-observer) and within the same radiologist over time (intra-observer) [2, 3].

Medical imaging technologies are improving and assisting in lung cancer detection. However, there is still an immediate need to minimise variability and the possibility of error due to human-driven diagnostic processes and improve the accuracy of emerging imaging technologies. Incorporating advanced automated tools is vital in improving early cancer detection in medical imaging [2, 3, 4].

With the development of Artificial Intelligence (AI), medical imaging analysis has witnessed a steady improvement, offering significant support in cancer detection and prevention. Machine Learning (ML), a subset of AI, encompasses a set of algorithms capable of identifying patterns within data, making them ideal for image classification. Deep Learning, a powerful branch of Machine Learning, leverages large amounts of data to mimic the human brain’s learning and improvement capabilities. Deep Learning architectures, particularly Convolutional Neural Networks (CNN) [4], have shown promise in computer-aided cancer detection (CADx) [5] and have achieved state-of-the-art performance in image analysis. Another deep-learning architecture, Transformer [6], is a self-attention-based mechanism prominently used for Natural Language Processing (NLP) and later adapted for Computer Vision, leading to the Vision Transformer’s (ViT) development [7]. ViT, the new deep-learning architecture, has emerged as a potent tool for medical image analysis, particularly for image classification, offering a new horizon of possibilities in this field.

This paper has several sections and subsections that introduce and explain the techniques and methods used to train

Vision Transformer models from the ground up for image classification to achieve state-of-the-art performance.

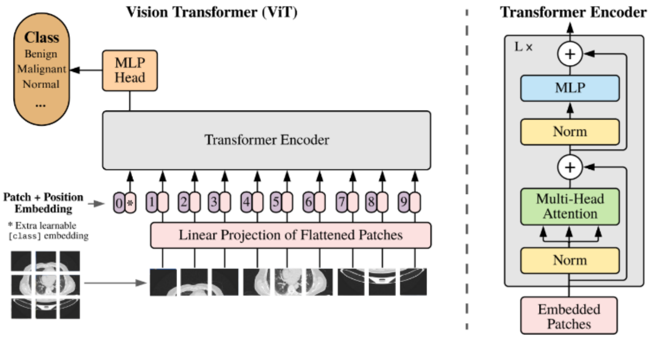


Figure 1: Illustration adapted from Vaswani et al. and Dosovitskiy et al. and shows a model overview of the Vision Transformer model for lung cancer detection [6, 7]. The input image, a CT scan image of a chest, is split into size-dependent patches (16px or 32px), and the sequence is linearly embedded. Each patch is concatenated to a position using the position embedding. An extra learnable [class] token is added to the sequence for classification.

2. BACKGROUND

The Vision Transformer architecture was proposed by Dosovitskiy et al. (2021) [7]. Different ViT architectures have been developed and deployed in many image classification and segmentation applications [8]. This research compared the performance of different ViT architectures with other state-of-the-art deep-learning architectures, such as CNN, for lung cancer detection via different medical imaging modalities. It focuses on image classification, aiming to classify three medical images: malignant images containing lung cancer, normal pictures with no signs of cancer, and benign images with suspicious lung abnormalities that are not cancerous, such as benign tumours or lung infections. This research aims to answer whether Vision Transformer models can achieve superior or competitive performance for lung cancer detection compared to existing state-of-the-art architectures and whether these models can use different imaging modalities while maintaining detection efficacy. This study experiments with and evaluates various training approaches to investigate which can maximise performance while also assessing the impact of data availability on model effectiveness.

In experimentation, implementing different Vision Transformer models can ensure a sounder study. This research implemented five Vision Transformer models (ViT, CVT, CCT ViT, Parallel ViT, and Efficient ViT) from the ground up and selected models explicitly finetuned for lung cancer detection. Selected models were trained on limited or abundant data to assess the effectiveness of data availability on model performance due to the steep cost of data quality and availability in the medical field—different image modalities assessed whether the Vision Transformer models are flexible [9]. The training approaches for each model included but were not limited to Supervised Learning and Transfer Learning [10, 11]. Established evaluations [12] for evaluating these models included but were not limited to accuracy, recall, precision, F1-score and Area Under the Receiver Operating Curve (AUC-ROC) [13], which assessed detection efficacy, data validity, and computational efficiency. Cost-sensitive evaluation metrics such as cost matrix and weighted loss analysed model performance by considering the real-world implications of different types of errors [14].

By determining the most effective ViT architectures combined with different training approaches, this research will help improve the accuracy of cancer detection, especially in cases where misdiagnosing a cancer case is far more critical. It will ultimately advance the role of Vision Transformers in medical imaging and build a foundation for more reliable AI-assisted healthcare solutions.

3. LITERATURE REVIEW & RELATED WORK

Deep Learning has paved the way for medical image analysis, including lung cancer detection [14]. Convolutional Neural Networks (CNN) are the leading state-of-the-art architecture for image analysis [4]. CNN proves to be better than ViT when data is insufficient, according to a study by Gai et al. [15]. However, sufficient data is necessary to improve model performance, especially for deep-learning architectures like CNN and ViT, an ongoing challenge in the medical domain [1]. However, with the recent improvements in ML techniques, Transfer Learning has proven helpful for insufficient data usage in model training [11]. Transfer Learning is an approach that is poorly adopted in medical research but is advantageous when we explore the usual scarcity or cost that comes with safely procuring medical data. Kim et al. demonstrated the effectiveness of Transfer Learning in data-scarce medical environments [9]. ViT architectures are models primarily constructed to take advantage of Transfer Learning, where we optimise source data for a more specific task to exuberate performance for that particular task [14]. In experimentation, using minimum data to expand on the performance of attention-based architectures can show that they are flexible and able to achieve state-of-the-performance.

