

## Chest X-ray image classification using transfer learning and hyperparameter customization for lung disease diagnosis

Thanh-An Pham & Van-Dung Hoang

**To cite this article:** Thanh-An Pham & Van-Dung Hoang (05 Mar 2024): Chest X-ray image classification using transfer learning and hyperparameter customization for lung disease diagnosis, Journal of Information and Telecommunication, DOI: [10.1080/24751839.2024.2317509](https://doi.org/10.1080/24751839.2024.2317509)

**To link to this article:** <https://doi.org/10.1080/24751839.2024.2317509>



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



Published online: 05 Mar 2024.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)



# Chest X-ray image classification using transfer learning and hyperparameter customization for lung disease diagnosis

Thanh-An Pham<sup>a\*</sup> and Van-Dung Hoang<sup>b</sup>

<sup>a</sup>University of Sciences, Hue University, Hue city, Vietnam; <sup>b</sup>Faculty of Information Technology, HCMC University of Technology and Education, Ho Chi Minh City, Vietnam

## ABSTRACT

Lung diseases often result in severe damage to the respiratory tract, and lead to a high risk of mortality within a short period of time. DL models based on ViT are considered to have promising advantages over CNN architectures in terms of computational efficiency, and accuracy when trained on large ImageNet datasets. In this study, we present a new DL approach based on the combination of CNN with ViT to improve the efficiency of pneumonia diagnosis using medical images. In the first stage, raw images are passed through a local filter to capture local relations on the inputs. The local filter block includes two convolutional layers with kernel  $3 \times 3$ . This local filtering method aims to enhance rich features before being fed into the patching layer of the ViT block. The proposed method is experimented on the benchmark chest X-ray dataset. The proposed method is evaluated and compared to some well-known models, which include ViT, VGG19, Resnet50, Densnet201. Experimental results demonstrated that the proposed approach based on CNN and ViT reaches higher efficiency with about 1% accuracy to the standard ViT model, and about 2% higher with VGG19, Resnet50, Densnet201 and smaller in model architecture.

## ARTICLE HISTORY

Received 15 June 2023

Accepted 7 February 2024

## KEYWORDS

Transfer learning; CNN; vision transformer; COVID-19; lung disease image

## 1. Introduction

Nowadays, diagnostic imaging in medicine is a common technique, which is used in medical examination and treatment, medical imaging includes X-ray image data, magnetic resonance imaging (MRI), imaging ultrasound images (Ultrasound), and endoscopic images (endoscopy). In the diagnosis of diseases based on images, doctors recognize the morphology and function of the internal structures of the patient's body. According to the results, doctors quickly determine the cause, area, and location of the lesion to give appropriate treatment indications.

The results of imaged diagnostics depend on the experience and abilities of the doctors, as well as their mental state and work pressure. Normally, it takes more time

**CONTACT** Van-Dung Hoang  [dunghv@hcmute.edu.vn](mailto:dunghv@hcmute.edu.vn)  1 Vo Van Ngan Street, Linh Chieu Ward, Thu Duc city, Ho Chi Minh City, Vietnam

\*Present address: Department of Information Technology Management, Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam

© 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.

to train a proficient doctor in interpreting medical imaging results. Brady et al. (2012) pointed out that there are significant differences in the interpretation of imaging results among physicians and expert radiologists, even when using the same image dataset.

As advances in information technology, particularly in the field of artificial intelligence and deep learning, over the past decade, computer-aided diagnosis (CAD) systems have increasingly played an important role in assisting doctors with disease diagnosis. The fields of machine learning, especially deep learning models have been researched and implemented in industry. CAD has become the major research domain in machine learning for medical imaging. CNN is an advanced deep learning architecture that has dominated the field of computer vision for the past decade and CNN is a major of the popular deep learning approaches for classification medical images.

In medical image processing, X-ray is a common and easily accessible diagnostic method due to its low cost, simple implementation, and accessibility. X-ray imaging is often the first indication for examination and treatment of diseases related to the lungs, breast cancer, skeletal joints, and heart failure. Differences to regular objects and abnormalities in a chest X-ray film are often challenging for experts in the field of diagnostic imaging to identify and classify consistently (Brady et al., 2012). For this reason, many studies in medical imaging were carried out to aid radiologists using CNN. These studies have achieved significant success in image classification tasks, including categorizing breast cancer (Khan et al., 2019), classifying skin cancer (Pham et al., 2018; Pham et al., 2020), and detecting brain tumours (Abiwinanda et al., 2019).

In the period of 2021–2022, Vision Transformer is a new deep learning architecture, attracting a lot of research interest from researchers around the world, ViT has a lot of potential compared to convolutional neural networks. The ViT-based deep learning models are evaluated to be more promising than the CNN architecture in terms of computational efficiency, resource usage and accuracy when trained on the large ImageNet dataset. The research community is increasingly focusing on ViT.

A multitude of intelligent diagnosis methods for pneumonia, pulmonary tuberculosis, and COVID-19 have been proposed, thanks to the accelerated advancements in image classification facilitated by deep learning architectures (Rajpurkar et al., 2017; Sharma et al., 2020). These studies have primarily concentrated on tackling the classification task between COVID-19 patients and normal cases, or more specifically, on distinguishing between COVID-19, normal, and pneumonia cases. Some of these models have been very successful in diagnosing pneumonia by analysing chest X-rays. However, due to the limitation of the public datasets for experiments, the size of the datasets used for studies is limited, lacking representativeness and generalizability. Therefore, further research is needed to evaluate the ability to detect and classify COVID-19 patients in comparison to other viral pneumonia cases, using sufficiently large datasets.

Within this study, we introduce a learning model that integrates local filters with ViT for chest disease classification. The proposed method is experimented on a benchmark dataset of chest X-ray images (Sait et al., 2020). The compared results of our model and the ViT\_B/16, VGG19, ResNet50, and DenseNet201 models demonstrated that the proposed method outperforms the classical models.

