electronics

MDPI

Article

CoroTrans-CL: A Novel Transformer-Based Continual Deep Learning Model for Image Recognition of Coronavirus Infections

Boyuan Wang, Du Zhang * and Zonggui Tian

School of Computer Science and Engineering, Faculty of Innovation Engineering, Macau University of Science and Technology, Taipa, Macau SAR 999078, China
* Correspondence: duzhang@must.edu.mo

Abstract: The rapid evolution of coronaviruses in respiratory diseases, including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), poses a significant challenge for deep learning models to accurately detect and adapt to new strains. To address this challenge, we propose a novel Continuous Learning approach, CoroTrans-CL, for the diagnosis and prevention of various coronavirus infections that cause severe respiratory diseases using chest radiography images. Our approach is based on the Swin Transformer architecture and uses a combination of the Elastic Weight Consolidation (EWC) and Herding Selection Replay (HSR) methods to mitigate the problem of catastrophic forgetting. We constructed an informative benchmark dataset containing multiple strains of coronaviruses and present the proposed approach in five successive learning stages representing the epidemic timeline of different coronaviruses (SARS, MERS, wild-type SARS-CoV-2, and the Omicron and Delta variants of SARS-CoV-2) in the real world. Our experiments showed that the proposed CoroTrans-CL model achieved a joint training accuracy of 95.34%, an F1 score of 92%, and an average accuracy of 83.40% while maintaining a balance between plasticity and stability. Our study demonstrates that CoroTrans-CL can accurately diagnose and detect the changes caused by new mutant viral strains in the lungs without forgetting existing strains, and it provides an effective solution for the ongoing diagnosis of mutant SARS-CoV-2 virus infections.

Keywords: continual learning; coronaviruses; swin transformer

Citation: Wang, B.; Zhang, D.; Tian, Z. CoroTrans-CL: A Novel Transformer-Based Continual Deep Learning Model for Image Recognition of Coronavirus Infections. *Electronics* **2023**, *12*, 866. https://doi.org/10.3390/electronics12040866

Academic Editors: Cheng Siong Chin, Kalyana C. Veluvolu, Mazdak Zamani and Len Gelman

Received: 14 January 2023
Revised: 31 January 2023
Accepted: 4 February 2023
Published: 8 February 2023

1. Introduction

Coronavirus disease 2019 (COVID-19) has become the most widespread respiratory infectious disease of the 21st century [1], infecting more than 657,060,111 people and causing 6,669,951 deaths worldwide as of 17 December 2022 [2]. The rapid mutation and emergence of immune escape variants such as Delta and Omicron have made testing for the virus a challenging task for public health workers. Currently, real-time reverse transcription-polymerase chain reaction (RT-PCR) testing have a false-negative rate in experimental testing, requiring repeat testing to reduce misdiagnosis [3,4]. Chest computed tomography (CT) can be used to improve sensitivity in diagnosing COVID-19 cases [5,6], with the main findings in chest CT being ground-glass opacities, pulmonary consolidation and 'leaving stone' signs after SARS-CoV-2 infection [7]. These findings, together with RT-PCR results, clinical symptoms, and epidemiological history, are the sole basis for the diagnosis or exclusion of COVID-19 pneumonia.

To achieve automatic early warning for COVID-19, some studies have attempted to develop models that can automatically identify COVID-19 patients by learning lesion characteristics using artificial intelligence technology. Most of these studies have used convolutional neural networks (CNNs) to automatically identify COVID-19 patients based on chest CT images [8,9]. Although CNNs have demonstrated their ability to solve various clas-

sification problems, they are not ideal for tasks requiring high-level categorization where global features such as patterns, multiplicity and distribution need to be considered [6].

Since 2017, there have been significant advances in deep learning algorithms, applications and technologies, such as the Transformer architecture proposed by Ashish et al. [10], which has become a highlight in the field of deep learning and deep neural networks. The Transformer architecture has evolved and expanded beyond its original language tasks to other domains such as bioinformatics, where it is the key technology for DeepMind's protein structure prediction model, AlphaFold [11]. More recently, the Transformer architecture entered the field of computer vision with the proposal of the Vision Transformer (ViT) by Dosovitskiy et al. [12], which has replaced CNNs in many complex tasks. Unlike the traditional convolution of CNNs, the main architecture of ViT consists of multiple stacks of Transformer blocks based on a self-attention mechanism. The multi-headed attention mechanism in the Transformer architecture can establish long-range dependencies on a target to extract more powerful features and capture global context information, making it suitable for disease detection in complex environments [13,14]. In addition, ViT's closer alignment with human cognitive features allows it to outperform CNNs in generalization under most distributed transformations (DS), with a top-1 correct rate that is 5–10% higher than the corresponding CNNs for the same or fewer number of parameters [15]. Other Transformer-based computer vision models include the Swin Transformer [16] and CrossViT [17]. Some recent studies using these architectures have achieved better classification results [13,18].

CNNs and ViT have demonstrated high performance in COVID-19 detection, outperforming general radiologists in certain tasks. However, many of these models are primarily based on fixed datasets and static environments that do not consider the gradual provision of information over time and therefore cannot adapt or learn new knowledge. In some cases, these models completely fail or show significant performance degradation on previously learned tasks, leading to problems of catastrophic forgetting [19] and limited intelligence. McCloskey and Cohen [20] were the first to identify catastrophic forgetting in neural network models. They found that when neural networks are trained on a new task or category, they often forget the information learned in the previous training task. The weights of the latest task can overwrite the weights of the previous task, leading to a decline in model performance, which is known as the stability–plasticity dilemma [21]. In contrast to neural networks, the human learning capacity consists of a rich set of neurocognitive and brain memory mechanisms that facilitate the development of learning skills and the consolidation of long-term memory [22].

Inspired by cognitive science, Continual Learning, also referred to as Lifelong Learning, is a research area that aims to address such problems in artificial intelligence [23]. Its goal is to increase the adaptive capacity of models so that they can learn different tasks at different times (plasticity) without forgetting the characteristics of previous tasks (stability), as well as to make trained models more general. Based on the method of historical information retention, Continual Learning methods can be classified into three categories: rehearsal methods, regularization methods, and parameter isolation methods. Rehearsal approaches work by retaining some historical data or high-level representations. When learning a new task, the old task data are simultaneously replayed to reduce model forgetting [24,25]. Regularization methods constrain the optimization direction of the model on the new task in a way that minimises catastrophic forgetting. This includes adding distillation losses to the old model as the target, optimizing constraints on essential model parameters, and projecting the gradient direction of the parameters [26,27]. Parameter isolation methods extend the old model to new tasks by isolating the parameters of the old and new models to reduce the occurrence of catastrophic forgetting [28,29].

Currently, there is a lack of research and application of Continual Learning methods for artificial intelligence-based solutions in COVID-19 detection based on CT and X-ray images. To address these issues, we undertook this study and summarise our main contributions as follows:

1. We established a benchmark dataset of CT and chest X-ray images of pneumonia for Continuous Learning in medical image classification tasks. The dataset included normal, wild-type SARS-CoV-2, SARS-CoV-2 Omicron and Delta variant, and other viral pneumonia infection CT images, as well as normal, MERS, SARS, wild-type SARS-CoV-2, SARS-CoV-2Omicron and Delta variant, other viral pneumonia, and bacterial pneumonia infection chest X-ray images.
2. We designed a five-stage incremental learning task based on the real-world epidemic timeline of different coronaviruses (SARS, MERS, wild-type SARS-CoV-2, and the Omicron and Delta variants of SARS-CoV-2). We compared several models using three deep learning architectures (CNNs, ViT and MLP) and found that the Swin Transformer model with the highest classification accuracy on learning was the most suitable feature extraction backbone for our study. We also propose a novel imaging approach, CoroTrans-CL, based on the Swin Transformer architecture and the Continuous Learning strategies of regularization-based and rehearsal-based learning approaches to recognise CT and chest X-ray images of coronavirus infections, which cause severe respiratory infections, and to mitigate the problem of catastrophic forgetting and performance stagnation.
3. To address the issue of representative sample selection, we used the Herding Selection strategy to minimise the feature centre distance of sub-samples from the full dataset. We also conducted extensive ablation experiments to compare the effects of different Continual Learning methods and different sample selection strategies on the results in order to verify the effectiveness of each key component.

This paper is structured as follows. We review some of the literature on artificial intelligence in the fight against the COVID-19 pandemic in Section 2, and we describe the sources and construction methods for our datasets and our proposed approach, as well as performance evaluation metrics and model parameter settings for our experimental study, in Section 3. The experimental results and discussion are presented in Sections 4 and 5, respectively. Finally, in Section 6, we conclude the paper with comments on future work.

2. Related Works

In this section, we review the primary research methods used for current COVID-19 case detection. CNNs are the most commonly used approaches for automated COVID-19 diagnosis. Previous studies have mainly used pre-trained networks such as variants of Very Deep Convolutional Networks (VGG) [30], Residual Network (ResNet) [31], Densely Connected Convolutional Networks (DenseNet) [32], Google Inception Network (Inception) [33], Deep Learning with Depth-Wise Separable Convolutions (Xception) [34], Efficient Convolutional Neural Networks for Mobile Vision Applications (MobileNet) [35], and Rethinking Model Scaling for Convolutional Neural Networks (EfficientNet) [36] as deep learning frameworks. These models adapt to the new task of COVID-19 patient detection and classification by modifying or adding custom layers and by transferring knowledge from previous experience. For example, Brunese et al. [37] proposed two models using the VGG-16 network as a backbone model based on transfer learning. The first network is used to identify whether the target is healthy or has pneumonia. If the first network gives a positive prediction, the second network is used to identify COVID-19. The VGG-16 network achieved 98% accuracy for three-class classification. ResNet is another common CNN architecture that avoids gradient disappearance problems compared with earlier architectures such as VGG. Using a Residual Network, Narin et al. [38] classified COVID-19 cases and healthy cases with ResNet-50, achieving the highest accuracy (98%) for binary classification. Other studies have used more efficient architectures such as DenseNet and EfficientNet. Wang et al. [39] developed a COVID-19 pneumonia classification pipeline using DenseNet-121, which achieved an AUC with an overall performance of 0.88–0.99 across different datasets. Shamila et al. [40] used the EfficientNet architecture to build a classification model with 95% accuracy and a 93% F1 score on the test set.

