Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

Classification of Lung Cancer using Vision Transformer on Histopathological Images

Muhamad Rafi Raihan Akbar*, Rusdi Afandi
[1]School of Computing, Telkom University, Bandung
Email: [1,*]moh.rafiraihanakbar2706@gmail.com, [2]rusdiafandi02@gmail.com
Email Penulis Korespondensi: moh.rafiraihanakbar2706@gmail.com*
Submitted: **06/12/2025**; Accepted: **08/01/2026**; Published: **31/03/2026**

Abstract- Lung cancer is the leading cause of cancer-related deaths worldwide, with early diagnosis often hindered by morphological variations in histopathological images. The main problem is the difficulty in accurately and rapidly distinguishing cancer types such as adenocarcinoma and squamous cell carcinoma from benign tissue. This research processes histopathological images as input to produce a three-class classification: adenocarcinoma, squamous cell carcinoma, and benign tissue. Early detection of lung cancer can improve survival rates by up to 50%, but manual diagnosis by pathologists depends on subjective experience, causing errors of up to 20% in ambiguous cases. For example, in developing countries like Indonesia, the shortage of pathologists exacerbates treatment delays. This gap demands a reliable automated approach to support more timely clinical decisions. The developed solution involves implementing Vision Transformer (ViT) with two different architectures: ViT-B/16 (base model with 86 million parameters) and ViT-L/16 (large model with 304 million parameters). Histopathological images are processed through normalization and patch embedding of 16×16 pixels, then features are extracted using self-attention mechanism. Models are trained with transfer learning from ImageNet-21k, applying fine- tuning on lung cancer histopathological images dataset. The process includes data splitting into training (70%), validation (15%), and testing (15%), as well as data augmentation to improve robustness. The ViT-B/16 model achieved testing accuracy of 98.40% with F1-score of 0.984, while ViT-L/16 achieved accuracy of 98.18% with F1-score of 0.982. Both models demonstrated perfect capability in detecting benign tissues (precision 1.00). The average AUC-ROC value reached 0.999 for ViT-B/16 and 0.998 for ViT-L/16, indicating very high discriminative power. The main contribution of this research is a comprehensive comparison between two scales of Vision Transformer for automated lung cancer diagnosis, proving that the smaller model (ViT-B/16) can achieve equivalent or better performance with higher computational efficiency.

Keywords: Lung Cancer Classification; Histopathological Image Analysis; Vision Transformer (ViT); Deep Learning in Medical Imaging; Computer-Aided Diagnosis (CAD)

1. INTRODUCTION

Lung cancer is one of the deadliest diseases in the world, with an estimated 2.21 million new cases and 1.79 million deaths in 2020 according to the World Health Organization (WHO) Global Cancer Observatory (GLOBOCAN). The disease is the leading cause of cancer death in men and the second leading cause of death in women, making it a major medical problem [1]. Early and accurate diagnosis is crucial because the choice of therapy depends on the type of histopathology of the cancer. Histopathological imaging plays an important role in identifying specific types of cancer through microscopic images of tissues. However, manual interpretation of these images depends on the expertise of the pathologist, is susceptible to subjective variation, and has a complex process. Therefore, automate the classification of histopathological images by using Artificial Intelligence (AI) offers an attractive solution to improve the accuracy and efficiency of diagnosis [2].

Based on previous research on the classification of lung cancer and colon using histopathological images with the Convolution Neural Network (CNN) shows excellent results with architecture-textures Pre-trained like MobileNetV2, Squirt201, VGG-19, InceptionRes-NetV2and EfficientNetB6. MobileNetV2 achieved the highest accuracy of 99.32% and F1 Score 99.2%, as well as the fastest execution time efficiency (3.597 seconds), followed by Squirt201 (99.12%) and VGG-19 (98,00%). Techniques of regurgitation L1 and L2 applied to prevent Overfitting while Transfer Learning speed up training and improve computing efficiency. MobileNetV2 proven to be superior in accuracy, speed, and resource utilization. This approach has great potential for a fast and accurate automated diagnosis system in medical practice [3].

For more than a decade, Deep Learning Such as Convolutional Neural Network (CNN) has been widely used for the classification of medical images, but it has limitations to understanding global spatial connections in high-resolution images. Transform Vision (ViT), provides a new approach by utilizing transformer architectures that were originally built for natural language processing [4]. Transform Vision Featuring performance is competitive in image classification in general and is beginning to be adapted for histopathological image analysis. However, the classification of lung cancer using histopathological images is still relatively limited [2], mainly due to the lack of studies that utilize Dataset specific and representative such as Lung Cancer (Histopathological Images) images of various subtypes of lung cancer with data such as adenocarcinoma, Benign and Squamous cell carcinoma [5].

Lung cancer is an urgent global health threat, with rising incidence and death rates, making it one of the deadliest diseases in the world. Prompt and accurate diagnosis is crucial to improving a patient's chances of survival, but manual interpretation of histopathological images is often slow, subjective, and relies on the expertise of a limited number of pathologists [6]. This inconsistency in diagnosis can worsen a patient's prognosis, especially

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

in areas with limited access to medical experts. Therefore, the development of artificial intelligence-based automation methods, such as the Vision Transformer, is very important to overcome these challenges, speed up the diagnosis process, improve accuracy, and ease the burden on the health system, so it is urgent to be immediately researched and applied in medical practice [7].