Among the diverse array of methodologies employed in medical imaging, CNNs have emerged as a highly potent deep learning technique for the classification of medical

images. Their effectiveness is derived from their capability to extract spatial features via convolutional computation. Motivated by CNN architecture, a lot of research on medical imaging has focused on brain tumours (Abiwinanda et al., 2019) breast cancer (Khan et al., 2019), pneumonia detection (Sharma et al., 2020), skin lesions (Pham et al., 2018), Tuberculosis Detection (Huy & Lin, 2023) are implemented.

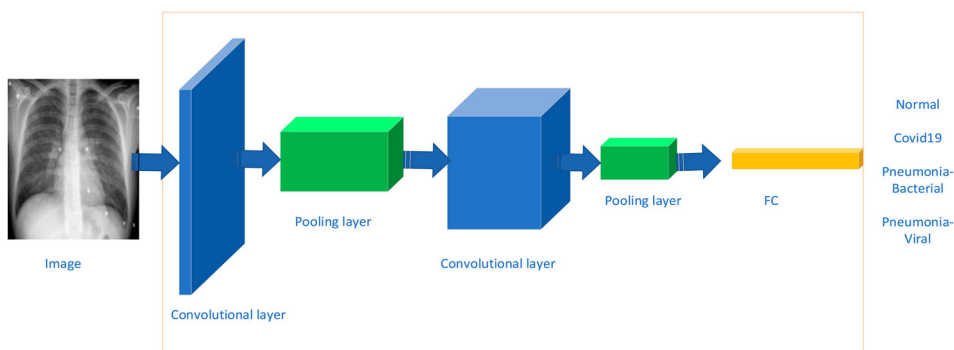
Classification of pneumonia together with COVID-19 also attracts special attention from researchers. Pertaining to pneumonia diagnosis, studies primarily emphasize the utilization of chest X-ray images for disease classification purposes. The efficacy of the aforementioned model is assessed and demonstrated through performance metrics including accuracy, sensitivity, precision and specificity. In 2017, Rajpurkar et al. (2017) proposed the CheXNet model to classify pneumonia including 100,000 chest X-ray images with 14 diseases, CheXNet is based on DenseNet (Huang et al., 2017), and this model achieved the state-of-the-art results on all of 14 diseases. CheXNet surpasses the average F1 metric performance of radiologists. In 2019, Ayan and Ünver (2019) employed two widely recognized CNN Backbones, namely VGG16 and Xception, for the purpose of pneumonia diagnosis. They incorporated transfer learning and fine-tuning methodologies during the training stage. The evaluation of their models revealed that the VGG16 network surpassed the Xception network in terms of accuracy, achieving an accuracy rate of 87%, and 82% respectively. Avola et al. (2022) used 12 well-known ImageNet pre-trained models to classify into 4 classes including healthy, bacterial, virus and COVID-19. Most models obtained significant performances. Recently, Hariri and Avşar (2023) introduced a compact diagnostic model utilizing convolutional neural networks (CNNs) using chest X-ray images for the classification of COVID-19 and pneumonia. The model was designed to classify images into four distinct classes: COVID-19, Healthy, viral and bacterial pneumonia. The proposed lightweight model, developed by Muhab Hariri and Ercan Avşar, demonstrated an impressive overall accuracy of 89.89%. In comparison, the pre-trained Efficient-Net B2 model attained an accuracy of 85.7%.

PneuNet model is based on Vision Transformer which is proposed by Wang et al. (2023) utilizing of X-ray images for pneumonia classification yielded remarkable results with the PneuNet model, achieving an impressive accuracy rate of 94.96% in the three-category classification tasks (COVID-19, None, Normal) and 90% for 4 categories classification (COVID-19, Normal, Bacterial, Viral), and 99.32% with binary classification (COVID-19, None). Park et al. (2021) integrated ViT and ResNet50 architectures to perform the classification of three distinct classes: normal, other infections (including tuberculosis, and bacterial pneumonia), and COVID-19, using a dataset comprising 17,548 chest X-ray images. The model they developed exhibited notable performance, with an average accuracy of 86.4%. 17,548 chest X-ray images. Their model attained an average accuracy of 86.4%.

## 2. Literature review

### 2.1. CNN-based methods

Convolutional neural network (CNN), which was originally proposed by LeCun et al. (1998), represents a class of neural network models that includes convolutional layers, pooling layers for reducing the dimensions of the feature maps, normalization layers, and the final component is fully connected layers. The general CNN architecture

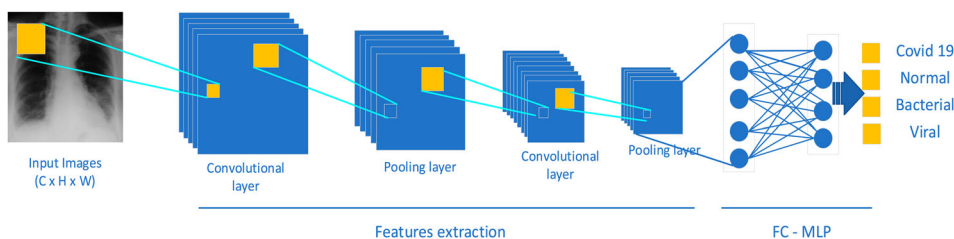


**Figure 1.** Overview of CNN architecture.

(LeCun et al., 1998) is represented in Figure 1. Among these components, the convolutional layer plays a pivotal role within the CNN architecture, facilitating the extraction of informative features from input images and thereby supporting the learning process of the CNN. In the CNN architecture, the input image passes through convolutional layers with filters to learn spatial features of the image, generating feature maps. The feature maps are enriched and gradually decrease in size according to the depth of the network, the pooling layers are used to reduce the size of feature maps. Finally, the features are passed through a fully connected layer, which is a multilayer perceptron (MLP) network. This structure is demonstrated in Figure 2, known as (LeCun et al., 2010). A classification layer utilizes the SoftMax function to assign the input to corresponding class labels. The architecture of deep learning based on CNNs differs from traditional machine learning methods in that the network learns features automatically during the training process. In CNNs, the convolutional layer plays an important role in capturing local relationships among neighbouring pixels in the image space.

