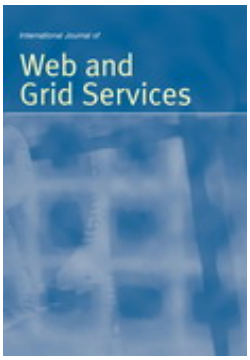
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Web and Grid Services

ISSN online: 1741-1114 - ISSN print: 1741-1106
https://www.inderscience.com/ijwgs

Simplified swarm optimisation for CNN hyperparameters: a sound classification approach

Zhenyao Liu, Wei-Chang Yeh

DOI: 10.1504/IJWGS.2024.10062236

Article History:

Received:	20 September 2023
Last revised:	31 October 2023
Accepted:	13 December 2023
Published online:	25 March 2024

Simplified swarm optimisation for CNN hyperparameters: a sound classification approach

Zhenyao Liu* and Wei-Chang Yeh

Integration and Collaboration Laboratory,
Department of Industrial Engineering and Management Engineering,
National Tsing Hua University,
Hsinchu, Taiwan
Email: liuzhenyao49@gmail.com
Email: yeh@ieee.org
*Corresponding author

Abstract: The pervasive integration of environmental sounds into diverse aspects of daily life – ranging from smart city management, accurate location pinpointing, surveillance mechanisms, auditory machine functionalities, to environmental monitoring – is evident. Central to this is environmental sound classification, gaining academic traction. However, sound classifications present challenges due to the variables causing noise. This research aimed to discern the convolutional neural network (CNN) model with optimal accuracy in ESC tasks via hyperparameter optimisation. Simplified swarm optimisation (SSO) algorithm was harnessed to encapsulate the CNN architecture, providing an untransformed representation of CNN hyperparameters during optimisation. Utilising the prominent datasets and applying data augmentation techniques, the CNN model designed via SSO achieved accuracies of 99.01%, 97.42%, and 98.96% respectively. Compared to prior studies, this denotes the highest accuracy from a pure CNN model, advancing automated CNN design for urban sound classification.

Keywords: convolutional neural network; CNN; simplified swarm optimisation; SSO; environmental sound classification; ESC; hyperparameter optimisation.

Reference to this paper should be made as follows: Liu, Z. and Yeh, W-C. (2024) 'Simplified swarm optimisation for CNN hyperparameters: a sound classification approach', *Int. J. Web and Grid Services*, Vol. 20, No. 1, pp.93–113.

Biographical notes: Zhenyao Liu is a PhD candidate in the Department of Industrial Engineering and Engineering Management at National Tsing Hua University in Taiwan. His research focuses on machine learning and soft computing.

Wei-Chang Yeh is Chair Professor in the Department of Industrial Engineering and Engineering Management at National Tsing Hua University in Taiwan, specialising in reliability, machine learning, and soft computing.

1 Introduction

In contrast to visual data, auditory data encapsulates a higher degree of semantic information (Thwe and War, 2017). Specifically, the significance of sound data has augmented in terms of garnering environmental insights. Implementations in everyday applications necessitate the utilisation of environmental sounds, as opposed to speech and musical sounds. Consequently, there has been an ascending trend in urban sound classification research. Environmental sound classification (ESC), recognised as a pivotal task in non-speech sound classification (Mu et al., 2021), plays an integral role in numerous applications including noise pollution analyses (Maisonneuve et al., 2010; Adapa, 2019), monitoring system (Arslan and Canbolat, 2018; Greco et al., 2019; Chandrakala and Jayalakshmi, 2019), situational awareness applications (Qu et al., 2022; Khan et al., 2020; Adidarma, 2023), machine sound testing (Lyon et al., 2010) and smart cities (Viveros-Muñoz et al., 2023). A wide array of statistical and machine learning methodologies (Park and Yoo, 2020; Madhu and Suresh, 2022; Peng et al., 2023; İnik, 2023) have been explored in the ESC literature. However, compared to deep learning-based research (Jahangir et al., 2023), these methods have exhibited relatively lower rates of success. The high success rates have stimulated extensive utilisation of deep learning models in recent ESC tasks (Nicholls, 2019; Abayomi-Alli et al., 2022; Tripathi and Paul, 2022; Presannakumar and Mohamed, 2023). Subsequent sections will delve into these studies in more detail.

2 Related work

A variety of datasets have been crafted to facilitate the ESC tasks. Piczak developed ESC-10/50 datasets in 2015 (Piczak, 2015). Ribeiro (2017) deployed a convolutional neural network (CNN) model to classify three distinct datasets. Research findings suggest that, when compared to alternative existing methodologies, the CNN model demonstrates a much better performance. In a subsequent study conducted by Davis and others (Davis and Suresh, 2018), a CNN was employed for ESC classification. Additionally, data augmentation methods were implemented during CNN training phase, resulting in getting a better result of the model with augmented data.

In a research endeavour by Su et al. (2020) a strategy was put forth to combine two distinct attributes within the ESC framework, aiming for a more holistic portrayal. The Urbansound8k was developed in 2014 (Salamon et al., 2014), and the outcome was the development of a specialised four-layer CNN architecture, denoted as TSCNN-DS, meticulously designed for effective classification. Impressively, this network yielded a remarkable accuracy of 97.2% when applied to the US8K dataset. Turning attention to the intricacies posed by sound overlapping, Mushtaq and Su (2020) addressed the complexities that arise when considering multiple sound sources in relation to microphone placement. Their solution involved the utilisation of a deep-CNN in conjunction with the ESC dataset. To improve feature extraction, an examination of three techniques was conducted that resulting in significant accuracy scores of 94.94% for ESC-10, 89.28% for ECS-50, and 95.37% for the US8K. They also utilised seven/nine-layer CNN models in conjunction with the ESC-10/50 and US8K datasets, this approach underscored the effectiveness of meaningful data augmentation in enhancing acoustic information processing. Notably, the application of this technique yielded

remarkable results, with ResNet-52 achieving accuracy of 99.04% for ESC-10, and the US8K reached the accuracy of 99.49%, and the dataset ESC-50 attaining 97.57% with DenseNet-161. Yazgaç and Kırcı (2022) introduced two fractional-order calculus-based data augmentation techniques for audio signals, demonstrating their effectiveness in enhancing classification accuracy by increasing the dataset size six-fold and outperforming non-augmented datasets, thus pioneering the use of fractional-order calculus in audio data augmentation.

Takahashi et al. (2018) put forth a CNN model for the detection of acoustic events, drawing inspiration from the VggNet framework. They also suggested a novel approach for augmenting the data utilised during the training process. In the sphere of deep learning models' training phase, Tokozume et al. (2018) introduced a unique strategy for data input into the model. Termed as between-class learning, this method generates sounds intermediate between classes by blending two sounds from distinct classes at random proportions. Based on the report, this cross-class learning approach exhibits strong performance in both speech recognition networks and data augmentation. Furthermore, the report delineates an ESC classification network that has been trained utilising the suggested methodology, with reported results indicating error rates in speech recognition lower than those achieved by human performance. Peng et al. (2020) detail the critical endangerment of the Chinese white dolphin population, highlighting the urgent need for conservation actions informed by reliable data. It critiques the traditional man-on-boat-watch method for dolphin observation as inefficient, proposing an internet of things (IoT) based mechanism involving hydrophones, UAVs, and a command control system. A Monte Carlo simulation demonstrates the IoT method's superior effectiveness over the traditional approach, emphasising the innovative contribution of the study while acknowledging its limitation of using off-the-shelf rather than high-end products (Peng et al., 2020).

