Applied Acoustics 172 (2021) 107581

Contents lists available at ScienceDirect

Applied Acoustics

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

Spectral images based environmental sound classification using CNN with meaningful data augmentation

Zohaib Mushtaq *, Shun-Feng Su, Quoc-Viet Tran

Department of Electrical Engineering, National Taiwan University of Science and Technology (NTUST), Taiwan

ARTICLE INFO

Article history:
Received 1 July 2020
Accepted 7 August 2020
Available online 28 August 2020

Keywords:
Environmental sound classification
Convolutional neural network
Spectrogram
Data augmentation
Transfer learning

ABSTRACT

In this study, an effective approach of spectral images based on environmental sound classification using Convolutional Neural Networks (CNN) with meaningful data augmentation is proposed. The feature used in this approach is the Mel spectrogram. Our approach is to define features from audio clips in the form of spectrogram images. The randomly selected CNN models used in this experiment are, a 7-layer or a 9-layer CNN learned from scratch. Also, various well-known deep learning structures with transfer learning and with a concept of freezing initial layers, training model, unfreezing the layers, again training the model with discriminative learning are considered. Three datasets, ESC-10, ESC-50, and Us8k are considered. As for the transfer learning methodology, 11 explicit pre-trained deep learning structures are used. In this study, instead of using those available data augmentation schemes for images, we proposed to have meaningful data augmentation by considering variations applied to the audio clips directly. The results show the effectiveness, robustness, and high accuracy of the proposed approach. The meaningful data augmentation can accomplish the highest accuracy with a lower error rate on all datasets by using transfer learning models. Among those used models, The ResNet-152 attained 99.04% for ESC-10 and 99.49% for Us8k datasets. DenseNet-161 gained 97.57% for ESC-50. From our understanding, they are the best-achieved results on these datasets.

1. Introduction

In recent years the classification of sound or audio recognition system has expanded its momentum in various fields like different animals voice recognition [1], automatic screams detection [2], the combination of video and audio for crime scene warning system in Ref. [3], IoT based solution for urban noise detection in smart cities [4], sound classification with detection for medical and health care problems [5], classification of distinct musical instruments [6] and many more. This shows the importance and scope of autonomous sound recognition systems in almost every aspect of not only humans but also other's living organisms like trees and animal's life. Most of the sound classification or detection of audio clips normally related to the following domains. Automatic speech recognition system [7], music information recognition [8], and sound event recognition famously known as environment sound classification (ESC) in Ref. [9]. However, after considering the essence and characteristics of the above disciplines, the targets to be recognized in ESC is much different from others and cannot easily be described as music or speech signals. One of the main issues behind this is that ESC does not have any specific audio scene or structure like music and speech signals [10,11]. The second reason which makes the ESC task harder, the signal to noise ratio is small due to the larger distance between audio clip recorder and voice generation source as in comparison with musical information retrieval and speech recognition systems [12]. The above symptoms make ESC much difficult as compared with others.

In recognition tasks, the basic issue is what to recognize. In other words, what the inputs of the system are. This process sometimes is called feature extraction or data preprocessing in data mining. Most of the prominent and traditional audio features extraction techniques used for ESC tasks are Mel Frequency Cepstral Coefficients (MFCC), [13], Log-Mel Spectrogram in Ref. [14], Gammatone features in Ref. [15] and Wavelet features in Ref. [16]. This study is to use Mel spectrogram features. The above methodologies are all to use audio clips directly to get the features to accomplish the recognition tasks. Our approach is to define features from audio clips in the form of spectrogram images. The spectral images can be viewed as a visible representation of the frequency spectrum for the audio signals. These images are further used to define classifications in the models learned. The

* corresponding author.
E-mail address: D10507809@mail.ntust.edu.tw (Z. Mushtaq).

https://doi.org/10.1016/j.apacoust.2020.107581
0003-682X/

advantages of using spectral images over sound clips are that the audio signals are less periodic, weak ambiance, short interval, and the addition of noise on audio signals is much easy as compared with images [17]. In the literature, by replacing the audio files with their spectral images indeed can achieve a state of the art performance with greater accuracy rates for ESC datasets as shown in Ref. [18].

With those features defined, the system needs to define a way of recognizing those sounds. Usually, this process includes the involvement of machine learning for classifiers or Convolutional Neural Network (CNN) architectures for sound recognition tasks. Different machine learning algorithms like K-nearest neighbor, Gaussian mixture, and support vector machine are employed in Ref. [9,21] for ESC tasks. On the other hand, in Ref. [19], CNN is considered for ESC. In Ref. [20], the authors also considered transfer learning by using pre-trained weights for acoustic scene detection. In this study, a simple CNN model learned from scratch and the idea of transfer learning are both considered for our datasets. In transfer learning, these pre-trained weights of that model have been already trained by millions of images. The earlier layers of these models are frozen to find the best and optimal discriminative or cyclic learning rates in Ref. [21]. The recognition effects as shown in our experiments are very nice when compared to the existing approaches.

In our experiments, three datasets are utilized; ESC-10 [25], ESC-50 [25] and Urbansound8k (Us8k) [26]. As for datasets used, it is very obvious it is better to use a large number of training samples to avoid the risk of overfitting problems [22]. It can be seen that those databases are not large enough for CNN models used. To overcome this issue, data augmentation becomes necessary for learning tasks. As stated in Refs. [14,15], data augmentation not only can provide, more training data for reducing the possibility of overfitting in the training, but also can increase the accuracy and performance of the models. The proposed methodology in this study involves the idea of meaningful data augmentation to improve the performance of the learning.

In this study, an effective approach of spectral images based on environmental sound classification using CNN with meaningful data augmentation is proposed. The feature used in this approach is the Mel spectrogram. There are various data transformation schemes available for image-based training systems. In this paper, instead of using those schemes, we proposed to have meaningful data augmentation for effective classification of environmental sounds, by considering variations applied to the audio clips directly. The proposed approach didn't require any fragmentation of whole audio recording into smaller windows or frames. After the transformation, the whole sound clip is converted into a spectrogram image. Such type of technique is much useful and effective as compared to traditional image augmentation approaches. This experimental study comprised of two stages. In the first stage two randomly selected a 7-layer or a 9-layer CNN models learned from scratch have been implemented. Also, various well-known deep learning structures with transfer learning are considered. As for the transfer learning methodology, 11 explicit pre-trained deep learning structures are used. The concept of training the models while freezing initial layers and retrained the models after unfreezing the layers with discriminative learning has been enforced. The results show the effectiveness, robustness, and high accuracy of the proposed approach. The meaningful data augmentation can accomplish the highest accuracy with a lower error rate on all datasets by using transfer learning models. Form our understanding, they are the best-achieved results on these datasets.

The remaining arrangement of this paper is as follows. Section 2 describes the relevant literature or work on the feature extraction, data augmentation, transfer learning related to environmental sound classification. The methodology structure is discussed in Section 3. Section 4 illustrates the details about the datasets and systems used. The detailed results and discussions on the comparison of the proposed method with others' published results presented in Section 5. In the last Section 6 conclusive remarks are given.

2. Related work

Nowadays, the environment sound classification (ESC) domain is becoming very popular. Recent studies promote the useage of CNN in ESC. Our study used three prominent ESC datasets, ESC-10, ESC-50, and Us8k, which are also considered by using CNN in Ref. [25] with an accuracy of 80.5%, 64.9%, and 73.7%, respectively. Two deep network compositions have been used in the time domain and the frequency domain on the ESC-10 dataset [26]. Results show the best performance given by using CNN is 89.9% in the frequency domain. Pillos *et al.* [27] used MFCC based real-time environment sound event recognition on the ESC-10 dataset. Random forest and Multilayer perceptron are both implemented in that work. The result shows the best performance by MLP with an accuracy rate of 74.50%. The author in Ref. [28] considered the classification of the ESC-50 dataset by using Mel-spectrogram feature extraction and online data augmentation. The best result attained by CNN was 71.2%. Li *et al.* introduced an ensemble stacked model by using CNN for ESC-10, ESC-50, and Us8k datasets [29]. The Dempster-Shafer theory of evidence is used to construct DS-CNN. The best accuracy achieved in this study is 82.8% for ESC-50, 92.1% for ESC-10, and 91.9% for Us8k. Huzaifah *et al.* demonstrate different signal processing techniques on the ESC-50 and Us8k datasets [19]. The methodologies involve Short-Time Fourier Transform (STFT), Constant Q Transform (CQT), and continuous wavelet transform (CWT). These techniques are combined with linear and Mel scales. The best results achieved by wideband on Mel-STFT on Us8k with an accuracy of 74.66% and wideband on Linear-STFT attained an accuracy of 55% for ESC-50 in Ref. [19]. In Ref. [30], Abdoli *et al.* proposed a 1D-CNN for the Us8k dataset and gained an accuracy of 89%. In Ref. [31] Agrawal *et al.* proposed a novel idea of Teager energy operator (TEO) based coefficients in various combinations with Mel filter cepstral coefficient (MFCC) and Gammatone spectral coefficient (GTSC) for the ESC-50 and Us8k datasets. The results indicate the combination of TEO-GTSC and GTSC achieved an accuracy of 81.95% for ESC-50 and 88.02% for Us8k. Aytar *et al.* [32] proposed a unique idea of transferring the discriminative visual information from a well-trained visual recognition model into audio modality by using unlabelled video. The author called this approach as SoundNet. It achieved an accuracy of 74.2% for ESC-50 and 92.2% for ESC-10. Zhao *et al.* proposed a combination of SoundNet and EnvNet and attained the result of 77.4% accuracy for the ESC-50 dataset in Ref. [33]. In Ref. [18], Boddapati *et al.* proposed a method of converting audio clips into spectral images and also implemented the well-known transfer learning models for image recognition tasks like (GoogleNet and AlexNet) to these spectral images. The experiment shows GoogleNet accomplished remarkable results of 86% accuracy for ESC-10, 73% for ESC-50 and 93% for Us8k. Another study which demonstrates the highest achieved results is proposed by Sharma *et al.* in Ref. [34]. He proposed a multifeatured approach with strong augmentation in combination with DCNN to get an accuracy of 97.25% for ESC-10, 95.50% for ESC-50 and 98.60% for Us8k. In this study, a new approach is proposed and it is a combination of transforming sound events into Mel-spectrogram images and strong data augmentation plus pre-trained weights with optimal learning rates based on cyclic learning.

