

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications



ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

KDViT: COVID-19 diagnosis on CT-scans with knowledge distillation of vision transformer

Yu Jie Lim, Kian Ming Lim, Roy Kwang Yang Chang & Chin Poo Lee

To cite this article: Yu Jie Lim, Kian Ming Lim, Roy Kwang Yang Chang & Chin Poo Lee (2024) KDViT: COVID-19 diagnosis on CT-scans with knowledge distillation of vision transformer, *Automatika*, 65:3, 1113-1126, DOI: [10.1080/00051144.2024.2349416](https://doi.org/10.1080/00051144.2024.2349416)

To link to this article: <https://doi.org/10.1080/00051144.2024.2349416>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 15 May 2024.



[Submit your article to this journal](#)



Article views: 501



[View related articles](#)



[View Crossmark data](#)



KDViT: COVID-19 diagnosis on CT-scans with knowledge distillation of vision transformer

Yu Jie Lim, Kian Ming Lim , Roy Kwang Yang Chang and Chin Poo Lee

Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia

ABSTRACT

This paper introduces Knowledge Distillation of Vision Transformer (KDViT), a novel approach for medical image classification. The Vision Transformer architecture incorporates a self-attention mechanism to autonomously learn image structure. The input medical image is segmented into patches and transformed into low-dimensional linear embeddings. Position information is integrated into each patch, and a learnable classification token is appended for classification, thereby preserving spatial relationships within the image. The output vectors are then fed into a Transformer encoder to extract both local and global features, leveraging the inherent attention mechanism for robust feature extraction across diverse medical imaging scenarios. Furthermore, knowledge distillation is employed to enhance performance by transferring insights from a large teacher model to a small student model. This approach reduces the computational requirements of the larger model and improves overall effectiveness. Integrating knowledge distillation with two Vision Transformer models not only showcases the novelty of the proposed solution for medical image classification but also enhances model interpretability, reduces computational complexity, and improves generalization capabilities. The proposed KDViT model achieved high accuracy rates of 98.39%, 88.57%, and 99.15% on the SARS-CoV-2-CT, COVID-CT, and iCTCF datasets respectively, surpassing the performance of other state-of-the-art methods.

ARTICLE HISTORY

Received 5 January 2024
Accepted 11 April 2024

KEYWORDS

Vision transformer;
knowledge distillation;
COVID-19 image
classification; CT scan images



1. Introduction

COVID-19, also known as the coronavirus disease, is a highly infectious respiratory illness caused by the SARS-CoV-2 virus [1]. The impact of COVID-19 has resulted in increased research across various fields, including epidemiology, immunology, virology, and medical imaging. While the initial surge of COVID-19 cases may have subsided, the importance of developing effective diagnostic tools remains crucial for managing potential future outbreaks and addressing lingering health concerns. In the Health Technology field, systems based on Artificial Intelligence (AI) have been widely implemented to enhance service quality and the efficiency of diagnosis and treatment processes [2]. Medical imaging, specifically, has been pivotal in diagnosing and treating COVID-19, with techniques such as chest CT scans and X-rays used to identify lung abnormalities associated with the disease. AI and machine learning are also being applied to medical imaging to assist with the detection and diagnosis of COVID-19. The practical application of the proposed model in COVID-19 diagnosis on CT scans lies in its ability to accurately and efficiently identify patterns associated with the disease, facilitating timely and accurate diagnoses even in post-pandemic scenarios. Furthermore, the model's capabilities extend beyond

COVID-19 to encompass various respiratory illnesses, enabling its continued relevance in ongoing healthcare efforts.

In this research, medical image analysis presents various challenges that hinder the development of effective AI models for classifying medical images. One problem encountered is the limited dataset size due to privacy considerations and the cost of data collection. Moreover, automatic detection of the regions of interest (ROIs) over the image global features is challenging due to the complexity and nonlinear nature of medical images. Extracting useful and important features from medical images can be challenging when datasets are collected from different sources and institutions, resulting in domain shift. Another challenge in training a good model with medical images is the class imbalance for each class or label in the dataset. Many medical conditions can be rare, leading to imbalanced datasets and causing the model's performance to be less accurate.

A deep learning framework is introduced, adopting the Vision Transformer (ViT) model with Knowledge Distillation to improve model performance and overcome challenges encountered in medical image analysis in this work. The Vision Transformer model was initially introduced in [2] as an alternative to Convolutional Neural Network (CNN) models for addressing

CONTACT Kian Ming Lim  kmlim@mmu.edu.my  Faculty of Information Science & Technology, Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang, 75450 Melaka, Malaysia.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

computer vision problems, such as image classification. The self-attention mechanism is the main component that makes this model distinct. This mechanism considers three components: query, key, and value, to compute the attention weight from the image features. Several patches are created as input from the image before passing into the transformer encoder, similar to the original transformer model used to solve natural language processing tasks. Besides utilizing the Vision Transformer as the backbone model, Knowledge Distillation [3] is also applied. The concept of Knowledge Distillation involves a teacher-student model where the teacher model guides the student model's learning process by transferring the rich knowledge encoded in the teacher model, which has better generalization ability. By distilling the knowledge from a larger ViT model into a smaller one, KDViT strikes a balance between model performance and computational efficiency, making it well-suited for resource-constrained medical imaging applications. Gou et al. [4] conducted a survey on knowledge distillation, identifying various types of knowledge distillation algorithms, including knowledge-based types, distillation schemes, teacher-student architectures, and more. The implementation of the proposed KDViT model primarily focuses on the teacher-student architecture. It is crucial to determine the right teacher model to effectively capture and distill knowledge for the student model. Additionally, data augmentation and class weighting are employed to address limited training samples and class imbalance problems. The implementation of fine-tuning and early stopping in the KDViT model lies in their synergistic enhancement of model performance. Fine-tuning allows the model to adapt its pre-trained parameters to better suit the specific characteristics of the dataset, thereby refining its ability to capture relevant features and patterns. Moreover, by stopping training at the optimal point, early stopping helps prevent the model from memorizing noise in the training data and promotes better generalization to unseen data.

The dominant contributions of this paper to perform image classification for medical images are:

- Vision Transformer model is employed as the foundation of the proposed KDViT allows for efficient representation learning in medical image classification. ViT's self-attention mechanism, considering query, key, and value components, allows the model to learn the complex patterns and dependencies present in medical images, contributing to enhanced feature extraction and classification accuracy.
- Knowledge Distillation in the proposed KDViT facilitates a seamless transfer of rich knowledge from a complex teacher model to a smaller student model. This process boosts the student model's generalization capability, enabling it to inherit the valuable insights encoded in the teacher model. The result

is a more compact yet proficient model in medical image classification tasks, contributing to improved performance and computational efficiency.

- The proposed KDViT incorporates data augmentation and class weight techniques, addressing challenges related to limited training samples and class imbalance in medical image datasets. Data augmentation diversifies the dataset, enhancing the model's ability to generalize to different variations. Meanwhile, class weight adjustments mitigate the impact of imbalanced class distributions, contributing to a more robust and accurate classification model.

2. Related works

Commonly, research in medical image classification can be broadly divided into two main directions: Convolutional Neural Network models (CNN-based models) and Non-Convolutional Neural Network models (Non-CNN models).