Gai et al. (2023) proposed similar research based on different training approaches for improving model performance, focusing on CT scans, overlooking the various imaging modalities' affordability and broader availability [16]. However, outstanding results were achieved based on model performance but lacked an account of using different imaging modalities or comparing hybrid architectures. Using multiple data types can assess the generalisation and robustness of a model. Cost-sensitive evaluation metrics are usually underutilised, crucial in medical analysis (misdiagnosis of a cancer case), and discounted from established evaluation metrics. A comprehensive analysis can ensure a higher efficacy in cancer detection by employing cost-sensitive metrics and using different data types. This research will also account for the misclassification and availability of medical data, where one class has a higher frequency ratio in the data; cost-sensitive metrics can analyse the implication of misclassification and which model can minimise the effect of missing cancer cases.

Xiaoyan et al. (2024) gave a detailed account of the applications and advancements of Vision Transformers (ViT) in cancer diagnosis [17]. It reviews 98 articles published since 2020, sourced from various research databases. The paper explores the various improved ViT models and their various applications. The review highlights the advantages of utilising Vision Transformer for cancer diagnosis. The paper suggests that future research should focus on more training approaches that can accompany the various ViT architectures, with a keen focus on self-and semi-supervised training approaches. This work will focus on multiple training approaches utilising cost-effective methods for improved efficiency and model performance.

Tran et al. (2021) discussed the recent developments of Vision Transformers in lung cancer detection and provided a comprehensive overview of how deep learning approaches can assist in image analysis [15]. Although this research provided the disadvantages and advantages of Vision Transformers, the study lacked data augmentation and analysis on model performance. Data augmentation techniques expand the training data and can increase feature extraction and classification ability. This study aims to optimise the efficacy of cancer detection using attention-based architectures and different data augmentation techniques to account for the various problems in the medical domain, such as data availability, data types, and missing data.

Demireriden et al. (2023) compared the various ViT architectures for general image classification [8]. Although this comparison analysis investigates the potential of ViT using different training approaches, validity is overlooked by not comparing ViT to state-of-the-art architectures. Using state-of-the-art architectures in this research acts as a control and can be compared to existing work, ensuring validity.

Khan et al. (2023) synthesised existing literature from 2020 to 2022 to see the utilisation of Vision Transformers for skin cancer detection using dermoscopy images [18]. ViT has demonstrated outstanding performance in diagnosing melanocytic lesions, which are challenging to identify due to their variable appearance. The research focused on the intrinsic visual ambiguities that pose a risk for Vision Transformers in clinical environments and highlights the best segmentation techniques that enhance melanoma diagnosis but require ViT-based approaches to improve diagnostic authenticity and decision-making in clinical practice. This research aims to improve diagnostic authenticity by minimising errors based on real-world consequences using cost-sensitive evaluations.

Hong et al. (2024) highlighted the efficacy of combining Low-Rank Adaptation techniques with ViT to enhance the detection of cervical cancer [19]. Limiting the data used in training overcomes data scarcity in medical imaging and can demonstrate the effectiveness of these architectures. Residual Networks (ResNets) and the LoRA-ViT architecture show improved generalisation and performance ability in data-limited environments, explaining that this approach achieves high accuracy in identifying cervical cancer and the innovative adaptation of improving model performance to overcome challenges in image analysis.

Ayana et al. (2024) presented an advanced colorectal cancer (CRC) detection approach using Vision and Spatial Transformers [20]. Traditional Methods primarily focused on Convolution Neural Networks (CNNs), which were limited in capturing global features from images. The research demonstrates excellent performance in detecting colorectal cancer by combining spatial and vision transformers for localising image features. With a focus on Vision Transformers, the study utilises different training approaches and preprocessing techniques that can maximise capturing local and global image features.

This work adds to the increasing source of research exploring image analysis and recognition at a more efficient scale than usual. This study examines the possibility of Vision Transformers exceeding the threshold in state-of-the-art performance, hence using additional models and datasets. Two specific datasets, balanced and imbalanced, each containing three classes, explore the effectiveness of these attention-based mechanisms in minimising misclassification and errors. Misclassification, emphasised using an imbalanced dataset, investigates whether attention-based mechanisms can classify on limited data while retaining performance in terms of detection efficacy. Using a single transformer encoder for each model can demonstrate a simple way of achieving state-of-the-art performance compared to convolution-based architectures. In terms of computational cost and efficiency, this study aims to use limited resources while optimising performance for each model during experimentation.

4. METHODS

4.1 VISION TRANSFORMER MODEL (VIT)

This model was based on the Transformer model by Vaswani et al., which has an encoder-decoder structure that explicitly operates on sequences and sets [6, 7]. This model is an over-generalised multi-layer perceptron (MLP), or feed-forward network, primarily used for Natural Language Processing and later adapted for Computer Vision [6, 7]. The architecture calculates the pairwise inner product of each pair of data in the set, called the attention, hence the attention-based mechanism, each data attends to another. Instead of using local attention over a kernel of pixels like the convolutional-based architectures, Vision Transformers use global attention, which requires the input image to be divided into patches. These patches are placed into a set for linear projection and then embedded with positions. An extra learnable [class] token is added to the sequence for image classification. The Transformer encoder in *Figure 1* is a standard Transformer encoder by Vaswani et al. [6, 7]. The architecture finally leads to a standard classifier.

In experimentation, the Vision Transformer models were built from the ground up, using a single transformer encoder to demonstrate a simple way of achieving state-of-the-art performance compared to convolution-based architectures.

4.2 CVT Model (Convolutional ViT)

This Transformer Model hybrid architecture combines convolutions with attention [21]. Specifically, the convolutions embed and down-sample the image or feature map in three stages. Depth-wise convolutions project the attention mechanism's queries, keys, and values [21].