3. Methodology

It is not an easy task to extract features and classify various sounds through short audio clips. Many audio recordings have background noises, very short intervals, and rapid changes in the clips. Those noises make it very hard for the DCNN model to classify. This study involves the classification of sounds from the environment after converting the audio clips into spectrogram images. The frequency spectrum of the audio signal is visually represented in the form of spectrogram images. It is a very rare approach to convert audio files into images for classification tasks. As discussed earlier, such conversion can provide a better classification accuracy and a less error rate. It can be found from Ref. [18] that with the use of CNN in environmental sound or acoustic scene classification, the performance of using the spectrogram image is much better in comparison with directly using audio files. Thus, in this study, spectrogram images are considered for ESC.

As mentioned, the CNN models obtained from scratch and transfer learning-based models are employed in our approach. To catch more abstract features, more layers will be used in the CNN model, and usually, it is called Deep Convolutional Neural Network (DCNN). The detailed information about DCNN will be introduced in the following. As stated in Ref. [24], data augmentation and transformation techniques are key factors for good performance for deep neural networks. Two distinct images augmentation approaches are considered; one is a traditional method available for various image-based training tasks [35] and the other one is the proposed approach specially designed for classification of different sounds by using spectrogram images. Since those data augmentation schemes have their physical meaning for sounds, it is called meaningful data augmentation in this study. Those details of data enhancement schemes are given in the following. Besides, transfer learning models and the discriminative learning method are considered in this study they are also introduced in the following subsections.

3.1. Deep convolutional neural network (DCNN)

In this study, two different deep convolutional neural networks (DCNN) architectures are employed. The first convolutional neural network architecture is CNN-1 with seven layers and the second architecture with nine layers is denoted as CNN-2. In Ref. [15], DCNN is also used for the Us8k dataset but only 5 layers. The Mel-spectrogram images extracted during the feature extraction process are different in sizes. The size of those input images is fixed to 128. Input is denoted as N, where Θ is a parameter of a composite nonlinear function denoted as $F(\cdot|\Theta)$. The purpose of this function is to maps N to the output predicted value Y:

$$Y = F(N|\Theta) = f_L(\cdots\cdots\cdots f_3(f_2(N|\theta_2)|\theta_3|\theta_L), \quad (1)$$

where $f_l(\cdot|\theta_l)$ is the l-layer of the convolutional neural network. L is the layer number, for CNN-1, L = 7, and CNN-2, L = 9. The parameter for the l-layer is $\theta_l = [X, b]$. Then the operations in the convolutional layers can be expressed as:

$$Y_l = f_l(N_l|\theta_l) = h(X * N_l + b), \quad (2)$$

where N_l *is the i*nput of the l-layer, X *r*epresents the corresponding filter, $*$ is the valid convolution, $h(\cdot)$ is the pointwise activation function, and b is the vector bias term. In this study, the following procedure is used for all experiments. The training lasts for 50 epochs, the Adam optimizer and the categorical cross-entropy loss function are used, and the batch size used for each dataset is 64. The description of layers with filters, max-pooling function, dropout, and activation function is discussed in the following.

3.1.1. Convolutional neural network architecture 1 (CNN-1) 7 layers

- L1: The first layer consists of 24 kernels with a 6*6 respective field. It is followed by a 4*2 strided max-pooling function over the last two time and frequency dimensions. The Rectified Linear Unit (ReLU) is utilized as the activation function.
- L2: The second layer consists of 48 kernels with a 5*5 respective field. The padding involved in this layer is "same". It is followed by a 4*2 strided max-pooling function. The Rectified Linear Unit (ReLU) is exploited as an activation function.
- L3: The third layer consists of 48 kernels with a 5*5 respective field. The padding involves in this layer is also "same". The max-pooling function is not involved in this layer. The Rectified Linear Unit (ReLU) is used as the activation function.
- L4: The fourth layer consists of 60 kernels with a 4*4 respective field. The padding involves in this layer is "same". The max-pooling function is not utilized in this layer. The activation function used is the Rectified Linear Unit (ReLU).
- L5: The fifth layer consists of 72 convolutional kernels with a 4*4 respective field. The padding involves in this layer is also "same" with no max-pooling. The activation function used is the Rectified Linear Unit (ReLU).
- L6: The sixth layer is the first dense layer that consists of 84 hidden units with Rectified Linear Unit (ReLU) as an activation function. The dropout rate of 0.5 is also used to avoid overfitting problems.
- L7: The last layer is the second dense layer and also known as the output layer. It consists of the output units, equal in numbers to the total number of classes in the dataset. The activation function used in this layer is softmax

3.1.2. Convolutional neural network architecture 2 (CNN-2) 9 layers

- L1: The first layer is comprising of 24 kernels with a 6*6 respective field. It is followed by 4*2 strided max-pooling function over the last two time and frequency dimensions. The Rectified Linear Unit (ReLU) is utilized as an activation function.
- L2: The second layer consists of 48 kernels with a 5*5 respective field. The padding involves in this layer is "same". It is followed by a 4*2 strided max-pooling function. The Rectified Linear Unit (ReLU) is exploited as an activation function.
- L3: The third layer comprises of 48 kernels with 5*5 respective field. The "same" padding layer is used. The max-pooling function is not involved in this layer. The Rectified Linear Unit (ReLU) is used as the activation function.
- L4: The fourth layer involves 60 kernels with a 4*4 respective field. The padding used is "same". The max-pooling function is not utilized in this layer. The activation function used is the Rectified Linear Unit (ReLU).
- L5: The fifth layer with 72 convolutional kernels and a 4*4 respective field. No max-pooling function and "same" padding is used. The activation function used is the Rectified Linear Unit (ReLU).
- L6: The sixth layer consists of 80 convolutional kernels with a 3*3 respective field. The padding utilized in this layer is also "same" with no max-pooling. The activation function used is the Rectified Linear Unit (ReLU).
- L7: The seventh layer contains 80 convolutional kernels with a 3*3 respective field and padding used is also "same" with no max-pooling. The activation function used is the Rectified Linear Unit (ReLU).
- L8: The eight-layer is the first dense layer that consists of 128 hidden units with Rectified Linear Unit (ReLU) as an activation function.0.5 dropout rate also applied in this layer to avoid overfitting.

- L9: The last layer is the second dense layer and also known as the output layer. It consists of the output units, equal in numbers to the total number of classes in the dataset. The activation function used in this layer is SoftMax.

The use of CNN-1 and CNN-2 is for the comparison of the accuracy, loss, and behavior of spectral images based on environmental sound classification with meaningful data augmentation. Of course, as mentioned, we also considered transfer learning models for spectral images based on environmental sound classification. Those transfer learning models will be introduced later.

3.2. Data augmentation approaches

DCNN can discriminate Spectro-temporal patterns and can make distinctions between sounds that are masked in frequency or time or by any other noise. The main drawback of DCNN is that it needs a large amount of data in training due to its large number of parameters to be tuned. When only an insufficient amount of data is used in training, there may have overfitting phenomena, which mean significant differences between the training accuracy and the testing accuracy. To overcome this deficiency, the technique of data augmentation is employed in practice. By data augmentation, training data are increased by artificially creating more training data in various ways [36]. Nowadays, data argumentation become a common approach in deep learning systems, especially for images and there are many methods and pre-build tools for augmenting images. Kera's library is one of the most common and widely used packages for image augmentation [35]. In this study, those frequently used image augmentation techniques are employed on spectrogram images and new sound-based augmentation approaches are also considered for spectrogram-based images. Those approaches are briefly introduced in the following.

Kera's package [35] is used in our study to perform the various augmentation operations on images. Their respective values used are **Zoom range**: 0.25, **Width shift**: 0.20, **Fill mode**: nearest, **Brightness range**: [0.5,1.5], **Rotation angle**: 30°, **Height shift**: 0.20, **Shear range**: 0.30, and **Horizontal flip**: True. This kind of approach is referred to as the Traditional augmentation approach (TAA) in this study. Those data augmentation approaches are very common and have been used widely for image classification tasks [39,40]. These transformations are randomly generated until the desired number of augmented images have been generated. The block diagram for this approach is shown in Fig. 1. In the variations introduced in the dataset are random and the total numbers of generated data are for ESC10, 2000 randomly generated audio files images + 400 original data set. Thus the total data number is 2400 for the training model. For ESC50, 10,000 randomly generated data from 2000 original images and there is a total of 12,000 images. For US8K, there is a total of 52,400 spectrogram images generated from 8732 original images.

The idea of this approach is that the basics of learning are to learn from those training patterns. Thus, what the inputs patterns provide are what the system learns. Thus, for data augmentation, instead of generating data from various mechanisms that may not have any physical meaning, we proposed to consider variations of data with a physical meaning. As mentioned, this study is to consider sound recognition in the view of spectrogram-based images. Thus, ways of increasing the number of images data in the form of spectrograms in the sound level are considered in this study. The methodology of our proposed system is shown in Fig. 2. Five distinct augmentation variations are conducted. Each augmentation technique is applied directly to the clips or audio recordings and those generated audio clips are converted into Mel-spectrogram images for training. The deformations parameters in each approach have been chosen such that the validity of the label is preserved. These deformations or augmentation techniques have been done by using the Librosa library [37]. The techniques used are described below:

- Positive pitch shift (PPS): In this approach, each sample of all datasets was positively pitch-shifted by the factor of (+2). Our earlier experiments indicated pitch-shifting as an effective augmentation technique.
- Negative pitch shift (NPS): The audio signals from each dataset were deformed by the factor of (−2).
- Slow time stretches (STS): In this approach pitch of the audio signals remains unchanged. Each audio sample was slow down by the factor of (0.7).
- Fast time stretches (FTS): The speed of the audio samples become fast by the value of (1.20).
- Trim silence (TS): This technique helps to trail or trim the silence part of the audio signals.

In this approach, the augmentation is in some deterministic order in such a way that each variation implemented on the whole original audio data set individually. For ESC10, there are 400 original audio clips, 400 generated from PPS, 400 generated from NPS, 400 generated from STS, and 400 generated from FTS. Thus, there is a total of 2400 audio files. For ESC50, there are 2000 original audio clips, 2000 generated from PPS, 2000 generated from NPS, 2000 generated from STS, 2000 generated from FTS, and 2000 generated from TS. Thus, there is a total of 12,000 audio files. For US8K, there are 8732 original audio, 8732 generated from PPS, 8732 generated from NPS, 8732 generated from STS, 8732 generated from FTS, and 8732 generated from TS. Thus. there is a total of 52,392 audio files.