Although CNNs are effective for image classification in deep learning, they have some conceptual limitations. During maximum pooling, CNNs lose information about the location of entities. They also do not consider some spatial relationships between simple objects and require a large receptive field to capture long-range dependencies, which leads to the development of large kernels or highly massive networks and results in a complex model that is difficult to train [13]. To overcome these limitations of CNNs, some researchers have used other architectures such as capsule neural networks (Capsnets) [41] and ViT [12] for COVID-19 classification. Sabour et al. proposed Capsnets [42], a new neural network architecture that uses location and orientation information to perform object recognition, to address the shortcomings of CNNs. Toraman et al. [41] proposed a five-convolutional layer Capsnets model with 16, 32, 64 and 128 kernels in the first four layers and 32 capsules in the fifth layer. After 10-fold cross-validation and 50 epochs of training, the model achieved 84.22% accuracy for multiclassification.

Recent research in COVID-19 detection has focused on the Transformer architecture [10]. Dosovitskiy et al. [12] applied the standard Transformer architecture to image recognition and proposed the use of self-attention in ViT to approach or outperform the state-of-the-art (SOTA) model on several image recognition benchmarks. A few studies have proposed the use of ViT in COVID-19 recognition algorithms. Shome et al. [14] created a dataset of 30,000 images and trained the ViT model on it, achieving 92% accuracy and 98% AUC, outperforming CNNs such as EfficientNet-B0, Inception-V3 and ResNet-50 in multiclassification. Mondal et al. [13] proposed a network based on the ViT-B/16 architecture and achieved the highest accuracy of 98.1%, outperforming most existing methods.

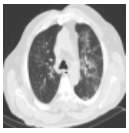
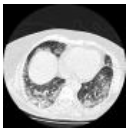
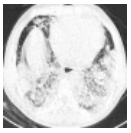
3. Materials and Methods

3.1. Datasets

The evaluation of our proposed approach involved the creation of a comprehensive benchmark image dataset, CL-COVIDset, specifically designed for continuous learning in medical image classification tasks. This dataset included a variety of images, including CT scans and X-rays, representing different types of coronavirus infections and other viral and bacterial pneumonia infections. The datasets used to create the CL-COVIDset are publicly available. The CT images in the CL-COVIDset consist of normal scans [43] and scans showing infections caused by the wild-type SARS-CoV-2 strain [43] and its Omicron and Delta variants [44]. The dataset also included CT images of other viral pneumonia infections [43]. X-ray images in the CL-COVIDset also included normal scans and those showing infections caused by MERS [45], SARS [45], wild-type SARS-CoV-2 [45], the Omicron and Delta variants [44], other viral pneumonia infections [46,47], and bacterial pneumonia [48] infections.

The CL-COVIDset consisted of three sets: a training set, a validation set, and an evaluation set (see Table 1). The use of the validation set allowed us to fine-tune our model, and the model was finally tested for performance on the evaluation set. The performance of the model on the evaluation set was assessed using observed and unobserved data and distributions. To facilitate the use of CL-COVIDset by the wider research community, we have made it publicly available on the Kaggle platform at the following link: https://www.kaggle.com/datasets/mustai/continual-learning-of-covid19 (accessed on 31 January 2023).

Table 1. Details of the CL-COVIDset dataset.

Image Types	Class	Images	Dataset		
			Train and Val	Evaluation	Total
CT	Normal		954	285	1239
	Wild-type SARS-CoV-2		450	120	570
	Omicron and Delta variants of SARS-CoV-2		538	191	729
	Other pneumonias		447	118	565
X-ray	Normal		470	60	530
	Bacterial pneumonia		283	60	343
	SARS		106	27	133
	MERS		135	49	184
	Wild-type SARS-CoV-2		205	52	257
	Omicron and Delta variants of SARS-CoV-2		111	50	161
	Other viral pneumonias		240	60	300
	Total		3939	1072	5011

3.2. Methods

3.2.1. Methodology of CoroTrans-CL

Figure 1 illustrates the proposed methodology, CoroTrans-CL, for disease detection. It is divided into three components: data augmentation, the CoroTrans model with a feature extraction backbone and a classification head, and the Continual Learning strategy.

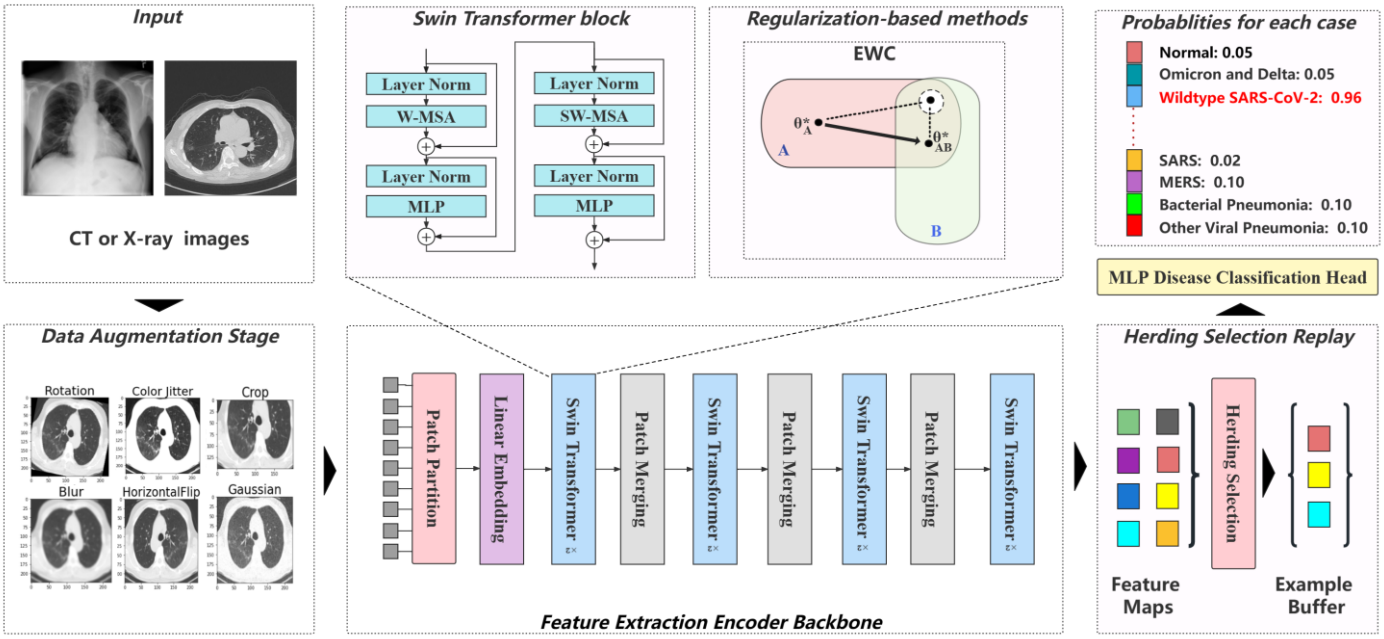


Figure 1. The CoroTrans-CL framework: a Continual Learning approach utilizing a Swin Transformer network for disease detection.

The data augmentation module is used to increase the diversity of the input data by applying various transformations to the images, including random rotation, cropping, blurring and noise addition. This helps to improve the generalizability of the model by exposing it to a wider range of data variations. To further increase the randomness of the data, the module applies a random order command to shuffle the order in which these transformations are applied. The resulting augmented dataset is then normalised using the mean and standard deviation.

The **CoroTrans model** is an artificial intelligence system designed to perform disease classification tasks. The architecture of the model consists of two main layers, namely, a feature extraction backbone network layer and a disease classifier layer. The feature extraction backbone network layer is responsible for encoding the input data into a feature representation, which is then used as input for the disease classifier layer.

This layer can be thought of as the core or foundation of the model, providing a basic structure upon which additional functionality can be built, similar to the concept of a 'backbone' in a network. The disease classifier layer, on the other hand, uses the feature representation generated by the feature extraction backbone network layer to perform the actual disease classification. This layer can be implemented using various techniques, such as using a neural network as the classifier, where the neural network can be trained to classify the input based on the feature representation generated by the feature extraction layer.