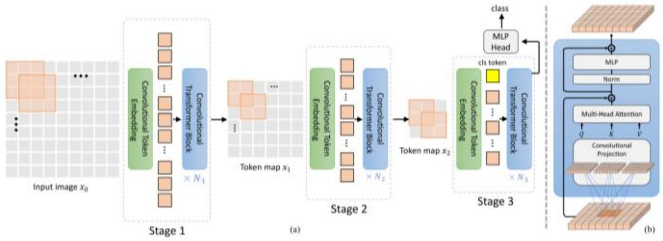


Figure 2: Illustration adapted from Wu et al., which illustrates the pipeline of the CVT architecture [21]. (a) Illustrates the hierarchical multi-stage structure facilitated by the convolutional token embedding layer. (b) Illustrates the convolutional transformer block, which contains the convolution projection as the first layer.

4.3 CCT Model (Compact Convolutional Transformer)

This model utilises compact transformers, which use convolutions instead of patching to perform sequence pooling [32]. This allows it to have fewer parameters while maintaining high performance and achieving high accuracy. It also has a low training time, making it ideal for quick deployment [32].

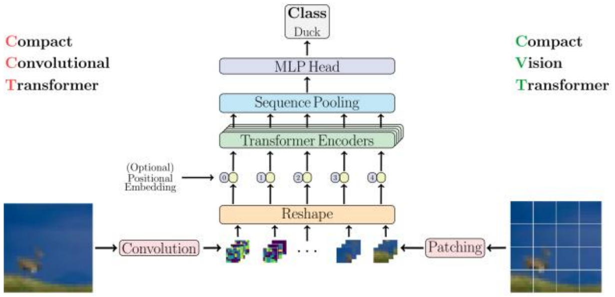


Figure 3: Illustrates the pipeline for the CCT model by Hassani et al.—the CCT Model performs well on small datasets and trains quickly [32]. On the left is the CCT Model, and on the right is the CVT Model [21, 32].

4.4 Parallel ViT Model

Touvron et al. proposed a model that can parallelise multi-headed self-attention and residual feedforward networks, making the ViT model more cost-efficient during training [22].

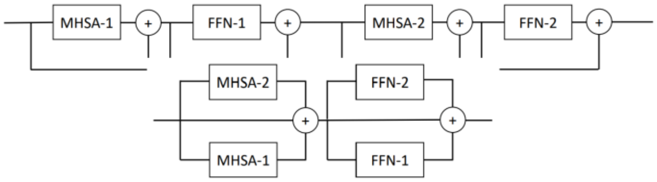


Figure 4: Illustration adapted from Touvron et al. and illustrates the pipeline for the Parallel Vision Transformer. MHSA denotes the multi-headed self-attention residual block, and FFN is the residual feedforward network [22].

4.5 Efficient ViT Model

Vision Transformers suffer from quadratic computational costs due to the self-attention mechanism on the input sequence length, limiting its application to longer sequences. Using efficient self-attention algorithms, like Nystromformer by Xiong et al. or Linformer by Wang et al., can reduce the quadratic complexity, optimising the self-attention mechanism for any Vision Transformer architecture [23, 24].

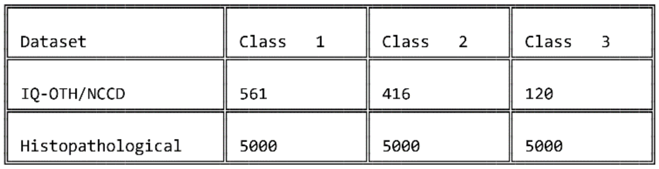
4.6 CNN Model (Convolutional Neural Network)

State-of-the-art image analysis architecture [16]. This research uses state-of-the-art architectures as a control and can be compared to existing work, ensuring validity. The Convolutional Neural Network ensures excellent performance on imaging applications and is used as a benchmark for a fair comparison of each architecture's overall performance [16].

5. TECHNIQUES

5.1 SETUP

Two datasets provided the training data for each model. IQ-OTH/NCCD, provided by Alyasriy et al., was used to train most models due to the imbalanced ratio of classes it offers [25]. IQ-OTH/NCCD dataset provides CT scan images of three classes: normal, malignant, and benign. The second dataset is a perfectly balanced large dataset using histopathological lung cancer images with three defined classes provided by Borkowski et al. [26].



Dataset	Class 1	Class 2	Class 3
IQ-OTH/NCCD	561	416	120
Histopathological	5000	5000	5000

Figure 5: Provides the size of each class in the IQ-OTH/NCCD and Histopathological dataset [25, 26].

5.2 PREPROCESSING THE DATA

The image dataset had to be normalised and converted to a suitable format. Normalisation and standardisation limited the range of each pixel value in each image contained in the dataset [27]. These techniques ensure that the pixel values are similar and can contribute equally to training the model.

5.2.1 IMAGE AUGMENTATION

Expands the dataset by creating variations of existing images [28]. It helps to model features that generalise better on unseen data. It prevents overfitting when a model familiarises itself too well with the training data, resulting in a model that inaccurately predicts unseen data that differs from the training data. These data augmentation techniques are used: random cropping and flipping to create image size and orientation variations, rotation and scaling to simulate different viewing angles and image sizes, and random brightness and contrast adjustments to develop variations in image lighting conditions [28].

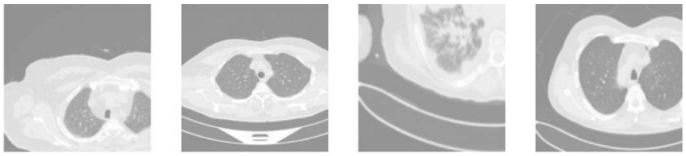
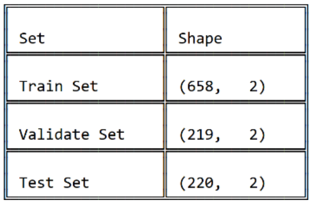


Figure 6: After data augmentation, each lung cancer image from the IQ-OTH/NCCD dataset was randomised in size, flipped, and cropped, which gave the model more data to train on [25].

