Expert Systems With Applications 187 (2022) 115834

Contents lists available at ScienceDirect

Expert Systems With Applications

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

# Efficient ensemble for image-based identification of Pneumonia utilizing deep CNN and SGD with warm restarts

Grega Vrbančič *, Vili Podgorelec

*University of Maribor, Faculty of Electrical Engineering and Computer Science, Koroška cesta 46, SI 2000, Maribor Slovenia*

ARTICLE INFO

*Keywords:*
Deep learning
Convolutional neural network
Ensemble model
Image classification
Stochastic gradient descent
Identification of pneumonia

ABSTRACT

Childhood pneumonia, the leading cause of children mortality globally, is most commonly diagnosed based on the radiographic data, which requires radiologic interpretation of X-ray images. With recent advancements in the field of deep learning, the convolutional neural networks (CNN) have proven to be able to achieve great performance in medical image segmentation, analysis and classification tasks. However, developing and training methods utilizing CNN is still a complex and time-consuming process with several open issues — generalization, demand for large datasets, and high time complexity. The optimization objective in the training of CNN models has multiple minima, which do not necessarily generalize well and may result in poor performance. Ensemble methods are commonly employed to address the generalization issue, but they require a group of diverse models and are generally even more time-consuming. To address the issues of generalization, dataset size and time complexity, we developed an ensemble method based on stochastic gradient descent with warm restarts (SGDRE) that exploits the generalization capabilities of ensemble methods and SGD with warm restarts mechanism, which is adopted to obtain a diverse group of classifiers, necessary for ensemble, in one single training process, spending the same or less training time than a single CNN classification model. The SGDRE method has been trained on publicly available pediatric chest X-ray images dataset and evaluated using 10-fold cross-validation approach. The experimental results show a significant improvement of SGDRE over the two compared baseline methods. With an achieved test accuracy of 96.26% and AUC of 95.15%, the proposed method proved to be a very competitive classification method.

## 1. Introduction

Childhood pneumonia, reported by Rudan, Boschi-Pinto, Biloglav, Mulholland, and Campbell (2008) is the leading cause of children mortality globally. It is responsible for about 19% of all deaths in children aged less than 5 years, the majority of which take place in developing countries, particularly in Southeast Asia and Africa (Rudan, et al., 2013). Two leading causes of pneumonia are bacterial and viral pathogens. In contrast to viral pneumonia, which is treated with supportive care, bacterial pneumonia requires urgent referral for immediate antibiotic treatment, therefore accurate and timely diagnosis is essential (Kermany, et al., 2018). The pneumonia is most commonly diagnosed based on the radiographic data, which requires radiologic interpretation of X-ray images.

The early attempts at computer-aided diagnosis (CAD) and analysis of medical images were made in the 1960s (Lodwick, Haun, Smith, Keller, & Robertson, 1963), when it was generally assumed that computers could replace radiologists in detecting abnormalities because computers are better performing at certain tasks than are human beings, but those attempts were sadly not successful. Two decades later, in the 1980s, the large-scale and systematic research and development of various CAD schemes were begun at the Kurt Rossman Laboratories for Radiologic Image Research in the Department of Radiology at the University of Chicago. At that time appeared to be extremely difficult to carry out a computer analysis on lesions involved in medical images and therefore, not easy to predict whether the development of CAD schemes would be successful or not. However, research studies took a turn with a different approach, which assumed that the computer output could be utilized by radiologists instead of replacing them. This approach is currently known as computer-aided diagnosis, which has spread widely and quickly (Doi, 2007).

Through the years, the growth of the computational capacity enabled researchers to utilize more complex methods to drive the expansion of CAD and analysis of medical images even further. One of such major breakthroughs occurred in 1998 with the LeCun's proposal of convolutional neural network (CNN) architecture LeNet5 (LeCun, Bottou, Bengio, Haffner, et al., 1998), which enabled researchers from various fields to tackle different image recognition tasks more easily.

* Corresponding author.
*E-mail addresses:* grega.vrbancic@um.si (G. Vrbančič), vili.podgorelec@um.si (V. Podgorelec).

https://doi.org/10.1016/j.eswa.2021.115834
Received 2 April 2020; Received in revised form 10 August 2021; Accepted 29 August 2021
Available online 7 September 2021

It also started an enormous expansion of the deep learning field. In recent years, deep learning has become a leading machine learning tool in medical image analysis (Qin, Yao, Shi, & Song, 2018), CAD systems (Yanase & Triantaphyllou, 2019), biomedical signal segmentation (Rouhi, Jafari, Kasaei, & Keshavarzian, 2015) and detection of various human organ activities (Vrbancic, Fister, & Podgorelec, 2019).

One of such tasks, where the exploitation of CNN capabilities could provide great results, is the detection of pneumonia from chest X-ray images. As the rapid radiologic interpretation of X-ray images is not always available, especially in low-resource countries where childhood pneumonia has the highest rates of mortality, it would be beneficial to utilize the modern CNN based methods to help with the identification of pneumonia disease in the early stages (Kermany, et al., 2018). However, developing and training methods utilizing CNNs is still a complex and time-consuming process with several open issues. One of the major problems is poor generalization. As CNN models are used for solving the most complex problems (image pixels represent several thousands of input attributes), its search space is huge, and the learning optimization objective has multiple minima, all of which do not necessarily generalize well. Therefore, picking the wrong minimum can lead to poor performance. Another problem is the necessity for large datasets (by an expert carefully labeled training instances) upon which the model could be trained and optimized. The third issue is the time complexity of the learning process.

In recent years, various CNN based methods has been presented in order to address the issue of detecting chest pathologies (Bardou, Zhang, & Ahmad, 2018; Chouhan, et al., 2020; Guan & Huang, 2020; Ke, et al., 2019; Lakhani & Sundaram, 2017; Liang & Zheng, 2020; Rajpurkar, et al., 2017; Shen, Han, Aberle, Bui, & Hsu, 2019; Taylor, Mielke, & Mongan, 2018), which are showing promising results, including the problem of detecting pneumonia from chest X-ray images. In Rajpurkar, et al. (2017), Siddiqi (2019), Stephen, Sain, Maduh, and Jeong (2019) the authors presented specialized CNN architectures for the purpose of identifying pneumonia from chest X-ray images which on the one hand deliver promising classification performance, while on the other hand does not address the problem of time complexity or the problem of generalization. Recently, various studies (Baltruschat, Nickisch, Grass, Knopp, & Saalbach, 2019; Kermany, et al., 2018) have shown that the utilization of transfer learning approaches provides us with high classification performance utilizing different pre-trained CNN architectures without the need of large labeled datasets. While utilizing such approaches is most definitely promising, we have to consider the problems introduced by using transfer learning, which are most commonly related to selecting a most suitable combination of fine-tunable layers (Guo, et al., 2018; Vrbančič & Podgorelec, 2020) as well as the problem of regularization when training complex CNN architecture on a small dataset. Chouhan, et al. (2020), also presented a novel transfer learning approach for pneumonia detection in chest X-ray images, where outputs of multiple CNN architecture models were combined in the ensemble method. While such approach provided state-of-the-art performance, the time-complexity when using such approach is increased. Most commonly, when utilizing the transfer learning approach, one would use a model pretrained on an ImageNet dataset. In contrast to such an approach, Liang and Zheng (2020) constructed a custom CNN architecture which was afterwards trained on a large chest X-ray dataset containing a total of 112,120 chest X-ray images. The obtained model was then used as a pre-trained model for utilization in the transfer learning approach. While such approach also delivered high classification performance, the problem of time complexity as well as the problem of generalization still remains. Guan and Huang (2020) in their research proposed a category-wises residual attention learning framework for multi-label chest X-ray image classification utilizing CNN as a feature embedding module and a residual attention module. While the proposed approach achieved state-of-the-art classification performance, the problem of generalization remains unaddressed. Another group of approaches, which is recently quite popular, is one where the evolutionary algorithms are utilized (Chandra, Verma, Singh, Jain, & Netam, 2021; El-Kenawy, et al., 2021; Khishe, Caraffini, & Kuhn, 2021; Radiuk & Kutucu, 2020; Singh, Kumar, Mahmud, Kaiser, & Kishore, 2021). Most commonly, the evolutionary algorithms are exploited for the purpose of designing deep neural network architectures (Khishe et al., 2021; Radiuk & Kutucu, 2020) or for the feature selection task (Chandra et al., 2021; El-Kenawy, et al., 2021; Singh et al., 2021) in which case the classification itself is conducted using conventional classifiers or deep learning approaches.

In general the transfer learning based methods do have an impact on the decreased time complexity in the process of training and do not suffer when trained on smaller dataset (assuming we have a well pre-trained model, which itself needs to be pre-trained on a huge amount of data), the issue of generalization (picking the right one from multiple minima) still remains. On the other hand, specialized CNN architectures still require rather large datasets and do not solve any of the mentioned issues well.

In order to address the mentioned issues, we developed an efficient stochastic gradient descent with warm restarts ensemble (SGDRE) method, which exploits the generalization capabilities of ensemble methods and stochastic gradient descent with warm restarts mechanism which is adopted to obtain a diverse group of classifiers, necessary for ensemble, in one single training process, limited by a given training budget not higher than for a single CNN classification model.

We summarize our contributions as follows:

- We present a novel ensemble method based on the stochastic gradient descent with warm restarts utilizing Deep CNN, which addresses the common deep learning issues such as generalization, size of dataset and time complexity of the training process.
- We presented new, sine based annealing function for stochastic gradient descent with warm restarts method.
- We conducted an empirical evaluation of the presented method against the problem of identification of childhood pneumonia.
- We performed an extensive performance analysis and comparison of the results obtained from conducted experiments.

The remaining of the paper is structured as follows. In Section 2 methods and materials are described, upon which our proposed SGDRE method is designed. The proposed method is explained in detail in Section 3. Section 4 covers the experimental setup and settings, while the results of the conducted experiments are presented and discussed in Section 5. Section 6 concludes our paper by summarizing the presented work and indicating possible directions for future studies.