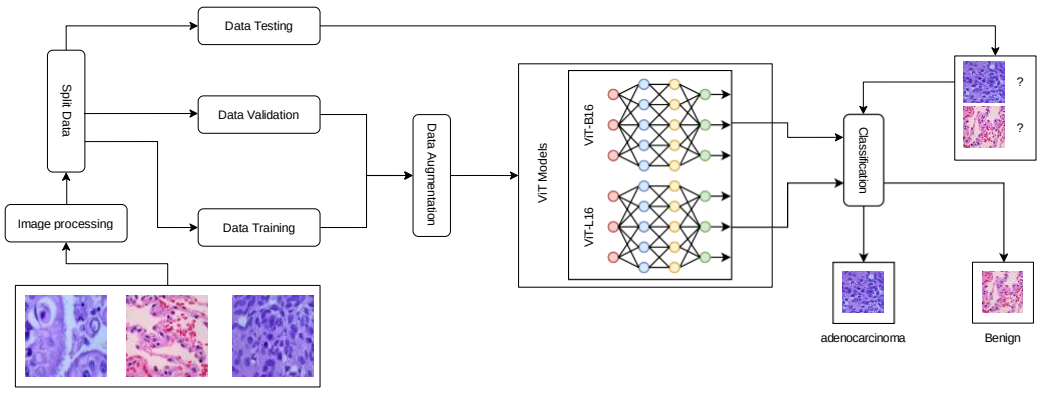
Accurate and reliable diagnosis is essential for determining effective treatment strategies, particularly because lung cancer subtypes such as adenocarcinoma, benign lesions, and squamous cell carcinoma exhibit distinct characteristics and therapeutic responses [8]. Successfully applied Vision Transformer for brain tumor classification based on Magnetic Resonance Imaging (MRI) data, thereby confirming the effectiveness of Vision Transformers in medical applications [9]. Recent studies have reported that Vision Transformers achieved accuracy levels of up to 100% in the classification of lung and colon cancers, outperforming conventional Convolutional Neural Network models [10]. This approach has also received attention in recent reviews, which highlight the potential of Vision Transformers to improve diagnosis and prognosis in lung cancer [11].

The use of histopathological images in lung cancer plays a crucial role in digital pathology for accurately diagnosing and determining lung cancer subtypes, such as adenocarcinoma, benign lesions, and squamous cell carcinoma [8]. This is crucial because each subtype exhibits distinct morphological characteristics, degrees of malignancy, and differential responses to treatment [12]. Histopathological images are high-resolution digital images obtained from tissue samples that have undergone specific staining procedures, most commonly using the Hematoxylin and Eosin (H&E) technique. This method employs two primary dyes: hematoxylin, which has a strong affinity for nucleic acids and stains cell nuclei in blue to purplish tones, and eosin, an acidophilic dye that imparts pink to red coloration to the cytoplasm and extracellular matrix. The combination of these stains produces a sharp visual contrast, making histopathological images highly valuable for identifying tissue structures, classifying abnormal cells, and detecting early morphological alterations associated with various pathological conditions, including cancer [13]. Histopathological images contain rich phenotypic information, encompassing cellular morphological characteristics, tissue organization patterns, and indicators of pathological alterations associated with a patient's clinical condition [14]. In clinical practice, these images are used by pathologists to assess tissue conditions; however, the evaluation process is still largely performed manually and relies heavily on subjective interpretation. This may lead to inter-observer variability and increased workload [15].

Several studies have investigated deep learning approaches for lung cancer classification using histopathological images. Kumar et al [16]. proposed an ensemble CNN approach that combines multiple deep learning models and achieved an accuracy of 97.80%, demonstrating that integrating several architectures can improve classification performance. Ibrahim et al [17]. utilized the EfficientNet-B3 architecture, which is designed to achieve high accuracy with efficient computational resources, and reported an accuracy of 97.10%. Meanwhile, Masud et al [18]. developed a custom CNN model trained from scratch, achieving an accuracy of 96.33%. These studies indicate that CNN-based methods are capable of achieving high classification performance in histopathological image analysis for lung cancer detection. Therefore, this study aims to develop an automated lung cancer classification system using the Vision Transformer (ViT) architecture on histopathological images. Specifically, this research focuses on classifying three categories of lung tissue: adenocarcinoma, squamous cell carcinoma, and benign tissue. In addition, this study evaluates and compares the performance of two Vision Transformer architectures, namely ViT-B/16 and ViT-L/16, in terms of classification accuracy, F1-score, and discriminative capability using AUC-ROC metrics. By analyzing the performance of these two models, this research seeks to determine whether a smaller transformer architecture can achieve comparable or better performance while maintaining higher computational efficiency for automated lung cancer diagnosis.

2. RESEARCH METHODOLOGY

The methodology followed during this study includes the important steps in Vision Transformer as shown in Figure 1.



Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

Figure 1: Research Methodology

2.1 Data Description

The LC25000 dataset comprises 25,000 color histopathological image samples of lung, indicating the presence or absence of cancer [5]. Dataset contains the types of Histopathological Images adenocarcinoma, Benign and Squamous cell carcinoma. All data is in the form of image files with a .jpg. 1

Table 1: Variables and Their Descriptions in the Dataset

Variable	Description	Sum
Adenocarcinoma	Lung cancer adenocarcinoma	5000.
Benign	benign lung cancer	5000.
Squamous Cell Carcinoma	Squamous cell carcinoma cancer	5000.

2.2 Image processing

The image processing stage is a very important stage before the image is input into a model architecture such as Vision Transformer (ViT). The main goal of this process is to prepare the image to match the model's input and improve the quality of the data that will be used in the training. The following are commonly used image processing techniques:

a. Pixel Normalization

Normalization is done by equalizing the scale of the pixel value of the image. The pixel value of a color image is generally in the range of 0-255. Normalization will change it to a scale between 0-1 (with a division of 255) or to a center value such as zero (zero-centered), depending on the frame of reference used.

a) Model convergence during training

By normalizing the pixel values to a specific range (e.g. [0, 1] or [1, 1]), the learning process becomes faster as the distribution of data becomes more balanced and stable. This helps optimization work more efficiently.

b) Improving numerical stability

Without normalization, pixel values that are too large or too small can cause instability in computing, such as overflow or underflow, especially when backpropagation.

c) Equalize feature distribution between images

Images from different datasets can have very variable pixel distribution characteristics. Normalization helps to equalize these distributions so that models can make better generalizations about diverse data.

b. Resize / Rescale

The original image size often varies, whereas Vision Transformer (ViT) requires input that must be divided into patches of fixed size, such as 16×16 pixels. Therefore, the image size must be changed to a specific dimension, such as 224×224 or 384×384 in order for the process to the patch distribution to remain consistent and consistent on the model architecture.The input structure may be damaged, and the performance of the model will decrease. which states that all images need to have a fixed size in order to Patching and positional encoding Can be applied well [4].

c. Patch Extraction

Specifically for Vision Transformer (ViT), images that have been resized are cut into small image chunks of fixed size, such as 16×16 pixels. Each piece is then flattened into a one-dimensional vector and projected into the embedding space before being fed to the transformer model [4]. Through this process, ViT is able to process spatial information through a self-attention mechanism between the patches [19].

