Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

IEViT: An enhanced vision transformer architecture for chest X-ray image classification

Gabriel Iluebe Okolo[a,1], Stamos Katsigiannis[b,2,*], Naeem Ramzan[a,3]

[a] University of the West of Scotland, High St., Paisley, PA1 2BE, UK
[b] Durham University, Stockton Road, Durham, DH1 3LE, UK

ARTICLE INFO

Article history:
Received 6 June 2022
Revised 2 August 2022
Accepted 14 September 2022

Keywords:
Chest radiography
X-Rays
Vision transformer
Image classification
Deep learning

ABSTRACT

Background and Objective: Chest X-ray imaging is a relatively cheap and accessible diagnostic tool that can assist in the diagnosis of various conditions, including pneumonia, tuberculosis, COVID-19, and others. However, the requirement for expert radiologists to view and interpret chest X-ray images can be a bottleneck, especially in remote and deprived areas. Recent advances in machine learning have made possible the automated diagnosis of chest X-ray scans. In this work, we examine the use of a novel Transformer-based deep learning model for the task of chest X-ray image classification.

Methods: We first examine the performance of the Vision Transformer (ViT) state-of-the-art image classification machine learning model for the task of chest X-ray image classification, and then propose and evaluate the Input Enhanced Vision Transformer (IEViT), a novel enhanced Vision Transformer model that can achieve improved performance on chest X-ray images associated with various pathologies.

Results: Experiments on four chest X-ray image data sets containing various pathologies (tuberculosis, pneumonia, COVID-19) demonstrated that the proposed IEViT model outperformed ViT for all the data sets and variants examined, achieving an F1-score between 96.39% and 100%, and an improvement over ViT of up to +5.82% in terms of F1-score across the four examined data sets. IEViT's maximum sensitivity (recall) ranged between 93.50% and 100% across the four data sets, with an improvement over ViT of up to +3%, whereas IEViT's maximum precision ranged between 97.96% and 100% across the four data sets, with an improvement over ViT of up to +6.41%.

Conclusions: Results showed that the proposed IEViT model outperformed all ViT's variants for all the examined chest X-ray image data sets, demonstrating its superiority and generalisation ability. Given the relatively low cost and the widespread accessibility of chest X-ray imaging, the use of the proposed IEViT model can potentially offer a powerful, but relatively cheap and accessible method for assisting diagnosis using chest X-ray images.

1. Introduction

Chest X-ray (CXR) imaging is one of the most widely utilised medical imaging techniques for detecting and diagnosing diseases [60], including pneumonia, tuberculosis, COVID-19, malignancy, and others [2,55,74]. Its great advantage lies in its relatively low cost, high accessibility, and easy operation [45]. Important information about a patient's health can be extracted from chest X-ray images, and manual analysis and detection by chest X-ray imaging of marks and signs of diseases is done by expert radiologists. Interpretations can be difficult and it is a long and complicated process. In addition, the requirement for expert radiologists can be a bottleneck, especially in remote or deprived areas. To address this issue, recent research has focused on the use of machine learning methods for automated diagnosis, with various approaches gaining popularity and aiming to become an important tool for clinicians [19,20,29,37].

The cutting edge development of general-purpose graphics processing unit (GPU) hardware [59], medical image analysis techniques [57], and deep learning techniques [17], has allowed sci-

[*] Corresponding author.
E-mail address: stamos.katsigiannis@durham.ac.uk (S. Katsigiannis).
[1] 0000-0002-1624-6668
[2] 0000-0001-9190-0941
[3] 0000-0002-5088-1462

https://doi.org/10.1016/j.cmpb.2022.107141

entists to automatically detect diseases using chest X-ray images [50], and design powerful computer-aided diagnosis (CAD) systems [5,38,70]. The potential gains from automated X-ray diagnosis are increased sensitivity for findings, automation of tedious daily tasks, prioritisation of time-sensitive cases, and solving the issue of radiologists not always being available in remote areas or developing countries [8]. Chest X-ray imaging is also relatively cheap and accessible, with modern digital radiography machines being affordable even in under-developed countries [45].

In recent years, the use of deep learning methods and convolutional neural networks (CNNs) has proven to be very effective in various computer vision-oriented tasks, including image classification, image segmentation, and object detection [35]. Based on this success, AI/machine learning systems have been extensively researched to automate image analysis in the clinical field, such as for tuberculosis diagnosis [40], and the detection of pneumonia [33], COVID-19 [41,43], pneumothorax [25], pneumoconiosis [69], lung cancer [28], as well as for other radiology analysis tasks [76].

Sharma et al. [56] and Stephen et al. [65] used CNNs for the purpose of detecting pneumonia in chest X-ray images, achieving an accuracy of 90.68% and 93.73% respectively. Khan et al.'s [32] CoroNet CNN-based architecture achieved an overall accuracy of 89.6% for the same task. Saraiva et al. [54] used a multilayer perceptron (MLP) and a CNN for the same task, achieving an average accuracy of 92.16% with the MLP, and 94.40% with the CNN, whereas Ayan et al. [4] used the well-known Xception and VGG16 CNN models, along with transfer learning and fine-tuning techniques for training, with test results showing that VGG16 outperformed Xception, reaching accuracies of 87% and 82% respectively. In a later work, Ayan et al. [3] proposed a CNN-based ensemble method, achieving an accuracy of 95.83%. Liang et al. [36] combined dilated convolutions and residual approaches, achieving an accuracy of 90%. Okolo et al. [41] examined the use of various CNN architectures to detect COVID-19 and viral pneumonia in chest X-rays, achieving a maximum accuracy of 98%.

Tuberculosis diagnosis from chest X-rays using machine learning has also been widely researched. Yadav et al. [75] implemented various deep learning techniques with the aim to detect Tuberculosis in chest X-ray images. Significant improvement was achieved using fine tuning and multiple data augmentation, reaching an accuracy of 94.89%. Deep learning was also used for the same task by Hooda et al. [24], achieving an accuracy of 82.09%, and by Pasa et al. [42], achieving an accuracy of 86.82%, whereas Evalgelista et al.'s [18] proposed CNN-based computer-aided diagnosis approach reached an accuracy of 88.76%. Rahman et al. [48] carried out an experiment for the detection of tuberculosis using various CNN models and transfer learning, achieving an F1-score for the best performing model of 96.47%.

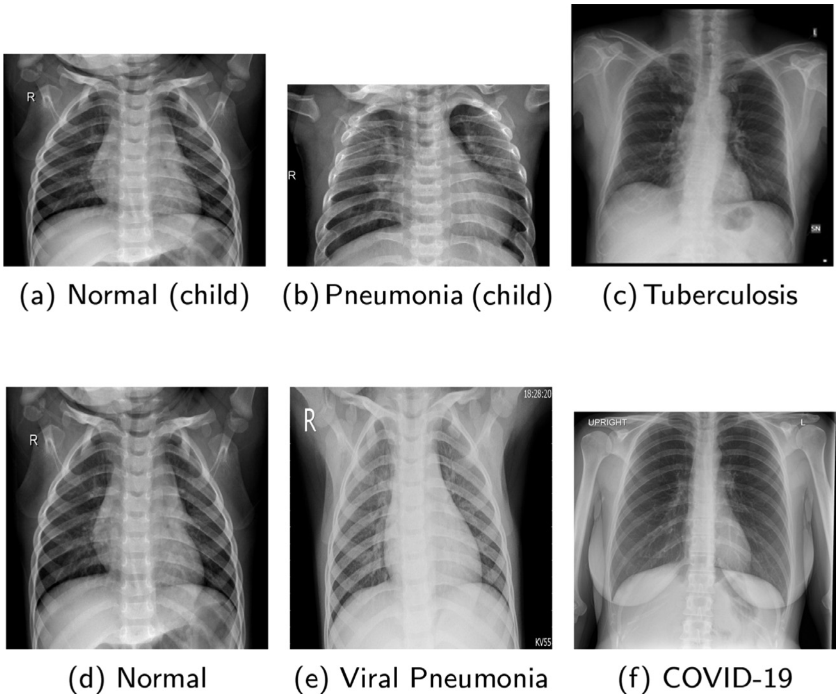
The success of CNNs in generic image classification tasks led to their widespread adoption for the classification of various types of specialised images with or without transfer learning, as explained above. Nevertheless, other architectures [16] have recently gained attention due to their enhanced performance compared to CNNs. Inspired by the success of Transformers in Natural Language Processing (NLP), researchers have recently tried to apply Transformers directly to images [21], using several approaches. Some works combined CNN architectures with self-attention. Ramachandran et al. [51] suggested substituting all convolutional layers for self-attention layers instead of using self-attention layers on top of them, whereas Bello et al. [7] proposed to improve CNNs by substituting some convolutional layers for self-attention layers. However, these approaches exhibited high computational cost due to the large size of the images that led to an enormous growth in the complexity of self-attention. Wu et al. [73] used convolutional layers to extract feature maps of the input images that were then fed to stacked Transformer layers for computing the final output.