## 2. Background

In general, the fundamental goal of learning is being capable of generalizing the knowledge learned from training data to the unseen instances (Zhou, 2012). Despite being a complex, non-convex optimization problem with numerous parameters, simple methods such as Stochastic Gradient Descent (SGD) and its variations (Adagrad, RMSprop, Adam) are able to achieve good solutions when training CNN models. The objective landscape in such complex problems has generally multiple minima, all minimizing the training error, but not necessarily all of them generalize well on unseen data. Picking the wrong minimum can lead to poor generalization (Neyshabur, Bhojanapalli, McAllester, & Srebro, 2017). One of the most common approaches to tackle the problem of generalization are ensemble methods. The cornerstone of each ensemble method is the diversity of the models, which can be achieved using various subsets of dataset or by utilizing various classification algorithms (Polikar, 2006). While the CNN based methods have proven to be the most successful methods for the image classification tasks (Simonyan & Zisserman, 2014; Szegedy, et al., 2015), the training process of such methods have high time complexity, therefore training multiple CNN based methods in order to acquire diverse models is even a more time-consuming process. While the strategy of creating an ensemble by utilizing the same type of classification algorithm on different subsets of dataset could be employed, commonly in the field of medical image analysis and CAD we are facing rather small datasets, which makes the utilization of such strategy more difficult.

Based on those grounds, in designing our SGDRE method we adopted an averaging ensemble method to address the generalization issue. For the purpose of collecting a diverse group of models, needed to construct an efficient ensemble, we exploited the SGD with warm restarts (SGDR) capability of exploring a larger part of search space. Exploring a larger part of search space gives us the possibility of visiting more local minima, which consequently gives us a chance to obtain diverse models from each of the visited minima and in that way provide us with a group of diverse models in one single training process. The mechanism in SGDR, which enables the exploration of larger part of search space, is achieved by occasionally restarting the learning rate to the higher value. With such action, the optimizer is trying to escape from the saddle point or local minimum and thus trying to find another, possibly better minimum.

In the following sections, the utilized ensembling method and SGDR mechanism are presented in a greater detail, after providing some brief background of CNNs.

### 2.1. Convolutional neural network

In recent years, the convolutional neural networks (CNN) achieved major breakthrough in various image recognition tasks in broad range of fields from astronomy to the medicine. Although their beginning dates back to the 1980s (Fukushima, 1980; LeCun et al., 1998), the CNNs became the go-to method for classification tasks just few years ago, with the increase of processing power, due to the development of powerful graphical processing units as well as due to the development of supporting libraries enabling the researchers and practitioners to easily adopt such methods.

Regardless of how simple or complex the architecture of CNN seems to be, they all have in common three architectural ideas: local receptive fields, shared weights (or weight replication) and spatial or temporal subsampling, presented in LeCun's work (LeCun et al., 1998), which are ensuring some degree of shift, scale and distortion invariance. Most commonly, the CNN architectures are composed of one or more sequentially connected convolutional and subsampling layers in alternative fashion and at least one fully connected layer (standard multi-layer perceptron layer) at the end.

The convolutional layer computes the convolution operation between the input matrix and a set of learnable filters, also known as kernels. Each filter is sliding across the input matrix in both directions performing the convolution with the local sub-block of input matrix and producing feature planes, also known as feature maps. Most commonly, a rectified linear unit (ReLU) function is applied to each feature map, to improve the computational efficiency and also to reduce the vanishing gradient effect (Morabito, Campolo, Ieracitano, & Mammone, 2019).

The convolutional layer is most commonly followed by a pooling or subsampling layer, which performs a maximum or average subsampling of the previously produced feature maps. The aim of the pooling layer is to achieve shift-invariance, by reducing the resolution of the feature maps and as a side effect, this operation also lowers the computational complexity of training such model (Gu, et al., 2018).

### 2.2. Ensemble methods

The problem of generalization has an impact on more or less any modern machine learning method. Good performance of a trained machine learning model on training data does not necessarily predict good generalization performance, where the generalization performance is defined as the performance of the classifier on data not seen during the training (test data). One of the most common solutions to tackle the problem of generalization is the usage of ensemble methods. A cornerstone of any ensemble method is the diversity of classifiers in an ensemble, where the diversity does not refer to the classifiers' diverse performance, but to their diverse knowledge — i.e., individual classifiers within an ensemble make errors on different training samples (Polikar, 2006). The second key component of any ensemble method is the strategy utilized in combining classifiers. Over the years, the huge amount of various strategies have been developed (Polikar, 2006; Zhou, 2012). One of the most straightforward, popular and fundamental combination strategy is the averaging of numeric outputs, which can be formally expressed as presented in (1):

$$H(x) = \frac{1}{T}\sum_{i=1}^{T} h_i(x) \tag{1}$$

where $H(x)$ represents the combined output of the ensemble, $T$ denotes the number of individual classifiers $\{h_1, \ldots, h_T\}$ and $h_i$ represents the classifiers' output (Zhou, 2012).

### 2.3. Stochastic Gradient Descent with warm restarts

Stochastic Gradient Descent (SGD) (Bottou, 2010) and its various improved variants such as Adam (Kingma & Ba, 2014), RMSprop (Tieleman & Hinton, 2012) and ADAGRAD (Duchi, Hazan, & Singer, 2011) have become in the recent years the de-facto approaches for optimizing different kinds of deep neural networks, mainly due to their ability to avoid and even escape spurious saddle-points and local minima as reported in study by Dauphin, et al. (2014).

The SGD is basically a drastic simplification of Gradient Descent (GD) algorithm proposed by Rumelhart, Hinton, Williams, et al. (1988). Gradient descent is a way to minimize an objective function $J(\theta)$ with model parameters $\theta \in \mathbb{R}^d$ by updating them in the opposite direction of the gradient objective function $\nabla_\theta J(\theta)$ with respect to the parameters (Ruder, 2016). In contrast to the GD where gradient is computed exactly for each sample, the SGD in each iteration estimates gradient on the basis of a single randomly picked example. Since the stochastic algorithm does not need to remember which examples were visited during the previous iterations, it can process examples on the fly, which results in faster computation (Bottou, 2010). In recent years, the majority of state-of-the-art methods utilizes the SGD variation known as mini-batch SGD, which represents a good compromise between GD's low variance in error updates and SGD's low time complexity. The core principle of the mini-batch SGD or mini-batch GD is to calculate the approximate gradient on the small number of random samples — mini-batch, in contrast to SGD where one random sample is picked for estimate gradient calculation and in contrast to GD where the gradient is computed for each sample. As presented by Keskar, Mudigere, Nocedal, Smelyanskiy, and Tang (2016), such approach also tends to avoid sharper minima, due to the computation of gradients is performed in small mini-batches and therefore inexact. The mini-batch GD can be formally expressed as presented in (2) where $\eta$ denotes the learning rate which determines the size of the steps we take to reach a (local) minimum, while $x^{(i)}$ represents a training sample and $y^{(i)}$ the corresponding label (Ruder, 2016).

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i:i+n)}; y^{(i:i+n)}) \tag{2}$$

Two key factors to achieve good convergence are choosing a proper learning rate and adjusting the learning rate during the training, known as the learning rate scheduling. In fact, the most widely used algorithms in the field of deep learning such as Adagrad, RMSprop and Adam are exploiting those key factors in order to achieve the best performance. For example, Adagrad adapts the learning rate of the parameters by applying larger updates to infrequent parameters and smaller updates to frequent parameters; as such, it is well-suited for dealing with sparse data. RMSprop is an extension of Adagrad, which seeks to reduce its aggressive, monotonically decreasing learning rate, dividing the learning rate by an exponentially decaying average of squared gradients. Adam, on the other hand, in addition to storing an exponentially decaying average of past squared gradients as RMSprop, also keeps an exponentially decaying average of past gradients similar to the momentum yielding the Adam update rule (Ruder, 2016).

As proven in recent studies (Huang, et al., 2017; Smith, 2017a), a quite effective method for training deep neural networks, presented by Loshchilov and Hutter (2016) is SGD with warm restarts (SGDR). The basic idea of the mentioned SGDR method is to periodically simulate warm restarts of SGD, where in each restart the learning rate is initialized to some value and is scheduled to decrease. The term "warm" in this context refers to continuing the training process of a deep neural network instead of training from scratch with different learning rate. As expected, the model performance after the increase of learning rate suffers, but only temporarily. Eventually, the performance of the model surpasses the previous one after the learning rate is being annealed. Conventional approaches for annealing of the learning rate dictates a monotonic decrease, while recent works suggest that cycling annealing of the learning rate perturbs the parameters of a converged model, which allows the model to find better local minimum and also more diverse models after each cycle (Huang, et al., 2017; Loshchilov & Hutter, 2016). Such behavior also enables the optimizer to explore a larger part of the search space, and is therefore useful when ensemble building is considered (Huang, et al., 2017; Loshchilov & Hutter, 2016). Using such method could be also beneficial from the time complexity standpoint, as results suggest that SGD with warm restarts requires 2 to 4 times fewer epochs than the commonly used learning rate schedule schemes to achieve comparable or even better results, as demonstrated in a study by Loshchilov and Hutter (2016).

## 3. SGDRE method

The SGDRE method is adopting the averaging ensemble method in order to address the issue of generalization, while for the purpose of obtaining the diverse group of classifiers needed to construct the ensemble, SGDR mechanism is exploited.

The basic concept of our proposed ensemble method is presented in Fig. 1. Essentially, our method consists of four phases and is designed to efficiently work under the limited training budget — limited number of epochs. The first three phases: *Train — initial phase*, *SGD restart 1* and *SGD restart 2* are phases where the training of our CNN model occurs, while the last, fourth phase *Ensembling phase* is the phase where we evaluate and select ensemble candidates and join them into final ensemble for classifying the given chest X-ray image instances.

