Neural Computing and Applications (2024) 36:18841–18862
https://doi.org/10.1007/s00521-024-10182-6

ORIGINAL ARTICLE

Deep multi-metric training: the need of multi-metric curve evaluation to avoid weak learning

Michail Mamalakis[1,2,3] · Abhirup Banerjee[8,9] · Surajit Ray[4] · Craig Wilkie[4] · Richard H. Clayton[2,3] · Andrew J. Swift[3,5] · George Panoutsos[6] · Bart Vorselaars[7]

Received: 28 August 2023 / Accepted: 1 July 2024 / Published online: 1 August 2024

Abstract

The development and application of artificial intelligence-based computer vision systems in medicine, environment, and industry are playing an increasingly prominent role. Hence, the need for optimal and efficient hyperparameter tuning strategies is more than crucial to deliver the highest performance of the deep learning networks in large and demanding datasets. In our study, we have developed and evaluated a new training methodology named deep multi-metric training (DMMT) for enhanced training performance. The DMMT delivers a *state of robust learning* for deep networks using a new important criterion of multi-metric performance evaluation. We have tested the DMMT methodology in multi-class (three, four, and ten), multi-vendors (different X-ray imaging devices), and multi-size (large, medium, and small) datasets. The validity of the DMMT methodology has been tested in three different classification problems: (i) medical disease classification, (ii) environmental classification, and (iii) ecological classification. For disease classification, we have used two large COVID-19 chest X-rays datasets, namely the BIMCV COVID-19+ and Sheffield hospital datasets. The environmental application is related to the classification of weather images in cloudy, rainy, shine or sunrise conditions. The ecological classification task involves a classification of three animal species (cat, dog, wild) and a classification of ten animals and transportation vehicles categories (CIFAR-10). We have used state-of-the-art networks of DenseNet-121, ResNet-50, VGG-16, VGG-19, and DenResCov-19 (DenRes-131) to verify that our novel methodology is applicable in a variety of different deep learning networks. To the best of our knowledge, this is the first work that proposes a training methodology to deliver *robust learning*, over a variety of deep learning networks and multi-field classification problems.

Keywords DenRes-131 · Classification · Chest X-rays · COVID-19 · Robust learning · Weak learning

1 Introduction

The development of AI-based medical systems, as well as their translation to medical practice, is playing an increasingly prominent role in the treatment and therapy of patients [13, 28]. Along with the automated methods that rely on blood test results or biomarkers for diagnosis [2, 3, 22, 35, 38], an increasing number of deep learning-based methods, specifically the convolution neural network (CNN)-based models [7, 14, 24, 29, 32], are being implemented and used to develop accurate, robust, and fast detection techniques to fight against COVID-19 and other respiratory diseases. In the environmental and industrial domains, there are studies that explore the utilisation of deep neural networks (DNNs) to approximate solutions for partial differential equations (PDEs) in computational mechanics, emphasising the energetic format of PDEs and demonstrating their efficacy in various engineering applications [36]. Furthermore, there are studies highlighting the use of CNNs and artificial intelligence in geoscientific, meteorology, and climate science applications [26, 27].

As the prevalence of deep learning applications continues to grow exponentially in medical, environmental, and industrial domains, the imperative for effective hyperparameter tuning strategies becomes crucial. Ensuring optimal performance of networks on large datasets, while concurrently managing training times, is essential for advancing the capabilities of these applications.

A widely used method for training neural networks is to apply a loss early stopping (LES) criterion and a maximum

Abhirup Banerjee and Bart Vorselaars have contributed equally to this work.

Extended author information available on the last page of the article

number of epochs for training [25, 31, 34, 46]. Typically, the dataset is divided into a training set, a validation set, and a test set. During training, it is common to observe that the validation set reaches a local (or even global) minimum of the network's loss function, indicating that further training may lead to overfitting. To prevent this, a criterion is applied to monitor the loss function for the validation set during training, with the user specifying the maximum number of epochs for training and the number of permitted epochs to continue without a change in the minimum loss value. Once either of these conditions is met, training is terminated, and the network's weights that lead to the minimum loss value for the validation set are used [34].

However, we argue in this paper that the LES approach may not always be the optimal solution. While the loss function can reach a minimum for the validation set, other evaluation metrics may continue to improve. For example, in certain medical applications it may be important to achieve a sensitivity threshold above a certain value, meaning that positive patients are correctly identified as positive. In practice, there are even requirements on multiple metrics, such as on both sensitivity and specificity [30]. In these cases, assuming overfitting based solely on a loss function may not be appropriate. Thus, we propose evaluating multiple metrics during training and advocate for the benefits of training for a longer duration. We develop a new method, called deep multi-metric training (DMMT) that utilises heuristics to automate the evaluation of multiple metrics. Our approach aims to optimise multiple criterion separately, rather than using a single loss function or aggregating multiple loss functions for optimisation. In case of combining multiple loss functions into a single function, changes in one component can interact with changes in another, leading to a stabilising effect on the overall criterion. Consequently, an aggregated criterion may exhibit early stopping behaviour, which should be mitigated if the loss functions are evaluated separately. Therefore, evaluating multiple metrics independently during training can yield more accurate and robust models. To facilitate this study, we introduce new terminology summarised in Table 1.

The proposed methodology introduces a new important criterion of multi-metric performance evaluation to deliver *robust learning* for a network in a dataset. Our methodology involves evaluating network performance using a protocol that incorporates both independent identical distribution (i.i.d.) cohorts and out-of-distribution (o.o.d.) cohorts. In medical applications, this evaluation protocol is crucial as it tests the network's ability to generalise and remain robust across different datasets. Our ultimate objective is to create a training methodology that delivers a reliable and robust AI network, capable of consistently providing precise results across a range of imaging scenarios (medical environmental, ecological, etc.). To achieve this, we propose testing the established training methodology, which employs the LES approach, alongside our own approach. To evaluate our methodology, we test it in a classification problem on four different kinds of image datasets (COVID chest X-rays from two different datasets, weather data, and animal species). Furthermore, to show the robustness, we apply five state-of-the-art deep learning networks, namely DenseNet-121 [17], ResNet-50 [15], VGG-16, VGG-19 [37], and DenResCov-19 [29]. The DenResCov-19 has consistently superior performance in all applications as compared to the other networks, and hence, to generalise its application we rename it to **DenRes-131**. Here, the '131' represents the total number of layers in the model.

To the best of our knowledge, this is the first development and utilisation of a deep multi-metric training methodology in a variety of different state-of-the-art deep learning networks. To this end, the main contributions of this study are:

1. Justifying the importance of multi-metric (AUC-ROC, recall, precision, F1, etc.) utilisation to achieve *robust learning* and avoid *state of weak learning* in deep learning networks;
2. Evaluating the performance and robustness of established deep learning networks over heterogeneous

Table 1 Terminology introduced in the current study

Concept	Definition
State of weak learning	When one or more metrics during training of the network are still in a transitional state of training, i.e. not fluctuating around a constant value for a specific number of epochs, based on monitoring each metric on the validation set
State of robust learning	When all metrics during training of the network are in a converged state of training, i.e. with small fluctuations around a constant value for a specific number of epochs, based on monitoring each metric on the validation set

medical imaging, environmental, and ecological datasets with multi-class labels;
3. Comparing the new DenRes-131 network with the DenseNet-121, ResNet-50, VGG-16, and VGG-19 established networks in multi-field, multi-size, multi-vendors, and multi-class validation schemes in both independent identical distribution (i.i.d.) cohorts and out-of-distribution (o.o.d.) cohorts; and
4. Finally, a proposed methodology that exhibits superior performance compared to the established training methodology that employs the LES criterion.

The rest of the paper is organised as follows: Sect. 2 presents a brief overview of the related works. Section 3 describes the proposed methodology and summarises its implementation, along with a brief description of the imaging datasets. Numerical results of the performance of proposed methodology are presented in Sect. 4, and a detailed discussion is provided in Sect. 5. The paper concludes in Sect. 6.

2 Related work

There are two main hyperparameter optimisation approaches: manual (e.g. grid search, random search) and automatic (e.g. Bayesian optimisation). More recently in the literature, new automatic strategies and approaches for optimal searching are developed.

[47] describe the Orthogonal Array Tuning Method and evaluate it by using recurrent neural networks and CNNs. Their method decreases the tuning time compared to previous state-of-the-art methods and delivers high performance of the results.

[20] describe a method utilising genetic programming to deliver both optimal activation functions and optimisation techniques. To evaluate their method, they implemented a neural network with the activation function and an optimisation technique that the algorithm chooses per iteration. Their method performed superior compared to conventional methods.

[49] determine a hyperparameter selection process with high diversity, investigating the optimal joint hyperparameter configuration on network structure and training to evaluate road image classification tasks. They showed that their approach can deliver an optimal architecture with an associated training configuration, to deliver a consistent and accurate performance of the network.

[10] propose a hyperparameter optimisation method, which searches for optimal hyperparameters based on an initial sequence and utilises an action-prediction network leveraged on continuous deep Q-learning. They evaluated their algorithm on different benchmarks, presenting its superior performance.

[39] introduce the application of the fractal decomposition-based algorithm to the optimisation of the hyperparameter of deep neural network architecture, in order to deliver state-of-the-art results.

[40] discuss empirical comparisons of the optimisers. Their investigation revealed that incorporating relationships between optimisers is crucial in practical scenarios, especially in adaptive gradient methods. Through their work, they raised some concerns about fairly benchmarking the optimisers for neural network training.

It is important to mention here that some of the studies discussed the importance of hyperparameter tuning in fine-tuning and not just during the training process [23, 39, 40].

New trends regarding optimisation approaches are the automated machine learning (AutoML) [11, 16, 44] and the no-new-UNet (nn-Unet) [18]. Both of these methodologies try to deliver the optimal accuracy solution in more than one step of deep learning training, such as pre-processing, post-processing, hyperparameters, and identification of the optimal structure. As COVID-19 has become an important area of research in the last years, there have been some attempts to apply hyperparameter strategies in COVID-19 classification and detection benchmarks [1, 4, 19, 41, 42]. These studies generally focus on efficient ways of searching the optimal values of hyperparameters.

[45] and [5] propose deep multi-metric learning methods, utilising cost functions involving multi-metric scores. The disadvantage of these studies is that they used only the cost function minimisation approach to determine the optimal solution.

On the contrary, here we advocate the involvement of more than one evaluation metric (multi-metric) score during the training process, in order to consider them separately, and a different total cost function minimisation criterion. To this end, the optimisation criterion of hyperparameters takes into consideration the performance of the network in terms of important evaluation metrics (AUC-ROC, recall, precision, and F1-score, as will be introduced later) depending on the computer vision application problem. As a result, the optimisation approach of the hyperparameter values, namely learning rate, epochs, batch number, patch number, etc., delivers *robust learning* results for the network. For our classification tasks, we have chosen the AUC-ROC, recall, precision, and F1-score evaluation metrics, due to their wide usage in the literature.

To the best of our knowledge, this study is the first to deliver the development and evaluation of a new training methodology combining multiple quantitative metrics and a cost function minimisation criterion.

3 Methods

In this section, we present the algorithm and associated implementation details of the proposed DMMT method. Furthermore, a description of the network architectures that we use to evaluate the training methodology is presented.

Algorithm 1 The deep multi-metric training

Initialisation:
$N \leftarrow 4$,
$\text{loss}^{\text{prev}} \leftarrow \infty$
for k in $\{1, \ldots, N\}$ **do**
  $\overline{M}_k^{\text{prev}} \leftarrow \text{SMA}_1^0(M_k)$
end for

Start training:
for t in $\{1, 2, \ldots, t_{\max}\}$ **do**
  comment: Check for convergence
  $N_s \leftarrow 0$
  for k in $\{1, \ldots, N\}$ **do**
    if $M_k \in [\overline{M}_k^{\text{prev}} - \Delta_k, \overline{M}_k^{\text{prev}} + \Delta_k]$ **then**
      $N_s \leftarrow N_s + 1$
    end if
  end for
  if $(N_s = N)$ & $(\text{loss}^t \geq \text{loss}^{\text{prev}})$ **then**
    break comment: end training
  end if

  comment: Update variables
  if $t \bmod \Delta t = 0$ **then**
    $\text{loss}^{\text{prev}} \leftarrow \text{loss}^t$
    for k in $\{1, \ldots, N\}$ **do**
      $\overline{M}_k^{\text{prev}} \leftarrow \text{SMA}_{\Delta t}^t(M_k)$
    end for
  end if
end for
save weights
End training

3.1 DMMT methodology

To explain the idea of DMMT methodology, we present the parameter and variable definitions in Table 2 and the algorithm in Algorithm 1. The DMMT algorithm requires choosing the N multiple metrics of the training $M_1, M_2, \ldots, M_N$, the epoch checkpoint interval Δt, the maximum epochs for training $t_{\max}$, the acceptable variation of the moving average metric value to define as equal Δ_k, and the loss cost function value at the t^{th} epoch, loss^t.