3.3. Transfer learning plus discriminative learning

3.3.1. Pre-trained models

Transfer learning is an idea of using pre-trained weights, models developed for some tasks for a specific task, for example like medical imaging in Ref. [38] or audio event detection from real-life scenes in Ref. [20]. In this study, this technique is used to improve the performance of our models. The pre-trained model is considered by using transfer learning techniques under two augmented approaches mentioned above. These pre-trained models already have very deep and dense layers. The perception of freezing the initial layers is used. These early layers involved fewer parameters but a high computational cost. In later steps, the models have been trained while unfreezing the initial layers with optimal learning rates based on cyclic learning technique. It is also found in the study that these transfer learning models also perform very well on spectral image recognition in a very less number of epochs. The general block diagram of the transfer learning implementation is shown in Fig. 3. These transfer learning models are already trained by a dataset of millions of images, like ImageNet, which contains thousands of classes of different objects. Nowadays this approach becoming very popular in various acoustic scenes and events detections. These transfer learning models are trained on images, extracted by different feature extraction techniques, and used in various applications like Refs. [39,40]. In this study, the same technique is used for the spectrogram images of the various environmental sounds related to our datasets. These pre-trained weights are trained with data augmentation approaches and discriminative learning with optimal learning rates is also considered. The overall block diagram of the proposed study through transfer learning is shown in Fig. 4.

The popular pre-trained weights used in this study are ResNet [41], DenseNet [42], SqueezeNet [43], AlexNet [47] and VGG [44]. The sub-categorization of these pre-trained models used in this study are shown below.

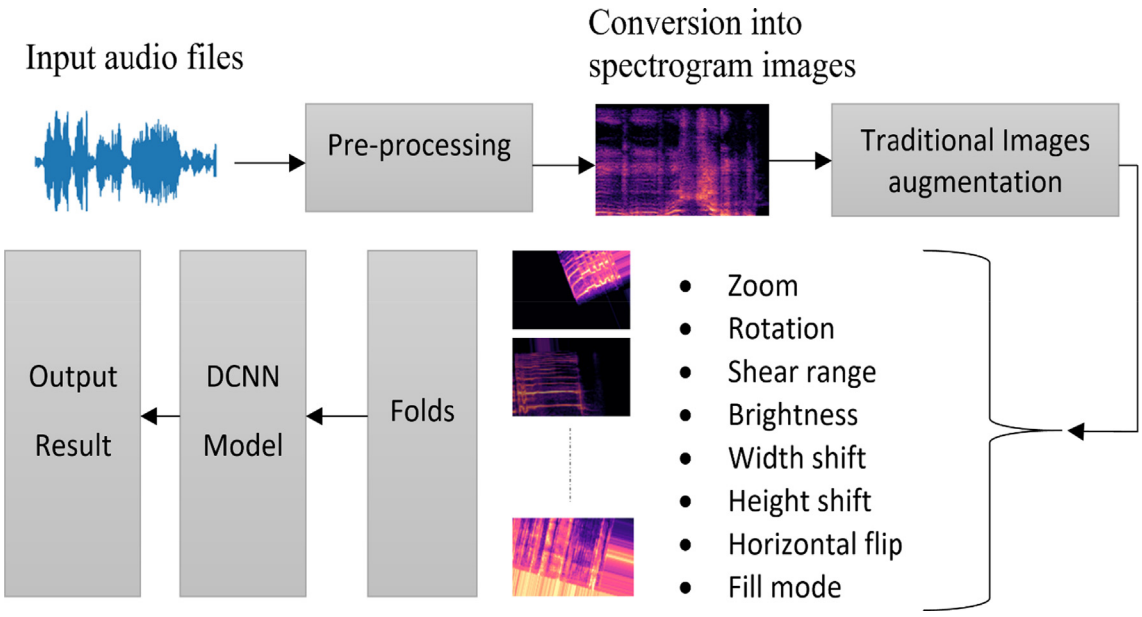

Fig.1. Traditional augmentation approach (TAA) block diagram.

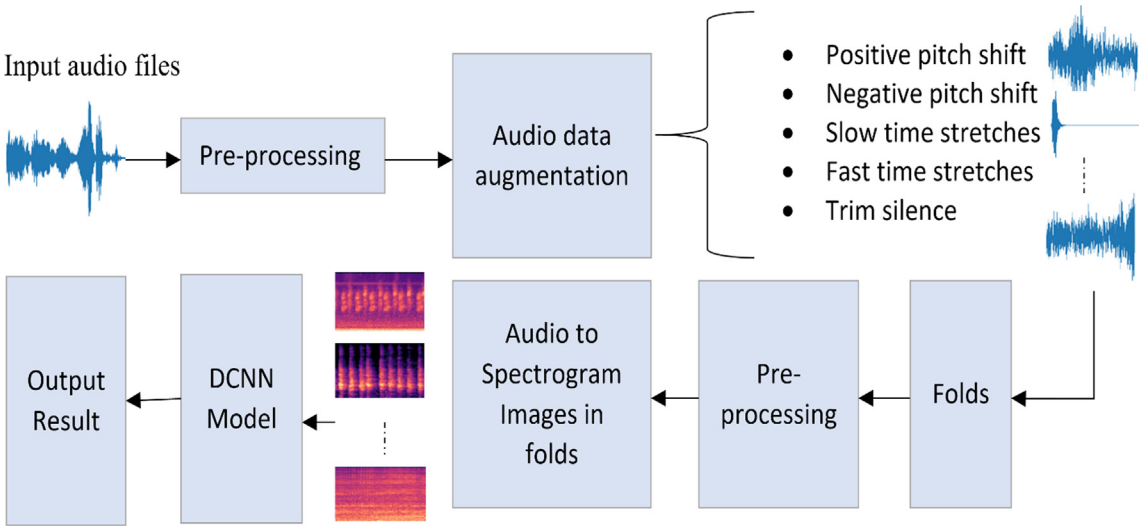

Fig.2. Novel augmentation approach (NAA) block diagram.

ResNet: ResNet-34, ResNet-50, ResNet-101, ResNet-152.
DenseNet: DenseNet-161, DenseNet-169, DenseNet-201.
Squeezenet: Squeezenet-1_1.
VGG: VGG-16, VGG-19.
AlexNet: Alex net.

3.3.2. Discriminative/Cyclic learning

The learning rate is one of the most crucial hyper-parameters while training neural networks. It not only helps the model to train quickly but also efficiently. Usually, it is suggested that the learning rate should be high at the starting point and then gradually decrease it as approaching to the global minimum. This type of approach only uses for building the model from scratch. Another way is to use a lower learning rate but this type of approach is not suitable as it will take a lot of time to converge. Using a high learning rate is not a good option because the loss function may go out of bounds and the error rate keeps increasing. There are a few other methods related to the learning rate that have been suggested by the researchers. The most famous is the adaptive learning rate and the cyclical learning rate [21]. The review of the adaptive learning methods has been discussed in detail in Refs. [45,46]. In this study, the cyclic learning rate is used to train pre-trained weights on the spectrogram-based datasets. In this technique instead of using increasing or decreasing learning rates, the model layers are grouped into three types of layers as shown in Fig. 5. The initial group of layers determines very small details and information like straight lines, edges, etc. Such types of layers are very helpful for most tasks. These layers will be trained on a lower learning rate so that the model can get more time to train on small details. The next group of layers describes or identifies complex and tricky patterns, like rectangles, squares, etc. and a relatively higher learning rate is used for this group. The last group of layers is to recognize patterns and shapes available in the images, and those layers are trained by an even larger but somehow to be the optimal learning rate. This type of using different learning rates for different groups of layers is called discriminative learning in our study.

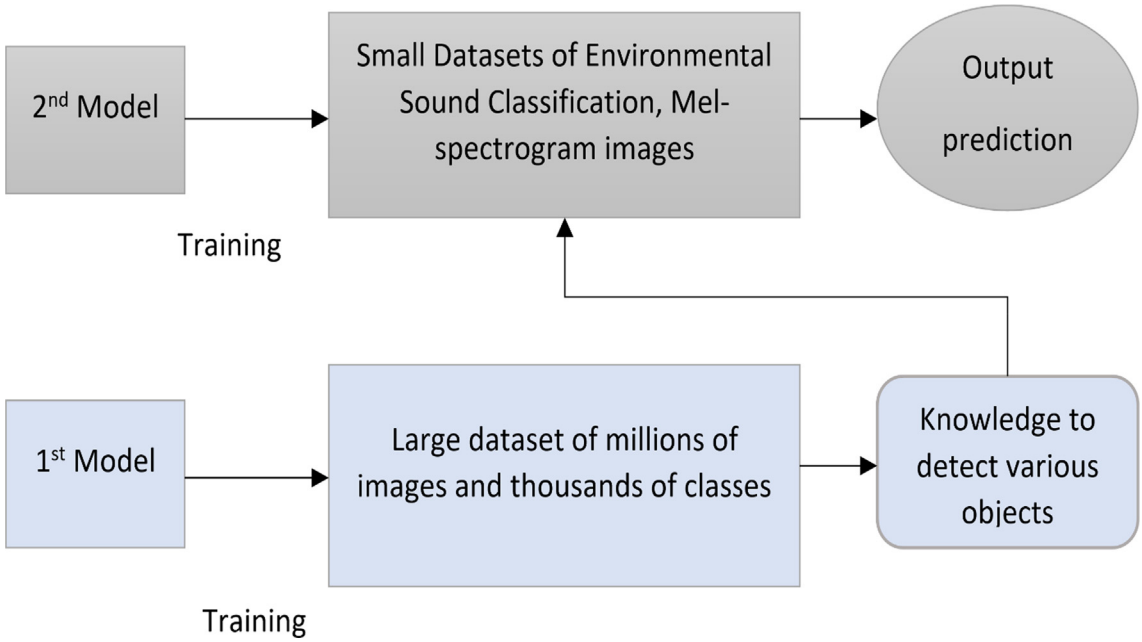

Fig.3. The general block diagram for transfer learning.