To obtain the most diverse models as possible, we adopted the SGDR method with combination of two known learning rate annealing functions or learning rate schedules: cosine annealing initially presented by Loshchilov and Hutter (2016) and formally characterized in Eq. (3), and linear decrease (Smith, 2017a) formally characterized in (4). Additionally, we introduce our own developed sine based annealing function, which can be formally expressed in Eq. (5), where $t$ denotes the iteration number, $T$ denotes the number of epochs and $\alpha_0$ denotes the initial learning rate. All three utilized learning rate scheduling functions can be observed on Fig. 2.

$$\alpha(t) = \alpha_0 \cdot (1 + \cos(\frac{t}{T}\pi)) \tag{3}$$

$$\alpha(t) = \alpha_0 \cdot (1 - \frac{t}{T}) \tag{4}$$

$$\alpha(t) = \alpha_0 \cdot (1 - \sin(\frac{\pi \cdot t}{2 \cdot T})) \tag{5}$$

With regard to the limited training budget (a given number of training epochs), the proposed method starts with the *Train — initial phase*, where we train our CNN for a maximum of half of the given epochs utilizing a linear annealing function with initial learning rate set to $1 \cdot 10^{-1}$. At this phase, the early stopping technique is employed in order to stop the training process before consuming all the given budget (epochs) if the model's performance is not improving anymore. In addition, model checkpoint technique during the training process is applied, which stores the best performing model throughout the whole process of training. For the early stopping and model checkpoint technique, we are monitoring the area under the ROC curve (AUC) metric, which is calculated after each training epoch. After the *Train — initial phase* is finished, the mentioned best performing model is obtained and passed to the next training phase. For the next two training phases, *SGD restart 1* and *SGD restart 2*, the remaining training budget is evenly distributed. Since each of the remaining training phases consist of two SGD warm restarts with different learning rates, the remaining training budget is divided by four. The distribution of the training budget over the all three training phases is precisely defined with Eqs. (6)–(8), where $B$ denotes the total training budget in epochs, $B_{init}$ denotes the budget allocated for *Train — initial phase*, $B_{consumed}$ denotes the number of consumed epochs in *Train — initial phase*, while $B_{remainder}$ and $B_{per_start}$ represents the remaining number of training budget in epochs and the number of epochs allocated for each of the SGD restart training respectively.

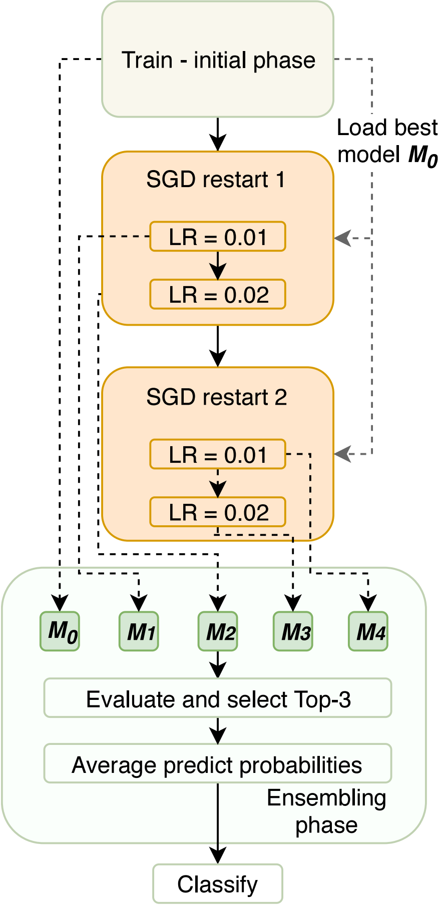

**Fig. 1.** The conceptual diagram of the proposed SGDRE method.

$$B_{init} = \lfloor \frac{B}{2} \rfloor \tag{6}$$

$$B_{remainder} = B - B_{consumed} \tag{7}$$

$$B_{per\_restart} = \lfloor \frac{B_{remainder}}{4} \rfloor \tag{8}$$

Taking the best performing model ($M_0$) from the first *Train — initial phase*, the training is restarted in *SGD restart 1* phase which is composed of two SGD warm restarts, each of which are allocated with the training budget of $B_{per\_restart}$ epochs. The *SGD restart 1* phase is utilizing the SGD optimizer with our own sine based learning rate function for the

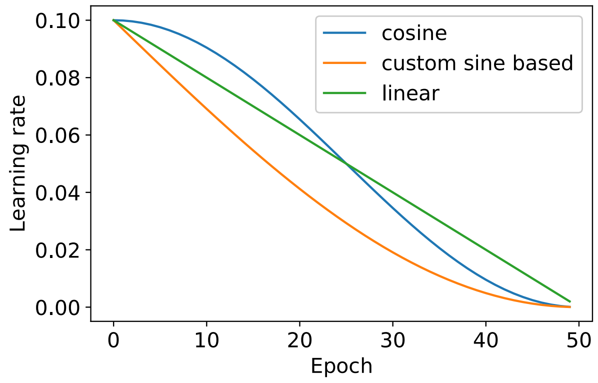

**Fig. 2.** Representation of different learning rate scheduling functions used in training phases of SGRE method.

**Table 1**
The parameter settings for each of the training phases of SGDRE method.

| Phase | Optimizer | Initial LR | Annealing func. |
|---|---|---|---|
| Train — initial phase | SGD | $1 \cdot 10^{-1}$ | linear |
| SGD restart 1 | SGD | $1 \cdot 10^{-2}$<br>$2 \cdot 10^{-2}$ | sine |
| SGD restart 2 | SGD | $10^{-2}$<br>$2 \cdot 10^{-2}$ | cosine |

purpose of decreasing the learning rate through the process of training. The first warm restart is performed with the initial learning rate set to $1 \cdot 10^{-2}$ as can be observed from Table 1, while the second warm restart is performed with initial learning rate set to $2 \cdot 10^{-2}$. The values for the initial learning rates are inferred from the initial study from Loshchilov and Hutter (2016), where the SGDR was first presented.

In the same manner as in the *SGD restart 1* phase, the *SGD restart 2* is performed. The only difference between *SGD restart 1* phase and the *SGD restart 2* phase is the employment of the sine annealing function instead of our custom cosine annealing function along with the SGD optimizer. Different annealing functions in each phases were selected in order to obtain as diverse models as possible. The usage of different annealing functions gives us the possibility to converge to the different minima as the learning rate is decreased by different schedules.

After each warm restart in the phases *SGD restart 1* and *SGD restart 2* is completed, the snapshot of the model is taken, therefore in *SGD restart 1* phase the $M_1$ and $M_2$ models are obtained, while in the *SGD restart 2* the $M_3$ and $M_4$ models are obtained as presented in Fig. 1.

The last, fourth phase is the *Ensembling phase*, where the obtained models ($M_0, M_1, M_2, M_3, M_4$) from the previously completed training phases are evaluated. The evaluation of the collected models is performed against the evaluation subset of the training set, observing the AUC metric. The best three performing models (models with the highest AUC value) are selected to construct the final ensemble. For the ensembling method, the model averaging was selected, where the prediction probabilities from each of the top-3 selected models are being averaged, and the resulting average value is then being used for classification as the final ensemble prediction probability.

From Fig. 3 we can observe the SGDRE training process in comparison to the more conventional training approach utilizing the training of a single CNN with the Adam optimizer (the baseline method $BaselineADAM$, presented later, is used here). If we observe the SGD restart lines, it is clearly shown that after each warm restart and increase of learning rate, the training accuracy initially drops and is then being improved towards the end of each SGD restart phase as expected. Focusing on the marked obtained models ($M_0, M_1, M_2, M_3, M_4$), we can see that none of them is achieving as high training accuracy as the $BaselineADAM$. Such behavior is expected, since the purpose of our SGDRE method is not to obtain the best performing model overall, but instead to obtain a group of diverse models, each with a bit of different knowledge, which will then be used to construct an ensemble and (hopefully) performed better.

## 4. Experiments

To properly evaluate the performance and efficiency of our proposed SGDRE method, we conducted experiments comparing our method to the baseline method (Stephen et al., 2019) utilizing the same CNN architecture and Adam optimizer, which has already proven to deliver great results on the task of identifying the pneumonia from chest X-ray images. Additionally, we conducted experiments based on the baseline method utilizing the SGD optimizer function in order to observe the behavior and the performance of the SGD optimizer function utilized using the baseline CNN architecture without any learning rate scheduling strategy being applied. Comparing the performance of the latter experiment also enables us to explore the performance capabilities of various learning rate strategies applied in our proposed method.

In the following sections, the dataset, data pre-processing, experimental settings, evaluation methods and used metrics are presented.

The whole experimental environment as well as the proposed SGDRE method was implemented in Python programming language. For the construction and training of CNNs, Keras (Chollet et al., 2015) framework with Tensorflow (Abadi et al., 2015) back-end was used. Various supportive routines were implemented with the help of the following external libraries: Numpy (Van Der Walt, Colbert, & Varoquaux, 2011), Pandas (McKinney, 2010) and scikit-learn (Pedregosa, et al., 2011).

All the conducted experiments were performed utilizing the Intel Core i7-6700K quad-core CPU running at 4 GHz clock speed with 64 GB of DDR4 memory, and three Nvidia GeForce Titan X Pascal GPUs each with 12 GB of GDDR5 memory, running the Linux Mint 19 operating system.

### 4.1. Chest X-ray images (Pneumonia) dataset

The Chest X-ray images dataset is publicly available dataset (Kermany & Goldbaum, 2018), originally collected and presented by Kermany, et al. (2018). The dataset consists in total of 5858 labeled chest X-ray images from pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. Out of all images, 4274 are characterized as depicting pneumonia and 1,584 as normal. The image samples of the used dataset are presented in Fig. 4. The collected image dataset in composed of images of various sizes, stored in the JPEG file format. All chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. Two expert physicians were employed for the task of grading the diagnoses of chest X-ray images before being cleared for training the CNN (Kermany, et al., 2018).

### 4.2. Dataset preprocessing

The dataset originally contains images of various sizes, which we have resized to the unified size of 200 × 200 px. For the baseline experiments, we have also applied several augmentation methods in accordance to Stephen et al. (2019) to artificially increase the number and quality of image instances in the dataset. The usage of various augmentation techniques and parameter settings for utilized techniques are presented in Table 2.

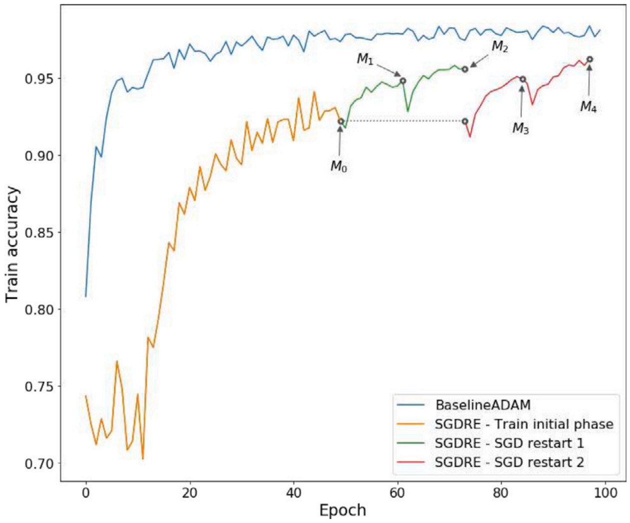

**Fig. 3.** Visualization of SGDRE training process in comparison to the conventional training of $BaselineADAM$. With $M$ are marked stored checkpoint models used for construction of ensemble.