2.1. Convolutional neural network

Convolutional Neural Networks (CNNs) have emerged as a powerful and widely utilized tool in medical image classification. These neural networks are particularly well-suited for processing visual data, making them widely applied in tasks like diagnosing diseases from medical images. Yadav and Jadhav [5] employed three distinct methods to train Convolutional Neural Network models using a chest X-ray dataset for pneumonia classification. These methods encompassed a Support Vector Machine (SVM) classifier utilizing oriented fast and rotated binary (ORB), transfer learning models, and the implementation of a capsule network. From the experimental result, it was found that CNN-based transfer learning produced the best result of 95.4% accuracy, followed by the capsule network and the SVM with ORB model. Later, a medical image dataset called the SARS-Cov-2 CT-scan dataset was introduced by [6]. This dataset contained only two cases, which were COVID and non-COVID cases. An eXplainable Deep Learning model (xDNN) was also proposed using CNN architecture to perform CT-Scan image classification and the result of 97.38% accuracy and 97.36% AUC was recorded as its best performance. A similar work by Yang et al. [7] also introduced the COVID-CT dataset which contained two labels. In their work, multi-task learning and self-supervised learning were utilized in their pre-trained transfer learning models, DenseNet-169 and ResNet-50. The combination of the transfer learning model and contrastive self-supervised learning was proved to achieve 89% accuracy and 98% of AUC score. Another larger dataset of CT-Scan medical images was presented by [8]. The chest CT images with three cases (non-informative CT, positive CT and negative CT)

and clinical features (CFs) were collected from 1,170 patients. A patient-centric resource, iCTCF, was developed to manage and share the data. A 13-layer CNN model was created to predict the disease of COVID-19 and the result of 97.8% AUC score was recorded. In [9], another transfer learning CNN model, GoogleNet, was used to classify a total of 749 chest CT Scan images. This solution obtained a validation accuracy of 82.14%. GoogleNet, which was known as Inception-V1, integrated multi-scale convolutional transformations through the concept of splitting, transforming, and merging. A later research by Saleh et al. [10] proposed combining CNN with the SVM algorithm to detect lung cancer when utilizing chest medical images as the input. The model produced 97.91% of accuracy and 1.0 of AUC score. Aytac et al. [11] introduced a novel adaptive momentum applied to a CNN model for testing and classifying over three different medical datasets, including brain cancer, chest X-Ray, and CT-scan images. Comparisons of adaptive momentum optimizers, such as Stochastic Gradient Descent (SGD) and Adam, revealed that SGD produced the best result by reducing the classification loss from 6.12% to 5.44%. Salehi et al. [12] published a study of CNN models which were used in medical imaging field. The study explained clearly for each of the components in CNN model and the application of CNN model in transfer learning technique. CNN model became the alternative solution of machine learning algorithm because of its capability to learn high-level features from the images. Wang and Yang [13] also made a survey on the application of CNN models in image classification, elucidating various model architectures and their performance on ImageNet and CIFAR10 datasets, which demonstrated improvements of at least 2% to 3%. Different network optimization methods were introduced as well. Some studies [14, 15] also demonstrated the ability of DenseNet model to solve object detection problem and achieved high accuracy score when the models were enhanced with the integration of YOLO algorithm. DenseNet model was able to improve feature propagation and support feature reuse.

2.2. Non-convolutional neural network

CNN in medical image classification may face challenges in instances of limited annotated medical datasets, hindering the model's ability to generalize effectively. Additionally, CNNs might struggle with interpretability, making it challenging to explain the reasoning behind specific classifications, a crucial aspect in the medical field for gaining trust and understanding from healthcare professionals. In light of these challenges, researchers have begun to explore alternative approaches to CNNs for medical image classification. Rajeev et al. [16] proposed a solution utilizing

Recurrent Neural Network (RNN) with Long-Short-Term Memory (LSTM) and batch normalization. The optimal batch size was determined using the Particle Swarm Optimization (PSO) algorithm, with RNN efficiently eliminating noise in medical CT scan images. Their model achieved better performance in Peak Signal-to-Noise Ratio (PSNR) and Mean Square Error (MSE) metrics. Furthermore, Zhou et al. [17] introduced a self-pretraining Masked Autoencoder (MAE) and segmentation network with a Vision Transformer model for medical segmentation and classification. Usman et al. [18] demonstrated that transformer-based models outperformed CNN models such as ResNet-50, Inception-V3, and VGG-16 in classifying X-ray images from two data sources, with the Vision Transformer achieving 87% accuracy and 86% F1-score. In another work by Almalik et al. [19], a Self-Ensembling Vision Transformer (SEViT) was proposed for classifying chest X-ray images and diabetic retinopathy. The ViT model achieved 96.38% and 97.64% accuracy for chest X-rays and diabetic retinopathy, respectively. Shaker and Xiong [20] employed LSTM and RNN to classify lung partial images, achieving 95.93% accuracy, surpassing VGG and Inception models. Leamons et al. [21] compared Vision Transformer, CNN, and Residual Neural Network models for medical image classification, with the transformer model producing superior results at 93% accuracy, compared to CNN (81.4%) and RNN (87.9%) models. Lee et al. [22] studied the use of Vision Transformer models on smaller datasets and proposed using Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) to address the problem of low locality inductive bias inherent in Vision Transformer models. Their experiments showed an improvement in accuracy scores of approximately 4.08% for both CIFAR100 and T-ImageNet datasets. Jamil and Roy [23] proposed a Non-CNN model which used Vision Transformer to detect Valvular Heart Diseases based on the cardiac phonocardiogram. Self-attention mechanism was proved to be effective when extracting the MFCC and LPCC features from the images and the model achieved a score of 99.90% for accuracy, specificity, sensitivity and F1-score. Jiang et al. [24] also proposed a Non-CNN model which utilizing Multiple Graph Learning Neural Networks (MGLNNs) model to perform semi-supervised classification.

Non-CNN based models have garnered considerable attention in recent years as a focus of study and exploration. Bi et al. [25] has studied various type of transformer models which are used in computer vision for image and video datasets. Transformer-based model has outperformed existing models in terms of accuracy metric. Therefore, vision transformer models are generally perceived as more intricate compared to traditional CNN models, particularly regarding the number of parameters and computational demands. For example,

employing Vision Transformer (ViT) models for classification tasks can present challenges, especially when working with smaller datasets. To overcome this obstacle, Knowledge Distillation is employed, involving the transfer of knowledge from a larger, more complex teacher model to a smaller student model with fewer parameters for training. Moreover, addressing the inherent limitations of smaller datasets, techniques like data augmentation and class weight adjustments play a crucial role. Data augmentation involves artificially expanding the dataset by performing several transformations on the existing images, helping the model generalize better to diverse scenarios. Additionally, incorporating class weights aids in mitigating the impact of imbalanced class distributions, ensuring that the model does not exhibit bias towards the dominant class and thus enhancing overall performance and accuracy.

3. Knowledge distillation of vision transformer (KDViT)

In this work, Knowledge Distillation of Vision Transformer, known as KDViT, is proposed for solving the problem of classifying COVID-19 CT scan images. The KDViT model, depicted in Figure 1, employs Vision Transformer as the backbone model (as shown in Figure 2) with a teacher-student architecture. Given the input CT scan image, the KDViT model starts with data augmentation, generating additional images with diverse characteristics such as random rotation by a specific degree and horizontal flipping. As pointed out by [26], training ViT models with larger datasets can lead to better results. These augmented images undergo training in the teacher model, facilitating the distillation of knowledge from the larger (teacher) to the smaller (student) model. The ViT-Base 32 (ViTB-32)