The Vision Transformer (ViT) [16] architecture is the first attempt for a pure Transformer architecture that achieved state–of-the-art results on image classification. ViT adapts the BERT [15] Transformer-based architecture for understanding language to image classification via some modifications. To this end, images are first divided into rectangular patches, which are then treated as tokens for which embeddings are computed. After the addition of positional embeddings to encode the structure of the image, the patch embeddings are fed to a series of Transformer layers for the creation of the final feature map. Experimental results showed that the various ViT variants are able to achieve better performance in ImageNet, CIFAR, and VTAB classification compared to common CNNs [16].

Following the success of the ViT model, various variants have been proposed. Touvron et al. [64] introduced the data-efficient image transformers (DeiT), which performed better than the original ViT architecture on ImageNet, achieving an accuracy of 85.2%. Yuan et al. [77] proposed a layer-wise Tokens-To-Token Vision Transformer (T2T-ViT), to encode the significant structure for every token, contrary to the simple tokenisation utilised in ViT [16]. They tested their method on ImageNet and achieved an 82.3% Top-1 accuracy, outperforming the original ViT architecture. Chen et al. [9] proposed an original double-branch vision transformer that also employed a cross-attention-based token fusion scheme, achieving a 82.8% Top-1 accuracy on ImageNet. Li et al. [34] introduced locality to vision transformers by incorporating 2D depthwise convolutions. This basic concept was inspired from comparisons between inverted residual blocks and feed-forward networks, and achieved a 94.2% Top-5 accuracy on ImageNet, the best among the other evaluated architectures. Wang et al. [68] proposed the Pyramid Vision Transformer (PVT), which incorporates the pyramid structure from CNNs to the vision transformer architecture, achieving a 18.3% Top-1 error on ImageNet. Liu et al. [39] proposed a hierarchical vision Transformer architecture that utilises shifted windows and has reduced computational complexity with respect to image size compared to the original ViT model.

Motivated from the importance of automating the diagnosis of chest X-ray images, from the success of self-attention in transformer models, and from the increasing interest in models utilising attention mechanisms within convnets [78], in this work, we introduce the *Input Enhanced Vision Transformer* (IEViT), an enhanced Vision Transformer architecture for classifying chest X-ray images. The proposed model builds on the ViT architecture and introduces a CNN block that is used to create an embedding of the full input image, which is then iteratively fed to each Transformer encoder layer by concatenating the image embedding to the output of each Transformer encoder layer. The proposed model, as well as the original ViT model, were then evaluated on four chest X-ray image data sets. Classification experiments demonstrated that IEViT outperformed ViT for all the variants and data sets examined, achieving a maximum F1-score of 98.48% for Normal vs. Children Pneumonia, 100% for Normal vs. Tuberculosis, 98.05% for Normal vs. Viral Pneumonia vs. COVID-19, and 96.39% for Normal vs. COVID-19.

The contribution of this work can be summarised as follows: **(i)** We evaluate the performance of the ViT "Base" and "Large" variants on four chest X-ray image data sets for the task of classifying various pathological conditions reflected on the images. **(ii)** We propose IEViT, a novel enhanced vision transformer architecture that improves the performance of both the "Base" and "Large" ViT variants on all the examined chest X-ray image data sets. **(iii)** We provide a detailed performance evaluation of the pro-



(a) Normal (child) (b) Pneumonia (child) (c) Tuberculosis

(d) Normal (e) Viral Pneumonia (f) COVID-19

Fig. 1. Sample X-ray images from the examined data sets: (a, b) Kermany, (c) Tuberculosis, (d, e) COVID-19 radiography, (f) COVIDx.

posed enhanced Vision Transformer architecture on the examined chest X-ray image data sets.

2. Methods

In this work, we propose the Input Enhanced ViT (IEViT), an enhanced vision transformer architecture for the classification of chest X-ray images. The proposed architecture was evaluated against the original ViT architecture on four chest X-ray image data sets that contained images associated with children pneumonia, tuberculosis, COVID-19, viral pneumonia, as well as with no pathology (healthy individuals). We opted to evaluate our proposed method on images originating from various sources, acquired using different radiography devices, and associated with various pathological conditions, in order to validate its robustness and generalisation in comparison to the original ViT architecture.

2.1. Data sets

The four data sets used in this work are the following: (i) The *Kermany* et al. data set [30], which consists of normal chest X-ray scans and scans associated with pneumonia of children patients (1–5 years old), acquired from Kermany et al. [31]. (ii) The *Tuberculosis Chest X-ray Database* [48], which contains normal chest X-ray images and chest X-ray images associated with Tuberculosis, collected from three publicly accessible databases [1,6,47]. (iii) The *COVID-19 radiography database* [11,49], which consists of normal chest X-ray images, chest X-ray images associated with viral pneumonia, and chest X-ray images associated with COVID-19, acquired from multiple sources [22,26,27,30,47,71]. (iv) The *COVIDx* data set [67], which contains normal and COVID-19-positive chest X-ray images, sourced from five different publicly available data repositories [12–14,46,47]. To the best of our knowledge, COVIDx has the most COVID-19 positive images out of all the publicly available data sets.

Sample X-ray images from the examined data sets are provided in Fig. 1, whereas the number of images per class in each data set is presented in Table 1.

Table 1
Data set details.

Data set	Number of images per class			
	Normal	Pneumonia	Tuberculosis	COVID-19
Kermany	1,349/234	3,883/624	-	-
Tuberculosis	3,500	-	700	-
COVID-19	10,202	1,345	-	3,615
COVIDx	13,992/200	-	-	16,490/200

Note: */* refers to the official train/test split. 80%/20% split used otherwise.

2.2. The original vision transformer (ViT) architecture

Proposed by Dosovitskiy et al. [16], the Vision Transformer (ViT) architecture is a pure transformer approach that can perform on par or even outperform common CNN architectures for image classification when trained on large amounts of image data. The input image to the ViT architecture is split into square patches, with each patch flattened and concatenated across the image's channels in order to create a vector representation of each image patch. An embedding of each vector is then created via linear projection to a specific dimension. A learnable position embedding is also added to each patch in order to allow the ViT model to learn about the structure of the input images. The patch embeddings are then fed to stacked Transformer layers and the output is passed to an MLP head for the final classification. Details about the various ViT variants are provided in Table 2.

2.3. The proposed input enhanced vision transformer (IEViT)

The proposed IEViT approach is partially motivated by the ResNet [23] architecture, which introduced the skip connection, i.e. the addition of the original input to the output of each convolutional block, which brought about the concept of a residual network. Our approach builds on this concept for modifying the ViT [16] architecture. To this end, a representation of the original input image is iteratively added to the output of each Transformer encoder layer. This is achieved by first designing a convolutional

Table 2
Details of the ViT and IEViT model variants.

Model	Patch size ($P \times P$)	Layers (L)	Hidden size (D)	MLP size	Heads	Params
ViT-B/16	16 × 16	12	768	3072	12	85.8M
ViT-B/32	32 × 32	12	768	3072	12	87.4M
ViT-L/16	16 × 16	24	1024	4096	16	303.3M
ViT-L/32	32 × 32	24	1024	4096	16	305.5M
IEViT-B/16	16 × 16	12	768	3072	12	90.8M
IEViT-B/32	32 × 32	12	768	3072	12	92.4M
IEViT-L/16	16 × 16	24	1024	4096	16	309.9M
IEViT-L/32	32 × 32	24	1024	4096	16	312.1M