Algorithm 1 presents the novel mathematical approach of the DMMT methodology for training the deep learning networks based on the multi-metric criterion. We utilise a combination of the cost function minimisation and multi-metric curve evaluation criterion. The training procedure initialises the model with random weights or transfer weights. The user sets the number of multi-metric evaluation scores, the epochs period Δt where the algorithm will check the convergence of the multi-metrics, and the maximum number of epochs for training. The convergence of the multi-metrics is achieved when the score of the metric is within Δ_k variation as defined by the user. The end of the training is achieved when either the algorithm reaches the maximum number of epochs or when all the multi-metrics converge and the loss value, loss^t, is higher than or equal to the previously stored value, $\text{loss}^{\text{prev}}$ (local minimum).

In Fig. 1, the second row and last column illustrate all the criterion employed in the DMMT methodology. We can observe that the loss function has been optimised, and all four evaluation metrics—M_1, M_2, M_3, and M_4—have stabilised. The results shown in Fig. 1 correspond to the performance of the Resnet-50 network in the weather evaluation dataset, as determined by the converged multi-metrics criterion and the loss function within the DMMT (green line). It is crucial to note that these outcomes differ from those obtained using the LES criterion (red line).

The metrics are often prone to large statistical fluctuations. To dampen these, we use an averaging procedure based on the simple moving average (SMA). For a quantity A, the SMA is defined as

$$\text{SMA}_n^t(A) = \frac{1}{n}\sum_{i=0}^{n-1} A_{t-i} \tag{1}$$

where A_t is the value of the quantity at epoch t and n is the number of instances averaged. We define the metrics recall, precision, and F1-score as

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP is the true positive results, TN is the true negative results, FP is the false positive results, and FN is the false negative results. We also define AUC-ROC as area under the receiver operating characteristic (ROC) curve that combines TP, TN, FP, and FN. In order to discretise the ROC curve, a set of thresholds evenly distributed along a linear scale is employed to determine pairs of recall and

Table 2 Hyperparameters, variables, and functions of the DMMT algorithm

User's hyperparameters:	• The optimisation algorithm of the loss cost function (Adam, SGD, etc.)
	• N: number of available metrics. The user also has to define each of the N metrics
	• Δt: stride of computing averages
	• Δ_k: acceptable variation of the moving average for each metric $k \in \{1, \ldots, N\}$
	• $t_{\max}$: maximum number of epochs for training
Algorithm variables	• N_s: number of stable metrics
	• t: current epoch
	• T: check point number of current epoch, $T = t/\Delta t$
	• k: current metric index
	• M_k: value of metric k
	• $\overline{M}_k$: value of the moving average of metric k
	• loss^t: loss cost function at epoch t

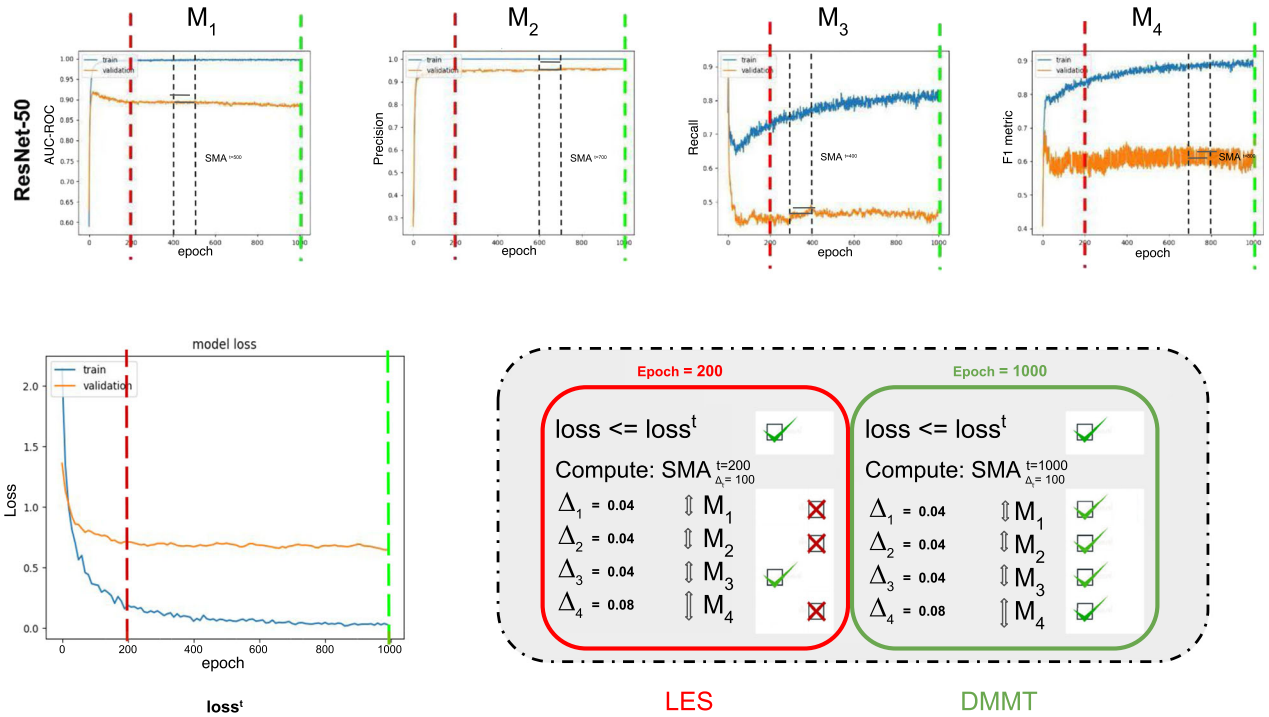


Fig. 1 Proposed DMMT methodology illustrated on the weather data: The user determines the number N and choice of multi-metric curve evaluation scores (here $N = 4$, with M_1: AUC-ROC, M_2: recall, M_3: precision, and M_4: F1-score), the epoch interval where the algorithm will check the convergence of the multi-metrics (here every 100 epochs), and the maximum number of epochs for training (here 1500). The function $\text{SMA}^t_{\Delta t}(M_k)$ is used to compute the simple moving average (SMA) between two checkpoints t and $t - \Delta t$ for each of the metrics M_k. A few sample SMAs are indicated in the graphs. The convergence of the multi-metrics is achieved when the score of the metric is within Δ_k variation of this average, as defined by the user (here $\Delta_1 = \Delta_2 = \Delta_3 = 0.04$ and $\Delta_4 = 0.08$). The end of training is achieved when either the algorithm reaches the maximum number of epochs or when all the multi-metrics converge and the loss value (loss^t) is higher than the previous stored (loss$^{\text{prev}}$). The red lines at 200 epochs are the result of the traditional technique of loss early stopping. The green lines at 1000 epochs are the result of the proposed DMMT algorithm

precision values. The height of the recall is multiplied by the FP to measure the final AUC-ROC metric. Equation (1) is used to compute the moving average of each of the metrics in Eqs. (2)–(4) and the AUC-ROC.

3.2 Network architectures

To test the DMMT methodology, we use four established networks, namely VGG-16, VGG-19, DenseNet-121, and ResNet-50, and a state-of-the-art deep learning model DenRes-131.

VGG-16 and VGG-19 are two well-established convolutional neural networks (CNNs) with a combination of pooling and convolution layers [37]. ResNet-50 is a deep network, in which all layers have the same number of filters as the number of the output feature size. In case the output feature size is halved, the number of filters is doubled, thus reducing the time complexity per layer [15]. DenseNet-121 is an efficient topology of convolutional network. The network comprises of deep layers, each of which implements a nonlinear transformation. [17] introduced a unique connectivity pattern information flow between layers to direct connecting any layer to all subsequent layers.

DenRes-131 network [29] is a concatenation of four blocks from ResNet-50 and DenseNet-121 with width, height, and frames of $58 \times 58 \times 256$, $28 \times 28 \times 512$, $14 \times 14 \times 1024$, and $7 \times 7 \times 2048$, respectively. Each of the four outputs feeds a block of convolution and average pooling layers. Thus, the initial concatenated information can be translated into the convolution space. [29] used some level of concatenation-CNN block techniques to create kernels that deliver a final layer of soft-max regression, so that the network can conclude in the classification decision.

3.3 Datasets

We evaluate our methodology on five different image datasets. We use two large datasets of COVID-19 and abnormal lung screening, two large datasets of animal species classification, and one relatively small dataset of weather classification. The evaluation tasks are: three-class classification (normal, abnormal, or COVID-19 in medical imaging dataset; cat, dog, or wild in ecological dataset), four-class classification (cloudy, rainy, shine, or sunrise in environmental dataset), and ten-class classification tasks (CIFAR-10 dataset, second ecological dataset).

The first dataset, which we refer to as BIMCV, is generated by combining the BIMCV COVID-19+ [8] and the BIMCV-COVID19-PADCHEST data [4] for medical imaging application. BIMCV COVID-19+ contains the normal and COVID-19 cases, while BIMCV-COVID19-PADCHEST contains the abnormal cases, which is a reorganisation of the PadChest dataset [4] related to COVID-19 pathology. In total, we use 4740 lung X-ray images classified as abnormal, 4456 as normal, and 2646 as COVID-19 positive.

For the second medical imaging dataset, named Sheffield hospital, we use a Sheffield hospital COVID-19 dataset of lung X-ray images. Here, we use 2011 chest X-ray images classified as abnormal, 2861 as normal, and 2263 images as COVID-19 positive.

The third dataset, concerning animals species, is a large collection of 16, 122 publicly available images for the three-class species classification into cats, dogs, and wild animals [6]. The dataset is a collection of 5153, 4731, and 4738 images of cats, dogs, and wild animals, respectively.

The fourth one, called the multi-class weather dataset, is a collection of images for environmental classification [12]. It consists of 357 sunrise, 253 shine, 215 rainy, and 300 cloudy images.

For the evaluation of the three classification tasks, we first split the total images into 70% and 30% as the training and testing datasets, respectively. The training dataset is further split into 70%:30% as the final training and validation datasets. As we need to evaluate the generalisation of our training algorithm, we test the deep learning networks in an identical independent distribution (i.i.d.) cohort of a collection of 500 images from each of cats, dogs, and wild animals (excluded before the splitting) and in an out of the distribution (o.o.d.) cohort by training on the BIMCV dataset and testing on the Sheffield hospital dataset. In this way, we verify that the DMMT can achieve highly accurate and robust results compared to the traditional LES criterion training technique [34].

We conduct a sensitivity analysis of the LES 'patience' (early stop criterion) hyperparameter and the DMMT Δ_t hyperparameter, using the publicly available CIFAR-10 dataset (https://paperswithcode.com/dataset/cifar-10). The CIFAR-10 dataset is a subset of the Tiny Images dataset and comprises 60,000 32×32 colour images. Each image is labelled with one of 10 mutually exclusive classes, including aeroplane, automobile (excluding trucks or pickup trucks), bird, cat, deer, dog, frog, horse, ship, and truck (excluding pickup trucks). The dataset is structured with 6,000 images per class, split into 5,000 training images and 1,000 testing images per class.

3.4 Datasets pre-processing image analysis

Image analysis techniques are applied to all slices to reduce the effect of noise and increase the signal-to-noise ratio (SNR). We use noise filters such as binomial deconvolution, Landweber deconvolution [43], and curvature anisotropic diffusion image filters [33] to reduce noise in the images. We normalise the images by subtracting the mean value from each image and dividing by its standard deviation. Finally, we use data augmentation techniques including rotation (around the centre of the image by a random angle from the range $[-15°, 15°]$), width shift (up to 20 pixels), height shift (up to 20 pixels), and ZCA whitening (add noise in each image) [21].

3.5 Hyperparameters initialisation

After random shuffling, each dataset is partitioned for training, validation, and testing of the models. We use the categorical cross-entropy as the loss function. The loss function is optimised using the stochastic gradient descent (SGD) method with a fixed learning rate of 0.001 for both the LES and DMMT methodologies. We apply transfer learning techniques to the networks using the ImageNet dataset [9] (https://www.image-net.org). The ImageNet dataset consists of over 14 million images, and the task is to classify the images into one of almost 22, 000 different categories (cat, sailboat, etc.).