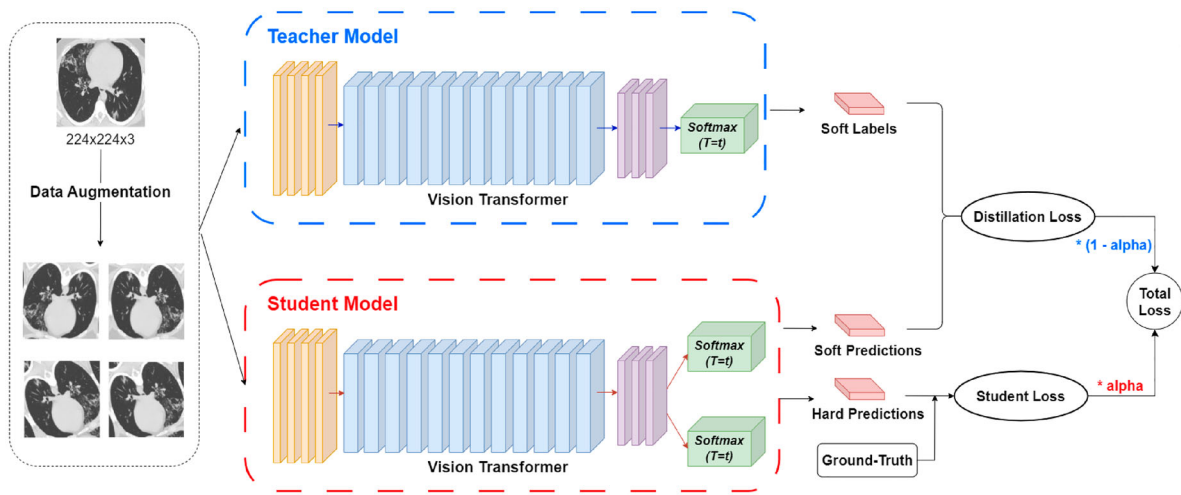


Figure 1. The overview of the proposed KDViT model.

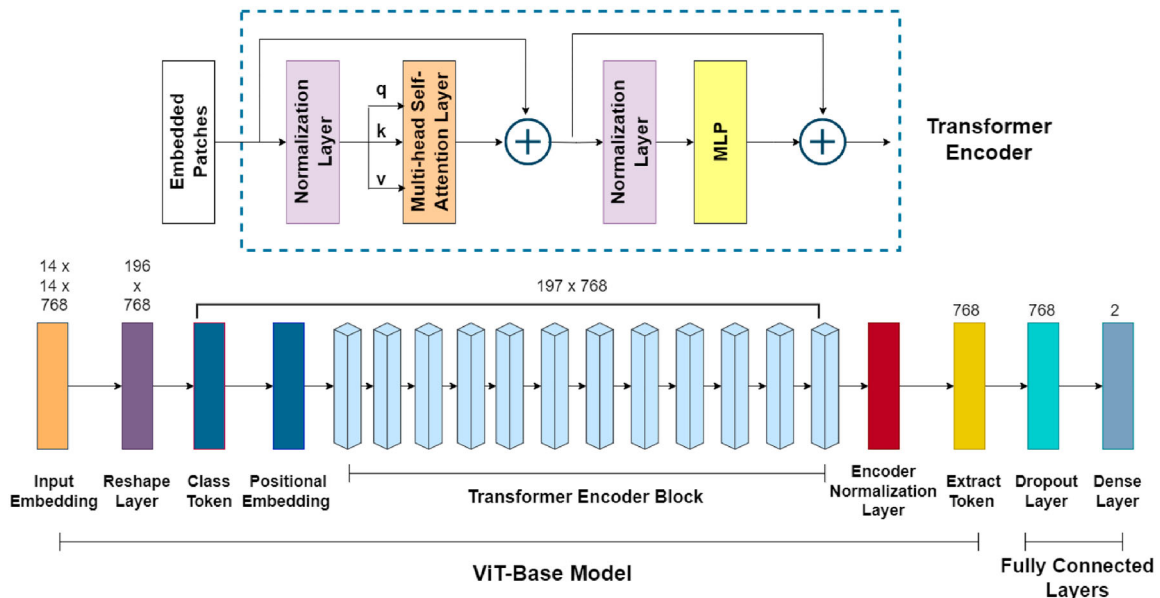


Figure 2. The illustration of the backbone model of Vision Transformer in the proposed KDViT model.

model serves as the teacher, while the ViT-Base 16 (ViTB-16) model functions as the student, both utilizing Vision Transformer as their feature extractor. Han et al. [26] performed a survey on Vision Transformer models. The study explained the architectures of Convolution and Attention with their backbone models. The transformer blocks in the Vision Transformer model leverage self-attention mechanisms to analyze complex relationships in CT-Scan images and capture both local and global information. A few layers of fully-connected layers are connected to the ViT-Base model, and the last layer is an output layer with a softmax activation function to compute class probabilities. For the classification task, a weighted average of two loss functions, distillation loss and student loss, is employed to enhance knowledge transfer between the student and teacher models. Algorithm 1 presents the comprehensive training steps for the KDViT model, encompassing the training of both the teacher and student models. In overview, the logic behind the KDViT model lies in transferring knowledge from a larger teacher model, such as a pre-trained Vision Transformer, to a smaller student model. This process aims to distil the rich information captured by the teacher model into a more compact student model, enabling it to achieve comparable performance with reduced computational complexity. By leveraging the soft labels and representations learned by the teacher model, the student model learns to mimic the teacher's behaviour, leading to improved generalization and performance on the target task.

Algorithm 1 Algorithm of training steps for the KDViT model.

Require: Teacher Model $M_{teacher}$, Student Model $M_{student}$, CT-Scan training data, D_{train}

- 1: Initialize temperature parameter: τ
- 2: **for** n epochs **do**
- 3: **for** b batch size **do**
- 4: $x, y \leftarrow D_{train}$
- 5: **while** Not converged **do**
- 6: Forward pass through teacher model:
 $M_{teacher}(x, y)$
- 7: Calculate teacher logits: $Z_{teacher}$
- 8: Forward pass through student model:
 $M_{student}(x, y)$
- 9: Calculate student logits: $Z_{student}$
- 10: Calculate knowledge distillation loss:
 $L_K =$
 $DistillationLoss(Z_{teacher}, Z_{student}, \tau)$
- 11: Backpropagate and update student weights: ∇L_K
- 12: **end while**
- 13: **end for**
- 14: **end for**

3.1. Data augmentation

In Figure 1, data augmentation is deployed to generate additional training samples, addressing the challenge of limited data availability in the dataset. This technique significantly expands the effective dataset size by introducing diverse transformations and alterations to the original data. The primary purpose is to enhance the proposed KDViT model, enabling it to handle a wide range of data scenarios during model training, leading to improved performance and resilience. The key advantage of data augmentation is its role in enriching the model's learning experience. Exposure to diverse data variations allows the model to develop a more comprehensive understanding of underlying patterns, resulting in enhanced generalization for accurate classification of new, unseen data. Garcea et al. [27] explained that medical images required suitable augmentation methods to generate more samples and allowed the deep learning model to improve the performance. In this paper, it is found that data augmentation appears to offer greater advantages in classification tasks compared to segmentation tasks after conducting the survey on many researches, and not every augmentation technique guarantees an improvement in the results, as its effectiveness still depends on its compatibility with the medical images. Goceri [28] introduced the data augmentation techniques for different medical image modalities such as MRI, CT images, mammography and eye fundus images. The effectiveness for some augmentation techniques are compared with the quantitative results which obtained from the experiments. The advantages and disadvantages are discussed as well.

From the study [28], transformation-based augmentation methods are easier to be implemented when compared to Generative Adversarial Network (GAN) based augmentation methods. Therefore, the proposed KDViT will be enhanced with transformation-based augmentation methods. Transformation-based augmentation methods encompasses an array of image manipulation techniques, including but not limited to rotations, cropping, scaling, flipping vertically or horizontally, and adjustments to brightness, contrast, and colour. These operations simulate real-world variations, such as different orientations and transformations, which the model should be equipped to handle. Figures 3 and 4 show the examples of the CT scan images when random flipping and random rotation are applied.

