Applied Acoustics 211 (2023) 109564

Contents lists available at ScienceDirect

Applied Acoustics

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

Investigation of the effectiveness of time-frequency domain images and acoustic features in urban sound classification

Turgut Özseven

Department of Computer Engineering, Tokat Gaziosmanpaşa University, Tokat 60100, Turkey

ARTICLE INFO

Keywords:
Smart city
Urban sound recognition
Deep learning
Audio-visual feature set
Acoustic analysis
Cepstral features
Environmental sound classification
Sound event recognition

ABSTRACT

Rapid urbanization and population growth worldwide seriously challenge building livable and sustainable cities. This increase causes the increase and diversification of urban sounds. They were transforming these sounds into information instead of just being heard, as noise plays an important role in the concept of smart cities. For this purpose, two basic methods are used to classify urban sounds. In the first of these, the sounds are processed by signal processing methods, and handcrafted features are obtained. In the other method, sounds are represented visually and classified with deep learning models. This study investigated the effect of the individual and hybrid use of features used in both approaches on the classification of urban sounds. In addition, a CNN model was created to classify hybrid features. The results obtained showed that both approaches produced successful results in classification. Among the visual representation methods, mel-spectrogram, scalogram, and spectrogram images achieved the highest classification success. Using mel-spectrogram and acoustic features and the SVM classifier positively affected accuracy. Experiments were performed on the ESC-10 and UrbanSound8k datasets. The highest accuracy for the ESC-10 was 98.33% when using the scalogram and acoustic features with the AVCNN model. The highest accuracy for UrbanSound8k was obtained as 97.70% by classifying the mel-spectrogram and acoustic features obtained from the AVCNN model with the SVM classifier.

1. Introduction

Many people live in urban areas, and this rate is increasing daily. Accordingly, people's quality of life and health are more affected by the sounds in the environment. Exposure to environmental sounds, especially noise pollution, significantly threatens human health. High noise levels (>85 dB(A)) are associated with auditory effects such as hearing loss [1] and tinnitus [2]. In comparison, prolonged exposure to low and medium levels (45–65 dB(A)) noises may cause non-auditory health problems [3–5]: sleep disorder [6,7], learning disability [8–10], behavioral and emotional disorders in children and adolescents [11], depression and anxiety [12], stronger physiological stress reactions [13], endocrine imbalance and cardiovascular disorders [7], hypertension [9,14,15], decreased work performance [16]. However, sound can also provide positive effects, such as improving a person's mood, triggering a pleasant memory of a previous experience, or encouraging one to relax and heal [17]. Regional noise maps are used to determine how much people are disturbed by noise and to evaluate environmental noise exposure. Different colored grading is used at 5 dB(A) intervals to represent different noise levels over an average time [5].

Technologies that will make cities more efficient, technologically advanced, greener, and socially inclusive are increasing daily. The smart city concept has become more popular with developing technology [18]. Along with the concept of smart cities, technological developments that facilitate human life are presented in areas such as reducing resource consumption and providing efficient urban services, environment, health, and transportation [19,20]. In the context of smart cities, environmental and human-induced urban sounds can provide important information about the city's needs. Examples of this information include traffic management by estimating the number of vehicles, air pollution forecasting, modeling virtual cities, and emergency response such as gunfire and screaming [19,21]. In future smart city buildings, more comprehensive intelligent sensing, automatic data acquisition, and the acquisition and classification of urban sound phenomena are significant [22]. In addition, smart cities offer a new approach to environmental noise assessment and monitoring by creating noise management strategies. In recent years, dynamic noise mapping, smart sensors, and soundscape approaches have provided smart noise solutions [23]. In this way, monitoring of urban sounds is performed dynamically with lower-cost IoT hardware [24] and wireless acoustic sensor networks [25,26].

E-mail address: turgut.ozseven@gop.edu.tr.

https://doi.org/10.1016/j.apacoust.2023.109564
Received 30 March 2023; Received in revised form 20 June 2023; Accepted 23 July 2023
Available online 29 July 2023

Both noise and other urban sounds are discussed under the main title of sound classification. Thus, signal processing, image processing methods, and noise or different sounds are categorized. The noise level is determined by measuring the level of the sounds detected as noise after classification.

Sound classification studies can be grouped under the main headings of automatic speech recognition, music genre classification, and available sound recognition. Classification of urban sounds is also considered within the scope of known sound recognition. However, urban sound classification is more complex due to the diversity of sound types, the distance between the sound source and the recording device, and the low signal-to-noise ratio [27]. Urban sound classification with traditional machine learning methods consists of feature extraction and classification steps. In recent years, with the use of deep learning models such as Convolutional Neural Networks (CNN), a separate action for feature extraction has yet to be used. In addition to this advantage, the increase in workload is seen as a disadvantage. The feature extraction step used with traditional methods involves processing the audio signal with digital signal processing techniques and obtaining acoustic and cepstral features of the signal. These features are classified with classifiers such as Support Vector Machine (SVM), k-NN, and Artificial Neural Networks (ANN) to categorize urban sounds [22]. The classification success of traditional methods is directly related to the feature extraction process. In addition, the signal filters used in the preprocessing step also change the values of the feature set and affect the classifier result. For CNN, the most significant factor in the classification success is the model's input data and layered architecture [28].

In this study, the performance of spectrogram (SPEC), MEL, GFCC, Constant Q-transform (CQT), and scalogram (SCL) images and acoustic/cepstral feature sets were compared for the classification of urban sounds. For this purpose, ESC-10 and UltraSound8k datasets were used. No preprocessing was applied to the sounds in the datasets while obtaining the feature sets. SVM and k-NN classifiers were used for classification with acoustic/cepstral parameters. For visual-based classification, transfer learning was performed via ShuffleNet and ResNet-18. Furthermore, a 5-layer CNN model was created, and images and acoustic features were used together as input to the model. Therefore, the contributions of this research paper are; 1) The impact of acoustic parameters used in speech recognition on urban sound classification is evaluated. 2) The classification success of image types derived from urban sounds is analyzed. 3) The performance of traditional classifiers and CNN architecture is compared. 4) Performance comparison of pre-trained networks was made. 5) For CNN, visual and acoustic features were used together for classification.