5.2.2 TRAIN, TEST, SPLIT, AND VALIDATE

A popular preprocessing technique involves splitting the dataset to ensure enough data is available to complete the model evaluation. This provides model robustness and generalisation and allows for a more straightforward model evaluation process.



Set	Shape
Train Set	(658, 2)
Validate Set	(219, 2)
Test Set	(220, 2)

Figure 7: Tabulates the dataset size after cross-validation on the IQ-OTH/NCCD dataset [25].

5.3 TRAINING APPROACHES

Training techniques and approaches are a strategy for training a model. They usually encompass algorithms and processes for adjusting a model's parameters/hyperparameters to learn from the training data and make an accurate decision or prediction. This research focuses on two training approaches commonly used for image classification tasks.

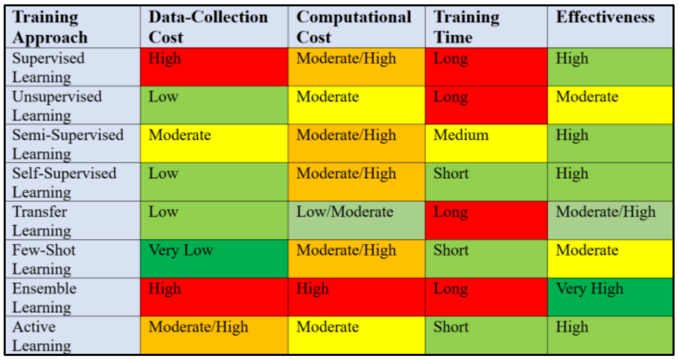
5.3.1 SUPERVISED LEARNING

Supervised Learning is a learning approach that uses a labelled training set to train a model [29]. Each image in the lung cancer detection training set is associated with a label the model must learn [29].

5.3.2 TRANSFER LEARNING

This technique was introduced by Virginia et al. [11]. Transfer Learning is a powerful training approach in Machine Learning in which knowledge gained from a pre-trained model on a source task is applied and configured to a specific task. Transfer learning is a cost-effective approach where data availability is limited [9, 11].

Tabulated summary of various training approaches with their pros and cons.



Training Approach	Data-Collection Cost	Computational Cost	Training Time	Effectiveness
Supervised Learning	High	Moderate/High	Long	High
Unsupervised Learning	Low	Moderate	Long	Moderate
Semi-Supervised Learning	Moderate	Moderate/High	Medium	High
Self-Supervised Learning	Low	Moderate/High	Short	High
Transfer Learning	Low	Low/Moderate	Long	Moderate/High
Few-Shot Learning	Very Low	Moderate/High	Short	Moderate
Ensemble Learning	High	High	Long	Very High
Active Learning	Moderate/High	Moderate	Short	High

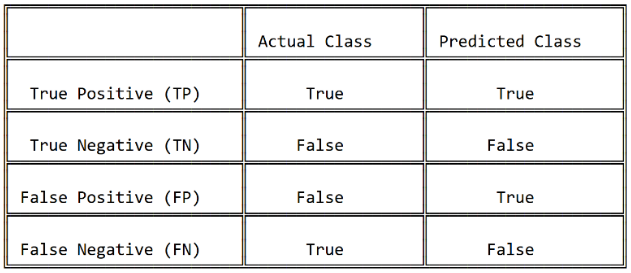
Figure 8: Tabulated summary of different training approaches considering their pros and cons [9, 10, 11, 29, 34]. The colour approximates the relative cost that arises with each training approach. Red signifies very costly, and dark green is less costly when describing the cost associated with each attribute. The other colours vary from very costly to less costly, respectively.

6. EVALUATION

6.1 Established Evaluation Metrics

Each model was evaluated using established evaluations such as accuracy, recall, precision, F1-score, and area under the ROC (AUC-ROC) [12, 13]. Established evaluation metrics offer a standardised comparison and can assess multiple architectures regardless of complexity, ensuring model fairness. Established evaluation promote comparison between existing work. Each evaluation metric is based on Binary Classification and later adapted for Multi-class Classification. To derive the following evaluation metrics, we can assume an image is represented as a patient in computation. Let the case where a patient has cancer be denoted as True, and the case where a patient does not have cancer be denoted as False. The actual class, y, is the ground truth label determined through expert annotation, clinical diagnosis, or another reliable source. The predicted class, $\hat{y}$, is the prediction label the model assigns based on its decision-making process. Based on the model's prediction and outcome, we can derive discriminator metrics [12].

These will count each instance where: If a patient has cancer (y is True) and the model predicts that the patient does have cancer ($\hat{y}$ is True), this is the True Positive (TP) discriminator; if a patient does not have cancer (y is False) and the model predicts that the patient does not have cancer ($\hat{y}$ is False), this is the True Negative (TN) discriminator; if a patient does not have cancer (y is False), the model predicts that the patient does ($\hat{y}$ is True), this is the False Positive (FP) discriminator; if a patient has cancer (y is True) and the model predicts that the patient does not ($\hat{y}$ is False), this is the False Negative (FN) discriminator, which needs to be minimised in computation.



	Actual Class	Predicted Class
True Positive (TP)	True	True
True Negative (TN)	False	False
False Positive (FP)	False	True
False Negative (FN)	True	False

Figure 9: Illustrates the discriminator metrics in a tabulated and summarised format [12, 13].