Table 2 summarises the main user's hyperparameters. We want to establish the efficiency of the algorithm for different hyperparameters to validate its robustness. To do so we vary the number of available metrics N from 3 to 4. Moreover, we use different values of Δ_k and epoch checkpoint T for each of the classification tasks. The parameters in the DMMT algorithm are taken to be $\Delta t = 10$ and $\Delta_k = 0.04$ for the considered metrics in the medical image datasets (recall, precision, and AUC-ROC). For the ecological and environmental datasets, the parameters are chosen as $\Delta t = 100$, $\Delta_1 = \Delta_2 = \Delta_3 = 0.04$, and $\Delta_4 = 0.08$ (Fig. 1) for the considered metrics recall, precision, AUC-ROC, and F1-score, respectively. The reason for the usage of $\Delta_4 = 0.08$ for the fourth metric (F1-score, Fig. 1) is that the F1-score metric produces large fluctuations and therefore, the DMMT does not converge earlier than the maximum epochs ($t_{\max}$) within the narrow window of $\Delta_4 = 0.04$. For LES, we use an early stopping of 10 continuous epochs ('patience'). For both methodologies, the maximum epochs for training $t_{\max}$ are 1500 for all datasets.

For the sensitivity analysis of the LES 'patience' and the DMMT Δ_t hyperparameters, we vary them over the values of 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, and 100 epochs, using the ResNet-50 network architecture.

3.6 Software

The code developed in this study is written in the Python programming language using Keras/TensorFlow (Python) libraries. For training and testing of deep learning networks, we use an NVIDIA cluster (JADE2) with 4 GPUs and 64 GB RAM memory.

4 Results

In this section, we examine the performance of the networks for the traditional LES criterion and proposed DMMT methodology. We present the performance of the established networks of VGG-16, VGG-19, ResNet-50, and DenseNet-121 and the new state-of-the-art network DenRes-131 [29].

4.1 Evaluate DMMT in multi-field classification

To generalise the applicability of the DMMT, we first need to verify the importance of quantitative multi-metric evaluation in different computer vision applications as compared to the commonly used LES criterion [34]. To this end, we compare both methodologies in multi-field classification problems, namely on (1) medical imaging, (2) environmental, and (3) ecological datasets.

4.1.1 Medical imaging computer vision task: chest X-rays classification

We first evaluate the recall, precision, and AUC-ROC metrics for the networks and test the stability of the training in these metrics (equilibrium point of a metric training/testing curve) on the medical imaging datasets, so that we can justify a weak or robust level of training performance (*state of weak learning* and *state of robust learning*).

Table 3 highlights the quantitative evaluation metrics on the test datasets of the BIMCV and Sheffield hospital datasets. Both VGG-16 and VGG-19 networks follow a specific pattern of high variability of the metric values (from 57.17 to 97.26%) with some high and some low values for the LES criterion. For the DMMT, this variability is smoothed, and the networks appear to converge for all evaluation metrics, with a small deviation of $\pm 5\%$. ResNet-50, DenseNet-121, and DenRes-131 follow a different pattern of performance compared to the previous two networks, with low values and low dispersion between the metrics during the LES, which increase significantly for the DMMT. Figures 2 and 3 present the behaviours of the training and validation curves for the recall, precision, and AUC-ROC metrics in BIMCV and Sheffield hospital datasets, respectively.

Based on the AUC-ROC metric alone, the network models for the Sheffield hospital dataset (Fig. 3) seem to have virtually converged after LES (as shown with the red dashed lines). However, for the precision and recall metrics the models are still in a transitional state of training (*state of weak learning*). Nevertheless, a converged state of the models is achieved by DMMT in all three metrics (green dashed line). The same pattern is observed in Fig. 2 for the BIMCV dataset. The number of epochs in which all metrics are in equilibrium (here in 800) determines the *state of robust learning*. Figures 4 and 5 illustrate the ROC curves of the deep learning networks on BIMCV dataset and Sheffield hospital dataset, respectively.

Table 3 Quantitative evaluation metrics of different networks on test datasets for medical image classification task

Classification performance on BIMCV dataset: COVID-19, abnormal, or normal					
Metric (%)	VGG-16 LES/DMMT	VGG-19 LES/DMMT	ResNet-50 LES/DMMT	DenseNet-121 LES/DMMT	DenRes-131 LES/DMMT
Recall	57.17/70.70	64.75/70.00	49.59/70.65	70.77/71.07	70.73/75.10
Precision	97.26/75.05	88.07/74.16	49.59/75.11	74.82/75.80	70.73/76.21
AUC-ROC sample	72.86/79.07	78.01/78.47	62.19/79.07	79.05/79.45	78.05/82.00
AUC-ROC macro	72.12/78.32	77.35/77.87	62.65/78.20	78.31/78.52	79.80/81.30
AUC-ROC micro	68.91/77.94	76.05/77.94	62.19/77.91	77.98/78.22	78.05/81.00
AUC-ROC weighted	68.81/76.69	74.67/76.14	60.51/76.73	77.14/77.03	76.56/80.00
F1 sample	70.46/72.15	72.52/71.34	49.59/72.14	72.13/72.65	70.73/76.16
F1 macro	72.12/73.86	73.75/73.24	48.91/73.71	73.40/74.23	73.31/77.50
F1 micro	68.82/77.94	71.30/71.10	49.59/71.85	77.98/72.32	70.73/75.71
F1 weighted	68.02/71.93	74.67/71.15	49.55/71.97	77.13/72.44	70.20/75.85
Classification performance on Sheffield hospital dataset: COVID-19, abnormal, or normal					
Metric (%)	VGG-16 LES/DMMT	VGG-19 LES/DMMT	ResNet-50 LES/DMMT	DenseNet-121 LES/DMMT	DenRes-131 LES/DMMT
Recall	58.91/64.69	43.38/66.09	53.73/55.17	65.43/64.97	64.39/67.01
Precision	81.82/65.83	92.46/65.59	53.73/55.17	65.85/65.43	64.39/67.32
AUC-ROC sample	69.35/73.68	66.61/74.69	65.30/66.38	74.17 /73.84	73.29/75.20
AUC-ROC macro	64.48/67.44	60.66/68.78	56.28/57.91	67.87/67.44	66.78/69.01
AUC-ROC micro	69.35/73.52	66.53/74.57	65.30/66.38	74.06/73.73	73.29/75.10
AUC-ROC weighted	67.52/69.68	64.41/70.99	58.60/59.27	70.74/70.09	64.71/72.03
F1 sample	65.95/64.80	58.66/66.26	53.73/55.17	65.57/65.13	64.39/67.03
F1 macro	52.04/53.91	49.81/54.28	37.97/38.86	52.70/52.60	53.36/55.11
F1 micro	63.26/64.81	58.07/66.26	53.73/55.17	52.76/65.13	64.39/66.67
F1 weighted	64.11/65.36	57.32/67.14	55.77/57.61	70.74/66.71	64.71/68.06

Values represent metrics after loss early stopping (LES)/DMMT criterion

To conclude, in this subsection we have justified the need to monitor more than one metric (recall, precision, and AUC-ROC) to determine the convergence of a network training in two medical image classification tasks.

4.1.2 Environmental computer vision task: weather classification

Table 4 shows the quantitative evaluation metrics of weather classification (cloudy, rainy, shine, or sunrise) for the LES and DMMT criterion. Figure 6 shows the behaviour of training and validation curves for the recall, precision, AUC-ROC, and F1 metrics. Even if precision and AUC-ROC metrics justify that the models converge at the LES (red dashed line), in the majority of the cases in the recall and F1 metrics the models are still in a transitional period of training (*state of weak learning*). However, the converge state of the models is achieved by DMMT for all metrics (green dashed line).

Figure 7 shows the confusion matrices of the environmental classification problem for the five networks using the LES and DMMT criterion. The DenRes-131 achieves recall of 76.7, 96.9, 90.8, and 92.5% during LES and 76.7, 98.4, 90.8, and 92.5% by DMMT, for the classification of cloudy, rainy, shine, and sunrise classes, respectively. DenseNet-121 achieves recall of 74.4, 92.2, 85.5, and 89.7% during LES and 75.6, 93.8, 86.8, and 88.8% during DMMT. ResNet-50 achieves recall of 72.2, 92.2, 80.3, and 86.9% at LES and 76.7, 95.3, 84.2, and 87.9% with DMMT. VGG-16 achieves recall of 72.2, 92.2, 85.5 and 90.7% by LES and 71.1, 89.1, 80.3, and 87.9% by DMMT. Finally, VGG-19 achieves the recall of 75.6, 89.1, 84.2, and 91.6% during LES and 73.3, 89.1, 84.2, and 86.9% with the DMMT criterion.

Figure 8 shows the barplots of recall, precision, and F1 metrics for the weather classification problem using five networks with LES and DMMT criterion. Both Figs. 7 and 8 show the improvement in the DMMT methodology

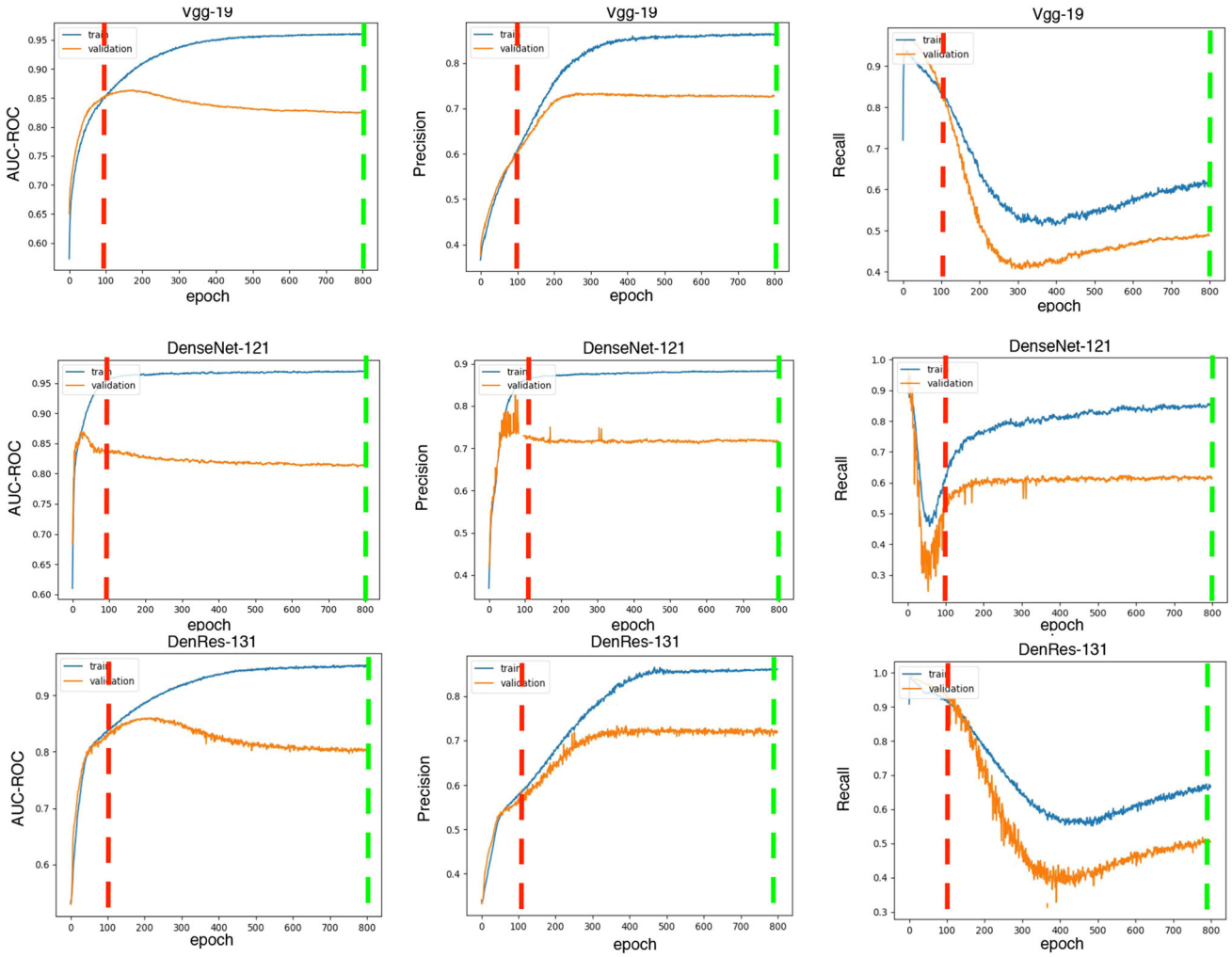


Fig. 2 Training and validation curves of the deep learning networks on BIMCV dataset for three metrics (AUC-ROC, precision, and recall). The red lines at 100 epochs are the result of the traditional technique of LES. The green lines at 800 epochs are the result of the DMMT algorithm. The red line represents a *state of weak learning* and the green a *state of robust learning*

regarding the need of a multi-metric performance evaluation, so that the network reaches a *state of robust learning* instead of a *state of weak learning*.