- Feature Extraction Backbone

The encoder backbone is a hierarchical structure consisting of four stages. The patch partition stage divides the input RGB image into non-overlapping patches, each of which is treated as a token. These patches are then processed through multiple Swin Transformer blocks [16] that consist of interconnected Window and Shift Window Multi-Head Self-

Attention (W-MSA and SW-MSA)-based Transformer blocks. These blocks enhance the computational performance of the Window and Shift Window methods and are governed by computational Equations (1) to (4), including the LayerNorm, Window Attention, Shifted Window Attention, and MLP modules. The Swin Transformer is a novel attention-based transform architecture specifically designed for the efficient processing of image data. It exploits the local structure of images by partitioning them into patches and only applying self-attention within each patch rather than over the entire image. This allows the Swin Transformer to effectively model the long-range dependencies present in images while maintaining a high degree of computational efficiency. In CoroTrans-CL, the patch size is 4×4 and the feature dimension of each patch is $4 \times 4 \times 3$, resulting in patch tokens with a (H/4, W/4, $4 \times 4 \times$ channel) shape. The use of the Swin Transformer has been shown to significantly improve the performance of image classification tasks compared with other Transformer architectures. This is due in part to its ability to effectively model local structure and long-range dependencies, as well as its high computational efficiency. The block design of the Swin Transformer, consisting of interconnected Window and Shift Window Multi-Head Self-Attention-based Transformer blocks, has also been shown to be an effective method for improving performance, particularly in the context of image processing.

$$\hat{x}^i = W - MSA\Big(LN(x^{i-1})\Big) + x^{i-1}, \tag{1}$$

$$x^i = MLP\Big(LN(\hat{x}^i)\Big) + \hat{x}^i, \tag{2}$$

$$\hat{x}^{i+1} = SW - MSA\Big(LN(x^{i-1})\Big) + x^i, \tag{3}$$

$$x^{i+1} = MLP\Big(LN(\hat{x}^{i+1})\Big) + \hat{x}^{i+1}, \tag{4}$$

where $\hat{x}^i$ is (S)W-MSA's output, x^i is the output of MLP, and i represents the block's position. The output shapes of the tokens are (224/8, 224/8, 2C), (224/16, 224/16, 4C), and (224/32, 224/32, 8C) for stages 2, 3, and 4, respectively. The resolution of the output features is 7×7, and the channel has 768 dimensions, as does the output of the encoder stage.

- Disease Classification Head

The CoroTrans model features a disease classification head that was specifically designed for the identification of various pathologies present in chest X-ray and CT images. The classifier is implemented as a multi-layer perceptron consisting of several linear layers. The input to the classifier is a one-dimensional feature vector of 768 dimensions derived from the image data, which is transformed by the linear layers to predict the target pathology among 11 classes. The final output of the classifier represents the prediction of the model, providing a diagnostic tool for physicians.

Continual Learning, also known as Lifelong Learning, refers to the ability of artificial intelligence models to sequentially adapt and learn new tasks without forgetting previously trained tasks [22]. This approach is particularly relevant in addressing the challenges posed by rapidly evolving environments, such as the COVID-19 pandemic, where new information and data are constantly being generated. Continual Learning strategies aim to improve the adaptability of AI models by allowing them to gain new knowledge for new tasks without forgetting previous information. Formally, sequential learning can be defined as the ability of a model to learn individual distributions $D_1, \ldots, D_n$, at $T_1, T^2 \ldots$, moments while being tested on a set containing all distributions. The goal of Continual Learning is to collect data from the new distribution D_{n+1} at T_{n+1} and to update the model parameters θ while the model simultaneously adapts to all distributions $D_1, \ldots, D_{n+1}$. In this study, we employed HSR [25] and EWC [27] as a combined COVID-19 Continual Learning strategy (HSR-EWC):

- Herding Selection Replay.

The classical replay strategy is a widely used technique to address the challenge of forgetting in Continuous Learning scenarios [49,50]. It involves storing and replaying relevant samples from previous training sets as experience, with the aim of improving the adaptability of the model to learn new tasks without forgetting previously acquired knowledge. A specific implementation of the classical replay strategy is the HSR method [25,51], which involves selecting representative examples that are as close as possible to the centre of the feature space. Algorithm 1 describes the representative image selection process of the HSR strategy.

Algorithm 1: Herding Selection Replay.

Input: Image set of CL-COVIDset, set $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ of class label y; The number of classes is n; Maximum size of the replay memory buffer; Swin Transformer backbone model **S** for feature extraction;

Training model **S** to obtain the feature maps θ and feature function φ of the extracted CT and X-ray images;

Initialise L to an empty list.

Calculate the mean μ of the samples in the class by θ;

for 1, 2, ... , m do

$$L_m \leftarrow \arg\min_{\alpha \in \boldsymbol{\alpha}} \| \mu - \frac{\varphi(\alpha) - \sum_{i=1}^{m-1} \varphi(L_i)}{n} \|$$

end for

L $\leftarrow (L_1, \ldots, L_m)$;

Output: Exemplar buffer list L;

The HSR method involves up-sampling using the feature extraction backbone during the training phase, followed by calculating the mean of the samples in each class. The distance of each individual sample to the class mean is then calculated and the closest distance ranking is generated. For each class, the top n most representative samples are selected based on this ranking to form a representative subset of samples that are stored in memory. The total number of representative samples in memory (set to N = 200 in this study) is equally divided between the learned classes, with the number of classes being dynamically adjusted according to the learning process. An advantage of the HSR method is that the final sample means are close to the actual class means, allowing the samples to better represent the classes to which they belong [52]. This contrasts with other methods, which can result in sample means that are far from the actual class means, leading to the less effective representation of the classes. The HSR approach has been shown to be effective in improving the Continuous Learning capabilities of artificial intelligence models.

- Elastic Weight Consolidation.

EWC enables Continual Learning by reducing the plasticity of synapses important to previous tasks [27]. As shown in Figure 2, the parameters (weights and biases) of tasks $\mathcal{T}_A$ and $\mathcal{T}_B$ are denoted as θ_A and θ_B, respectively, and the sets of parameters that reduce the errors for tasks A and B are Θ_A^* and Θ_B^*, respectively. It is possible to find a solution with $\theta_A \in \Theta_A$ and $\theta_B \in \Theta_B$.

Under a Bayesian perspective, if the data are divided into two independent parts, the D_A of $\mathcal{T}_A$ and D_B of $\mathcal{T}_B$, the posterior distribution $p(\theta|D)$ is estimated with (5).

$$\log p(\theta|D) = \log p(D_B|\theta) + \log p(\theta|D_A) - \log p(D_B), \quad (5)$$

As it is not possible to compute the true posterior probability, the EWC assumes that it is a Gaussian distribution with a mean given by the parameter θ_A^* and the Fisher

information matrix F is used to estimate diagonal precision [53]. The loss function is built in (6).

$$\updownarrow(\theta) = \updownarrow_{\mathrm{B}}(\theta) + \sum_{i} \frac{\lambda}{2} F_i (\theta_i - \theta^*_{A,\,i})^2, \tag{6}$$

where $\updownarrow_{\mathrm{B}}(\theta)$ is the loss function specific to task $\mathcal{T}_B$, $\rangle$ is the weight vector of the index, $\theta^*_{A,\,i}$ denotes the parameter after learning task $\mathcal{T}_A$, F denotes the Fisher information matrix, and λ is the parameter which determines the relative importance of the old and new tasks. [54].

When task B arrives, EWC uses parameter θ close to θ^*_A in Equation (6), and when a third task $\mathcal{T}_c$ arrives, EWC continues to make the parameter θ close to θ^*_{AB}, where $\theta^*_{A,B}$ is the parameter learned from tasks $\mathcal{T}_A$ and $\mathcal{T}_B$. Extending to all $\mathcal{T}$ tasks, the optimization objectives of EWC are given in (7).

$$\theta^*_{\mathrm{T}} = \underset{\theta}{\operatorname{argmin}} \left\{ -\log p(D_{\mathcal{T}}|\theta) - \frac{1}{2} \sum_{i} \Big(\sum_{t<T} (\lambda_{\mathrm{t}} F_{t,i}) (\theta_i - \theta^*_{T-1,\,i})^2 \right\} \tag{7}$$

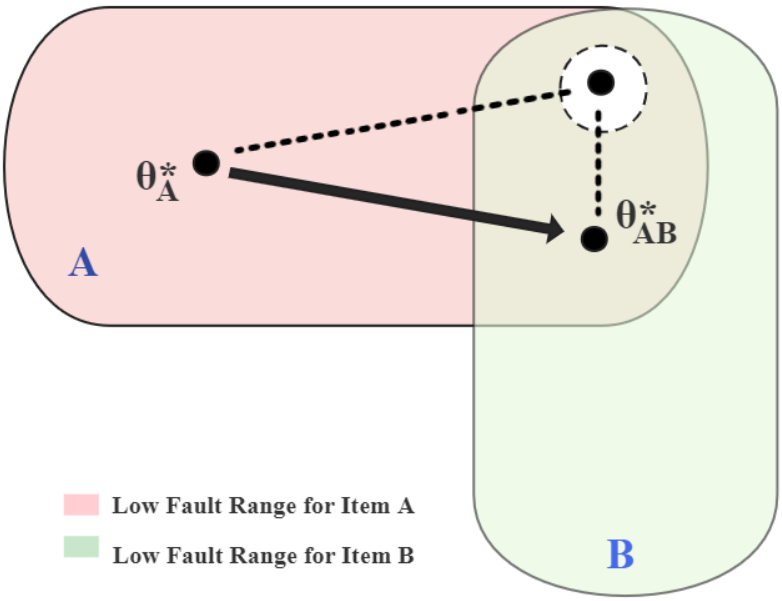


Figure 2. Search space for Elastic Weight Consolidation strategies.

3.2.2. Performance Evaluation Metrics

In this study, we used a number of Continual Learning evaluation metrics to assess the performance of the models. One such metric was average accuracy [55], which measures the average accuracy of a model after class-incremental training for the first task up to T. This metric is calculated with formula (8). Average accuracy is a widely used metric in the Continual Learning literature, as it provides a comprehensive view of a model's performance across all tasks. Unlike other metrics, such as per-task accuracy, which only consider performance on individual tasks, average accuracy takes a model's performance on all tasks into account, providing a more comprehensive assessment of a model's ability to continuously learn new tasks without forgetting previously acquired knowledge.

$$Average\ Accuracy = \frac{1}{\mathcal{T}} \sum_{i=1}^{\mathcal{T}} p_{\mathcal{T},\,i}, \tag{8}$$

where $p_{j,\,i}$ is the Top1 accuracy of the model on the held-out test set of task $\mathcal{T}_i$ after the model is trained on task $\mathcal{T}_j$. The precision, sensitivity and F1 score are defined as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positive}, \tag{9}$$

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negative} \tag{10}$$

$$F1 - score = \frac{2 \times (Sensitivity \times Precision)}{Sensitivity + Precision} \tag{11}$$

To assess the overall performance of the models in this study, the average of each metric was calculated using the MACRO method [39]. The confusion matrix is a valuable tool for analysing the error types of a model, as it provides a breakdown of the number of true positive, true negative, false positive and false negative predictions made by a model. This information is useful for identifying patterns in a model's errors and for developing strategies to improve a model's performance.