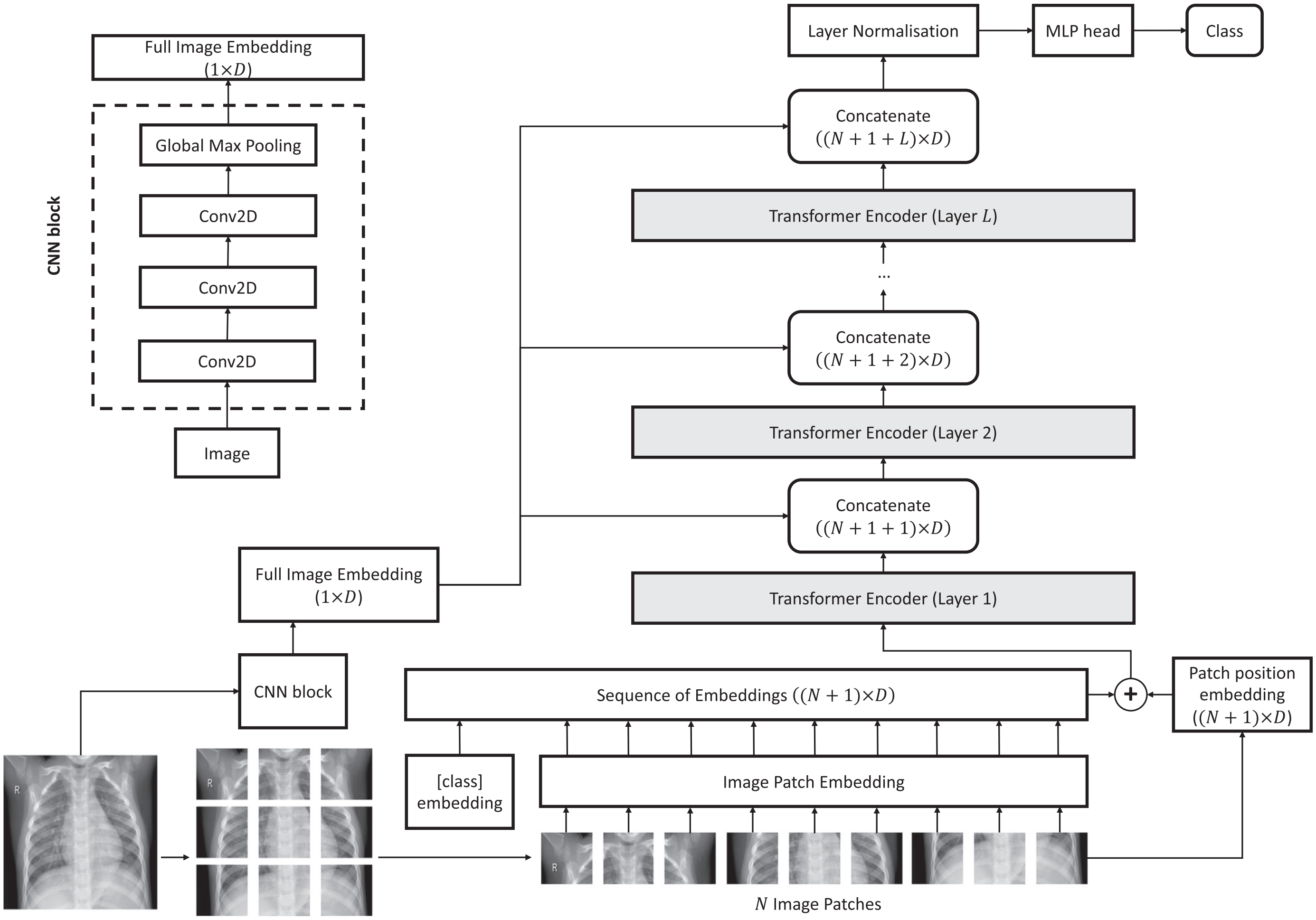

Full Image Embedding (1×D)
Global Max Pooling
Conv2D
Conv2D
Conv2D
CNN block
Image
Layer Normalisation
MLP head
Class
Concatenate ((N + 1 + L)×D)
Transformer Encoder (Layer L)
...
Concatenate ((N + 1 + 2)×D)
Transformer Encoder (Layer 2)
Concatenate ((N + 1 + 1)×D)
Full Image Embedding (1×D)
Transformer Encoder (Layer 1)
CNN block
Sequence of Embeddings ((N + 1)×D)
+
Patch position embedding ((N + 1)×D)
[class] embedding
Image Patch Embedding
R
N Image Patches

Fig. 2. Overview of the proposed IEViT model.

block in parallel with the ViT network. The CNN block takes as an input the whole input image and outputs an embedding of the whole image, which is then iteratively concatenated to the output of each Transformer encoder layer, thereby making the network to always "remember" the full image at the end of each transformer block output. An overview of the proposed IEViT architecture is provided in Fig. 2.

The proposed CNN block consists of stacked 2D convolutional layers, followed by a 1D global maximum pooling layer, as shown in Fig. 2. Three 2D convolutional layers were used in our experiments, with the first using 16 filters with a kernel size of 3, the second using 256 filters with a kernel size of 5, and the third using D filters with a kernel size of 5. The input to the CNN block is an image of size $H \times W \times C$, where H is the height of the image, W the width, and C the number of channels. Then, the max pooling layer is used in order to compute the output vector x_{img} of size D, which is a mapping of the input image to D dimensions.

For the vision transformer part of the architecture, similar to ViT [16], the input image is divided into $N = \frac{H \cdot W}{P^2}$ patches, where (P, P) is the resolution of each patch, which are then flattened across all dimensions to create a sequence of flattened patches x_p of total size $N \times (P^2 \cdot C)$. Then, the patch embeddings are created by mapping the patches to D dimensions using a trainable linear projection, as follows:

$$\mathbf{z}_p = [x_p^1 E, x_p^2 E, \ldots, x_p^N E], \quad E \in \mathbb{R}^{(P^2 \cdot C) \times D} \tag{1}$$

Then, as in ViT [16], a learnable embedding x_{class} of size D, similar to BERT's [class] token [15], is prepended to the sequence of patch embeddings, while position embeddings are added to the sequence of patch embeddings in order to incorporate information about the position of each patch in the original image. A standard learnable 1D position embedding of size D is used for each of of the $N+1$ embeddings in the patch embedding matrix, with the position for the additional x_{class} embedding set as 0. The initial patch embeddings $\mathbf{z}_0$ are computed as follows:

$$\mathbf{z}_0 = [x_{class}, \mathbf{z}_p] + E_{pos}, \quad E_{pos} \in \mathbb{R}^{(N+1)\times D} \tag{2}$$

Then, L transformer encoder layers [16,66] are stacked, with $\mathbf{z}_0$ being the input of the first layer. $\hat{\mathbf{z}}_l$ is then computed by concatenating the image embedding x_{img} to the output $\mathbf{z}_l$ of each Transformer encoder layer l, $l = 1, 2, ., L$, as follows:

$$\hat{\mathbf{z}}_l = [\mathbf{z}_l, x_{img}], \quad \hat{\mathbf{z}}_l \in \mathbb{R}^{(N+1+l)\times D} \tag{3}$$

$\hat{\mathbf{z}}_l$ is then fed to the next Transformer encoder layer, or to the next layer of the architecture if $l = L$. As a result, contrary to ViT, each Transformer encoder layer in IEViT is fed a representation of the whole input image in addition to the output of the previous Transformer encoder layer, as also shown in Fig. 2.

As shown in Table 2, the original ViT architecture supported different configurations that were adopted from BERT [15]. Following the same approach, the proposed IEViT architecture adopts the "Base" and "Large" models using a similar notation to ViT. To this end, IEViT-B/16 refers to the "Base" variant with an image patch size of 16 × 16, whereas similarly, IEViT-L/32 refers to the "Large" variant with an image patch size of 32 × 32. Details about the proposed model's variants, as well as the variants of the original ViT are provided in Table 2.

It must also be noted that Keras and the *vit-keras*[4] implementation of ViT were used for all our experiments and as the backbone for the implementation of IEViT.

2.4. Training and classification

To evaluate the proposed IEViT architecture against the original ViT architecture for the four examined chest X-ray data sets, we compared each of the B/16, B/32, L/16, L/32 variants of the proposed architecture against the same variants of the original ViT architecture. Transfer learning was used for both architectures, with the ViT models and the ViT parts of the proposed models being initialised with weights pre-trained on the ImageNet [53] data set. For the proposed architecture, the additional parts were initialised randomly and their weights were trained during the fine-tuning process. For each data set, the classifier on top of each model was set according to the number of classes in the data set, and end-to-end training was used for the fine-tuning.

Training was performed using Cross-Entropy as the loss function (Eq. (4)), the Adam optimiser, a batch size of 16 and a 0.0001 learning rate. The low learning rate was selected in order to better adapt the pre-trained weights to the new data.

$$L_{CE} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{4}$$

with M being the number of classes ($M_{\text{Kermany}} = 2$, $M_{\text{Tuberculosis}} = 2$, $M_{\text{COVID-19}} = 3$, $M_{\text{COVIDx}} = 2$), $y_{o,c}$ having a value of 1 if observation o belongs to class c and a value of 0 if not, and $p_{o,c}$ being the predicted probability that observation o belongs to class c.

Label smoothing was also used for the training process in order to reduce over-fitting and encourage the model to be less confident, thus leading to better generalisation. Label smoothing is a regularisation technique, proposed by Szegedy et al. [62] for improving the performance of the Inception architecture on the ImageNet data set, which has since been adopted by many state-of-the-art deep learning classification approaches [52,79]. When using cross-entropy as the loss function, training aims to minimise L_{CE}, with $y_{o,c}$ being a "hard" target $\in \{0, 1\}$. When using label smoothing, a "soft" target $y^{LS}_{o,c}$ is used instead, computed as:

$$y^{LS}_{o,c} = y_{o,c}(1-\alpha) + \frac{\alpha}{M} \tag{5}$$

where M is the number of classes and α the label smoothing parameter, which was set to $\alpha = 0.1$ in this work.