4.1.3 Ecological computer vision task: animal species classification

Table 5 shows the quantitative evaluation metrics of animals species classification (cat, dog, or wild) by LES and DMMT criterion. Once again, the same trend as in the medical and ecological applications is observed here. The models initially deliver a *state of weak learning* after LES and more *robust learning* after DMMT. All networks deliver higher performance for all metrics in DMMT, as compared to the LES. Figure 9 shows the behaviour of the training and validation curves for AUC-ROC, precision, recall, and F1 metrics. Even if the AUC-ROC and precision curves (Fig. 9, columns 1-2) show that the models have converged during the LES (indicated by the red dashed line), in the majority of the cases in the recall and F1 metrics the models are still in a transitional period of training (*state of weak learning*). However, the convergence state of the models is achieved at DMMT for all metrics (green line). Hence, we justify the need to observe curves for more than one metric (specifically recall, precision, AUC-ROC, and F1 here) to determine the *state of robust learning* for a deep network in an environmental classification task.

Figure 10 shows the confusion matrices of the ecological classification task for the five networks using the LES and DMMT criterion. DenRes-131 achieves recall of 93.0, 96.8, and 88.4% at LES and 92.7, 97.3, and 88.3% at DMMT. DenseNet-121 achieves recall of 92.7, 95.3, and 86.4% after LES and 93.2, 96.0, and 87.8% after DMMT. ResNet-50 achieves recall of 91.1, 95.8, and 84.9% using LES and 92.6, 95.5, and 86.6% using DMMT. VGG-16

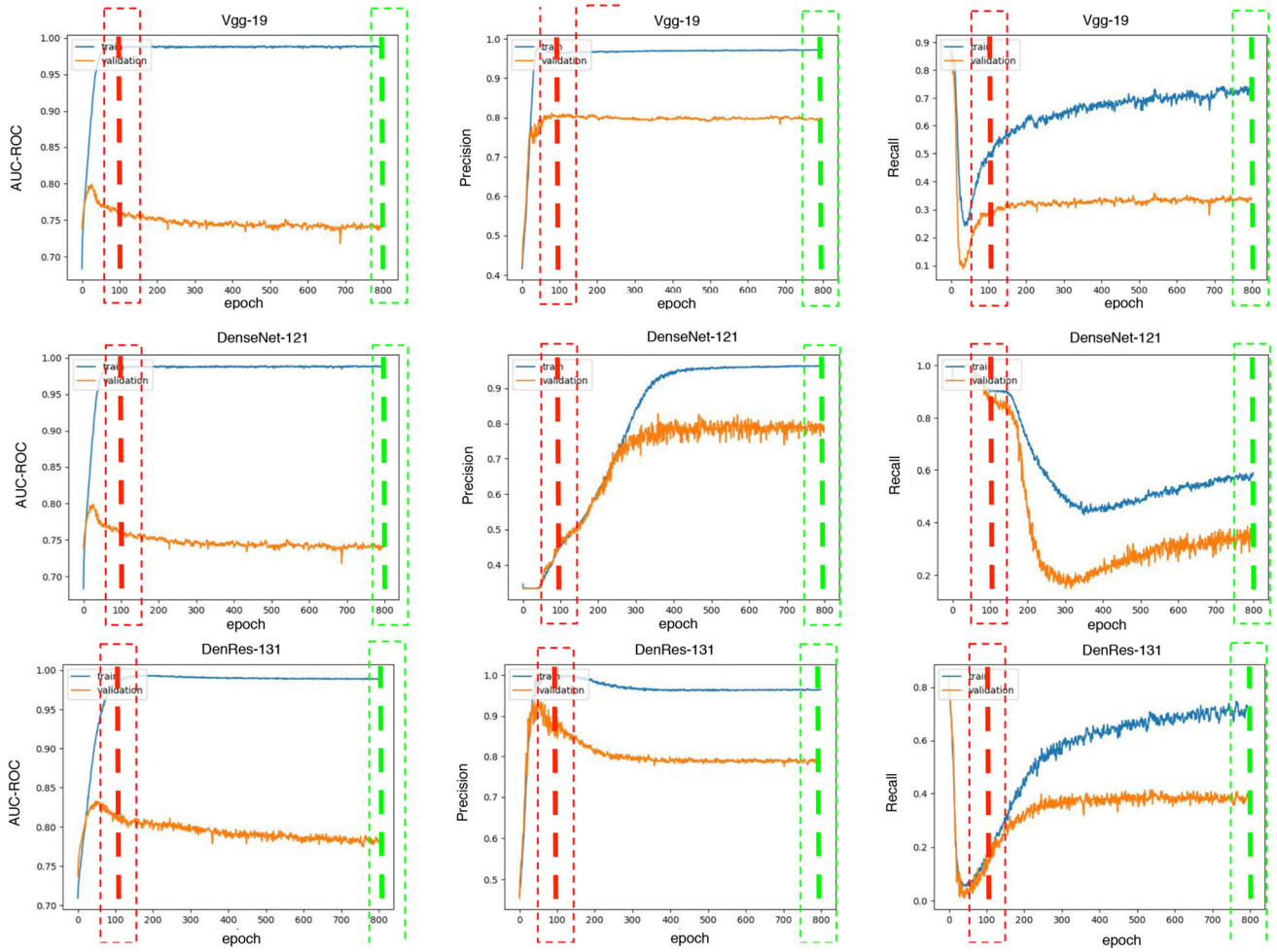


Fig. 3 Training and validation curves of the deep learning networks on Sheffield hospital dataset for three metrics (AUC-ROC, precision, and recall). The thick dashed red line at 100 epochs is the results of the established technique of LES. The thick green dashed lined at 800 epochs is the results of the DMMT algorithm. The thick red dashed line presents a *state of weak learning* and the green a *state of robust learning*

achieves recall of 94.0, 97.6, and 91.3% at LES and 94.8, 97.3, and 90.9% at DMMT. Finally, VGG-19 achieves the recall of 94.0, 97.5, and 89.4% using the LES and 94.3, 97.3, and 90.3% using the DMMT criterion.

Figure 11 highlights the barplots of the animal species classification for the five networks using the LES and DMMT. Both Figs. 10 and 11 show the proposed criterion of DMMT methodology, regarding the need of a multi-metric performance evaluation so that the network reaches a *state of robust learning*.

4.2 Evaluation of networks' generalisation: effect of DMMT

In this section, we present the results of two evaluation tests in an i.i.d. and o.o.d cohorts for the LES and DMMT criterion, in order to study their generalisation.

4.2.1 Evaluation of networks in i.i.d. cohorts: effect of DMMT

The first evaluation to examine the generalisation of the DMMT algorithm is an i.i.d. evaluation of the deep learning models in the animals testing dataset with 500 images per class. Table 6 shows the quantitative evaluation metrics without meta-learning or domain adaptation techniques in the unseen cohort of animals dataset for both LES and DMMT criterion. Once again, the networks follow the same performance patterns as in the test cohort of the animals dataset (Table 5) described in the previous subsection.

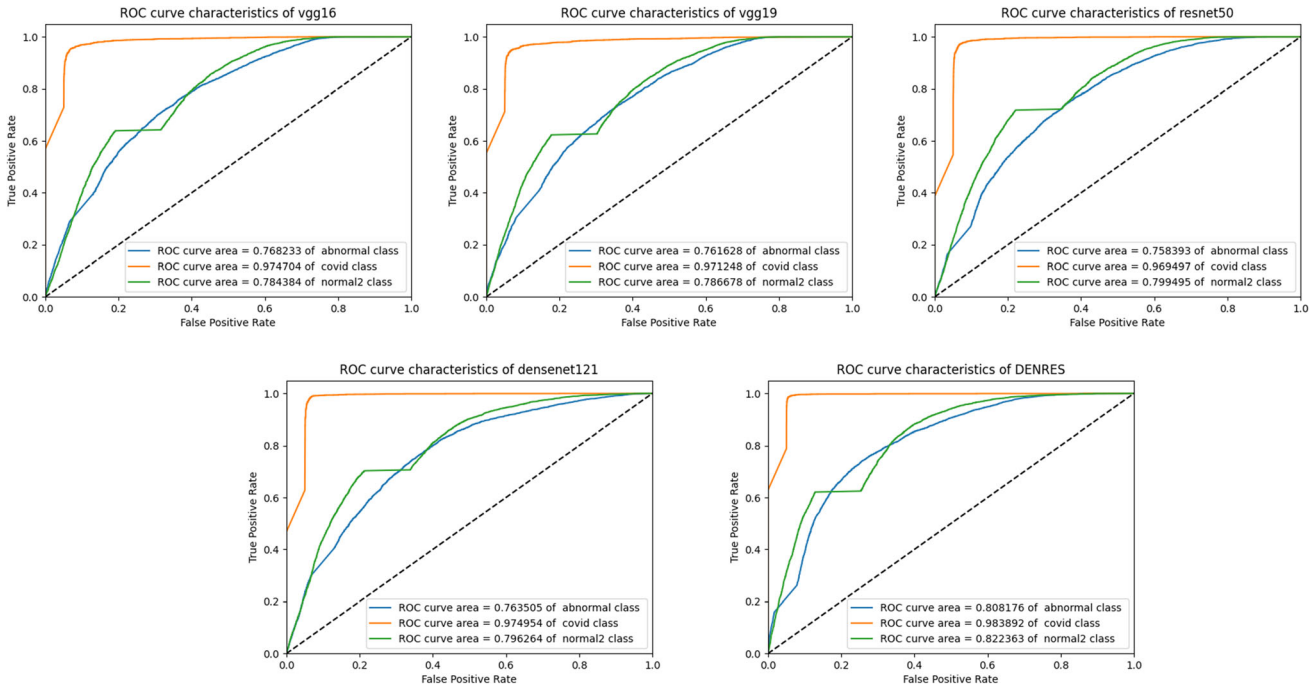


Fig. 4 ROC curves of the deep learning networks on BIMCV dataset. Row 1: VGG-16, VGG-19, and ResNet-50; row 2: DenseNet-121 and DenRes-131

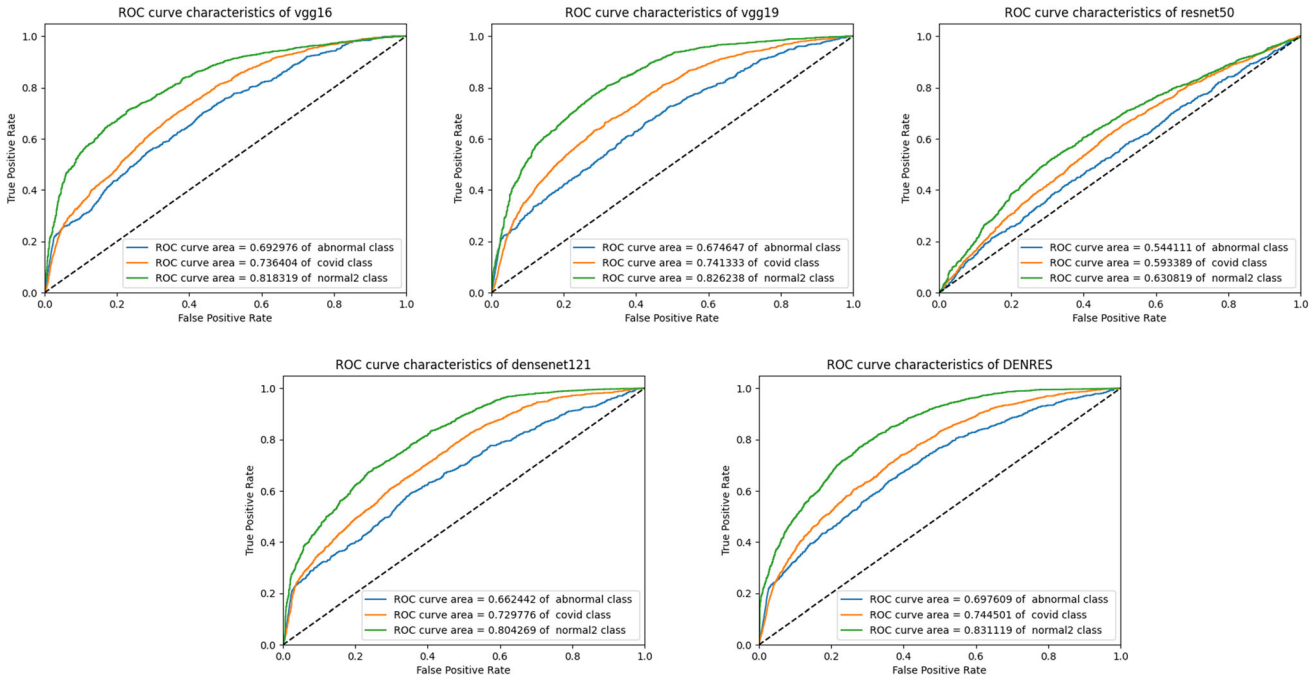


Fig. 5 ROC curves of the deep learning networks on Sheffield hospital dataset. Row 1: VGG-16, VGG-19, and ResNet-50; row 2: DenseNet-121 and DenRes-131