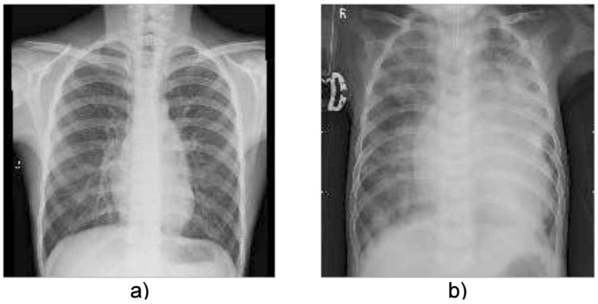

**Fig. 4.** Two image samples from Chest X-ray images dataset: (a) represents X-ray image of normal lungs, (b) represent X-ray image of pneumonia lungs.

**Table 2**
Utilized image augmentation techniques with corresponding values used for the purpose of generating augmented training samples.

| Technique | Setting |
|---|---|
| Rescale | 1/255 |
| Rotation range | 40° |
| Height shift | 0.2% |
| Width Shift | 0.2% |
| Shear range | 0.2% |
| Zoom range | 0.2% |
| Horizontal flip | $True$ |

**Table 3**
Utilized CNN architecture layers with the corresponding settings and output shape. For the convolutional layers the settings represent the kernel size and number of kernels used, while for the maximization pooling layers the settings denotes the window size.

| Layer | Settings | Output shape |
|---|---|---|
| Convolution | $3 \times 3, 32$ | (None, 198, 198, 32) |
| Max pooling | 2 × 2 | (None, 99, 99, 32) |
| Convolution | $3 \times 3, 64$ | (None, 97, 97, 64) |
| Max pooling | 2 × 2 | (None, 48, 48, 64) |
| Convolution | $3 \times 3, 128$ | (None, 46, 46, 128) |
| Max pooling | 2 × 2 | (None, 23, 23, 128) |
| Convolution | $3 \times 3, 128$ | (None, 21, 21, 128) |
| Max pooling | 2 × 2 | (None, 10, 10, 128) |
| Flatten | – | (None, 12800) |
| Dropout | probability 0.5 | (None, 12800) |
| Dense | 512 units, ReLU | (None, 512) |
| Dense | 2 units, Softmax | (None, 2) |

**Table 4**
Parameter settings for baseline (with Adam optimizer), baseline SGD and SGDRE methods.

| Parameter | Baseline | Baseline SGD | SGDRE |
|---|---|---|---|
| No. of epochs | 100 | 100 | 100 |
| Batch size | 32 | 32 | 32 |
| Optimizer | Adam | SGD | SGD |
| Learning rate | $1 \cdot 10^{-3}$ | $1 \cdot 10^{-1}$ | $1 \cdot 10^{-1}$ |

### 4.3. Baseline CNN

For the baseline CNN, we adopted the architecture specifically developed for the task of identifying the pneumonia, which was initially presented in a research study by Stephen et al. (2019). The detailed listing of utilized CNN architecture layers is presented in Table 3. The architecture (see Fig. 5) is composed of four sequentially connected pairs of convolutional and maximization pooling layers, followed by a flatten layer, dropout layer and two fully connected layers. The convolutional layers are comprised of the kernels with size 3 × 3 with ReLU activation function and connected to the maximization pooling layers with kernel size 2 × 2. The first convolutional layer employs 32 filters, second 64 filters, while third and fourth convolutional layers employs 128 filters. Last maximization pooling layer is connected to the flatten layer, which converts the extracted 2-dimensional feature planes into 1-dimensional feature vector, which is classified using dropout layer with a dropout probability set to 0.5 and two fully connected (dense) layers of size 512 and 1, respectively. The last dense layer is utilizing the Sigmoid activation function, which performs the classification task.

### 4.4. Parameter settings

In Table 4 the parameter setting for training the baseline and SGDRE methods are presented. As can be observed from the table, both methods share the parameter settings for number of epochs and batch size, while the optimizer function used in Baseline method is Adam, while on the other hand, Baseline SGD and proposed SGDRE method are utilizing the conventional SGD optimizer. The selected learning rate value for the Baseline method with Adam optimizer is set based on its initial research study by Kingma and Ba (2014), where the authors proposed that the good default setting for learning rate is $1 \cdot 10^{-3}$. The learning rate value for the other two methods was set based on the research study by Loshchilov and Hutter (2016) where the SGDR method was initially presented.

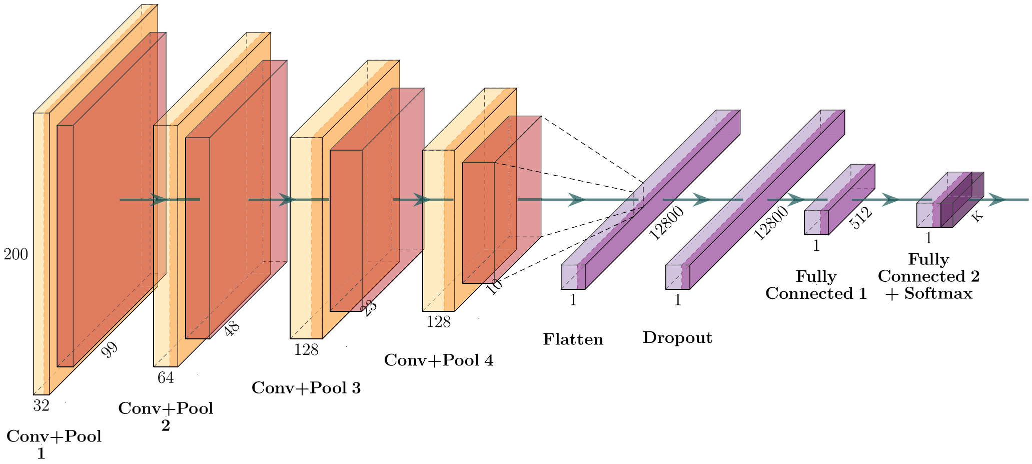

**Fig. 5.** The graphical representation of utilized CNN architecture, initially proposed by Stephen et al. (2019).

### 4.5. Evaluation method and metrics

With the presented experimental settings and methods' parameter setting, we conducted three experiments. First experiment was conducted utilizing the known successful method from Stephen et al. (2019) reported as *BaselineADAM*. In the second experiment, reported as *BaselineSGD*, we adopted the baseline method, but we trained it with the basic SGD optimizer instead of Adam optimizer. The third experiment was conducted with our proposed SGDRE method. All of the experiments are run against Chest X-ray images dataset, tackling the task of classifying the normal and pneumonia images.

To objectively evaluate the performance of each method, we adopted a well-established 10-fold cross-validation methodology, where the initial dataset is evenly divided into 10 subsets (folds). Nine of those folds are then used for the training phase (training set) and the remaining one for the performance testing (test set) of the trained model. In the same manner, the process is repeated in total 10 times, each time leaving different fold out for testing. Due to the nature of our proposed method, we had to further split our training set even further in ratio 80:20, as presented in Fig. 6, in order to perform the evaluation and selection of our models, and then test the performance of the ensemble on the test set.

In each repetition of 10-fold cross-validation methodology, we calculated the following metrics: general accuracy, AUC, precision, recall/sensitivity, specificity, F-1 measure, and Cohen's kappa coefficient. Given the adaptive mechanisms in SGDRE method, at the train time, we have also recorded how many epochs it consumed and how much time was needed to complete each fold.

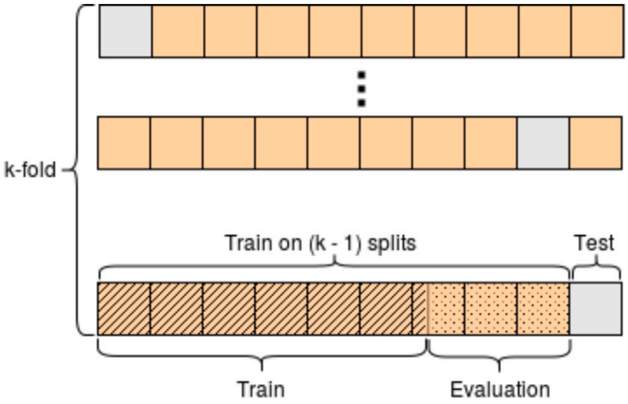

**Fig. 6.** Conceptual overview of the $k$-fold cross-validation methodology adopted in our work. The $k$ denotes the number of folds, in our case $k$ is set to 10.

## 5. Results

In this section, we present the obtained experimental results, in accordance with the defined research questions, we investigated in our study:

- RQ 1: Can the classification performance of childhood pneumonia based on chest X-ray images be improved with our proposed method SGDRE?
  - RQ 1.1: What is the classification performance of SGDRE, and how does it compare with existing methods utilizing CNNs?
  - RQ 1.2: What is the influence of SGDRE on generalization of learned models?
  - RQ 1.3: Can SGDRE improve classification performance without enlarging the initial dataset?
  - RQ 1.4: Can SGDRE provide an ensemble of several classification models without increasing the time-complexity of training?

In order to evaluate our proposed method objectively, we compared it with two baseline methods — *BaselineADAM* (a method presented by Stephen et al. (2019) that uses Adam as an optimization function for training the CNN model) and *BaselineSGD* (a method that uses the same CNN architecture but SGD as an optimization function instead of Adam). All three compared methods use the same CNN architecture as proposed in Stephen et al. (2019). In this manner, the differences among the obtained predictive performance results can be contributed solely to the consequence of different learning method used. For the sake of comparison, we performed a series of experiments on the Chest X-ray dataset using the 10-fold cross-validation approach.

Results, obtained from the conducted experiments, are summarized in Table 5. As can be observed from the table, our proposed SGDRE method showed the best performance among the three compared methods, regardless of the selected compared metric. In general, the second-best results were obtained by the *BaselineADAM* method, while the worst results were obtained by the *BaselineSGD* method.