2. Related works

The urban sound classification consists of feature extraction and classification steps with traditional machine learning methods. Fast Fourier Transform (FFT), acoustic analysis, and Mel-Frequency Cepstral Coefficient (MFCC) are used to obtain feature sets. SVM classifier is mostly used with these feature sets. Also, Random Forest (RF), k-NN, Decision Tree (DT), Naive Bayes (NB), and ANN classifiers were used in a limited number of studies. Stoeckle et al. (2001) classified sounds from different sources using ANN and FFT and showed that FFT could be used to classify urban sounds [29]. SVM and ANN classifiers, Principle Component Analysis (PCA) for feature selection, and Sequential Minimal Optimization (SMO) for parameter optimization were used in the studies using acoustic analysis [30,31]. Since the datasets used in these studies were self-recorded, no performance comparison was made. The MFCC feature set was mostly used in the studies involving UrbanSound8K [32], ESC-10 [33], and ESC-50 [33] datasets, which are widely used in the literature. The study's results using these three datasets, the MFCC feature set, and five different classifiers show that the classifier performance varies according to the dataset [34]. The k-NN classifier performed better on the ESC dataset, and the NB classifier performed better on the UrbanSound8K dataset [34]. A different way of using traditional classifiers is to derive the feature set from deep learning models. Also, traditional classifiers are used by adding acoustic features to the features obtained from the deep learning model. Luz et al. (2021) used deep and acoustic feature sets with RF and SVM classifiers on ESC-50 and UrbanSound8K datasets [28]. Recent studies on urban sound classification have focused on CNN architecture. In studies using CNN, the size of the dataset is an important factor in the classification success of the model. Therefore, data augmentation methods are used to increase the size of the dataset. Data augmentation is applied to images or sounds. One of the highest achievement results in the literature was obtained using CNN and Generative adversarial network (GAN) for data augmentation on ESC-10 and UrbanSound8k dataset [35]. Piczak (2015) used a log-scaled mel-spectrogram and two-layer CNN to classify urban sounds in ESC-10, ESC-50, and UrbanSound8k datasets [36]. In addition, time-stretching and pitch-shifting methods were used for data augmentation [36]. Salamon and Bello (2017) proposed a CNN architecture for classifying environmental sounds and applied data augmentation (time stretching, pitch shifting, dynamic range compression, background noise) to the Urbansound8k dataset. Data augmentation increased classification success according to the results obtained [37]. Ye et al. (2017) extracted local and global features from spectrogram images and classified these features with SVM. The classification success rate obtained with the proposed model is higher than in the literature [19]. The classification was made with mel-filterbank, phase-encoded filterbank, and CNN in the ESC-10 dataset, and the combination of both features increased the classification success by 10% [38]. In the study where Teager Energy Operator (TEO) and gamma tone filter bank are used, the results are compared with MFCC, GMM, and the CNN architecture used in the literature [36] are used [39]. The results show that TEO-based methods have a higher classification success rate than MFCC. In studies using CNN architecture, models proposed by researchers or pre-trained CNN architectures are used. AlexNet, GoogLeNet, DenseNet161, and VGG deep learning models have been used in various studies with the UrbanSound8K dataset [22,27,40–42]. In these studies, spectrogram, mel-spectrogram, log mel-spectrogram, Gammatone Frequency Cepstral Coefficients (GFCC), log2 mel-spectrogram, and log3 mel spectrogram were used as input to the model. The study's results using three different datasets, spectrogram images, AlexNet, and GoogLeNet, showed that GoogleNet achieved higher classification success [40]. When no preprocessing was used on the data, the DenseNet-16 model performed poorly [22,41]. However, in the study using data augmentation, mel-spectrogram, log mel-spectrogram, log2 mel-spectrogram, and log3 mel spectrogram were used, and high classification success was achieved in all three datasets [27]. Many researchers have performed classification over mel-spectrogram images with their proposed CNN models. The most important difference in the studies is the number of layers in the CNN architecture. The CNN architectures used usually have 3-layers [43,44], 4-layers [45,46], or 5-layers [43]. In the study comparing the performance of CNN and LSTM models, the mel-spectrogram was used as the model's input. It was determined that the LSTM model showed higher classification success than the CNN model [47]. The mel-spectrogram and the mel-spectrogram of the inverse of the audio signal were used as the input of the CNN model, and over 90% of classification success was achieved [48]. According to the results of the study investigating the effect of noise on mel-spectrogram images with 4-layer CNN architecture, adding noise increased the classification success of the model [46]. According to the study comparing two different CNN models, using the pooling layer negatively affects the classification success of the model [49]. Another method used to input the CNN architecture is combining multiple image types or classifying them with two different CNN architectures and combining the output. The study used two parallel 5-layer CNN architectures, mel-spectrogram, chroma, spectral contrast, and tonnetz were input for one CNN architecture. The other CNN architecture used log mel-spectrogram, chroma, spectral contrast, and tonnetz [50]. Another

visual data used by the researchers is waveforms of urban sounds. In the study using SoundNet, an 8- and 5-layer CNN architecture, classification was performed based on waveforms [51]. The study used the waveform as the model input, where CNN and RNN were used together [52]. Two-, five-, eight- and twelve-layer architectures were compared, and the highest classification success was obtained with the 8-layer architecture [52]. The aim of the study, in which an 8-layer CNN architecture called ACDNet using waveforms is proposed, is to achieve high classification success with low hardware resources [53]. In another study, a hybrid model combining mel-spectrogram and waveform was proposed [54]. The proposed model uses 5-layer architecture for mel-spectrogram and 7-layer architecture for waveform and combines the classification results of these architectures with evidence theory.

If the research on the classification of urban sounds is generally evaluated, most studies used the researchers' proposed CNN architecture. Mostly spectrogram, mel-spectrogram, and log mel-spectrogram were used as the input of CNN architectures. ESC-10 [33] and UrbanSound8k [32] datasets were used. However, data augmentation methods were used since ESC datasets' class-based samples are low. After the literature review, it has been observed that the number of studies using acoustic parameters is limited, updated versions of traditional methods are not used, and there needs to be a comparison between feature sets. In addition, pre-trained networks, used in the literature for many image-based classifications, are used in a minimal number of studies.

3. Materials and methods

3.1. Dataset description

This study used the ESC-10 and UrbanSound8k datasets, widely used in the literature. The ESC-10 dataset contains 400 recordings with ten different classes and 40 samples in each class. The duration of each recording is 5 s, and the sampling rate is 44.1 kHz. The UrbanSound8K dataset is one of the most extensive urban sound datasets in the literature. This dataset consists of 8732 labeled and ten classes of urban sound data with a duration of up to 4 s, totaling 9.7 h [52]. All audio recordings are in "wav" format and have a sampling rate of 44.1 kHz. Most of the classes contain 1000 samples.

Class-based distributions are balanced in the ESC-10 dataset. In UrbanSound8K, on the other hand, the number of samples in only two classes is low compared to other classes, and the dataset has a balanced distribution of 80%. Therefore, data augmentation was not used.

3.2. Feature sets

In this study, acoustic analysis and cepstral feature extraction were used. F1, F2, F3, MFCC, and LPCC features were extracted in this context. SPEC, SCL, MEL, GFCC, and QCT images of each audio recording were used for CNN models. F1, F2, and F3 values are called formant frequencies. A formant is a resonance in the vocal tract. It provides information about sound production and the vocal tract's quantitative properties [55,56]. There are infinite formats in theory, but in practice, only the first three or four formants contain essential information. Linear Predictive Coding (LPC) is the most widely used method for determining formant frequencies and bandwidths.

LPC is the expression of the n_{th} sample of the speech signal as a linear combination of the previous p samples, as given in Eq. (1) [56].

$$X_{LPC}[n] = \sum_{i=1}^{p} b_i x[n-i] \tag{1}$$

b_i 's are considered constant over the frame duration. The difference between the actual and predicted value of speech is the prediction error (Eq. (2)).

$$e[n] = x[n] - X_{LPC}[n] \tag{2}$$

With the error signal, the LPC equation will be as in Eq.3.

$$X_{LPC}[n] = \sum_{i=1}^{p} b_i x[n-i] + e[n] \tag{3}$$

If the z-transform transfer function is calculated from Equation (4), a finite length impulse response filter is obtained (Eq. (4)).

$$H(z) = \frac{E(z)}{S(z)} = 1 - \sum_{i=1}^{p} x^i z^i = A(z) \tag{4}$$

A(z) in Eq. (4) is the filter used for vocal tract modeling.

Polynomial roots are obtained to calculate formant frequencies after LPC analysis. When these roots are ordered in the polar coordinate system, the first three values are expressed as F1, F2, and F3. Cepstral coefficients are obtained when the Fourier transform is applied to the LPC coefficients. These coefficients are called LPCC [58].

To obtain the cepstral features of an audio signal, the audio signal is represented by band filters and their energies. MFCCs are obtained when the band filters are defined as a triangular array of filters in mel scale. The mel used here is a unit for modeling the human hearing system [56,57].

Mel's frequency scale is given in Equation (1) [56].

$$f_{MEL} = 2595\log\left(1 + \frac{F_{Hz}}{700}\right) \tag{5}$$

This study used a frame size of 25 ms and an overlap of 50% for acoustic and cepstral features. Formant frequencies, MFCC, and LPCC features were extracted from each frame. For MFCC and LPCC, coefficient 13 was used. Since a sound signal contains more than one frame, these features were extracted from each frame. By statistically calculating these values obtained from the frames, 99 features were obtained. Their distribution is F1 (7 features), F2 (7 features), F3 (7 features), MFCC (39 features), and LPCC (39 features).

A spectrogram contains a visual representation of the frequency spectrum of a time-varying signal. Fourier transform is applied to the signal to obtain the spectrogram images. Mel spectrogram is a spectrogram in which frequencies are converted to mel scale. Considering greater sensitivity to noise, a gamma tone filter bank is used instead of a mel-scale filter bank, and GFCC is obtained [22]. When spectrogram images are obtained, SCL images are obtained when wavelet transform is used instead of Fourier transform. Another method, CQT, converts by creating a logarithmic gap between the STFT and frequency transitions [59]. Fig. 1 shows the sample spectral images obtained from the dataset.