4.2.2 Evaluation of networks in o.o.d. cohorts: effect of DMMT

To strengthen the justification and the generalisation of the importance of multi-metric evaluation, we validate the deep learning networks using the LES and DMMT criterion on an unseen test dataset (trained on BIMCV cohort and tested on Sheffield hospital cohort) to examine their classification performance. Table 7 shows the quantitative evaluation metrics without meta-learning or domain

Table 4 Quantitative evaluation metrics of different networks on test set of the weather dataset

Classification performance on weather dataset: cloudy, rainy, shine, or sunrise					
Metric (%)	VGG-16 LES/DMMT	VGG-19 LES/DMMT	ResNet-50 LES/DMMT	DenseNet-121 LES/DMMT	DenRes-131 LES/DMMT
Recall	82.78/82.20	83.08/85.16	82.19/85.16	86.64/87.24	87.01/87.83
Precision	82.78/82.20	83.08/85.16	82.19/85.16	86.64/87.24	87.01/87.83
AUC-ROC sample	88.52/88.13	88.72/90.10	88.13/90.10	91.09/91.49	91.43/91.90
AUC-ROC macro	88.78/88.32	88.70/89.86	88.67/90.04	91.43/91.49	91.51/92.11
AUC-ROC micro	88.52/88.13	88.72/90.10	88.13/90.10	91.09/91.49	91.42/91.90
AUC-ROC weighted	88.46/88.16	88.70/90.13	88.20/90.08	91.09/91.40	91.48/91.90
F1 sample	82.79/82.19	83.08/85.16	82.19/85.16	86.64/87.24	87.23/87.89
F1 macro	82.78/81.89	82.77/84.66	82.13/84.91	86.75/87.24	87.19/87.83
F1 micro	82.78/82.19	83.08/85.16	82.19/85.16	86.64/87.24	87.21/87.85
F1 weighted	82.49/82.37	82.90/85.15	81.89/85.04	86.52/87.12	87.12/87.70

Values represent metrics using LES/DMMT criterion

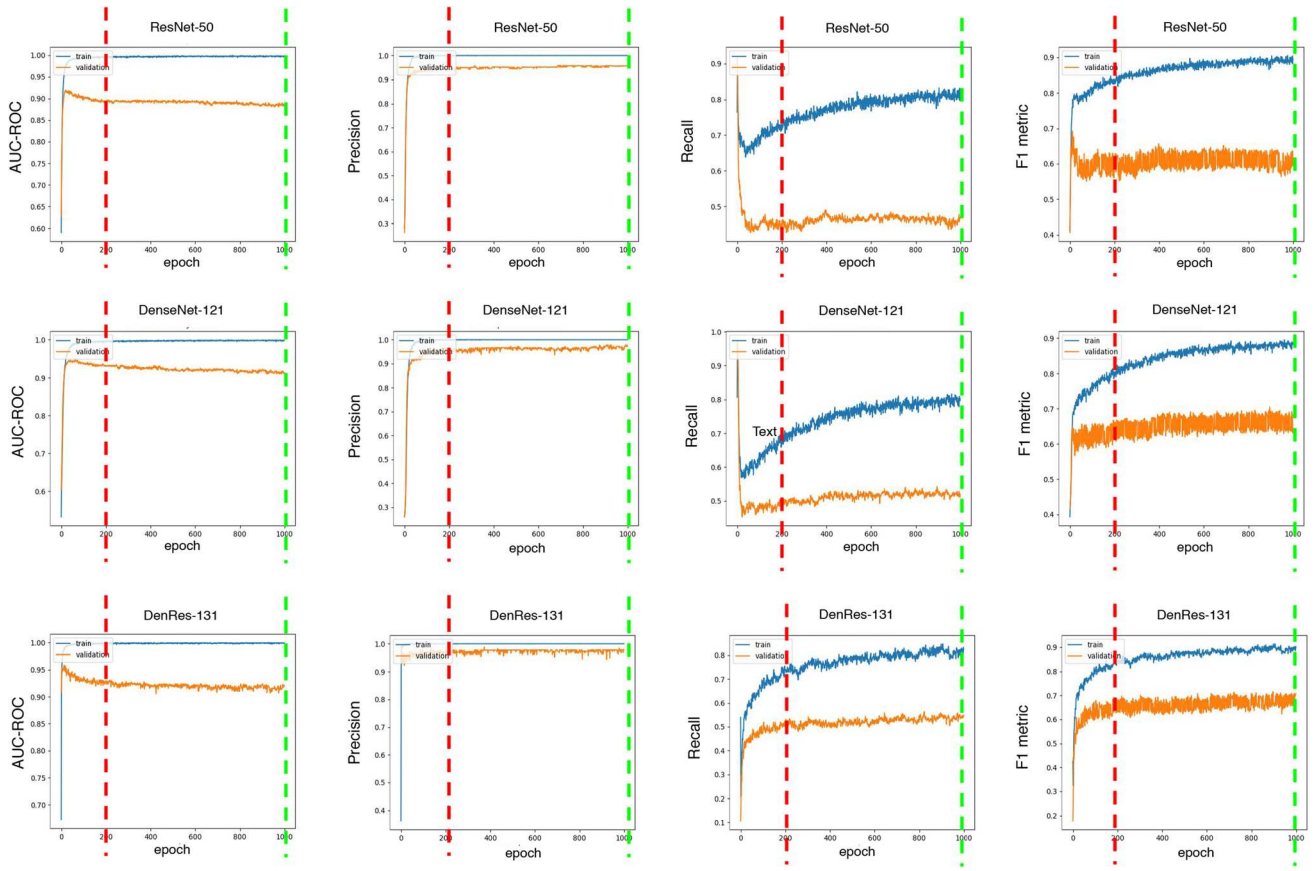
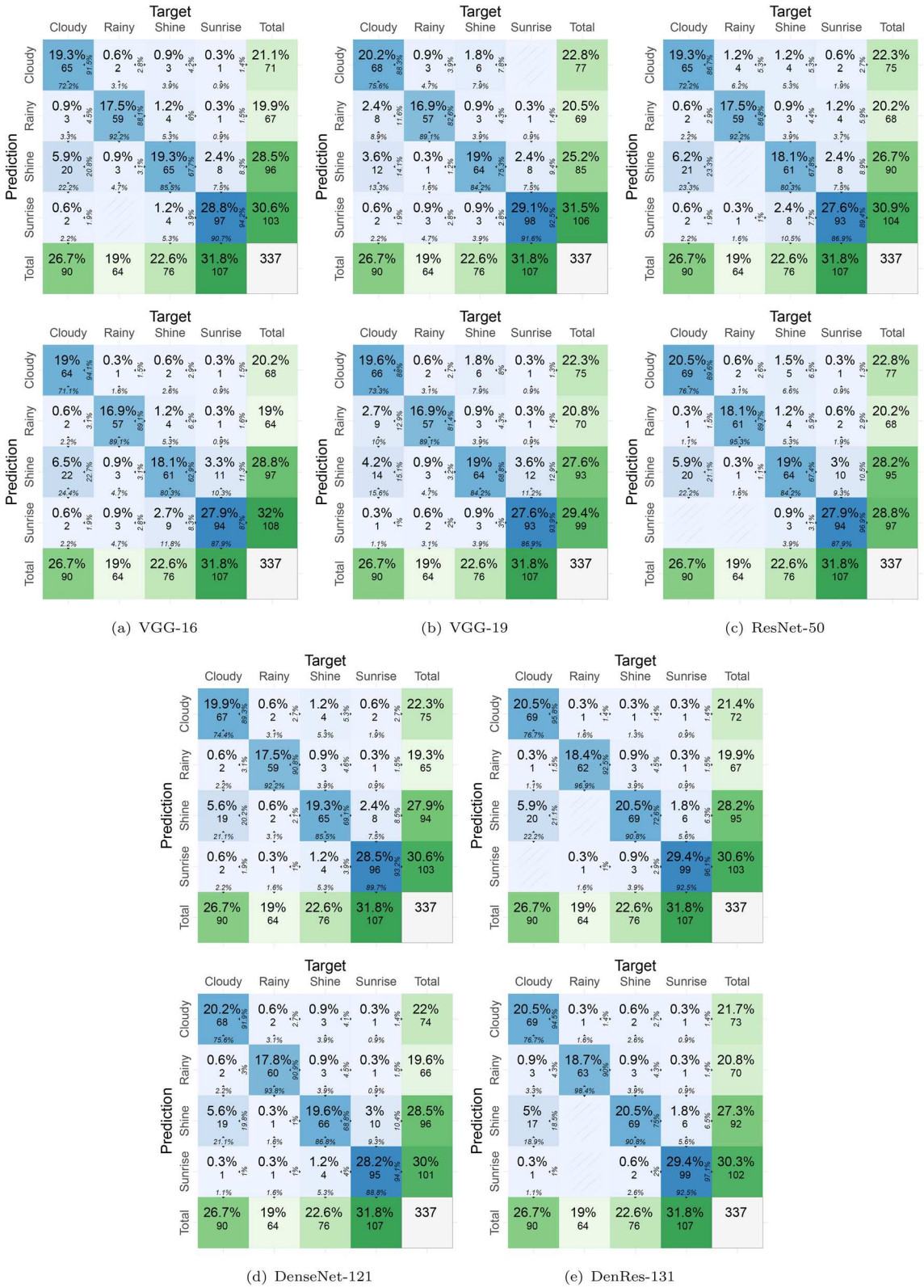


Fig. 6 Training and validation curves of the deep learning networks on the weather dataset for four metrics (AUC-ROC, precision, recall, and F1). The red dashed lines are the convergence results of the traditional LES technique. The green lines are the results of the DMMT algorithm

adaptation technique in the unseen Sheffield hospital dataset for both LES and DMMT. Once again, the networks follow the same performance patterns as in the test set of BIMCV dataset (Table 3 top) described in the previous subsection.