The ESC challenge was addressed comprehensively by Roy et al. (2022) through the deployment of an extended CNN architecture, a notable departure from the customary max-pooling approach adopted post-convolution. In parallel, an exploration into the influence of diverse expansion ratios and convolution layer quantities was undertaken to dissect their impact on the obtained outcomes. Their findings showcased the superiority of the extended CNN over its max-pooling counterpart when applied to the ESC problem. However, it was noteworthy that an undue augmentation in filter quantities and ratios negatively impacted the classification accuracy. Turning to the realm of one-dimensional (1D) CNN networks, Abdoli (2021) tackled ESC classification with a unique perspective. Their approach involved the extraction of frames from audio signals to serve as input data for the network. Thorough experimentation yielded noteworthy outcomes, evident in an average rate of 89% on the US8K. Crucially, opting for raw input data has surfaced as the prime selection among end-to-end techniques, yielding unparalleled performance. Remarkably, the proposed methodology not only outperformed several existing models in the literature but also exhibited a parsimonious parameter count, further bolstering its appeal.

In a parallel study by Ashurov et al. (2022) attention was drawn to the convolutional layer's filter sizes and activation functions within CNNs, dissecting their impact on ESC tasks. This exploration led to the formulation of a model, which demonstrated remarkable prowess across three distinct sound datasets, particularly excelling with fewer errors on the US8K dataset, setting it apart from alternative approaches. In a different vein, Lim

et al. (2017) presented a CNN-centric methodology tailored for sound event classification. Their approach effectively categorised 30 different sound events across a diverse array of datasets, culminating in an impressive accuracy rate of 81.5%. Shifting focus to spatial aspects, Tuncer et al. (2020) embarked on ESC classification using innovative spiral patterns and a unique 2D-4M pooling approach. Kuang et al. (2021) introduced an innovative web of things (WoT) concept aimed at enhancing police anti-terrorism operations, validated through Monte Carlo simulations comparing the traditional and WoT-based police rescue squad forces (RSF). The findings demonstrate that the WoT-based RSF significantly outperforms the current approach, offering commanders a more effective, less risky, and cost-efficient method for executing anti-terrorist and hostage rescue missions (Kuang et al., 2021). The impact of their method was assessed across the ESC-10 and ESC-50 datasets, revealing elevated levels. Concurrently, Dogan et al.'s (2020) work offered a robust feature extraction strategy targeting the identification of activity locations within environmental sounds. This multi-stage method encompassed feature creation, selection, and classification. Deployed on the ESC-10 dataset, this approach achieved a commendable accuracy rate of 90.25%. Venturing into noise-affected domains, Gontier et al. (2021) conducted an intricate exploration to mitigate the distortive effects of noise on high-accuracy deep learning models. Their study encompassed analyses of attacks on CNN-classified ESC data and led to the creation of benchmark datasets tailored for such investigations. Lastly, İnik's contribution introduced a PSO-based CNN hyperparameter optimisation method applicable to ESC and US8K datasets. Impressively, this technique yielded accuracy rates of 98.64%, 93.71%, and 98.45% respectively, showcasing its efficacy across diverse datasets (İnik, 2016). Sangaiah et al. (2023) introduce an intelligent method for dynamic resource allocation in cloud computing using Takagi-Sugeno-Kang (TSK) neural-fuzzy systems and ant colony optimisation (ACO) to enhance efficiency and reduce energy consumption. The method optimises virtual machine migration and resource allocation, demonstrating improved performance in various criteria, including energy efficiency and request handling, compared to existing approaches (Sangaiah et al., 2023). Khanduzi and Sangaiah (2023) introduce a recurrent neural network (RNN) to effectively solve the continuous defensive location problem (CDLP), a complex bilevel mathematical model for placing defence facilities. The RNN outperforms existing methods like tabu search and imperialist competitive algorithm in terms of execution time, precision, and keeping attackers further from critical points, particularly in smaller-sized problems where it matches the results of exact methods with less runtime (Khanduzi and Sangaiah, 2023). Sakamoto et al. (2019) discusses the evolution of networks, particularly the emergence of wireless mesh networks (WMNs) which, despite their robustness and ease of maintenance, face challenges like the NP-hard node placement problem. The work introduces WMN-PSOSA, an intelligent hybrid simulation system combining particle swarm optimisation (PSO) and simulated annealing (SA), and demonstrates its effectiveness, particularly for Weibull distributions, through performance evaluations involving different mesh client distributions (Sakamoto et al., 2019).

Mkrtchian and Furletov (2022) explored the application of AlexNet and GoogLeNet architectures for the classification tasks involving the ESC and US8K datasets. Their methodology involved the transformation of audio signals within these datasets into image representations through the utilisation of spectrogram, MFCC, and CRP techniques. These transformed images then underwent classification processes using deep learning models. In a distinct vein, Ahmed et al. (2020) introduced an inventive stacked

CNN model that leveraged a unique approach. Their model incorporated multiple convolutional layers with reduced filter numbers, forming a distinctive network structure. Notably, their method entailed training two separate CNN networks using the original raw audio signal data. The integration of these networks culminating in the creation of a novel CNN model termed DS-CNN. This innovative model exhibited remarkable performance improvements in contrast to other CNN-based architectures across the ESC and US8K datasets.

In general, deep learning models have demonstrated superior performance in achieving successful results for ESC compared to other artificial intelligence approaches. This is primarily attributed to the innate capacity of DL models to autonomously uncover pertinent features from the input data. However, the design of these models requires the adjustment of numerous parameters, which can be divided into two parts: optimisation parameters and model design parameters. On the positive side, this parameter adjustability allows different researchers to design diverse models tailored to address specific problems. Conversely, it also presents a challenge in finding the optimal model configuration for a given problem. To address this, efforts have been made to enhance CNN models for ESC by employing various algorithms for CNN parameter optimisation. Notably, simplified swarm optimisation (SSO), GA and PSO have all proven effective in fine-tuning CNN model parameters, yielding highly satisfactory results (İnik, 2023; Yeh et al., 2023; Sun et al., 2020). While CNN models have garnered popularity for ESC tasks, there is a notable dearth of studies investigating the optimal parameter configurations and layers configurations for CNNs in the realm of ESC (Mohaimenuzzaman et al., 2023; Bahmei et al., 2022; Fang et al., 2022; Zhang et al., 2017, 2021; Mushtaq et al., 2021; Li et al., 2019). As a result, the primary goal of this research is to optimise the CNN hyperparameter s to achieve the highest accuracy for urban sound classification. In this paper, the primary objective is to compare with İnik's method proposed in 2023 and take reference from it for the audio transformation methodology and CNN hyperparameter configuration.

Beyond the utilisation of signal processing approaches, traditional machine learning methods in urban ESC, this study presents the following noteworthy contributions:

1 This research is the first one to employ SSO to optimise hyperparameters of CNN applied in classification of urban sound.

2 The SSO algorithm is demonstrated to be suitable for CNN parameter optimisation. SSO has been proven to be more effective and simpler than PSO in many fields (Yeh et al., 2023).

3 By optimising the parameters in this research, we have developed an automatic CNN model capable of classifying environmental sound data presented in various image formats.

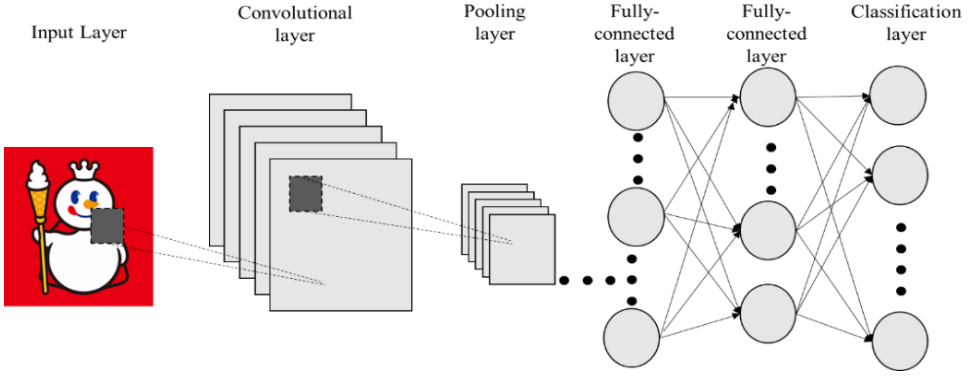
This research is organised as follows. Section 3 offers insights into CNN, SSO, and details about the datasets. Section 4 outlines the proposed method, elucidating the application of SSO of optimisation of the CNN model. This section also highlights our approach's achievement of the highest classification accuracy for each dataset. Section 5 shows the results of the experiments. Last, Section 6 provides an explanation of the conclusion.