3.3. Classifiers

This study used SVM and k-NN classifiers, transfer learning, and CNN models to classify urban sounds. The basic operation of the SVM classifier is based on finding the hyper-plane that separates the classes [60]. The kernel function is the most significant factor in the SVM classifier's performance. The data is moved to a multidimensional space with the kernel function, and classification is made. Commonly used kernel functions are linear, polynomial, radial basis, and sigmoid functions. Another kernel function used recently is the cubic function [61,62]. In this study, cubic kernel functions are used for the SVM classifier.

K-NN is an example-based learning algorithm based on the distances of the observations. First, the k nearest neighbors of the training set are calculated, then the similarities to the k nearest neighbors of a sample from the test data are clustered according to the classes of the neighbors. An advantage of k-NN is that it is suitable for multiple classes since the classification decision is based on small neighborhoods of similar objects. In general, the choice of the best k parameter is data dependent, and large values of k reduce the noise on the classification but minimize the clarity of the boundaries between the classes. A good k value can be selected using different heuristic techniques, such as cross-validation [63]. When determining the class of a new instance, the distance to

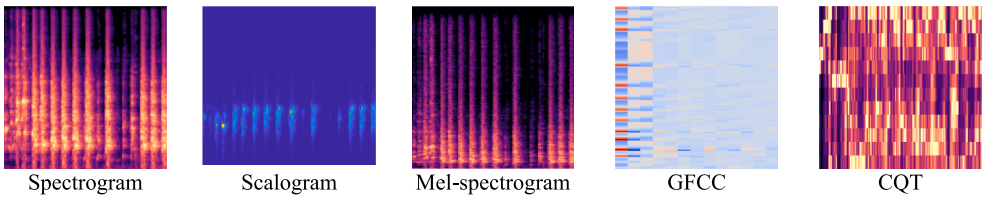

Fig. 1. Sample spectral images from the dataset.

some other cases is calculated, and the new instance is given the class information of the closest instance. Distances from multiple instances can also be used in this process [64–66].

DNN is a multi-layered and neuron form of artificial neural network in which the feature extraction process is carried out in the model instead of handcrafted features. It is a type of CNN multilayer perceptron mostly used in deep learning architecture. Thanks to its layered architecture, CNN provides the attributes that represent the image in layers starting from the general features. Current CNN models in the literature; include LeNet-5 [67], AlexNet [68], VGG-16 [69], GoogLeNet [70], and ResNet [71].

Transfer learning uses machine learning methods to use the knowledge learned to solve a problem in other problems. Thus, using previous knowledge, successful and fast learning models are obtained with less training data. There are many trained CNN models in the literature. By changing the input data, input parameters, and output parameters of these models according to the new dataset, the learning of the existing model is transferred to the new problem. In this study, transfer learning was carried out with ShuffleNet [72] and ResNet-18 [71] architectures. In addition, a new CNN model was created within the scope of this study, and audio-visual features were used together. Also, Adam was chosen for optimization, 50 for epoch, and 0.001 for learning rate. Since the sample size of the ESC-10 dataset is small, the mini-batch value of 32 was used. For the UrbanSound8k dataset, this value is 256. The model (AVCNN) used to perform classification over audio-visual features is given in Fig. 2. The figure given is representative, and not all layers are shown. Details of all layers are given in Table 1.

AVCNN uses spectral images and handcrafted features together. Visual CNN passes the image through a 5-layer structure to obtain features. These features are combined with handcrafted features, and the classification is performed by passing the image through the dense layer again. Drop-out layers are included in the model to prevent the model from overfitting. Pooling layers minimize the image by reducing the number of parameters.

Table 1
Layers of the AVCNN model.

Layer	Size	Stride	Filters	Output
Input	300x300x3	–	–	300x300x3
Conv2D	3x3	2x2	64	150x150x64
Batch Normalization	–	–	–	–
ReLU	–	–	–	–
Max Pooling	5x5	2x2	–	75x75x64
Conv2D	3x3	2x2	128	38x38x128
ReLU	–	–	–	–
Conv2D	3x3	2x2	128	19x19x128
ReLU	–	–	–	–
Max Pooling	5x5	2x2	–	10x10x128
Dropout (0.5)	–	–	–	10x10x128
Conv2D	3x3	2x2	256	5x5x256
ReLU	–	–	–	–
Conv2D	3x3	2x2	256	3x3x256
Max Pooling	5x5	2x2	–	2x2x256
Drop Out (0.5)	–	–	–	2x2x256
Flatten	–	–	–	1024x1
Feature Input	99x1	–	Ones(1024x1)	1024x1
Concatenate	–	–	–	2048x1
Dense	–	–	–	10x1
Softmax	–	–	–	–
Output	–	–	–	10x1

4. Experimental results

In this study, the inputs to the classifiers are feature vectors obtained from feature extraction steps and CNN models. The visual dataset includes SPEC, SCL, MEL, GFCC, and CQT images. Three experiments were

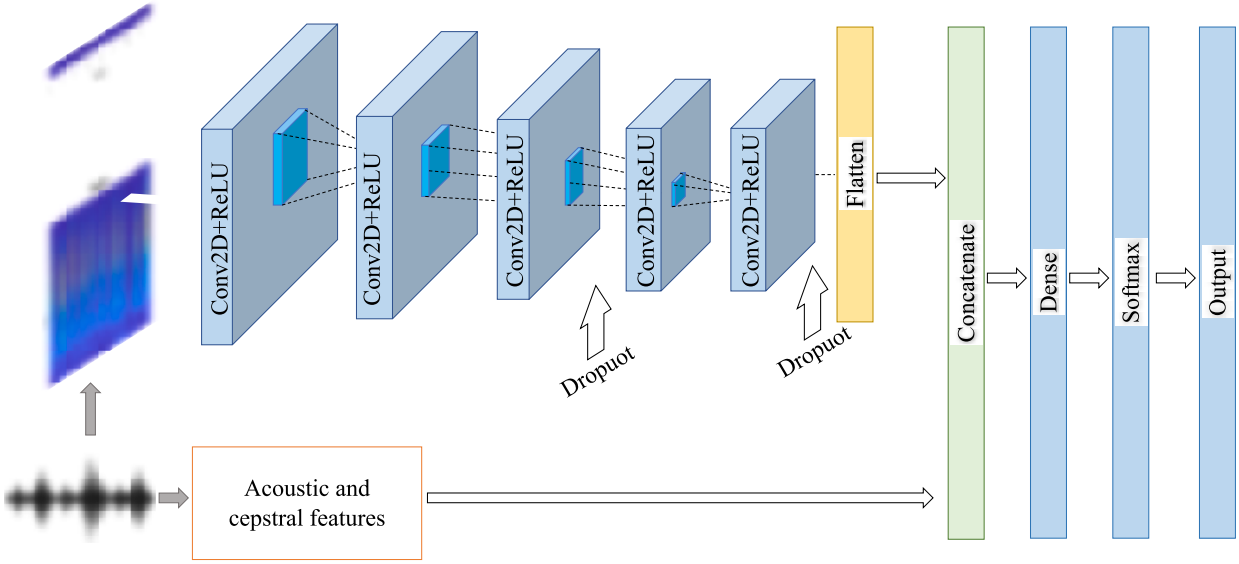

Fig. 2. Proposed CNN model for classification with audiovisual features (AVCNN).

conducted to classify urban sounds. In the first experiment (Exp1), traditional classifiers and acoustic/cepstral features were used. The second experiment (Exp2), classification was performed with CNN models using transfer learning. It also includes the results of the AVCNN model using only the visual module. In the last experiment (Exp3), classification was performed with AVCNN. Also, the features obtained from AVCNN were classified with SVM and k-NN. Specificity, accuracy, recall, and f-score were used to evaluate its performance. In all experiments, the dataset was randomly divided into 70% training, 15% validation, and 15% testing. Classification success was evaluated using 10-fold cross-validation.

4.1. Results of Exp1

In Exp1, 99 features of urban sounds were used. These features (F1, F2, F3, MFCC, and LPCC) were obtained with Praat [73]. SVM and k-NN classification results were compared using these feature sets. Table 2 shows the performance metrics obtained for Exp1, and Table 3 shows the class-based accuracy rates.