3.3.3. Finding optimal learning rates for used datasets

With the assistance of a learning rate finder in the FASTAI [47] library, the optimal learning rate can be obtained. Repeat the experiment with each dataset. Fig. 6. shows the best learning rates found in the experiments for our datasets. It is a relationship of loss versus the learning rate. It is used to find the best starting learning rate without making random guesses. To get the results, firstly the training lasts a few repetitions. Start with a very low learning rate and continue up until a high learning rate. The recorder will be used to record a loss over each iteration. The suggested learning rate is shown by the red dot which is the steepest value of the gradient. This red point is a first suggested guess for a learning rate. For a discriminative learning process, the initial layers of the models trained on the learning rate shown by the red point then middle layers of the models trained slightly higher learning rate towards the lowest loss (optimal) learning rate and the last bunch of the layers trained through that optimal learning rate which indicates least loss value which is also slightly higher learning rate than the starting value. Fig. 6(a) shows the learning rate for the ESC-10 dataset. The red circle indicates the starting learning rate as in the case of ESC-10 dataset the initial learning rate is $1e^{-4}$ and the final maximum value is $1e^{-3}$. It means if we have six layers in our model then the first two layers will be trained on the lowest learning rate in case of ESC-50 it is $1e^{-5}$, the next two layers will follow $1e^{-4}$ learning rate and the final and last two layers trained on maximum suggested learning the rate of $1e^{-3}$. In the same scenario, Fig. 6(b) and (c) demonstrate the best learning rate as ($1e^{-5}$, $1e^{-3}$) for ESC-50 and ($1e^{-6}$, $1e^{-4}$) for Us8k datasets

4. Materials and experimental setup

4.1. Datasets

In this experiment, three different audio datasets are used. These datasets are not related to any speech recognition or musical instrument classification. These datasets involved nonoverlapping, short audio clips of environmental sounds. Each sound file related to all datasets consists of a single event. Among these datasets, two are weak labeled ESC-10 and ESC-50 [48] data sets. The third one is a strong labeled Urbansound8k [26] dataset. The weakly labeled terminology means, the major portion of the audio clip is empty, only a minor part consists of a sound. Similarly, strong labeled comprises, most of the period of clip consists of voice or sound. Those datasets are briefly introduced in the following.

4.1.1. ESC-10

This dataset consists of 400 short clips recordings with an average time span of 5 s each. These small clips involve 10 different classes with a total time duration of 33 min. The class distribution in this data set is a uniform with a rate of 50 clips for each class. This dataset is divided into five-fold cross-validation and publicly available and annotated by Piczak [48]. Each fold consists of 80 clips with a random distribution of classes. The average human classification for this dataset is 95.7% as described by the author.

4.1.2. ESC-50

The collection of this dataset is 2000 short recordings of 50 separate classes, which are grouped into 5 major categories. These are animal sounds, a human sound which are non-speech sounds, urban or outdoor noises, indoor noises, and various natural soundscapes. The total duration of these sound clips is 168 min. This dataset is also dispersed into 5-fold cross-validation with 400 sound prunes in each fold. The average time duration of each sound clip is five seconds with a sampling frequency rate of 44,100 Hz. This dataset is searched, verified by Piczak [48] with an average human accuracy of 81.30%.

4.1.3. Urbansound8k

The Urbansound8k (Us8k) includes 8732 sound clips. The average period of these short clips is up to four seconds each. A total of 10 classes of various indoor and outdoor environmental sounds are a part of this dataset. This dataset is unequally distributed into 10-fold cross-validation. It is also a publicly available dataset in Ref. [49]. The comprehensive information related to each dataset is shown in Table 1.

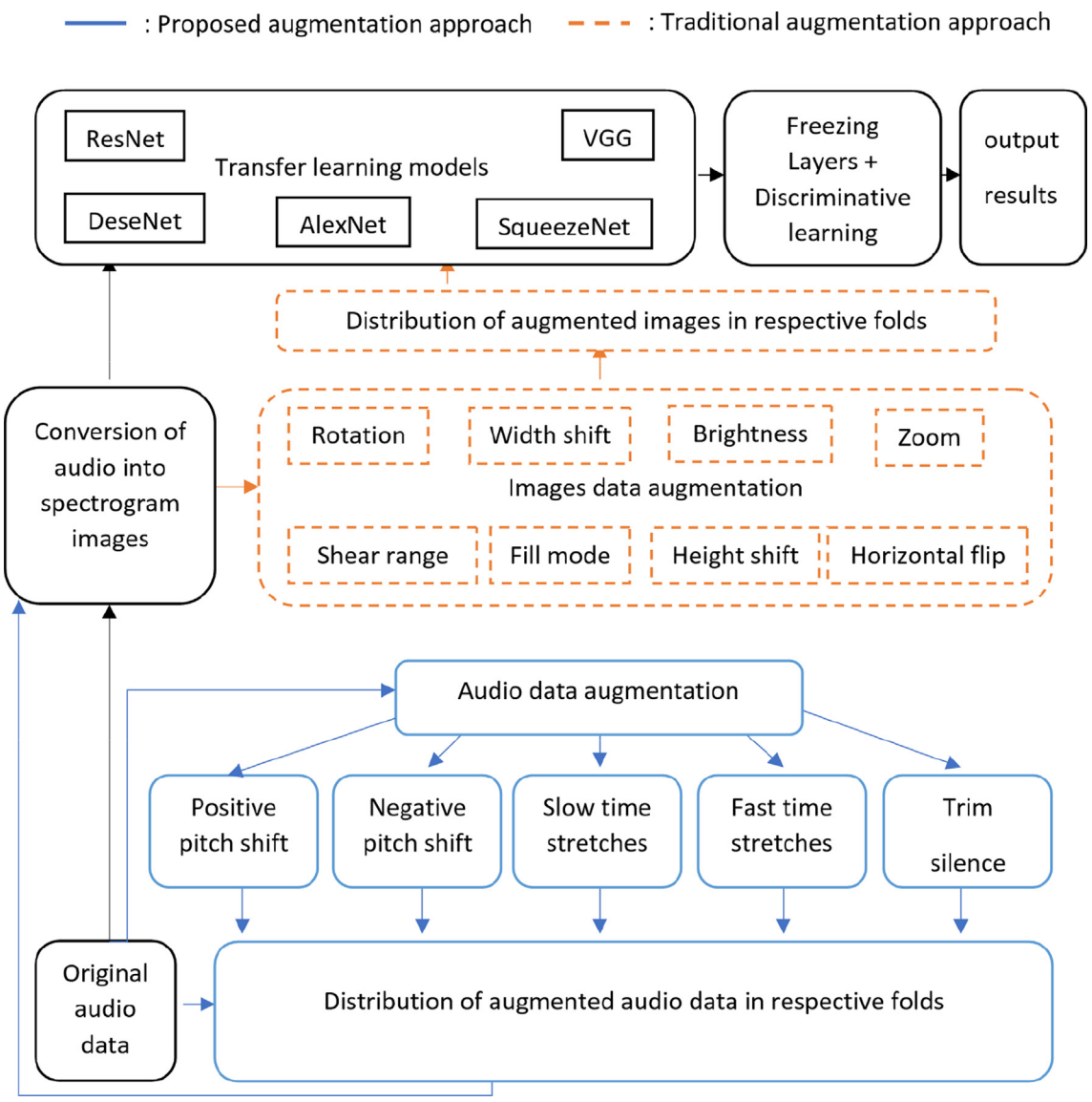

Fig.4. The general block diagram of the proposed methodology by using transfer learning.

4.2. Hardware specifications of the system

In this experiment we used two separate systems, having different hardware and almost identical software specification excluding the operating system. The first system is used to train on ESC-10 and ESC-50 datasets. It is a desktop system with Intel(R) Core™ i7-4770 CPU @ 3.40 GHz. RAM of the system is 16 GB and Graphics card used in the system is Nvidia GeForce, GTX 1080 with 8 GB VRAM. The hard disk of this system is 256 GB SSD + 1 TB HDD. The second system used is Intel(R) Core™ i9-7900X CPU @ 3.30 GHz with 64 GB RAM. The Graphics card of the system is Nvidia GeForce, GTX 1080 with 22 GB VRAM. The hard drive of the second system is 1 TB SSD.

4.3. Software specifications of the system

Different software and API libraries and packages are used in this experiment to train and build pre-trained models and convolutional neural networks from scratch. The first system has the Windows 10 operating system and the second system is Ubuntu 18.0.4. All the experiments are done in python language with the assistance of various installed packages and libraries. A few of the main APIs are as follows.

4.3.1. Anaconda

It is an open-source distribution to perform tasks in python with a lot of pre-installed libraries which are helpful in the visualization and analysis of data. It is not only operational in Windows and Linux but also easily useable in Mac operating systems.

4.3.2. Librosa

It is a python package normally used for the analysis of audio and music signal processing [37]. Its various functions involve feature extraction, decomposition of spectrograms, filters, temporal segmentation of spectrograms, and much more. In this study, this package used as a feature extraction technique, to extract Mel-spectrogram images from audio files or clips.

4.3.3. Keras

It is a high-level API specially designed to deal with deep neural networks [35]. It can run the python code on top of Theano, CNTK, and TensorFlow. In this study, it is used to build and implement a convolutional neural network from scratch. The traditional aug-

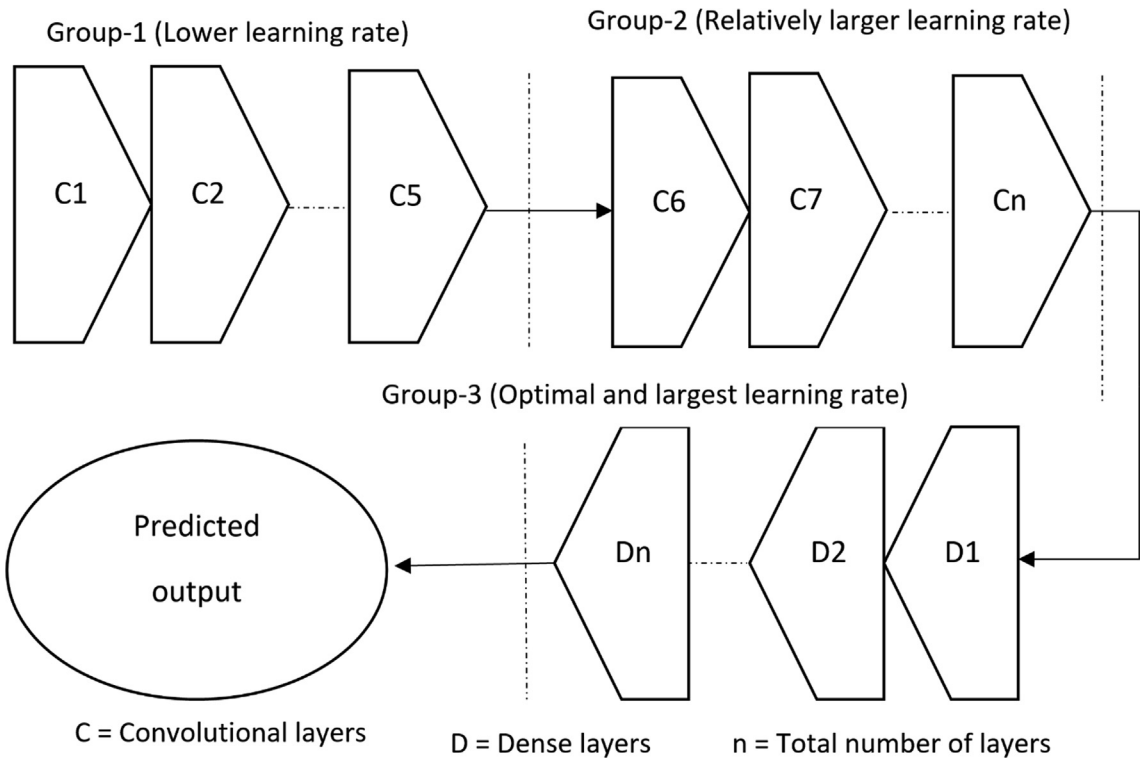

Fig. 5. The block diagram to demonstrate the idea of the cyclic learning technique.

