IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 17, 2020, accepted March 29, 2020, date of publication April 1, 2020, date of current version April 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984903

A New Deep CNN Model for Environmental Sound Classification

FATIH DEMIR[1], DABAN ABDULSALAM ABDULLAH[2], AND ABDULKADIR SENGUR[1]
[1]Electrical and Electronics Engineering Department, Technology Faculty, Firat University, 23100 Elazig, Turkey
[2]Research Center, Sulaimani Polytechnic University, Sulaimanyah 46001, Iraq

Corresponding author: Abdulkadir Sengur (ksengur@firat.edu.tr)

ABSTRACT Cognitive prediction in the complicated and active environments is of great importance role in artificial learning. Classification accuracy of sound events has a robust relation with the feature extraction. In this paper, deep features are used in the environmental sound classification (ESC) problem. The deep features are extracted by using the fully connected layers of a newly developed Convolutional Neural Networks (CNN) model, which is trained in the end-to-end fashion with the spectrogram images. The feature vector is constituted with concatenating of the fully connected layers of the proposed CNN model. For testing the performance of the proposed method, the feature set is conveyed as input to the random subspaces K Nearest Neighbor (KNN) ensembles classifier. The experimental studies, which are carried out on the DCASE-2017 ASC and the UrbanSound8K datasets, show that the proposed CNN model achieves classification accuracies 96.23% and 86.70%, respectively.

INDEX TERMS Environmental sound classification, spectrogram images, CNN model, deep features.

I. INTRODUCTION

Smart sound recognition (SSR) is a modern technique for detecting sound events that exist in the real life. The SSR is principally based on analyzing human hearing systems and embedding such perception capability in artificial intelligence applications [1]. Environmental sound classification (ESC) takes part as a basic and necessary step of SSR. The key target of ESC is to exactly detect the truth category of a perceived sound, such as doorbell, horn and jackhammer. With the practical applications of SSR in audio surveillance systems, smart device applications and healthcare [2], the ESC problem has taken very much interest in recent times. For automatic speech recognition (ASR) [3] and music information recognition (MIR) [7], it has been achieved great improvements with advances in machine learning. Because of greatly non-stationary characteristics of environmental sounds, these signals cannot be categorized as speech or music only. In other words, the models constituted for ASR and MIR will be poor when applying to ESC problems. Therefore, it is important to develop the efficient machine learning algorithm for ESC problems.

ESC is formed two main parts: audio based features and classifiers. For feature extraction, audio signals are first divided into frames with a window function such as Hamming or Hann window. Then, this set of features extracted from each frame is used in training or testing processing [8]. Features based on Mel filters (Mel Frequency Cepstral Coefficients (MFCC)) are commonly used features in ESC with acceptable efficiency, although they are actually developed for ASR [9], [10]. Also, a notable number of studies demonstrate that concatenated features performed better than only use one feature set in ESC missions. However, more concatenated conventional features cannot increase the classification performance. Therefore, an appropriate feature concatenation strategy is a vital part of sound classification. Artificial Neural Network (ANN), Support Vector Machines (SVM), Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) are greatly used classifiers in sound and other category. However, these conventional classifiers are designed to classify apparent changes which conclude in the absence of time and frequency invariability.

In recent years, deep learning (DL) models have been demonstrated to be more capable than conventional classifiers in resolving complicated classification problems. The convolutional neural network (CNN) is one of the most widely used models of DL, which could tackle the prior restrictions by learning parameters, which is including the time and frequency representations [10], [11]. The CNN is constituted to process data that get in the shape of multiple arrays: 1D signals, such as speech and biomedical signals, and 2D for image or audio spectrograms [12], [13]. The CNN

The associate editor coordinating the review of this manuscript and approving it for publication was Victor S. Sheng.

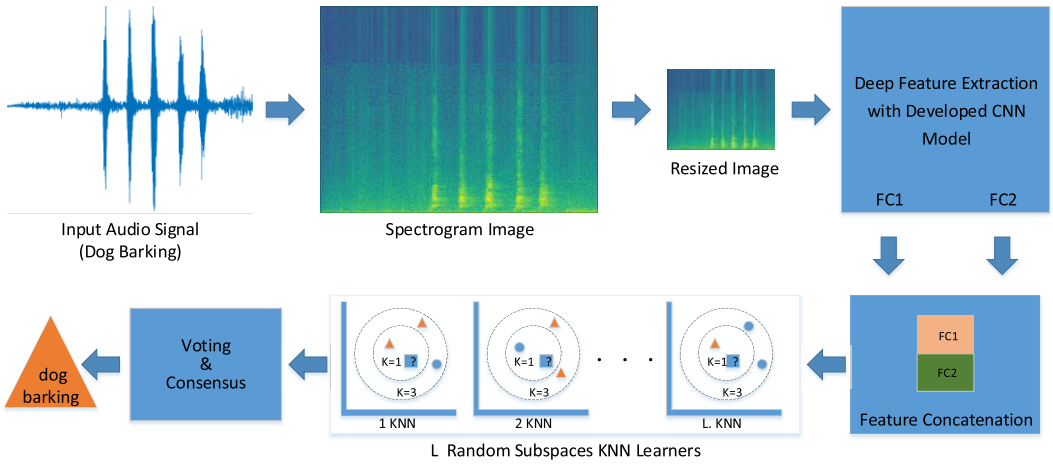

FIGURE 1. Illustration of the proposed method.