3.2. Vision transformer

The augmented samples are input into the presented Knowledge Distillation of Vision Transformer (KDViT) model, where both teacher and student models utilize Vision Transformer (ViT) as the backbone. This type

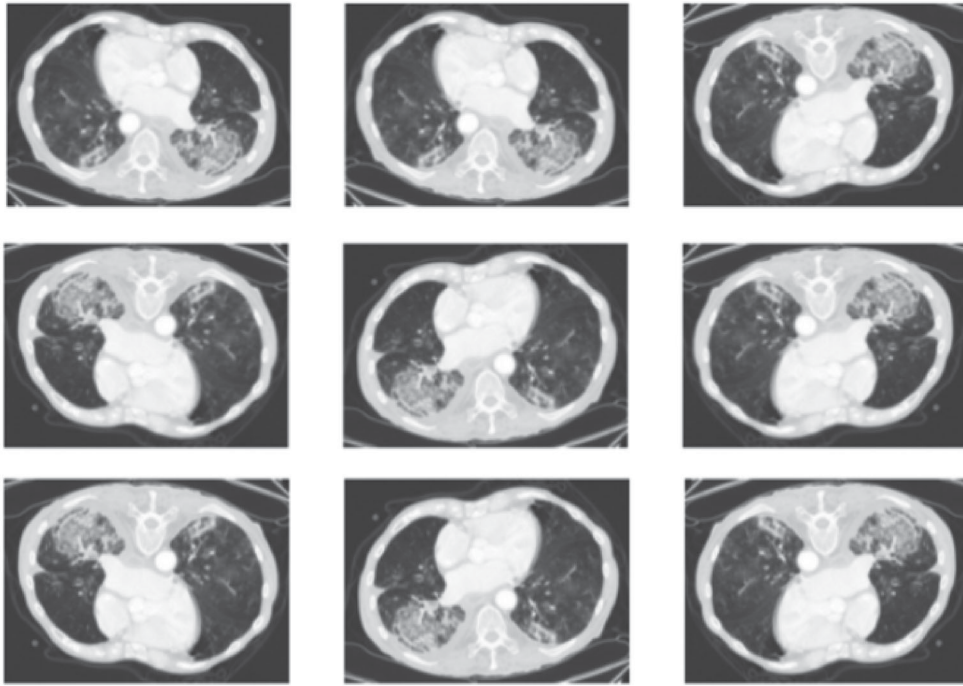


Figure 3. Sample augmented CT scan images when random flip technique is applied.

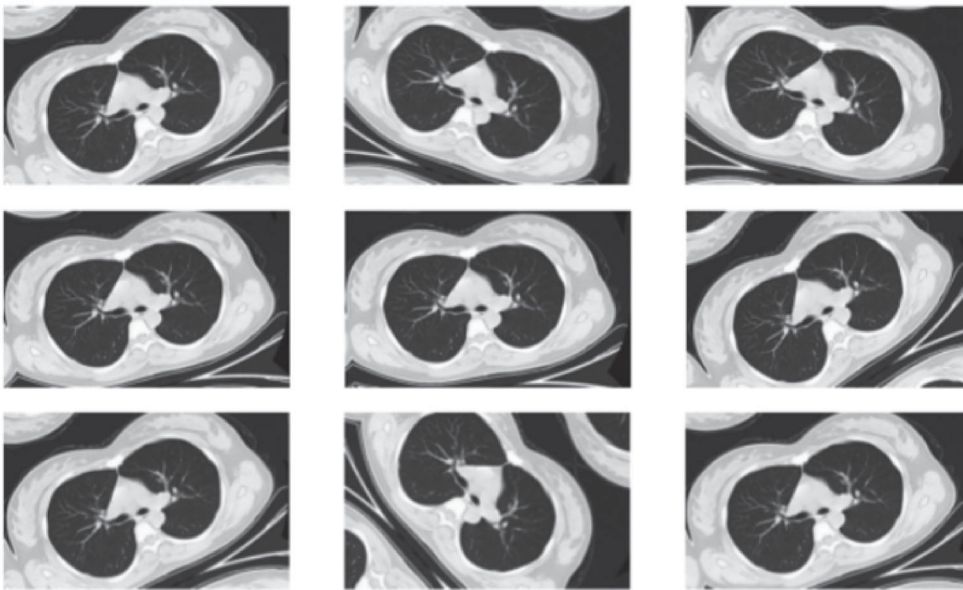


Figure 4. Sample augmented CT scan images when random rotation technique is applied.

of model is initially used for Natural Language Processing (NLP), has transitioned into Computer Vision tasks. The augmented images are tokenized by patching process as shown in Figure 5.

Firstly, the 3-colour channel image with width w and height h , denoted as $I \in \mathbb{R}^{w \times h \times 3}$, will be chunked into smaller patches with a shape of (p, p, c) and a predefined patch size p , considering the channel value of the image, as described in (1) and (2). The patches will then be flattened into a sequence of vectors with a certain dimension, producing linear embeddings of dimension d as the output, E . The position of each patch will be recorded by a positional embedding, E_{pos} , which

is added to the vector. This addition enables the model to not just encapsulate the content but also the spatial position of these patches within the image. A dimension of $(1, d)$ learnable class token, x_{cls} , is also attached at the beginning of the patch embeddings' sequence. The equations involved in this process are expressed in (3) and (4).

$$I_p \in \mathbb{R}^{n \times p^2 \times 3} \quad (1)$$

$$n = \frac{w \times h}{p^2} \quad (2)$$

$$E \in \mathbb{R}^{d \times (p^2 \times 3)}, \quad E_{pos} \in \mathbb{R}^{d \times n} \quad (3)$$

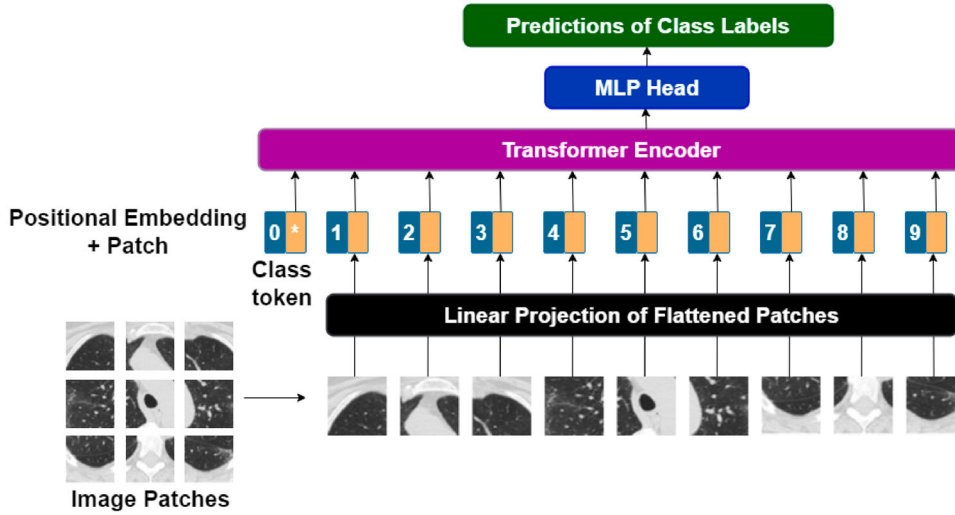


Figure 5. The processes in the Vision Transformer.

$$z_0 = [x_{cls}; x_p^1 E; x_p^2 E; \dots; x_p^n E] + E_{pos} \quad (4)$$

where n is the number of patches created from an image, z_0 is the sequence of embedded patches which will be fed into transformer encoder, and x_p^n is the n th patch.

The encoder will receive embedded patches as the input, as depicted in Figure 5. According to [29], a transformer encoder consists of three essential layers: the Multi-Head Self-Attention layer (MSA), Layer Normalization (LN), and Multi-Layer Perceptions Layer (MLP). Residual connections are used between the layers. The Layer Norm layer is implemented to help stabilize and improve the model's training. Its primary function is to normalize the activation (outputs) of neurons within each layer of the network. Layer Normalization involves scaling the features of each training sample using their mean and standard deviation. These scaled features are then multiplied by learnable scaling and added with shifting factors during the training process. In contrast, Residual connections provide alternate paths for gradients, addressing the issue of vanishing gradients in extremely deep architectures.