3.2.3. Experimental Setup

In this study, we compared the performance of several state-of-the-art deep learning models as the backbones for CT and chest X-ray image classifiers. The models used for comparison included an all-MLP Architecture for Vision (MLP-Mixer) B/16 [56], a multi-layer perceptron model with a mixer block architecture (ResNet-50) [31], a widely used convolutional neural network (CNN) model with a residual architecture (Efficientnet-b4) [36], a CNN model with an efficient architecture designed to improve performance while reducing the number of parameters and computational complexity (ViT-S/16) [12], a vision Transformer model with a small patch size and BERT Pre-Training of Image Transformers (BeiT) v2 [57], and a hybrid Transformer model that combines the strengths of both CNNs and Transformers. In addition to these models, we also included our own model for comparison.

To evaluate the classification ability of our proposed model (CoroTrans) and the other models, we performed a joint learning experiment. The experiment involved simultaneously training all models on all classes of images. In this setup, each model is trained for a certain number of iterations, called epochs. In our experiment, the models were trained for 30 epochs. The training process for the models involved updating their parameters to minimise the difference between the predicted output and the actual output. This process as conducted using an optimisation algorithm. In our experiment, we used the Adaptive Moment Estimation Decoupling Weight Decay (AdamW) optimiser. The Adam optimiser [58] is a popular optimisation algorithm that is widely used in deep learning and is particularly well-suited to training large neural networks. It is a combination of two other optimisation techniques, namely, the Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp). The Adam optimiser has several advantages over other optimisation techniques, including the ability to adaptively adjust the learning rate on a per-parameter basis. This means that the optimiser can adjust the learning rate for different parameters based on the past gradients of the parameters, resulting in faster convergence. In addition, Adam has a momentum term that helps to smooth the gradients, allowing the optimiser to converge faster and more stably. AdamW is a modified version of the Adam stochastic optimisation algorithm that improves upon the traditional implementation of weight decay. This is achieved by decoupling the weight decay calculation from the gradient update operations, allowing the weights to be updated in a more effective and efficient manner. The learning rate is another important hyperparameter in the training process. It controls the step size at which the optimiser updates the model parameters. A high learning rate causes the model to make rapid updates to the parameters and the training process can quickly converge, though with the risk of overshooting the optimal solution, while a low learning rate causes the model to make smaller updates, leading to

slower convergence but with less risk of overshooting. In our experiment, we used an initial learning rate of 3×10^{-5}. To further assess the performance of each model, we used a 5-fold cross-validation strategy to derive the average scoring accuracy of each model on the validation set, providing a robust assessment of their performance.

We then conducted a series of class incremental learning experiments, as shown in Figure 3, in which the models were trained on successive sets of images corresponding to different types of coronavirus infection. To this end, we conducted experiments on five levels of Continuous Learning that accurately reflected the epidemic timeline of different coronaviruses (namely, SARS, MERS, wild-type SARS-CoV-2, and the Omicron and Delta variants of SARS-CoV-2) in the real world. The results of these experiments were compared with the baseline results obtained by training the models on all previous classes in addition to the new class each time (the upper bound). The aim of these experiments was to assess the ability of the models to adapt to new information without forgetting previously learned knowledge.

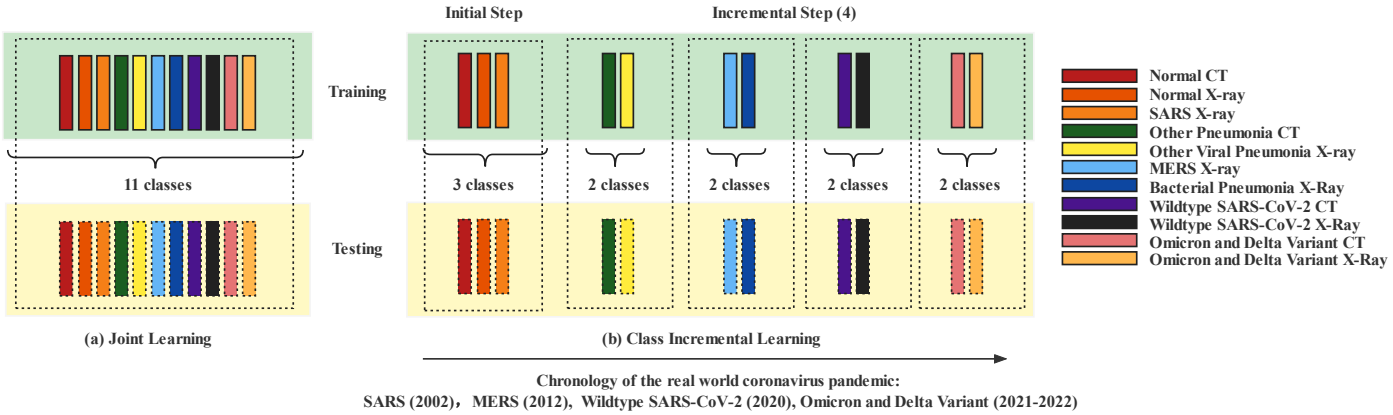


Figure 3. Using AI to simulate the discovery of coronaviruses: a contrast between joint and class-incremental learning experiments: (**a**) joint learning; (**b**) class-incremental learning experiment in which trials were conducted in five successive learning phases reflecting the actual epidemic timeline of different coronaviruses (namely, SARS, MERS, wild-type SARS-CoV-2, and the Omicron and Delta variants of SARS-CoV-2) in the real world.

Finally, we compared several state-of-the-art Continual Learning methods for their performance in adapting to new information without forgetting previously learned knowledge. The methods we considered included Gradient Episodic Memory (GEM) [59], a method that stores past gradients and adjusts the learning rate for each example based on the distance between the current gradient and the past gradients; GDumb [24], a method that stores a fixed number of examples from each task and performs gradient descent on these examples at the beginning of each task; Average Gradient Episodic Memory (AGEM) [60], a method that stores and averages previous slopes to provide a continuous representation of the understanding acquired in previous tasks—this representation is then used to streamline the learning process for the current task, ensuring that the network retains essential data from previous tasks while avoiding harmful interference; CopyWeights with Re-init Star (CWRStar) [61], a rehearsal-free Continuous Learning approach in deep learning, which is a notable way of dealing with forgetting in the Single-Incremental Task–New Classes situation that involves the implementation of a double memory in the fully connected layer; Cumulative (the upper bound), a method that stores all past examples and trains on them in addition to the current task; and Random Replay, a method that randomly samples past examples to train on at the beginning of each task. These methods represent three different strategies for Continuous Learning: gradient-based, example-based, and random sampling. We compared the performance of these methods using a five-step incremental learning task involving the sequential learning of different coronavirus classes. The capacity

hyperparameter of the buffer for all replay methods was uniformly set to store a maximum of 200 samples.

The experiments were conducted using the Tesla A100, V100 and P100 GPU graphics cards, and they were implemented using PyTorch (https://pytorch.org/ (accessed on 31 January 2023)), the PyTorch image model library (https://github.com/fastai/timmdocs/ (accessed on 31 January 2023)), and the PyTorch Continuous Learning framework (https://avalanche.continualai.org/ (accessed on 30 January 2023)) [52]. The use of GPU technology has become increasingly common in deep learning due to its ability to increase the efficiency of the training and inference process through parallel processing. PyTorch, a widely used open-source deep learning framework, provides a high-level interface for training and deploying deep learning models. The timmdocs library developed by fastai is a suite of PyTorch utilities and callbacks for training image classification models. In addition, the avalanche framework is a PyTorch-based toolkit for the design and evaluation of Continuous Learning algorithms.

4. Results

4.1. Joint Training Results

4.1.1. Accuracy Results

The results presented in Table 2 show the superiority of our approach, CoroTrans, a Swing Transformer Network-based model, over the other evaluated models in terms of precision, recall, F1 score and accuracy as performance metrics in joint training. Our approach, CoroTrans, showed exceptional performance with a precision of 97.18%; in addition, our approach achieved an accuracy of 95.34%, further highlighting the effectiveness of CoroTrans in identifying coronavirus-infected respiratory diseases in medical images in joint training. Notably, the performance gain of CoroTrans over other models was significant, with average improvements of 20% in precision, 20% in recall, 20% in F1 score and 16% in accuracy over other models. These results illustrate the potential of our proposed approach to provide a comprehensive and accurate prediction model for Coronavirus-infected respiratory disease.

Table 2. Comparison of baseline models and proposed approach.

Backbone Model	Evaluation Metrics			
	Precision	**Recall**	**F1 Score**	**Accuracy**
MLP-Mixer B/16	0.7994	0.7724	0.7806	0.7919
Efficientnet-b4	0.8967	0.8855	0.8887	0.8843
ResNet-50	0.9202	0.9146	0.9167	0.9244
ViT-B/16	0.9371	0.9276	0.9312	0.9300
BeiT-v2	0.9443	0.9323	0.9373	0.9412
CoroTrans(Ours)	0.9718	0.9716	0.9716	0.9534

4.1.2. Various Diseases Classification Results

The results, as shown in Figure 4 and Table 3, demonstrate the robustness and validity of our proposed model, CoroTrans, for the classification of different diseases. The F1 scores for all diseases were higher than 92%, highlighting the model's ability to accurately discriminate between different disease classes. In particular, CoroTrans showed a superior ability to detect wild-type SARS-CoV-2 and other viral pneumonias in both CT and X-ray images.