When the results for Exp1 are analyzed, it is seen that the test's success varies according to the dataset. SVM for ESC-10 and k-NN for UrbanSound8k provided higher accuracy. However, when the class-based accuracies are analyzed, the class-based accuracy distribution is more balanced in the SVM classifier.

4.2. Results of Exp2

Exp2 involves the application of the ShuffleNet and Resnet-18 CNN models in the literature to urban datasets with transfer learning. In this experiment, five different images were used. The training results of the image datasets for each CNN model according to the epochs are given in Fig. 3 and Fig. 4.

When the results in Fig. 3 are analyzed, higher accuracies are obtained with the ShuffleNet model compared to ResNet-18. Regarding the dataset, the success of training the SCL images is higher in both models. The closest training success to the SCL images was obtained with the MEL images. Training results for UrbanSound8k are given in Fig. 4.

When the results for UrbanSound8k are analyzed, the results obtained for ShuffleNet are like the ESC-10 results, and the highest training success is obtained with the SCL images. However, for ResNet-18 in UrbanSound8k dataset, the highest training success was obtained with SPEC, followed by MEL.

Another value classification process performed within Exp2 is the classification with only the visual part of the AVCNN model. The feature input part of the AVCNN model was removed, dense, softmax, and output layers were added after the last convolutional layer, and the model was trained with the visual dataset. The results of this training are given in Fig. 5.

The visual part of the AVCNN model has similar results to the other two transfer learning processes. The highest success of training was obtained in the SCL and MEL images. To test the test success of the models used in Exp2, 15% of the datasets were randomly selected and classified. The results obtained are given in Table 4.

Table 4 contains the classification accuracy of randomly selected samples from the datasets. According to the results in the table, SCL, MEL, and SPEC images have achieved high success in most of the tests, as in the training process. Although the test success rates of GFCC and CQT images are lower than others, they are not at a level that cannot be used. The class-based truth table for the highest test success rates listed in Table 4 is given in Table 5.

Table 3
Class-based accuracy rates for Exp1 in the test dataset.

Classes	ESC-10		UrbanSound8k	
	SVM	k-NN	SVM	k-NN
Class 1	85.00	77.50	97.50	98.60
Class 2	65.00	42.50	81.12	79.72
Class 3	82.50	85.00	91.10	95.70
Class 4	90.00	90.00	87.60	90.20
Class 5	75.00	45.00	91.80	93.70
Class 6	90.00	82.50	96.80	98.00
Class 7	65.00	65.00	92.25	95.45
Class 8	65.00	52.50	96.80	97.20
Class 9	85.00	72.50	95.05	97.09
Class 10	77.50	85.00	87.60	92.60
Overall	**78.00**	69.75	92.40	**94.61**

4.3. Results of Exp3

Exp3 includes the performance results of the AVCNN model. In this experiment, two different approaches were used. The first is to obtain the performance results of the AVCNN model in its current form. The second is the classification process with SVM and k-NN by changing the classifier layer of the AVCNN model. The training results of the AVCNN model are given in Fig. 6. The test results of the AVCNN model with randomly selected samples from the datasets are shown in Table 6.

When the training results of the AVCNN model are examined, the training process of the SPEC, SCL, MEL, and GFCC images for the ESC-10 has been completed with 100% training success. The training success rate varies between 82% and 100%. For UrbanSound8k, training success is over 95% with SPEC, SCL, and MEL images. The training success rate for UrbanSound8k with CQT and GFCC images is similar. The graph starts at low values in Fig. 6.a because the success of training is common in the first epochs of the training process and increases in the next epochs. SPEC, SCL, and MEL graphs in Fig. 6.b show that the model has achieved high training success since the first epochs. The most important reason for this difference is the small number of samples in the ESC-10 dataset. To evaluate the test success of the AVCNN model, 15% of the data allocated for testing from the data set was used. The results obtained are listed in Table 6.

When the test results of the AVCNN model are analyzed, AVCNN increased the test's success by approximately 3%-10% for all classifications. Only for the AVCNN + MEL experiment was no change in the test's success. AVCNN results, like the other models, achieved the highest test success with SPEC, SCL, and MEL images. Unlike the other models, the AVCNN model significantly improved the test success of GFCC and CQT images. The last experiment performed in this study is the classification of the feature set obtained from the AVCNN model with SVM and k-NN classifier. Table 7 shows the results of the AVCNN model features with SVM and k-NN.

When the classification results of AVCNN feature sets with SVM and k-NN are analyzed, they are similar to the AVCNN model. While the test success rate decreased slightly for the ESC-10 dataset, the test success rate increased for the UrbanSound8k dataset. Compared to the results in Exp1, the test success rate for the ESC-10 dataset increased by about

Table 2
Performance metrics for Exp1 in the test dataset.

Model	ESC-10				UrbanSound8k			
	Recall	Accuracy	Specificity	F1score	Recall	Accuracy	Specificity	F1score
SVM	0.7800	0.7800	0.9756	0.7794	0.9176	0.9240	0.9915	0.9214
k-NN	0.6975	0.6975	0.9664	0.6926	0.9383	0.9461	0.9939	0.9421

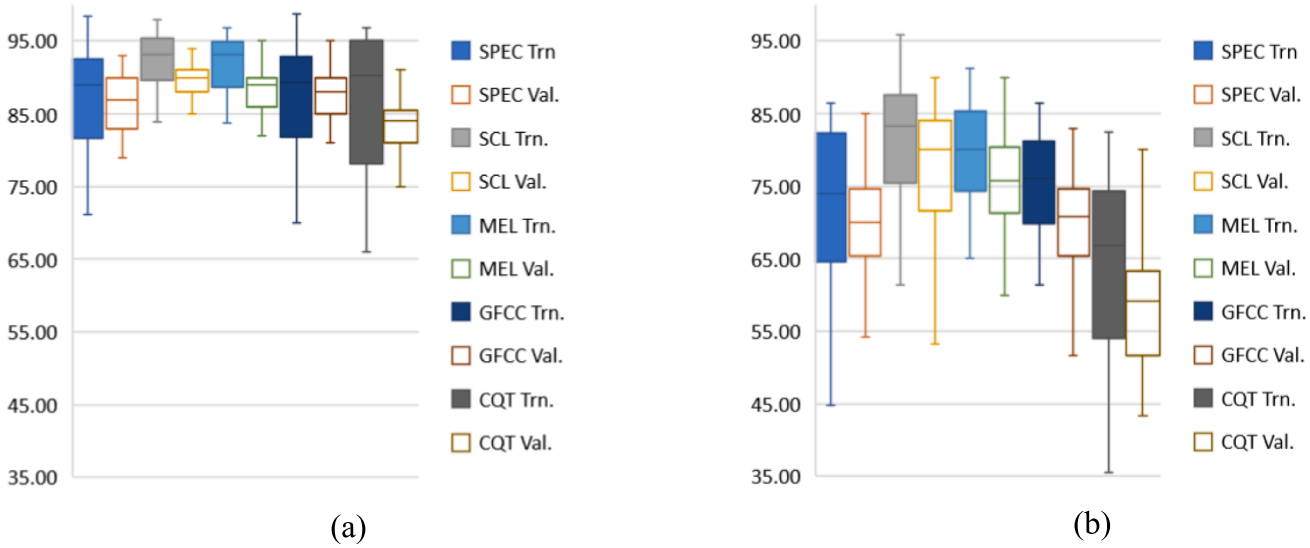

Fig. 3. Training and validation results for ESC-10 dataset (a) ShuffleNet (b) ResNet-18.