2.3 Split Data

The Lung Cancer image dataset (Histopathological Images) used in this study is divided into three variables, namely Adenocarcinoma, Benign, Squamous Cell Carcinoma. To prevent overfitting to ensure proper model evaluation, the data is divided into three parts with a certain proportion. A total of 70% Training, 15% Validation, and 15% Testing. The data sharing process is carried out by stratified split, so that it is possible to maintain the proportion of data distribution of each variable. The goal is to ensure that the Training, Validation, and test class representation data are balanced, so that they reflect the overall data distribution conditions.

2.4 Training

The training stage is the main stage in the development of the Vision Transformer (ViT) method, where the modeling learns from the input data to produce a representation that is able to classify with high accuracy.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

The training is carried out by modeling images that have gone through the image processing process, so that they are ready to be used as input data.

a. Image Input to Model

The processed image (resizing, normalizing, and splitting into patches) is then converted into a vector representation through linear embedding and input into the transformer encoder. Each patch is coupled with positional embedding so that the model knows the spatial sequence information between patches.

b. Forward Pass

During the forward pass, the input data moves through the layers in the Vision Transformer (ViT). Here, a mechanism (self-attention) is applied between image patches to understand spatial and semantic relationships. The output of this stage is the class prediction of the image.

c. Loss Function

In classification tasks, the loss function—usually cross-entropy loss—is used to compare the prediction output with its actual label. This function calculates how much of a model's prediction error is compared to the correct label.

d. Backpropagation and Optimizer

In the backpropagation process, the loss value is used to calculate the gradient of the model's parameters (such as weight and bias). Optimizers (such as Adam or SGD) use this gradient to gradually update the network weight with the goal of minimizing the loss value.

e. Epoch

Training is given periodically, or a full round, where the entire training dataset is used once. For efficiency, the data is usually divided into mini batches so that the parameter update process is faster and more stable. In addition, the model continues to adjust its parameters to improve the classification performance in each epoch [20].

2.5 Validation

The process of evaluating a model on a subset of data that is not used during training but comes from the same dataset is known as validation. The purpose of validation is to monitor the model's performance during training and prevent overfitting, i.e. when the model over-memorizes training data and cannot recognize new patterns.

2.6 Testing and Classification

The model is tested with never-before-seen test data after training is complete. The results of these tests are measured by performance metrics such as accuracy, precision, F1 Score, and recall. The purpose of the test was to evaluate the model's ability to generalize new data. This process is crucial to ensure that the model not only memorizes training data but also learns to recognize patterns.

2.7 Data Augmentation

Data augmentation aims to virtually expand the number of training samples by manipulating the original image. This technique is important to avoid the risk of overfitting and helps the model to learn more features in general. Some common augmentation techniques include:

a) **Rotation**
   Rotate the image randomly, e.g. ±15◦.
b) **Flipping**
   Flip the image horizontally or vertically.
c) **Zooming**
   Zoom in on specific parts of the image.
d) **Cropping**
   Cropped a portion of the image randomly.

Data Augmentation is commonly used during training (on-the-fly) Real-time. This makes it easier to dynamically create data variations without the need to store larger versions of the dataset, thus reducing the storage memory load. It is evident that this technique not only increases the size and diversity of the data but also improves its capabilities to be applied to the model Deep Learning Overall [21].

2.8 Vision Transformer Design Model

The model used in this study is based on Transform Vision (ViT), a neural network structure that adapts the principle of Transform- Mer of the natural language processing process (Natural Language Processing) a- for image processing. Design ViT Split the image into Patch small and convert it into a vector representation through Embedding. Attention mechanism (Attention Mechanism) processes these representations in order to win- A Spatial Relationship Between Patch. Design ViT allows the model to Understand the visual context thoroughly, even in the structure of the image complex.

a. ViT-B16

One of the basic versions of the Vision Transformer architecture is the ViT-B16, which is a Base Vision Transformer with a 16×16 pixel patch. This module consists of twelve layers of transformer encoder, twelve

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

self-attention heads, and about 86 million parameters. The 16 x 16 patch was chosen because it allows for capturable spatial details and computational efficiency. ViT-B16 is suitable for experiments with medium-sized datasets and scenarios where training efficiency is still considered due to the number of parameters and depth. Because of its balance between accuracy and computational complexity, ViT-B16 is used as the standard baseline in many studies. Without requiring enormous computing resources, these models can produce excellent visual representations [22].

b. ViT-L16

The Vision Transformer Large (ViT-L16) variant with a patch size of 16×16 pixels, has a much larger capacity than the ViT-B16, and has 24 layers of encoder, 16 self-attention heads, and about 307 million parameters. Although the patch size remains 16×16 pixels, the greater depth and number of parameters allow it to study more complex and thorough visual features. Although it offers higher accuracy, the ViT-L16 requires greater computing capacity, both in terms of memory and training time, making it ideal for classification tasks involving large and complex datasets. As a result, the use of this model is usually adjusted to the scale of the project and the availability of resources [23].