3 Overview of CNN, SSO and datasets

3.1 Convolutional neural networks

CNNs are fundamental to deep learning, celebrated for their ability to uncover unique features in unprocessed data during the learning phase. They excel in a spectrum of tasks encompassing classification, recognition, and segmentation, making them indispensable across various domains, including engineering, medicine, and defence industries. The ascendancy of CNNs has surged, particularly in the era of big data, where their innate ability for automatic feature discovery has proven invaluable. CNNs constitute deep networks characterised by a cascade of layers that progressively extract feature maps during network training. Moreover, researchers enjoy the flexibility to tailor the number of layers and model parameters, granting them extensive experimental avenues to optimise model performance. The fundamental CNN structure is shown in Figure 1. Within the CNN architecture, the input, convolutional and pooling layer are dedicated to extract the features, while the fully connected layers are responsible for classification. Within the operational framework of a CNN, the input traverses through segments composed of convolutional and pooling layers before ultimately arriving at the fully connected layer. After that, output is produced, which is subsequently contrasted with the anticipated results, with deviations being interpreted as errors. Weight updates are iteratively executed via a gradient-based backpropagation algorithm to mitigate these discrepancies. Network training persists until the predefined epoch value is attained.

Figure 1 Basic CNN architectures (see online version for colours)



3.2 Simplified swarm optimisation

The CNN model can bring significant improvements in many image classification tasks. To achieve the best model performance, a substantial number of hyperparameters within the CNN architecture need to be adjusted and optimised (Tuba et al., 2021). The optimisation of these hyperparameters is an NP-hard problem, and in the past, it was often tackled through guessing and estimating methods or empirical rules. However, these approaches are typically complex and do not guarantee finding the optimal solution. By utilising heuristic algorithms from swarm intelligence, it is possible to find approximate optimal solutions within reasonable time.

SSO (Yeh, 2009) is a heuristic algorithm proposed by scholar Yeh in 2009. It builds upon the concept of PSO introduced by Poli et al. (2007). The main goal of SSO is to overcome the limitations of PSO in solving problems with discrete variables and improve its issue of premature convergence.

Each particle in SSO is encoded with positive integers and possesses a feasible system structure. For different problems, different solution update mechanisms can be generated through step functions, as shown in equation (1). In the equation, C_g, C_p, C_w are hyperparameters, and their relationship is $0 < C_g < C_p < C_w < 1$. Additionally, ρ represents a uniformly distributed random variable between 0 and 1, while x_{ij}^{t+1} represents the j^{th} variable of the i^{th} solution in the $(t+1)^{\text{th}}$ iteration.

$$x_{i,j}^{t+1} = \begin{cases} x_{i,j}^{t} \ if \ \rho_{[0,1]} \in [0, C_g) \\ p_{i,j}^{t} \ if \ \rho_{[0,1]} \in [C_g, C_p) \\ g_j \ if \ \rho_{[0,1]} \in [C_p, C_w) \\ x \ if \ \rho_{[0,1]} \in [C_w, 1) \end{cases} \tag{1}$$

The updating mechanism of SSO encompasses four scenarios: when ρ falls in the range $[0, C_w)$, the value of x_{ij}^{t+1} remains the same as x_{ij}^{t}, which represents the value of the j^{th} variable in the i^{th} solution during the t^{th} iteration. When ρ falls in the range $[C_w, C_p)$, x_{ij}^{t+1} takes the value of $p_{i,j}^{t}$, denoting the personal best (*pbest*) solution for the j^{th} variable within the current i^{th} solution. When ρ falls in the range $[C_p, C_g)$, x_{ij}^{t+1} is set to g_i, representing the global best (*gbest*) solution for the j^{th} variable across all solutions. Lastly, when ρ falls in the range $[C_g, 1)$, x_{ij}^{t+1} is assigned a randomly generated value x, aiming to enhance the diversity of solutions and allow them to move towards the global best solution.

By selecting different values of the hyperparameters C_w, C_p and C_g distinct updating outcomes can be generated, leading to diverse optimisation solutions. Considering the significance of parameter selection in determining solution quality, SSO often employs orthogonal arrays (OA) to select the optimal parameter combinations (Yeh, 2014; Dusberger and Raidl, 2015; Chen et al., 2016). SSO has found various applications in different fields, such as breast cancer feature classification problems (Yeh et al., 2009), training of neural network models (Yeh, 2013; Yeh et al., 2021), quantum computing (Su et al., 2022), reliability redundancy allocation problems (Jiang et al., 2023) and close-loop supply chain network, etc. (Liu et al., 2023; Yeh et al., 2023). These studies have demonstrated that is effective in handling discrete variable problems and possesses the ability to generate high-quality solutions.

3.3 Datasets

In recent years, the classification of environmental sounds has extensively relied on three distinct datasets, namely ESC10, ESC50, and UrbanSound8K. Below, we provide detailed information about each of these datasets:

- UrbanSound8K (US8K): curated by Salamon et al. (2014) comprising 8,732 labelled sound samples. Each audio recording within this dataset spans approximately 4

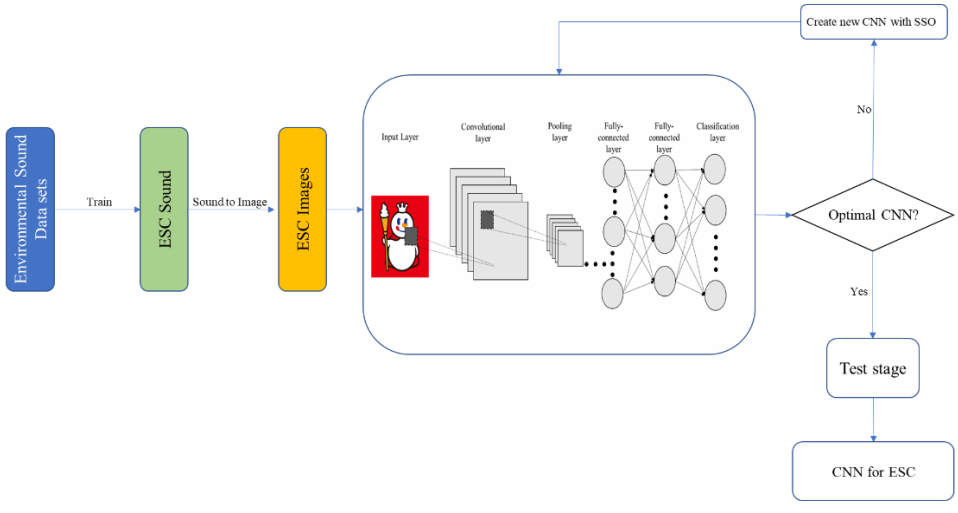
seconds. US8K encompasses ten distinct sound classes, including gunshot, police siren, car horn, jackhammer, drilling, engine idling, air conditioner, dog barking, playing children, and street music.

- ESC-10: developed by Piczak (2015) comprises ten classes of environmental sounds. On average, each entry in this dataset has a duration of 5 seconds, with an average of 40 entries per category. It's important to highlight that the ESC-10 is the subset of the ESC-50.
- ESC-50: an expanded version of ESC-10, offering a more diverse soundscape for classification. It encompasses 50 distinct sound classes, each with 40 recordings, each lasting 5 seconds. This result in a total of 2,000 sound segments sampled from various urban environments. These categories provide a comprehensive and diverse set of sound data for research and analysis.

4 Proposed approach

The structure of the method we propose is illustrated in Figure 2. In our approach, we initiate the process by transforming the audio data into image representation using the Scalogram technique (Peng and Chu, 2004). Elaboration on the fine-tuning of hyperparameters within the CNN model can be found in Section 4.2.