model constituted by Krizhevsky *et al.* [22] outperformed all the conventional methods in the ImageNet Large Scale

Visual Recognition Challenge (ILSVRC-2012). This CNN model known as AlexNet has pioneered the other popular CNN models, such as VGGNet and ResNet. The pre-trained CNN models, which are shared the learnable parameters, have shown good performances in almost all classification applications [4], [14], [15]. Moreover, hybrid approaches, which consist of the pre-trained CNN models and conventional classifiers, have been used to improve the classification performance. In [16], the deep features are extracted by using the pre-trained CNN model. The SVM and KNN algorithms are used for hyperspectral images classification. In [17], the pre-trained CNN models such as AlexNet and VGG16 are utilized to extract deep features from EMG signals. The best accuracy is achieved with SVM classifier. In [18], a new approach is proposed for the brain MRI classification. The feature set is constituted by combining the AlexNet and VGG16 models with hypercolumn technique. The evaluation is performed by the SVM classifier. In [19], the deep features are extracted by the last fully connected layer of the ResNet50 CNN model by using videocapsule endoscopy (VCE) images for diagnosing celiac disease. In the classification stage, the SVM, the KNN, the LDA and the softmax classifiers are evaluated on a dataset. The best accuracy is achieved by the SVM classifier. However, the popular pre-trained CNN models for feature extraction cannot fully represent the sound characteristics as they are only trained with images. In addition, the large input size and the very deep network structure, which are needed for recognition of high-resolution images may not be always required for ESC problems. In this state, it is obtained the low computational cost because of decreasing the learnable parameters.

In the paper, an approach, which consists of the deep feature extraction and the classification stage, is proposed for ESC problem. To this end, an end-to-end CNN model is constructed and trained with the spectrogram images. Thus, we obtain our own pre-trained CNN model. Then, the fully connected layers of the constructed CNN model are discarded for feature extraction. Thus, a flexible CNN architecture is obtained where the sizes and numbers of all layers are freely changed by the authors. In the classification stage of the proposed study, the random subspace KNN ensembles model is used, which uses the vote of many prediction scores in the subspace-feature sets. The classification accuracy is used to evaluate the performance of our proposed method. We further compare the performance of the proposed method with other pre-trained CNN models and classifiers for classification performance. The classification accuracies have been significantly improved by the proposed method compared to the other studies on the UrbanSound8K [5] and the DCASE-2017 ASC [6] datasets.

The main contribution of this paper is that a new CNN architecture for ESC classification is proposed. The proposed CNN model is not too deep which does not necessitates too much training time. In addition, the achievement of the proposed new CNN model is comparable with the pre-trained CNN models.

II. THE METHODOLOGY

The illustration of the proposed method is shown in Fig. 1. According to the method, the input sound signals are initially converted into time-frequency images by using the spectrogram method. The spectrogram parameters such as the window type, window length and the overlap size are adjusted during the experimental works. Later, the spectrogram images are saved by using the viridis colour map and are resized to fit them for the input of the proposed CNN model. The proposed CNN model, which is shown in Fig. 2, is constituted of three convolution, three max-pooling and normalization and three fully connected layers. The softmax and classification layers were followed the last fully connected layer. The rest part of the used datasets is utilized for the feature extraction and the testing process. The feature set is achieved with concatenating the outputs of the first and second fully connected layers of the proposed CNN. Finally, the performance of proposed method is tested with the random subspace KNN ensembles, which are used a robust classification algorithm.

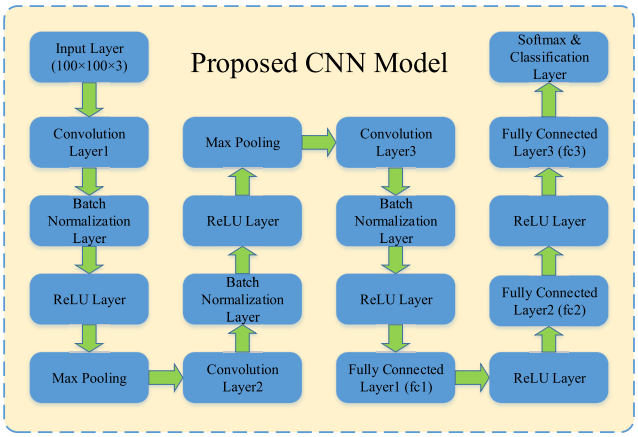

FIGURE 2. Illustration of the proposed method.

A. SPECTROGRAM IMAGES

The spectrogram method converts the signals into time frequency images or loudness of a signal over time at different frequencies existing in a specific waveform. The spectrograms also show how energy levels vary over time. Spectrogram of an input signal can be described as the square of the Short Time Fourier Transform (STFT) magnitude. The STFT formulation is given as follows.

$$F(n, \omega) = \sum_{i=-\infty}^{\infty} x(i)\, \omega(\text{n-i})\, e^{-jwn} \tag{1}$$

where x(i) is input signal, and ω(i) that is generally centered at the time n is a window function such as Hamming and Hanning window. Then, the spectrogram images are saved via viridis colour map which is a homogenous colour map changing from blue to green to yellow [20], [21].