In addition to label smoothing, data augmentation was used during training, as it has been proven to be an effective tool for image classification [44], and is mostly used in deep learning approaches to increase the amount of training data and assist in avoiding over-fitting [72]. To this end, more training images were created using the original training images by following the augmentation steps described in Table 3. A uniform probability distribution was used to create the random values for the data augmentation procedure, and batches of augmented images were created in real-time during each training procedure using the *Keras ImageDataGenerator* class. It must also be pointed out that data augmentation was used only for training. Consequently, the reported results on the test data refer to original images only.

Table 3
Data augmentation procedure.

Augmentation step	Options/Range
Rotation at random angle	[-40, 40] degrees
Random flipping	{horizontally, vertically}
Random shifting across the height	10% of the total height
Random shifting across the width	10% of the total width
Random shifting of brightness	[0.5, 1.5]
Random shearing in counter-clockwise direction	[0, 0.1] radians
Random zooming within a range	[0.9, 1.1]
Rescaling by factor	1/255

Note: "nearest" image filling mode used where needed.

3. Results

3.1. Classification experiments

The proposed IEViT models, as well as the original ViT models, for the B/16, B/32, L/16, and L/32 variants were evaluated via supervised classification experiments on the four examined chest X-ray data sets for the respective 2-class and 3-class problems, i.e. Normal vs. Pneumonia for the Kermany data set, Normal vs. Tuberculosis for the Tuberculosis data set, Normal vs. COVID-19 vs. Viral Pneumonia for the COVID-19 data set, and Normal vs. COVID-19 for the COVIDx data set. The training set of each data set was further split into a training (90%) and validation (10%) set, whereas final performance results were reported on the unseen test sets in order to provide a fair estimate of classification performance and reduce over-fitting.

For all models, classification performance was measured using the following metrics: Accuracy, Recall (Sensitivity), Precision, and F1-score, which is the harmonic mean of Precision and Recall, defined as:

$$\text{F1-score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

In the case of F1-score, recall, and precision, the metrics were computed for all classes and the average across classes (macro average) was reported as the final value of the metric. It must also be noted that due to the class imbalance of some data sets, F1-score was used as the primary performance benchmark, as classification accuracy can be biased in cases of unbalanced data sets. Tensorflow and the Keras API were employed for the proposed experimental evaluation. It must also be noted that input images were all rescaled to 224 × 224. The classification performance achieved

[4] https://github.com/faustomorales/vit-keras

Table 4
Classification performance (%) on the Kermany et al. children pneumonia data set.

Model	Accuracy	Recall	Precision	F1-score
IEViT-B/16	97.76	98.46	**97.96**	98.21
IEViT-B/32	97.12	98.98	96.50	97.72
IEViT-L/16	97.60	99.23	96.99	98.10
IEViT-L/32	**98.08**	**99.74**	97.25	**98.48**
ViT-B/16	92.63	97.18	91.55	94.28
ViT-B/32	90.71	98.21	88.25	92.96
ViT-L/16	91.99	98.97	89.35	93.92
ViT-L/32	90.22	98.72	87.30	92.66
CNN-block-B	69.55	91.28	69.53	78.94
CNN-block-L	67.95	88.21	69.08	77.48

Note: Results in bold denote the highest value per metric.

Table 5
Classification performance (%) on the Tuberculosis data set.

Model	Accuracy	Recall	Precision	F1-score
IEViT-B/16	99.76	**100.0**	98.59	99.29
IEViT-B/32	**100.0**	**100.0**	**100.0**	**100.0**
IEViT-L/16	99.76	**100.0**	98.59	99.29
IEViT-L/32	99.29	97.14	98.55	97.84
ViT-B/16	99.64	**100.0**	97.90	98.94
ViT-B/32	99.41	96.43	**100.0**	98.94
ViT-L/16	98.81	98.57	94.52	96.50
ViT-L/32	98.69	92.14	**100.0**	95.91
CNN-block-B	89.43	68.21	94.37	73.72
CNN-block-L	88.48	65.36	93.93	70.27

Note: Results in bold denote the highest value per metric.

Table 6
Classification performance (%) on the COVID-19 radiography data set.

Model	Accuracy	Recall	Precision	F1-score
IEViT-B/16	97.93	95.54	98.29	96.82
IEViT-B/32	97.37	95.28	98.34	96.74
IEViT-L/16	96.18	91.17	97.26	93.90
IEViT-L/32	**98.59**	**97.09**	**99.06**	**98.05**
ViT-B/16	97.96	94.17	98.77	96.27
ViT-B/32	97.01	94.36	98.17	96.16
ViT-L/16	96.54	90.02	97.10	92.95
ViT-L/32	96.08	91.40	97.64	94.21
CNN-block-B	81.54	70.24	74.38	71.99
CNN-block-L	83.68	72.00	77.41	74.24

Note: Results in bold denote the highest value per metric.

Table 7
Classification performance (%) on the COVIDx data set.

Model	Accuracy	Recall	Precision	F1-score
IEViT-B/16	**96.50**	**93.50**	**99.47**	**96.39**
IEViT-B/32	95.00	91.00	98.91	94.79
IEViT-L/16	95.25	91.00	99.45	95.04
IEViT-L/32	95.00	90.50	99.45	94.76
ViT-B/16	95.00	90.50	99.45	94.76
ViT-B/32	92.00	85.00	98.84	91.40
ViT-L/16	93.25	87.00	99.43	92.80
ViT-L/32	92.75	88.00	97.24	92.39
CNN-block-B	90.50	90.50	90.50	90.50
CNN-block-L	91.00	91.00	91.00	91.00

Note: Results in bold denote the highest value per metric.

by the examined IEViT and ViT variants on the Kermany, Tuberculosis, COVID-19 and COVIDx data sets is reported in Tables 4, 5, 6, and 7 respectively, in terms of the examined metrics, whereas confusion matrices for the IEViT variants are provided in Fig. 3.

3.2. Classification results per data set

Classification results for the proposed model's variants, as well as for the original ViT model variants, on the Kermany et al. children pneumonia data set are presented in Table 4. From this table, as well as from Table 8 that shows the per variant performance difference in terms of F1-score, it is evident that all the proposed IEViT model's variants outperformed their respective ViT model variants for all the examined metrics. The achieved improvement in terms of F1-score spans from +3.93% for the B/16 variant to +5.82% for the L/32 variant, with an average improvement of +4.67% across all variants. Results demonstrated stability across the examined models, with an average F1-score within the range of 92.66%-94.28% for the original ViT variants, compared to 97.72%-98.48% for the proposed IEViT model's variants. The best performance for the Kermany et al. children pneumonia data set was achieved by the IEViT-L/32 model, reaching an F1-score of 98.48%, an accuracy of 98.08%, a recall of 99.74%, and a precision of 97.25%, as shown in Table 4.

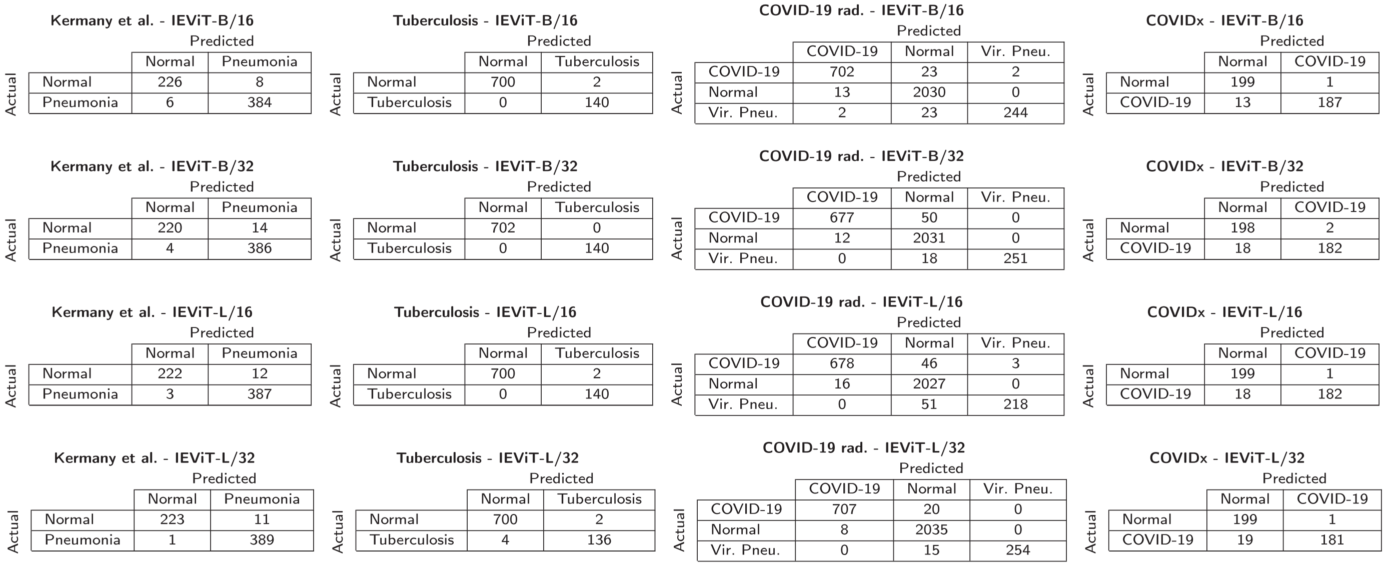
Results for the proposed model's variants, as well as for the original ViT model variants, on the Tuberculosis data set are presented in Table 5. From this table and from Table 8, it is evident that all the proposed IEViT model's variants outperformed their respective ViT model variants for all the examined metrics. The achieved improvement in terms of F1-score spans from +0.35% for the B/16 variant to +2.79% for the L/16 variant, with an average improvement of +1.72% across all variants. F1-scores for the ViT model variants ranged from 95.91%-98.94%, compared to 97.84%-100% for the proposed IEViT model's variants. The best performance for the Tuberculosis data set was achieved using the IEViT-B/32 model, reaching an F1-score, accuracy, recall, and precision of 100%, as shown in Table 5.