The model achieved F1 scores of 98.32% for CT images and 100% for X-rays for the identification of wild-type SARS-CoV-2, a significant improvement over the other compared models. In addition, the model achieved high F1 scores of 98.73% and 100% for CT and X-ray images, respectively, for the identification of other viral pneumonias. Furthermore, the model's performance in identifying the Omicron and Delta variants of SARS-CoV-2 (with F1 scores of 89.06% and 96.97% respectively), while not as high as its performance in

other disease classes, demonstrates its ability to accurately identify these variants, which are known to be more difficult to diagnose.

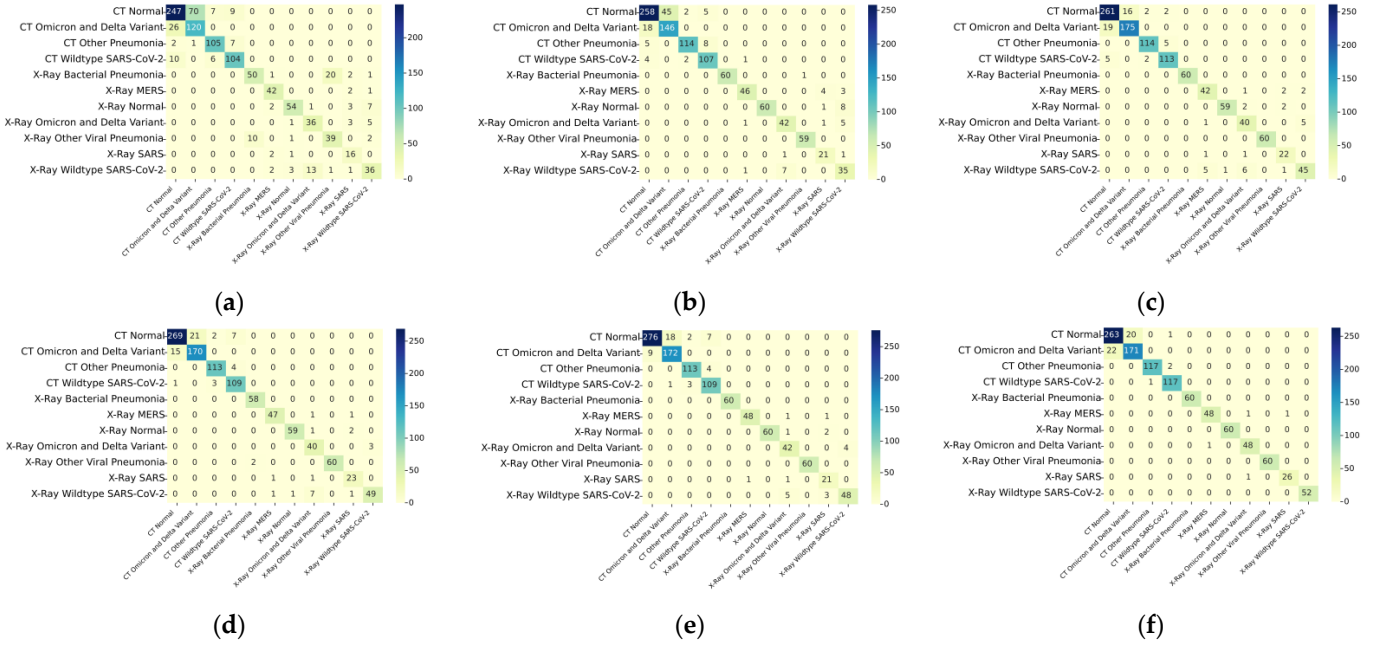


Figure 4. Confusion matrix illustrating the classification results of the models: (**a**) MLP-Mixer B/16; (**b**) Efficientnet-b4; (**c**) ResNet-50; (**d**) ViT-B/16; (**e**) BeiT-v2; (**f**) CoroTrans (ours).

Table 3. Results of CoroTrans classification for different diseases.

Image Types	Class	Evaluation Metrics		
		Precision	Sensitivity	F1 Score
CT	Normal	0.9261	0.9228	0.9244
	Wild-type SARS-CoV-2	0.9915	0.9750	0.9832
	Omicron and Delta Variants of SARS-CoV-2	0.8860	0.8953	0.8906
	Other Pneumonias	0.9832	0.9915	0.9873
X-Ray	Normal	0.9677	1.0000	0.9836
	Wild-type SARS-CoV-2	1.0000	1.0000	1.0000
	Omicron and Delta Variants of SARS-CoV-2	0.9796	0.9600	0.9697
	SARS	0.9630	0.9630	0.9630
	MERS	0.9600	0.9796	0.9697
	Other Viral Pneumonias	1.0000	1.0000	1.0000
	Bacterial Pneumonia	1.0000	1.0000	1.0000

4.1.3. Comparison of Model Feature Extraction

t-Distributed Stochastic Neighbour Embedding (t-SNE) [62] is a dimensionality reduction and visualization technique for high-dimensional data that is particularly useful for visualizing the structure of complex datasets and has been widely used in the field of machine learning to visualise the representations learned by deep neural networks. In our study, we used t-SNE to visualise the disease features learned by our proposed model, CoroTrans, as well as other models based on three major deep learning architectures: CNNs, ViT, and MLP.

We extracted the output of the last layer of the feature extractor in each model to obtain a multidimensional feature vector and projected it into a two-dimensional space using t-SNE. The resulting scatter plots, shown in Figure 5, indicate that the feature distributions obtained by the MLP-Mixer B/16, EfficientNet-b4 and ResNet-50 models did not result in a clear boundary between different disease classes.

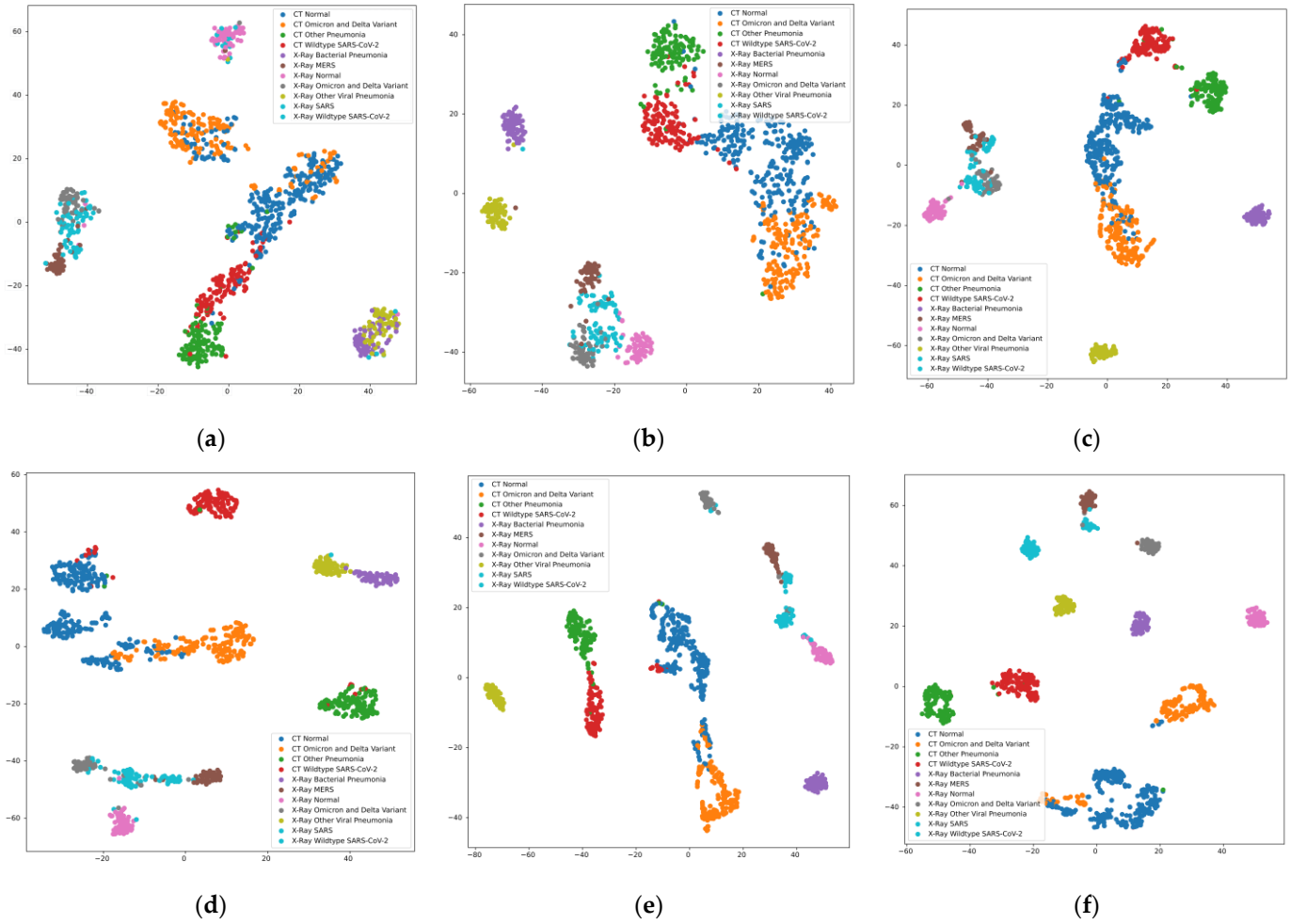


Figure 5. Results of t-SNE visualization for high-dimensional data clustering: (**a**) MLP-Mixer B/16; (**b**) Efficientnet-b4; (**c**) ResNet-50; (**d**) ViT-B/16; (**e**) BeiT-v2; (**f**) CoroTrans (ours).