B. CNN LAYERS

The CNN is designed to process data, which is taken from the multidimensional data, i.e., a colour image composed of three 2D data including pixel density in the 3D channels. CNNs use the properties of natural signals organized at four key ideas that consist of shared weights, local connections, pooling and other layers [22], [23]. Convolutional layer, ReLU layer and pooling layer are the most used CNN layers.

The basic aim of the convolutional layers is to determine local connections of features from the previous layers and mapping their information to particular feature maps. The convolution of the input I with filter F ($F \epsilon \mathbb{R}^{2a_1+2a_2}$) is given as follows.

$$(I * F)_{n,m} = \sum_{k=-a_1}^{a} {}_1 \sum_{l=-a_2}^{a} {}_2 F_{k,l} I_{n-k,m-l} \tag{2}$$

ReLU ($g(z) = \max(0, z)$) which is a non-linearity activation function, is applied the feature maps created with the convolutional layers. The task of the max-pooling layers is to combine similar features conveyed from the previous layer. The max-pooling layers realize down-sampling operation by calculating the maximum value of the field on the feature map overlapping with the filter [23].

CNN structure, which is from the fully connected (fc) layer to classification layer, is in general similar to the multi-layer perceptron neural network (MLP). The task of the fc layers is the same as the hidden layers in the MLP. One or more the fc layer can be in a CNN structure. The fc layer connects each neuron in next layer to each neuron in previous layer.

Softmax function is generally utilized in CNNs, to match the non-normalized values of previous layer to a possibility distribution over predicted class scores [24].

$$\boldsymbol{\sigma}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}, \quad j = 1, \ldots, K \tag{3}$$

where $\boldsymbol{\sigma}(\boldsymbol{x_i})$ is the softmax output for each $\boldsymbol{x_i}$, and $\boldsymbol{x_j}$ represents values of the input vector.

The batch normalization layers are used to decrease training time of CNNs and the sensitivity to network initialization [27]. Therefore, this layer is chosen for the normalization process in the proposed CNN architecture. The normalized activations with input (x_i), mini-batch mean (m_b) and mini-batch variance (v_b) variables is computed as

$$\hat{x}_i = \frac{x_i - m_b}{\sqrt{v_b^2 + \epsilon}} \tag{4}$$

where ϵ is constant and develops the numerical state in case thev_b is very small. The m_b and the v_b calculations are also shown in equations (5) and (6), respectively.

$$m_b = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5}$$

$$v_b = \frac{1}{n} \sum_{i=1}^{n} (x_i - m_b)^2 \tag{6}$$

Finally, the activations in the batch normalization layer is concluded with shift and scale operation as

$$y_i = a\hat{x}_i + b \tag{7}$$

where a and b are balance and scale factors, respectively. These factors are learnable variables updated to the most appropriate values during training process.

C. DEEP FEATURE EXTRACTION WITH THE PROPOSED CNN MODEL

The feature extraction processing with the pre-trained CNN models is called as deep feature extraction in literature [16], [25], [26]. For deep feature extraction, it is used the fc layers of the pre-trained CNN models. In the paper, instead of the pre-trained CNN models such as VGGNet and AlexNet, the fc layers of the proposed CNN are utilized for deep feature extraction. The layer numbers of the proposed CNN, Alexnet, VGG16, VGG19 and ResNet-50 is given in Tab 1.

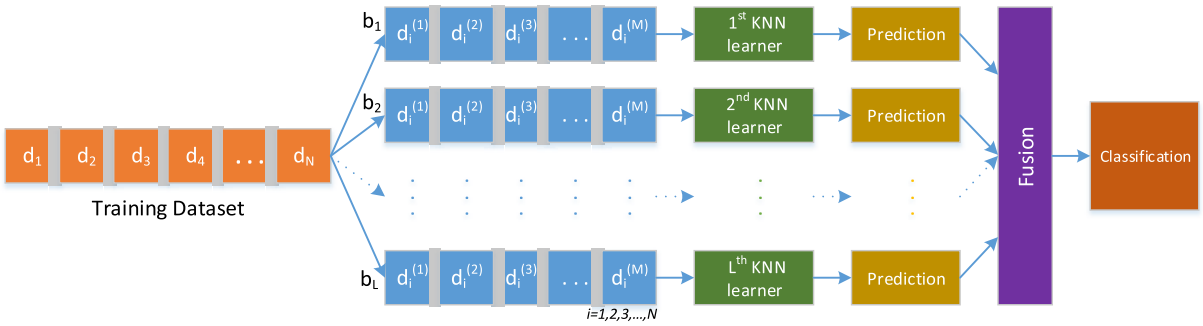

FIGURE 3. Representation of the random space ensemble.

TABLE 1. The layer info of the proposed CNN and the pre-trained CNN models.