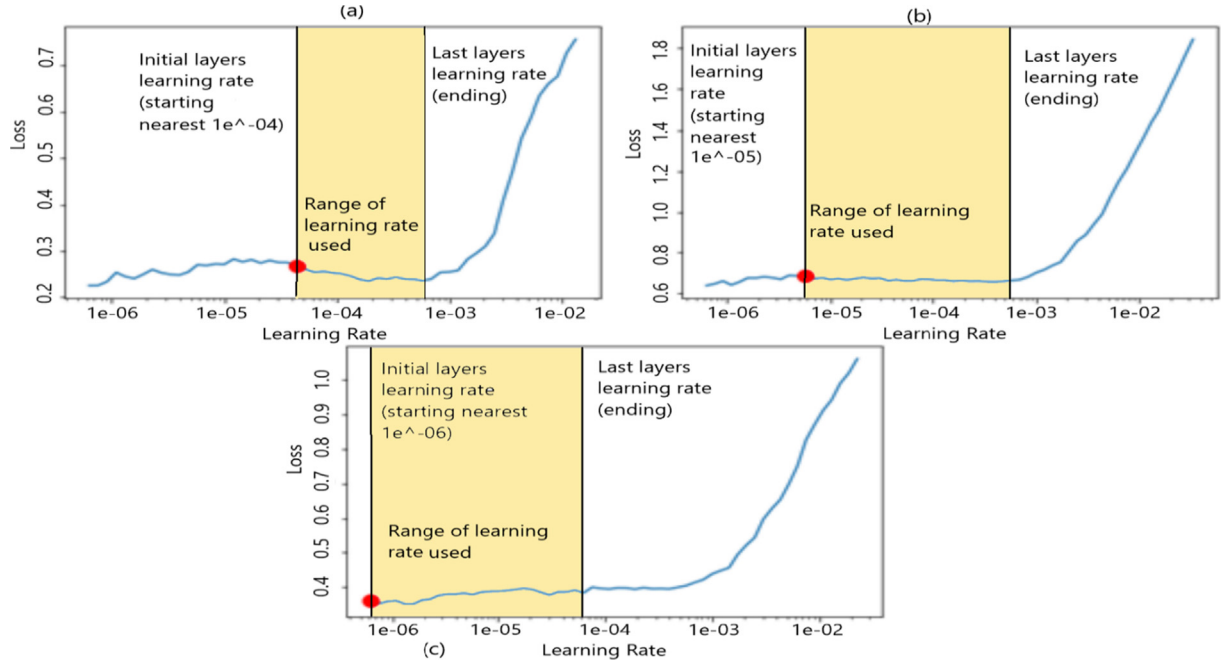

Fig. 6. The optimal learning rates of each dataset used. The red circle indicates the optimal starting learning rate. (a) ESC-10 ($1e^{-4}$, $1e^{-3}$), (b) ESC-50 ($1e^{-5}$, $1e^{-3}$), (c) Us8k ($1e^{-6}$, $1e^{-4}$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Description of used datasets.

Datasets	Classes	Total duration (mins)	Folds	No's of samples
ESC-10	10	33	05	400
ESC-50	50	168	05	2000
Urbansound8k	10	582	10	8732

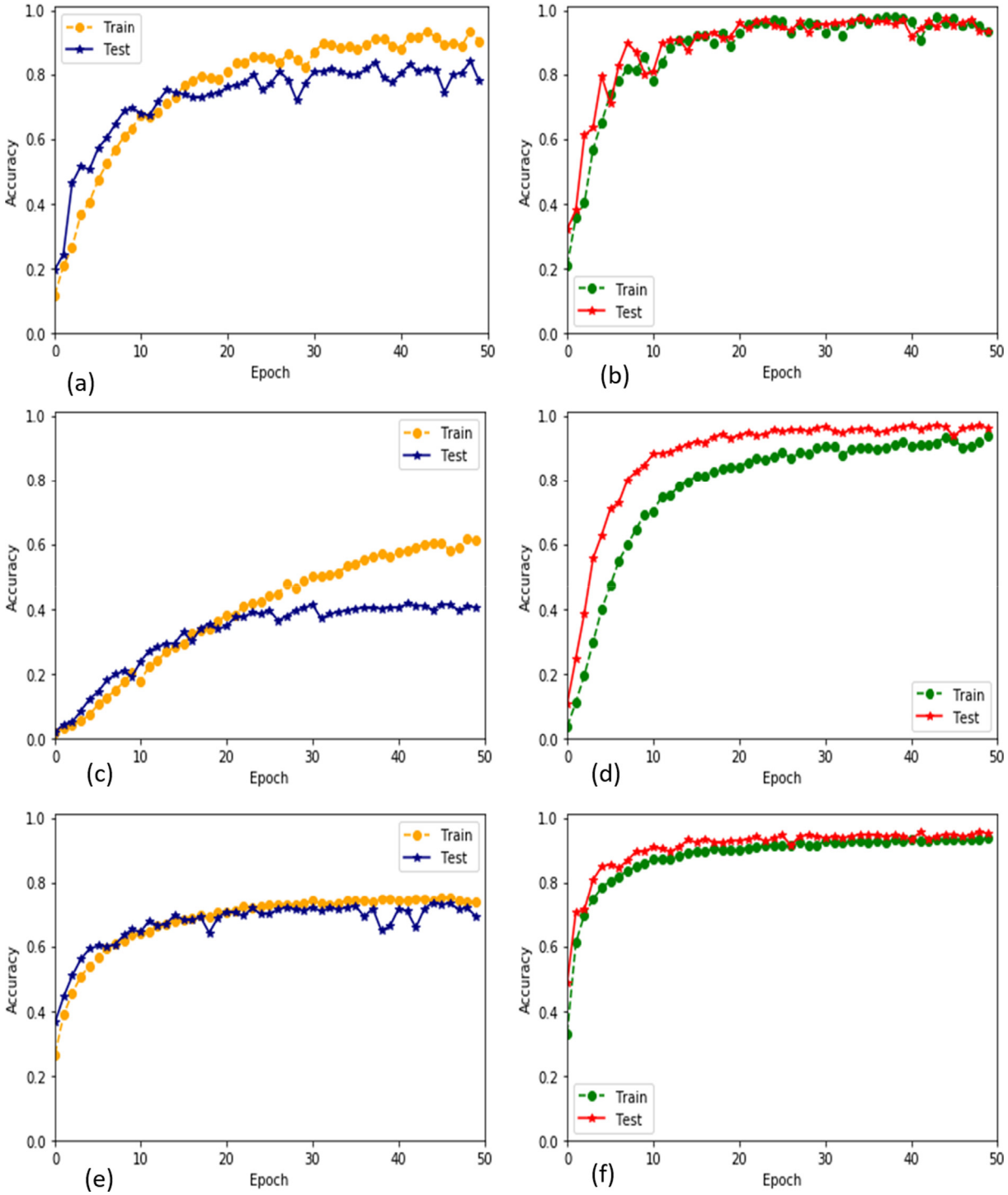

Fig. 7. Training accuracy versus validation accuracy results for 7-layers proposed CNN architecture. (a), (b) related to ESC-10, (c), (d) belongs to ESC-50 and (e), (f) associated with Us8k datasets. The left side of the figure (a), (c), (e) parts are the results implemented on TAA and the right side of the figure (b), (d), (f) belongs to NAA.

mentation approach for images normally used in this library. In our experiment, the traditional augmentation approach for Mel-spectrogram images also implemented in this package.

4.3.4. Fastai

One of the recently developed API, which makes it very easy for everyone to use deep neural networks is Fastai [47]. In this setup the transfer learning techniques with a concept of freezing layers and discriminative learning implemented in this library.

5. Results and discussions

This section illustrates the performance evaluation of those techniques discussed in the above on environment sound classification with different augmentation approaches. Our experimentation involves two parts. The first part is to consider DCNN. CNN-1 and CNN-2 are the deep convolutional neural networks used. The second part is to consider the pre-trained models by using transfer learning techniques on augmented data. Discriminative learning is also involved in transfer learning approaches.