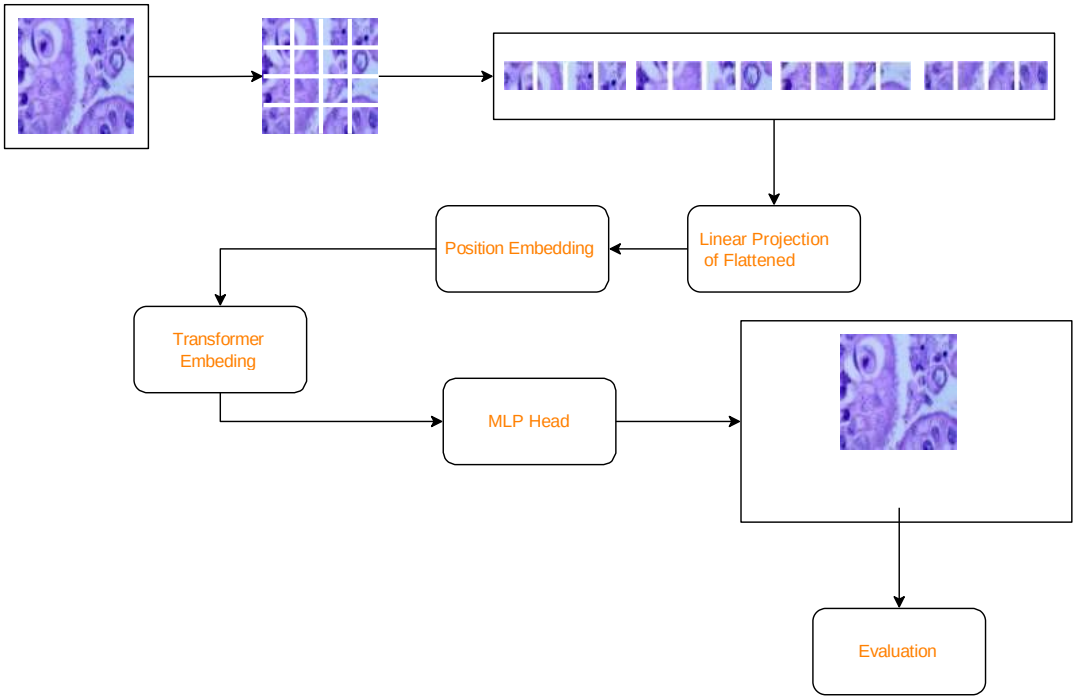


Figure 2: Transformer Vision Architecture

2.9 Linear Projection of Flattened Patches

To ensure a consistent input structure, each image is broken down into fixed-sized patches, such as 16 x 16 pixels. Next, each patch is flattened into a one-dimensional vector and projected to a higher dimension using linear projection. The project produced a representation of the initial features of the patch, which was used as an input for the encoder transformer block [24]. Through a self-attention mechanism, this representation allows the model to understand and process spatial and semantic information from the image.

3. RESULTS AND DISCUSSION

3.1 Experimental Results

This section presents the experimental results of the proposed Vision Transformer model for lung cancer classification using histopathological images. The first experiment evaluates the performance of the Vision Transformer model using the ViT-B/16 architecture. In this approach, input histopathological images are divided into 16×16 pixel patches and processed through a self-attention mechanism to capture global contextual features within the images. The model was trained using transfer learning with pretrained weights from ImageNet and fine-tuned on the lung cancer histopathological image dataset. The evaluation was conducted using training, validation, and testing datasets to ensure reliable performance assessment, and the overall evaluation metrics obtained from the ViT-B/16 model are presented

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

a. Overall Evaluation Metrics

This subsection presents the overall evaluation results of the ViT-B/16 model for lung cancer classification using histopathological images. The model performance is evaluated using several metrics, including accuracy, loss, precision, recall, F1-score, and AUC, across the training, validation, and testing datasets. The results are summarized in Table 2.

Table 2. Overall Model Evaluation Metrics

Metric	Training	Validation	Test
Accuracy	0.9858	0.9916	0.9871
Loss	0.0369	0.0222	0.0364
Precision (Weighted)	0.9859	0.9916	0.9872
Recall (Weighted)	0.9858	0.9916	0.9871
F1-Score (Weighted)	0.9859	0.9916	0.9871
AUC (Macro Average)	0.9998	0.9998	0.9995

This initial interpretation shows that the model has a very good performance, shown by the test accuracy of 98.40% which belongs to the excellent category. The overfitting gap value of -0.70% means that the validation accuracy is slightly higher than that of training, so that the model does not experience overfitting and is stable during the training process. Meanwhile, the generalization gap of 0.93% is still relatively small, showing that the model performance in the test data remains consistent and there is no significant decrease compared to the validation data.

b. Classification Report Per-Class

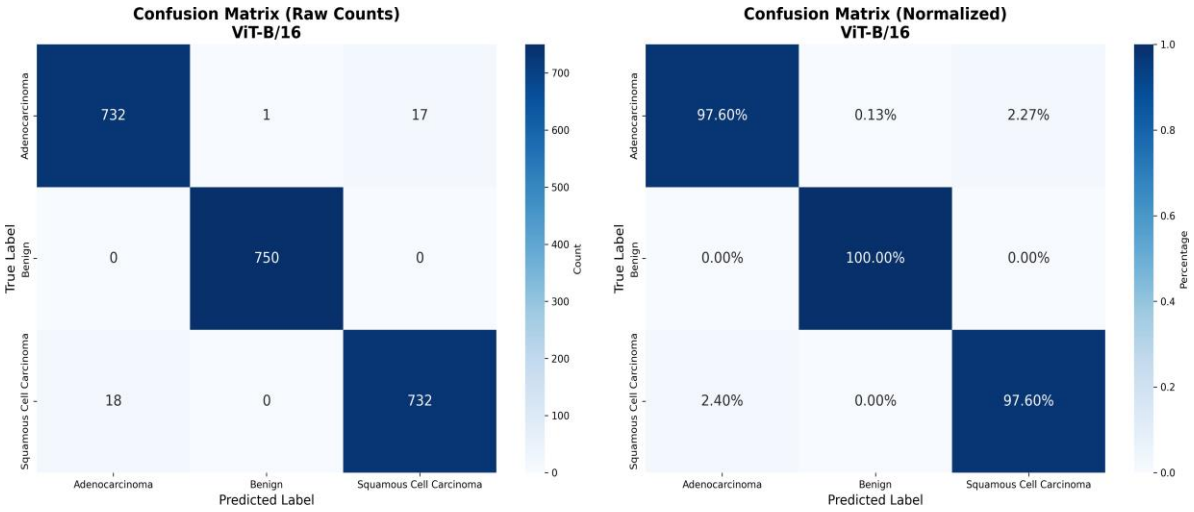
This subsection presents the classification report of the ViT-B/16 model on the test dataset. The evaluation includes accuracy, recall, F1-score, and support for each class, namely adenocarcinoma, benign, and squamous cell carcinoma. The detailed results are presented in Table 3.