Layers	AlexNet	VGG16	VGG19	ResNet-50	Proposed CNN
convolution	5	13	16	53	3
pooling	5	5	5	2	2
ReLU	7	15	18	49	5
Input	227×227	224×224	224×224	224×224	100×100
First fc	4096	4096	4096	-	500
Second fc	1000	1000	1000	1000	450

D. KNN ENSEMBLES WITH RANDOM SUBSPACE METHOD

The random subspace method is used random subspace ensembles to boost the classification accuracy of k-nearest neighbor (KNN) classifiers. The method bases on a stochastic operation that randomly chooses a number of components of the learning model in creating of each classifier [28]. In the method, the training dataset is sub-divided into random subspaces and distance calculations such as Euclidean and Chebyshev are performed by using the test samples on training set constituting with the random subspaces. According to the number of nearest neighbors (K), the most appropriate subspace class membership is determined by the distance and majority voting [29]. Then, class memberships coming with each subspace ensemble is assembled in a class vector (C). The classification is achieved with highest average score in C. The base random subspace method implements the following items:

- Step 1: Select without changing a stochastic set of the M-size from training dataset (M<N).
- Step 2: Train a KNN learner using only the selected predictors (b).
- Step 3: Repeat step1 and step2 until there are L KNN learners.
- Step 4: Constitute by averaging prediction values of the KNN learners
- Step 5: Classify the test dataset with the highest average value.

Where d is numeric values in the training dataset, b is the selected subspace predictor, M is length of the b predictors, and L is the number of learners in the ensemble. In Fig. 3, representation of the random subspace ensemble method is shown for the KNN classifier.

III. EXPERIMENTAL WORKS

A. DATASETS

In this work, two popular datasets are considered to evaluate the ESC problem. UrbanSound8K dataset is organized with ten class labels consisting of air conditioner, car horn, children, dog bark drilling, engine idling, gun shot, jackhammer, siren, and street music. The record duration for an audio file of the dataset, which contains 8732 audio files, is up to 4 seconds and the audio files are recorded with 22.05 KHz sampled frequency. Also the record lengths of the audio file and the number of files in each class are not same. The DCASE-2017 ASC dataset is constituted of two part including the development dataset with 4680 audio files and the evaluation dataset with 1620 audio files. The duration of each audio file is 10 second. The file numbers of each class are balance, and all audio files are recorded with 44.1 KHz sampled frequency. The dataset contains fifteen classes of which labels are beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, park, residential area, train, tram. The performances in the DCASE-2017 challenge have been ranked to classification accuracy on the evaluation data.

B. EVALUATION METHOD AND CRITERIA

The development and the evaluation datasets, which the DCASE-2017 ASC dataset contains, are used for the proposed CNN training and the evaluation processes, respectively. On the other hand, the UrbanSound8K dataset is randomly divided for the proposed CNN training process with a ratio of 0.9 of the full dataset, and the evaluation process is performed with the rest part of the full dataset. The classification performances on the UrbanSound8K dataset is tested with 10-fold cross-validation. The evaluation criteria consist of accuracy, specificity, sensitivity, precision, and F-score. These criterions are computed by using the confusion matrix values as given the following equations.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

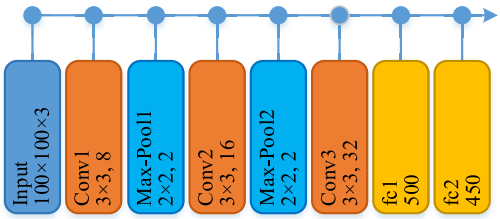

FIGURE 4. Dimensional parameters related to the proposed CNN structure.

TABLE 2. The feature extraction duration for all CNN models.

Datasets	Alexnet	VGG16	VGG19	ResNet-50	Proposed CNN
DCASE-2017 ASC	2432 sec	2835 sec	3145 sec	4670 sec	576 sec
UrbanSound8K	2657 sec	3015 sec	3467 sec	4876 sec	650 sec

TABLE 3. The classification accuracy on the evaluation data according to the sizes of the FC layers in the proposed CNN.

The sizes of the FC layers		The used datasets			
		UrbanSound8K (%Acc)		DCASE-2017 ASC (%Acc)	
fc1	fc2	fc1	fc2	fc1	fc2
100	50	78.6	77.4	84.5	84.1
150	100	79.2	78.3	85.7	85.2
200	150	80.3	79.4	86.9	85.6
250	200	81.1	80.4	88.7	87.5
300	250	82.6	81.3	90.2	89.4
350	300	83.1	82.9	91.3	90.8
400	350	83.4	83.3	92.8	91.9
450	400	84.1	83.7	94.7	93.8
500	**450**	**85.5**	**84.6**	**95.3**	**94.2**
550	500	84.2	83.9	93.7	92.5
600	550	83.7	82.5	91.4	90.6
650	600	82.8	81.5	86.5	84.2

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$F - score = 2x\frac{PrecisionxSensitivity}{Precision + Sensitivity} \tag{12}$$