LeNet was the first CNN network and it comprises two primary parts, the first part consists of two convolutional layers followed by a pooling layer to reduce the size of the features, and the second part is an MLP network - consisting of two fully connected layers followed by SoftMax layer to classify handwriting images (LeCun et al., 1998).

In 2012, the AlexNet network was introduced (Krizhevsky et al., 2017), and it won the ImageNet LSVRC competition. The classification performance of AlexNet was significantly better than the subsequent ranked networks. AlexNet opened a new phase in the field of computer vision. It is a much deeper neural network compared to the LeNet network,



**Figure 2.** CNN based classification.

consisting of five convolutional layers. With an input image size of  $224 \times 224 \times 3$ , the first and second convolutional layers of AlexNet used filters with kernel sizes of  $11 \times 11$  and  $5 \times 5$ , respectively, while the last three convolutional layers used a kernel size of  $3 \times 3$ . AlexNet utilized the ReLU non-linear activation function. The VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), and DenseNet (Huang et al., 2017) networks were introduced in 2014, 2015, and 2016, respectively, following the debut of AlexNet. Several improvements were made in the VGG, ResNet, and DenseNet networks, including increasing the depth of the network, incorporating skip connections, and determining the combination of convolutional and pooling layers. These changes led to significant advancements in image classification on the ImageNet dataset. Many deep learning models for analysing medical images have been focused on and developed based on the CNN architecture. Examples include breast cancer classification (Khan et al., 2019) and brain tumour classification (Abiwinanda et al., 2019).

## 2.2. Self-attention based methods

Self-Attention is a key component in the Transformer Architecture (Dosovitskiy et al., 2020; Vaswani et al., 2017). The self-attention mechanism was developed to address the limitations of sequential processing methods in recurrent neural networks (RNNs) and LSTM (Hochreiter & Schmidhuber, 1997) in natural language processing (NLP). The Self-Attention mechanism aims to determine the relative importance of one token with respect to other tokens in a sequence.

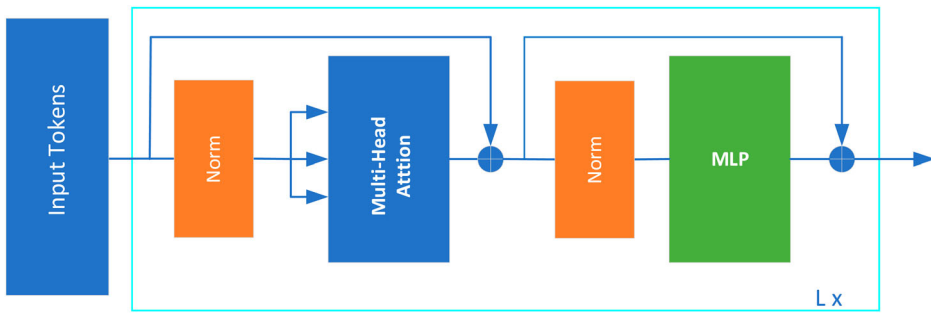
In the Self-Attention mechanism, each input vector is transformed into three separate vectors through linear projection, including query vector  $q$ , key vector  $k$ , and value vector  $v$ , all of which have a fixed size. These vectors are then packed together to form three different weight matrices, namely  $W^Q$ ,  $W^K$ , and  $W^V$ .  $W^Q$ ,  $W^K$ , and  $W^V$  are learnable parameters during the model training process. In practice, for computational convenience, the sets of queries, keys, and values are combined into respective matrices  $Q$ ,  $K$ , and  $V$ . For each input vector  $X$ , the calculations for  $Q$ ,  $K$ , and  $V$  values are performed in the following manner:  $K = W^K X$ ;  $Q = W^Q X$ ;  $V = W^V X$ .

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, d_k \text{ is the dimension of } K$$

Multi-Head Self-Attention (MSA):  $\text{MultiHead}(Q, K, V) = [\text{Concat}(\text{head}_1; \dots; \text{head}_h)]W^O$  MSA is an Attention extension, which includes many Attention mechanisms to learn many features on the data set,  $h$  is the number of heads.

The architecture based on Self-Attention has been proposed and applied in the field of computer vision (CV), achieving classification effectiveness compared to CNN-based models. Vision Transformer (ViT) (Dosovitskiy et al., 2020) utilizes only the Transformer Encoder block, which consists of multiple identical Encoder Layers. The flowchart of the Transformer Encoder in Dosovitskiy et al. (2020) is represented in Figure 3 and each Encoder Layer comprises two sub-layers: Multihead-Self Attention and Multi-layer Perceptron (MLP) as shown in Figure 3.