Figure 2 Architecture of the proposed method (see online version for colours)

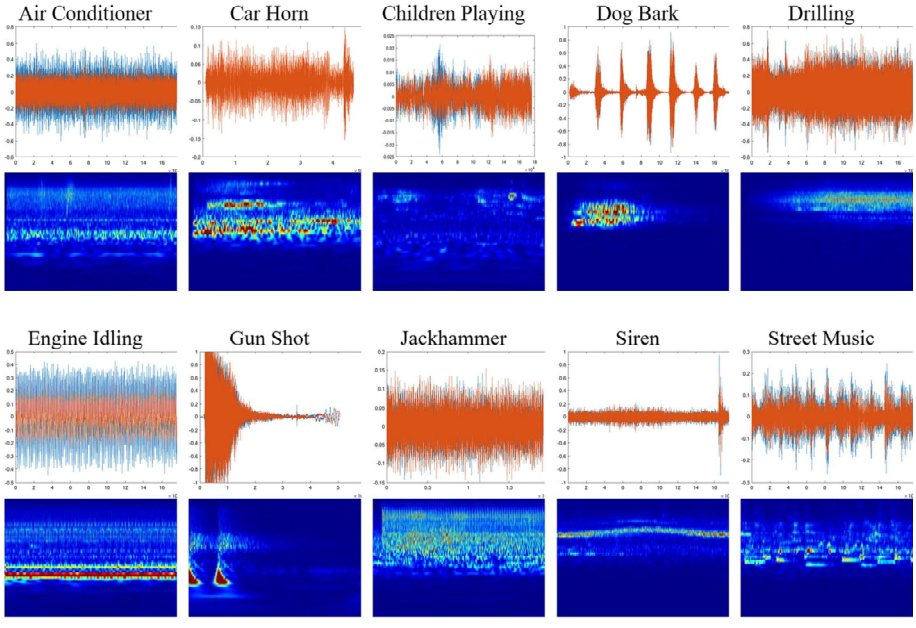


4.1 Dataset pre-processing

Within the investigation, sound data underwent transformation from a signal to an image configuration utilising the scalogram approach. Scalogram, which portrays the absolute magnitude of a signal's continuous wavelet transform in relation to both time and frequency, was harnessed for this purpose. The conversion procedure was facilitated through the application of the wavelet toolbox software suite. To illustrate, Figure 3

showcases instances of sound recordings from the US8K dataset that have been transformed into the image domain (İnik, 2023). Since the models obtained from comparative literature's experimental studies have the increased depth, a similar need for more data for training these models has arisen. As a result, the models in this article have been trained not only using the original dataset but also employing augmented datasets. Data augmentation techniques of translation and flipping were applied. By utilising these techniques in both vertical, horizontal, and vertical-horizontal directions, each approach led to a threefold increase in the dataset, resulting in an overall six-fold increase.

Figure 3 Example of sound (top) and transformed image (bottom) from the US8K dataset categories (see online version for colours)



Source: İnik (2023)

4.2 CNN hyperparameter optimisation

The aim of optimising CNN parameters is to identify the most appropriate settings, ultimately leading to the best attainable accuracy that suitable for a specific task. Nonetheless, this endeavour poses significant challenges owing to the vast array of parameters necessitating fine-tuning, as well as the computational intensity associated with the process. Consequently, it becomes imperative to adopt optimisation algorithms that can minimise the number of iterations. In this research, SSO algorithm is employed to identify the optimal CNN model that achieves peak accuracy in urban sound classification.

Through SSO, the six hyperparameters to be optimised can be represented as a set of solutions, where each solution consists of seven variables, each representing a different hyperparameter. The research's solution encoding is illustrated in Figure 4. During the SSO iteration and solution updating process, N_{sol} sets of solutions will be generated. Each

set of solutions comprises six variables, resulting in a 6-dimensional solution space. Furthermore, different variables have defined upper and lower bounds. $\boldsymbol{x_1}$ represents the number of layers, and their variable range is [3, 15]. $\boldsymbol{x_2}$ represents the number of filters in a convolutional layer, and the variable range is [16, 256]. $\boldsymbol{x_3}$ represents the size of the kernel in a convolutional layer, and the variable range is [2, 11]. $\boldsymbol{x_4}$ represents the size of the kernel in a pooling layer, and the variable range is [2, 7]. $\boldsymbol{x_5}$ represents the size of the stride in a pooling layer, and the variable range is [2, 7]. Randomly deactivating neurons and input connections in the fully connected layers helps to avoid overfitting. $\boldsymbol{x_6}$ is a value between 10 and 1,024 used to determine the neuron number in fully-connected layers, a higher number of neurons allows the model to represent a larger function space, leading to better data fitting. However, increasing the number of neurons also comes with higher computational costs and overfitting risks. Specifically, a broad range of parameters has been employed to explore CNN models extensively, aiming to identify the optimal architecture. The model architecture includes a maximum of 15 layers, with a minimum requirement of three layers, including an obligatory fully connected layer. Notably, the final position in the layer sequence is always reserved for the fully connected layer. The layers encompass convolutional, pooling, and fully connected elements. Complementary ReLu layers follow the convolutional stages, while dropout layers are automatically integrated after the fully connected layers. The purpose of dropout layers is to prevent overfitting. The probability in this layer indicates the ratio of deactivated neurons in the fully connected layers. The hyperparameters and ranges are shown in Table 1.

Figure 4 Encoding of SSO

x_1	x_2	x_3	x_4	x_5	x_6

Table 1 Hyperparameters and ranges represented by SSO solution variables

Variables	*Hyperparameters*	*Range*
$\boldsymbol{x_1}$	Number of layers	[3, 15]
$\boldsymbol{x_2}$	Filter numbers in the convolutional layers	[16, 256]
$\boldsymbol{x_3}$	Size of the kernels in the convolutional layers	[2, 11]
$\boldsymbol{x_4}$	Size of the kernels in the pooling layers	[2, 7]
$\boldsymbol{x_5}$	Size of the strides in the pooling layers	[2, 7]
$\boldsymbol{X_6}$	Number of the neurons in the fully connected layers	[10, 1024]

Fitness value for SSO is computed by the formula presented in equations (2)–(3). Based on the design of its fitness function, it belongs to a maximisation problem. Initially, the seven variables are randomly initialised, and the fitness function value of the initial solution is computed to obtain the initial global best solution, denoted as 'gbest'. Subsequently, the SSO iteratively updates and searches for the optimal solution. To enhance the efficiency of the classification of sound, the maximum number of iterations, denoted as N_{gen}, will be used as the termination condition for updating in SSO. When the number of iterations exceeds the predefined maximum number of iterations, SSO will be terminated.

$$Fitness\ function = Accuracy = \sum_{i=1}^{N_{\text{test}}} \frac{a_i}{N_{\text{test}}}, \tag{2}$$

where

$$a_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ sample predicted correctly} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