C. EXPER MENTAL SETUP AND RESULTS

As it was mentioned earlier, the spectrogram method was applied to all the audio signals to convert the input audio signals to the time-frequency images. Window size, window type, overlap and FFT size parameters of the spectrogram method were chosen as 1024, Hamming, 256, 3000, respectively. These spectrogram parameters were selected according to the optimum resolution of the spectrogram images. The dimensions of the spectrogram images were 875 × 656 × 3 and then were re-sized to 100 × 100 × 3 for the input of the proposed CNN model. The re-sized spectrogram images were fed into the proposed CNN model. The dimensional parameters in the proposed CNN layers are shown in Fig. 4. For example, the filter size and the filter number in the first convolutional layer were assigned as 3×3 and 8, respectively. And, the pixel block size and the stride were selected as 2 × 2 and 2 for the max-pooling layers,

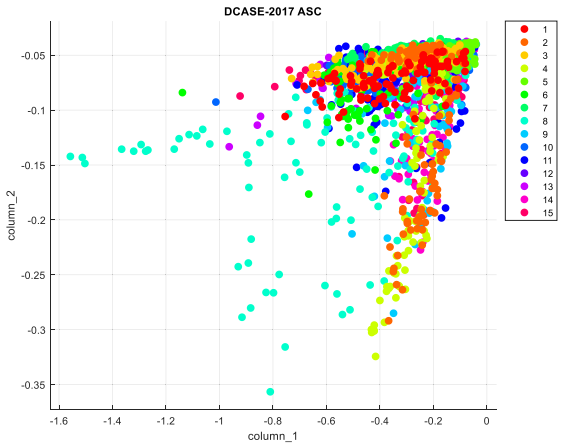

FIGURE 5. The scatter plot of the concatenated features for the DCASE2017-ASC dataset.

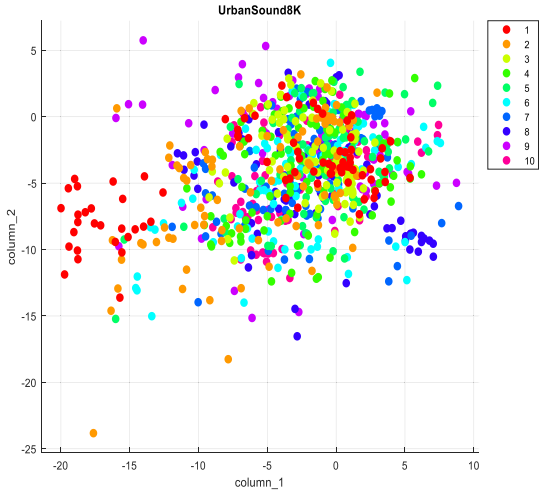

FIGURE 6. The scatter plot of the concatenated features for the UrbanSound8K dataset.

TABLE 4. The effect of input size on the performance of the proposed method.

The input size	UrbanSound8k (Acc%)	DCASE-2017 ASC (Acc%)
20×20	72.10%	79.60%
50×50	79.20%	88.10%
100×100	86.70%	96.23%
200×200	83.10%	92.10%

TABLE 5. Comparison for the DCASE-2017 ASC to the proposed method with the other CNN models and classifiers.

Classifiers	CNN models				
	AlexNet	VGG16	VGG19	ResNet-50	Proposed CNN
Fine Tree	61.56%	62.32%	61.44%	60.10%	67.20%
KNN	69.18%	70.78%	67.16%	67.54%	79.10%
SVM	80.35%	81.20%	79.34%	79.10%	85.40%
Boosted Trees Ensembles	55.92%	58.56%	57.30%	56.10%	68.40%
Bagged Trees Ensembles	55.30%	56.18%	55.68%	55.45%	75.40%
Subspace Discriminant Ensembles	68.20%	69.26%	67.96%	66.12%	88.20%
Subspace KNN Ensembles	76.15%	79.34.%	75.82%	74.53%	**96.23%**

TABLE 6. Comparison for the Urbansound8k to the proposed method with the other CNN models and classifiers.

Classifiers	CNN models				
	AlexNet	VGG16	VGG19	ResNet-50	Proposed CNN
Fine Tree	35.70%	38.10%	34.50%	35.20%	44.10%
KNN	70.35%	71.35%	69.80%	70.80%	84.20%
SVM	76.20%	77.10%	75.00%	75.95%	78.00%
Boosted Trees Ensembles	44.60%	45.80%	42.90%	43.85%	50.00%
Bagged Trees Ensembles	63.10%	64.55%	62.10%	63.20%	67.80%
Subspace Discriminant Ensembles	60.50%	61.60%	59.20%	60.65%	65.40%
Subspace KNN Ensembles	71.35%	72.35%	70.65%	70.90%	**86.70%**

respectively. The mini-batch size with 128, the initial learning rate with 0.005 and the '*adam*' optimizer were the option parameters used in the training process of the proposed CNN model. The training durations on both datasets are given for the feature extraction of all CNN models in Tab. 2. For both datasets, the feature extraction of the proposed CNN has been completed in less time than other CNN models. According to the results given in Tab. 3, the sizes of the first and second fully connected layers (fc1, fc2) are selected as 500 and 450, respectively. The scatter plots of the features concatenated with the fc1 (the size of 500) and fc2 (the size of 450) are shown in Figs. 5 and 6 for both datasets. These parameters are selected during the experiments and the configuration is selected which yields the best accuracy score.