In contrast, the feature distributions obtained by the ViT-B/16, BeiT-v2 and CoroTrans models had clear boundaries between different classes of features, with CoroTrans showing the clearest separation of each class compared with other techniques. These results indicate that our model effectively captured the relevant features of the images, leading to improved classification performance.

4.2. Continual Learning Results

4.2.1. Average Accuracy and Confusion Matrix Results

The results presented in Table 4 demonstrate the effectiveness of our proposed method, CoroTrans-CL, in mitigating catastrophic forgetting in Continuous Learning tasks. Using the HSR-EWC strategy, our method achieved superior performance compared with other evaluated strategies such as AGEM, CWRStar, GEM, GDumb and Random Replay. In particular, when applied to the Swin Transformer backbone, our method achieved an accuracy of 83.40%, while the next best strategy, Random Replay, achieved an accuracy of 66.04%. This result indicated a significant improvement in performance, around 30% better than other models. In addition, it is worth noting that the average accuracy of our method was the closest to the upper performance limit of the Cumulative method of all evaluated methods. These results highlight the effectiveness of our proposed approach in dealing with catastrophic forgetting and maintaining performance in Continuous Learning tasks.

Table 4. A comparison of the average accuracy of different strategies.

Architecture	Baseline	Evaluation Strategies					
	Cumulative (the Upper Bound)	AGEM	CWRStar	GEM	GDumb	Random Replay	HSR-EWC (Our CL Strategy)
ResNet-50	0.4776	0.2323	0.3209	0.2155	0.3256	0.3269	0.2323
Efficientnet-b4	0.4813	0.2668	0.3218	0.2780	0.0914	0.2351	0.3461
BeiT-v2	0.8834	0.2267	0.3256	0.4646	0.6772	0.3414	0.5718
MLP-Mixer B/16	0.8657	0.3461	0.3358	0.3563	0.4039	0.5765	0.6604
ViT-B/16	0.8983	0.2257	0.3619	0.7453	0.6922	0.653	0.7724
CoroTrans (Our Model)	0.9375	0.2304	0.3479	0.4403	0.5373	0.694	0.8340

4.2.2. Incremental Learning Processes Results

The results of our study, as depicted in Figure 6, illustrate the incremental learning process of the CoroTrans-CL across several Continual Learning strategies. As seen in the figure, other methods such as AGEM, CWRStar, GEM, GDumb, and Random Replay experienced sharp declines in average accuracy, indicating catastrophic forgetting. This is particularly problematic in the context of a medical deep learning model, as it renders the model ineffective in adapting to new, unseen classes of data. On the other hand, our proposed CoroTrans-CL approach, the Swin Transformer backbone combined with the HSR strategy, exhibited a slow degradation of performance, similar to the way the human brain learns. This slow degradation in accuracy maintained the model's ability to recognise new, unseen classes of data without compromising performance on previous classes. This ability to continuously learn and adapt to new virus strains without forgetting previous strains is critical in the fight against coronavirus mutations and pandemics, as it allows for real-time adaptation to new strains while maintaining performance on previous strains.

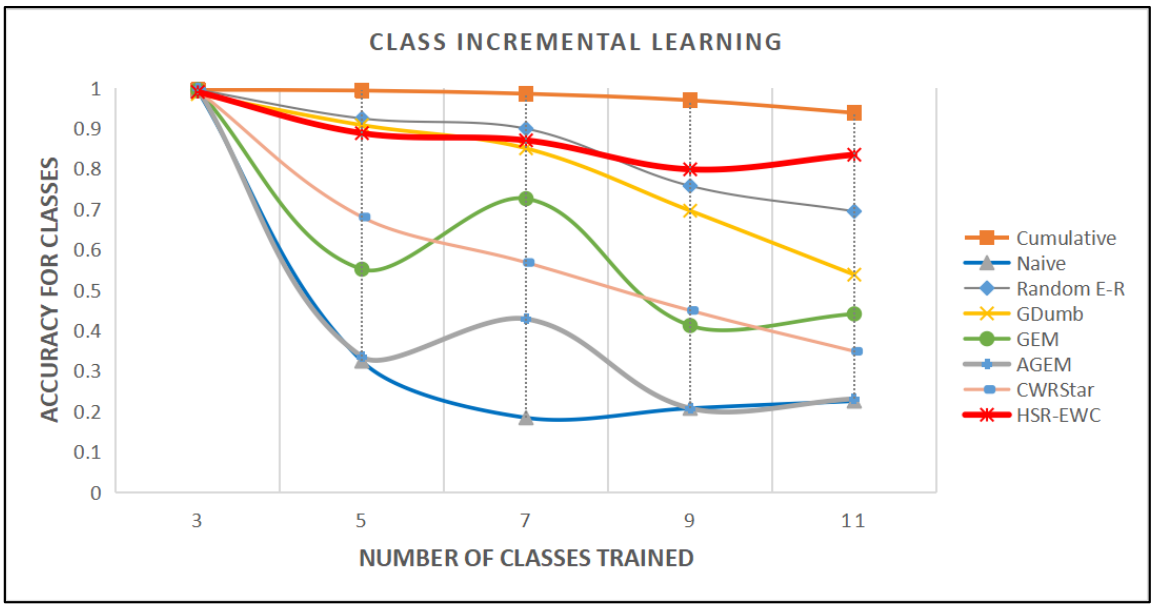


Figure 6. Accuracy on CoroTrans-learnt classes with different strategies.

4.2.3. Ablation Experiments

In our study, we conducted ablation experiments to evaluate the contribution of each component of our proposed method, which combined EWC and HSR to mitigate catastrophic forgetting. Specifically, we compared the performance of three different combinations of methods:

Hybrid 1, where the model was trained using only EWC.

Hybrid 2, where the model was trained using EWC and Replay.

Our method, where the model was trained using EWC, Replay, and Herding Selection.

These ablation experiments were conducted to gain insight into the individual contributions of each component of our proposed method and to determine which combination of methods resulted in the most effective performance. The results presented in Table 5 are essential for understanding the architecture of the model. Understanding the precise impact of each component of the proposed method on incremental learning performance is crucial for making informed decisions about the structure and design of the model.

Table 5. Ablation analysis of average accuracy results.

Methods	Average Incremental Accuracy
Hybrid 1 (EWC)	0.2248
Hybrid 2 (EWC + Replay)	0.7220
Our method (EWC + Replay + Herding Selection)	0.8340

Hybrid 1, using only EWC, was able to maintain the weights of the previously learned classes. However, as the gap between the new and old classes increased, so did the confusion between the old and new classes, leading to a significant decrease in the final incremental average accuracy.

Hybrid 2, which combined EWC and Replay, showed improved performance over Hybrid 1, as the increased storage of example samples helped to reduce confusion between old and new classes. However, as the selection of examples was random, the samples lacked feature representativeness and recognition accuracy was not high after several stages of incremental training.

Our method, which included the addition of the Herding Selection method, improved on Hybrid 2 by using a Herding Selection algorithm to select representative samples near the mean feature after averaging the features extracted by the backbone network. This resulted in a limited number of subsamples that effectively represented the entire sample and led to significantly improved performance and slow forgetting during incremental learning, resulting in the highest final incremental average accuracy.

4.2.4. Comparison of Different Sample Selection Strategies

To analyse the effects of different exemplar selection strategies on incremental learning performance, we compared three different exemplar selection strategies.

Hybrid 1: Random Exemplar Selection strategy, which randomly selected the exemplars in the dataset.

Hybrid 2: Closest to Centre strategy, which is a greedy algorithm that selected the remaining exemplar that is closest to the centre of the feature space based on the already selected elements.

Our method: Herding Selection strategy, which selected the remaining exemplar that brought the centre of the already selected exemplars as close as possible to the overall centre of the feature space by iteratively adjusting the selection criteria.

The results of our experiments, shown in Figure 7, show that the Herding Selection strategy was the most effective method for selecting exemplars in incremental learning tasks. The strategy, which is based on a Herd Selection algorithm, minimised the distance between the feature centres of the selected exemplars and the overall feature centres of the sample. This resulted in a better representation of the replayed samples, which consequently led to a higher average accuracy in the final increment. In contrast, the Random Exemplar Selection strategy, which used stochastic sampling to retain information from old classes, suffered from a lack of representativeness as the gap between the random sub-sample space and the total space gradually increased with increment. The Closest to Centre strategy performed well in the initial stages, but as the incremental training stage gradually increased, the representativeness of the samples selected by the algorithm decreased and recognition accuracy accordingly decreased.