For the COVID-19 radiography data set, results for the proposed model's variants, as well as for the original ViT model variants, are presented in Table 6. As shown in this table, as well as in Table 8, all the proposed IEViT model's variants outperformed their respective ViT model variants for all the examined metrics. The improvement achieved in terms of F1-score ranged between +0.58% for the B/16 variant to +3.84% for the L/32 variant, with an average improvement of +1.49% across all variants. F1-scores for the ViT model variants ranged from 92.95%-96.27%, compared to 93.90%-98.05% for the proposed IEViT model's variants. The best performance for the COVID-19 radiography data set was achieved using the IEViT-L/32 model, reaching an F1-score of 98.05%, an accuracy of 98.59%, a recall of 97.09%, and a precision of 99.06%, as shown in Table 6.

Results for the COVIDx data set for the original ViT variants, as well as the proposed IEViT variants are reported in Table 7. From this table and from Table 8, it is evident that all the proposed IEViT model's variants outperformed their respective ViT model variants for all the examined metrics. The achieved improvement in terms of F1-score ranged from +1.63% for the B/16 variant to +3.39% for the B/32 variant, with an average improvement of +2.41% across all variants. For the ViT model variants, F1-scores ranged from 91.40%-94.76%, compared to 94.76%-96.39% for the proposed IEViT model's variants. The best performance for the COVIDx data set was achieved using the IEViT-B/16 model, reaching an F1-score of 96.39%, an accuracy of 96.50%, a recall of 93.50%, and a precision of 99.47%, as shown in Table 7.

3.3. Ablation study

The acquired experimental results demonstrate that the addition of the CNN block in the ViT architecture and the concatenation



Kermany et al. - IEViT-B/16

		Predicted	
		Normal	Pneumonia
Actual	Normal	226	8
	Pneumonia	6	384

Tuberculosis - IEViT-B/16

		Predicted	
		Normal	Tuberculosis
Actual	Normal	700	2
	Tuberculosis	0	140

COVID-19 rad. - IEViT-B/16

		Predicted		
		COVID-19	Normal	Vir. Pneu.
Actual	COVID-19	702	23	2
	Normal	13	2030	0
	Vir. Pneu.	2	23	244

COVIDx - IEViT-B/16

		Predicted	
		Normal	COVID-19
Actual	Normal	199	1
	COVID-19	13	187

Kermany et al. - IEViT-B/32

		Predicted	
		Normal	Pneumonia
Actual	Normal	220	14
	Pneumonia	4	386

Tuberculosis - IEViT-B/32

		Predicted	
		Normal	Tuberculosis
Actual	Normal	702	0
	Tuberculosis	0	140

COVID-19 rad. - IEViT-B/32

		Predicted		
		COVID-19	Normal	Vir. Pneu.
Actual	COVID-19	677	50	0
	Normal	12	2031	0
	Vir. Pneu.	0	18	251

COVIDx - IEViT-B/32

		Predicted	
		Normal	COVID-19
Actual	Normal	198	2
	COVID-19	18	182

Kermany et al. - IEViT-L/16

		Predicted	
		Normal	Pneumonia
Actual	Normal	222	12
	Pneumonia	3	387

Tuberculosis - IEViT-L/16

		Predicted	
		Normal	Tuberculosis
Actual	Normal	700	2
	Tuberculosis	0	140

COVID-19 rad. - IEViT-L/16

		Predicted		
		COVID-19	Normal	Vir. Pneu.
Actual	COVID-19	678	46	3
	Normal	16	2027	0
	Vir. Pneu.	0	51	218

COVIDx - IEViT-L/16

		Predicted	
		Normal	COVID-19
Actual	Normal	199	1
	COVID-19	18	182

Kermany et al. - IEViT-L/32

		Predicted	
		Normal	Pneumonia
Actual	Normal	223	11
	Pneumonia	1	389

Tuberculosis - IEViT-L/32

		Predicted	
		Normal	Tuberculosis
Actual	Normal	700	2
	Tuberculosis	4	136

COVID-19 rad. - IEViT-L/32

		Predicted		
		COVID-19	Normal	Vir. Pneu.
Actual	COVID-19	707	20	0
	Normal	8	2035	0
	Vir. Pneu.	0	15	254

COVIDx - IEViT-L/32

		Predicted	
		Normal	COVID-19
Actual	Normal	199	1
	COVID-19	19	181

Fig. 3. Confusion matrices for all the IEViT variants for the four examined data sets. Results refer to the test set of each dataset.

Table 8
F1-scores (%) and their difference (Δ) for IEViT and ViT for all examined data sets.

Data set	ViT-B/16	IEViT-B/16	**Δ**	ViT-L/16	IEViT-L/16	**Δ**	ViT-B/32	IEViT-B/32	**Δ**	ViT-L/32	IEViT-L/32	**Δ**	**Avg Δ**
Kermany	94.28	98.21	3.93	93.92	98.10	4.18	92.96	97.72	4.76	92.66	98.48	5.82	4.67
Tuberculosis	98.94	99.29	0.35	96.50	99.29	2.79	98.18	100.0	1.82	95.91	97.84	1.93	1.72
COVID-19	96.27	96.85	0.58	92.95	93.90	0.95	96.16	96.74	0.59	94.21	98.05	3.84	1.49
COVIDx	94.76	96.39	1.63	92.80	95.04	2.24	91.40	94.79	3.39	92.39	94.76	2.38	2.41
Average	96.06	97.69	1.62	94.04	96.58	2.54	94.67	97.31	2.64	93.79	97.28	3.49	2.57

Table 9
F1-scores (%) for each data set using IEViT variants and the examined CNN models.

Model	Data set			
	Kermany	Tuberculosis	COVID-19	COVIDx
InceptionV3	94.76	99.28	97.86	97.19
Xception	95.24	96.68	96.92	97.70
ResNet50V2	95.13	98.55	97.70	98.48
EfficientNetB4	94.15	94.74	**98.05**	**98.48**
InceptionResNetV2	95.20	97.06	96.26	98.22
IEViT-B/16	98.21	99.29	96.85	96.39
IEViT-B/32	97.72	**100.0**	96.74	94.79
IEViT-L/16	98.10	99.29	93.90	95.04
IEViT-L/32	**98.48**	97.84	**98.05**	94.76

Note: Results in bold denote the best performance for each dataset.