As shown in Tab. 4, The input size of the proposed CNN model is changed by basing the classification accuracy. The best result is reached with the size of 100×100.

TABLE 7. The effect of cross validation on the performance of the proposed method.

Dataset	5-fold cross validation	10-fold cross validation
UrbanSound8K	84.20%	86.70%

TABLE 8. The scores of the other performance criteria for the DCASE-2017 ASC dataset.

Classes	The other performance criteria			
	Sensitivity	Specificity	Precision	F-score
Beach	0.9722	0.9980	0.9722	0.9722
Bus	0.9722	0.9987	0.9813	0.9767
Cafe/Restaurant	0.9630	0.9947	0.9286	0.9455
Car	0.9630	0.9987	0.9811	0.9720
City center	0.9630	0.9954	0.9369	0.9498
Forest Path	0.9630	0.9967	0.9541	0.9585
Grocery store	0.9722	0.9967	0.9545	0.9633
Home	0.9259	0.9974	0.9615	0.9434
Library	0.9352	0.9947	0.9266	0.9309
Metro station	1.000	1.000	1.000	1.000
Office	0.9630	0.9974	0.9630	0.9630
Park	0.9722	1.000	1.000	0.9859
Residential Area	0.8981	0.9954	0.9327	0.9151
Train	0.9722	1.000	1.000	0.9859
Tram	1.000	0.9960	0.9474	0.9730

TABLE 9. The scores of the other performance criteria for the Urbansound8k dataset.

Classes	The other performance criteria			
	Sensitivity	Specificity	Precision	F-score
Air conditioner	0.8810	0.9899	0.9024	0.8916
Car horn	0.8506	0.9771	0.8043	0.8268
Children playing	0.8090	0.9847	0.8571	0.8324
Dog bark	0.9022	0.9885	0.9022	0.9022
Drilling	0.8889	0.9819	0.8627	0.8756
Engine idling	0.8617	0.9846	0.8710	0.8663
Gun shot	0.8902	0.9848	0.8588	0.8743
Jackhammer	0.8095	0.9797	0.8095	0.8095
Siren	0.9136	0.9950	0.9487	0.9308
Street music	0.8659	0.9861	0.8659	0.8659

The k, which is the nearest neighbor number, and f, which is the size of the subspace feature vector, is the most important parameters of the random subspace k-NN ensembles.

TABLE 10. Classification accuracies of the proposed methods for the DCASE-2017 ASC.

class index	Classification accuracies (%) for the used methods						
	ISPL [15]	JKU_All_ca [14]	BUETBOSCH1 [13]	DCNN_SVM [12]	F1EnsemSel [11]	GAN_SKMUN [10]	The Proposed Method
1	54.6	87.0	87.0	71.3	78.7	83.3	**97.2**
2	59.3	66.7	59.3	84.3	71.3	74.1	**97.2**
3	71.3	88.9	91.7	79.6	83.3	88.0	**96.3**
4	79.6	80.6	92.6	85.2	93.5	93.5	**96.3**
5	91.7	92.6	94.4	82.4	88.9	94.4	**96.3**
6	85.2	92.6	91.7	78.7	**98.1**	95.4	96.3
7	75.0	76.9	81.5	80.6	79.6	82.4	**97.2**
8	98.1	88.9	**97.2**	73.1	94.4	88.0	92.6
9	44.4	49.1	47.2	59.3	53.7	75.9	**93.5**
10	98.1	79.6	76.9	97.2	100	88.0	**100**
11	84.3	65.7	49.1	81.5	86.1	92.6	**96.3**
12	23.Oca	45.4	38.0	57.4	44.4	75.9	**97.2**
13	76.9	55.6	58.3	85.2	75.9	86.1	**90**
14	82.4	84.3	81.5	92.6	90.7	67.6	**97.2**
15	64.8	53.7	65.7	57.4	66.7	63.9	**100**
Average Accuracy	72.6	73.8	74.1	77.7	80.4	83.3	**96.2**

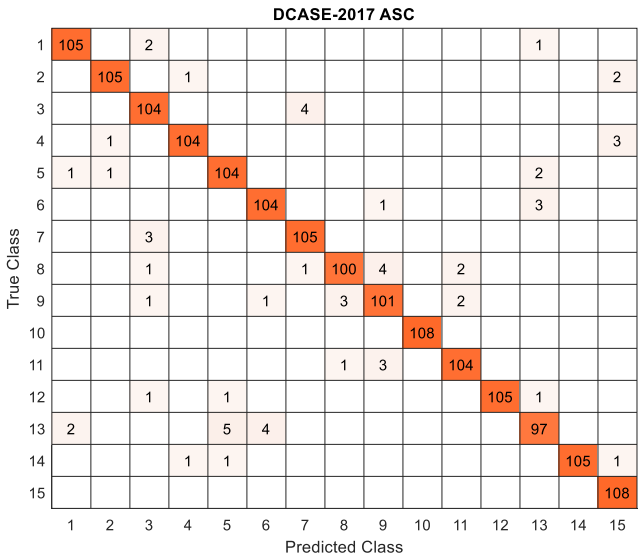

FIGURE 7. The scores of the other performance criteria for the DCASE-2017 ASC dataset.