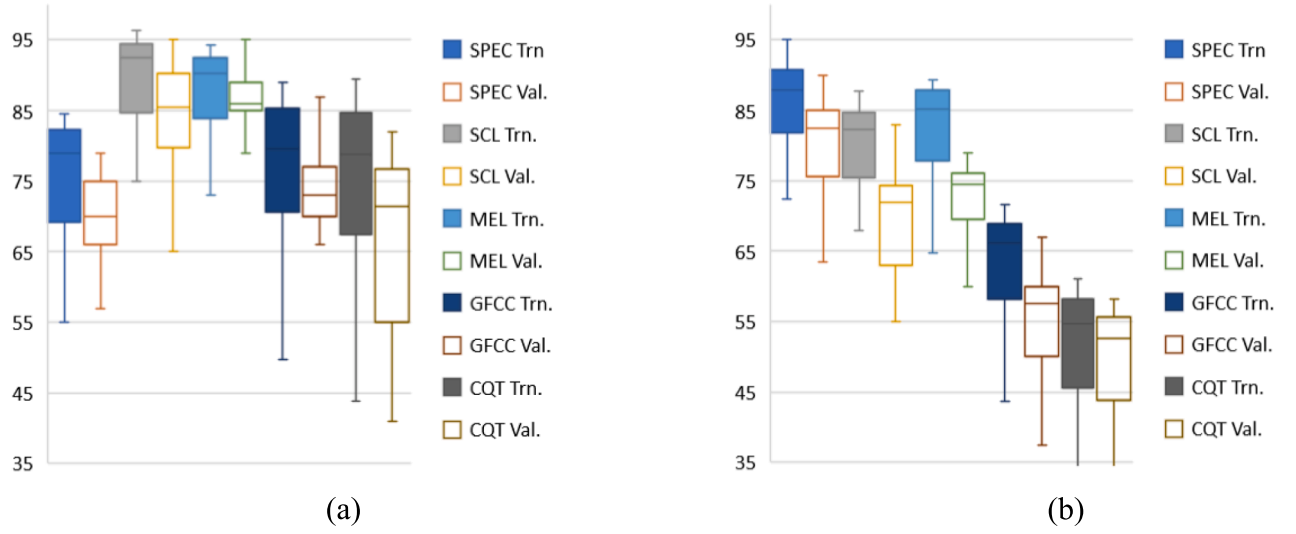

Fig. 4. Training and validation results for UrbanSound8k dataset (a) ShuffleNet (b) ResNet-18.

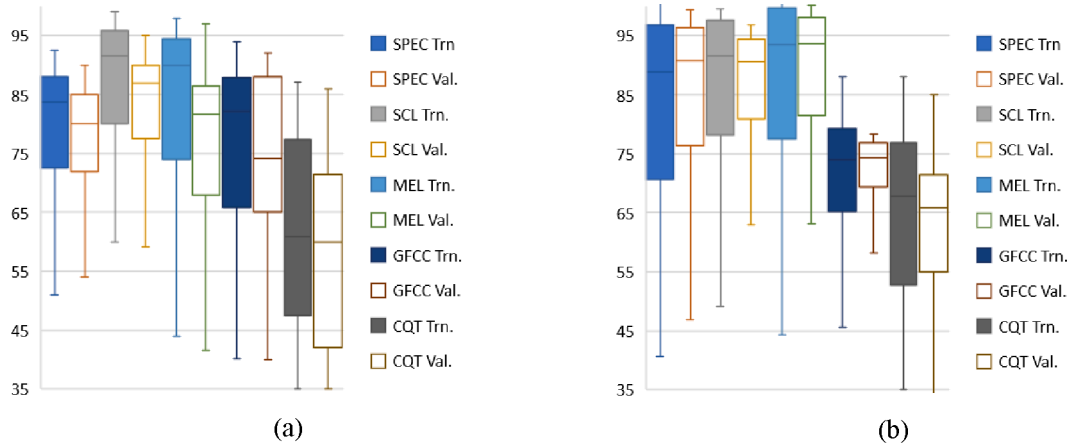

Fig. 5. Training and validation results with only the visual part of the AVCNN model (a) ESC-10 (b) UrbanSound8k.

Table 4
Test performance metrics obtained in Exp2.

Model		ESC-10				UrbanSound8k			
		Recall	Accuracy	Specificity	F1score	Recall	Accuracy	Specificity	F1score
ShuffleNet	SPEC	0.9388	0.9286	0.9921	0.9297	0.8213	0.7865	0.9764	0.7940
	SCL	0.9426	**0.9333**	0.9929	0.9272	0.9466	**0.9430**	0.9936	0.9445
	MEL	0.9027	0.8833	0.9873	0.8820	0.9205	0.9116	0.9901	0.9135
	GFCC	0.9213	0.9167	0.9908	0.9156	0.8203	0.8079	0.9785	0.8143
	CQT	0.8283	0.7917	0.9773	0.7931	0.8315	0.8064	0.9784	0.8158
ResNet-18	SPEC	0.7620	0.7583	0.9740	0.7368	0.9008	**0.8465**	0.9831	0.8657
	SCL	0.8934	**0.8917**	0.9881	0.8899	0.7866	0.7528	0.9725	0.7539
	MEL	0.8893	0.8750	0.9862	0.8757	0.7063	0.6502	0.9618	0.6498
	GFCC	0.8651	0.8417	0.9826	0.8430	0.6782	0.6372	0.9598	0.6369
	CQT	0.7279	0.7083	0.9679	0.7057	0.6164	0.5817	0.9532	0.5935
AVCNN (only visual)	SPEC	0.9269	0.9167	0.9908	0.9175	0.9615	**0.9560**	0.9951	0.9581
	SCL	0.9419	0.9333	0.9927	0.9326	0.9569	0.9541	0.9949	0.9545
	MEL	0.9723	**0.9667**	0.9963	0.9675	0.9477	0.9365	0.9929	0.9424
	GFCC	0.9413	0.9250	0.9919	0.9252	0.8037	0.7727	0.9745	0.7823
	CQT	0.9011	0.9000	0.9891	0.8959	0.8289	0.8067	0.9784	0.8120

Table 5
Class-based accuracy rates of the models with the highest test success in Exp3.

Classes	ESC-10 + AVCNN + MEL	UrbanSound8k + AVCNN + SPEC
Class 1	100.00	99.29
Class 2	100.00	99.15
Class 3	100.00	90.76
Class 4	100.00	96.13
Class 5	100.00	96.70
Class 6	80.00	96.33
Class 7	100.00	100.00
Class 8	100.00	96.41
Class 9	92.31	93.27
Class 10	100.00	93.40
Overall	**96.67**	**95.60**

20%. For the UrbanSound8k dataset, the test success rate increased by about 2%. In general, the SVM classifier was more accurate than the k-NN classifier.

5. Discussion

This study analyzes the accuracy rates of the features obtained by signal processing methods and spectral images to classify urban sounds. For this purpose, SVM and k-NN classifiers were used for numeric values. In addition, spectrogram, scalogram, mel-spectrogram, GFCC, and CQT images were used to evaluate the performance of spectral images. Transfer learning was performed with ShuffleNet and ResNet18 models to classify these images. In addition, a 5-layer model named AVCNN is proposed. The proposed model uses handcrafted features along with spectral images. Within the scope of the study, three

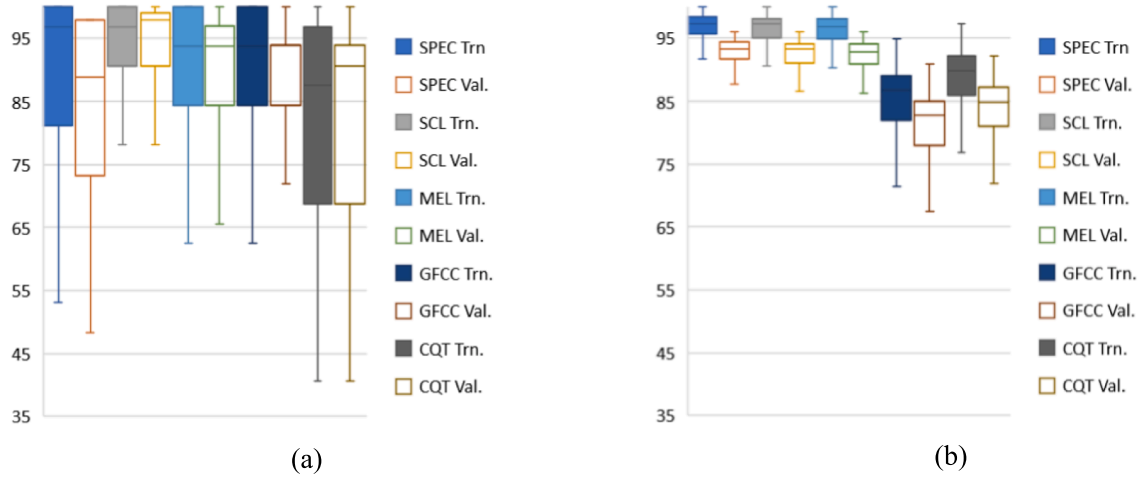

Fig. 6. Training and validation results of the AVCNN model (a) ESC-10 (b) UrbanSound8k.