4.3 Statistical significance analysis of DMMT criterion

To demonstrate the effectiveness of the proposed DMMT methodology, we perform a statistical significance analysis between the metrics of the performance criterion in the



(a) VGG-16 (b) VGG-19 (c) ResNet-50

(d) DenseNet-121 (e) DenRes-131

Fig. 7 Confusion matrices of the classification performance of five deep learning networks on Weather dataset. Row 1: LES criterion, row 2: DMMT criterion

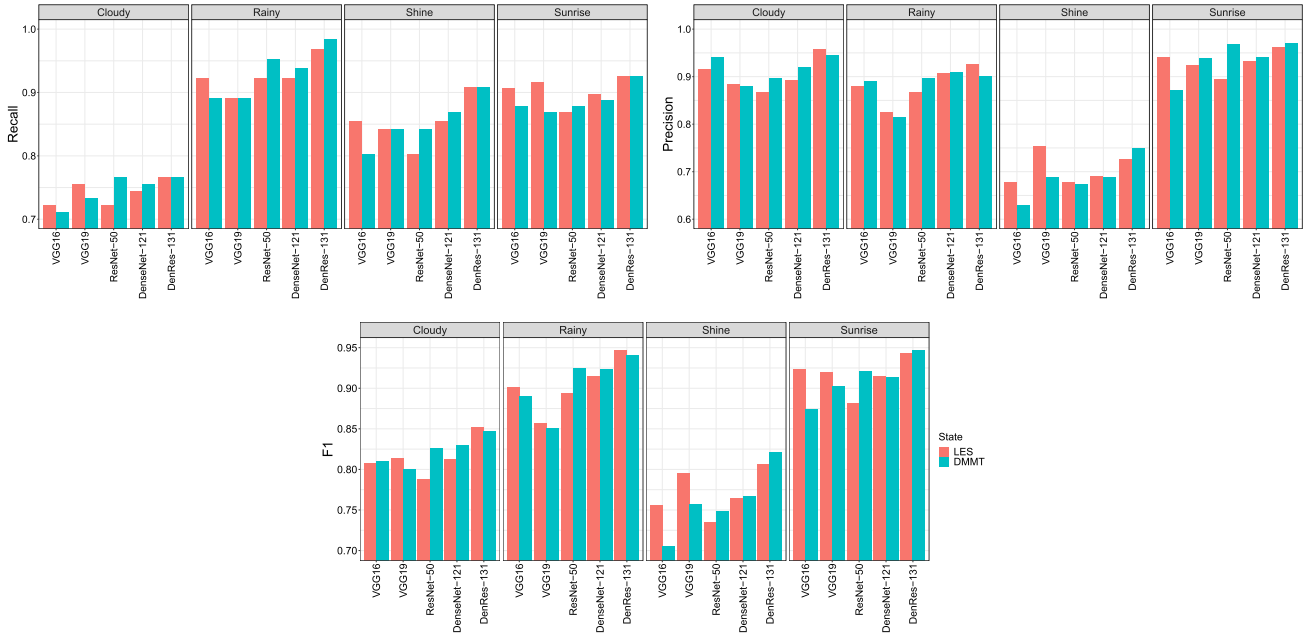


Fig. 8 Barplots for the classification performance of five deep learning networks on the Weather dataset

Table 5 Quantitative evaluation metrics of different networks on test set of the Animals dataset at the LES and DMMT criterion

Classification performance on the Animals dataset: cat, dog, or wild					
Metric (%)	VGG-16 LES/DMMT	VGG-19 LES/DMMT	ResNet-50 LES/DMMT	DenseNet-121 LES/DMMT	DenRes-131 LES/DMMT
Recall	94.00/94.44	93.70/94.04	90.70/91.67	91.50/92.45	92.78/93.01
Precision	94.00/94.44	93.70/94.04	90.71/91.67	91.50/92.45	92.78/93.01
AUC-ROC sample	95.80/95.83	95.28/95.53	93.04/93.75	93.60/94.34	94.57/95.03
AUC-ROC macro	95.81/95.86	95.34/95.57	93.10/93.80	93.72/94.40	94.65/95.11
AUC-ROC micro	95.80/95.83	95.28/95.53	93.04/93.75	93.61/94.34	94.57/95.03
AUC-ROC weighted	95.83/95.88	95.36/95.59	93.15/93.84	93.73/94.42	94.67/95.14
F1 sample	94.00/94.44	93.70/94.04	90.71/91.67	91.54/92.45	92.76/93.01
F1 macro	93.96/94.38	93.63/93.97	90.60/91.58	91.45/92.37	92.68/92.93
F1 micro	94.00/94.44	93.70/94.05	90.71/91.67	91.55/92.45	92.78/93.03
F1 weighted	94.23/94.45	93.70/94.06	90.74/91.69	91.56/92.47	92.79/93.07

Values represent metrics at the LES/DMMT criterion

state of weak learning (LES criterion) and the *state of robust learning* (DMMT criterion). We present our results as boxplots in Fig. 12, with red boxplots showing LES results and cyan boxplots the DMMT, for all quantitative metrics in Tables 3 and 7. The cyan boxplots show significant difference from the red, with reduced quartile deviation and higher median value for the majority of the metrics. This can verify our criterion of multi-metric convergence in the proposed DMMT methodology for all metrics. For the quantitative evaluation of the statistical significance analysis, we incorporate the one-tailed paired *t* test, with level of significance 0.05.

Table 8 shows the results of statistical significance analysis using the paired *t* test between the *state of weak learning* and *state of robust learning* for the recall, precision, AUC-ROC, F1, and the combination of all metrics. We only consider the medical imaging application here, since we have more samples for the statistical significance analysis (two large datasets for five networks). From the results, we can see that the *state of robust learning* is providing statistically significant improvement over the *state of weak learning* for the recall, AUC-ROC, F1, and combined metrics (with p values 0.009, 0.014, 0.04, and 0.009), while no significant difference is observed for the precision metric (with p value 0.253). Therefore, it justifies

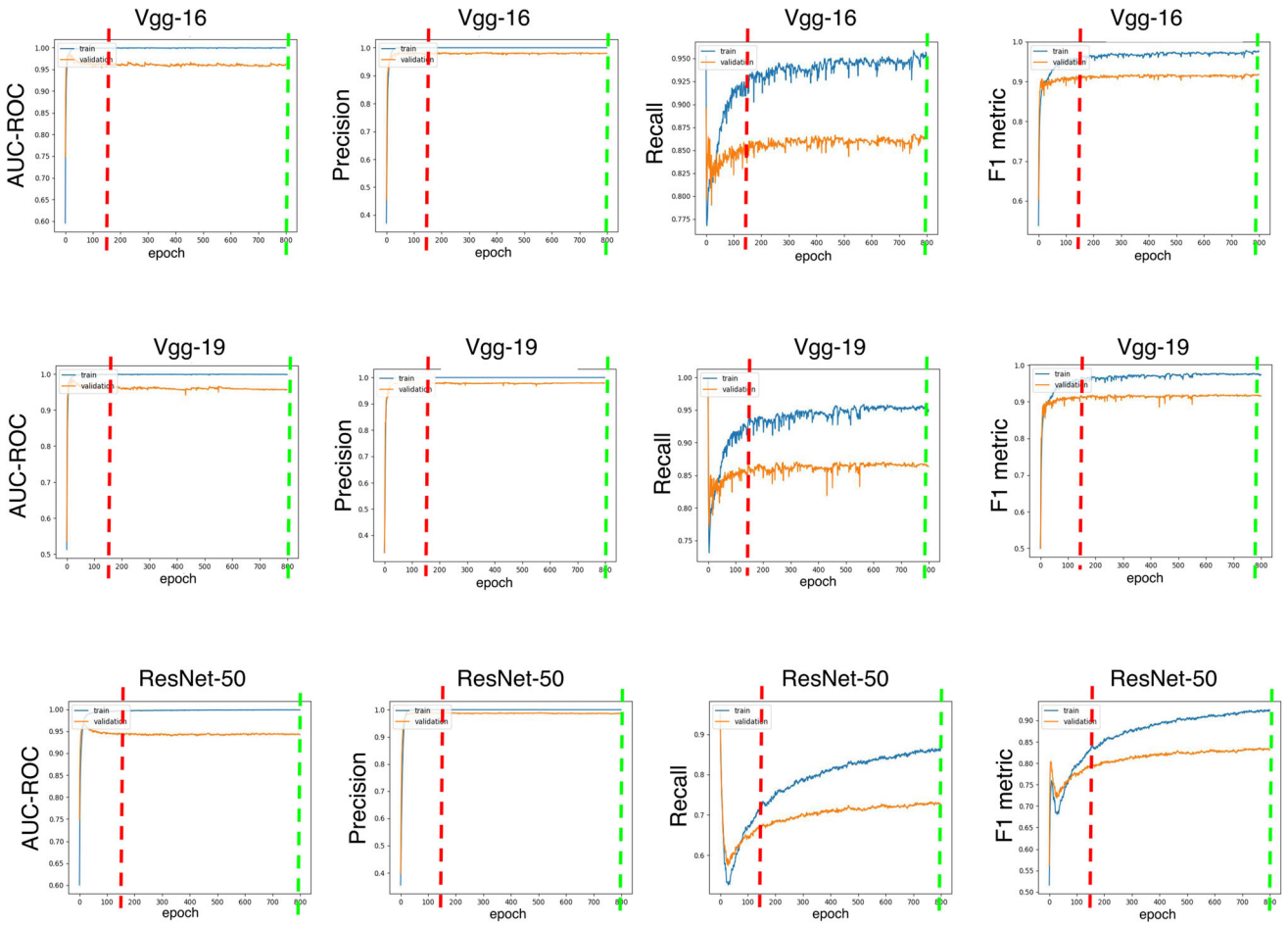


Fig. 9 Training and validation curves of the deep learning networks on Animals dataset for four metrics (AUC-ROC, precision, recall, and F1). The red lines are the convergence results of the traditional LES technique. The green lines are the results of the DMMT algorithm

the need for multi-metric evaluation in order to achieve robust learning.

4.4 Sensitivity analysis of DMMT and LES parameters

The detailed sensitivity analysis of the LES 'patience' and the DMMT Δ_t hyperparameters using the ResNet-50 network over the CIFAR-10 dataset are presented in Supplementary Material (Figures 1, 2, 3 and 4 and Tables 1, 2). For LES, the best performance metrics are: 1.429 (validation loss), 0.845 (F1), 0.213 (AUC-ROC), 0.196 (sensitivity), 0.129 (precision), 0.621 (accuracy), and 0.645 (specificity). For DMMT, the corresponding metrics are: 1.425, 0.860, 0.245, 0.220, 0.134, 0.455, and 0.395. Overall, DMMT outperforms LES in five out of seven metrics. However, these results reflect the best performance from each methodology, rather than a single trained model. To assess robustness, we consider the parameter settings that consistently yields top results across multiple metrics. The best outcomes for LES are observed with a 'patience' setting of 15, whereas for DMMT, the optimal results come with a Δ_t parameter of 20. Notably, DMMT achieves superior performance in five out of the seven metrics at this setting, indicating greater robustness and consistency.

Furthermore, DMMT shows improved results in key metrics such as F1, AUC-ROC, sensitivity, and precision. This improvement in terms of both consistency and performance metrics indicates a more robust learning state for the network when employing the DMMT methodology. Although LES occasionally achieves higher results in specific metrics (like accuracy and specificity), its performance is less consistent across different parameter settings, thus highlighting the robustness and overall reliability of DMMT over LES.

4.5 DenRes-131: a superior network again?

DenRes-131 is a new network introduced by [29], with promising state-of-the-art performance. The authors claimed that the network provides superior performance

(a) VGG-16 (b) VGG-19 (c) ResNet-50

(d) DenseNet-121 (e) DenRes-131

Fig. 10 Confusion matrices of the classification performance of five deep learning networks on the Animals dataset. Row 1: LES criterion, row 2: DMMT criterion