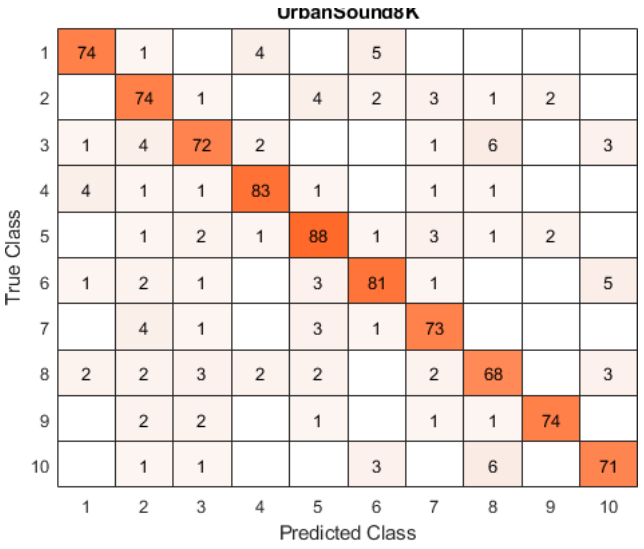

FIGURE 8. The scores of the other performance criteria for the DCASE-2017 ASC dataset.

According to the experiments in [28], k and f give the best performances for 1 and 64, respectively.

For both the datasets in Tabs. 5 and 6, the proposed method is compared with the pre-trained CNN models and the other classifiers. The obtained results showed that classification accuracy of the proposed method was better than the other CNN model-classifier structures.

The average classification accuracies for the DCASE-2017 ASC and the UrbanSound8K datasets has been increased by 15% and 9.6% compared to the other best CNN model-classifier structure, respectively. The other performance

TABLE 11. Classification accuracies of the proposed methods for the Urbansound8k.

class index	Classification accuracies (%) for the used methods				
	Baseline System [5]	Piczak CNN [4]	SKM [31]	SB-CNN [8]	The Proposed Method
1	50.40	55.70	51.30	48.90	**88,1**
2	46.97	78.60	63.25	**88.13**	85,1
3	71.60	82.00	76.60	**83.00**	80,9
4	74.90	84.00	79.50	90.00	**90,2**
5	76.60	66.30	79.90	80.20	**88,9**
6	64.10	67.90	77.20	79.80	**86,2**
7	90.64	92.51	91.71	**94.11**	89
8	63.00	62.70	70.20	67.30	**81**
9	74.00	81.05	75.67	85.79	**91,4**
10	75.00	76.00	77.00	84.40	**86,6**
average accuracy	68.57	73.09	73,69	78,65	**86,7**

criteria including sensitivity, specificity, precision and F-score is separately given in Tabs. 8 and 9 for each class of both datasets. The average scores of the sensitivity, specificity, precision, and F-score for the DCASE-2017 ASC dataset are 0.9623, 0.9973, 0.9626, and 0.9623, respectively. The same scores for the Urbansound8K dataset are 0.8672, 0.9852, 0.8682 and 0.8675, respectively. In Figs. 7 and 8, the states of TP, TN, FP, and FN in both datasets are shown for each class on the confusion matrices. In Tabs. 10 and 11, the proposed method is compared with the other method using the same datasets. The first ten works achieving the best classification accuracy in the DCASE-2017 ASC challenge is used for the comparison. The average classification accuracy with the proposed method has been boosted by 12.93% compared to the best challenge score [10]. In addition, the best classification accuracy has been achieved in 13 out of 15 classes, with a significant difference in most. For the UrbanSound8K dataset, the average classification accuracy has been improved by 8.05% compared to the best score [8] in the used other methods, and the best classification accuracy has been achieved in 8 out of 10 classes. For the UrbanSound8K dataset, the 5-fold cross validation test is also applied and the obtained result is given Tab. 7. As seen in Tab 7, when 5-fold cross validation test is used in evaluation of the proposed method, 84.20% average accuracy score is obtained, that score is 86.70% for 10-fold cross validation.

From this comparison, it is observed that an increase in fold number causes an increase in the accuracy score. It is also worth to mentioning that smaller amount of training data causes low achievement, as it is obvious almost in pattern recognition problems.

IV. CONCLUSION

In this a paper, a new CNN model was developed and trained in end-to-end fashion in order to produced deep feature vectors for efficient classification of the environmental sounds. The developed CNN model was consisted of three convolution, three max-pooling and normalization and three fully connected layers. The softmax and classification layers were followed the last fully connected layer. The proposed new CNN model was quite effective in both classification and running time. After training of the proposed new CNN model, instead of using the softmax and classification layers, we opted to used deep feature extraction. These deep features were then used as input to the random subspaces K Nearest Neighbor (KNN) classifier. This classifier was chosen due to its robustness against various dataset. The DCASE-2017 ASC and the UrbanSound8K datasets were considered in experimental works and the classification accuracies were calculated for performance evaluation. The obtained results show that the proposed CNN model and subsequent deep features were quite successful in characterization of the environmental sounds. The performance of the proposed method was also compared with the state-of-the-art results. The comparison results showed that the proposed method outperformed in all compared methods.