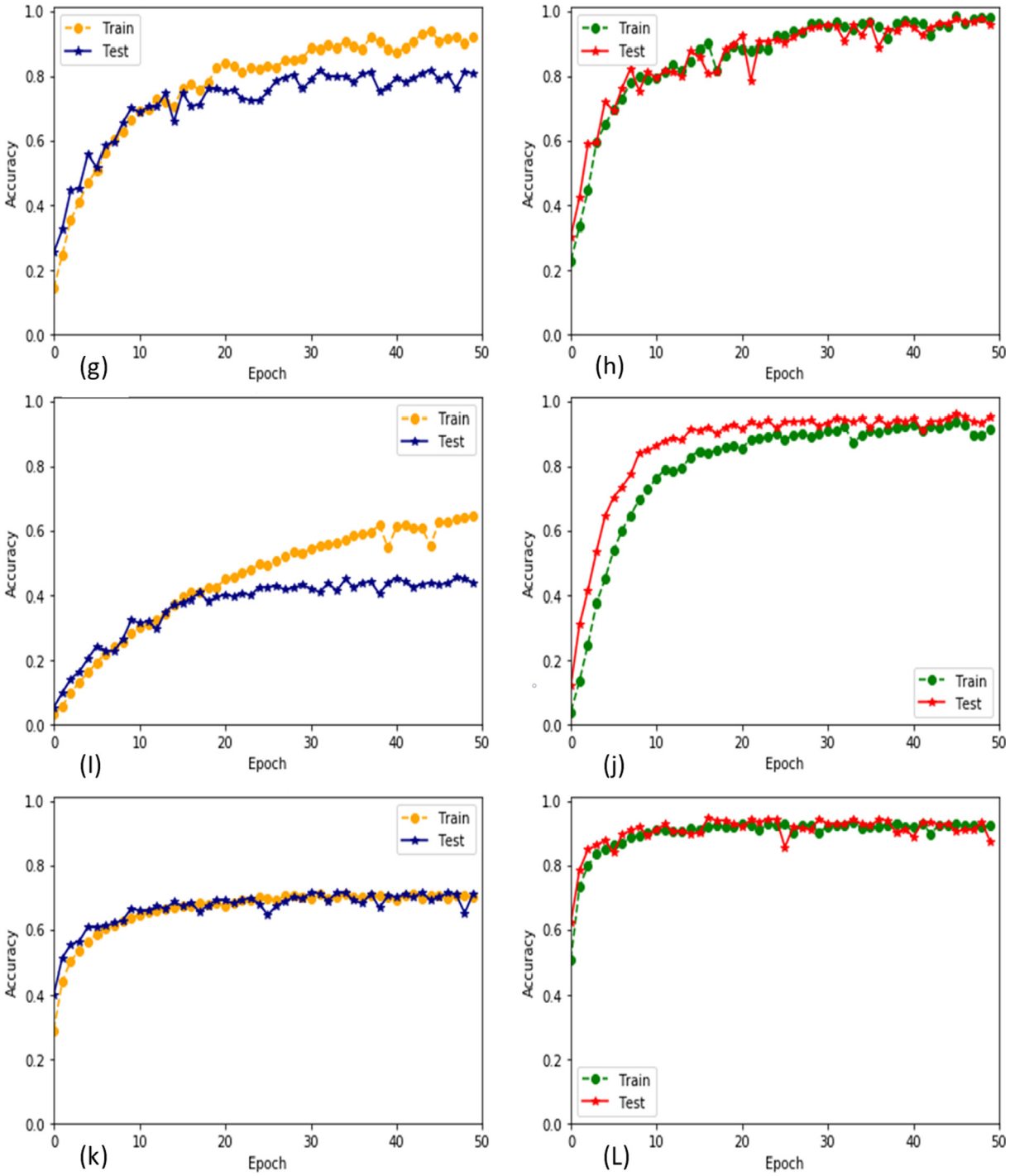

Fig. 8. Training accuracy versus validation accuracy results for 9-layers proposed CNN architecture. (g),(h) related to ESC-10, (I),(j) belongs to ESC-50 and (k),(l) associated with Us8k datasets. The left side of the figure (g),(I),(k) parts are the results implemented on TAA and the right side of the figure (h),(j),(L) belongs to NAA.

5.1. Results of using DCNN architectures

This section shows the performance evaluation of the DCNN architectures (CNN-1 and CNN-2) for augmented data. The total numbers of these augmented images dataset files are the same in number. For ESC-10, the total number after augmentation is 2400, for ESC-50, the total number is 12,000, and for Us8k, the number is 52,487. Fig. 7. shows the validation and training accuracy results of using the 7-layers CNN architecture for TAA and NAA on all datasets. As it can be seen that the proposed augmentation approach NAA outperforms with a huge validation accuracy difference in comparison with the traditional augmentation approach, TAA. The accuracy of the TAA on the ESC-10 dataset is 77.86% and for NAA is 93.50%. In the case of the ESC-50 dataset, the accuracy for TAA is 40.46% and for NAA is 96.10%. The last dataset Us8k, the accuracy for TAA is 69.13% and for NAA is 95.05%. To confirm the robustness and effectiveness of the proposed augmentation approach NAA, the same experiment conducted on the 9-

Table 2
The performance comparison of NAA with TAA by using 7- & 9-layers CNN.

Datasets	Augmentation Techniques	CNN layers	Validation loss	Accuracy In %	Training Time in Sec's	Epochs
ESC-10	TAA	07	0.9068	77.86	77.26	50
	NAA (proposed)	07	0.2454	93.50	65.46	50
ESC-50	TAA	07	2.4306	40.46	415.52	50
	NAA (proposed)	07	0.2068	96.10	316.05	50
Us8k	TAA	07	0.9675	69.13	3311.85	50
	NAA (proposed)	07	0.1777	95.05	1992.45	50
ESC-10	TAA	09	0.7794	80.86	82.11	50
	NAA (proposed)	09	0.1853	95.50	69.21	50
ESC-50	TAA	09	2.2773	43.66	424.23	50
	NAA (proposed)	09	0.2424	95.13	350.52	50
Us8k	TAA	09	0.8725	71.21	3532.22	50
	NAA (proposed)	09	0.4722	87.08	2311.65	50

layer CNN model. The results are shown in Fig. 8. The proposed NAA gets better results with a huge difference in accuracy in each dataset. The detailed results are summarized in Table 2. From the table, it is evident that the proposed NAA augmentation approach for spectrogram-based images classification concedes the best performance not only in accuracy but also in terms of validation loss and training time.

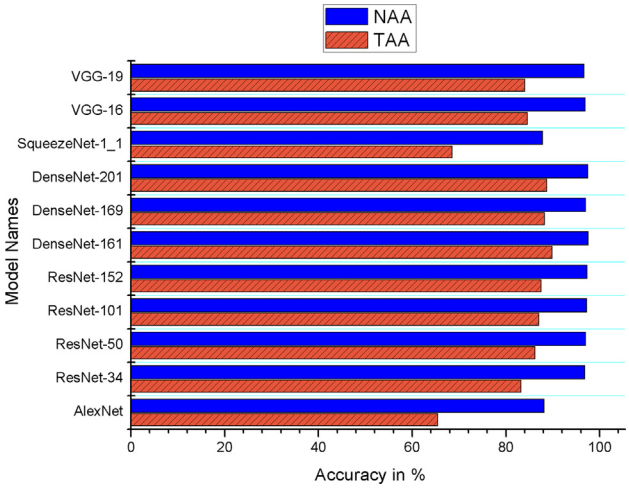

Fig. 9. Comparison of the accuracies of Transfer learning models with optimal learning rates for three epochs on TAA and NAA augmented data for the ESC-50 dataset.

5.2. Results with transfer learning

In this part, the transfer learning models are considered. The K-fold cross-validation strategy is implemented for all datasets. in this study, K = 5 is used for ESC-10, and ESC-50 and K = 10 are used for Us8k. The whole training process can be seen in Fig. 4. This experimental process carried out for only three epochs. The results for those three datasets are given in the following.

The comparison of these transfer learning models with the best discriminative learning rates on data augmentation approaches for the ESC-50 dataset is shown in Fig. 9. The clear difference between the accuracies of the models can be seen. The proposed data augmentation approach NAA again achieves better results not only beating the TAA approach but also accomplish the best and the highest accuracy of 97.57% achieved by any method published earlier on the ESC-50 dataset. This highest development in the accuracy for the ESC-50 dataset achieved by DenseNet-161 in three epochs only. Table 3. summarizes all results for the ESC-50 dataset. The highest accuracy is achieved by using DenseNet-161 for NAA

Table 3
The performance comparison of NAA with TAA by using Transfer learning models for the ESC-50 dataset.

ESC-50						
Model names	Epochs	Augmentation Techniques	Error-rate in %	Validation Loss	Accuracy In %	Training time (mins: secs)
Alex net	03	TAA	34.590	1.1436	65.410	69:13
	03	NAA (proposed)	11.858	0.4154	88.141	51:42
ResNet-34	03	TAA	16.794	0.5631	83.206	71:54
	03	NAA (proposed)	3.175	0.1104	96.824	54:55
ResNet-50	03	TAA	13.843	0.4659	86.157	76:38
	03	NAA (proposed)	3.008	0.1031	96.991	57:06
ResNet-101	03	TAA	12.975	0.4247	87.025	87:27
	03	NAA (proposed)	2.808	0.1721	97.191	102:47
ResNet-152	03	TAA	12.515	0.4134	87.485	99:30
	03	NAA (proposed)	2.700	0.1749	97.299	83:56
DenseNet-161	03	TAA	10.186	0.3513	89.813	107:16
	03	NAA (proposed)	2.425	0.0835	97.574	92:34
DenseNet-169	03	TAA	11.765	0.3849	88.235	107:29
	03	NAA (proposed)	3.008	0.1105	96.991	92:09
DenseNet-201	03	TAA	11.270	0.3735	88.730	117:12
	03	NAA (proposed)	2.558	0.0890	97.441	102:33
SqueezeNet-1_1	03	TAA	31.459	1.0399	68.541	70:23
	03	NAA (proposed)	12.158	0.4200	87.841	54:11
VGG-16	03	TAA	15.450	0.5241	84.55	75:11
	03	NAA (proposed)	3.102	0.1096	96.897	58:51
VGG-19	03	TAA	15.980	0.5329	84.02	71:05
	03	NAA (proposed)	3.308	0.1126	96.691	62:50

and has a shorter training time compared with TAA. For the validation loss, the proposed approach NAA achieves the lowest loss value of 0.0835. To the best of our knowledge, this is the highest accuracy achieved by any model or methodology for the ESC-50 dataset.

Next, the performance evaluation for the ESC-10 dataset is illustrated in Fig. 10. The highest accuracy is achieved by ResNet-152 which is 99.041% through the proposed NAA approach. This accuracy is the highest achieved accuracy for the ESC-10 dataset. Again our proposed augmentation approach accomplished the best results by using transfer learning models for the ESC-10 dataset. It is exhibited that the proposed augmentation scheme NAA again outperforms and provides the best results, especially for weakly labeled data. Table 4. describes the analogy of the prominent evaluation metrics used in transfer learning models. In terms of the training time, the highest accuracy achiever ResNet-152 is identical for both techniques, but the accuracy difference margin is around 3.80%.

The final dataset used is strongly labeled Urbansound8k (Us8k). The analysis of the accuracies between using NAA and TAA is shown in Fig. 11. NAA has better performance over TAA in terms of accuracy, error rate, and validation loss. The proposed methodology NAA attained the highest accuracy for the Us8k dataset by using ResNet-152 (99.49% accuracy. It is clearly illustrated in the figure that the proposed augmentation approach not only is valid for deeper pre-trained models like DenseNet-201, ResNet-152, etc. but illustrates a huge difference in terms of accuracy for a less deep model like AlexNet. All results are summarized in Table 5. This part of the study involves a huge number of images (more than 52,000). Thus, the second computer system with high computational power and GPU memory is used as mentioned in Section 3.