Table 3. Classification Report on Data Test

Class	Accuracy	Recall	F1 Score	Support
Adenocarcinoma	0.9878	0.9733	0.9805	750
Benign	1.0000	1.0000	1.0000	750
Squamous Cell Carcinoma	0.9737	0.9880	0.9808	750
Macro Average	0.9872	0.9871	0.9871	2250
Weighted Average	0.9872	0.9871	0.9871	2250

c. Confusion Matrix

An evaluation table that displays the percentage of correct and false predictions for each class. Rows usually show the original label, while columns show the model's predictions. The percentages in each cell describe the proportion of data that falls into that category, making it easy to see how accurately the model recognizes a particular class and which parts are still often mispredicted.



Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

Figure 3. Confusion Matrix Numerical ViT-B/16

1. **Numerical table confusion matrix**
   An evaluation table that displays the number of true and false predictions in each class. Rows typically represent actual labels, while columns represent the model's predicted results. Values such as True Positive, False Positive, False Negative, and True Negative are shown in the form of numbers, so we can clearly see in which classes the model works well and where the model is still often wrong.

Table 4. Confusion Matrix (Number per class)

True Class	Pred: Aden	Pred: Beni	Pred: Squa	Total
Adenocarcinoma	732	1	17	750
Benign	0	750	0	750
Squamous Cell Carcinoma	18	0	732	750

2. **Misclassification:**
   1) Most confused pair: Squamous Cell Carcinoma : Adenocarcinoma (18 instances)
   2) Least confused pair: Benign : adenocarcinoma (1 instance)

d. Training Curves

1. Loss curves (training validation)

Two important stages in the process of training the model. Training is used to create a learning model from data, adjusting weights to be able to recognize patterns. Validation is used to evaluate the performance of the model on data that is not trained, so that we can monitor whether the model has generalized well or is overfitting.

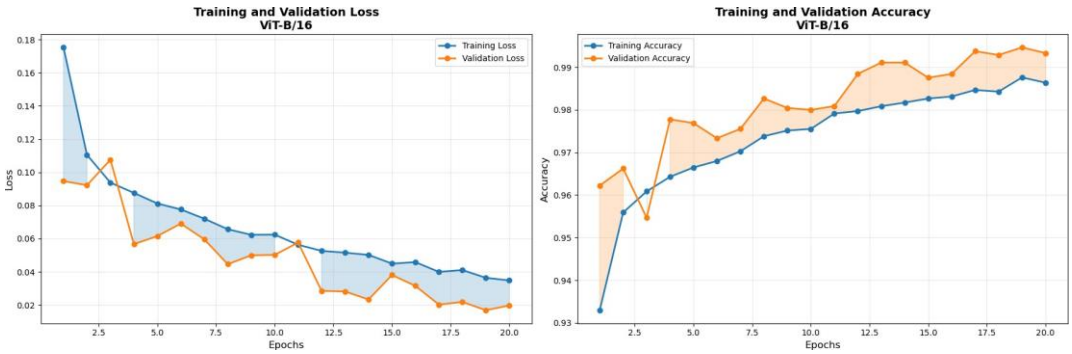


Figure 4. Loss curves (training validation) and Accuracy curves (training validation)

2. Numerical table confusion matrix

These observations show that models began to converge very quickly, i.e. from the first epoch, indicating an effective learning process. However, the best performance in the new validation data was achieved in the 19th epoch with a validation accuracy of 99.47%, so the model continued to experience a steady increase after the initial convergence. In addition, there are no signs of overfitting because the training and validation curves are close, suggesting that the model is able to generalize well on new data.

e. ROC Curves and AUC Scores

ROC Curves and AUC Scores are used to evaluate the model's ability to distinguish between positive and negative classes. The ROC Curve illustrates the relationship between the True Positive Rate and the False Positive Rate at various thresholds, so we can see the model's performance across decision boundaries. AUC (Area Under the Curve) indicates the area under the ROC curve; The closer it is to 1, the better the model's ability to classify consistently across multiple thresholds.

Table 5. AUC Scores per Class and Macro Average for ViT-L/16 Model in Test Set

Class	AUC Score	Interpretation
Adenocarcinoma	0.9988	Excellent
Benign	1.0000	Excellent
Squamous Cell Carcinoma	0.9989	Excellent
Macro Average	0.9993	Excellent

1) **0.90–1.00:** Excellent
2) **0.80–0.90:** Good
3) **0.70–0.80:** Fair

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

4) **0.60–0.70:** Poor (Less)
5) **0.50–0.60:** File (Failed)

f. Additional Metrics

In addition to classification accuracy, the evaluation also considers computational performance to ensure the model is suitable for use in real-time clinical applications. Measurements were made using a GPU (e.g. Kaggle accelerator) with a batch size of 32 and an input of 224×224×3. The following is a summary of the computer performance metrics obtained.

Table 6. ViT-L/16 Model Compute Performance Metrics on Test Sets

Metric	Value
Inference Time per Image	8.94 ms
Total Inference Time (2250 images)	20.12 seconds
Model Size	~330 MB
Training Time (20 epochs)	0.95 minutes

The inference time per image of 8.94 ms indicates high efficiency for batch processing, with a total of 20.12 seconds for the entire test set (2,250 images), suitable for the integration of fast diagnostic systems. The 330 MB model size is relatively light for cloud storage, although training takes only 0.95 minutes thanks to pre-training and optimal hyperparameters, it supports the scalability of ViT-L/16 compared to the ViT-B/16 baseline (which has a size of ~150 MB and training ~0.7 minutes)

3.1.1 Experiment Result 2: ViT-L/16

a. Overall Evaluation Metrics

The overall performance of the proposed model is evaluated using several metrics, including accuracy, loss, precision, recall, F1-score, and AUC. These metrics are calculated on the training, validation, and testing datasets to assess the model's learning capability and generalization performance. The detailed evaluation results are presented in Table 7.