REFERENCES

[1] R. V. Sharan and T. J. Moir, "Robust acoustic event classification using deep neural networks," *Inf. Sci.*, vol. 396, pp. 24–32, Aug. 2017, doi: 10.1016/j.ins.2017.02.013.

[2] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," 2016, *arXiv:1604.07160*. [Online]. Available: http://arxiv.org/abs/1604.07160

[3] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.

[4] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. for Signal Process. (MLSP)*, Sep. 2015, pp. 1–6.

[5] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 1041–1044.

[6] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2016, pp. 1128–1132.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[8] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017, doi: 10.1109/LSP.2017.2657381.

[9] Y. Chen, Q. Guo, X. Liang, J. Wang, and Y. Qian, "Environmental sound classification with dilated convolutions," *Appl. Acoust.*, vol. 148, pp. 123–132, May 2019.

[10] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," in *Proc. DCASE*, 2017, pp. 93–97.

[11] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proc. Detection Classification Acoustic Scenes Events Workshop*, 2017, pp. 1–5.

[12] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2017, pp. 1–5.

[13] R. Hyder, S. Ghaffarzadegan, Z. Feng, and T. Hasan, "Buet Bosch consortium (B2C) acoustic scene classification systems for DCASE 2017 challenge," in *Proc. Detect. Classif. Acoust. Scenes Events*, 2017, pp. 1–4.

[14] B. Lehner, H. Eghbal-Zadeh, M. Dorfer, F. Korzeniowski, K. Koutini, and G. Widmer, "Classifying short acoustic scenes with I-vectors and CNNs: Challenges and optimisations for the 2017 DCASE ASC task," in *Proc. IEEE AASP Chall. Detect. Classif. Acoust. Scen Events*, 2017, pp. 1–5.

[15] S. Park, S. Mun, Y. Lee, and H. Ko, "Acoustic scene classification based on convolutional neural network using double image features," in *Proc. Detection Classification Acoustic Scenes Events Workshop*, Munich, Germany, 2017, pp. 1–5.

[16] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[17] F. Demir, V. Bajaj, M. C. Ince, S. Taran, and A. engür, "Surface EMG signals and deep transfer learning-based physical action classification," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 8455–8462, Dec. 2019.

[18] M. To aćar, Z. Cömert, and B. Ergen, "Classification of brain MRI using hyper column technique with convolutional neural network and feature selection method," *Expert Syst. Appl.*, vol. 149, Jul. 2020, Art. no. 113274.

[19] X. Wang, H. Qian, E. J. Ciaccio, S. K. Lewis, G. Bhagat, P. H. Green, S. Xu, L. Huang, R. Gao, and Y. Liu, "Celiac disease diagnosis from videocapsule endoscopy images with residual learning and deep feature extraction," *Comput. Methods Programs Biomed.*, vol. 187, Apr. 2020, Art. no. 105236.

[20] Y. Zhou, H. Nejati, T.-T. Do, N.-M. Cheung, and L. Cheah, "Image-based vehicle analysis using deep neural network: A systematic study," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Oct. 2016, pp. 276–280.

[21] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jun. 2015, pp. 1–6.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[24] F. Zang and J.-S. Zhang, "Softmax discriminant classifier," in *Proc. 3rd Int. Conf. Multimedia Inf. Netw. Secur.*, Nov. 2011, pp. 16–19.

[25] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

[26] Y. Guo, H. Cao, J. Bai, and Y. Bai, "High efficient deep feature extraction and classification of spectral-spatial hyperspectral image using cross domain convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 345–356, Jan. 2019.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[28] T. K. Ho, "Nearest neighbors in random subspaces," in *Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern Recognit. (SPR) Struct. Syntactic Pattern Recognit. (SSPR)*, 1998, pp. 640–648.

[29] G. Tremblay, R. Sabourin, and P. Maupin, "Optimizing nearest neighbour in random subspaces using a multi-objective genetic algorithm," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 1, 2004, pp. 208–211.

[30] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 171–175.

FATIH DEMIR received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Firat University, Turkey, in 2007 and 2010, respectively, where he is currently pursuing the Ph.D. degree in signal and sound processing. His research interests include environmental sound classification, biomedical signals, and deep learning.

DABAN ABDULSALAM ABDULLAH received the B.Sc. degree in computer science from the University of Human and Development, Iraq, in 2012, and the M.Sc. degree in applied mathematics and computer science from Eastern Mediterranean University, Cyprus, in 2015. He became a Lecturer with the Technical College of Informatics, Sulaimani Polytechnic University, in 2018. His research interests include data mining, association rules mining, and pattern recognition.

ABDULKADIR SENGUR received the B.Sc. degree in electronics and computer education, the M.Sc. degree in electronics education, and the Ph.D. degree in electrical and electronics engineering from Firat University, Turkey, in 1999, 2003, and 2006, respectively. He became a Research Assistant with the Faculty of Technical Education, Firat University, in February 2001, where he is currently a Professor with the Faculty of Technical Education. His research interests include signal processing, image segmentation, pattern recognition, medical image processing, and computer vision.