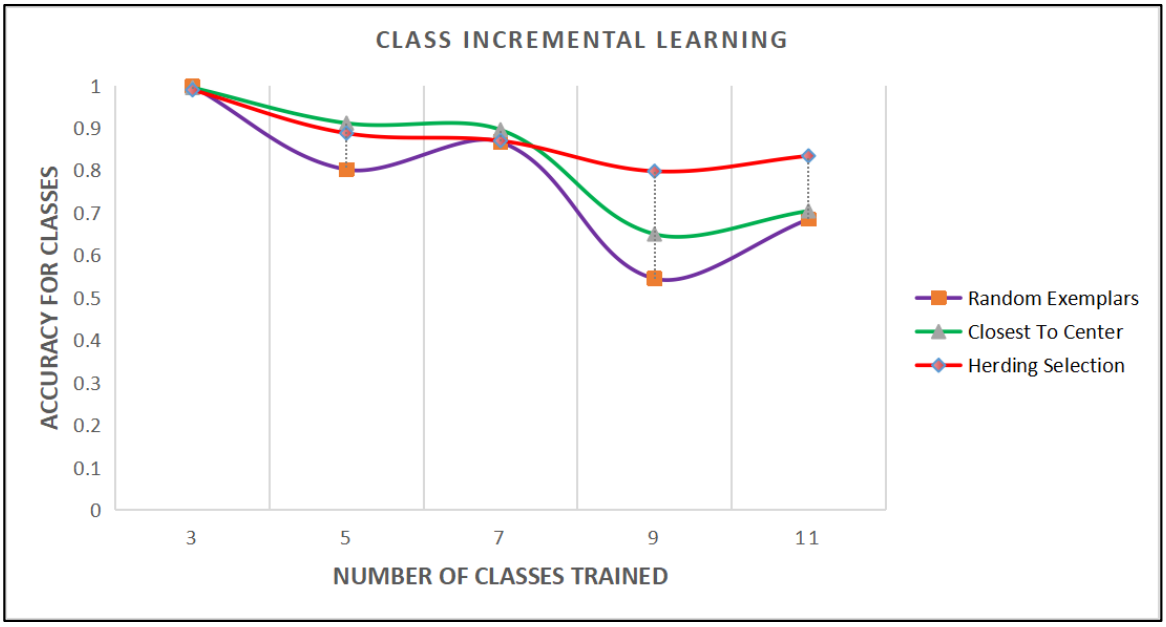


Figure 7. Incremental learning evaluation: comparison of sample selection strategies.

5. Discussion

Our proposed Continuous Learning approach, CoroTrans-CL, based on the Swin Transformer architecture, was developed for the diagnosis and prevention of coronavirus infection using chest radiography images. During the joint training phase, we employed a robust evaluation strategy by using 5-fold cross-validation, a widely accepted method for evaluating the performance of machine learning models. In this approach, the training data were divided into five equally sized subsets called 'folds', and the model was trained on four of these folds, with the remaining fold used as a validation set. This process was repeated five times, with each fold being used as a validation set only once. The advantage of this method, especially for smaller datasets, is that it allows for a more comprehensive evaluation of a model's performance. The performance of our proposed model CoroTrans was validated by a 5-fold cross-validation on the evaluation set, achieving an accuracy of 95.34% in the conclusive results. This outstanding result demonstrates the superior feature extraction capability of our model.

The results presented in Table 2 show that the CoroTrans-CL model had high precision and recall values, with a precision of 97.18%, a recall of 97.16%, and an F1 score of 97.16%, indicating low numbers of both false positive and false negative predictions. These results demonstrate the model's ability to accurately identify infected images while minimizing the number of uninfected images misclassified as infected. In addition, Table 3 illustrates the performance of the CoroTrans model for different diseases, such as normal scans, wild-type SARS-CoV-2, the Omicron and Delta variants of SARS-CoV-2, and other pneumonia infections. The F1 scores for all diseases were greater than 92%, indicating the ability of the model to accurately discriminate between different disease classes and to minimise the number of false positive and false negative predictions. It is worth noting that the high performance of the CoroTrans-CL model in terms of precision, recall and accuracy suggests that it is capable of effectively identifying coronavirus-infected images while minimizing the number of false positive and false negative predictions. Our confusion matrix results further demonstrate the robustness and validity of CoroTrans for the classification of various diseases.

The utilization of the ViT architecture in medical image analysis has been found to be particularly advantageous in this task, as it allows for the processing of input images of arbitrary resolution. The self-attention mechanism enables the model to focus on the most informative regions of the input images, leading to better performance. Additionally,

the Transformer architecture allows for the utilization of a vast number of parameters, leading to improved representation capability. A comparison of the ability of model feature extraction (t-sne) indicated that our model outperformed the comparative CNNs and MLP model architectures in terms of feature extraction, as evidenced by the t-SNE visualization. This emphasises the significance of using Swin Transformer as a feature extraction backbone in our model, as its hierarchical architecture enables the extraction of features at multiple scales, thus leading to improved performance in medical image recognition tasks. These factors collectively contributed to the superior performance of our model CoroTrans in identifying images of different coronavirus infections.

The results of our proposed method in the context of Continual Learning, as presented in Section 4.2, reveal its superior performance compared with other comparative approaches. In particular, the mean accuracy of 83.40% achieved by our proposed method in the Continual Learning setting highlights its ability to effectively adapt to new data and task variations while maintaining its performance on previously learned tasks. In addition, to gain insight into the influence of image resolution on the precision of our proposed approach, we evaluated the classification results of Continuous Learning using a standard resolution image (224×224) and an enhanced resolution image obtained by adapting it to a larger resolution image (384×384). As expected, the higher resolution resulted in a further increase in the overall accuracy of the model recognition, reaching an overall accuracy of 84.70% after five stages of Continuous Learning, an increase of 1.3 percentage points compared with the mean accuracy for the standard resolutions.

Rehearsal-based methods, such as our proposed HSR strategy, have been found to be more effective in addressing the issue of catastrophic forgetting as they actively store and replay examples from previous tasks, allowing models to retain previous knowledge. In contrast, regularization-based methods, such as EWC, primarily focus on constraining a model's parameters to prevent excessive changes but do not actively store previous knowledge. The Random Replay strategy, while being a rehearsal-based approach, still suffers from severe forgetting due to the random selection of examples that may not be representative of the previous tasks. Our HSR approach, on the other hand, utilises a Herding Selection algorithm, which selects a small representative sample of examples based on their similarity to the previous tasks, thus allowing the model to effectively retain previous knowledge and minimise forgetting during the learning process. Furthermore, the results of the incremental learning processes presented in Section 4.2.2 further support the effectiveness of our proposed method in addressing the issue of catastrophic forgetting. The gradual decline in performance observed in our proposed method, in contrast to the severe forgetting exhibited by other comparative methods, aligns with the process of memory learning in the human brain and helps to maintain a balance of plasticity and stability.

The results of our ablation experiments and comparison of different sample selection strategies demonstrate the effectiveness of our proposed method, which combines EWC, Replay, and Herding Selection to mitigate catastrophic forgetting in incremental learning tasks. Our study shows that a hybrid approach using multiple strategies is more effective in addressing the problem of catastrophic forgetting than relying on a single strategy alone. In particular, our proposed method improved on the performance of other studied methods by using a Herding Selection algorithm to select representative samples that effectively represent the entire sample. This led to significantly improved performance and slow forgetting during incremental learning, resulting in the highest final average incremental accuracy. Furthermore, the comparison of different sample selection strategies showed that HSR was a more effective method for selecting exemplars in incremental learning tasks, as it minimised the distance between the feature centres of the selected exemplars and the feature centres of the whole sample. The results of our study provide valuable insights into the design and structure of incremental learning models and demonstrate the potential of our proposed method to address the challenge of catastrophic forgetting in such models.

6. Conclusions

In this paper, we propose a novel approach to detect lung CT and X-ray images of different coronaviruses that cause major respiratory diseases, such as SARS, MERS, wild-type SARS-CoV-2, and the Omicron and Delta variants of SARS-CoV-2, using a Transformer-based deep learning model. We combined regularization-based and rehearsal-based methods to address the challenge of Continuous Learning. Our approach achieved impressive performance, with a joint training accuracy of 0.9534, an F1 score of over 92%, and an average accuracy of 83.40% in the Continuous Learning environment. The proposed approach is a promising solution to address the challenges posed by continuously mutating viruses. In future work, we plan to further investigate the segmentation and lesion detection tasks in the CT and X-ray imaging of coronavirus-infected lungs based on the Continuous Learning methods proposed in this paper.

Author Contributions: Conceptualization, B.W. and D.Z.; methodology, B.W. and Z.T.; software, B.W.; validation, Z.T.; resources, B.W.; writing—original draft preparation, B.W.; writing—review and editing, D.Z. and Z.T.; visualization, B.W.; supervision, D.Z.; project administration, B.W.; funding acquisition, D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Science and Technology Development Fund, Macao SAR, under the Macao Funding Scheme for Key R&D Projects, grant number 0025/2019/AKP.

Data Availability Statement: The datasets utilised during the present investigation are accessible on Kaggle, https://www.kaggle.com/datasets/mustai/continual-learning-of-covid19 (accessed on 23 December 2022).

Acknowledgments: This research was supported by the Research Centre for Intelligent Prediction and Warning System for Mass Epidemics at Macau University of Science and Technology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **2020**, *395*, 497–506. [CrossRef] [PubMed]
2. Worldometer. "COVID Live—Coronavirus Statistics—Worldometer". Available online: https://www.worldometers.info/coronavirus/ (accessed on 17 December 2022).
3. Fang, Y.; Zhang, H.; Xie, J.; Lin, M.; Ying, L.; Pang, P.; Ji, W. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. *Radiology* **2020**, *296*, E115–E117. [CrossRef] [PubMed]
4. Xie, X.; Zhong, Z.; Zhao, W.; Zheng, C.; Wang, F.; Liu, J. Chest CT for Typical Coronavirus Disease 2019 (COVID-19) Pneumonia: Relationship to Negative RT-PCR Testing. *Radiology* **2020**, *296*, E41–E45. [CrossRef] [PubMed]
5. Das, A. Adaptive UNet-based Lung Segmentation and Ensemble Learning with CNN-based Deep Features for Automated COVID-19 Diagnosis. *Multimed. Tools Appl.* **2022**, *81*, 5407–5441. [CrossRef] [PubMed]
6. Park, S.; Kim, G.; Oh, Y.; Seo, J.B.; Lee, S.M.; Kim, J.H.; Moon, S.; Lim, J.K.; Ye, J.C. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. *Med. Image Anal.* **2022**, *75*, 102299. [CrossRef]
7. Guan, C.S.; Lv, Z.B.; Yan, S.; Du, Y.N.; Chen, H.; Wei, L.G.; Xie, R.M.; Chen, B.D. Imaging Features of Coronavirus disease 2019 (COVID-19): Evaluation on Thin-Section CT. *Acad. Radiol.* **2020**, *27*, 609–613. [CrossRef]
8. Shorfuzzaman, M.; Masud, M.; Alhumyani, H.; Anand, D.; Singh, A. Artificial Neural Network-Based Deep Learning Model for COVID-19 Patient Detection Using X-Ray Chest Images. *J. Healthc. Eng.* **2021**, *2021*, 100340. [CrossRef]
9. Yang, D.; Martinez, C.; Visuña, L.; Khandhar, H.; Bhatt, C.; Carretero, J. Detection and analysis of COVID-19 in medical images using deep learning techniques. *Sci. Rep.* **2021**, *11*, 19638. [CrossRef]
10. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *arXiv* **2017**, arXiv:1706.03762.
11. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
12. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
13. Mondal, A.K.; Bhattacharjee, A.; Singla, P.; Prathosh, A.P. xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography. *IEEE J. Transl. Eng. Health Med.* **2022**, *10*, 1100110. [CrossRef]
14. Shome, D.; Kar, T.; Mohanty, S.N.; Tiwari, P.; Muhammad, K. COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11086. [CrossRef]