The transformer (Dosovitskiy et al., 2020) is applied in computer vision involving the following steps as represented in Figure 4. The 2D image data,  $x \in \mathbb{R}^{H \times W \times 3}$ , is divided into non-overlapping patches (tokens), where  $N = (H/p) \times (W/p)$ , with  $H$  and  $W$  being



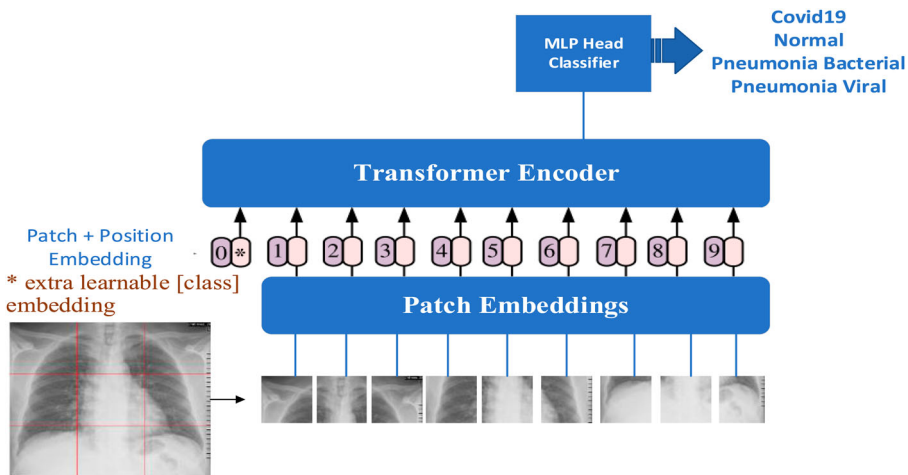
**Figure 3.** Transformer encoder.

the height and width of the image, and  $p$  being the patch size (in standard ViT,  $p$  is chosen as 16). These patches are flattened into 1-dimensional vectors of size  $D$  through a linear projection, where  $D$  is chosen as 768, and the number of patches is  $N$ . The set of patches is then packed into a matrix  $X = \mathbb{R}^{N \times D}$ ,  $X$  is called the embedding matrix. A special classification token (CLS) and position vectors are added to the embedding matrix, which is then fed into the Transformer Encoder block and processed similarly to the NLP domain.

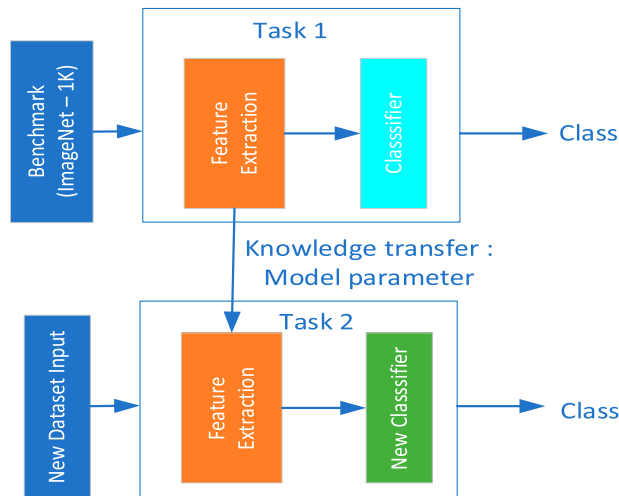
The deep learning model based on CNN focuses on learning and extracting the relationship between local features through convolutional layers, and the ViT model focuses on describing the long dependencies in the global spatial domain of the input image.

### 2.3. Transfer learning method

Transfer Learning is a commonly used method that aims to leverage the knowledge and experience gained by a model from a previously trained task to a new learning task. The goal of transfer learning is to reduce the training time of a model and improve its accuracy



**Figure 4.** Vision transformer architecture.



**Figure 5.** Fine-tuning approach for new dataset.

in conditions where there are not enough large datasets to train machine learning models from scratch.

The transfer learning method is applied in two stages. In the first stage, we train a deep learning model on a large dataset like ImageNet. In the final stage, the trained model is applied to a new problem using two approaches: feature extraction, and fine-tuning.

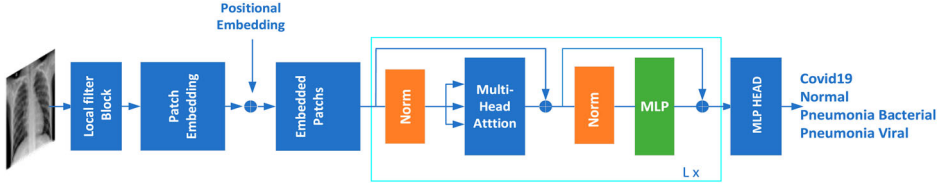
The feature extraction approach involves using the pre-trained model and freezing the model's parameters up to the feature extraction block. Then, we modify the classification block using a Multilayer Perceptron or Support Vector Machine (SVM), and then, we train the model to classify on the new Dataset. The feature extraction approach significantly reduces training time as the parameters of the feature extraction block remain fixed and are not retrained on the new dataset.

In the fine-tuning approach, we inherit the weights of the pre-trained model and modify the fully connected layers in the classification block. Then, training the previously trained model on the new dataset to leverage the learned parameters and knowledge from the pre-training. Transfer learning is widely used in medical image analysis research (Avola et al., 2022; Khan et al., 2019). The fine-tuning technique allows the model to adapt to the new task while still benefiting from the pre-trained weights, as demonstrated in Figure 5.

### 3. Proposed methodology

The proposed method is described in this section, Figure 6 presents the general model architecture and Figure 7 describes our implementation. The proposed method consists of three stages as follows: (1) first stage, neural network-based convolutional filters, known as the local filter - CovEnBlock; (2) rich feature map is fed into the Transformer Encoder module to enrich feature map; (3) MLP Head Classifier is used for pattern classification.





**Figure 6.** Overview of the proposed method – CovEnViT model.

This local filter takes raw images of size  $H \times W \times C$ , where  $(H, W)$  is the height and width of images, and  $C$  is colour channels. This implies that the input image undergoes a transformation by the local filter block to extract local features while maintaining the same image size as the original input.

In the standard ViT model (Dosovitskiy et al., 2020), raw images of size  $H \times W \times 3$  are divided into patches of size  $16 \times 16 \times 3$ . Each patch is treated as a token and fed into the transformer encoder block. In the CovEnViT architecture, to capture local relationships within the spatial domain, we introduce a CovEn Block module. The input images are processed through the local filter block to learn the local features of the image using convolutional layers as shown in Figure 8. The feature maps generated after passing through the local filter block are then divided into patches of size  $16 \times 16 \times 3$  and processed by the Transformer Encoder.