Table 6
Test performance metrics with the AVCNN model.

Model		ESC-10				UrbanSound8k			
		Recall	Accuracy	Specificity	F1score	Recall	Accuracy	Specificity	F1score
AVCNN	SPEC	0.9250	0.9500	0.9944	0.9269	0.9736	**0.9725**	0.9969	0.9724
	SCL	0.9889	**0.9833**	0.9982	0.9798	0.9747	**0.9725**	0.9969	0.9747
	MEL	0.9750	0.9667	0.9965	0.9617	0.9636	0.9656	0.9961	0.9649
	GFCC	0.9217	0.9333	0.9925	0.9283	0.8891	0.8944	0.9881	0.8915
	CQT	0.9405	0.9333	0.9924	0.9407	0.8944	0.8929	0.9880	0.8961

Table 7
Test performance metrics of AVCNN feature sets with SVM and k-NN.

Model		ESC-10				UrbanSound8k			
		Recall	Accuracy	Specificity	F1score	Recall	Accuracy	Specificity	F1score
AVCNN + SVM	SPEC	0.9425	0.9425	0.9936	0.9425	0.9731	0.9746	0.9971	0.9750
	SCL	0.9600	0.9600	0.9956	0.9600	0.9717	0.9715	0.9968	0.9722
	MEL	0.9724	**0.9724**	0.9969	0.9722	0.9768	**0.9770**	0.9974	0.9771
	GFCC	0.9300	0.9300	0.9922	0.9296	0.9000	0.8977	0.9885	0.9022
	CQT	0.8800	0.8800	0.9867	0.8803	0.9130	0.9120	0.9901	0.9142
AVCNN + k-NN	SPEC	0.9225	0.9225	0.9914	0.9223	0.9522	0.9511	0.9945	0.9539
	SCL	0.9397	0.9398	0.9933	0.9396	0.9351	0.9367	0.9929	0.9368
	MEL	0.9400	**0.9400**	0.9933	0.9399	0.9678	**0.9680**	0.9964	0.9690
	GFCC	0.9000	0.9000	0.9889	0.8985	0.8069	0.8031	0.9779	0.8065
	CQT	0.7850	0.7850	0.9761	0.7824	0.8009	0.7963	0.9771	0.8006

experiments were conducted using ESC-10 and UrbanSound8k datasets.

In Exp1, 99 acoustic and cepstral features were classified with SVM and k-NN. The classification resulted in an accuracy of 78.00% for the ESC-10 dataset and 94.61% for the UrbanSound8k dataset. Exp2 obtained five spectral images for each dataset and compared their classification performance with CNN models. In all the analyses performed in this experiment, scalogram and mel-spectrogram images are more successful than others. The classification success rate of spectrogram images is close to these two images. Although the classification success rates obtained in GFCC and CQT datasets are lower than the other images, the classification success rates are above 80%. Another analysis performed in Exp2 is the analysis of the AVCNN model with only image datasets. The AVCNN model showed higher classification success than the other two transfer learning processes. The highest accuracy for ESC-10 in Exp2 is 96.67% with AVCNN (visual only) and mel-spectrogram images. For UrbanSound8k, the highest accuracy is 95.60% with AVCNN (only visual) and spectrogram images. In the last experiment, the classification success of the AVCNN model using both spectral images and handcrafted features and the classification success of AVCNN features with traditional classifiers were analyzed. According to the analysis results, the AVCNN model increased the classification success rate compared to the other models in all image datasets. The highest classification success rate was obtained for scalogram and spectrogram images as a feature set. The highest accuracy for ESC-10 is 98.33% with scalogram images. For URbanSound8k, the highest accuracy is 97.25% with both spectrogram and scalogram. The SVM and k-NN classification results of the features obtained from the AVCNN model are generally like the AVCNN. However, the classification success rate increased by about 20% compared to Exp1, especially on the ESC-10 dataset. This analysis again obtained the highest classification success with the mel-spectrogram and scalogram. The comparison of the results obtained in all experiments with the literature is given in Table 8.

When the results obtained in this study are compared with those in the literature, both datasets have higher accuracy than those in the literature. Only the study results using DenseNet [27] are higher than ours. However, data augmentation and many features were used in that study. A similar situation exists in the study using GAN [35].

Table 8
Comparison of the results obtained in the study with the literature.

Classifier	Input	Data Aug.	Accuracy (%) on Datasets	
			ESC-10	UrbanSound8k
SoundNet (CNN) [51]	Waveform	–	92.20	–
CNN [39]	TEO-GTSC	–	–	88.02
GoogleNet [40]	SPEC, MFCC, CRP	–	86.00	93.00
TSCNN-DS [54]	MEL + Waveform	✓	–	97.20
VGG [42]	MEL + GFCC	–	91.70	83.70
CNN [50]	MLMC	✓	–	97.20
CNN [74]	SPEC	✓	94.00	–
LSTM [59]	MEL	–	–	84.25
ESResNet [75]	Log SPEC	–	97.00	85.42
DenseNet-161 [27]	Hybrid1	✓	**99.22**	97.98
CNN [76]	MFCC + GFCC + CQT + Chromagram	✓	94.75	97.52
CNN [48]	LM + MFCC	✓	–	94.30
DCNN [49]	MEL + MFCC + LM	✓	94.94	95.37
LSTM [59]	MFCC + Chroma STFT	✓	–	98.81
CNN [77]	Spectrogram	–	–	86.70
SVM [78]	Spectrogram	✓	94.80	78.14
RF [28]	Hybrid2	–	–	96.10
CNN [79]	SPEC	✓	92.00	–
CNN + GAN [35]	–	✓	96.50	**98.97**
ACDNet [53]	Waveform	✓	96.65	84.45
Results of this study SVM and k-NN	Acoustic + Cepstral	–	78.00	94.61
AVCNN	MEL	–	96.67	–
AVCNN	SPEC	–	–	95.60
AVCNN	SCL + Acoustic	–	**98.33**	97.25
AVCNN + SVM	MEL + Acoustic	–	97.24	**97.70**

LM: Log-Mel Scale Spectrogram, TEO-GTSC: TEO-based Gammatone spectral coefficients, CRP: Cross Recurrence Plot, MLMC: MFCC + LM + Chroma + Spectral Contrast + Tonnetz, Hybrid1: LM + MEL + Log2Mel + Log3Mel + ZCR + Chroma + Tonnetz + Spectral contrast, Hybrid2: Deep features (Mel) + MFCC + DeltaMFCC + Delta2MFCC + ZCR + Spectral rolloff + Spectral centroid + RMS + Chroma STFT + Chroma CQT + Chroma CENS + Entropy + Spectral Flatness + Spectral bandwidth + Spectral contrast + Poly + Tempogram + Tonnez.

6. Conclusion

Urban sounds have both positive and negative effects on people. If these sounds are noise pollution, they can negatively impact people physically and mentally. On the other hand, urban sounds trigger a previous memory and create positive psychological effects. Moreover, with the development of technology, sound classification is an important field of study for descriptive tasks such as traffic density and city safety in smart cities. This study uses traditional classifiers and CNN models to examine the effect of signal processing and spectral-based methods on the classification of urban sounds. In this context, it was found that acoustic parameters used in speech recognition can also be used to classify urban sounds. Among the spectrogram, scalogram, mel-spectrogram, Gammatone Frequency Cepstral Coefficients, and Constant Q-transform images used to visually represent the audio signal, especially mel-spectrogram, scalogram, and spectrogram images were found to produce more successful results. Using CNN models with these images produced even more positive results on classification success. Although the images and acoustic features separately classified urban sounds at a high level, their hybrid use positively affected the classification success results.