15. Zhang, C.; Zhang, M.; Zhang, S.; Jin, D.; Zhou, Q.; Cai, Z.; Zhao, H.; Liu, X.; Liu, Z. Delving Deep Into the Generalization of Vision Transformers Under Distribution Shifts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 14 June 2021; pp. 7277–7286.
16. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 10–17 October 2021; pp. 9992–10002.
17. Chen CF, R.; Fan, Q.; Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 27 March 2021; pp. 347–356.
18. Ukwuoma, C.C.; Qin, Z.; Heyat, M.B.B.; Akhtar, F.; Smahi, A.; Jackson, J.K. Automated Lung-Related Pneumonia and COVID-19 Detection Based on Novel Feature Extraction Framework and Vision Transformer Approaches Using Chest X-ray Images. *Bioengineering* **2022**, *9*, 709. [CrossRef]
19. Hadsell, R.; Rao, D.; Rusu, A.A.; Pascanu, R. Embracing Change: Continual Learning in Deep Neural Networks. *Trends Cogn. Sci.* **2020**, *24*, 1028–1040. [CrossRef]
20. McCloskey, M.; Cohen, N.J. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*; Bower, G.H., Ed.; Academic Press: Cambridge, MA, USA, 1989; Volume 24, pp. 109–165.
21. Abraham, W.C.; Robins, A. Memory retention–the synaptic stability versus plasticity dilemma. *Trends Neurosci.* **2005**, *28*, 73–78. [CrossRef]
22. Parisi, G.I.; Kemker, R.; Part, J.L.; Kanan, C.; Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Netw.* **2019**, *113*, 54–71. [CrossRef]
23. Lesort, T.; Lomonaco, V.; Stoian, A.; Maltoni, D.; Filliat, D.; Díaz-Rodríguez, N. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Inf. Fusion* **2020**, *58*, 52–68. [CrossRef]
24. Prabhu, A.; Torr, P.H.S.; Dokania, P.K. GDumb: A Simple Approach that Questions Our Progress in Continual Learning. In Proceedings of the Computer Vision—ECCV 2020, Cham, Switzerland, 23–28 August 2020; pp. 524–540.
25. Rebuffi, S.; Kolesnikov, A.; Sperl, G.; Lampert, C.H. iCaRL: Incremental Classifier and Representation Learning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5533–5542.
26. Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; Tuytelaars, T. Memory Aware Synapses: Learning What (not) to Forget. In Proceedings of the Computer Vision—ECCV 2018, Cham, Switzerland, 5 October 2018; pp. 144–161.
27. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef]
28. Ergün, E.; Töreyin, B.U. Sparse Progressive Neural Networks for Continual Learning. In Proceedings of the International Conference on Computational Collective Intelligence, Cham, Switzerland, 29 September–1 October 2021; pp. 715–725.
29. Li, Z.; Hoiem, D. Learning without Forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2935–2947. [CrossRef]
30. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
33. Szegedy, C.; Wei, L.; Yangqing, J.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
34. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
35. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
36. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
37. Brunese, L.; Mercaldo, F.; Reginelli, A.; Santone, A. Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. *Comput. Methods Programs Biomed.* **2020**, *196*, 105608. [CrossRef]
38. Narin, A.; Kaya, C.; Pamuk, Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Anal. Appl.* **2021**, *24*, 1207–1220. [CrossRef]
39. Wang, G.; Liu, X.; Shen, J.; Wang, C.; Li, Z.; Ye, L.; Wu, X.; Chen, T. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nat. Biomed. Eng.* **2021**, *5*, 509–521. [CrossRef]
40. Shamila Ebenezer, A.; Deepa Kanmani, S.; Sivakumar, M.; Jeba Priya, S. Effect of image transformation on EfficientNet model for COVID-19 CT image classification. *Mater. Today Proc.* **2022**, *51*, 2512–2519. [CrossRef]
41. Toraman, S.; Alakus, T.B.; Turkoglu, I. Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks. *Chaos Solitons Fractals* **2020**, *140*, 110122. [CrossRef]
42. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic Routing Between Capsules. *arXiv* **2017**, arXiv:1710.09829.

43. Gunraj, H.; Wang, L.; Wong, A. COVIDNet-CT: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases From Chest CT Images. *Front. Med.* **2020**, *7*, 608525. [CrossRef]
44. Ghaderzadeh, M.; Eshraghi, M.A.; Asadi, F.; Hosseini, A.; Jafari, R.; Bashash, D.; Abolghasemi, H. Efficient Framework for Detection of COVID-19 Omicron and Delta Variants Based on Two Intelligent Phases of CNN Models. *Comput. Math. Methods Med.* **2022**, *2022*, 4838009. [CrossRef]
45. Tahir, A.M.; Qiblawey, Y.; Khandakar, A.; Rahman, T.; Khurshid, U.; Musharavati, F.; Islam, M.T.; Kiranyaz, S.; Al-Maadeed, S.; Chowdhury, M.E.H. Deep Learning for Reliable Classification of COVID-19, MERS, and SARS from Chest X-ray Images. *Cogn. Comput.* **2022**, *14*, 1752–1772. [CrossRef]
46. Anas, M.; Tahir, M.E.H.C.; Qiblawey, Y.; Khandakar, A.; Rahman, T.; Kiranyaz, S.; Khurshid, U.; Ibtehaz, N.; Mahmud, S.; Ezeddin, M. *COVID-QU-Ex*; Kaggle: San Francisco, CA, USA, 2021. [CrossRef]
47. Tahir, A.M.; Chowdhury, M.E.H.; Khandakar, A.; Rahman, T.; Qiblawey, Y.; Khurshid, U.; Kiranyaz, S.; Ibtehaz, N.; Rahman, M.S.; Al-Maadeed, S.; et al. COVID-19 infection localization and severity grading from chest X-ray images. *Comput. Biol. Med.* **2021**, *139*, 105002. [CrossRef]
48. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **2018**, *172*, 1122–1131.e1129. [CrossRef]
49. Merlin, G.; Lomonaco, V.; Cossu, A.; Carta, A.; Bacciu, D. Practical Recommendations for Replay-Based Continual Learning Methods. In Proceedings of the Image Analysis and Processing, ICIAP 2022 Workshops, Cham, Switzerland, 23–27 May 2022; pp. 548–559.
50. Robins, A.V. Catastrophic Forgetting, Rehearsal and Pseudorehearsal. *Connect. Sci.* **1995**, *7*, 123–146. [CrossRef]
51. Zhou, Y.; Zhang, S.; Sun, X.; Ma, F.; Zhang, F. SAR Target Incremental Recognition Based on Hybrid Loss Function and Class-Bias Correction. *Appl. Sci.* **2022**, *12*, 1279. [CrossRef]
52. Lomonaco, V.; Pellegrini, L.; Cossu, A.; Carta, A.; Graffieti, G.; Hayes, T.L.; Lange, M.D.; Masana, M.; Pomponi, J.; Ven, G.M.v.d.; et al. Avalanche: An End-to-End Library for Continual Learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 3595–3605.
53. Grossberg, S. A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action. *Front. Neurorobot.* **2020**, *14*, 36. [CrossRef]
54. Amalapuram, S.K.; Tadwai, A.; Vinta, R.; Channappayya, S.S.; Tamma, B.R. Continual Learning for Anomaly based Network Intrusion Detection. In Proceedings of the 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 4–8 January 2022; pp. 497–505.
55. Biesialska, M.; Biesialska, K.; Costa-jussà, M.R. Continual Lifelong Learning in Natural Language Processing: A Survey. *arXiv* **2020**, arXiv:2012.09823.
56. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Keysers, D.; Uszkoreit, J.; Lucic, M.; et al. MLP-Mixer: An all-MLP Architecture for Vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
57. Peng, Z.; Dong, L.; Bao, H.; Ye, Q.; Wei, F. BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. *arXiv* **2022**, arXiv:2208.06366.
58. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
59. David Lopez-Paz, M.A.R. Gradient Episodic Memory for Continual Learning. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1–17.
60. Chaudhry, A.; Ranzato, M.A.; Rohrbach, M.; Elhoseiny, M. Efficient Lifelong Learning with A-GEM. *arXiv* **2018**, arXiv:1812.00420.
61. Lomonaco, V.; Maltoni, D.; Pellegrini, L. Rehearsal-Free Continual Learning over Small Non-I.I.D. Batches. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 989–998.
62. Van der Maaten, L.; Hinton, G. Viualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.