The CovEn Block comprises two convolutional layers using a  $3 \times 3$  kernel size. The Rectified Linear Unit (ReLU) (Agarap, 1803) activation function is used after each convolutional layer,  $f(x) = \max(0, x)$ . The design of the CovEn Block is based on the idea of using two 2D-convolutional layers to capture local relations on raw images. The CovEn Block Module is presented in Figure 8 and PyTorch-like pseudo code is demonstrated in Algorithm 1. In our suggested method, the transformer encoder takes the embedded patches from the high-level feature map as its inputs.

---

**Algorithm 1:** ConvEnViT

---

**Data:**  $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$

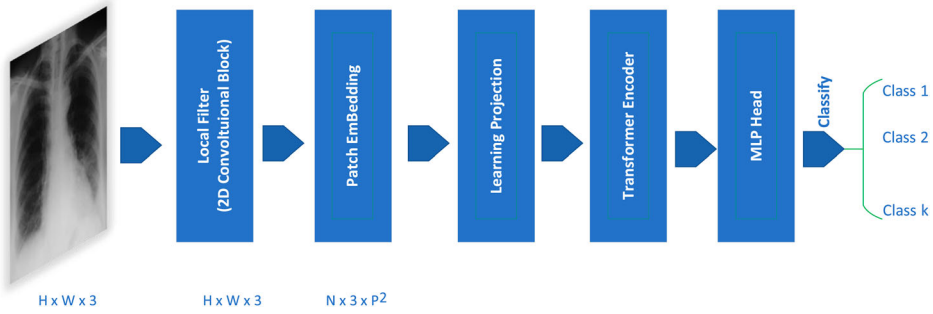
**Result:** class of  $\mathbf{X}$

```

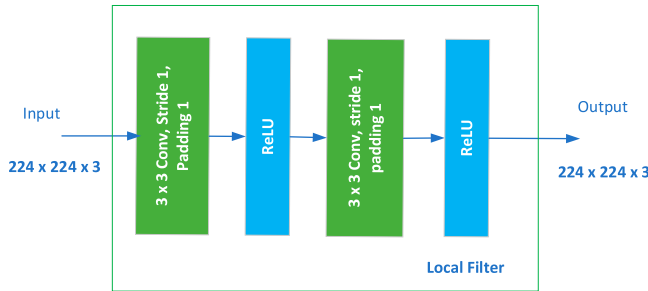
1 procedure: Local Filter Block
2  $\sigma \leftarrow$  Relu function
3  $f^{3 \times 3} \leftarrow$  2D convolution operation with filter size of  $3 \times 3$ 
4 for  $0 \leq i \leq 2$  do
5    $\mathbf{X} \leftarrow \sigma(f^{3 \times 3}(\mathbf{X}))$ 
6 procedure: Vision Transformer by Dosovitskiy et.al.
7  $\text{MSA} \leftarrow$  Multi-head self-attention
8  $\text{MLP} \leftarrow$  Multi-layer Perceptron.
9  $\text{NL} \leftarrow$  Normalization
10  $\mathbf{X} \leftarrow$  Patch and Embedding( $\mathbf{X}$ )
11  $\mathbf{X} \leftarrow \text{cls\_token} \cup \mathbf{X}$ 
12  $\mathbf{X} \leftarrow \mathbf{X} + \text{pos\_embedding}$ 
13  $\mathbf{Z}_0 \leftarrow \mathbf{X}$ 
14 for  $1 \leq i \leq L$  do
15    $\mathbf{Z}_i \leftarrow \text{MSA}(\text{LN}(\mathbf{Z}_{i-1})) + \mathbf{Z}_{i-1}$ 
16    $\mathbf{Z}_i \leftarrow \text{MLP}(\text{LN}(\mathbf{Z}_i)) + \mathbf{Z}_i$ 
17  $\mathbf{y} = \mathbf{Z}_L[0]$ 
18 Return MLP_Head( $\mathbf{y}$ )

```

---



**Figure 7.** The description of implementation the model processing.



**Figure 8.** Local filter – CovEn block module.

where

MLP\_Head(y): MLP\_Head takes the output feature vector of the cls\_token and maps it to a classification prediction.

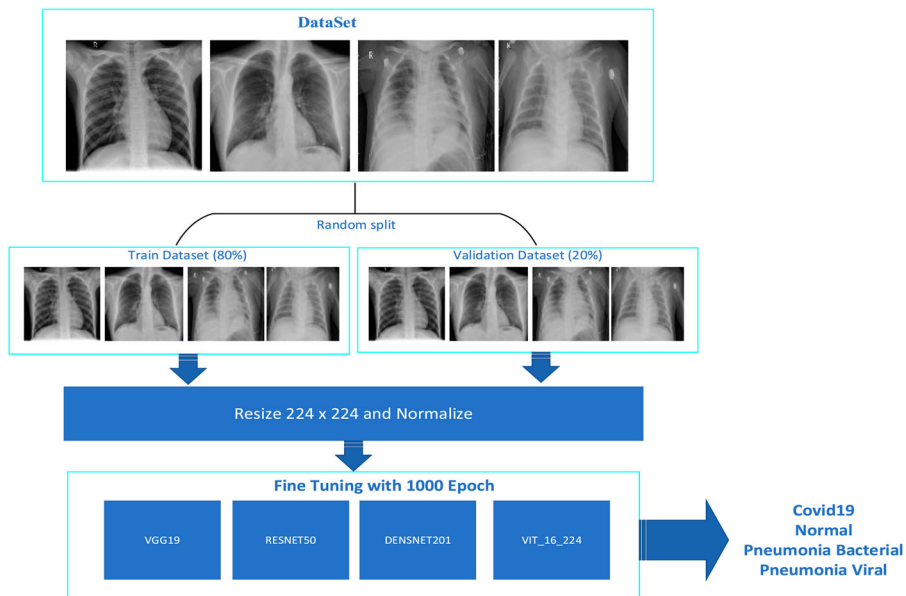
The transfer learning method with the fine-tuning approach is chosen to be implemented in this study to evaluate and compare the effectiveness of the proposed model with ViT and other CNN-based models.