The output from the Layer Norm is then passed into the MSA Layer. The purpose of the MSA layer, introduced and explained in detail by [30], is to capture the relationships and dependencies between the patches and learn contextual representations by updating the weight based on the similarities of the features. This process is completed by mapping each patch to three vectors, denoted as q for query, k for key, and v for value. This self-attention mechanism is represented as scaled dot-product attention, expressed in (5). Furthermore, the attention weight of a Query vector and Key vector is multiplied by the Value vector. This calculation of three vectors is considered as single-head attention. For the multi-head attention mechanism, a single attention with d -dimensional queries (d_q), keys (d_k), and values (d_v) is carried out for h times, as denoted in (6).

Utilizing multiple attention heads enables the model to simultaneously focus on different segments of the input and learn diverse features and representations. Each attention head can focus on a different aspect or relationship within the data, making the model more expressive. A residual connection is implemented to add the original input to the result obtained from the MSA layer. The last step in the transformer encoder is image classification through a Layer Norm and MLP layer. The classification token, cls , is extracted and used as the classification head in the MLP layer. This layer is crucial for producing an accurate classification result.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

where T is the length of patched embeddings, and d_k is the hidden dimensional for the key.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)$$

$$W^0 \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (6)$$

3.3. Knowledge distillation

In this work, we propose Knowledge Distillation of Vision Transformer (KDViT) models, a method involving two different-sized models. The larger model, referred to as the teacher model, disseminates its learned information and knowledge to the smaller student model by minimizing the loss function. The ViT-Base 32 (ViTB-32) model acts as the teacher, whereas the ViT-Base 16 (ViTB-16) model serves as the student.

During training, the student model learns not only from the original dataset but also from a modified dataset containing soft labels generated by the teacher model. These soft labels, represented as probability distributions over classes, replace traditional hard one-hot encoded labels. Two loss functions are considered during training: Distillation loss, measuring the disparity

between the student and teacher model's predictions on the soft labels, encourages mimicking the teacher's behaviour. Simultaneously, Student loss, the standard loss function for the original task (e.g. cross-entropy loss for classification), ensures the student model performs well on the primary task. The overall loss is a linear combination of both, with a hyperparameter, the temperature (τ), controlling the relative weight. The softmax layer converts logits (z_i) into probabilities (q_i), as expressed in (7).

$$q_i = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)} \quad (7)$$

The temperature parameter influences the smoothness of soft labels; a higher temperature leads to softer labels providing more training information, while a lower temperature makes labels closer to one-hot distributions, yielding sharper predictions.

Knowledge distillation finds application in various machine learning domains, including computer vision, natural language processing, and speech recognition. It enables accurate model deployment in resource-constrained scenarios by compressing large models into smaller ones while maintaining performance and accuracy. Smaller models are more efficient for inference, making them suitable for deployment on devices with limited computational resources.

3.4. Class weight

Class weight is a well-established strategy extensively used to address the challenges presented by imbalanced datasets. The three datasets used in this study manifest distinct distributions of data across their labelled classes. Notably, the iCTCF dataset [8] illustrates an imbalance in image count among the three classes, with the largest class comprising 9979 images, while the other two classes exhibit disparities of 5978 and 4274 images, respectively. These imbalances result in an uneven training set, disproportionately favouring the majority class and compromising the model's ability to generalize effectively across all classes.

To effectively tackle this problem, the class weight technique is applied, acting as a mechanism to rebalance the impact of different classes during the overall training process. For the minority class, its contribution to the total loss in model training is strategically enhanced. This adjustment counteracts the disproportionate influence of the majority class, helping to prevent the pitfalls of overfitting driven by an overwhelming majority class representation.

Through the incorporation of the class weight technique, the proposed KDViT model can distribute its learning efforts more equitably across all classes. This promotes more balanced and accurate classification outcomes within the context of imbalanced datasets,

ultimately contributing to a more robust and reliable model performance.

3.5. Fine tuning

In the proposed Knowledge Distillation of Vision Transformer (KDViT) models, fine-tuning significantly enhances the model's performance on specific tasks. Fine-tuning involves refining a pre-trained Vision Transformer model by training it with a downstream dataset, tailoring its knowledge to the intricacies of the target problem, in this case, the classification of COVID-19 CT scan images. The fine-tuning process begins by loading the pre-trained ViT-Base 32 (ViTB-32) model, which serves as the teacher, and the ViT-Base 16 (ViTB-16) model, which functions as the student. These models have already acquired general features from a broader dataset, providing a solid foundation for understanding complex patterns in images. The pre-trained models are then fine-tuned on a dataset specific to COVID-19 CT scan images. During this phase, several hyperparameters are adjusted to optimize the model's performance, such as the learning rate of the optimizer, batch size, and the number of neurons in the fully-connected layers.

Spolaôr et al. [31] conducted a study on the effectiveness of applying fine-tuning to the VGG16 transfer learning model to address the challenge of learning from small medical datasets. The study demonstrated that fine-tuning a pre-trained model by transferring generic features learned to the medical domain features can achieve better results than some existing models. One of the key advantages of fine-tuning in the context of the proposed KDViT models is the efficient transfer of knowledge from the pre-trained models to the task-specific COVID-19 CT scan classification. Pre-trained models bring a wealth of knowledge about general image features, significantly accelerating the learning process on the new dataset. This approach allows the models to leverage the pre-existing understanding of complex relationships in images, facilitating quicker convergence and enhancing their ability to generalize to COVID-19 CT scan patterns. Additionally, fine-tuning enables the models to adapt to the unique characteristics of the medical imaging domain. COVID-19 CT scans may exhibit distinct features that are not present in the original dataset used for pre-training. Fine-tuning allows the models to refine their representations, capturing nuances specific to COVID-19 cases. This adaptability is crucial in medical image classification, where precise identification of disease-related patterns is essential.

In short, fine-tuning in the proposed KDViT models optimizes the knowledge transfer from pre-trained Vision Transformer models to the task of COVID-19 CT scan classification. It expedites model training, enhances generalization capabilities, and enables

adaptation to the specific nuances of medical imaging datasets, contributing to more accurate and robust classification outcomes.

3.6. Early stopping

In this work, early stopping is employed as a strategic approach to enhance training efficiency and avoid overfitting during the learning process. It entails continuously evaluating the model's performance, such as accuracy and loss, on a validation set throughout training and halting the training procedure when signs of performance degradation emerge.

One of the primary advantages of early stopping in KDViT models is its role in preventing overfitting. Overfitting occurs when a model not only learns inherent patterns within the training data but also noise or specific characteristics unique to the training set, resulting in poor generalization to new data. By monitoring validation performance and stopping training when the model's accuracy or loss on the validation set starts to decline, early stopping ensures that the model does not become excessively specialized to the training samples. Moreover, early stopping enhances resource efficiency by halting the training process upon detecting diminishing returns on the validation set and avoiding unnecessary computational costs associated with prolonged training.

In the context of classifying COVID-19 CT scan images in the medical field, where datasets may be limited, early stopping is particularly valuable. Limited datasets increase the risk of overfitting, and early stopping acts as a regularization method to mitigate this risk, striking a balance between model complexity and dataset size. This ensures that KDViT models achieve optimal performance without compromising their ability to generalize to new, unseen COVID-19 CT scan images.

4. Experiment and result

This paper assesses the proposed KDViT using three benchmark datasets: the SARS-Cov-2 CT-scan dataset [6], COVID-CT dataset [7], and integrative CT images and CFs for COVID-19 (iCTCF) dataset [8]. Additionally, the performance of the KDViT model is compared and contrasted against several existing models.