These established evaluation metrics are defined from the above discriminator metrics. Each evaluation metric measures the performance of how well a model can achieve a particular result. These metrics are called established evaluation metrics because they can be used to measure other models outside the scope of lung cancer detection and act as a benchmark for a fairer comparison analysis of the performance of each architecture.

6.1.1 Accuracy

Accuracy measures the ratio of the correct number of predictions over the total number of predictions evaluated. Accuracy evaluates the overall percentage of accurate predictions. Accuracy is calculated as follows:

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + FP + TN + FN}}$$

6.1.2 Precision

Precision calculates the number of positive instances correctly predicted from the total number of positive predictions made by the model [12]. Precision evaluates whether a model could identify relevant cases of the data [13]. Precision, established as the Positive Predicted Value, is calculated as follows:

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

6.1.3 Recall

Recall calculates the number of positive instances correctly predicted by the model from the total number of positive cases in the test set [12]. Recall evaluates whether a model could capture relevant cases of the data. Established as the Negative Predicted Value [13]. Recall is calculated as follows:

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

6.1.4 F1-Score

F1-Score calculates the harmonic mean between recall and precision. F1-Score evaluates whether a model was able to capture and identify relevant instances of the data.

$$\boldsymbol{F1\text{–}Score = \frac{2 * Precision * Recall}{Precision + Recall}}$$

6.1.5 Area Under the ROC Curve (AUC-ROC)

The following equations must be introduced to establish the foundation for AUC-ROC.

6.1.5.1 Sensitivity

Sensitivity [12] calculates the fraction of positive instances correctly predicted from the total number of positive cases, known as the True Positive Fraction [13]. Sensitivity determines whether a model can capture positive cases. Sensitivity is calculated as follows:

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

6.1.5.2 1-Sensitivity

Known as the False Negative Fraction [13]. 1-Sensitivity is calculated as follows:

$$\boldsymbol{1\text{–}Sensitivity} = \frac{\boldsymbol{FN}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

6.1.5.3 Specificity

Specificity [12] calculates the fraction of negative instances correctly predicted from the total number of negative cases, known as the True Negative Fraction [13]. Specificity determines whether a model can capture negative cases. Specificity is calculated as follows:

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN} + \boldsymbol{FP}}$$

6.1.5.4 1-Specificity

Known as the False Positive Fraction [13]. 1-Specificity is calculated as follows:

$$\boldsymbol{1\text{–}Specificity} = \frac{\boldsymbol{FP}}{\boldsymbol{TN} + \boldsymbol{FP}}$$

AUC-ROC quantifies discrimination between classes [13]. The ROC curve evaluates whether a model can distinguish between these classes when they overlap. The ROC curve plots the distribution of the True Positive Fraction (dependent axis) and the False Positive Fraction (independent axis). The trade-off between these distributions is evaluated. If these distributions overlap entirely, it implies no discrimination in the evaluation, whilst a complete separation of these distributions implies a perfectly discriminating evaluation [13]. The AUC determines the overall ranking performance of a classifier [12].

6.2 Cost-Sensitive Evaluation Metrics

In lung cancer detection, misdiagnosing a patient who has cancer (False Negative) is far more critical than misdiagnosing a patient without cancer (False Positive) [9, 14, 17]. Cost-sensitive metrics will help analyse which models minimised False Negatives. False negatives may occur due to class imbalance when a class has a higher frequency ratio in the corpus. A more balanced dataset may help with class imbalance, but medical data is usually scarce and often limited [14]. For Cost-Sensitive Evaluation on Multi-class Classification, the following metrics will be used:

6.2.1 Real-World Weight Cross-Entropy Loss

This evaluation is the real-world weight that models the cost of misclassification, allowing the model to account for the class with a lower representation in the dataset [30]. The standard weighted binary cross-entropy loss function is given by:

$$\boldsymbol{J_{wbce}} = -\frac{\mathbf{1}}{\boldsymbol{M}}\sum_{\boldsymbol{m}=\mathbf{1}}^{\boldsymbol{M}}\left[\boldsymbol{w} \times \boldsymbol{y_m} \times \mathbf{log}\left(\boldsymbol{h_\theta}(\boldsymbol{x_m})\right)\right] + (\mathbf{1} - \boldsymbol{y_m}) \times \mathbf{log}\left(\mathbf{1} - \boldsymbol{h_\theta}(\boldsymbol{x_m})\right)$$

Where $\boldsymbol{M}$ is the number of training samples, $\boldsymbol{w}$ is weight, $\boldsymbol{y_m}$ is target label for training sample $\boldsymbol{m}$, $\boldsymbol{x_m}$ is the input for training sample $\boldsymbol{m}$, and $\boldsymbol{h_\theta}$ is the model with neural network weights $\boldsymbol{\Theta}$.

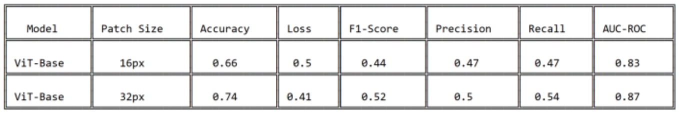
6.2.2 Cost-Sensitive Matrix

The Cost-Sensitive Matrix measures the probability of misclassification given a class. It helps understand the trade-offs between errors, which may entail real-world consequences [14].

7. RESULTS

7.1 Patch Size

The first result computed was the hyperparameter, the patch size, which contributed to the overall performance of the Vision Transformer model. Vision Transformers use global attention, which requires the input image to be divided into patches [6, 7]. The best patch size for this experiment was 32px. Using a larger patch size, in this case 32px, takes advantage of the global self-attention mechanism [6, 7].