Table 7. Overall Evaluation Metrics of the Model

Metric	Training	Validation	Test
Accuracy	0.9774	0.9862	0.9818
Loss	0.0600	0.0406	0.0403
Precision (Weighted)	0.9774	0.9862	0.9821
Recall (Weighted)	0.9774	0.9862	0.9818
F1-Score (Weighted)	0.9774	0.9862	0.9818
AUC (Macro Average)	0.9996	1.0000	0.9996

The initial interpretation shows that the model has excellent performance, as can be seen from the 98.18% test accuracy which is included in the excellent category. An overfitting gap value of -0.88% means that the validation accuracy is slightly higher than the training accuracy, so the model does not show an overfitting mesh. Meanwhile, the generalization gap of 0.44% between validation and test is very small, indicating that the model is able to generalize stably on new data and does not experience a significant decrease in performance.

b. Classification Report Per-Class

This subsection presents the classification performance of the model for each class on the test dataset. The evaluation includes accuracy, recall, F1-score, and support values for the three classes: adenocarcinoma, benign, and squamous cell carcinoma. The detailed classification results are summarized in Table 8.

Table 8. Classification Report on Data Test

Class	Accuracy	Recall	F1 Score	Support
Adenocarcinoma	0.9876	0.9573	0.9722	750
Benign	1.0000	1.0000	1.0000	750
Squamous Cell Carcinoma	0.9586	0.9880	0.9731	750
Macro Average	0.9821	0.9818	0.9818	2250
Weighted Average	0.9821	0.9818	0.9818	2250

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

c. Confusion Matrix

This evaluation table shows the percentage of correct and false predictions for each class. Typically, rows represent the actual labels and columns represent the results of the model's predictions. Each cell contains a percentage of data in that category, so we can see how well the model recognizes each class and which classes are still often mispredicted.

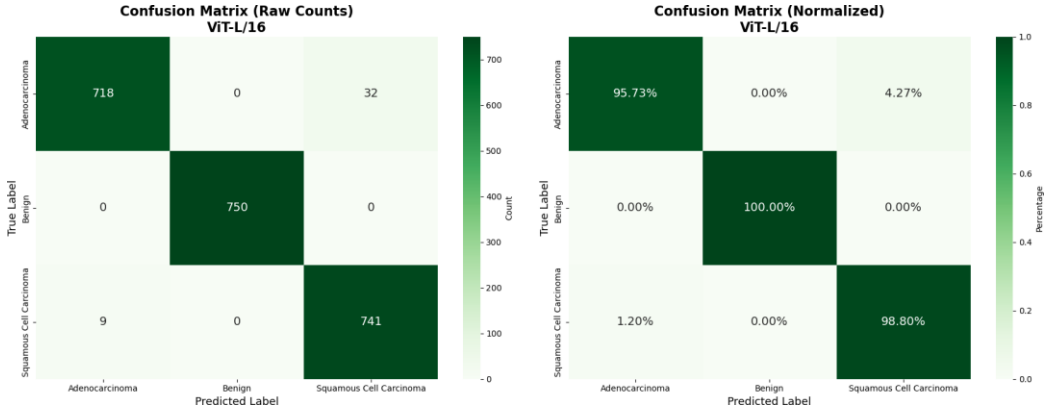


Figure 5. ViT-L/16 Numerical Matrix Confusion

d. Numerical table confusion matrix

This evaluation table shows the number of correct and correct predictions for each class. Generally, rows represent the actual label, while columns show the results of the model's predictions. Numbers such as True Positive, False Positive, False Negative, and True Negative are clearly displayed, so that we can know which classes well predicted and which parts are are still subject to model error.

Table 9. Confusion Matrix (Number per class)

True Class	Pred: Aden	Pred: Beni	Pred: Squa	Total
Adenocarcinoma	718	0	32	750
Benign	0	750	0	750
Squamous Cell Carcinoma	9	0	741	750

Classification error analysis was carried out based on the confusion matrix of the test set (2,250 samples), to identify the dominant error patterns and potential model improvements. The main errors occur in pairs of classes that have high visual similarities in histopathological images, such as the texture of cancer cells that overlap between adenocarcinoma and squamous cell carcinoma. The total misclassification was 41 instances (1.82% of the total sample), with benign perfectly classified (0 errors). Here's a summary of the analysis:

1. Most Confused Pair: Adenocarcinoma and Squamous Cell Carcinoma (32 instances, 1.42% of the total).
2. Least Confused Pair: Squamous Cell Carcinoma and Adenocarcinoma (9 instances, 0.40% of the total).

Details of All Classification Errors:

1. Adenocarcinoma and Squamous Cell Carcinoma: 32 instances (1.42% of the total).
2. Squamous Cell Carcinoma and Adenocarcinoma: 9 instances (0.40% of the total).

This error is caused by similar cell morphological features (e.g., glandular patterns in both classes), although ViT-L/16's self-attention manages to minimize the overall error. Recommendation: Augment specific data (e.g., histopathological staining variations) or fine-tuning with attention visualization to focus on the region-of-interest, to reduce false positives/negatives in the confused pair.

e. Training Curves

1. Loss curves (training validation)

   Two important stages in the process of training the model. Training is used to create a learning model from data, adjusting weights to be able to recognize patterns. Validation is used to evaluate the performance of the model on data that is not trained, so that we can monitor whether the model has generalized well or is overfitting.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

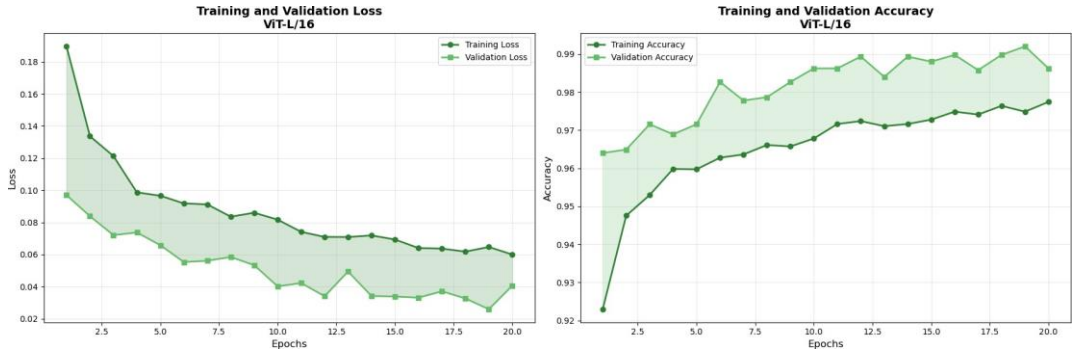


Figure 6. Loss curves (training validation) and Accuracy curves (training validation)