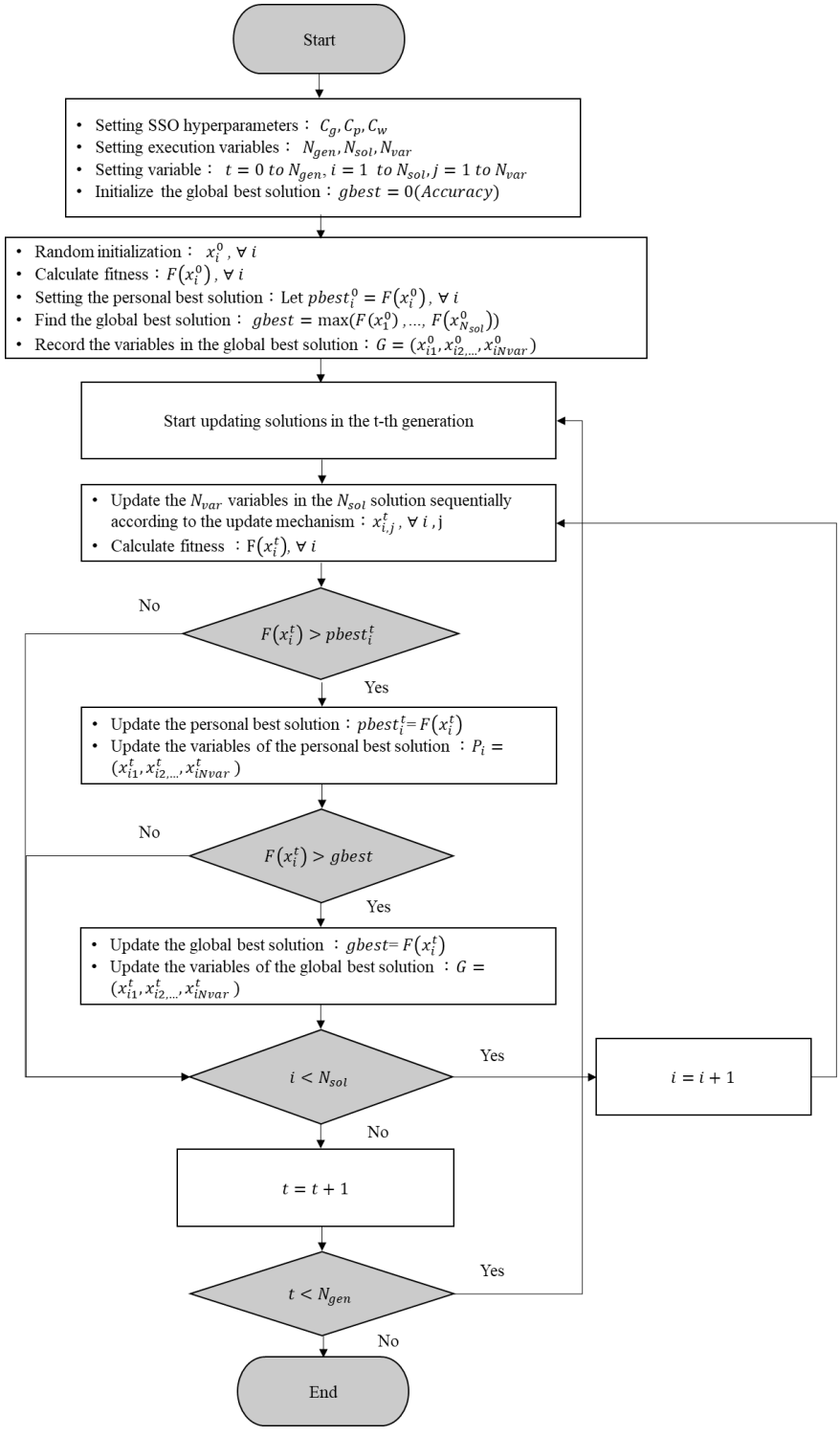
And N_{test} represent the size of the testing data.

This study employs the global best (*gbest*) and personal best (*pbest*) mechanisms with SSO to update all variables in each iteration for every solution. Following the update of all variables, the fitness function value is computed. Subsequently, a comparison is made with *gbest* and *pbest*, and finally, *gbest* and *pbest* are updated. The notations and flowchart of SSO are presented in Table 2 and Figure 5.

Table 2 SSO notations

Notations	*Definition*
N_{var}	The number of variables
N_{sol}	The number of solutions
N_{gen}	The number of generations
N_{run}	The number of experiments
t	Denotes N_{gen}, $t = 0, 1, 2 \ldots, N_{gen}$
i	Denotes N_{sol}, $I = 1, 2 \ldots N_{sol}$
j	Denotes N_{var}, $j = 1, 2 \ldots N_{var}$
$x_{i,j}^t$	The j^{th} variable of Xi in the t^{th} generation.
x_i^t	$x_i^t = \left(x_{i1}^t, x_{i2}^t, \ldots x_{iNvar}^t\right)$ denotes the variables of the i^{th} solution in the t^{th} generation.
$F(x_i^t)$	A fitness function is used to calculate the fitness value of each solution.
gbest	Denotes the value of the global best solution.
$pbest_i^t$	Denotes the value of the best i^{th} solution in the t^{th} generation.
G	$G = \left(x_{i1}^t, x_{i2}^t, \ldots x_{iNvar}^t\right)$ denotes the variables of the global best solution in its evolutionary history.
P_i	$P_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,Nvar})$ denotes the variables of the best i^{th} solution in its evolutionary history.
ρ_i^t	$\rho_i^t = \left(\rho_{i1}^t, \rho_{i2}^t, \ldots \rho_{iNvar}^t\right)$, which ρ_{iNvar}^t means a random number generated within [0, 1] uniformly.
Cg, Cp, Cw	The predefined hyperparameters.

Figure 5 SSO flowchart



5 Experiment results and comparisons

Section 5 provides details regarding the training of CNN models using SSO and presents the corresponding test results achieved by these models. The experimental investigations were conducted on a computer equipped with Windows 11, an Intel® Core™ i7-11700KF processor operating at 3.60 GHz with 16 cores, 64 GB of RAM, and GPU is an NVDIA GeForce RTX3060Ti. The software platform employed for the experiments was MATLAB R2023a 64-bit (win64).

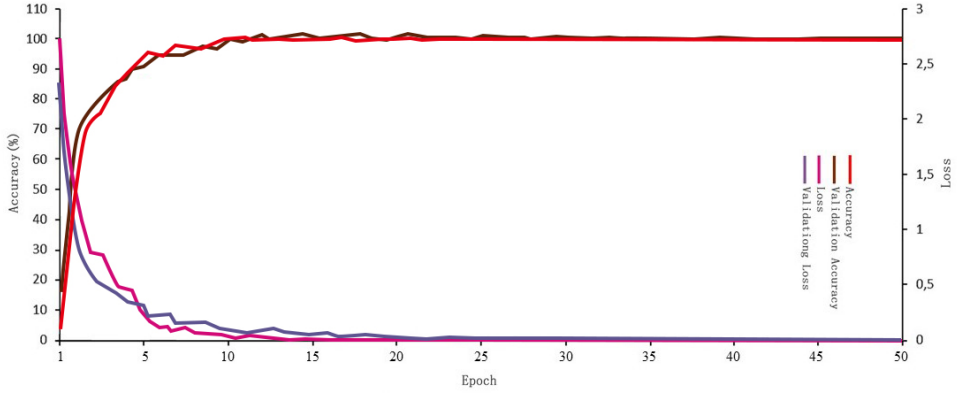
5.1 Parameters configuration in the proposed approach

The best hyperparameters of SSO, C_g, C_p and C_w will be set to 0.4, 0.6, and 0.9, respectively. And the other settings are presented in Table 3.

Table 3 Training configuration

Parameter	*Values*
Optimiser	SGD with momentum
Epoch	10
Dropout	0.5
Batch size	256
Learning rate	1×10^{-3}
Loss function	Cross entropy
(C_g, C_p, C_w)	(0.4, 0.6, 0.9)
N_{sol}	50
N_{gen}	20
N_{run}	30

Figure 6 Convergence graph of SSOC-ESC10 training model (see online version for colours)



5.2 Experiment training and results

The CNN models underwent a training regimen utilising a 10-fold cross-validation method. Figures 6, 7, and 8 illustrate the convergence trajectories for SSOC-ESC10, SSOC-ESC 50, and SSOC-US8K, respectively, during this training phase. A comparative analysis of these figures reveals a striking proximity between the accuracy/validation and error/validation error curves, suggesting that the models exhibit minimal overfitting throughout the training process. Notably, the CNN-ESC10 model demonstrated a more rapid convergence compared to its counterparts.