5.3. Analysis and discussion

Finally, Table 6. illustrates the detailed comparison of our results (of using ResNet-152 and DenseNet-161) with the use of those

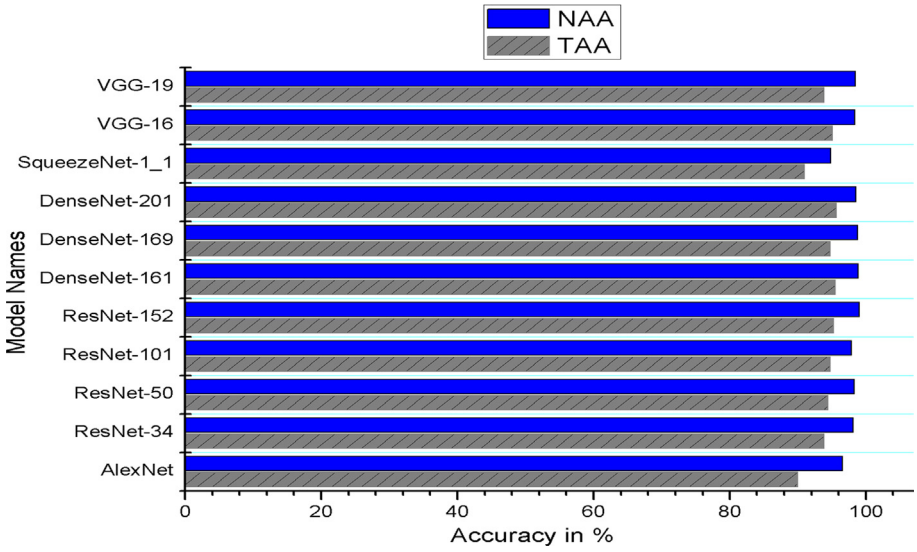

Fig. 10. Comparison of the accuracies of Transfer learning models with optimal learning rates for three epochs on TAA and NAA augmented data for the ESC-10 dataset.

Table 4
The performance comparison of NAA with TAA by using Transfer learning models for the ESC-10 dataset.

ESC-10						
Model names	Epochs	Augmentation Techniques	Error-rate in %	Validation Loss	Accuracy In %	Training time (mins: secs)
Alex net	03	TAA	10.042	0.2981	89.957	43:00
	03	NAA (proposed)	3.4583	0.0934	96.541	41:51
ResNet-34	03	TAA	6.1278	0.1711	93.872	44:54
	03	NAA (proposed)	1.8751	0.0530	98.124	47:41
ResNet-50	03	TAA	5.6170	0.1818	94.382	44:23
	03	NAA (proposed)	1.7500	0.0434	98.250	43:33
ResNet-101	03	TAA	5.2766	0.1651	94.723	46:31
	03	NAA (proposed)	2.1250	0.0613	97.874	45:52
ResNet-152	03	TAA	4.7661	0.1410	95.233	48:20
	03	NAA (proposed)	0.9583	0.0332	99.041	48:22
DenseNet-161	03	TAA	4.5414	0.1406	95.458	51:07
	03	NAA (proposed)	1.1250	0.0264	98.874	50:57
DenseNet-169	03	TAA	5.2768	0.1611	94.723	50:51
	03	NAA (proposed)	1.2084	0.0372	98.791	51:01
DenseNet-201	03	TAA	4.384	0.1346	95.616	52:30
	03	NAA (proposed)	1.458	0.0333	98.541	52:39
SqueezeNet-1_1	03	TAA	9.0212	0.2664	90.978	42:46
	03	NAA (proposed)	5.168	0.1494	94.832	41:53
VGG-16	03	TAA	4.8940	0.1477	95.105	44:11
	03	NAA (proposed)	1.6660	0.03718	98.333	44:08
VGG-19	03	TAA	6.1277	0.1768	93.872	44:45
	03	NAA (proposed)	1.5417	0.0413	98.458	44:32

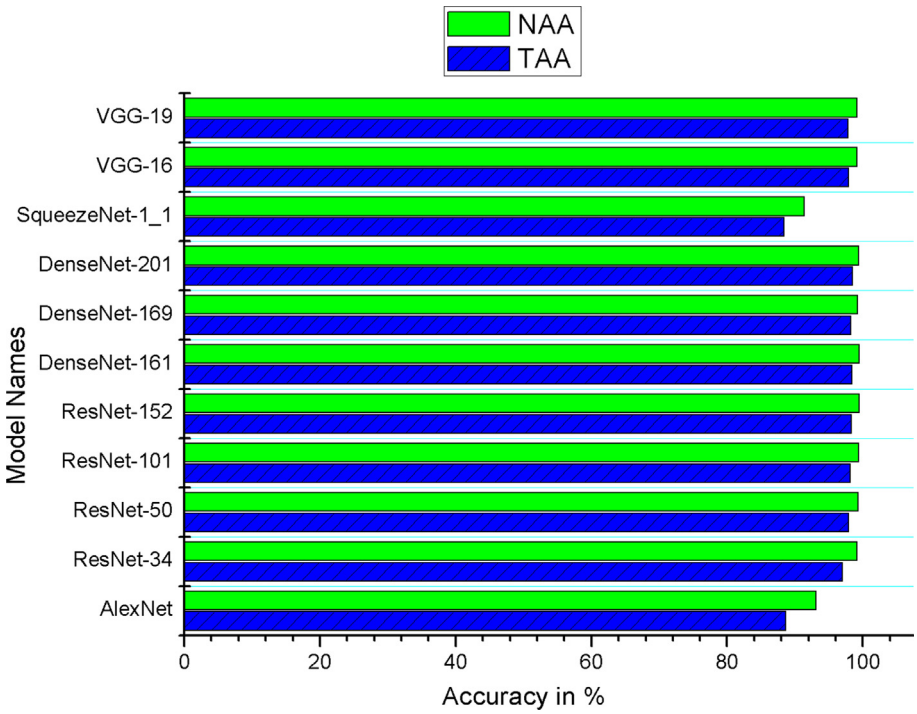

Fig. 11. Comparison of the accuracies of Transfer learning models with optimal learning rates for three epochs on TAA and NAA augmented data for the Urbansound8k (Us8K) dataset.

Table 5
The performance comparison of NAA with TAA by using Transfer learning models for Us8K (Urbansound8k) dataset.

Us8K (Urbansound8k)						
Model names	Epochs	Augmentation Techniques	Error-rate in %	Validation Loss	Accuracy In %	Training time (mins: secs)
Alex net	03	TAA	11.324	0.3423	88.675	53:13
	03	NAA (proposed)	6.943	0.2102	93.056	41:18
ResNet-34	03	TAA	2.9827	0.0847	97.017	66:54
	03	NAA (proposed)	0.8586	0.0254	99.141	56:49
ResNet-50	03	TAA	2.1376	0.0625	97.862	105:35
	03	NAA (proposed)	0.7266	0.0216	99.273	91:05
ResNet-101	03	TAA	1.861	0.0534	98.139	168:14
	03	NAA (proposed)	0.6197	0.0173	99.380	143:43
ResNet-152	03	TAA	1.707	0.0484	98.292	234:38
	03	NAA (proposed)	0.5023	0.0149	99.497	197:16
DenseNet-161	03	TAA	1.579	0.0449	98.420	246:05
	03	NAA (proposed)	0.5549	0.0166	99.445	210:01
DenseNet-169	03	TAA	1.776	0.0527	98.224	181:38
	03	NAA (proposed)	0.7456	0.0204	99.254	155:21
DenseNet-201	03	TAA	1.508	0.0433	98.491	221:18
	03	NAA (proposed)	0.596	0.0171	99.404	188:56
SqueezeNet-1_1	03	TAA	11.578	0.3793	88.421	52:55
	03	NAA (proposed)	8.621	0.3727	91.378	41:48
VGG-16	03	TAA	2.127	0.0641	97.872	147:30
	03	NAA (proposed)	0.8786	0.0261	99.121	126:52
VGG-19	03	TAA	2.193	0.0649	97.806	167:18
	03	NAA (proposed)	0.819	0.0248	99.181	143:45

existing approaches. Those approaches include Human, Baseline models, techniques involving data augmentation, images-based feature extraction models, and audio-based various feature combinations techniques. Table 6. It is the evidence that other's published studies have a huge marginal difference in accuracy with those studies which used augmentation techniques. The simple use of different prediction algorithms, including ensemble techniques, CNN models, and transfer learning models can not overcome the problem of overfitting which led to an extensive disparity among the training and testing accuracy. There are many possible ways to conquer this issue, one of them is regularization discussed in Ref. [50] and augmentation. The process of data enhancement still was not much success to get the state-of-the-art results on these diversified datasets as mentioned in Refs. [24,51]. The reason is that there is no specific augmentation technique has been developed for spectrogram images to accomplish exceptional results. The TAA transformation approaches are not suitable for such a task, as spectrogram images don't have any specific face, front, back, etc. This study recommends NAA methodology in combination with transfer learning and using optimal discriminative learning. This research exhibits the state of the art and the highest accuracy achieved for ESC-10 (99.04%), ESC-50 (97.57%) and Us8k (99.49%).

Table 6
Comparison of the proposed approach NAA with other existing published studies on evaluated datasets.