4.1. Dataset

The first dataset is called SARS-CoV-2 CT-scan dataset, introduced by [6]. The total number of images in this dataset are 2482, with two cases, COVID and Non-COVID. Among all the 2482 CT scan images, there are a total of 1252 SARS-CoV-2 cases belonging to infected patients, further broken down into 32 males and 28 females. Another 1230 CT scan images correspond to non-SARS-CoV-2 infected patients, consisting of 30 males and females, respectively.

Yang et al. [7] introduced the second dataset named COVID-CT dataset, which consists of a total of 746 images from resources available online, such as medRxiv, bioRxiv, LUNA, Radiopaedia website, and PubMed Central (PMC). The ratio of males to females in this dataset is 86 and 51, respectively, and the age range covers from 1 to 81 years old. Ages between 31–41 years old contributed the most cases, followed by the 41–51 age group, whereas 11 to 21 years old has the least cases collected.

The third dataset, which is also the largest dataset, was introduced by [8]. In this dataset, it included three classes: positive CT images (pCT), negative CT images (nCT), and non-informative CT images (NiCT), with a total of 19,685 images available in the dataset. All the data collected from 1170 patients in this third dataset are all under the approval of institutional ethical committees of Union Hospital and Liyuan hospital in China.

An overview of the three datasets used in this research work is outlined in Table 1.

4.2. Experimental settings and evaluation metrics

In this work, each dataset is divided into training, validation, and testing sets for model training, following the standard protocol in existing works. The image dimensions used for each dataset vary, as summarized in Table 2. During data splitting, stratified data splitting is employed to create subsets of data while maintaining a similar class distribution to the original dataset. This ensures that the model's performance evaluation is not skewed by class imbalances that may exist in the data, a critical consideration when dealing with imbalanced datasets such as the iCTCF dataset, where some classes may have significantly fewer samples than others. Additionally, data augmentation is applied to every image, and various transformation techniques are

Table 1. Details of the three datasets used.

Dataset	Type of Labels	Number of image	Total Images
SARS-CoV-2 CT-scan	COVID-19	1252	2482
	NON-COVID-19	1230	
COVID-CT	COVID-19	349	746
	Non-COVID-19	397	
iCTCF	pCT	4001	19,685
	nCT	9979	
	NiCT	5705	

Table 2. The experimental settings for the three datasets.

Dataset	Training ratio	Validation ratio	Testing ratio	Image dimension
SARS-CoV-2 CT-scan	60	20	20	224 × 224
COVID-CT	80	10	10	224 × 224
iCTCF	60	10	30	96 × 96

tested to identify the combination that yields the best results, including random rotation, horizontal or vertical flipping, random cropping, brightness adjustments, and others. Moreover, class weights are computed to address class imbalance issues in the dataset, ensuring that the performance metrics accurately reflect the model's ability to classify all classes, including minority classes.

In the experiment, the ViTB-32 model, acting as the teacher model, has a total of 87,455,232 trainable parameters, with a patch size set to 32. Meanwhile, the student model, ViTB-16, has a total of 85,798,656 trainable parameters, and a patch size of 16. Fine-tuning is applied to both the teacher and student models to enhance classification results, with the last 5 to 12 layers of ViT models unfrozen to enable modification and updates of pre-trained weights over epochs, empowering the model to adapt to specific tasks or data distributions. The dropout rate is set to 0.5, and the Adam optimizer is used with a learning rate of 0.001 and a clipping value of 0.5. In the Knowledge Distillation setting, the temperature parameter is also tested with different values to obtain the optimal value. Both teacher and student models are implemented with binary cross-entropy loss for binary-class datasets and categorical cross-entropy loss for three-class datasets. For the distillation loss function, Kullback-Leibler divergence is computed to measure the loss between the true label and predicted label, teaching the student to make predictions similar to the teacher. The teacher model is trained for 100 epochs, followed by training the student model for 80 epochs. The batch size is set to 64 for the SARS-CoV-2 CT-scan and COVID-CT datasets and 128 for the iCTCF dataset. Early stopping is implemented to track validation set accuracy throughout the training process and preserve the model's best-trained weights.

Implementing the proposed KDViT model to solve medical image classification presents several challenges. Firstly, ensuring access to high-quality and diverse CT scan datasets accurately representing COVID-19 cases is challenging due to variations in image acquisition protocols and data labelling inconsistencies, necessitating data cleaning and preprocessing before training the model. Limited RAM memory and GPU capacity are critical challenges that can significantly hinder the training process, particularly with large datasets and high-resolution images. Insufficient RAM may lead to out-of-memory errors, preventing

the model from loading the entire dataset into memory for training. Similarly, running out of GPU memory mid-task, often referred to as an Out-of-Memory (OOM) error, is a common issue, with 8.8% of failed deep learning jobs attributed to GPU memory exhaustion. This makes it the leading cause of OOM failures in deep learning tasks. Consequently, smaller image sizes and batch sizes are employed to complete model training and testing processes.

To assess performance, various metrics such as accuracy, precision, recall, and F1-score are calculated when the KDViT model is tested by predicting the testing set of each dataset. These metrics are computed based on the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Each metric is denoted in (8)–(11).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (11)$$

4.3. Experiment results and analysis

Within this section, we showcase the experimental outcomes and result analysis of COVID-19 classification utilizing CT-scan images employing our proposed model, KDViT. We conduct an ablation analysis of the model's performance over three benchmark datasets, discussing the impact of each enhancement technique used to formulate the KDViT model, including class weight, data augmentation, fine-tuning, and early stopping. The classification accuracy, precision, recall, F1-Score, and AUC score achieved by the KDViT model with the applied enhancements are recorded in Table 3 for the SARS-CoV-2 CT-scan dataset, Table 4 for the COVID-CT dataset, and Table 5 for the iCTCF dataset. The performance evaluation of the KDViT model, incorporating various techniques, reveals that slight improvements in classification accuracy are observed when using class weight and data augmentation to address dataset limitations, such as class imbalance and a small dataset size. Significantly, fine-tuning contributes to enhanced KDViT performance as the model effectively learns features from CT scan images. In the SARS-CoV-2 CT-scan and COVID-CT datasets, KDViT model shows an improvement of approximately more than 10% in accuracy after applying fine-tuning by unfreezing the last few layers in the student model in KDViT.

We also compare the classification results between the teacher model and the student model, as shown

Table 3. The classification result of KDViT model on SARS-CoV-2 CT-scan dataset with different enhancements (DA = data augmentation, CW = class weights, FT = fine-tuning and ES = early stopping).

Enhancement	Accuracy	Precision	Recall	F1-score	AUC
KDViT	88.10	88.00	88.00	88.00	88.00
KDViT + CW	88.30	88.00	88.00	88.00	88.00
KDViT + CW + DA	88.91	89.00	89.00	89.00	89.00
KDViT + CW + DA + FT	96.57	97.00	97.00	97.00	97.00
KDViT + CW + DA + FT + ES	98.39	98.00	98.00	98.00	98.00

Table 4. The classification result of KDViT model on COVID-CT dataset with different enhancements (DA = data augmentation, CW = class weights, FT = fine-tuning and ES = early stopping).