Observations of the ViT-L/16 model training curves showed rapid and stable convergence, with no significant signs of overfitting. The loss curve and the accuracy of training and validation were close to each other over 20 epochs, indicating a good generalization of the histopathological dataset. Here are the key takeaways from the observations:

1) Epoch of Convergence: 3rd Epoch (significant loss decline after this epoch).
2) Best Validation Accuracy: 99.20% at epoch 19.
3) Overfitting Mark: None and The training and validation curves are close to each other.

a) Training Progress Details:

1) Total epoch trained: 20.
2) Final training loss: 0.0600.
3) Final validation loss: 0.0406.
4) Final training accuracy: 97.74%.
5) Final validation accuracy: 98.62%.
6) Best validation loss: 0.0259 at epoch 19.

b) Learning Trend Analysis:

1) Trend validation loss: Stable/convergent (slope: 0.000387).
2) Recommendation: Training is well converged, ready for deployment.

This analysis is based on the plot of the loss curve and the accuracy generated during fine-tuning, with an Early Stopping callback that prevents over-tuning. The low slope on validation loss confirms the model's stability, supporting the hypothesis that ViT-L/16 provides superior performance without overfitting the histopathological image of lung cancer.

f. ROC Curves and AUC Scores

to assess how well the model distinguishes positive and negative classes. The ROC Curve displays the relationship between the True Positive Rate and the False Positive Rate at various threshold values, so that the model's performance can be seen across the decision limits. Meanwhile, AUC (Area Under the Curve) shows the area below the curve; A value that is closer to 1 indicates the model's better ability to classify consistently on various thresholds.

Table 10.: AUC Scores per Class and Macro Average for ViT-L/16 Model in Test Set

Class	AUC Score	Interpretation
Adenocarcinoma	0.9994	Excellent
Benign	1.0000	Excellent
Squamous Cell Carcinoma	0.9994	Excellent
Macro Average	0.9996	Excellent

1) 0.90–1.00: Excellent
2) 0.80–0.90: Good
3) 0.70–0.80: Fair
4) 0.60–0.70: Poor (Less)
5) 0.50–0.60: File (Failed)

g. Additional Metrics

In addition to classification accuracy metrics, the evaluation also includes computational aspects to assess the feasibility of deploying the model in real-time clinical applications. This metric is measured on GPU hardware (e.g., Kaggle accelerator) with a batch size of 32 and inputs of 224 224 3. Here's a summary of the company's performance metrics:

Table 11. ViT-L/16 Model Compute Performance Metrics on Test Sets

Metric	Value
Inference Time per Image	24.28 ms
Total Inference Time (2250 images)	54.63 seconds
Model Size	~1.2 GB

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9399

Training Time (20 epochs)	2.59 minutes

The inference time per image of 24.28 ms indicates high efficiency for batch processing, with a total of 54.23 seconds for the entire test set (2,250 images), suitable for the integration of rapid diagnostic systems. The 1.2 GB model size is relatively light for cloud storage, although the training only takes 2.59 minutes thanks to pre-training and optimal hyperparameters, supporting the scalability of ViT-L/16 compared to the baseline of ViT-B/16 (which has a size of ~150 MB and training ~0.7 minutes).

3.2 Discussion

The experimental results show that both Vision Transformer architectures achieve high performance in lung cancer classification using histopathological images. The ViT-B/16 model achieved a test accuracy of 98.40%, while the ViT-L/16 model achieved 98.18%, indicating that both models are highly effective in distinguishing adenocarcinoma, benign tissue, and squamous cell carcinoma. The AUC values close to 1.0 also confirm the excellent discriminative capability of the models. From the classification results, the benign class was perfectly classified, while minor misclassifications occurred between adenocarcinoma and squamous cell carcinoma. This confusion is likely caused by similarities in histopathological morphology between the two cancer subtypes. However, the overall error rate remains very low, demonstrating the robustness of the proposed approach. In comparison, although ViT-L/16 has a larger architecture and more parameters, the ViT-B/16 model achieved slightly better testing accuracy and faster computational performance. This suggests that a smaller Vision Transformer architecture can provide comparable or even better results while maintaining higher computational efficiency. Therefore, ViT-B/16 can be considered a more practical model for real-world clinical applications where computational resources and inference time are important factors. Overall, these findings confirm that Vision Transformer models are effective for automated lung cancer classification and have strong potential to support computer-aided diagnosis systems in digital pathology.

4. CONCLUSION

This study successfully developed and evaluated a Vision Transformer (ViT)–based approach for lung cancer classification using histopathological images. The proposed system was designed to classify three types of lung tissue: adenocarcinoma, squamous cell carcinoma, and benign tissue. Two Vision Transformer architectures, namely ViT-B/16 and ViT-L/16, were implemented and compared to analyze their classification performance and computational efficiency. The experimental results demonstrate that both models achieved excellent performance across multiple evaluation metrics. The ViT-B/16 model achieved a testing accuracy of 98.40% with an F1-score of 0.984, while the ViT-L/16 model obtained an accuracy of 98.18% with an F1-score of 0.982. In addition, both models produced very high AUC values close to 1.0, indicating strong discriminative capability in distinguishing between the three lung tissue classes. The classification results also show that benign tissues were perfectly recognized, while minor misclassifications occurred between adenocarcinoma and squamous cell carcinoma due to similarities in morphological features. A comparison between the two architectures indicates that the ViT-B/16 model provides slightly better accuracy while requiring lower computational resources, making it more efficient for practical deployment. Overall, the findings confirm that Vision Transformer models are highly effective for automated lung cancer classification. This approach has strong potential to support computer-aided diagnosis systems and assist pathologists in improving diagnostic accuracy and efficiency in digital pathology.