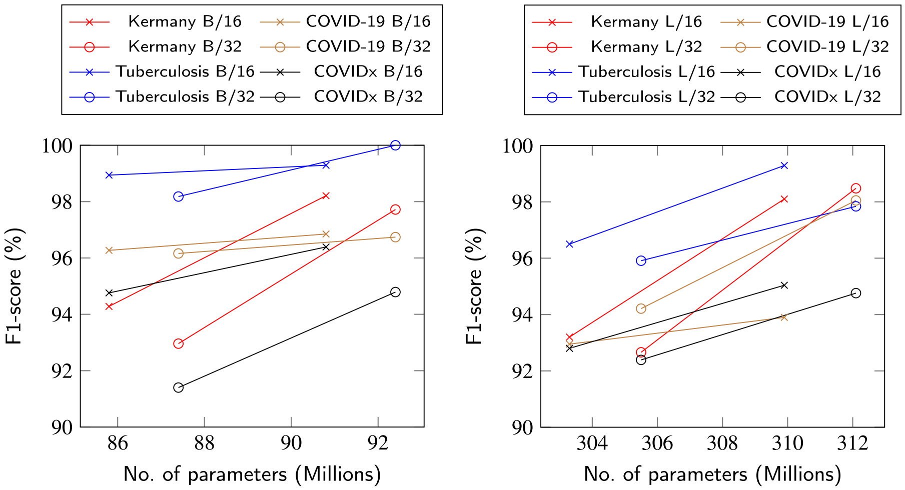
of its output to the output of each Transformer encoder layer led to a consistently improved classification performance of IEViT over ViT for all the examined data sets and variants. To further validate the contribution of the CNN block to the improvement in performance and to examine whether the combination of the CNN block and ViT is justified or the CNN block on its own is able to perform on par with IEViT, we evaluated the performance of the CNN block by creating a model consisting of only the CNN block and the MLP head. The "Base" ($D = 768$) and "Large" ($D = 1024$) variants of the CNN block were evaluated on the four examined data sets, using the same training parameters (including data augmentation) as for the ViT and the IEViT experiments.

Classification results for the CNN block variants are provided in Tables 4, 5, 6, and 7 for each data set respectively. From these tables, it is evident that the CNN block variants on their own provide significantly worse performance than the respective IEViT and ViT variants. For the Kermany et al. children pneumonia data set, the CNN block achieved a maximum F1-score of 78.94%, compared to 98.48% for IEViT and 94.28% for ViT, whereas for the Tuberculosis data set, the CNN block achieved a maximum F1-score of 73.72%, compared to 100% for IEViT and 98.94% for ViT. For the COVID-19 radiography data set, the CNN block achieved a maximum F1-score of 74.24%, compared to 98.05% for IEViT and 96.27% for ViT, while for the COVIDx data set, the CNN block achieved a maximum F1-score of 91%, compared to 96.39% for IEViT and 94.76% for ViT. Consequently, the superiority of the IEViT variants over their respective ViT variants indicates that the introduction of the CNN block is significant for boosting the performance of the proposed IEViT model.

3.4. Comparison to established CNN models

To further assess the suitability of the proposed model, the performance of five well-established CNN models was evaluated on the four examined data sets. The CNN models used for the comparison with IEViT were InceptionV3 [62], Xception [10], ResNet50V2 [23], EfficientNetB4 [63], and InceptionResNetV2 [61]. All models were initialised with weights pre-trained on the ImageNet data set and were then fine-tuned on the examined data sets using end-to-end training and the same training parameters (including data augmentation) as for the ViT and the IEViT experiments. It must be noted that the default Keras implementations of the examined CNN models were used for the experimental evaluation.

The F1-scores achieved by the examined CNN models, as well as by the four IEViT variants, are provided for each data set in Table 9. The IEViT-L/32 variant achieved the best performance for the Kermany et al. children pneumonia data set, with an F1-score of 98.48% compared to 95.24% achieved by the best performing CNN model (Xception). Regarding the Tuberculosis data set, the IEViT-B/32 variant performed the best, achieving an F1-score of 100% compared to 99.28% for the best performing InceptionV3 CNN



(a) (b)

Fig. 4. F1-score (%) achieved for each data set and model variant vs. the number of parameters (millions) of the model. (a) “Base” variant, (b) “Large” variant.

Table 10
Best performing IEViT model per data set.

Data set	Model	Accuracy	Recall	Precision	F1-score
Kermany	IEViT-L/32	98.08	99.74	97.25	98.48
Tuberculosis	IEViT-B/32	100.0	100.0	100.0	100.0
COVID-19	IEViT-L/32	98.59	97.09	99.06	98.05
COVIDx	IEViT-B/16	96.50	93.50	99.47	96.39

model. For the COVID-19 radiography data set, the best performance was provided by both the IEViT-L/32 variant and the EfficientNetB4 CNN model, with both achieving an F1-score of 98.05%. Finally, the EfficientNetB4 model provided the best performance for the COVIDx data set, achieving an F1-score of 98.48% compared to 96.39% achieved by the best performing IEViT variant (IEViT-B/16).

4. Discussion

By examining the experimental results, it is evident that the proposed IEViT enhancement of the ViT architecture consistently results in improved performance for all the examined variants and data sets. As shown in Table 8, the average improvement across data sets over the original ViT model in terms of the F1-score was +1.62% for the B/16 variant, +2.64% for B/32, +2.54% for L/16, and +3.49% for L/32. For the B/16, B/32, and L/16 variants, the lowest improvement was achieved for the COVID-19 radiography data set, ranging from +0.58% to +0.95%, whereas the lowest improvement for the L/32 variant was achieved for the Tuberculosis data set (+1.93%).

It is also worth mentioning that different IEViT variants performed the best for each data set, as shown in Table 10. Variant L/32 provided the best performance for the Kermany and COVID-19 radiography data sets, B/32 for the Tuberculosis data set, and B/16 for the COVIDx data set. Consequently, the selection of the most suitable variant depends on the task at hand and would require experimentation and validation experiments on the related data. Furthermore, as shown in Table 9, well-established CNN models can achieve comparable (COVID-19 data set) or better (COVIDx data set) performance than IEViT in some cases, whereas IEViT performs better in other cases (Kermany and Tuberculosis data sets). Consequently, the task at hand and the computational complexity of the models used must be taken into consideration when selecting the most appropriate model. Nevertheless, as shown in Table 8, the IEViT-B/16 variant achieved the highest average F1-score (97.69%) across all data sets, indicating that it can be a good generic solution for X-ray image classification tasks. To this end, the B/16 variant can be selected when there are computational or other constrains that do not allow a thorough experimentation with all the available variants.

The proposed IEViT enhancement of the ViT architecture comes at a cost in terms of computational complexity. The addition of convolution layers, as well as the increase in size of the input to the Transformer encoder layers, results in an increase in the number of trainable parameters of the proposed models compared to the original ViT models. As shown in Table 2, the IEViT “Base” variants have approximately 5 million more trainable parameters compared to the respective ViT “Base” variants, whereas the IEViT “Large” variants have approximately 6.5 million more trainable parameters compared to the respective ViT “Large” variants. Nevertheless, as shown in Fig. 4, this increase in complexity (number of parameters) led in all cases to improved performance, with some cases exhibiting a substantial improvement (up to +5.82% in F1-score for the L/32 variant and the Kermany data set), thus this increase in complexity can be justified.

To ensure the generalisation ability of the proposed models, various measures were taken to avoid overfitting. For each data set, the models were trained and tested on independent sets, where the final test set was completely “unseen” to the training process. Models were optimised for a validation set which was selected out of the training set. Furthermore, a stratified random sampling was used to create the train/validation/test splits when no official split was available, in order to ensure that the class distribution within the various sets was the same. The use of data augmentation during the training procedure further helped in addressing overfitting as it introduces variation in the training data and has been shown to improve the generalisation ability of deep learning models and reduce overfitting [58]. Similarly, the use of the employed label smoothing technique has also been shown to reduce overfitting and lead to better generalisation [62]. The proposed IEViT variants, as well as the original ViT variants were trained and evaluated on four diverse X-ray image data sets. Performance was consistent across data sets for both models, with IEViT consistently outperforming ViT for all the data sets. Finally, the F1-score was used as the benchmark metric for the performance of the examined models in order to ensure a fair performance evaluation, as the accuracy metric can be heavily biased towards the majority class in the data set and thus lead to erroneous conclusions.

The proposed IEViT model built upon the state-of-the-art original ViT model for image classification and achieved consistently better performance than ViT for the classification of chest X-ray images, using transfer learning via weights pre-trained on ImageNet. Given their widespread availability, the ability to use weights pre-trained for the original ViT model is also an advantageous characteristic of IEViT, as IEViT models can be easily trained using transfer learning for various image classification tasks, without the need to train the model on huge data sets. Furthermore, the use of various data sets that contained chest X-ray images acquired from various sources, using different radiography devices, indicates that IEViT is able to generalise well and is not limited to images acquired under specific settings and using specific radiography devices. Considering this and given the reported experimental results, it is evident that the proposed IEViT model constitutes a powerful solution for chest X-ray image classification that is able to consistently outperform the original ViT model variants on a set of diverse chest X-ray image data sets.

5. Conclusion

In this work, we evaluated the performance of the state-of-the-art Vision Transformer (ViT) architecture for the task of classifying various pathological conditions in chest X-ray images, and proposed the novel IEViT architecture that outperformed the original ViT for all the variants and data sets examined. Experiments on a data set for children pneumonia, a data set for Tuberculosis, a data set for COVID-19 and viral pneumonia, as well as a data set for COVID-19, demonstrated the consistent improvement in classification performance of the proposed IEViT model's variants over the respective original ViT model's variants. Classification performance in terms of F1-score reached 98.4% for the children pneumonia data set using the IEViT-L/32 variant, 100% for the Tuberculosis data set using the IEViT-B/32 variant, 98.05% for the COVID-19 and viral pneumonia data set using the IEViT-L/32 variant, and 96.39% for the COVID-19 data set using the IEViT-B/16 variant.