The strengths of this study are the classification success rates obtained, the comparison of the visual datasets used, and the creation of

the AVCNN model for hybrid classification. The limitations of this study are the small number of samples in the used datasets and the unbalanced distribution between the classes. In future studies, these limitations can be removed, and the classification success can be compared. Also, the classification performance can be tested on mixed sounds from the environment. Another area to focus on is noise-type modeling by using hybrid noise maps and CNN models.

Compliance with ethical standards

Informed consent

Informed consent was not required as no humans or animals were involved.

Human and animal rights

This article does not contain any studies with human or animal subjects performed by the author.

Data availability statement

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

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.

Data availability

Data will be made available on request.

References

[1] Themann CL, Masterson EA. Occupational noise exposure: A review of its effects, epidemiology, and impact with recommendations for reducing its burden. J Acoust Soc Am Nov. 2019;146(5):3879. https://doi.org/10.1121/1.5134465.

[2] Pienkowski M. 'Loud music and leisure noise is a common cause of chronic hearing loss, tinnitus and hyperacusis'. Int J Environ Res Public Health Apr. 2021;18(8):4236. https://doi.org/10.3390/ijerph18084236.

[3] Alsouda Y, Pllana S, Kurti A. A machine learning driven IoT solution for noise classification in smart cities, arXiv preprint arXiv:1809.00238, 2018.

[4] Fredianelli L . et al., 'Traffic flow detection using camera images and machine learning methods in ITS for noise map and action plan optimization', Sensors, vol. 22, no. 5, Art. no. 5, Jan. 2022, doi: 10.3390/s22051929.

[5] Licitra G, Bolognese M, Chiari C, Carpita S, Fredianelli L. Noise source predominance map: a new representation for strategic noise maps. Noise Mapping Jan. 2022;9(1):269–79. https://doi.org/10.1515/noise-2022-0163.

[6] Muzet A. Environmental noise, sleep and health. Sleep Med Rev Apr. 2007;11(2):135–42. https://doi.org/10.1016/j.smrv.2006.09.001.

[7] Basner M, McGuire S. WHO environmental noise guidelines for the European region: a systematic review on environmental noise and effects on sleep. Int J Environ Res Public Health Mar. 2018;15(3):519. https://doi.org/10.3390/ijerph15030519.

[8] Minichilli F, Gorini F, Ascari E, Bianchi F, Coi A, Fredianelli L, et al. Annoyance judgment and measurements of environmental noise: a focus on Italian secondary schools. Int J Environ Res Public Health Feb. 2018;15(2):208.

[9] Petri D, Licitra G, Vigotti MA, Fredianelli L. Effects of exposure to road, railway, airport and recreational noise on blood pressure and hypertension. Int J Environ Res Public Health Aug. 2021;18(17):9145. https://doi.org/10.3390/ijerph18179145.

[10] Thompson R, Smith RB, Bou Karim Y, Shen C, Drummond K, Teng C, et al. Noise pollution and human cognition: An updated systematic review and meta-analysis of recent evidence. Environ Int 2022;158:106905.

[11] Schubert M, Hegewald J, Freiberg A, Starke K, Augustin F, Riedel-Heller S, et al. Behavioral and emotional disorders and transportation noise among children and adolescents: a systematic review and meta-analysis. Int J Environ Res Public Health Sep. 2019;16(18):3336.

[12] Dzhambov AM, Lercher P. Road traffic noise exposure and depression/anxiety: an updated systematic review and meta-analysis. Int J Environ Res Public Health Oct. 2019;16(21):4134. https://doi.org/10.3390/ijerph16214134.

[13] Daiber A, Kröller-Schön S, Frenis K, Oelze M, Kalinovic S, Vujacic-Mirski K, et al. Environmental noise induces the release of stress hormones and inflammatory signaling molecules leading to oxidative stress and vascular dysfunction-Signatures of the internal exposome. Biofactors Jul. 2019. https://doi.org/10.1002/biof.1506.

[14] Dratva J, Phuleria HC, Foraster M, Gaspoz J-M, Keidel D, Künzli N, et al. Transportation noise and blood pressure in a population-based sample of adults. Environ Health Perspect 2012;120(1):50–5.

[15] Lee PJ, Park SH, Jeong JH, Choung T, Kim KY. Association between transportation noise and blood pressure in adults living in multi-storey residential buildings. Environ Int Nov. 2019;132:105101. https://doi.org/10.1016/j.envint.2019.105101.

[16] Vukić L, Mihanović V, Fredianelli L, Plazibat V. Seafarers' perception and attitudes towards noise emission on board ships. Int J Environ Res Public Health Jun. 2021;18(12):6671. https://doi.org/10.3390/ijerph18126671.

[17] Sun K, De Coensel B, Filipan K, Aletta F, Van Renterghem T, De Pessemier T, et al. Classification of soundscapes of urban public open spaces. Landsc Urban Plan 2019;189:139–55.

[18] Yıldırım M. Automatic classification of environmental sounds with the MFCC method and the proposed deep model. Fırat University Journal of Engineering Science 2022;34(1):449–57.

[19] Ye J, Kobayashi T, Murakawa M. Urban sound event classification based on local and global features aggregation. Appl Acoust 2017;117:246–56.

[20] Ascari E, Cerchiai M, Fredianelli L, Licitra G. 'Statistical pass-by for unattended road traffic noise measurement in an urban environment', Sensors, vol. 22, no. 22, Art. no. 22, Jan. 2022, doi: 10.3390/s22228767.

[21] Fan X, Sun T, Chen W, Fan Q. Deep neural network based environment sound classification and its implementation on hearing aid app. Measurement 2020;159:107790.

[22] Huang Z, Liu C, Fei H, Li W, Yu J, Cao Y. Urban sound classification based on 2-order dense convolutional network using dual features. Appl Acoust 2020;164:107243.

[23] Asdrubali F, D'Alessandro F. Innovative approaches for noise management in smart cities: a review. Curr Pollution Rep Jun. 2018;4(2):143–53. https://doi.org/10.1007/s40726-018-0090-z.

[24] López JM, Alonso J, Asensio C, Pavón I, Gascó L, de Arcas G. 'A digital signal processor based acoustic sensor for outdoor noise monitoring in smart cities', Sensors, vol. 20, no. 3, Art. no. 3, Jan. 2020, doi: 10.3390/s20030605.

[25] Alías F, Alsina-Pagès RM. Review of wireless acoustic sensor networks for environmental noise monitoring in smart cities. J Sensors 2019;2019:1–13.

[26] Liu Ye, Ma X, Shu L, Yang Q, Zhang Yu, Huo Z, et al. Internet of things for noise mapping in smart cities: state of the art and future directions. IEEE Netw 2020;34(4):112–8.

[27] Mushtaq Z, Su S-F. Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images. Symmetry 2020;12(11):1822.

[28] Luz JS, Oliveira MC, Araujo FH, Magalhães DM. Ensemble of handcrafted and deep features for urban sound classification. Appl Acoust 2021;175:107819.

[29] Stoeckle S, Pah N, Kumar DK, McLachlan N. 'Environmental sound sources classification using neural networks'. In: The Seventh Australian and New Zealand Intelligent Information Systems Conference, 2001, IEEE, 2001, pp. 399–403.

[30] Torija AJ, Ruiz DP, Ramos-Ridao ÁF. A tool for urban soundscape evaluation applying support vector machines for developing a soundscape classification model. Sci Total Environ 2014;482:440–51.

[31] Jeon JY, Hong JY. Classification of urban park soundscapes through perceptions of the acoustical environments. Landsc Urban Plan 2015;141:100–11.

[32] Salamon J, Jacoby C, Bello JP. 'A dataset and taxonomy for urban sound research'. In: Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 1041–1044.

[33] Piczak KJ. 'ESC: Dataset for environmental sound classification'. In: Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018.

[34] da Silva B, Happi AW, Braeken A, Touhafi A. Evaluation of classical machine learning techniques towards urban sound recognition on embedded systems. Appl Sci 2019;9(18):3885.

[35] Madhu A, K. S. EnvGAN: a GAN-based augmentation to improve environmental sound classification. Artif Intell Rev 2022;55(8):6301–20.

[36] Piczak KJ. 'Environmental sound classification with convolutional neural networks'. In: 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP), IEEE; 2015. p. 1–6.