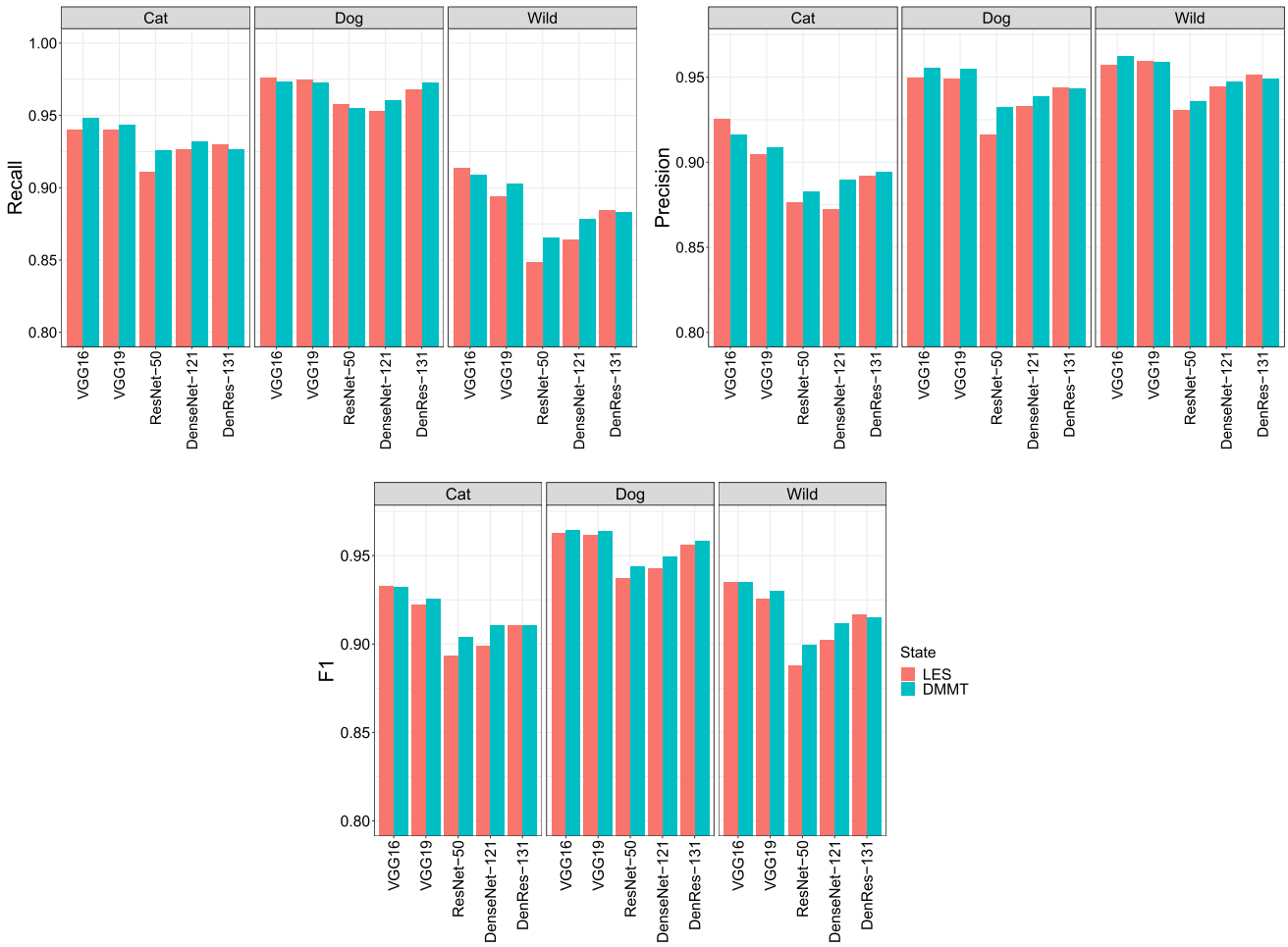


Fig. 11 Barplots for the classification performance of five deep learning networks on the Animals dataset

Table 6 Quantitative evaluation metrics of different networks on i.i.d. test set of the Animals dataset using the LES and DMMT criterion

Classification performance on i.i.d. Animals dataset: cat, dog, or wild					
Metric (%)	VGG-16 LES/DMMT	VGG-19 LES/DMMT	ResNet-50 LES/DMMT	DenseNet-121 LES/DMMT	DenRes-131 LES/DMMT
Recall	93.77/93.55	85.87/93.88	90.11/90.44	90.33/91.88	89.88/93.43
Precision	93.77/93.55	85.87/93.88	90.11/90.44	90.33/91.88	89.88/93.43
AUC-ROC sample	94.25/95.28	88.83/95.16	92.33/92.84	92.50/93.66	92.16/95.14
AUC-ROC macro	94.25/95.28	88.87/95.21	92.35/92.83	92.49/93.85	92.24/95.24
AUC-ROC micro	94.23/95.16	88.83/95.16	92.33/92.84	92.49/93.66	92.16/95.13
AUC-ROC weighted	94.23/93.58	88.83/95.16	92.35/92.83	92.53/93.80	92.19/95.14
F1 sample	93.77/93.58	85.44/93.88	90.11/90.44	90.54/91.88	89.88/93.33
F1 macro	93.78/93.52	85.46/93.91	90.10/90.44	90.45/91.86	89.90/93.33
F1 micro	93.77/93.55	85.42/93.88	90.11/90.44	90.55/91.88	89.88/93.33
F1 weighted	93.77/93.58	85.42/95.16	90.17/90.44	90.56/91.91	89.87/93.34

Values represent metrics at the LES/DMMT criterion

over established networks such as ResNet-50, DenseNet-121, and VGG-16. In this study, we justify the claim, since the DenRes-131 achieves superior performance in two medical imaging cohorts (BIMCV and Sheffield hospital) and in the o.o.d. evaluation scheme (Sect. 4.2.2) for all evaluation metrics, as presented in Tables 3 and 7. The DenRes-131 network achieves better results in terms of the ROC curve in Figs. 4 and 5, with 80.81, 98.38, and 82.23%

Table 7 Quantitative evaluation metrics of different networks for medical image classification task without meta-learning on the o.o.d. Sheffield hospital dataset

Classification performance on o.o.d. dataset: COVID-19, abnormal, or normal					
Metric (%)	VGG-16 LES/DMMT	VGG-19 LES/DMMT	ResNet-50 LES/DMMT	DenseNet-121 LES/DMMT	DenRes-131 LES/DMMT
Recall	34.61/35.69	32.92/39.01	31.66/31.11	38.03/38.76	28.26/39.33
Precision	76.06/35.69	62.67/39.01	31.66/31.11	38.03/38.76	28.26/39.33
AUC-ROC sample	51.58/51.77	50.06/54.26	48.74/48.33	53.52/54.07	50.41/55.25
AUC-ROC macro	51.85/48.50	46.80/48.98	46.34/48.15	49.75/49.22	46.20/51.44
AUC-ROC micro	51.59/51.77	50.06/54.26	48.74/48.33	53.52/54.07	50.42/55.25
AUC-ROC weighted	52.48/48.23	48.73/52.11	46.72/48.31	49.99/49.09	46.20/52.14
F1 sample	46.83/35.69	42.81/31.14	38.04/38.76	28.26/32.33	27.32/33.43
F1 macro	44.63/28.72	32.86/26.06	16.14/21.03	27.66/23.70	15.29/28.50
F1 micro	47.14/35.69	43.56/39.01	31.66/31.11	38.04/38.76	28.26/39.33
F1 weighted	47.77/40.48	48.52/49.42	47.93/42.52	45.52/50.77	43.53/46.74

Values represent metrics at the LES/DMMT criterion

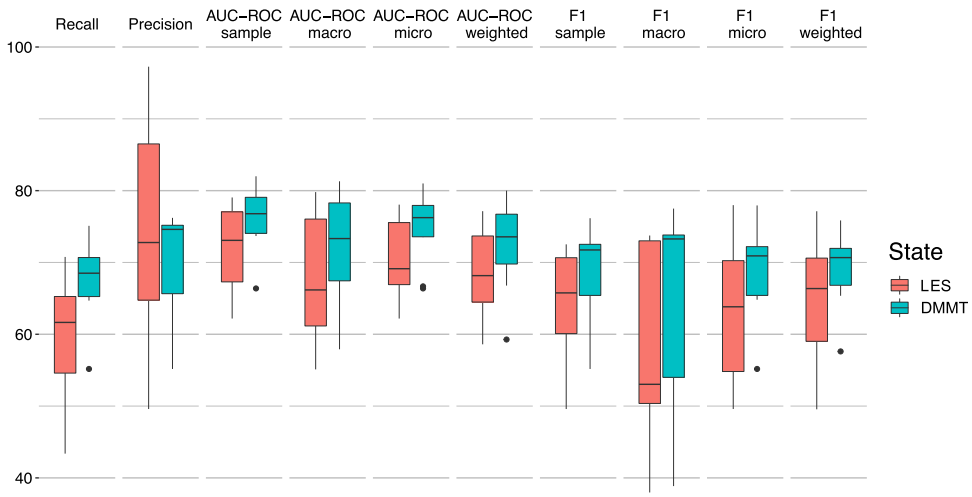


Fig. 12 Boxplots for the state of *weak* (LES criterion, red boxplots) and *robust* training (DMMT criterion, cyan boxplots) in deep learning networks. The results include the performance of all five deep learning networks on two large medical imaging datasets

Table 8 Statistical significance analysis between the *state of weak learning* of LES criterion and *state of robust learning* of DMMT criterion on the medical imaging application

Parameter	Recall	Precision	AUC-ROC	F1	Combined
Mean*	59.88/67.54	73.87/70.35	71.88/76.17	64.37/68.90	59.88/67.54
Standard deviation*	9.04/5.43	16.05/7.03	5.84/4.44	7.97/6.14	9.04/5.43
t-statistic	−2.861	−0.689	−2.602	−1.969	−2.861
Critical value one-tailed	1.833	1.833	1.833	1.833	1.833
p value one-tailed	0.009	0.253	0.014	0.040	0.009

*Values represent metrics at *weak learning* (LES)/*robust learning* (DMMT)

AUC-ROC in the BIMCV dataset and 69.76, 74.45, 83.11% AUC-ROC in the Sheffield hospital dataset for abnormal, COVID-19, and normal classes, respectively.

Furthermore, the DenRes-131 attains superior performance for the classification tasks in the environmental and ecological cohorts. Tables 4 and 5 show that DenRes-131 delivers state-of-the-art results and outperforms the other deep learning networks. More thoroughly, the DenRes-131 achieves 87.83% recall and precision, 91.90% AUC-ROC, and 87.70% F1 metric values in the Weather cohort and 93.01% recall and precision, 95.14% AUC-ROC, and 93.07% F1 metric values in the Animal species cohort. For the ROC curves, DenRes-131 outperforms all classes' scores compared to the VGG-16, VGG-19, ResNet-50, and DenseNet-121 networks for the environmental and ecological classification problems.

Figures 7, 8, 10, and 11 show the performance of the networks based on true positive and true negative

predictions and recall, precision, and F1 metrics for the *state of weak learning* (LES criterion) and the *state of robust learning* (DMMT criterion). The DenRes-131 outperforms the other established networks in the environmental classification problem and achieves the similar level of performance compared to the leading VGG networks in the ecological classification problem. We did not expect DenRes-131 to outperform the VGG networks in this cohort, as the VGG networks perform significantly better than both ResNet-50 and DenseNet-121 networks in this dataset. This probably happens because the VGG structures outperform the complex structures of ResNet and DenseNet for less complicated classification problems such as the animal species classification [48].

5 Discussion

We have developed a new deep multi-metric training (DMMT) methodology to avoid the *state of weak learning* of a deep learning network for medical, environmental, and ecological classification tasks. The convergence criterion of the DMMT methodology is defined as the optimal number of epochs for achieving equilibrium in the user-defined multi-metric performance (recall, precision, AUC-ROC, F1, etc.). One important limitation of this study is the utilisation of one computer vision task, namely classification, to verify the optimal training methodology. To generalise the proposed methodology, a study involving different computer vision tasks (e.g. semantic segmentation, regression, object detection, etc.) is required. Another less important limitation of this study is that the classification experiment has been applied on medical, environmental, and ecological datasets. A further investigation on some other fields such as automation and industrial classification problems could be beneficial. The main advantage of this study is the simplicity of the converge criterion to deliver *state of robust learning* performance for a deep network (criterion of multi-metric performance evaluation).

In the second part of this study, we have examined the performance of DenRes-131 compared to other established networks of VGG-16, VGG-19, ResNet-50, and DenseNet-121. DenRes-131 was first introduced in [29] with promising state-of-the-art performance, and it provided superior results compared to established networks of ResNet-50, DenseNet-121, and VGG-16. The DenRes-131 was initially tested in small size cohorts due to the lack of available large COVID-19 datasets. Thus, one of the aims of this study has been to further evaluate its performance in larger COVID-19 datasets (BIMCV COVID-19+ and Sheffield hospital datasets). In addition, we are interested to study the performance of the network in multi-field classification problems such as environmental and ecological classification tasks. The network outperforms the established networks in the environmental problem and provides similar performance with the leading VGG-16 and VGG-19 networks in the ecological task.

In our future study, we want to focus on the generalisation of the DMMT methodology for *robust learning* in different computer vision tasks such as semantic segmentation, regression, and object detection. We wish to evaluate the performance of DenRes-131 in industrial classification problems and present an ablation analysis study of the network structure. We are also interested in evaluating the performance of Bayesian optimisation when combined with the DMMT.