Given the relatively low cost and the widespread accessibility of chest X-ray imaging, the use of the proposed IEViT model can potentially offer a powerful, but relatively cheap and accessible method for assisting diagnosis using chest X-ray images. Furthermore, the variety of data sets and image sources within the data sets indicates that the proposed solution is not constrained to images acquired using a specific device under specific settings, but can be generalised. Finally, the successful use of transfer learning, by using weights pre-trained on generic images, suggests that the proposed approach can be easily extended and re-trained for the classification of chest X-ray images with various pathologies. Future work will focus on examining the performance of the proposed IEViT model on multiple pathologies, on exploring ways to reduce the overall computational complexity of the model, and on examining the use of transfer learning based on X-ray images instead of generic natural images.

CRediT authorship contribution statement

Gabriel Iluebe Okolo: Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Visualization. **Stamos Katsigiannis:** Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Naeem Ramzan:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration.

Acknowledgment

This research was partially supported by the SAFE_RH project at the University of the West of Scotland under Grant No. ERASMUS+ CBHE - 619483-EPP-1-2020-1-UK-EPPKA2-CBHE.

References

[1] A. Abbas, M.M. Abdelsamea, M.M. Gaber, DeTrac: transfer learning of class decomposed medical images in convolutional neural networks, IEEE Access 8 (2020) 74901–74913, doi:10.1109/ACCESS.2020.2989273.

[2] D. Avola, A. Bacciu, L. Cinque, A. Fagioli, M.R. Marini, R. Taiello, Study on transfer learning capabilities for pneumonia classification in chest-X-rays images, Comput. Methods Programs Biomed. 221 (2022) 106833, doi:10.1016/j.cmpb.2022.106833.

[3] E. Ayan, B. Karabulut, H.M. Ünver, Diagnosis of pediatric pneumonia with ensemble of deep convolutional neural networks in chest X-ray images, Arabian J. Sci. Eng. 47 (2021) 2123–2139, doi:10.1007/s13369-021-06127-z.

[4] E. Ayan, H.M. Ünver, Diagnosis of pneumonia from chest X-ray images using deep learning, in: Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), IEEE, 2019, pp. 1–5, doi:10.1109/EBBT.2019.8741582.

[5] H. Behzadi-khormouji, H. Rostami, S. Salehi, T. Derakhshande-Rishehri, M. Masoumi, S. Salemi, A. Keshavarz, A. Gholamrezanezhad, M. Assadi, A. Batouli, Deep learning, reusable and problem-based architectures for detection of consolidation on chest X-ray images, Comput. Methods Programs Biomed. 185 (2020) 105162, doi:10.1016/j.cmpb.2019.105162.

[6] B.T. Portal, Belarus Tuberculosis data set. Accessed: Sep. 9, 2020. http://tuberculosis.by/.

[7] I. Bello, B. Zoph, A. Vaswani, J. Shlens, Q.V. Le, Attention augmented convolutional networks, in: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3286–3295, doi:10.1109/ICCV.2019.00338.

[8] E. Çallı, E. Sogancioglu, B. van Ginneken, K.G. van Leeuwen, K. Murphy, Deep learning for chest X-ray analysis: a survey, Med. Image Anal. 72 (2021) 102125, doi:10.1016/j.media.2021.102125.

[9] C.F.R. Chen, Q. Fan, R. Panda, CrossViT: cross-attention multi-scale vision transformer for image classification, in: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 357–366, doi:10.1109/ICCV48922.2021.00041.

[10] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258, doi:10.1109/CVPR.2017.195.

[11] M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K.R. Islam, M.S. Khan, A. Iqbal, N.A. Emadi, M.B.I. Reaz, M.T. Islam, Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8 (2020) 132665–132676, doi:10.1109/ACCESS.2020.3010287.

[12] A. Chung, Actualmed COVID-19 chest X-ray data initiative, 2020a. Accessed: Nov. 2021 https://github.com/agchung/Actualmed-COVID-chestxray-dataset.

[13] A. Chung, COVID-19 chest X-ray data initiative, 2020b. Accessed: Nov. 2021. https://github.com/agchung/Figure1-COVID-chestxray-dataset.

[14] J.P. Cohen, P. Morrison, L. Dao, COVID-19 image data collection, arXiv preprint arXiv:2003.11597(2020).

[15] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, doi:10.18653/v1/N19-1423.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, (2020), doi:10.48550/arXiv.2010.11929.

[17] J. Egger, C. Gsaxner, A. Pepe, K.L. Pomykala, F. Jonske, M. Kurz, J. Li, J. Kleesiek, Medical deep learning–a systematic meta-review, Comput. Methods Programs Biomed. 221 (2022) 106874, doi:10.1016/j.cmpb.2022.106874.

[18] L.G.C. Evalgelista, E.B. Guedes, Computer-aided tuberculosis detection from chest X-ray images with convolutional neural networks, SBC, 2018. Anais do XV Encontro Nacional de Inteligência Artificial e Computacional, 518–527, 10.5753/eniac.2018.4444.

[19] O. Faust, Y. Hagiwara, T.J. Hong, O.S. Lih, U.R. Acharya, Deep learning for healthcare applications based on physiological signals: a review, Comput. Methods Programs Biomed. 161 (2018) 1–13, doi:10.1016/j.cmpb.2018.04.005.

[20] S. Govindarajan, R. Swaminathan, Extreme learning machine based differentiation of pulmonary tuberculosis in chest radiographs using integrated local feature descriptors, Comput. Methods Programs Biomed. 204 (2021) 106058, doi:10.1016/j.cmpb.2021.106058.

[21] M.H. Guo, T.X. Xu, J.J. Liu, Z.N. Liu, P.T. Jiang, T.J. Mu, S.H. Zhang, R.R. Martin, M.M. Cheng, S.M. Hu, Attention mechanisms in computer vision: a survey, Comput. Visual Media 4 (2022), doi:10.1007/s41095-022-0271-y.

[22] A. Haghanifar, M.M. Majdabadi, S. Ko, COVID-CXNet: Detecting COVID-19 in frontal chest X-ray images using deep learning, 2020. Accessed: Nov. 30, 2021. https://github.com/armiro/COVID-CXNet.

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi:10.1109/CVPR.2016.90.

[24] R. Hooda, S. Sofat, S. Kaur, A. Mittal, F. Meriaudeau, Deep-learning: a potential method for tuberculosis detection using chest radiography, IEEE, 2017. IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 497–502, doi:10.1109/ICSIPA.2017.8120663.

[25] E.J. Hwang, J.H. Hong, K.H. Lee, J.I. Kim, J.G. Nam, D.S. Kim, H. Choi, S.J. Yoo, J.M. Goo, C.M. Park, Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study, Eur. Radiol. 30 (2020) 3660–3671, doi:10.1007/s00330-020-06771-3.

[26] M. de la Iglesia Vayá, J.M. Saborit-Torres, J.A. Montell Serrano, E. Oliver-Garcia, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García, M. Caparrós, G. González, J.M. Salinas, BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients, 2021, doi:10.21227/w3aw-rv39.

[27] Italian Society of Medical and Interventional Radiology (SIRM), COVID-19 database 2020, 2020. Accessed: Nov. 30, 2021. https://www.sirm.org/en/category/articles/covid-19-database/.

[28] S. Jang, H. Song, Y.J. Shin, J. Kim, J. Kim, K.W. Lee, S.S. Lee, W. Lee, S. Lee, K.H. Lee, Deep learning–based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs, Radiology 296 (2020) 652–661, doi:10.1148/radiol.2020200165.

[29] J. Ker, L. Wang, J. Rao, T. Lim, Deep learning applications in medical image analysis, IEEE Access 6 (2017) 9375–9389, doi:10.1109/ACCESS.2017.2788044.

[30] D. Kermany, K. Zhang, M. Goldbaum, et al., Labeled optical coherence tomography (OCT) and chest X-ray images for classification, Mendeley Data 2 (2018), doi:10.17632/rscbjbr9sj.2.

[31] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (2018) 1122–1131, doi:10.1016/j.cell.2018.02.010.

[32] A.I. Khan, J.L. Shah, M.M. Bhat, Coronet: a deep neural network for detection and diagnosis of COVID-19 from chest X-ray images, Comput. Methods Programs Biomed. 196 (2020) 105581, doi:10.1016/j.cmpb.2020.105581.