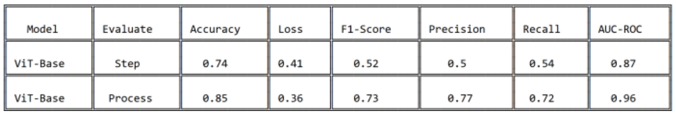


Model	Patch Size	Accuracy	Loss	F1-Score	Precision	Recall	AUC-ROC
ViT-Base	16px	0.66	0.5	0.44	0.47	0.47	0.83
ViT-Base	32px	0.74	0.41	0.52	0.5	0.54	0.87

Figure 10: The experiment results with the different patch size hyperparameters.

7.2 Loss Function

While fitting a model with specific data, it evaluates itself during three steps using the loss function to see whether it is improving [31]. These steps are the training step, the validation step, and the test step. Using a loss function for each step seems counter-intuitive, but it can be done to see whether the model evaluates itself correctly during each step. The next set of results shows the difference between using a loss function for each step and a loss function for the entire process during training and evaluation. Using a loss function during training as the whole evaluation process can show where the model can improve overall, ensuring better performance.



Model	Evaluate	Accuracy	Loss	F1-Score	Precision	Recall	AUC-ROC
ViT-Base	Step	0.74	0.41	0.52	0.5	0.54	0.87
ViT-Base	Process	0.85	0.36	0.73	0.77	0.72	0.96

Figure 11: The experiment results using the loss function in each step or process.

7.3 Weighted Loss Function

While training a model, loss functions can be weighted or unweighted to handle misclassification [30]. In medical datasets, data is usually scarce and unavailable [9]. This limited availability often leads to imbalanced distributions of images, which can result in biased or inaccurate model performance. For instance, if one class (Normal) significantly outnumbers other classes (Malignant or Benign), a model may struggle to detect the minority classes accurately (in this case, Malignant), leading to poor generalisation. This issue is particularly problematic in medical image classification tasks, where accurate detection of rare cases, such as early-stage cancer, is critical for diagnosis and treatment.

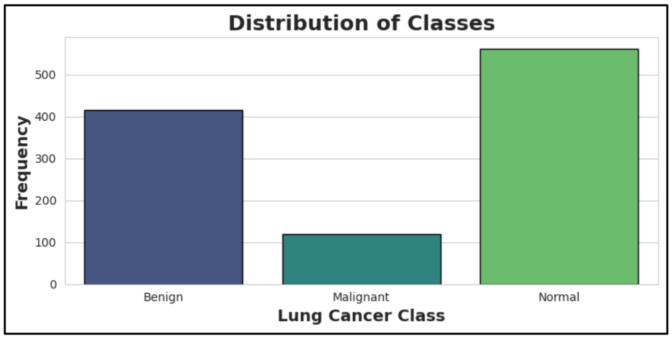
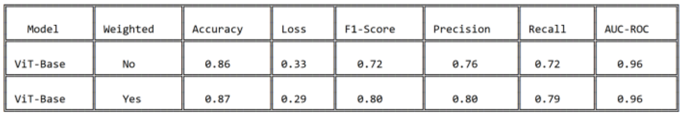


Figure 12: Illustrates the distribution and imbalance of classes in the IQ-OTH/NCCD dataset [25].

Weighted loss function can give higher importance to minority classes, accounting for the above imbalance. The weights were calculated by taking the inverse of each class's frequency, ensuring that classes with fewer samples receive higher weights and, hence, higher importance, and then averaging the weights across training, validation, and test sets to create a universal weight. The results below show the difference between weighted and unweighted loss functions. Using a weighted loss function ensures that the model achieves higher performance.



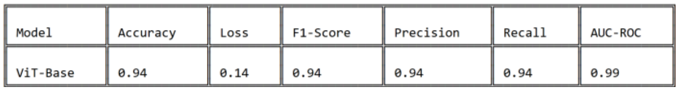
Model	Weighted	Accuracy	Loss	F1-Score	Precision	Recall	AUC-ROC
ViT-Base	No	0.86	0.33	0.72	0.76	0.72	0.96
ViT-Base	Yes	0.87	0.29	0.80	0.80	0.79	0.96

Figure 13: Illustrates the results of using a weighted and unweighted loss function.

The above results show the process of going from subpar to state-of-the-art performance by implementing small changes to maximise the overall performance. These results were saved, and the parameters were distributed for each other model.

7.4 Balanced Dataset

Vision Transformers are architectures that thrive in data-abundant environments. Using the same architecture with a new, perfectly balanced dataset can achieve state-of-the-art performance with fewer computational resources. The results below show the ViT-Base after training using five epochs.



Model	Accuracy	Loss	F1-Score	Precision	Recall	AUC-ROC
ViT-Base	0.94	0.14	0.94	0.94	0.94	0.99

Figure 14: Illustrates the results of the ViT-Base model after using a balanced dataset.

The above model was trained on Histopathological lung cancer images with a perfectly balanced distribution of classes [26]. The distribution graph of the dataset is below.

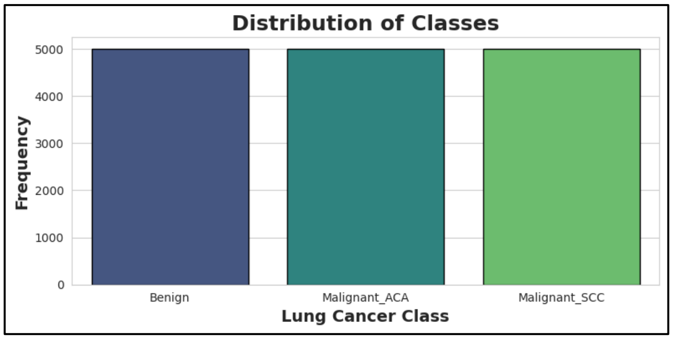
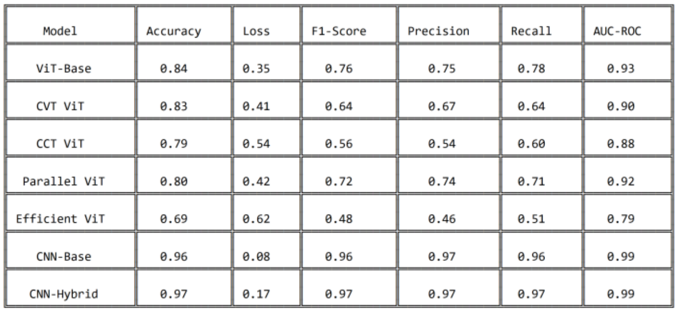


Figure 15: Illustrates the distribution graph of the Histopathological dataset by Borkowski et al. [26].