Figure 7 Convergence graph of SSOC-ESC50 training model (see online version for colours)

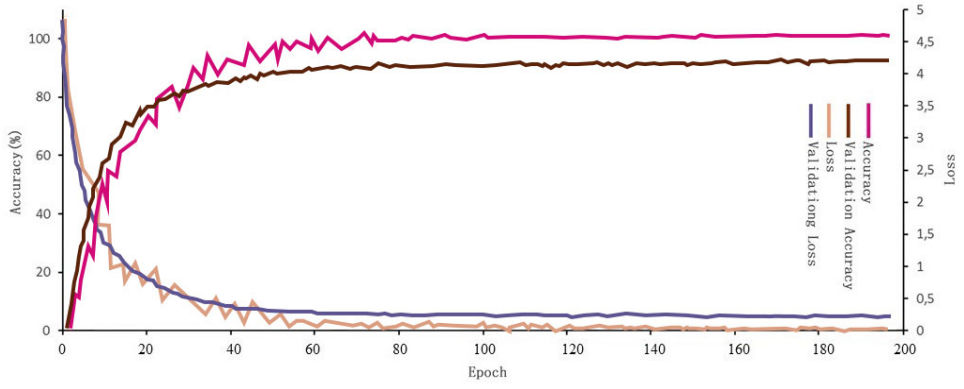


Figure 8 Convergence graph of SSOC-US8K training model (see online version for colours)

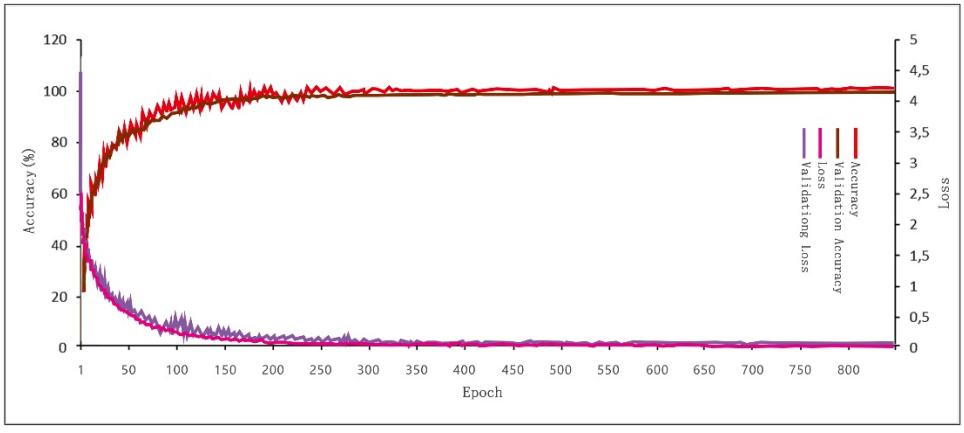


Table 4 presents comparison of results before and after data augmentation of the models tested on various datasets. It details results both with and without data augmentation for each dataset. For ESC and US8K datasets, the averages without data augmentation stood at 89.61%, 77.82%, and 93.03% respectively. However, with data augmentation, these values rose to 99.37%, 97.69%, and 99.23%. Notably, data augmentation led to significantly improved outcomes, with the most pronounced improvement observed in the ESC-50 dataset. This can be attributed to its 50 classes and limited training data,

restricting the model's ability to optimally update its weights. As the data volume increased, the performance for the 50 classes notably improved.

Table 4 Comparison of results before and after data augmentation

Proposed model	*Data augmentation*	*Mean accuracy*	*Max accuracy*	*Min accuracy*	*Standard deviation*
SSOC-ESC10	No	89.61	96.33	81.23	3.98
SSOC-ESC10	Yes	99.01	99.21	98.88	0.38
SSOC-ESC50	No	77.82	84.12	73.54	2.76
SSOC-ESC50	Yes	97.42	98.03	97.29	0.29
SSOC-US8K	No	93.03	94.96	90.11	1.12
SSOC-US8K	Yes	98.96	99.27	98.43	0.12

5.3 Comparison with other researches

Several deep learning studies have been undertaken using the ESC dataset. A comparison of the average accuracy from our proposed method against other researches is showcased in Table 5. In this table, the symbol '/' indicates that a particular method was not applied to the corresponding dataset. Our observations reveal that the proposed CNN models outshine the baseline models referenced across all datasets (Piczak, 2015; Salamon et al., 2014). The use of transfer learning in our study likely contributes to this elevated accuracy. In this technique, models trained on a separate dataset with millions of entries are subsequently applied to urban sounds, boosting their performance.

Table 5 Experimental results comparison table

Ref	*Approach*	*US8K*	*ESC10*	*ESC50*
Salamon et al. (2014)	Support vector machine (baseline)	68.00	/	/
Piczak (2015)	RFE (Baseline)	/	72.70	44.30
Piczak (2015)	HP	/	95.70	81.30
Boddapati et al. (2017)	AlexNet, GoogLeNet, CRNN	93.00	91.00	73.00
Zhang et al. (2017)	d-CNN(LeakyReLU)	81.90	/	68.10
Salamon and Bello (2017)	DCNN + data augmentation	79.00	/	/
Mushtaq et al. (2021)	ResNet-152, DenseNet-161	99.49	99.04	97.57
Li et al. (2019)	MS CNN	/	93.70	83.50
Luz et al. (2021)	Handcrafted + Deep	96.80	/	86.20
Zhang et al. (2021)	RNN	/	93.70	86.10
Medhat et al. (2020)	MCNN	74.22	85.25	66.60
İnik (2023)	CNN-PSO	98.45	98.64	96.77
This paper (SSO)	CNN-SSO	93.03	89.61	77.82
SSO with data augmentation	CNN-SSO	98.96	99.01	97.42

6 Discussion and conclusions

6.1 Discussion

Lately, the utilisation of AI techniques for urban sound classification has gained momentum. The effectiveness of models across diverse data scenarios has prompted their frequent adoption within this field. While there's an ongoing exploration into crafting more efficient CNN models for urban sound classification, a significant limitation has been the manual design of these models. The sheer number of parameters in a CNN, from layer order permutations that could number in the millions to individual layer parameters, makes manual optimisation a Herculean task. The vastness of this solution space suggests the necessity of optimisation algorithms.

In this research, we leveraged SSO algorithm to identify the most efficient CNN model for urban sound classification. A key challenge in the optimisation of CNN parameters is how to formulate a representative model. In this regard, we employed the SSO algorithm for hyperparameter optimisation of CNNs. In the context of urban sound classification, we compared the SSO-enhanced CNN model with a pure CNN model, as outlined in related studies (İnik, 2023; Zhang et al., 2017; Salamon and Bello, 2017; Chen et al., 2019). The results indicate the superiority of our CNN model over these alternatives. It is noteworthy that these models exhibit greater depth, suggesting a demand for more training data to optimise their weights. This insight prompted us to generate synthetic data, particularly for datasets like ESC-50, which possess numerous categories but limited training samples per category. For instance, upon augmenting the dataset for the enhanced SSOC-ESC50 model, accuracy escalated from 77.82% to 97.42%. This trend implies that larger training datasets can further enhance the accuracy of CNN models tailored for ESC classification. However, if exclusively trained on ESC sounds, our model lags others, notably those leveraging transfer learning. Interestingly, as observed in studies (Luz et al., 2021), ensemble strategies that incorporate diverse features outperform our proposed model in the absence of data augmentation. This commendable performance might be attributed to their integration of traditional machine learning and handcrafted features with CNN-derived features. Looking ahead, we anticipate forthcoming research to delve into more effective transfer learning techniques. This will aid in deploying our SSO-optimised CNN model on larger datasets, while minimising computational expenses during the training process.

6.2 Conclusions

In this study, SSO algorithm was leveraged to achieve optimal models for ESC and Urbansound8k, both of which are universally recognised as premier benchmarks in urban sound classification. The methodology introduced a customised adaptation of the SSO algorithm to intricately adjust CNN parameters. Remarkably, the highest performing models were identified as SSOC-ESC10, SSOC-ESC50, and SSOC-US8K, registering accuracy rates of 89.61%, 77.82%, and 93.03% for their respective datasets. It is significant to mention that these figures were obtained using raw data. To enhance the robustness of the results, a meticulous sevenfold augmentation process was applied to the datasets, after which the models were retrained. This resulted in a marked improvement in performance, with accuracy rates climbing to 99.01%, 97.42%, and 98.96% for the ESC-10/50 and US8k datasets, respectively. When juxtaposed with contemporary

research efforts, the CNN models achieved in this study for urban sound classification emerge as singularly outstanding.