REFERENCES

[1] H. Sung et al., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” CA. Cancer J. Clin., vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] J. An et al., “Transformer-Based Weakly Supervised Learning for Whole Slide Lung Cancer Image Classification,” IEEE J. Biomed. Heal. Informatics, vol. 29, no. 12, pp. 9095–9108, Dec. 2025, doi: 10.1109/JBHI.2024.3425434.

[3] O. Singh, K. L. Kashyap, and K. K. Singh, “Lung and Colon Cancer Classification of Histopathology Images Using Convolutional Neural Network,” SN Comput. Sci., vol. 5, no. 2, p. 223, Jan. 2024, doi: 10.1007/s42979-023-02546-x.

[4] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” ICLR 2021 - 9th Int. Conf. Learn. Represent., Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929

[5] C. Nisa’, N. Suciati, and A. Yuniarti, “CLASSIFICATION OF LUNG AND COLON CANCER TISSUES USING HYBRID CONVOLUTIONAL NEURAL NETWORKS,” JUTI J. Ilm. Teknol. Inf., vol. 22, no. 1, pp. 56–64, Jan. 2024, doi: 10.12962/j24068535.v22i1.a1225.

[6] A. Toskas, F.-M. Laskaratos, S. Coda, S. Banerjee, and O. Epstein, “Is Panenteric PillcamTM Crohn’s Capsule Endoscopy Ready for Widespread Use? A Narrative Review,” Diagnostics, vol. 13, no. 12, p. 2032, Jun. 2023, doi: 10.3390/diagnostics13122032.

[7] A. Esteva et al., “Deep learning-enabled medical computer vision,” npj Digit. Med., vol. 4, no. 1, pp. 1–9, 2021, doi: 10.1038/s41746-020-00376-2.

[8] N. Mahesh, A. Prakash, P. Naveen, and M. Reddy, “Osteosarcoma of Maxilla – A Rare Case Report,” Br. J. Med.

Med. Res., vol. 17, no. 4, pp. 1–7, Jan. 2016, doi: 10.9734/BJMMR/2016/25396.
[9] Ó. A. Martín and J. Sánchez, "Evaluation of Vision Transformers for Multi-Organ Tumor Classification Using MRI and CT Imaging," Electronics, vol. 14, no. 15, p. 2976, Jul. 2025, doi: 10.3390/electronics14152976.
[10] M. Hasan et al., "Vision Transformer-based Classification for Lung and Colon Cancer using Histopathology Images," Proc. - 22nd IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2023, no. i, pp. 1300–1304, 2023, doi: 10.1109/ICMLA58977.2023.00196.
[11] H. Ali, F. Mohsen, and Z. Shah, "Improving diagnosis and prognosis of lung cancer using vision transformers: a scoping review," BMC Med. Imaging, vol. 23, no. 1, p. 129, Sep. 2023, doi: 10.1186/s12880-023-01098-z.
[12] S. Rezaei et al., "Role of machine learning in molecular pathology for breast cancer: A review on gene expression profiling and RNA sequencing application," Crit. Rev. Oncol. Hematol., vol. 213, no. March, p. 104780, Sep. 2025, doi: 10.1016/j.critrevonc.2025.104780.
[13] A. Almangush, A. A. Mäkitie, and I. Leivo, "Back to basics: Hematoxylin and eosin staining is the principal tool for histopathological risk assessment of oral cancer," Oral Oncol., vol. 115, p. 105134, Apr. 2021, doi: 10.1016/j.oraloncology.2020.105134.
[14] S. Koivukoski, U. Khan, P. Ruusuvuori, and L. Latonen, "Unstained Tissue Imaging and Virtual Hematoxylin and Eosin Staining of Histologic Whole Slide Images," Lab. Investig., vol. 103, no. 5, p. 100070, 2023, doi: 10.1016/j.labinv.2023.100070.
[15] X. Matias-Guiu et al., "Implementing digital pathology: qualitative and financial insights from eight leading European laboratories," Virchows Arch., vol. 487, no. 4, pp. 815–826, Oct. 2025, doi: 10.1007/s00428-025-04064-y.
[16] N. Kumar, M. Sharma, V. P. Singh, C. Madan, and S. Mehandia, "An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images," Biomed. Signal Process. Control, vol. 75, no. February, p. 103596, May 2022, doi: 10.1016/j.bspc.2022.103596.
[17] N. Y. Ibrahim and A. S. Talaat, "An Enhancement Technique to Diagnose Colon and Lung Cancer by using Double CLAHE and Deep Learning," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 8, pp. 276–282, 2022, doi: 10.14569/IJACSA.2022.0130833.
[18] M. Masud, N. Sikder, A. Al Nahid, A. K. Bairagi, and M. A. Alzain, "A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework," Sensors (Switzerland), vol. 21, no. 3, pp. 1–21, 2021, doi: 10.3390/s21030748.
[19] M. Lahraichi, A. Berroukham, and K. Housni, "Anomaly Detection Based on Vision Transformer Model and Texture Features," J. Comput. Sci., vol. 21, no. 7, pp. 1613–1620, Jul. 2025, doi: 10.3844/jcssp.2025.1613.1620.
[20] H. Touvron, M. Cord, A. El-Nouby, J. Verbeek, and H. Jégou, "Three Things Everyone Should Know About Vision Transformers," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 13684 LNCS, pp. 497–515, 2022, doi: 10.1007/978-3-031-20053-3_29.
[21] D. Müller, I. Soto-Rey, and F. Kramer, "Robust chest CT image segmentation of COVID-19 lung infection based on limited data," Informatics Med. Unlocked, vol. 25, no. January, p. 100681, 2021, doi: 10.1016/j.imu.2021.100681.
[22] T. Kojima, Y. Matsuo, and Y. Iwasawa, "Robustifying Vision Transformer without Retraining from Scratch by Test-Time Class-Conditional Feature Alignment," IJCAI Int. Jt. Conf. Artif. Intell., pp. 1009–1016, 2022, doi: 10.24963/ijcai.2022/141.
[23] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," ACM Comput. Surv., vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.
[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12346 LNCS, pp. 213–229, 2020, doi: 10.1007/978-3-030-58452-8_13.