[References] year	Methodology	Accuracy on ESC-10 in %	Accuracy on ESC-50 in %	Accuracy on Us8k in %
Results of Human Accuracy				
[48] 2015	Human Accuracy	95.7	81.3	–
Results of all used datasets Baseline models accuracy				
[48] 2015	Random forest ensemble	72.7	44.3	–
[25] 2015	CNN	80.5	64.9	73.7
Results of other's data augmentation techniques and models accuracy				
[24] 2017	DCNN + augmentation	–	–	79.0
[51] 2018	CNN + Augmentation + mix-up	91.7	83.9	83.7
[34] 2019	Multichannel input + DCNN-8 + strong augmentation	97.25	95.50	98.60
[23] 2018	EnvNet-v2 + augmentation	91.4	84.9	78.3
Results of other's Image-based methodology accuracies				
[18] 2017	Images (Combined features + Google Net)	91	73	93
[17] 2019	Images (CNN + TDSN)	56	49	–
[55] 2020	Pyramidal concatenated CNN	94.8	81.4	78.1
Results of other's audio-based methodology models accuracy				
[29] 2018	Pro-CNN (Combine features)	92.1	82.8	91.9
[32] 2016	Sound Net	92.2	74.2	–
[52] 2019	TSCNN-DS	–	–	97.2
[53] 2020	ISHMM	74	–	85.4
[54] 2018	WaveMsNet	93.7	79.1	–
[50] 2020	DCNN (no max-pool) + Log-Mel + Augmentation	94.9	89.2	95.3
Results of this study best-adopted approach NAA accuracies				
This study 2020	Proposed NAA (ResNet-152 + Freezing/Unfreezing + Discriminative learning)	99.04	97.30	99.49
This study 2020	Proposed NAA (DenseNet-161 + Freezing/Unfreezing + Discriminative learning)	98.87	97.57	99.46

6. Conclusion

The main contribution of this research includes meaningful data augmentation for the spectral images obtained from the Mel spectrogram by using DCNN and with transfer learning models together with discriminative learning for environmental sound classification. Three publicly available environmental sound classification datasets, ESC-10, ESC-50, and Us8k, are used. The experimental study involves the use of two DCNN architectures with 7-layers and 9-layers, respectively. The study also considers the implementation of transfer learning models in combination with optimal discriminative learning. The distinct pre-trained weights from (ResNet, DenseNet, AlexNet, SqueezeNet, VGG) are successfully trained on these spectral images generated by the Mel-spectrograms feature extraction for environmental sounds. This study also considered data augmentation for the above deep neural networks. In addition to the traditional data augmentation approaches used for images, meaningful data argumentation approaches are also considered in this study.

From the results obtained, the transfer learning models, ResNet-152 and DenseNet-161, together with discriminative learning based on the proposed meaningful data argumentation (NAA) can have the best performance among all. The model trained by using NAA data is much better than those trained by TAA data. For ESC-10 and Us8k datasets, the best accuracies are 99.04% and 99.49% respectively, by using ResnNet-152 for NAA data. For the ESC-50 dataset, DenseNet-161 shows 97.57% accuracy by using NAA data. As far as we know these results are the highest and the best results ever published on these environmental sound classification datasets.

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.

References

[1] Weninger F, Schuller B. Audio recognition in the wild: static and dynamic classification on a real-world database of animal vocalizations. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings. p. 337–40.
[2] Laffitte P, Sodoyer D, Tatkeu C, Girin L. Deep neural networks for automatic detection of screams and shouted speech in subway trains. In: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc. p. 6460–4.
[3] Intani P, Orachon T. Crime warning system using image and sound processing no. Iccas. In: International Conference on Control, Automation and Systems. p. 1751–3.
[4] Alsouda Y, Pllana S, Kurti A. IoT-based urban noise identification using machine learning: performance of SVM, KNN, bagging, and random forest. In: ACM Int. Conf. Proceeding Ser.. p. 62–7.
[5] Vacher M, Istrate D, Besacier L, Serignat J, Castelli E. Sound detection and classification for medical telesurvey. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014.
[6] Deng JD, Simmermacher C, Cranefield S. A study on feature analysis for musical instrument classification. IEEE Trans Syst Man Cybern B Cybern 2008;38 (2):429–38.
[7] Ali H, Tran SN, Benetos E, d'Avila Garcez AS. Speaker recognition with hybrid features from a deep belief network. Neural Comput Appl 2018;29(6):13–9.
[8] Choi K, Fazekas G, Sandler M, Cho K. Transfer learning for music classification and regression tasks. Proceedings of the 18th ISMIR Conference, Suzhou, China, 2017.
[9] Chachada S, Kuo CCJ. Environmental sound recognition: a survey. APSIPA Trans Signal Inf Process 2014;3(2014).
[10] Lagrange M, Lafay G, Défréville B, Aucouturier J-J. The bag-of-frames approach: a not so sufficient model for urban soundscapes. J Acoust Soc Am 2015;138(5): EL487–92.
[11] Phan H, Hertel L, Maass M, Mazur R, Mertins A. Learning representations for nonspeech audio events through their similarities to speech patterns. IEEE/ACM Trans Audio Speech Lang Process 2016;24(4):807–22.
[12] Crocco M, Cristani M, Trucco A, Murino V. Audio surveillance. ACM Comput Surv 2016;48(4):1–46.
[13] Cotton CV, Ellis DPW. Spectral vs. spectro-temporal features for acoustic event detection. In: IEEE workshop on applications of signal processing to audio and acoustics. p. 69–72.
[14] Li J, Dai W, Metze F, Qu S, Das S. A comparison of Deep Learning methods for environmental sound detection. In: ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings. p. 126–30.
[15] Valero X, Alias F. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. IEEE Trans Multimed 2012;14 (6):1684–9.
[16] Geiger JT, Helwani K. Improving event detection for audio surveillance using Gabor filterbank features. In: 2015 23rd Eur. Signal Process. Conf. EUSIPCO 2015. p. 714–8.

[17] Khamparia A, Gupta D, Nguyen NG, Khanna A, Pandey B, Tiwari P. Sound classification using convolutional neural network and tensor deep stacking network. IEEE Access 2019;7:7717–27.
[18] Boddapati V, Petef A, Rasmusson J, Lundberg L. Classifying environmental sounds using image recognition networks. Procedia Comput Sci 2017;112:2048–56.
[19] Huzaifah M. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. In: arXiv e-prints; 2017, pp. 1–5.
[20] Arora P, Haeb-Umbach R. A study on transfer learning for acoustic event detection in a real life scenario. In: 2017 IEEE 19th International Workshop on Multimedia Signal Processing, MMSP 2017. p. 1–6.
[21] Smith LN. Cyclical learning rates for training neural networks no. April. In: Proceedings – 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017. p. 464–72.
[22] Ying X. An overview of overfitting and its solutions. J Phys Conf Ser 2019;1168 (2).
[23] Tokozume Y, Ushiku Y, Harada T. Learning from between-class examples for deep sound recognition. In: ICLR. p. 1–13.
[24] Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process Lett 2017;24(3):279–83.
[25] Piczak J. Environmental sound classification with convolutional neural networks. IEEE International Workshop on Machine Learning for Signal Processing, Boston, USA, 2015.
[26] Hertel L, Phan H, Mertins A. Comparing time and frequency domain for audio event recognition using deep learning. In: Proc. Int. Jt. Conf. Neural Networks. p. 3407–11.
[27] Pillos A, Alghamidi K, Alzamel N, Pavlov V, Machanavajhala S. A real-time environmental sound recognition system for the Android Os no. September. Detect. Classif. Acoust. Scenes Events 2016, 2016.
[28] Emmanouilidou D, Gamper H. The effect of room acoustics on audio event classification. Proceedings of the 23rd International Congress on Acoustics, 9–13 September, 2019.
[29] Li S, Yao Y, Hu J, Liu G, Yao X, Hu J. An ensemble stacked convolutional neural network model for environmental event sound recognition. Appl Sci 2018;8 (7).
[30] Abdoli S, Cardinal P, Lameiras Koerich A. End-to-end environmental sound classification using a 1D convolutional neural network. Expert Syst Appl 2019;136:252–63.
[31] Agrawal DM, Sailor HB, Soni MH, Patil HA. Novel TEO-based gammatone features for environmental sound classification. In: 25th European Signal Processing Conference, EUSIPCO 2017. p. 1809–13.
[32] Aytar Y, Vondrick C, Torralba A. SoundNet: learning sound representations from unlabeled video no. Nips. In: Adv. Neural Inf. Process. Syst.. p. 892–900.
[33] Zhao H, Huang X, Liu W, Yang L. Environmental sound classification based on feature fusion. In: MATEC Web of Conferences. p. 1–5.
[34] Sharma J, Granmo O-C, Goodwin M. Environment sound classification using multiple feature channels and deep convolutional neural networks. J Latex Cl Files 2019;14(8):1–11.
[35] Chollet F. Image preprocessing – Keras documentation. GitHub. [Online]. Available: <https://keras.io/preprocessing/image/>; 2015. [accessed: 16-Nov-2019].
[36] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019;6(1).
[37] McFee B et al. librosa: Audio and music signal analysis in Python no. Scipy. In: Proceedings of the 14th Python in Science Conference. p. 18–24.
[38] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging no. NeurIPS. In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). p. 1–11.
[39] Hershey S et al. CNN architectures for large-scale audio classification. In: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc.. p. 131–5.
[40] Arandjelović R, Zisserman A. Objects that sound. In: Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). LNCS; 2018. p. 451–66.
[41] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.. p. 770–8.
[42] Huang G, Liu Z, Van DerMaaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. p. 2261–9.
[43] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. 50 X fewer parameters and <0.5Mb model size. In: Iclr. p. 1–13.
[44] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: ICLR. p. 1–14.
[45] George AP, Powell WB. Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming. Mach Learn 2006;65 (1):167–98.
[46] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 2011;12:2121–59.
[47] Howard J, others. vision.learner | fastai, GitHub. [Online]. Available: <https://docs.fast.ai/vision.learner.html>; 2018. [accessed: 26-Feb-2020].
[48] Piczak KJ. ESC: dataset for environmental sound classification. In: MM 2015 – Proceedings of the 2015 ACM Multimedia Conference. p. 1015–8.
[49] Salamon Justin, Jacoby Christopher, Bello Juan Pablo. A dataset and taxonomy for urban sound research no. 3. In: MM'14 Proceedings of the 22nd ACM International Conference on Multimedia. p. 1041–4.
[50] Mushtaq Z, Su SF. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Appl Acoust 2020;167:107389.
[51] Zhang Z, Xu S, Cao S, Zhang S. Deep Convolutional Neural Network with mixup for environmental sound classification. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). p. 356–67.
[52] Su Y, Zhang K, Wang J, Madani K. Environment sound classification using a two-stream CNN based on decision-level fusion. Sensors (Switzerland) 2019;19(7):1–15.
[53] Chandrakala S, Jayalakshmi SL. Generative model-driven representation learning in a hybrid framework for environmental audio scene and sound event recognition no. c. In: IEEE Trans. Multimed.. p. 1.
[54] Zhu B et al. Learning environmental sounds with multi-scale convolutional neural network. Proceedings of the International Joint Conference on Neural Networks, 2018.
[55] Demir Fatih, Turkoglu Muammer, Aslan Muzaffer, Sengur Abdulkadir. A new payramidal concatenated CNN approach for environmental sound classification. Applied Acoustics 2020;170:107520. https://doi.org/10.1016/j.apacoust.2020.107520.