Enhancement	Accuracy	Precision	Recall	F1-Score	AUC
KDViT	71.42	72.00	71.00	71.00	71.00
KDViT + CW	73.33	73.00	73.00	73.00	73.00
KDViT + CW + DA	74.29	74.00	74.00	74.00	74.00
KDViT + CW + DA + FT	87.61	88.00	88.00	88.00	88.00
KDViT + CW + DA + FT + ES	88.57	89.00	89.00	89.00	89.00

in Table 6. Noticeably, it is evident that all the student models perform better after receiving knowledge distillation from the teacher model, resulting in higher accuracy scores. This highlights the significant enhancement that knowledge distillation can bring to the performance of smaller student models. Specifically, in the COVID-CT dataset, the student model ViTB-16 achieves a notable improvement in classification accuracy, increasing from 83.81% to 88.57% after receiving knowledge distillation from the teacher model. In the iCTCF dataset, while the improvement is not as substantial, the student model still maintains a classification accuracy of over 99.10%. Similarly, in the SARS-CoV-2 CT-scan dataset, the student model demonstrates a modest accuracy improvement of 0.41%, rising from 97.98% to 98.39%. This observation underscores the suitability of simpler, lower-complexity models with fewer parameters for smaller datasets, as they tend to generalize better and are less susceptible to overfitting issues.

Furthermore, to evaluate the classification result based on the testing set of each dataset, we have plotted and presented confusion matrices in Figure 6.

These confusion matrices illustrate the model's ability to classify images accurately for all three datasets. For instance, in Figure 6(a), the SARS-CoV-2 CT-scan dataset, only 8 out of 496 images are classified incorrectly, and in Figure 6(b), the COVID-CT dataset, only 12 out of 105 images are predicted incorrectly. For the multi-class classification in the iCTCF dataset, KD-ViT achieves highly accurate predictions, correctly identifying approximately 98.89% of non-informative CT-Scan cases, 99.16% of negative cases, and 99.5% of positive cases in the testing set.

Furthermore, we analyse and determine the classification performance of the KDViT model on the testing set, benchmarking it with some state-of-the-art solutions [6, 32–39], as presented in Table 7. It is noteworthy that the proposed KDViT model outperforms state-of-the-art methods for COVID-19 diagnosis on CT-scan images. In the SARS-CoV-2 CT-scan dataset, KDViT demonstrates outstanding performance with an accuracy of 98.39% and achieves 98% for precision, recall, F1 score, and AUC. Similarly, in the smaller COVID-CT dataset, the proposed model still delivers commendable results, achieving an accuracy of 88.57% and 89%

Table 5. The classification result of KDViT model on iCTCF dataset with different enhancements (DA = data augmentation, CW = class weights, FT = fine-tuning and ES = early stopping).

Enhancement	Accuracy	Precision	Recall	F1-Score
KDViT	90.98	90.89	90.98	90.82
KDViT + CW	92.41	92.62	92.41	92.45
KDViT + CW + DA	92.80	92.99	92.80	92.84
KDViT + CW + DA + FT	98.61	98.62	98.61	98.61
KDViT + CW + DA + FT + ES	99.15	99.15	99.15	99.15

Table 6. The classification result of the teacher and student models in the KD-ViT model.

Dataset	KDViT Model	Trainable Parameters	Classification Result	
			Accuracy	Loss
SARS-CoV-2 CT-scan	Teacher model: ViTB-32	35,440,896	0.9798	0.1303
	Student model: ViTB-16	21,265,152	0.9839	0.0882
COVID-CT	Teacher model: ViTB-32	35,440,896	0.8381	0.9822
	Student model: ViTB-16	21,265,152	0.8857	1.0446
iCTCF	Teacher model: ViTB-32	70,880,256	0.9912	0.0426
	Student model: ViTB-16	35,440,896	0.9915	0.0403

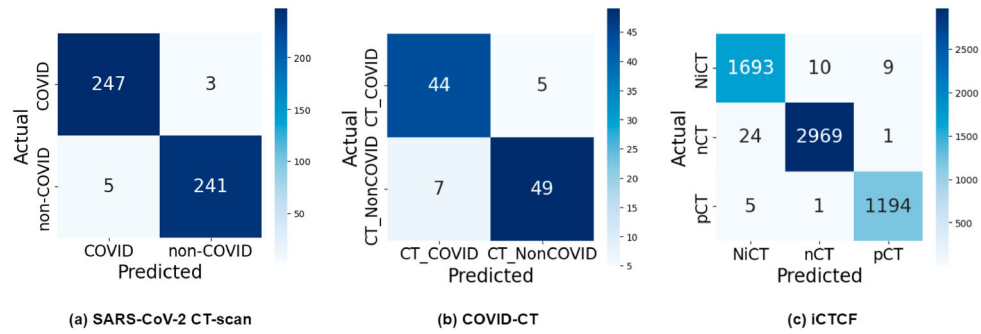


Figure 6. Confusion matrix of the proposed KDViT model on all the datasets.

Table 7. Classification result comparison of the proposed KDViT with some existing methods.

Dataset	Source of Method	Method	Performance Result (%)				
			Accuracy	Precision	Recall	F1-Score	AUC
SARS-CoV-2 CT-scan	[32]	SepNorm + Contrastive COVID-Net	90.83	95.75	85.89	90.87	96.24
	[33]	CNN	91.13	95.00	94.00	95.00	–
	[34]	VGG-19	95.00	95.00	94.00	95.00	–
	[35]	DenseNet201	96.25	96.29	96.29	96.29	–
	[6]	xDNN	97.38	99.16	95.53	97.31	97.36
	Proposed KDViT	KD-ViT	98.39	98.00	98.00	98.00	98.00
COVID-CT	[32]	SepNorm + Contrastive COVID-Net	78.69	78.02	79.71	78.83	85.32
	[36]	ResNet50	82.91	–	77.66	–	–
	[37]	Ensemble model	85.00	85.70	85.20	–	–
	[38]	Targeted self supervision with DenseNet169	86.21	–	–	87.04	86.09
	Proposed KDViT	KD-ViT	88.57	89.00	89.00	89.00	89.00
iCTCF	[39]	Deep learning 1-Simple CNN	98.83	98.83	98.85	98.84	–
		Deep learning 2-Multiheaded	98.49	98.49	98.52	98.50	–
		Fusion deep learning model	99.08	99.08	99.08	99.08	1.00
	Proposed KDViT	KD-ViT	99.15	99.15	99.15	99.15	–

for the remaining evaluation metrics. In the iCTCF dataset, our proposed method attains an impressive classification accuracy of 99.15%, slightly surpassing the performance of the compared models.

5. Conclusion

In summary, the proposed KDViT model leverages the Vision Transformer as its backbone, integrating Knowledge Distillation to transfer knowledge from a larger, high-parameter model to a more compact one. Vision Transformers, equipped with image patching and self-attention mechanisms within transformers, play a pivotal role in interpreting and learning robust representations and features from images. Knowledge Distillation enhances training efficiency, especially in resource-limited scenarios, without compromising performance. Techniques like data augmentation and class weighting substantially improve model performance by addressing limited samples and class imbalances. Data augmentation augments the dataset by generating additional samples through diverse transformations, such as rotation and random flipping, thereby enriching the learning process. Class weight adjusts the contribution of each class to the total loss during training, giving more weight to minority classes. This helps to mitigate the effects of class imbalance and prevent the model from being biased towards the majority class, resulting in more balanced and accurate predictions.

Fine-tuning allows the model to adapt to the specific data by updating its weights during training, improving its ability to learn relevant features from the CT scan images. Early stopping prevents overfitting by monitoring the model's performance on a validation set and stopping the training process when performance no longer improves, thus helping to achieve better generalization and prevent the model from memorizing noise in the training data.

Through experiments, optimal hyperparameter settings are identified to achieve peak performance in terms of accuracy, precision, recall, F1 score, and AUC score. The KD-ViT model demonstrates its prowess in classifying COVID CT-scan images, surpassing several state-of-the-art methods with the highest accuracy rates of 98.39%, 88.57%, and 99.15% for the SARS-CoV-2 CT-scan dataset, COVID-CT dataset, and iCTCF dataset, respectively.