#### 4. Materials and evaluation metrics

The public X-Ray dataset (Sait et al., 2020) was selected for experimentation and evaluation of the classification performance. The experimental dataset consists of 9209 samples with four categories Normal, COVID-19, Pneumonia-Viral and Pneumonia-Bacterial. The dataset undergoes normalization and resizing to achieve size  $224 \times 224$  and is randomly split into separate training sets and evaluation sets in an 80% to 20% ratio, respectively. Four models are used fine-tuning process as shown in Figure 9. The specifics of the dataset within each class are shown in Table 1.

In this study, confusion matrix and common metrics including accuracy, precision, recall, specificity, and F1-score were employed to assess both the pre-trained models and our proposed model.

In the realm of multi-class classification, accuracy serves as a prevalent metric for assessing the performance of deep learning models. It quantifies the proportion of accurately classified samples relative to the total number of samples and is derived by dividing the



**Figure 9.** Transfer learning for X-ray.

sum of true positive and true negative predictions by the total number of prediction samples (Grandini et al., 2020).

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall(Sensitive)} = TP / (TP + FN) \quad (3)$$

$$\text{Specificity} = TN / (TN + FP) \quad (4)$$

$$F1\text{-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

Precision measures the proportion of predictions belonging to a specific class that belongs to that class. Recall evaluates the percentage of positive samples, which is computed as the fraction of true positives (TP) over the sum of true positives (TP) and false negatives (FN) samples. Specificity calculates the proportion of negative samples by comparing the number of true negatives (TN) to the sum of true negatives (TN) and false positives (FP) samples.

Concepts are used for calculating metrics including True Positive – TP, False Positive – FP, True Negative – TN, False Negative – FN. Specifically, True Positive are the samples

**Table 1.** Train and validation dataset.

	Train	Validation	Total
COVID-19	1025	256	1281
Normal	2617	654	3271
Pneumonia-bacterial	2401	600	3001
Pneumonia-viral	1325	331	1656
Total	7368	1841	9209

that have been predicted as positive by the model and are positive, while False Positives are instances that the model has incorrectly classified as positive, even though they are negative. True Negative are the samples that have been labelled as negative by the model and they are True Negative, False Negative are the samples that have been labelled as negative by the model, but they are True positives. Metric indexes are calculated from the confusion matrix as follows: confusion\_matrix is the confusion matrix,  $i$  denotes class  $i^{\text{th}}$ .

$$TP_{\{i\}} = \text{confusion\_matrix}_{\{i, i\}} \quad (6)$$

$$FN_{\{i\}} = \text{sum}(\text{confusion\_matrix}_{\{i, :\}}) - \text{confusion\_matrix}_{\{i, i\}} \quad (7)$$

$$FP_{\{i\}} = \text{sum}(\text{confusion\_matrix}_{\{:, i\}}) - \text{confusion\_matrix}_{\{i, i\}} \quad (8)$$

$$TN_{\{i\}} = \text{sum}(\text{confusion\_matrix}_{\{:, :\}}) - TP_{\{i\}} - FP_{\{i\}} - FN_{\{i\}} \quad (9)$$

In the context of multi-class classification, due to the number of samples of each class is not equal. We define a weight ratio for each class based on the samples of the class. The weight ratio of  $i^{\text{th}}$  class = (the number of samples  $i^{\text{th}}$  class/total samples) for calculating the accuracy, precision, sensitivity, and specificity of the models. Precision, Recall, and Specificity are calculated the same as model accuracy.

$$\text{Weight\_ratio}_{\{i\}} = \text{samples of class } i^{\text{th}} / \text{Total samples} \quad (10)$$

$$\text{Model Accuracy} = \sum (\text{Acc}_{\{i\}} * \text{Weight\_ratio}_{\{i\}}) \quad (11)$$

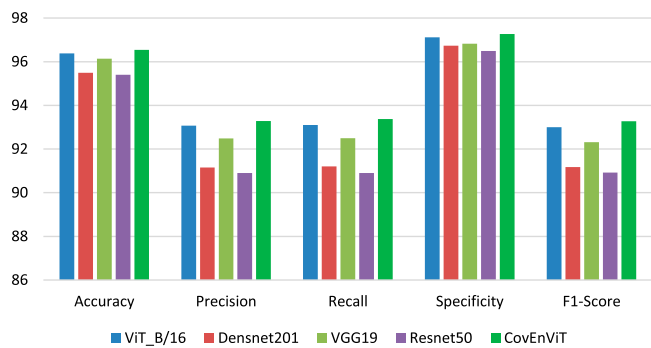
$$\text{Acc}_{\{i\}} = TP_{\{i\}} + TN_{\{i\}} / \text{Total of samples} \quad (12)$$

## 5. Experimental results

In our study, we conducted experiments and comparisons involving the proposed approach and several well-known backbones such as the original ViT (Dosovitskiy et al., 2020), Densnet201 (Huang et al., 2017), VGG19 (Simonyan & Zisserman, 2014), Resnet50 (He et al., 2016). The proposed CovEnViT model was trained from scratch on the ImageNet-1 K dataset with about 200 epochs. Then, the best-performing model on the validation set is used for fine-tuning the X-ray dataset with 1000 epochs. We adopted well-known backbones such as VGG19, Resnet50, and Densenet201, modifying only the number of output neurones to suit our specific requirements. The cross-entropy loss function is employed for training in all approaches. The fine-tuning models are trained on the X-ray dataset with 1000 epochs. The parameters set for training task are as follows: learning rate = 0.00002, batch size = 16, optimizer with AdamW. The Transfer Learning for X-ray processing is presented in Figure 9.