### 5.1. An in-depth analysis of classification results from the three compared methods

#### 5.1.1. General accuracy

First, we compared the obtained accuracy results. The accuracy metric is the most general metric in classification, expressing a share of

**Table 5**
Comparison of average accuracies, AUCs, $F$-1 scores, precisions, recalls, kappa coefficient, consumed numbers of epochs and time (in seconds) with standard deviations over 10-fold cross-validation.

| Metrics | BaselineADAM | BaselineSGD | SGDRE |
|---|---|---|---|
| Accuracy | 92.81 ± 1.48 | 90.97 ± 6.51 | **96.26 ± 0.94** |
| AUC | 91.98 ± 3.14 | 87.60 ± 13.33 | **95.15 ± 1.32** |
| Precision | 96.38 ± 2.69 | 93.71 ± 7.50 | **97.32 ± 0.79** |
| Rec./Sens. | 93.80 ± 2.66 | 94.92 ± 3.65 | **97.57 ± 0.93** |
| Specificity | 90.15 ± 8.01 | 80.28 ± 28.69 | **92.74 ± 2.20** |
| $F$-1 | 95.01 ± 1.04 | 94.05 ± 3.59 | **97.44 ± 0.64** |
| Kappa | 0.82 ± 0.04 | 0.74 ± 0.26 | **0.91 ± 0.02** |
| Epoch | 100.00 ± 0.00 | 100.00 ± 0.00 | **98.10** ± 0.02 |
| Time | 528.70 ± 15.75 | 490.10 ± 12.78 | **440.30** ± 9.59 |

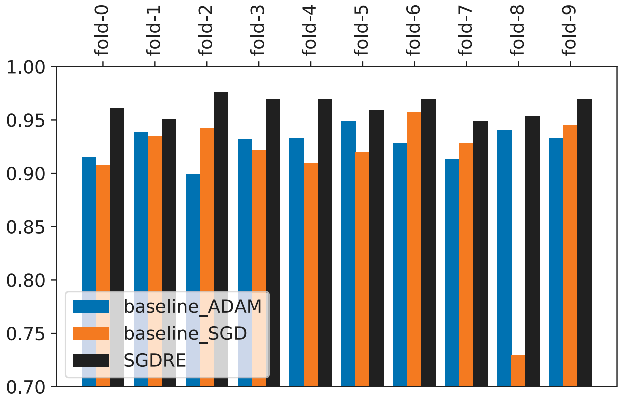

**Fig. 7.** Accuracy of compared methods on 10 folds.

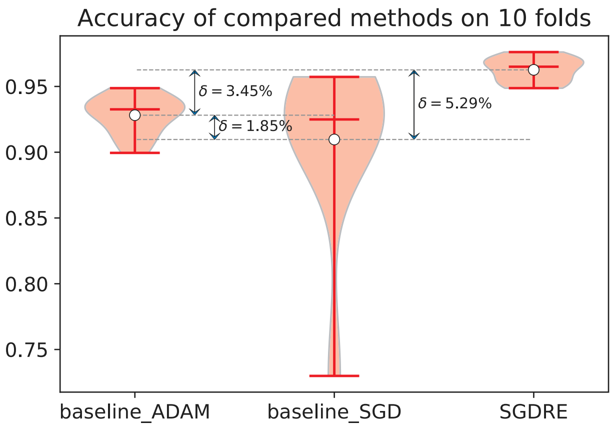

**Fig. 8.** Violin plots presenting the accuracy of the compared three methods.

instances among all, which have been correctly classified by a classifier model. Fig. 7 shows a comparison of test accuracy results obtained by the three compared methods for all 10 folds on the Chest X-ray dataset. As we can see, the SGDRE method achieved the highest accuracy in all 10 folds, while the results of two baseline methods alternate with the exception of fold-8, in which the $BaselineSGD$ seems not to converge well.

To evaluate the statistical significance of these results, we first applied the Friedman test as suggested by Demsar (2006) by calculating the asymptotic significance for the three compared methods on all 10 folds. As the results are not normally distributed, the Friedman test was applied, which is a non-parametric statistical test used to detect differences in the results of various methods across multiple test attempts. The results of the performed Friedman test with regard to accuracy show that differences between the three methods are statistically significant ($p < 0.001$).

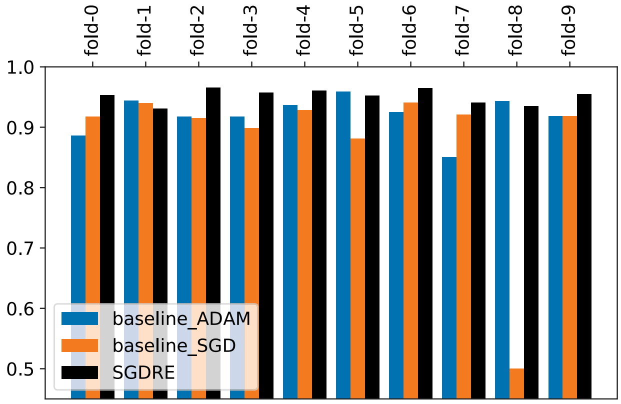

**Fig. 9.** AUC of compared methods on 10 folds.

To test further whether the results of our proposed method SGDRE, which obtained the highest average accuracy, are indeed significant, the Wilcoxon signed-rank test was applied next, as suggested by Demsar (2006), to compare SGDRE with both baseline methods (the Holm–Bonferroni correction was applied). The Wilcoxon signed rank test is a non-parametric alternative to the paired T-Test, which can be used to compare the statistical equality of two methods over the same sample. It tests whether the difference of achieved ranks of the two methods is statistically significant (Moore, McCabe, & Craig, 2009). In our case, the Wilcoxon test was used to compare the results, achieved by our proposed method SGDRE, with both baseline methods $BaselineADAM$ and $Baseline - SGD$, one by one. If the Wilcoxon test resulted in a statistically significant difference between the two methods, the one with the higher average rank can be regarded as the better method. The method SGDRE achieved the best average rank among three methods ($rank = 3.0$ on the scale from 1 being the worst result up to 3 being the perfect score), and it turned out that the results of SGDRE are indeed significantly better than the results from the other two methods (in both cases, when compared either to $BaselinADAM$ and $BaselineSGD$, the resulting $p < 0.001$).

The domination of our SGDRE with regard to achieved accuracy can be easily observed also on a violin plot of the comparison of the three methods (see Fig. 8). Beside having the highest average accuracy, the SGDRE's accuracy results across 10 folds also had by far the lowest standard deviation, demonstrating the method's good generalization capability.

### 5.1.2. AUC

As in the case of accuracy, the analyses of other predictive performance metrics have been performed similarly. In this manner, Fig. 9 shows a comparison of AUC (Area Under Curve) results obtained by the three compared methods on test data for all 10 folds. Beside accuracy, AUC is one of the most important metric to consider when evaluating a classification model, especially in medicine. The AUC represents the probability that a classifier will rank a randomly chosen positive instance (a patient with pneumonia in our case) higher (i.e. with greater suspicion) than a randomly chosen negative one (a patient who does not have a pneumonia) (Hanley & McNeil, 1982). AUC has the attractive property that it side-steps the need to specify the costs of the different kinds of misclassification (Hand & Till, 2001).

We can see that SGDRE achieved the best AUC results in 7 out of 10 folds, while the remaining three wins were achieved by the $BaselineADAM$ method. Again, the SGDRE method achieved the most similar AUC results in all 10 folds, while the AUC results of the two baseline methods vary more.

The performed Friedman test revealed that the differences among the three methods are statistically significant ($p = 0.045$). The highest

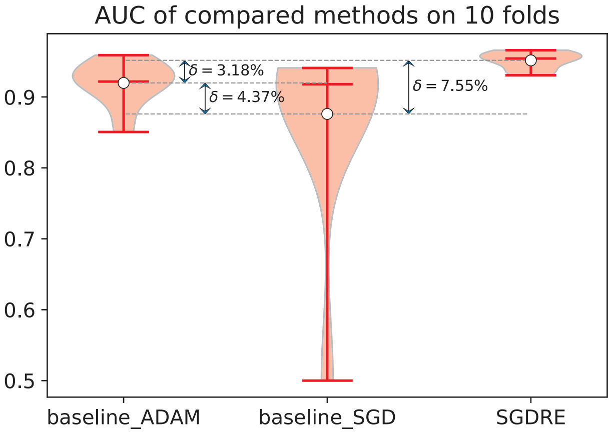

**Fig. 10.** Violin plots presenting the AUC of the compared three methods.

rank were achieved by the SGDRE ($rank$ = 2.6), following by the $BaselineADAM$ ($rank$ = 1.9) and $BaselineSGD$ ($rank$ = 1.5). The performed Wilcoxon test confirmed significant dominance of SGDRE ($p$ = 0.0069 when compared to $BaselineSGD$, and $p$ = 0.0284 when compared to $BaselineADAM$).

Fig. 10 shows the distribution density of the AUC values, the minimum, maximum, mean and median values for each of the three compared methods, as well as the $\delta$ difference between AUC mean values. As we can see, again, the SGDRE not only outperformed both compared methods, but also has the smallest standard deviation. The small variation of AUC results across different folds, achieved by our SGDRE method, complies with the presumption that our method will generalize better, by being able to obtain diverse models due to the exploitation of the SGDR capabilities.

### 5.1.3. Precision, recall/sensitivity, specificity, and $F$-1 measure

Similarly, we also performed analyses of precision, recall (also named sensitivity), specificity, and $F$-1 measure (see Fig. 11).

Precision can be defined as the fraction of relevant instances (patients of certain class — i.e. having pneumonia or not in our case) among the retrieved instances (patients, classified to a certain class by a classification model). Recall, on the other hand, can be defined as the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are usually considered as a pair, as the improvement of one generally decrease the other and vice versa. We can see, that or proposed SGDRE method once again achieved the most stable results (with the lowest variance) with the highest average value for both two metrics.