7.5 Vision Transformers and State-of-the-Art Architecture Results

The below illustrates the results of each model trained on the IQ-OTH/NCCD dataset from the ground up [25]. Each model used the same hyperparameters and parameters except for the CNN models, which were trained on 50 epochs instead of 20. The criterion or loss function used for the CNN model was unweighted.



Model	Accuracy	Loss	F1-Score	Precision	Recall	AUC-ROC
ViT-Base	0.84	0.35	0.76	0.75	0.78	0.93
CVT ViT	0.83	0.41	0.64	0.67	0.64	0.90
CCT ViT	0.79	0.54	0.56	0.54	0.60	0.88
Parallel ViT	0.80	0.42	0.72	0.74	0.71	0.92
Efficient ViT	0.69	0.62	0.48	0.46	0.51	0.79
CNN-Base	0.96	0.08	0.96	0.97	0.96	0.99
CNN-Hybrid	0.97	0.17	0.97	0.97	0.97	0.99

Figure 16: Illustrates the performances of various Vision Transformers and state-of-the-art models trained on the IQ-OTH/NCCD dataset from the ground up [25].

The above results showcase that the Vision Transformer model can achieve state-of-the-art performance using only one transformer encoder. Combining convolutions and Vision Transformers (CVT, CCT) is a noteworthy innovation but can be revisited for its computational complexity [21, 22]. The idea behind combining convolutions with Vision Transformers is to capture localised and globalised attention in the data, improving performance [21, 22]. Vision Transformer models are based on simplicity, which makes them perform exceptionally well on downstream tasks like classification [6, 7]. The more complex the model is, the more computational resources are required to train it.

The Efficient Vision Transformer, which is supposed to optimise the quadratic complexity of the self-attention mechanism, led to poorer results than expected [22, 23, 32]. This could be due to insufficient data and the more complex architecture. Optimising a simple architecture can make it more complicated; hence, computational resources may be needed to achieve a higher performance. The average loss is also reasonably high, meaning the model could have been trained longer and required more computing time to generalise the data well [23, 24].

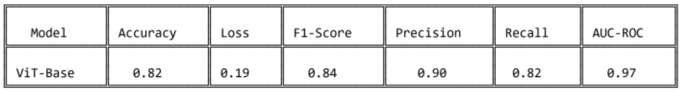
Parallel ViT has shown exceptional performance in training time and evaluation metrics. Parallelising multi-head attention can achieve state-of-the-art performance while significantly reducing the cost of computation [22].

7.6 Comparison to state-of-the-art

The Vision Transformer model was compared to state-of-the-art CNN models (CNN-Base and CNN-Hybrid). Although the CNN achieved high accuracy, the Vision Transformer models achieved high performance using fewer computational resources. Each Vision Transformer model achieved an average accuracy of 80% with fewer parameters and compute time. Compared with each Vision Transformer model, which required a training time of 20 epochs, CNN required a training time of 50. This demonstrates that the multi-head self-attention architecture could generalise and converge much faster to the actual solution than the convolutional architecture.

7.7 Transfer Learning with Vision Transformers

Transfer Learning is vital in data-scarce environments. The ViT-Base architecture, fine-tuned on the IQ-OTH/NCCD dataset, achieved state-of-the-art performance [25, 33]. The results below show the performance of the fine-tuned Vision Transformer Model.



Model	Accuracy	Loss	F1-Score	Precision	Recall	AUC-ROC
ViT-Base	0.82	0.19	0.84	0.90	0.82	0.97

Figure 17: Illustrates the performance of the fine-tuned Vision Transformer model.

Google Research developed the model vit-base-patch16-224-in21k, fine-tuned for lung cancer images [33]. It achieved exceptional precision and AUC-ROC performance, demonstrating the model's ability to capture relevant data instances. Ideally, the model fine-tuned on more significant amounts of data and computational resources can improve detection efficacy and overall performance.

8. FUTURE WORK

Using more computational resources and accessing more extensive datasets can contribute to much sounder results, ensuring validity. Training these models on large datasets can improve results and validate their advantages. This research can be expanded by accessing the attention mechanism for visualisation to see which model's mechanism can be optimised for improved performance.

The Efficient ViT can be trained on larger datasets, improving its overall performance instead of small datasets [22, 23, 32]. This study used smaller datasets to see which model can generalise on limited data [32]. A further exploration of algorithms that can accompany Vision Transformers should be implemented. DINO, self-distillation with no labels, a self-supervised learning algorithm, can infer state-of-the-art performance from ordinary training approaches [10, 34]. We can use a variety of training approaches or a combination of them to see which can lead to more outstanding performance [10, 34].

Accessing the embedding from these Vision Transformer models can infer insight about feature representation and similarity between images [6, 7]. Future work will focus on understanding the embedding of Vision Transformer Models.

These models can be more efficient by employing algorithms that can optimise the quadratic self-attention mechanism, such as Nystromformer by Xiong et al. or Linformer by Wang et al., which can reduce the quadratic complexity, optimising the self-attention mechanism for any Vision Transformer architecture [23, 24]. This study can be further expanded by focusing on optimising the self-attention mechanism of the Vision Transformer model for computational efficiency [23, 24].