**Table 2.** Performance metrics of the different models.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)
ViT_B/16	96.38	93.07	93.10	97.12	93.00
Densnet201	95.49	91.15	91.20	96.73	91.17
VGG19	96.14	92.48	92.50	96.82	92.31
Resnet50	95.40	90.90	90.90	96.49	90.92
CovEnViT	96.54	93.28	93.37	97.27	93.27



**Figure 10.** Evaluation metrics on the validation set using different models.

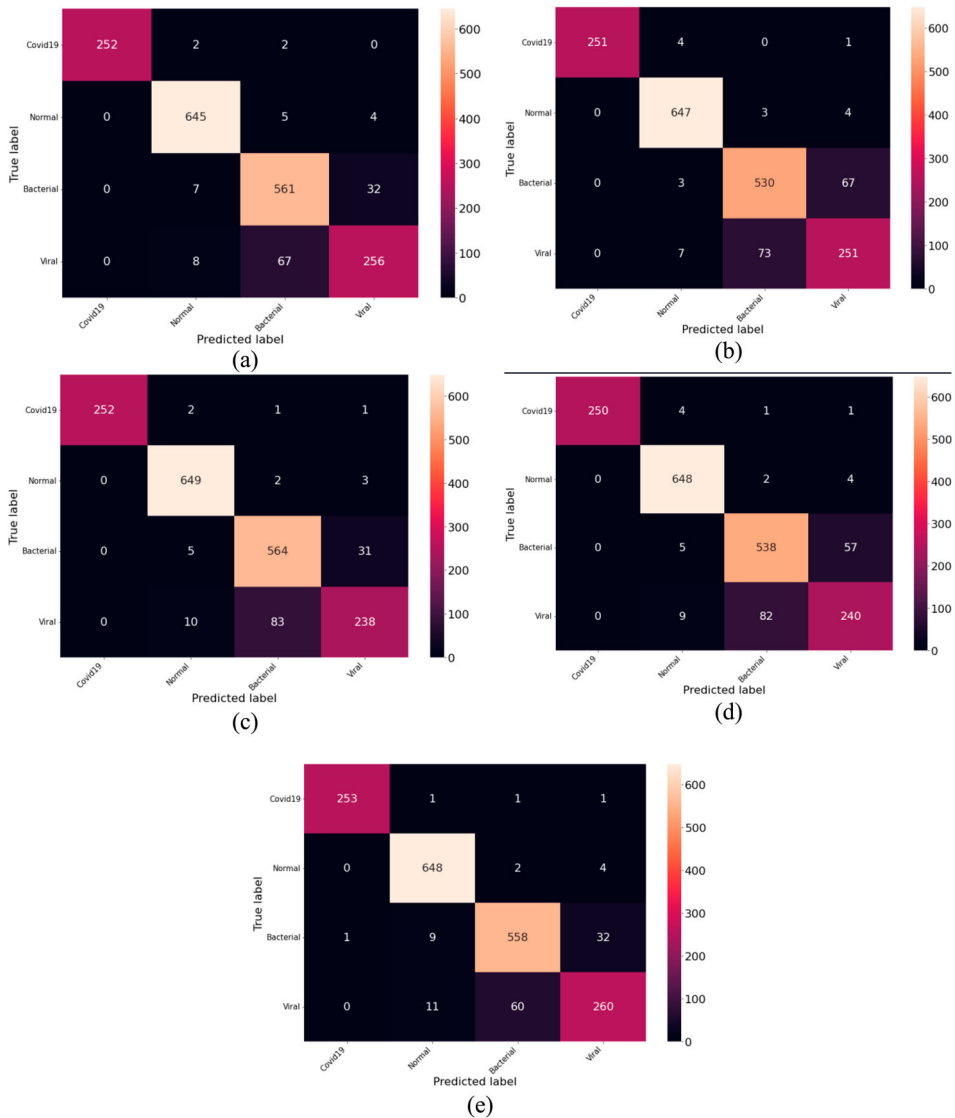
**Table 3.** Details of classified results of the different models.

Model	Class	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1_Score (%)
ViT_B/16	COVID-19	99.78	100.0	98.44	100.0	99.21
	Normal	98.59	97.43	98.62	98.57	98.02
	Pneumonia Bacterial	93.86	88.35	93.50	94.04	90.85
	Pneumonia Viral	93.97	87.67	77.34	97.62	82.18
Densenet201	COVID-19	99.73	100.0	98.05	100.0	99.01
	Normal	98.86	97.88	98.93	98.82	98.40
	Pneumonia Bacterial	92.07	87.46	88.33	93.88	87.89
	Pneumonia Viral	91.74	77.71	75.83	95.23	76.76
VGG19	COVID-19	99.78	100.0	98.44	100.0	99.21
	Normal	98.80	97.45	99.24	98.57	98.33
	Pneumonia Bacterial	93.37	86.77	94.00	93.07	90.24
	Pneumonia Viral	93.05	87.18	71.90	97.68	78.81
Resnet50	COVID-19	99.67	100.0	97.66	100.0	98.81
	Normal	98.70	97.30	99.08	98.48	98.18
	Pneumonia Bacterial	92.02	86.36	89.67	93.15	87.98
	Pneumonia Viral	91.69	79.47	72.51	95.89	75.83
CovEnViT	COVID-19	99.78	99.61	98.83	99.94	99.22
	Normal	98.53	96.86	99.08	98.23	97.96
	Pneumonia Bacterial	94.30	89.86	93.00	94.92	91.40
	Pneumonia Viral	94.13	87.54	78.55	97.55	82.80

The evaluation results, along with the solutions, are presented in Table 2. The results demonstrated that CovEnViT model achieved the best overall prediction accuracy.