Similar to precision and recall, also sensitivity and specificity metrics are usually observed as a pair. Sensitivity is a metric, equal to recall, that measures the proportion of actual positive instances (patients with actual pneumonia) that are correctly identified as such (patients classified as having pneumonia). Specificity, on the other hand, measures the proportion of actual negative instances (patients not having pneumonia) that are correctly identified as such (patients classified as not having pneumonia). Like it was the case already with recall (sensitivity), also with regard to specificity our proposed SGDRE metric performed the best on average with (substantially) lower deviation. The improvement of SGDRE with regard to the two baselines are particularly high in the case of specificity, indicating the better generalization of SGDRE that is able to classify also negative instances much more correctly (note that the amount of negative instances within the dataset is much lower). All these can be well observed on the violin plots of these metrics (see Fig. 12).

The performed Friedman tests revealed that differences among the three methods are statistically not significant in the case of precision ($p$ = 0.1225) and specificity ($p$ = 0.1905). However, in both cases SGDRE achieved the highest average rank (in case of precision SGDRE achieved $rank$ = 2.4, $BaselineADAM$ $rank$ = 2.1 and $BaselineSGD$ $rank$ = 1.5; in case of specificity SGDRE achieved $rank$ = 2.3, followed by $BaselineADAM$ with $rank$ = 2.15 and $BaselineSGD$ with $rank$ = 1.55). In the case of recall/sensitivity, however, the differences turned out to be significant (Friedman test: $p = 0.0208$), with SGDRE being the best method (Wilcoxon test: $p = 0.0093$ when compared to $BaselineADAM$ and $p = 0.0367$ when compared to $BaselineSGD$).

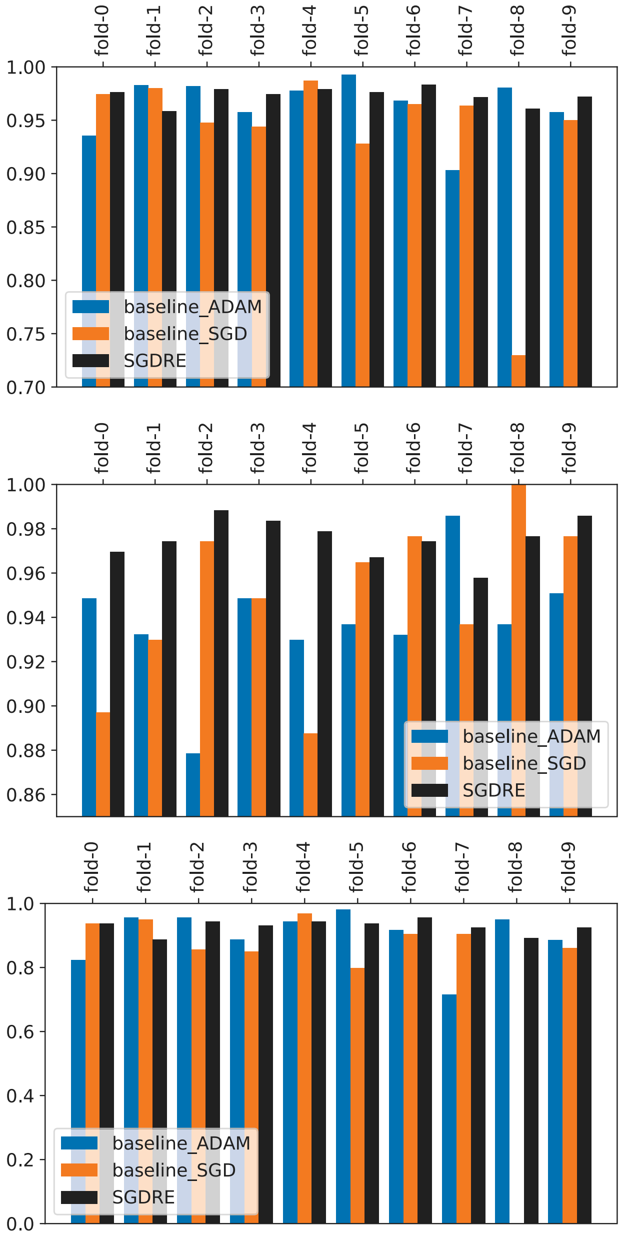

**Fig. 11.** Precision (top), recall/sensitivity (middle) and specificity (bottom) of compared methods on 10 folds.

The $F$-1 measure can be interpreted as a harmonic mean of the precision and recall. It is considered as a kind of substitution for general accuracy and is preferred to accuracy, especially in cases of imbalanced datasets. In our case, the obtained results of $F$-1 measure were very similar to accuracy, with the statistical evaluation being completely the same (the ranks and the results of Friedman and Wilcoxon test).

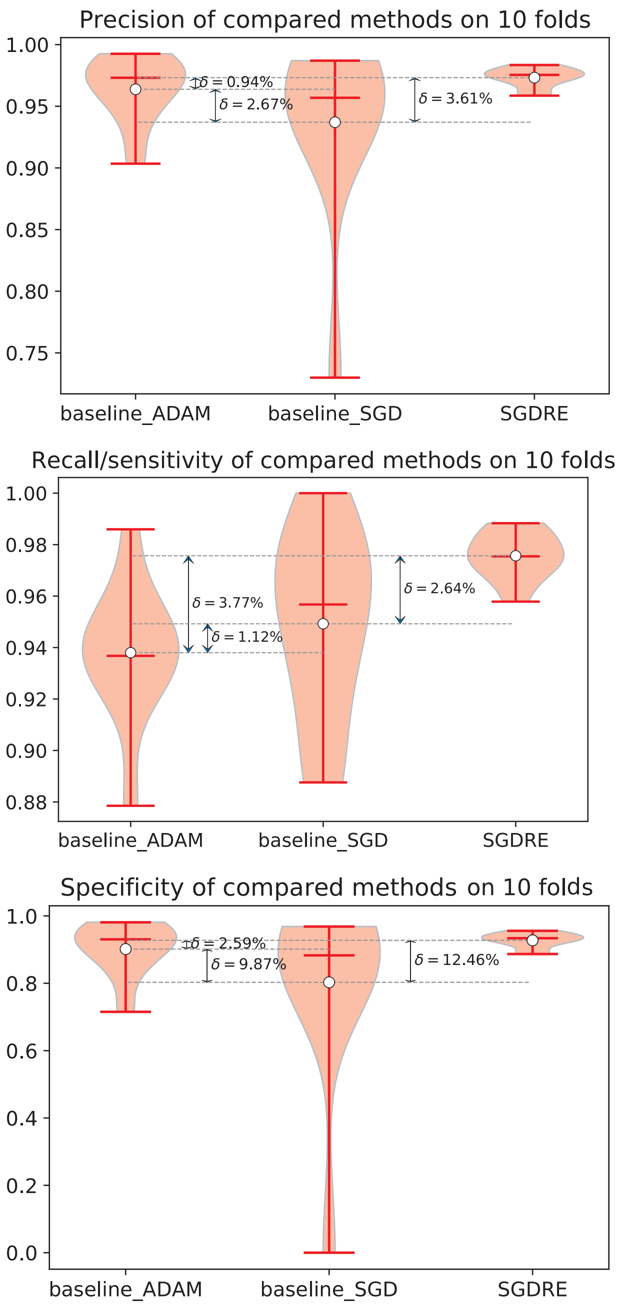

**Fig. 12.** Comparison of precision (top), recall/sensitivity (middle) and specificity (bottom) for the three methods.

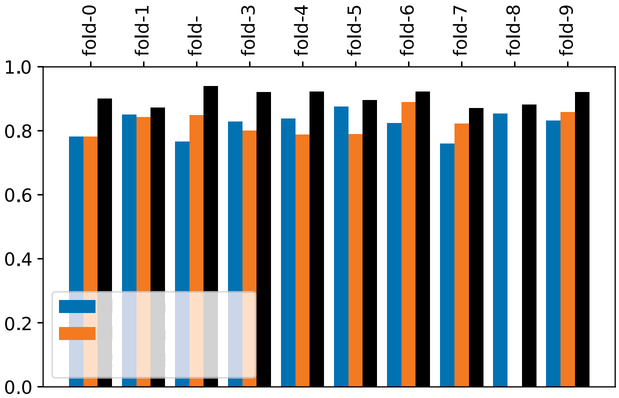

**Fig. 13.** Cohen's kappa of compared methods on 10 folds.

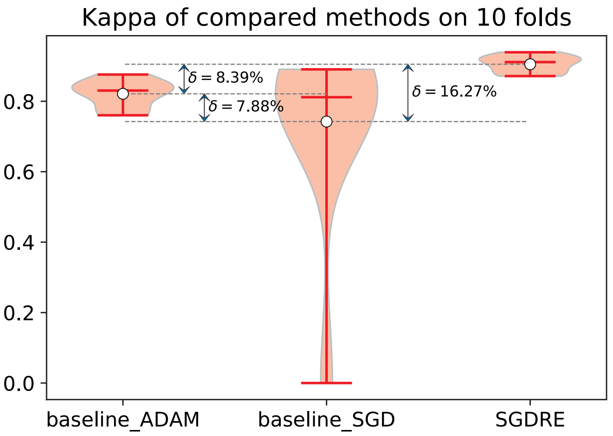

**Fig. 14.** Violin plots presenting the Cohen's kappa values for the compared three methods.

### 5.1.4. Cohen's kappa

Fig. 13 shows a comparison of Cohen's kappa results obtained by the three compared methods on test data for all 10 folds. Cohen's kappa is a metric that compares an observed accuracy with an expected accuracy (random chance). It takes into account random chance, which generally means that it is less misleading than simply using accuracy as a metric. When comparing different classification models, an interesting aspect to be considered is the interpretation of magnitude, although existing magnitude guidelines are not universally accepted. Thus, Landis and Koch (1977) characterized values <0 as indicating no agreement, 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement. Similarly, Fleiss, Levin, and Paik (2003) equally arbitrary guidelines characterize kappas $> 0.75$ as excellent, 0.40–0.75 as fair to good, and $< 0.40$ as poor. We can see that SGDRE achieved the highest kappa in all 10 folds, with the lowest variance in results (see also Fig. 14). In each fold, the SGDRE's kappa value was at least 0.87.

The performed Friedman test revealed that the differences among the three methods are statistically significant ($p < 0.001$). The highest average rank were achieved by the SGDRE method ($rank = 3.0$), with the two baselines following (both achieved $rank = 1.5$). Also, the performed Wilcoxon test confirmed significant dominance of SGDRE ($p = 0.0051$ when compared to both baselines).