References

Abayomi-Alli, O.O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M. and Misra, S. (2022) 'Data augmentation and deep learning methods in sound classification: a systematic review', *Electronics*, Vol. 11, No. 22, p.3795.

Abdoli, S. (2021) *End-to-end Deep Learning for Audio Classification: From Waveforms to a Security Perspective*, Doctoral dissertation, École de technologie supérieure).

Adapa, S. (2019) *Urban Sound Tagging using Convolutional Neural Networks*, arXiv preprint arXiv:1909.12699.

Adidarma, A.S. (2023) *Parallelizing CNN and Transformer Encoders for Audio Based Emotion Recognition in English Language*.

Ahmed, M.R., Robin, T.I. and Shafin, A.A. (2020) 'Automatic environmental sound recognition (AESR) using convolutional neural network', *International Journal of Modern Education & Computer Science*, Vol. 12, No. 5, pp.41–54.

Arslan, Y. and Canbolat, H. (2018) 'Performance of deep neural networks in audio surveillance', IEEE, pp.1–5.

Ashurov, A., Zhou, Y., Shi, L., Zhao, Y. and Liu, H. (2022) 'Classification of environmental sounds through spectrogram-dilation-based CNN', in *2022 IEEE 8th International Conference Communications (ICCC)*, IEEE, pp.1934–1938.

Bahmei, B., Birmingham, E. and Arzanpour, S. (2022) 'CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification', *IEEE Signal Processing Letters*, Vol. 29, pp.682–686.

Boddapati, V., Petef, A., Rasmusson, J. and Lundberg, L. (2017) 'Classifying environmental sounds using image recognition networks', *Procedia Computer Science*, Vol. 112, pp.2048–2056.

Chandrakala, S. and Jayalakshmi, S.L. (2019) 'Environmental audio scene and sound event recognition for autonomous surveillance: a survey and comparative studies', *ACM Computing Surveys (CSUR)*, Vol. 52, No. 3, pp.1–34.

Chen, Q., Cui, Y. and Chen, Y. (2016) 'Sequential value correction heuristic for the two-dimensional cutting stock problem with three-staged homogenous patterns', *Optimization Methods and Software*, Vol. 31, No. 1, pp.68–87.

Chen, Y., Guo, Q., Liang, X., Wang, J. and Qian, Y. (2019) 'Environmental sound classification with dilated convolutions', *Applied Acoustics*, Vol. 148, pp.123–132.

Davis, N. and Suresh, K. (2018) 'Environmental sound classification using deep convolutional neural networks and data augmentation', in *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, IEEE, December, pp.41–45.

Dogan, S., Akbal, E. and Tuncer, T. (2020) 'A novel ternary and signum kernelled linear hexadecimal pattern and hybrid feature selection based environmental sound classification method', *Measurement*, Vol. 166, p.108151.

Dusberger, F. and Raidl, G.R. (2015) 'Solving the 3-staged 2-dimensional cutting stock problem by dynamic programming and variable neighborhood search', *Electronic Notes in Discrete Mathematics*, Vol. 47, pp.133–140.

Fang, Z. Yin, B., Du, Z. and Huang, X. (2022) 'Fast environmental sound classification based on resource adaptive convolutional neural network', *Scientific Reports*, Vol. 12, No. 1, p.6599.

Gontier, F., Lostanlen, V., Lagrange, M., Fortin, N., Lavandier, C. and Petiot, J-F. (2021) 'Polyphonic training set synthesis improves self-supervised urban sound classification', *The Journal of the Acoustical Society of America*, Vol. 149, No. 6, pp.4309–4326.

Greco, A., Saggese, A., Vento, M. and Vigilante, V. (2019) 'SoReNet: a novel deep network for audio surveillance applications', in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, IEEE, October, pp.546–551.

İnik, Ö. (2023) 'CNN hyper-parameter optimization for environmental sound classification', *Applied Acoustics*, 2023/01/01, Vol. 202, p.109168, DOI: https://doi.org/10.1016/j.apacoust.2022.109168.

Jahangir, R., Nauman, M.A., Alroobaea, R., Almotiri, J., Malik, M.M. and Alzahrani, S.M. (2023) 'Deep learning-based environmental sound classification using feature fusion and data enhancement', *Computers, Materials & Continua*, Vol. 75, No. 1, pp.1069–1091.

Jiang, Y., Liu, Z., Chen, J-H., Yeh, W-C. and Huang, C-L. (2023) 'A novel binary-addition simplified swarm optimization for generalized reliability redundancy allocation problem', *Journal of Computational Design and Engineering*, Vol. 10, No. 2, pp.758–772.

Khan, A.A., Shao, J., Ali, W. and Tumrani, S. (2020) 'Content-aware summarization of broadcast sports videos: an audio-visual feature extraction approach', *Neural Processing Letters*, Vol. 52, pp.1945–1968.

Khanduzi, R. and Sangaiah, A.K. (2023) 'An efficient recurrent neural network for defensive Stackelberg game', *Journal of Computational Science*, Vol. 67, p.101970.

Kuang, C-C., Wang, K.M., Hui, L., Chang, C-Y. and Chiu, K.H. (2021) 'Data analysis of simulated WoT-based anti-crime scenario', *International Journal of Web and Grid Services*, Vol. 17, No. 1, pp.3–19.

Li, X., Chebiyyam, V. and Kirchhoff, K. (2019) *Multi-Stream Network with Temporal Attention for Environmental Sound Classification*, arXiv preprint arXiv:1901.08608.

Lim, M., Lee, D., Park, H. and Kim, J-H. (2017) 'Audio event detection using deep neural networks', *Journal of Digital Contents Society*, Vol. 18, No. 1, pp.183–190.

Liu, Z., Hu, L-M. and Yeh, W-C. (2023) 'Risk-averse two-stage stochastic programming-based closed-loop supply chain network design under uncertain demand', *Applied Soft Computing*, 1 November, Vol. 147, p.110743, DOI: https://doi.org/10.1016/j.asoc.2023.110743.

Luz, J.S., Oliveira, M.C., Araújo, F.H.D. and Magalhães, D.M.V. (2021) 'Ensemble of handcrafted and deep features for urban sound classification', *Applied Acoustics*, 1 April, Vol. 175, p.107819, DOI: https://doi.org/10.1016/j.apacoust.2020.107819.

Lyon, R.F., Rehn, M., Bengio, S., Walters, T.C. and Chechik, G. (2010) 'Sound retrieval and ranking using sparse auditory representations', *Neural Computation*, Vol. 22, No. 9, pp.2390–2416.

Madhu, A. and Suresh, K. (2022) 'EnvGAN: a GAN-based augmentation to improve environmental sound classification', *Artificial Intelligence Review*, Vol. 55, No. 8, pp.6301–6320.

Maisonneuve, N., Stevens, M. and Ochab, B. (2010) 'Participatory noise pollution monitoring using mobile phones', *Information Polity*, Vol. 15, Nos. 1–2, pp.51–71.

Medhat, F., Chesmore, D. and Robinson, J. (2020) 'Masked conditional neural networks for sound classification', *Applied Soft Computing*, Vol. 90, p.106073.

Mkrtchian, G. and Furletov, Y. (2022) 'Classification of environmental sounds using neural networks', in *2022 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO)*, IEEE, June, pp.1–4.

Mohaimenuzzaman, M., Bergmeir, C., West, I. and Meyer, B. (2023) 'Environmental sound classification on the edge: a pipeline for deep acoustic networks on extremely resource-constrained devices', *Pattern Recognition*, Vol. 133, p.109025.

Mu, W., Yin, B., Huang, X., Xu, J. and Du, Z. (2021) 'Environmental sound classification using temporal-frequency attention based convolutional neural network', *Scientific Reports*, Vol. 11, No. 1, p.21552.