Within CNN-based deep learning methods, VGG19 demonstrates the highest accuracy (see Figure 10), while the Densenet201 model outperforms the ResNet50 but falls short of the VGG19 model in terms of results.

As shown in Table 3, evaluation metrics for COVID-19 and Normal classes are the highest metrics in all models. In contrast, the evaluation metrics for classes Bacterial and Viral are much lower. All models have difficulty distinguishing between Bacterial and Viral infections as shown in confusion matrices in Figure 11. The sensitivity metric for classes Bacterial and Viral is much lower than COVID-19 and Normal.



**Figure 11.** Confusion matrices were computed using (a) ViT, (b) DenseNet201, (c) VGG19, (d) Resnet50, (e) CovEnViT.

## 6. Conclusions

In this study, we presented a new model that synergistically combines the power of CNN and ViT to tackle the challenging task of pneumonia classification. By integrating these two distinct yet complementary architectures, we aim to leverage the exceptional feature extraction capabilities of CNNs alongside the self-attention mechanisms inherent in ViT, thus harnessing the collective strengths of both approaches. The proposed approach aspires to enhance the accuracy and efficacy of pneumonia classification, ultimately contributing to improved diagnostic outcomes and patient care in the realm of respiratory health. The proposed model combines a local filter with ViT, by adding

CovEn blocks to capture local relations. CovEn Blocks consists of convolutional layers, to enable the model to learn local relationships within the spatial domain of the image. The results showed that this model improves the classification performance compared to the standard ViT. Additionally, the models VGG19, ResNet50, DenseNet201, and ViT-B/16 were fine-tuned to classify pneumonia on the public X-ray images. The experiment demonstrated that this study unveiled noteworthy classification performance with the ViT achieving better classification performance than the CNN-based models. The proposed approach achieved higher classification performance compared to the standard ViT.

## Disclosure statement

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

## References

- Abiwinanda, N., Hanif, M., Hesaputra, S. T., Handayani, A., & Mengko, T. R. (2019). Brain tumor classification using convolutional neural network. In *World Congress on Medical Physics and Biomedical Engineering 2018: June 3–8, 2018, Prague, Czech Republic* (Vol. 1, pp. 183–189). Springer Singapore.
- Agarap, A. F. (1803). Deep learning using rectified linear units (relu). *arXiv* 2018.
- Avola, D., Bacciu, A., Cinque, L., Fagioli, A., Marini, M. R., & Taiello, R. (2022). Study on transfer learning capabilities for pneumonia classification in chest-X-rays images. *Computer Methods and Programs in Biomedicine*, 221, 106833. <https://doi.org/10.1016/j.cmpb.2022.106833>
- Ayan, E., & Ünver, H. M. (2019). Diagnosis of pneumonia from chest X-ray images using deep learning. In *2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science* (pp. 1–5). IEEE
- Brady, A., Laoide, RÓ, McCarthy, P., & McDermott, R. (2012). Discrepancy and error in radiology: Concepts, causes and consequences. *The Ulster Medical Journal*, 81(1), 3.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*, arXiv:2010.11929.
- Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint*, arXiv:2008.05756.
- Hariri, M., & Avşar, E. (2023). COVID-19 and pneumonia diagnosis from chest X-ray images using convolutional neural networks. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 12(1), 17. <https://doi.org/10.1007/s13721-023-00413-6>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).
- Huy, V. T. Q., & Lin, C.-M. (2023). An improved densenet deep neural network model for tuberculosis detection using chest X-ray images. *IEEE Access*.
- Khan, S., Islam, N., Jan, Z., Din, I. U., & Rodrigues, J. J. (2019). A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 125, 1–6. <https://doi.org/10.1016/j.patrec.2019.03.022>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 60(6), 84–90.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. <https://doi.org/10.1109/5.726791>

- LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems* (pp. 253–256). IEEE.
- Park, S., Kim, G., Oh, Y., Seo, J. B., Lee, S. M., Kim, J. H., Moon, S., Lim, J.-K., & Ye, J. C. (2021). Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus. *arXiv preprint*, arXiv:2103.07055.
- Pham, T.-C., Luong, C.-M., Visani, M., & Hoang, V.-D. (2018). Deep CNN and data augmentation for skin lesion classification. In *Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19–21, 2018, Proceedings, Part II 10* (pp. 573–582). Springer.
- Pham, T. C., Tran, C. T., Luu, M. S. K., Mai, D. A., Doucet, A., & Luong, C. M. (2020). Improving binary skin cancer classification based on best model selection method combined with optimizing full connected layers of deep CNN. In *2020 International conference on multimedia analysis and pattern recognition (MAPR)* (pp. 1–6). IEEE.
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., & Shpanskaya, K. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint*, arXiv:1711.05225.
- Sait, U., Lal, K., Prajapati, S., Bhaumik, R., Kumar, T., Sanjana, S., & Bhalla, K. (2020). Curated dataset for COVID-19 posterior-anterior chest radiography images (X-rays). *Mendeley Data*, 1.
- Sharma, H., Jain, J. S., Bansal, P., & Gupta, S. (2020). Feature extraction and classification of chest x-ray images using CNN to detect pneumonia. In *2020 10th international conference on cloud computing, data science & engineering (Confluence)* (pp. 227–231). IEEE.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv:1409.1556.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wang, T., Nie, Z., Wang, R., Xu, Q., Huang, H., Xu, H., Xie, F., & Liu, X.-J. (2023). Pneunet: Deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using vision transformer. *Medical & Biological Engineering & Computing*, 1–14.