### 5.1.5. Training time and epochs

Finally, Fig. 15 shows a comparison of total training time, used to train all the models for all ten folds. Please note, that the total training time for the SGDRE method includes the training of all single CNNs, which are needed to construct an ensemble, while the two baseline methods each trains only a single model for each fold. Notwithstanding this fact, we can see that SGDRE needed the least amount of time to train all the models, being the fastest in each single fold as well, followed by the $BaselineSGD$ method, while $BaselineADAM$ needed the most time in each of 10 folds (see also Fig. 16).

The performed Friedman test confirmed the significance of differences among the three methods ($p < 0.001$). The average ranks were: SGDRE ($rank = 3.0$), $BaselineSGD$ ($rank = 2.0$), and $BaselineADAM$ ($rank = 1.0$). Performed Wilcoxon test confirmed significant dominance of SGDRE ($p = 0.005$ in each pair-wise comparison).

Focusing on the values of consumed epochs, we can observe that the two baseline methods consumed the maximum number of epochs available for training while the SGDRE method, on average, consumed 98 epochs. Since for all compared methods, the number of epochs at the

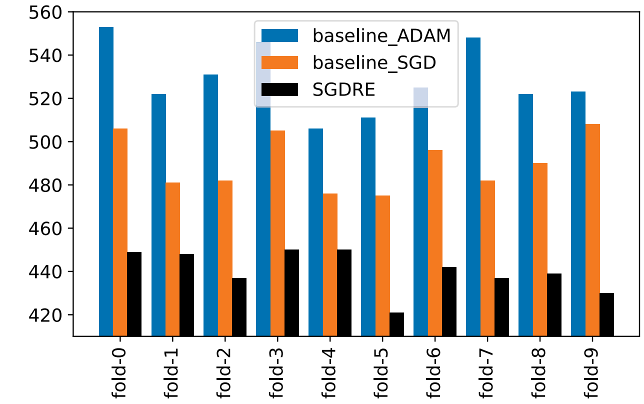

**Fig. 15.** Total time, used to train the models for all 10 folds (in seconds).

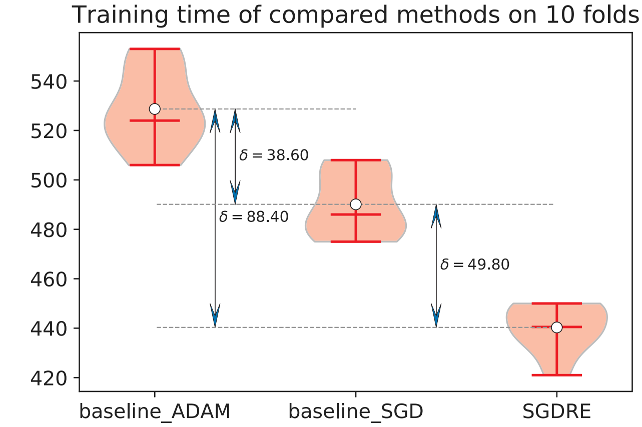

**Fig. 16.** Violin plots presenting the total time used to train the models for the compared three methods.

beginning of training was set to 100, the lower number of consumed epochs for the SGDRE method is due to the early stopping technique employed in the initial training phase.

### 5.2. Worst-case analysis of classification metrics

In the previous section, we were focusing on best average values of compared methods, and also presented the median values as well as the standard deviations for each metric. However, in many specific domains such as medicine, where the poor classifier's performance could lead to catastrophic consequences, it is an imperative to also evaluate the worst-case performance of classifiers. Therefore, we have compared the worst values of classification metrics for the three compared methods obtained from conducted 10-fold cross-validation. The values for each metric of compared classifiers are presented in a form of radar chart in Fig. 17.

Observing the radar chart, we can see that for each classification metric, the proposed SGDRE method achieved the highest score, while the worst performing method is $Baseline - SGD$, which in the case of specificity and kappa value achieved value 0. The remaining $BaselineADAM$ method is noticeably better in the aspects of each classification metric, while also lagging behind the proposed SGDRE method quite a bit.

### 5.3. Analysis of commonly misclassified samples

In addition to the statistical analysis of the obtained results and worst-case analysis, we also conducted a more qualitative analysis of some common misclassified samples. Fig. 18 represents three chest X-ray samples which were misclassified. The top two samples are in the dataset labeled as "normal" while the SGDRE method classified them as "pneumonia". On the other hand, the bottom one is in a dataset labeled as "pneumonia" while the SGDRE method classified it as "normal". For the presented samples, we cross-validated the pathology prediction with the utilization of Chester (Cohen, Bertin, & Frappier, 2019) application, developed by Machine Learning and Medicine Lab, and Mila - Quebec AI institute. For each of the three presented cases, the Chester tool predicted the highest probability for lung opacity pathology (which should not be considered as a diagnosis, but rather as supportive information for a medical expert). The blue regions on each sample indicate the areas of chest X-ray images which have the highest impact on predicted pathology, while the transparent pixels have negligible impact on the prediction. The lung opacity, from the medical domain standpoint, refers to any area that preferentially attenuates the X-ray beam and therefore appears more opaque than the surrounding area. In that sense, it is a nonspecific term that does not indicate the size or the pathologic nature of the abnormality (Goodman, 2014). Lung opacity is commonly present with pneumonia, since the body's immune response fills the sacks in the lungs (termed alveoli) with fluids instead with air. However, the diagnosis of the pneumonia is, in general, based on the chest X-ray image combined with additional medical data and lab results. When inspecting only the chest X-ray image, there might be images which look similar to pneumonia, but with different diagnosis. Therefore, such cases as presented in Fig. 18 are challenging to classify, since a machine learning method can only make a prediction based on the given chest X-ray image.

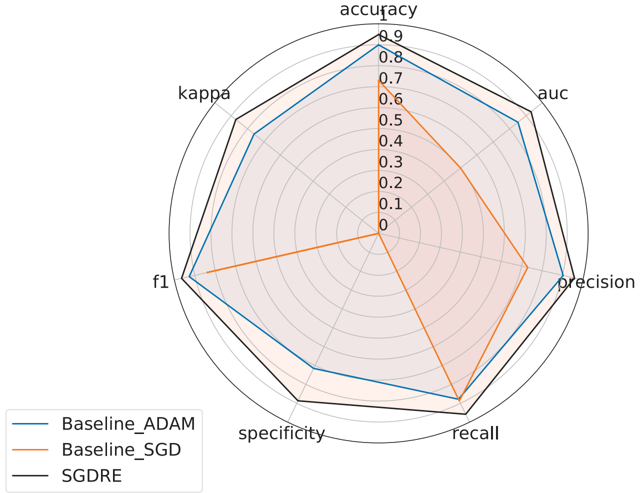

**Fig. 17.** Radar chart presenting worst-case values of classification metrics for the compared three methods.

### 5.4. Comparison of SGDRE to some other published results

There are some publications, in which the authors have used the same Chest X-ray dataset to classify the childhood pneumonia. The work that we based our proposed method SGDRE on is Stephen et al. (2019), where the authors report the accuracy of 93.7%, which is inline with the results, we obtained when re-implementing the method and using it as $BaselineADAM$ in our experiments (the obtained average accuracy was 92.8%; the somewhat lower score than originally reported can be contributed to the fact that we used the 10-fold cross validation methodology instead of a simpler split approach).

Kermany, et al. (2018) presented a method that exploited the transfer learning approach utilizing the Inception V3 CNN architecture — the authors report the accuracy of 92.8%, lagging behind SGDRE by a

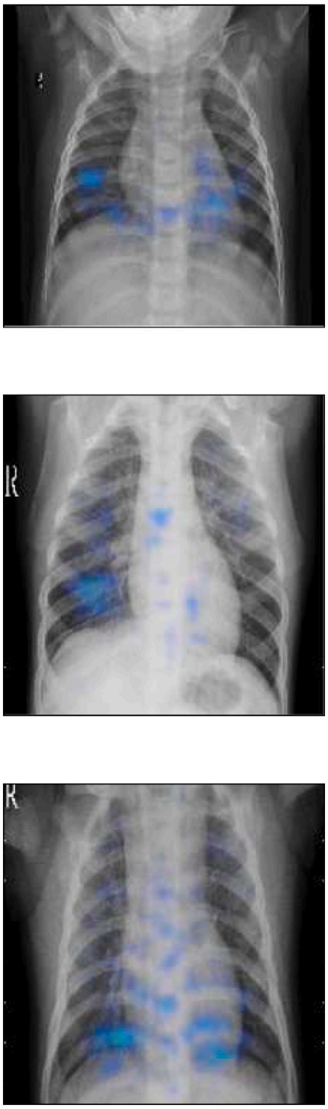

**Fig. 18.** Misclassified examples of chest X-ray images.

margin of 3.46%. In another recently published work by Siddiqi (2019), the authors presented a custom sequential CNN architecture, achieving the accuracy of 94.39% which is an improvement over the method from Kermany, et al. (2018), but still lags behind the SGDRE method by a margin of 1.87%.

Kermany, et al. (2018) reported also sensitivity (93.2%) and specificity (90.1%), which both lag behind SGDRE's results (sensitivity = 97.6%, specificity = 92.7%). On the other hand, the reported AUC of 96.8% seems to be a bit better than SGDRE's (AUC = 95.2%).

Siddiqi (2019) reported beside accuracy also sensitivity (99.0%) and specificity (86.0%). We can see a big difference among the two values, indicating the bias of the trained model toward correctly classifying positive subjects (children having pneumonia). In this manner, the SGDRE lag slightly behind with regard to sensitivity (by a margin of 1.4%), but having a substantial advantage with regard to specificity (by a margin of 6.7%), indicating a much better balance of SGDRE.

### 5.5. Answering the research questions

After inspecting all the results, we can say that our proposed SGDRE method improved the classification performance of childhood pneumonia from Chest X-ray images. All the observed performance metrics were better than the two compared baseline methods, most of the improvements, with exception of precision and specificity, which however achieved the highest average rank, were also statistically significant. A comparison of obtained results with published works, who reported the classification results upon the same problem, showed that SGDRE can indeed be considered as a very competitive classification method.

Using the SGDRE method, which exploits the SGDR's capabilities of exploring a larger part of a search space and therefore visits more minima, we were able to obtain a more diverse group of models, from which we were able to construct an ensemble using the averaging strategy. Based on the achieved the smallest standard deviation for all of the observed performance metrics as well as their highest average values, we can infer that the models obtained and combined in such manner are achieving a high level of generalization.