[33] J.H. Kim, J.Y. Kim, G.H. Kim, D. Kang, I.J. Kim, J. Seo, J.R. Andrews, C.M. Park, Clinical validation of a deep learning algorithm for detection of pneumonia on chest radiographs in emergency department patients with acute febrile respiratory illness, J. Clin. Med. 9 (2020) 1981, doi:10.3390/jcm9061981.

[34] Y. Li, K. Zhang, J. Cao, R. Timofte, L. Van Gool, LocalViT: bringing locality to vision transformers, (2021), doi:10.48550/arXiv.2104.05707.

[35] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Trans. Neural Netw. Learn. Syst. (2021). doi: 10.1109/TNNLS.2021.3084827(Early Access).

[36] G. Liang, L. Zheng, A transfer learning method with deep residual network for pediatric pneumonia diagnosis, Comput. Methods Programs Biomed. 187 (2020) 104964, doi:10.1016/j.cmpb.2019.06.023.

[37] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, doi:10.1016/j.media.2017.07.005.

[38] X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, et al., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, Lancet Digital Health 1 (2019) e271–e297, doi:10.1016/S2589-7500(19)30123-2.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022, doi:10.1109/ICCV48922.2021.00986.

[40] S.Y. Lu, S.H. Wang, X. Zhang, Y.D. Zhang, TBNet: a context-aware graph network for tuberculosis diagnosis, Comput. Methods Programs Biomed. 214 (2022) 106587, doi:10.1016/j.cmpb.2021.106587.

[41] G.I. Okolo, S. Katsigiannis, T. Althobaiti, N. Ramzan, On the use of deep learning for imaging-based COVID-19 detection using chest X-rays, Sensors 21 (2021), doi:10.3390/s21175702.

[42] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, D. Pfeiffer, Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization, Sci. Rep. 9 (2019) 1–9, doi:10.1038/s41598-019-42557-4.

[43] R.M. Pereira, D. Bertolini, L.O. Teixeira, C.N. Silla, Y.M. Costa, COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios, Comput. Methods Programs Biomed. 194 (2020) 105532, doi:10.1016/j.cmpb.2020.105532.

[44] L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning, (2017), doi:10.48550/arXiv.1712.04621.

[45] C. Qin, D. Yao, Y. Shi, Z. Song, Computer-aided detection in chest radiography based on artificial intelligence: a survey, Biomed. Eng. Online 17 (2018) 1–23, doi:10.1186/s12938-018-0544-y.

[46] Radiological Society of North America, COVID-19 radiography database, 2019a. Accessed: Nov. 2021. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

[47] Radiological Society of North America, RSNA pneumonia detection challenge, 2019b. Accessed: Nov. 2021. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data.

[48] T. Rahman, A. Khandakar, M.A. Kadir, K.R. Islam, K.F. Islam, R. Mazhar, T. Hamid, M.T. Islam, S. Kashem, Z.B. Mahbub, et al., Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization, IEEE Access 8 (2020) 191586–191601, doi:10.1109/ACCESS.2020.3031384.

[49] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S.B. Abul Kashem, M.T. Islam, S. Al Maadeed, S.M. Zughaier, M.S. Khan, M.E. Chowdhury, Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images, Comput. Biol. Med. 132 (2021) 104319, doi:10.1016/j.compbiomed.2021.104319.

[50] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning, (2017), doi:10.48550/arXiv.1711.05225.

[51] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, in: 33rd Conference on Neural Information Processing Systems (NeurIPS), 2019.

[52] E. Real, A. Aggarwal, Y. Huang, Q.V. Le, Regularized evolution for image classifier architecture search, in: Proc. AAAI Conference on Artificial Intelligence, 2019, pp. 4780–4789, doi:10.1609/aaai.v33i01.33014780.

[53] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252 , doi:10.1007/s11263-015-0816-y.

[54] A.A. Saraiva, D. Santos, N.J.C. Costa, J.V.M. Sousa, N.M.F. Ferreira, A. Valente, S. Soares, Models of learning to classify X-ray images for the detection of pneumonia using neural networks, in: Proc. 12th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOIMAGING,, 2019, pp. 76–83, doi:10.5220/0007346600760083.

[55] J.C. Seah, C.H. Tang, Q.D. Buchlak, X.G. Holt, J.B. Wardman, A. Aimoldin, N. Esmaili, H. Ahmad, H. Pham, J.F. Lambert, et al., Effect of a comprehensive deep-learning model on the accuracy of chest X-ray interpretation by radiologists: a retrospective, multireader multicase study, Lancet Digit. Health 3 (2021) e496–e506, doi:10.1016/S2589-7500(21)00106-0.

[56] H. Sharma, J.S. Jain, P. Bansal, S. Gupta, Feature extraction and classification of chest X-ray images using CNN to detect pneumonia, IEEE, 2020. 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 227–231, 10.1109/Confluence47617.2020.9057809.

[57] D. Shen, G. Wu, H.I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248, doi:10.1146/annurev-bioeng-071516-044442.

[58] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019) 60, doi:10.1186/s40537-019-0197-0.

[59] E. Smistad, T.L. Falch, M. Bozorgi, A.C. Elster, F. Lindseth, Medical image segmentation on GPUs–a comprehensive review, Med. Image Anal. 20 (2015) 1–18, doi:10.1016/j.media.2014.10.012.

[60] J.C. Souza, J.O.B. Diniz, J.L. Ferreira, G.L.F. da Silva, A.C. Silva, A.C. de Paiva, An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks, Comput. Methods Programs Biomed. 177 (2019) 285–296, doi:10.1016/j.cmpb.2019.06.005.

[61] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proc. AAAI Conference on Artificial Intelligence, 2017, doi:10.1609/aaai.v31i1.11231.

[62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826, doi:10.1109/CVPR.2016.308.

[63] M. Tan, Q. Le, EfficientNet: rethinking model scaling for convolutional neural networks, PMLR, 2019. International Conference on Machine Learning, 6105–6114.

[64] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, PMLR, 2021. International Conference on Machine Learning, 10347–10357.

[65] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, A. Mittal, Pneumonia detection using CNN based feature extraction, IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), (2019), doi:10.1109/ICECCT.2019.8869364.

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NIPS), 2017, pp. 5998–6008.

[67] L. Wang, Z.Q. Lin, A. Wong, COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images, Sci. Rep. 10 (2020) 1–12, doi:10.1038/s41598-020-76550-z.

[68] W. Wang, E. Xie, X. Li, D.P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: a versatile backbone for dense prediction without convolutions, in: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 568–578, doi:10.1109/ICCV48922.2021.00061.

[69] X. Wang, J. Yu, Q. Zhu, S. Li, Z. Zhao, B. Yang, J. Pu, Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography, Occup. Environ. Med. 77 (2020) 597–602, doi:10.1136/oemed-2019-106386.

[70] Y. Wang, N. Wang, M. Xu, J. Yu, C. Qin, X. Luo, X. Yang, T. Wang, A. Li, D. Ni, Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound, IEEE Trans. Med. Imaging 39 (2019) 866–876, doi:10.1109/TMI.2019.2936500.

[71] H.B. Winther, H. Laser, S. Gerbel, S.K. Maschke, J.B. Hinrichs, J. Vogel-Claussen, F.K. Wacker, M.M. Höper, B.C. Meyer, COVID-19 image repository, (2020), doi:10.25835/0090041.

[72] S.C. Wong, A. Gatt, V. Stamatescu, M.D. McDonnell, Understanding data augmentation for classification: when to warp? in: Proc. International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2016, doi:10.1109/DICTA.2016.7797091.

[73] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, P. Vajda, Visual transformers: token-based image representation and processing for computer vision, (2020). doi:10.48550/arXiv.2006.03677.

[74] W. Xing, W. He, X. Li, J. Chen, Y. Cao, W. Zhou, Q. Shen, X. Zhang, D. Ta, Early severity prediction of BPD for premature infants from chest X-ray images using deep learning: a study at the 28th day of oxygen inhalation, Comput. Methods Programs Biomed. 221 (2022) 106869, doi:10.1016/j.cmpb.2022.106869.

[75] O. Yadav, K. Passi, C.K. Jain, Using deep learning to classify X-ray images of potential tuberculosis patients, IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2018) 2368–2375, doi:10.1109/BIBM.2018.8621525.

[76] R. Yamashita, M. Nishio, R.K.G. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insights Imaging 9 (2018) 611–629, doi:10.1007/s13244-018-0639-9.

[77] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.H. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token ViT: training vision transformers from scratch on ImageNet, in: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 558–567, doi:10.1109/ICCV48922.2021.00060.

[78] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., ResNeSt: split-attention networks (2020), doi:10.48550/arXiv.2004.08955.

[79] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8697–8710, doi:10.1109/CVPR.2018.00907.