9. CONCLUSIONS

In this work, we explored the Vision Transformer, the first sequence transduction model based entirely on attention. It replaces the recurrent layers most used in encoder-decoder architectures with multi-headed self-attention. These models have achieved state-of-the-art lung cancer image classification performance, outperforming some conventional image classification models. The Vision Transformer base model with only a single transformer encoder has achieved an above average of 80% in performance metrics. These models show that simplifying existing Machine Learning architectures can vastly improve performance while being computationally efficient.

10. REFERENCES

[1] Zhang B, Shi H, Wang H. Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. PMCID: PMC10312208, 2023.

[2] Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. PMCID: PMC8477474, 2021.

[3] Hanaa Al-Khawari, Reji P. Athyal, Osama Al-Saeed, Prio N. Sada, Sana Al-Muthairi, and Adel Al-Awadhi. Ann Saudi Med. Inter- and intraobserver variation between radiologists in the detecting abnormal parenchymal lung changes on high-resolution computed tomography. PMCID: PMC2855063, 2010.

[4] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. DOI: s40537-021-00444-8, 2021.

[5] Firmino M, Angelo G, Morais H, Dantas MR, Valentim R. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. Biomed Eng Online. PMCID: PMC5002110, 2016.

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017.

[7] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.

[8] Demireriden, Asuman & Gumus, Abdurrahman. Comparative Analysis of Vision Transformer-Based Architectures for Image Classification. 2023.

[9] Hee, Kim & Cosa, Alejandro & Santhanam, Nandhini & Jannesari, Mahboubeh & Maros, Mate & Ganslandt, Thomas. (2022). Transfer learning for medical image classification: a literature review. BMC Medical Imaging. 22. 10.1186/s12880-022-00793-7.

[10] Gui J, Chen T, Zhang J, Cao Q, Sun Z, Luo H, Tao D. A survey on self-supervised learning: Algorithms, applications, and future trends. arXiv preprint arXiv:2301.05712. 2023.

[11] Cowan, In & G, Tesauro & de Sa, Virginia. (1997). Learning Classification with Unlabeled Data.

[12] Hossin, Mohammad & M.N, Sulaiman. A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201, 2015.

[13] Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. PMCID: PMC3755824, 2013.

[14] Ibomoiye Domor Mienye, Yanxia Sun, Performance analysis of cost-sensitive learning methods with application to imbalanced medical data, Informatics in Medicine Unlocked, Volume 25, 2021, 100690, ISSN 2352-9148.

[15] Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. PMCID: PMC8477474, 2021.

[16] Gai, L., Xing, M., Chen, W. et al. Comparing CNN-based and transformer-based models for identifying lung cancer: which is more effective?. Multimed Tools Appl, 2023.

[17] Vision transformer promotes cancer diagnosis: A comprehensive review, Expert Systems with Applications, Xiaoyan Jiang, Shuihua Wang, Yudong Zhang, Volume 252, Part A, 2024, 124113, ISSN 0957-4174.

[18] Khan S, Ali H, Shah Z. Identifying the role of vision transformer for skin cancer-A scoping review. Front Artif Intell. 2023 Jul 17; 6:1202990. doi: 10.3389/frai.2023.1202990. PMID: 37529760; PMCID: PMC10388102.

[19] Hong, Z.; Xiong, J.; Yang, H.; Mo, Y.K. Lightweight Low-Rank Adaptation Vision Transformer Framework for Cervical Cancer Detection and Cervix Type Classification. Bioengineering 2024, 11, 468.

[20] Ayana, G.; Barki, H.; Choe, S.-w. Pathological Insights: Enhanced Vision Transformers for the Early Detection of Colorectal Cancer. Cancers 2024, 16, 1441.

[21] Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, 2021.

[22] Touvron H, Cord M, El-Nouby A, Verbeek J, Jégou H. Three things everyone should know about vision transformers. InEuropean Conference on Computer Vision 2022 Oct 23 (pp. 497-515). Cham: Springer Nature Switzerland.

[23] Wang S, Li BZ, Khabsa M, Fang H, Ma H. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768. 2020 Jun 8.

[24] Xiong Y, Zeng Z, Chakraborty R, Tan M, Fung G, Li Y, Singh V. Nyströmformer: A nyström-based algorithm for approximating self-attention. InProceedings of the AAAI Conference on Artificial Intelligence 2021 May 18 (Vol. 35, No. 16, pp. 14138-14148).

[25] alyasriy, hamdalla; AL-Huseiny, Muayed (2023), “The IQ-OTH/NCCD lung cancer dataset”, Mendeley Data, V4, doi: 10.17632/bhmdr45bh2.4

[26] Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142v1 [eess.IV], 2019

[27] Patro SG, Sahu KK. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462. 2015 Mar 19.

[28] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019).

[29] Liu, Qiong & Wu, Ying. (2012). Supervised Learning. 10.1007/978-1-4419-1428-6_451.

[30] Y. Ho and S. Wookey, "The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling," in *IEEE Access*, vol. 8, pp. 4806-4813, 2020, doi: 10.1109/ACCESS.2019.2962617.

[31] Ciampiconi L, Elwood A, Leonardi M, Mohamed A, Rozza A. A survey and taxonomy of loss functions in machine learning. arXiv preprint arXiv:2301.05579. 2023 Jan 13.

[32] Lee SH, Lee S, Song BC. Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492. 2021 Dec 27.

[33] Zhai X, Kolesnikov A, Houlsby N, Beyer L. Scaling vision transformers. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022 (pp. 12104-12113).

[34] Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A. Emerging properties in self-supervised vision transformers. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 9650-9660).