Observing the classification performance results, achieved by compared methods, we can see that the SGDRE method outperformed the two compared methods. In contrast to SGDRE, the compared baseline methods are in fact utilizing the dataset augmentation technique, which artificially enlarges the initial dataset and therefore the SGDRE did improve the classification performance without enlarging the initial dataset.

Based on the reported empirical results, we can confirm that the SGDRE method is capable of providing a group of diverse classification models in a single training process, consuming even less time as the compared baseline methods, while delivering overall better classification performance.

### 5.6. Threats to validity

When utilizing the machine learning methods and techniques, the validity threats most commonly relate to the diversity, quality and quantity of the data. Therefore, it is crucial how the data used for training is gathered, pre-processed and labeled or pre-classified. In our case, we used the Chest X-ray images (Pneumonia) dataset, which was initially screened for quality control by removing all low quality or unreadable scans. Afterwards, two expert physicians were employed for the task of grading the diagnoses of chest X-ray images before being cleared for using for the purpose of CNN training. Nevertheless, our obtained results and findings may not be generalized to all specific situations since the used dataset may not be representative.

Additionally, splitting the data into training and test set could also be a potential threat to validity, since the methods may over-fit and bias the results. Therefore, instead of splitting the dataset randomly using the simple train-test split (70% or 80% for training and 30% or 20% for testing), we adopted a well-known 10-fold cross-validation procedure in which the initial dataset is split into 10 subsets, 9 of which are used for training while the remaining set is used for the evaluation. The process was repeated in total for 10 times, each time evaluating the method against a different subset.

In this study, we utilized deep CNN and SGD with warm restarts for the task of identifying pneumonia from chest X-ray images. While the proposed SGDRE method delivered promising results for the given task and could be, in practice, easily adapted to different image classification domain problems, the reported performance could not be generalized to different datasets. Therefore, extending the research to different domain problems might be part of our future work.

## 6. Conclusion

In this paper, we presented a new SGDRE method, an ensemble method based on stochastic gradient descent with warm restarts mechanism for the identification of childhood pneumonia from chest X-ray images. The method exploits the averaging ensemble method and SGDR capabilities to converge and escape from the local minima with warm restarts of the learning rate, which enables the method to visit several local minima and therefore to obtain a more diverse group of models, needed for the construction of a well performing ensemble. During the training phases, the SGDRE method employs three different learning rate schedules, one of which is the new custom sine based schedule first presented in this work. Employment of different learning rate annealing functions enables our method to obtain even more diverse models in the training phase as the learning rates are decreased in different manner. From three training phases, five CNN models are obtained, which are evaluated against the evaluation subset in the fourth, ensemble construction phase. The top-3 best performing models are then combined using the averaging ensemble strategy, which outputs the final classification prediction of the SGDRE method.

Through the conducted experiments, the presented SGDRE method has proven to perform well under the limited training budget. The SGDRE method outperformed the compared methods in all the measured metrics. With the exception of precision and specificity, the improvements of classification metrics when compared to other methods are also statistically significant. The achieved proposed methods' average accuracy of 96.26% is also better than the highest reported accuracy (94.39%) from published works, using the same Chest X-ray image dataset.

In the future, we would like to apply our proposed SGDRE method to other medical classification tasks using different image datasets and assess SGDRE's predictive performance. Regarding the nature of SGDRE, we would only need to choose an appropriate CNN architecture tailored to a specific problem. We presume that obtained results would be competitive with existing state-of-the-art methods for the chosen classification task.

## CRediT authorship contribution statement

**Grega Vrbančič:** Conceptualization, Methodology, Software, Writing – original draft, Visualization. **Vili Podgorelec:** Validation, Formal analysis, Writing – review & editing, Visualization, Supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

The authors acknowledge the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0057).

## References

Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. URL: https://www.tensorflow.org/. Software available from tensorflow.org.

Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2019). Comparison of deep learning approaches for multi-label chest X-ray classification. *Scientific Reports*, *9*(1), 6381.

Bardou, D., Zhang, K., & Ahmad, S. M. (2018). Lung sounds classification using convolutional neural networks. *Artificial Intelligence in Medicine*, *88*, 58–69.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier, & G. Saporta (Eds.), *Proceedings of COMPSTAT'2010* (pp. 177–186). Heidelberg: Physica-Verlag HD.

Chandra, T. B., Verma, K., Singh, B. K., Jain, D., & Netam, S. S. (2021). Coronavirus disease (COVID-19) detection in chest X-ray images using majority voting based classifier ensemble. *Expert Systems with Applications*, *165*, Article 113909.

Chollet, F., et al. (2015). Keras. URL: https://keras.io.

Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., et al. (2020). A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Applied Sciences*, *10*(2), 559.

Cohen, J. P., Bertin, P., & Frappier, V. (2019). Chester: A web delivered locally computed chest x-ray disease prediction system. arXiv preprint arXiv:1901.11210.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems* (pp. 2933–2941).

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, http://dx.doi.org/10.1016/j.jecp.2010.03.005, URL: arXiv:arXiv:1011.1669v3.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, *31*(4–5), 198–211.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121–2159.

El-Kenawy, E.-S. M., Mirjalili, S., Ibrahim, A., Alrahmawy, M., El-Said, M., Zaki, R. M., et al. (2021). Advanced meta-heuristics, convolutional neural networks, and feature selectors for efficient COVID-19 X-Ray chest image classification. *IEEE Access*, *9*, 36019–36037.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Wiley-Interscience.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202, S. Shiotani Et Al./Neurocomputing 9 (1995) Ill-130, 130.

Goodman, L. R. (2014). *Felson's principles of chest roentgenology, a programmed text*. Elsevier Health Sciences.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377.

Guan, Q., & Huang, Y. (2020). Multi-label chest X-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, *130*, 259–266.

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., & Feris, R. S. (2018). SpotTune: Transfer learning through adaptive fine-tuning. CoRR, abs/1811.08737. URL: arXiv:1811.08737.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*(2), 171–186.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., & Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get m for free. arXiv preprint arXiv:1704.00109.

Ke, Q., Zhang, J., Wei, W., Połap, D., Woźniak, M., Kośmider, L., et al. (2019). A neuro-heuristic approach for recognition of lung diseases from X-ray images. *Expert Systems with Applications*, *126*, 218–232.

Kermany, D., & Goldbaum, M. (2018). Labeled optical coherence tomography (OCT) and chest X-Ray images for classification. *Mendeley Data*, *2*, http://dx.doi.org/10.17632/rscbjbr9sj.2.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, *172*(5), 1122–1131.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836.

Khishe, M., Caraffini, F., & Kuhn, S. (2021). Evolving deep learning convolutional neural networks for early COVID-19 detection in chest X-ray images. *Mathematics*, *9*(9), 1002.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, *284*(2), 574–582.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Liang, G., & Zheng, L. (2020). A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer Methods and Programs in Biomedicine*, *187*, Article 104964.

Lodwick, G. S., Haun, C. L., Smith, W. E., Keller, R. F., & Robertson, E. D. (1963). Computer diagnosis of primary bone tumors: A preliminary report. *Radiology*, *80*(2), 273–275.

Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt, & J. Millman (Eds.), *Proceedings of the 9th Python in science conference* (pp. 51–56).

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). *Introduction to the practice of statistics*. New York: W.H. Freeman.

Morabito, F. C., Campolo, M., Ieracitano, C., & Mammone, N. (2019). Deep learning approaches to electrophysiological multivariate time-series analysis. In *Artificial intelligence in the age of neural networks and brain computing* (pp. 219–243). Elsevier.

Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. In *Advances in neural information processing systems* (pp. 5947–5956).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45.

Qin, C., Yao, D., Shi, Y., & Song, Z. (2018). Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomedical Engineering Online*, *17*(1), 113.

Radiuk, P., & Kutucu, H. (2020). Heuristic architecture search using network morphism for chest X-Ray classification. In *IntelITSIS* (pp. 107–121).

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.

Rouhi, R., Jafari, M., Kasaei, S., & Keshavarzian, P. (2015). Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications*, *42*(3), 990–1002.

Rudan, I., Boschi-Pinto, C., Biloglav, Z., Mulholland, K., & Campbell, H. (2008). Epidemiology and etiology of childhood pneumonia. *Bulletin of the World Health Organization*, *86*, 408–416B.

Rudan, I., O'brien, K. L., Nair, H., Liu, L., Theodoratou, E., Qazi, S., et al. (2013). Epidemiology and etiology of childhood pneumonia in 2010: estimates of incidence, severe morbidity, mortality, underlying risk factors and causative pathogens for 192 countries. *Journal of Global Health*, *3*(1).

Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, *5*(3), 1.

Shen, S., Han, S. X., Aberle, D. R., Bui, A. A., & Hsu, W. (2019). An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, *128*, 84–95.

Siddiqi, R. (2019). Automated pneumonia diagnosis using a customized sequential convolutional neural network. In *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies* (pp. 64–70). New York, NY, USA: ACM, http://dx.doi.org/10.1145/3342999.3343001.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Singh, A. K., Kumar, A., Mahmud, M., Kaiser, M. S., & Kishore, A. (2021). COVID-19 infection detection from chest X-ray images using hybrid social group optimization and support vector classifier. *Cognitive Computation*, 1–13.

Smith, L. N. (2017a). Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision* (pp. 464–472). IEEE.

Stephen, O., Sain, M., Maduh, U. J., & Jeong, D.-U. (2019). An efficient deep learning approach to pneumonia classification in healthcare. *Journal of Healthcare Engineering*, *2019*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Taylor, A. G., Mielke, C., & Mongan, J. (2018). Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study. *PLoS Medicine*, *15*(11), Article e1002697.

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, *4*(2), 26–31.

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, *13*(2), 22.

Vrbancic, G., Fister, I. J., & Podgorelec, V. (2019). Automatic detection of heartbeats in heart sound signals using deep convolutional neural networks. *Elektronika Ir Elektrotechnika*, *25*(3), 71–76. http://dx.doi.org/10.5755/j01.eie.25.3.23680.

Vrbančič, G., & Podgorelec, V. (2020). Transfer learning with adaptive fine-tuning. *IEEE Access*, *8*, 196197–196211.

Yanase, J., & Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, Article 112821.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.