Mushtaq, Z. and Su, S-F. (2020) 'Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images', *Symmetry*, Vol. 12, No. 11, p.1822,

Mushtaq, Z., Su, S-F. and Tran, Q-V. (2021) 'Spectral images based environmental sound classification using CNN with meaningful data augmentation', *Applied Acoustics*, Vol. 172, p.107581.

Nicholls, M. (2019) 'Machine learning – state of the art: the critical role that machine learning can play in advancing cardiology was outlined at a packed session at ESC 2019', *European Heart Journal*, Vol. 40, No. 45, pp.3668–3669.

Park, H. and Yoo, C.D. (2020) 'CNN-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification', *IEEE Signal Processing Letters*, Vol. 27, pp.411–415.

Peng, L. et al. (2023) 'A high accuracy and low power CNN-based environmental sound classification processor', *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 70, No. 12.

Peng, Q.M., Hui, L., Wang, K.M. and Chang, L.C. (2020) 'Effectiveness analysis of an IoT mechanism in support of monitoring Chinese white dolphins by simulation model', *The Journal of Supercomputing*, Vol. 76, pp.3847–3865.

Peng, Z.K. and Chu, F.L. (2004) 'Application of the wavelet transform in machine condition monitoring and fault diagnostics: a review with bibliography', *Mechanical Systems and Signal Processing*, 1 March, Vol. 18, No. 2, pp.199–221, DOI: https://doi.org/10.1016/S0888-3270(03)00075-X.

Piczak, K.J. (2015) 'ESC: dataset for environmental sound classification', in *Proceedings of the 23rd ACM International Conference on Multimedia*, pp.1015–1018.

Poli, R., Kennedy, J. and Blackwell, T. (2007) 'Particle swarm optimization', *Swarm Intelligence*, Vol. 1, No. 1, pp.33–57.

Presannakumar, K. and Mohamed, A. (2023) 'Deep learning based source identification of environmental audio signals using optimized convolutional neural networks', *Applied Soft Computing*, Vol. 143, p.110423.

Qu, Y., Li, X., Qin, Z. and Lu, Q. (2022) 'Acoustic scene classification based on three-dimensional multi-channel feature-correlated deep learning networks', *Scientific Reports*, Vol. 12, No. 1, p.13730.

Ribeiro, P. (2017) 'Sound classification with convolutional neural networks', *Celebration of Undergraduate Research*, Vol. 3, https://digitalcommons.oberlin.edu/cour/2017/panel_04/.

Roy, S.S., Mihalache, S.F., Pricop, E. and Rodrigues, N. (2022) 'Deep convolutional neural network for environmental sound classification via dilation', *Journal of Intelligent & Fuzzy Systems*, Vol. 43, No. 2, pp.1827–1833.

Sakamoto, S., Barolli, L. and Okamoto, S. (2019) 'WMN-PSOSA: an intelligent hybrid simulation system for WMNs and its performance evaluations', *International Journal of Web and Grid Services*, Vol. 15, No. 4, pp.353–366.

Salamon, J. and Bello, J.P. (2017) 'Deep convolutional neural networks and data augmentation for environmental sound classification', *IEEE Signal Processing Letters*, Vol. 24, No. 3, pp.279–283.

Salamon, J., Jacoby, C. and Bello, J.P. (2014) 'A dataset and taxonomy for urban sound research', in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp.1041–1044.

Sangaiah, A.K., Javadpour, A., Pinto, P., Rezaei, S. and Zhang, W. (2023) 'Enhanced resource allocation in distributed cloud using fuzzy meta-heuristics optimization', *Computer Communications*, Vol. 209, pp.14–25.

Su, P-C., Tan, S-Y., Liu, Z. and Yeh, W-C. (2022) 'A mixed-heuristic quantum-inspired simplified swarm optimization algorithm for scheduling of real-time tasks in the multiprocessor system', *Applied Soft Computing*, Vol. 131, p.109807.

Su, Y., Zhang, K., Wang, J., Zhou, D. and Madani, K. (2020) 'Performance analysis of multiple aggregated acoustic features for environment sound classification', *Applied Acoustics*, Vol. 158, p.107050.

Sun, Y., Xue, B., Zhang, M., Yen, G.G. and Lv, J. (2020) 'Automatically designing CNN architectures using the genetic algorithm for image classification', *IEEE Transactions on Cybernetics*, Vol. 50, No. 9, pp.3840–3854.

Takahashi, N., Goswami, N. and Mitsufuji, Y. (2018) 'Mmdenselstm: an efficient combination of convolutional and recurrent neural networks for audio source separation', in *2018 16th International workshop on acoustic signal enhancement (IWAENC)*, IEEE, September, pp.106–110.

Thwe, K.Z. and War, N. (2017) 'Environmental sound classification based on time-frequency representation', in *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, IEEE, June, pp.251–255.

Tokozume, Y., Ushiku, Y. and Harada, T. (2018) 'Between class learning for image classification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5486–5494.

Tripathi, A.M. and Paul, K. (2022) 'Data augmentation guided knowledge distillation for environmental sound classification', *Neurocomputing*, Vol. 489, pp.59–77.

Tuba, E., Bačanin, N., Strumberger, I. and Tuba, M. (2021) 'Convolutional neural networks hyperparameters tuning', in *Artificial Intelligence: Theory and Applications*, pp.65–84, Springer International Publishing, Cham.

Tuncer, T., Aydemir, E. and Dogan, S. (2020) 'Automated ambient recognition method based on dynamic center mirror local binary pattern: DCMLBP', *Applied Acoustics*, Vol. 161, p.107165.

Viveros-Muñoz, R. et al. (2023) 'Dataset for polyphonic sound event detection tasks in urban soundscapes: the synthetic polyphonic ambient sound source (SPASS) dataset', *Data in Brief*, Vol. 50, p.109552.

Yazgaç, B.G. and Kırcı, M. (2022) 'Fractional-order calculus-based data augmentation methods for environmental sound classification with deep learning', *Fractal and Fractional*, Vol. 6, No. 10, p.555.

Yeh, W., Liu, Z. and Tseng, K. (2023) 'Bi-objective simplified swarm optimization for fog computing task scheduling', *International Journal of Industrial Engineering Computations*, Vol. 14, No. 4, pp.723–748.

Yeh, W-C. (2009) 'A two-stage discrete particle swarm optimization for the problem of multiple multi-level redundancy allocation in series systems', *Expert Systems with Applications*, Vol. 36, No. 5, pp.9192–9200.

Yeh, W-C. (2013) 'New parameter-free simplified swarm optimization for artificial neural network training and its application in the prediction of time series', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 24, No. 4, pp.661–665.

Yeh, W-C. (2014) 'Orthogonal simplified swarm optimization for the series-parallel redundancy allocation problem with a mix of components', *Knowledge-Based Systems*, Vol. 64, pp.1–12.

Yeh, W-C., Chang, W-W. and Chung, Y.Y. (2009) 'A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method', *Expert Systems with Applications*, Vol. 36, No. 4, pp.8204–8211.

Yeh, W-C., Lin, Y-P., Liang, Y-C. and Lai, C-M. (2021) *Convolution Neural Network Hyperparameter Optimization Using Simplified Swarm Optimization*, arXiv preprint arXiv:2103.03995.

Yeh, W-C., Lin, Y-P., Liang, Y-C., Lai, C-M. and Huang, C-L. (2023) 'Simplified swarm optimization for hyperparameters of convolutional neural networks', *Computers & Industrial Engineering*, Vol. 177, p.109076.

Zhang, X., Zou, Y. and Shi, W. (2017) 'Dilated convolution neural network with LeakyReLU for environmental sound classification', in *2017 22nd International Conference on Digital Signal Processing (DSP)*, IEEE, pp.1–5.

Zhang, Z., Xu, S., Zhang, S., Qiao, T. and Cao, S. (2021) 'Attention based convolutional recurrent neural network for environmental sound classification', *Neurocomputing*, Vol. 453, pp.896–903, 17 September, DOI: https://doi.org/10.1016/j.neucom.2020.08.069.