However, there are some limitations of the proposed KDViT model. The proposed model has high sensitivity to teacher model quality. The effectiveness of knowledge distillation heavily depends on the quality and architecture of the teacher model. If the teacher model is not well-trained or does not capture relevant features, the distilled knowledge may not be beneficial for the student model. Another limitation of the proposed solution is having longer training time of the models. Knowledge distillation involves training both the teacher and student models simultaneously,

which can increase the overall training time required compared to training a single model. Additionally, fine-tuning the student model to distill knowledge from the teacher model may necessitate more iterations and epochs to converge effectively, further extending the training duration.

For forthcoming work, exploring recent deep learning models to achieve better results in reduced training time and optimized computational and memory resources is recommended. Additionally, studying and comparing different knowledge distillation methods, such as Self Distillation and Cross-Domain Knowledge Distillation, can provide insights into further improving model performance.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research in this work was supported by the Telekom Malaysia Research Development under grant number RDTC/231084.

ORCID

Kian Ming Lim  <http://orcid.org/0000-0003-1929-7978>

References

- [1] World Health Organization, Coronavirus disease (COVID-19). 2023; Available from: <https://www.who.int/health-topics/coronavirus>.
- [2] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16 × 16 Words: transformers for image recognition at scale; 2020.
- [3] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network; 2015.
- [4] Gou J, Yu B, Maybank SJ, et al. Knowledge distillation: a survey. *Int J Comput Vis.* 2020;129:1789–1819.
- [5] Yadav SS, Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. *J Big Data.* 2019;6:1–18.
- [6] Soares E, Angelov P, Biaso S, et al. SARS-CoV-2 CT-scan dataset: a large dataset of real patients CT scans for SARS-CoV-2 identification. *medRxiv*; 2020. p. 1–8.
- [7] Yang X, He X, Zhao J, et al. COVID-CT-Dataset: a CT Scan Dataset about COVID-19. *arXiv preprint arXiv:2003.13865*; 2020. p. 1–14.
- [8] Ning W, Lei S, Yang J, et al. Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. *Nat Biomed Eng.* 2020;4:1197–1207.
- [9] Alsharman N, Jawarneh I. GoogleNet CNN neural network towards chest CT-coronavirus medical image classification. *J Comput Sci.* 2020;16:620–625.
- [10] Saleh AY, Chin CK, Penshie V, et al. Lung cancer medical images classification using hybrid cnn-svm. *Int J Adv Intell Informatics.* 2021;7:151–162.
- [11] Aytaç UC, Güneş A, Ajlouni N. A novel adaptive momentum method for medical image classification using convolutional neural network. *BMC Med Imaging.* 2022;22:34.
- [12] Salehi AW, Khan S, Gupta G, et al. A study of CNN and transfer learning in medical imaging: advantages, challenges, future scope. *Sustainability.* 2023;15:5930.
- [13] Wang W, Yang Y. Development of convolutional neural network and its application in image classification: a survey. *Opt Eng.* 2019;58:1–19.
- [14] Roy AM, Bose R, Bhaduri J. A fast accurate fine-grain object detection model based on YOLOv4 deep neural network. *Neural Comput Appl.* 2022;34:3895–3921.
- [15] Roy AM, Bhaduri J. DenseSPH-YOLOv5: an automated damage detection model based on DenseNet and swin-transformer prediction head-enabled YOLOv5 with attention mechanism. *Adv Eng Inform.* 2023;56:102007.
- [16] Rajeev R, Samath JA, Karthikeyan NK. An intelligent recurrent neural network with long short-term memory (LSTM) BASED batch normalization for medical image denoising. *J Med Syst.* 2019;43:234.
- [17] Zhou L, Liu H, Bae J, et al. Self pre-training with masked autoencoders for medical image classification and segmentation; 2022.
- [18] Usman M, Zia T, Tariq A. Analyzing transfer learning of vision transformers for interpreting chest radiography. *J Digit Imaging.* 2022;35:1445–1462.
- [19] Almalik F, Yaqub M, Nandakumar K. Self-ensembling vision transformer (SEViT) for robust medical image classification. In: Wang L., Dou Q, Fletcher PT, Speidel S, Li S, editors. *Medical image computing and computer assisted intervention - MICCAI 2022.* MICCAI 2022. Lecture notes in computer science. Cham: Springer; 2022. vol. 13433. p. 376–386.
- [20] Shaker AM, Xiong S. Lung image classification based on long-short term memory recurrent neural network. *J Phys Conf Ser.* 2023;2467:012007.
- [21] Leamons R, Cheng H, Al Shami A. Vision transformers for medical images classifications. In: Arai K, editor. *Intelligent systems and applications.* IntelliSys 2022. Lecture notes in networks and systems, Cham: Springer; 2023. vol. 544. p. 319–325.
- [22] Lee SH, Lee S, Song BC. Vision transformer for small-size datasets; 2021.
- [23] Jamil S, Roy AM. An efficient and robust phonocardiography (PCG)-based valvular heart diseases (VHD) detection framework using vision transformer (ViT). *Comput Biol Med.* 2023;158:106734.
- [24] Jiang B, Chen S, Wang B, et al. MGLNN: semi-supervised learning via multiple graph cooperative learning neural networks. *Neural Netw.* 2022;153:204–214.
- [25] Bi J, Zhu Z, Meng Q. Transformer in computer vision. In 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology, Fuzhou, China, CEI 2021; 2021. p. 178–188.
- [26] Han K, Wang Y, Chen H, et al. A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell.* 2023;45:87–110.
- [27] Garcea F, Serra A, Lamberti F, et al. Data augmentation for medical imaging: a systematic literature review. *Comput Biol Med.* 2023;152:106391.
- [28] Goceri E. Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev.* 2023;56:12561–12605.
- [29] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need; 2017.

- [30] Park N, Kim S. How do vision transformers work? In ICLR 2022 – 10th International Conference on Learning Representations; 2022.
- [31] Spolaôr N, Lee HD, Mendes AI, et al. Fine-tuning pre-trained neural networks for medical image classification in small clinical datasets. *Multimed Tools Appl.* **2023**;83:27305–27329.
- [32] Wang Z, Liu Q, Dou Q. Contrastive cross-site learning with redesigned net for COVID-19 CT classification. *IEEE J Biomed Health Inform.* **2020**;24:2806–2813.
- [33] Aguirre-Alvarez PA, Diaz-Carmona J, Arredondo-Velázquez M. Flexible Systolic Hardware Architecture for Computing a Custom Lightweight CNN in CT Images Processing for Automated COVID-19 Diagnosis. In: Mahmud M, Mendoza-Barrera C, Kaiser MS, Bandyopadhyay A, Ray K, Lugo E, editors. *Proceedings of trends in electronics and health informatics. TEHI 2022. Lecture notes in networks and systems.* Singapore: Springer; **2023**. vol. 675. p. 17–34.
- [34] Carvalho ED, Silva RR, Araújo FH, et al. An approach to the classification of COVID-19 based on CT scans using convolutional features and genetic algorithms. *Comput Biol Med.* **2021**;136:104744.
- [35] Jaiswal A, Gianchandani N, Singh D, et al. Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *J Biomol Struct Dyn.* **2021**;39:5682–5689.
- [36] Loey M, Manogaran G, Khalifa NEM. A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images. *Neural Comput Appl.* **2020**;0123456789:1–13.
- [37] Gifani P, Shalhaf A, Vafaezadeh M. Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans. *Int J Comput Assist Radiol Surg.* **2021**;16:115–123.
- [38] Ewen N, Khan N. Targeted self supervision for classification on a small covid-19 ct scan dataset. In *Proceedings – International Symposium on Biomedical Imaging, Nice, France; 2021 Apr.* p. 1481–1485.
- [39] Abdar M, Salari S, Qahremani S, et al. Uncertainty-FuseNet: robust uncertainty-aware hierarchical feature fusion with ensemble Monte Carlo dropout for COVID-19 detection; 2021. p. 1–16.