6 Is faster always better? Concluding remarks on DMMT methodology

In this study, we have proposed the DMMT methodology, which incorporates a convergence criterion that defines the optimal number of epochs for achieving an equilibrium point in multi-metric performance, including recall, AUC-ROC, precision, F1, and others. Unlike most existing methodologies, which rely on loss early stopping (LES) or evaluation of the network's training based solely on accuracy metric results, our approach demonstrates a distinct advantage. In validation protocols, we have demonstrated that our proposed methodology outperforms the established training methodology that employs the LES criterion. Our findings indicate that achieving the point of equilibrium for the multi-metrics evaluation methodology may require deeper epochs, suggesting that faster training is not always the optimal solution. Overall, our research offers a valuable contribution by providing a more effective and efficient methodology for achieving generalised and robust performance of deep learning networks. Moreover, we have verified the superior performance of the deep learning network DenRes-131 [29] on four large imaging datasets.

Our study has revealed that in our analysis faster training is not the best approach for achieving optimal accuracy performance in multi-metrics evaluation. We have observed the point of equilibrium may only be reached after training for deeper epochs, suggesting that a slower and more deliberate approach to training may be more effective.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-024-10182-6.

Acknowledgements The authors acknowledge the use of the facilities of the Research Software Engineering Sheffield (RSE), UK, and the JADE2 Tier 2 HPC UK system specification.

Author contributions Michail Mamalakis contributed to conceptualisation, data curation, methodology, visualisation, supervision, investigation, formal analysis, writing—original draft preparation, writing—reviewing and editing, and validation and provided software. Abhirup Banerjee was involved in conceptualisation, visualisation, co-supervision, statistical analysis, writing—original draft preparation, writing—reviewing and editing, and validation. Surajit Ray contributed to writing—reviewing and editing, and validation. Craig Wilkie was involved in visualisation and writing—reviewing and editing. Richard H. Clayton contributed to writing—reviewing and editing. Andrew J. Swift was involved in resources and data curation. George Panoutsos contributed to writing—reviewing and editing, and validation. Bart Vorselaars was involved in conceptualisation, methodology, writing—original draft preparation, writing—reviewing and editing, and validation.

Funding Abhirup Banerjee is a Royal Society University Research Fellow and is supported by the Royal Society Grant No. URFR1221314. The work of Andrew J. Swift was supported by the Wellcome Trust fellowship grant 205188/Z/16/Z. The work of George Panoutsos was supported by EPSRC grant EP/P006566/1.

Data availability The non-public Sheffield hospital medical imaging dataset can be provided upon request from the corresponding author.

Code availability The code is available upon request from the corresponding author.

Declaration

Conflict of interest The authors express no conflict of interest.

Ethics approval All medical datasets used in this study are anonymised with ethics approval.

References

1. Adedigba AP, Adeshina SA, Aina OE, Aibinu AM (2021) Optimal hyperparameter selection of deep learning models for COVID-19 chest x-ray classification. Intell Based Med 5:100034
2. Banerjee A, Ray S, Vorselaars B, Kitson J, Mamalakis M, Weeks S, Mackenzie LS (2020) Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. Int Immunopharmacol 86:106705
3. Boss AN, Banerjee A, Mamalakis M, Ray S, Swift AJ, Wilkie C, Mackenzie LS (2022) Development of a mortality prediction model in hospitalised SARS-CoV-2 positive patients based on routine kidney biomarkers. Int J Mol Sci 23:13
4. Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M (2020) PadChest: A large chest x-ray image dataset with multi-label annotated reports. Med Image Anal 66:101797
5. Chen Z, Liu Y, Zhu J, Zhang Y, Li Q, Jin R, He X (2021) Deep multiple metric learning for time series classification. IEEE Access 9:17829–17842
6. Choi Y , Uh Y , Yoo J, Ha JW (2020) StarGAN v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8188–8197). Data available via https://www.kaggle.com/andrewmvd/animal-faces/metadata
7. Das D, Santosh KC, Pal U (2020) Truncated inception net: COVID-19 outbreak screening using chest x-rays. Phys Eng Sci Med 43:915–925
8. de la Iglesia Vayá M, Saborit JM, Montell JA, Pertusa A, Bustos A, Cazorla M, Salinas JM (2020) BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. CoRRabs/2006.01174
9. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition pp 248–255
10. Dong X, Shen J, Wang W, Liu Y, Shao L, Porikli F (2018) Hyperparameter optimization for tracking with continuous deep Q-learning. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (pp. 518–527)
11. Feurer M, Klein A, Eggensperger K, Springenberg JT, Blum M, Hutter F (2019) Auto-sklearn: Efficient and robust automated machine learning (F. Hutter, L. Kotthoff, J. Vanschoren, eds.). Cham: Springer International Publishing
12. Gbeminiyi Oluwafemi A, Zenghui W (2019) Multi-class weather classification from still image using said ensemble method. In: 2019 Southern African universities power engineering conference/robotics and mechatronics/pattern recognition association of South Africa (SAUPEC/RobMech/PRASA) (pp. 135–140)
13. Greenspan H, Estépar RSJ, Niessen WJ, Siegel E, Nielsen M (2020) Position paper on COVID-19 imaging and AI: from the clinical needs and technological challenges to initial AI solutions at the lab and national level towards a new era for AI in healthcare. Med Image Anal 66:101800
14. Harmon SA, Sanford TH, Xu S, Turkbey EB, Roth H, Xu Z, Turkbey B (2020) Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat Commun 11(1):4080
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR) pp 770–778
16. He X, Zhao K, Chu X (2021) AutoML: A survey of the state-of-the-art. Knowl-Based Syst 212:106622
17. Huang G, Liu Z, Maaten LVD, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR) pp 2261–2269
18. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, Maier-Hein KH (2018) nnU-Net: Self-adapting framework for U-net-based medical image segmentation. CoRRabs/1809.10486
19. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJ (2019) Identifying pneumonia in chest x-rays: A deep learning approach. Measurement 145:511–518
20. Kim JY, Cho SB (2019) Evolutionary optimization of hyperparameters in deep learning models. In: 2019 IEEE congress on evolutionary computation (CEC) pp 831–837

21. Koivunen AC, Kostinski AB (1999) The feasibility of data whitening to improve performance of weather radar. J Appl Meteorol 38(6):741–749
22. Lalmuanawma S, Hussain J, Chhakchhuak L (2020) Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons Fractals 139:110059
23. Li H, Chaudhari P, Yang H , Lam M, Ravichandran A, Bhotika R, Soatto S (2020) Rethinking the hyperparameters for fine-tuning. CoRRabs/2002.11770
24. Li K, Fang Y, Li W, Pan C, Qin P, Zhong Y, Li S (2020) CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). Eur Radiol 30(8):4407–4416
25. Mahsereci M, Balles L, Lassner C, Hennig P (2017) Early stopping without a validation set. CoRRabs/1703.09580. http://arxiv.org/abs/1703.09580
26. Mamalakis A, Barnes EA, Ebert-Uphoff I (2022) Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. Art Intell Earth Syst 1(4):e220012. https://doi.org/10.1175/AIES-D-22-0012.1
27. Mamalakis A, Ebert-Uphoff I, Barnes E (2022) Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. In: A. Holzinger, R. Goebel, R. Fong, T. Moon, K. R. Müller., W. Samek (eds.), xxai - beyond explainable ai: International workshop, held in conjunction with icml 2020, july 18, 2020, vienna, austria, revised and extended papers (pp. 315–339). Cham:Springer International Publishing. doi: https://doi.org/10.1007/978-3-031-04083-2_16
28. Mamalakis M, Garg P, Nelson T, Lee J, Wild JM, Clayton RH (2021) MA-SOCRATIS: An automatic pipeline for robust segmentation of the left ventricle and scar. Comput Med Imaging Graph 93:101982
29. Mamalakis M, Swift AJ, Vorselaars B, Ray S, Weeks S, Ding W, Banerjee A (2021) DenResCov-19: A deep transfer learning network for robust automatic classification of COVID-19, pneumonia, and tuberculosis from x-rays. Comput Med Imaging Graph 94:102008
30. MHRA UMHRA (2022) Guidance: Target product profile: Point of care sars-cov-2 detection tests. https://www.gov.uk/government/publications/how-tests-and-testing-kits-for-coronavirus-covid-19-work/target-product-profile-point-of-care-sars-cov-2-detection-tests
31. Mohamed B, Daoud M, Mohamed B, Ahmed A (2022) Improvement of emotion recognition from facial images using deep learning and early stopping cross validation. Multimed Tools Appl. https://doi.org/10.1007/s11042-022-12058-0
32. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U (2020) Automated detection of COVID-19 cases using deep neural networks with x-ray images. Comput Biol Med 121:103792
33. Perona P, Malik J (1990) Scale-space and edge detection using anisotropic diffusion. IEEE Trans Pattern Anal Mach Intell 12(7):629–639
34. Prechelt L (1998) Early stopping - but when? In: G.B. Orr, K. R. Müller (eds.), Neural networks: Tricks of the trade (pp. 55–69). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: https://doi.org/10.1007/3-540-49430-8_3
35. Ray S, Swift A, Fanstone JW, Banerjee A, Mamalakis M, Vorselaars B, Weeks S (2021) LUCAS: A highly accurate yet simple risk calculator that predicts survival of COVID-19 patients using rapid routine tests. medRxiv
36. Samaniego E, Anitescu C, Goswami S, Nguyen-Thanh V, Guo H, Hamdia K, Rabczuk T (2020) An energy approach to the solution of partial differential equations in computational mechanics via machine learning: Concepts, implementation and applications. Comput Methods Appl Mech Eng 362:112790. https://doi.org/10.1016/j.cma.2019.112790
37. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. CoRRabs/1409.1556
38. Song JW, Lam SM, Fan X, Cao WJ, Wang SY, Tian H, Shui G (2020) Omics-driven systems interrogation of metabolic dysregulation in COVID-19 pathogenesis. Cell Metab 32(2):188-202.e5
39. Souquet L, Shvai N, Llanza A, Nakib A (2020) Hyperparameters optimization for neural network training using fractal decomposition-based algorithm. 2020 IEEE congress on evolutionary computation (CEC) (pp. 1–6)
40. van Rijn JN, Hutter F (2018) Hyperparameter importance across datasets. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2367–2376)
41. Varela-Santos S, Melin P (2021) A new approach for classifying coronavirus COVID-19 based on its manifestation on chest x-rays using texture features and neural networks. Inf Sci 545:403–414
42. Varshni D, Thakral K, Agarwal L, Nijhawan R, Mittal A (2019) Pneumonia detection using CNN based feature extraction. In: IEEE international conference on electrical, computer and communication technologies (pp. 1–7)
43. Vonesch C, Unser M (2008) A fast thresholded landweber algorithm for wavelet-regularized multidimensional deconvolution. IEEE Trans Image Process 17(4):539–549
44. Waring J, Lindvall C, Umeton R (2020) Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med 104:101822
45. Xu J, Wang X, Feng B, Liu W (2020) Deep multi-metric learning for text-independent speaker verification. Neurocomputing 410:394–400
46. Zhang T, Zhu T, Gao K, Zhou W, Yu PS (2021) Balancing learning model privacy, fairness, and accuracy with early stopping criteria. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2021.3129592
47. Zhang X, Chen X, Yao L, Ge C, Dong M (2019) Deep neural network hyperparameter optimization with orthogonal array tuning. CoRRabs/1907.13359
48. Zhongqi M, Jiayun W, Ziwei L, Oliver M (2019) Insights and approaches using deep learning to classify wildlife. Sci Rep 9:8137
49. Zhou S, Song W (2020) Deep learning-based roadway crack classification using laser-scanned range images: A comparative study on hyperparameter selection. Autom Constr 114:103171

Authors and Affiliations

Michail Mamalakis[1,2,3] · Abhirup Banerjee[8,9] · Surajit Ray[4] · Craig Wilkie[4] · Richard H. Clayton[2,3] · Andrew J. Swift[3,5] · George Panoutsos[6] · Bart Vorselaars[7]

✉ Michail Mamalakis
mm2703@cam.ac.uk

Abhirup Banerjee
abhirup.banerjee@eng.ox.ac.uk

Surajit Ray
surajit.ray@glasgow.ac.uk

Bart Vorselaars
bvorselaars@lincoln.ac.uk

[1] Department of Psychiatry, University of Cambridge, Herchel Smith Building, Robinson Way, Cambridge CB2 0SZ, UK

[2] Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield S1 4DP, UK

[3] Insigneo Institute for in silico Medicine, University of Sheffield, The Pam Liversidge Building, Sheffield S1 3JD, UK

[4] School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK

[5] Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Sheffield S10 2HQ, UK

[6] Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S10 2TN, UK

[7] School of Mathematics and Physics, University of Lincoln, Brayford Pool, Lincoln LN6 7TS, UK

[8] Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford OX3 7DQ, UK

[9] Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 9DU, UK