[37] Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process Lett 2017;24(3):279–83.

[38] Tak RN, Agrawal DM, Patil HA. Novel phase encoded mel filterbank energies for environmental sound classification. In: International Conference on Pattern Recognition and Machine Intelligence. Springer; 2017. p. 317–25.

[39] Agrawal DM, Sailor HB, Soni MH, Patil HA. 'Novel TEO-based Gammatone features for environmental sound classification'. In: 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, 2017. p. 1809–13.

[40] Boddapati V, Petef A, Rasmusson J, Lundberg L. Classifying environmental sounds using image recognition networks. Procedia Comput Sci 2017;112:2048–56.

[41] McMahan B, Rao D. 'Listening to the world improves speech command recognition'. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[42] 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). Springer; 2018. p. 356–67.

[43] Shu H, Song Y, Zhou H. Time-frequency performance study on urban sound classification with convolutional neural network. TENCON 2018–2018 IEEE Region 10 Conference, IEEE 2018:1713–7.

[44] Medhat F, Chesmore D, Robinson J. Masked Conditional Neural Networks for sound classification. Appl Soft Comput May 2020;90:106073. https://doi.org/10.1016/j.asoc.2020.106073.

[45] Massoudi M, Verma S, Jain R. Urban sound classification using CNN. In: in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. IEEE; 2021. p. 583–9.

[46] Zhao W, Yin B. Environmental sound classification based on adding noise. In: in *2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*. IEEE; 2021. p. 887–92.
[47] Lezhenin I, Bogach N, Pyshkin E. 'Urban sound classification using long short-term memory neural network'. In: 2019 federated conference on computer science and information systems (FedCSIS), Sep. 2019, pp. 57–60. doi: 10.15439/2019F185.
[48] Peng N, Chen A, Zhou G, Chen W, Zhang W, Liu J, et al. Environment sound classification based on visual multi-feature fusion and GRU-AWS. IEEE Access 2020;8:191100–14.
[49] Mushtaq Z, Su S-F. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Appl Acoust Oct. 2020;167:107389. https://doi.org/10.1016/j.apacoust.2020.107389.
[50] Su Yu, Zhang Ke, Wang J, Madani K. Environment sound classification using a two-stream CNN based on decision-level fusion. Sensors Jan. 2019;19(7):1733.
[51] Aytar Y, Vondrick C, Torralba A. 'SoundNet: learning sound representations from unlabeled video'. In: Advances in neural information processing systems, Curran Associates, Inc., 2016. Accessed: Dec. 09, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/7dcd340d84f762eba80aa538b0c527f7-Abstract.html.
[52] Sang J, Park S, Lee J. 'Convolutional recurrent neural networks for urban sound classification using raw waveforms'. In: 2018 26th European Signal Processing Conference (EUSIPCO), Sep. 2018, pp. 2444–2448. doi: 10.23919/EUSIPCO.2018.8553247.
[53] Mohaimenuzzaman M, Bergmeir C, West I, Meyer B. Environmental sound classification on the edge: a pipeline for deep acoustic networks on extremely resource-constrained devices. Pattern Recogn Jan. 2023;133:109025. https://doi.org/10.1016/j.patcog.2022.109025.
[54] 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 Jul. 2018;8(7):1152.
[55] Rabiner LR. 'Digital-formant synthesizer for speech-synthesis studies'. J Acoust Soc Am, vol. 43, no. 4, pp. 822–828, 1968.
[56] Ozseven T. Detection of acoustic parameters in sound analysis and investigation of the relationship between anxiety disorder and acoustic parameters. Karabük University; 2017. PhD Thesis,.
[57] Vergin R, O'Shaughnessy D, Gupta V. Compensated mel frequency cepstrum coefficients. In: in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. IEEE; 1996. p. 323–6.
[58] Makhoul J. Linear prediction: A tutorial review. Proc IEEE 1975;63(4):561–80.
[59] Das JK, Ghosh A, Pal AK, Dutta S, Chakrabarty A. 'Urban sound classification using convolutional neural network and long short term memory based on multiple features'. In: 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS), Oct. 2020, pp. 1–9. doi: 10.1109/ICDS50568.2020.9268723.
[60] Vapnik V. The nature of statistical learning theory. Springer Science & Business Media, 2000. Accessed: Apr. 13, 2015. [Online]. Available: http://www.google.com/books?hl=tr&lr=&id=sna9BaxVbj8C&oi=fnd&pg=PR7&dq=The+Nature+of+Statistical+Learning+Theory&ots=ooIfP-imff&sig=2l0THEvc8K3GQFrxMDT0QI6fn7k.
[61] Ozyurt F, Sert E, Avci D. Ensemble residual network features and cubic-SVM based tomato leaves disease classification system. TS Feb. 2022;39(1):71–7. https://doi.org/10.18280/ts.390107.
[62] Jain U, Nathani K, Ruban N, Joseph Raj AN, Zhuang Z, Mahesh VGV. 'Cubic SVM classifier based feature extraction and emotion detection from speech signals'. In: 2018 international conference on sensor networks and signal processing (SNSP), Oct. 2018, pp. 386–391. doi: 10.1109/SNSP.2018.00081.
[63] Yuan C, Yang H. Research on K-value selection method of K-means clustering algorithm. J — Multidisciplinary Scientific Journal 2019;2(2):226–35.
[64] Albornoz EM, Milone DH, Rufiner HL. Spoken emotion recognition using hierarchical classifiers. Comput Speech Lang Jul. 2011;25(3):556–70. https://doi.org/10.1016/j.csl.2010.10.001.
[65] Huang C, Liang R, Wang Q, Xi J, Zha C, Zhao L. Practical speech emotion recognition based on online learning: from acted data to elicited data. Math Probl Eng 2013;2013:1–9. https://doi.org/10.1155/2013/265819.
[66] Gharavian D, Sheikhan M, Ashoftedel F. Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model. Neural Comput & Applic May 2013;22(6):1181–91. https://doi.org/10.1007/s00521-012-0884-7.
[67] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–324.
[68] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems; 2012. p. 1097–105.
[69] Zeiler MD, Fergus R. 'Visualizing and understanding convolutional networks'. In: *E*uropean conference on computer vision, Springer, 2014. p. 818–33.
[70] M. Lin, Q. Chen, and S. Yan, 'Network in network', *arXiv preprint arXiv:1312.4400*, 2013.
[71] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.
[72] Zhang X, Zhou X, Lin M, Sun J. 'ShuffleNet: An extremely efficient convolutional neural network for mobile devices'. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856. Accessed: Dec. 09, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_ShuffleNet_An_Extremely_CVPR_2018_paper.html.
[73] Boersma P. Praat, a system for doing phonetics by computer. Glot Int 2001;5(9):341–5.
[74] Zhang Z, Xu S, Zhang S, Qiao T, Cao S. Learning Attentive representations for environmental sound classification. IEEE Access 2019;7:130327–39. https://doi.org/10.1109/ACCESS.2019.2939495.
[75] Guzhov A, Raue F, Hees J, Dengel A. 'ESResNet: environmental sound classification based on visual domain models'. In: 2020 25th International Conference on Pattern Recognition (ICPR), Jan. 2021, pp. 4933–4940. doi: 10.1109/ICPR48806.2021.9413035.
[76] Sharma J, Granmo O-C, Goodwin M. 'Environment sound classification using multiple feature channels and attention based deep convolutional neural network'. In: Interspeech 2020, ISCA, Oct. 2020, pp. 1186–1190. doi: 10.21437/Interspeech.2020-1303.
[77] Demir F, Abdullah DA, Sengur A. A new deep CNN model for environmental sound classification. IEEE Access 2020;8:66529–37. https://doi.org/10.1109/ACCESS.2020.2984903.
[78] Demir F, Turkoglu M, Aslan M, Sengur A. A new pyramidal concatenated CNN approach for environmental sound classification. Appl Acoust Dec. 2020;170:107520. https://doi.org/10.1016/j.apacoust.2020.107520.
[79] Tripathi AM, Mishra A. Environment sound classification using an attention-based residual neural network. Neurocomputing Oct. 2021;460:409–23. https://doi.org/10.1016/j.neucom.2021.06.031.