Chapter 11

AN EXAMINATION OF VGGISH EMBEDDINGS USAGE IN ENVIRONMENTAL SOUND CLASSIFICATION

Ilker Ali OZKAN[1]

1 Department of Computer Engineering, Selcuk University, Konya, Turkey. E-mail: ilkerozkan@selcuk.edu.tr, ORCID: 0000-0002-5715-1040

Introduction

Environmental sounds usually include daily sound events in their structure. These sound events can occur from live sources such as cats, birds, bee sounds or non-living sources such as wind, rain, engine, siren. It does not consist of musical or speech data and has in its structure daily sound events which are often more diverse and complex. Unlike speech sounds, these sounds, which are made up of many different sources, also naturally have a noise (Piczak, 2015b; Z. Zhang, Xu, Zhang, Qiao, & Cao, 2019). Analyzing images and videos is often used to describe objects or events in our environment. In addition to these, phonetic information includes a lot of information about our environment (X. Zhang, Zou, & Shi, 2017). By analyzing these environmental sounds, we can obtain information about many events in our environment.

Environmental Sound Classification (ESC) has an increasing interest by researchers (Dogan, Akbal, & Tuncer, 2020). In the classification of the sound, firstly, the process of extracting its features is done. In general, Mel Frequency Cepstrum Coefficients (MFCC) is considered a good manual feature extraction method for the classification of sound signals. Although MFCC is very good for the classification of speech signals, they do not fully reflect the characteristics of non-speech signals (Uzkent, Barkana, & Cevikalp, 2012). Depending on this, researchers continued to study over developing various feature extraction methods as well as MFCC features. For example, Bansal et al. stated that MFCC features are frequently used in the classification of environmental sounds. In their study, they have proposed EMD (empirical mode decomposition) as a new feature extraction approach. In the study that they did with different machine learning methods, they stated that the use of MFCC + EMD features together increases the performance on the classification of environmental sounds (Bansal, Shukla, Goyal, & Kumar, 2021). Chu et al. have proposed several features to fulfill the task of recognizing environmental sounds, including the popular MFCC which describes the spectral shape of sound. They also performed an empirical feature analysis for the characterization of the sound environment and proposed the Matching Pursuit (MP) algorithm to obtain effective time-frequency features. By combining the MP and MFCC features, they achieved an accuracy of 83.9% in distinguishing fourteen classes (Chu, Narayanan, & Kuo, 2009). Wang et al. have obtained the sound properties with spectrum centroid, spectrum spread, and spectrum flatness on a 12-class non-speech database. They achieved 85.10% accuracy in a new sound classification architecture that they obtained by using the hybrid support vector machine (SVM) and k-nearest neighbor (k-NN) delimiters (Jia-Ching Wang, Jhing-Fa Wang, Kuok Wai He, & Cheng-Shu Hsu, 2006).

In addition, deep learning methods have been used in the classification of environmental sounds in recent years. Convolutional neural networks (CNN), with their ability to learn spectro-temporal patterns, are well suited to the environmental sound classification problem (Salamon & Bello, 2017). In a study to evaluate whether convolutional neural networks can be successfully applied to environmental sound classification tasks, they determined that a convolutional model outperforms hand-crafted feature approaches (Piczak, 2015a). As a result of training the CNN model with 5-fold cross-validation on the ESC-50 dataset, it outperformed applications using manually designed features with 64.5% accuracy performance in the best network (Piczak, 2015a). In another study, AlexNet and GoogleNet convolutional neural networks, which are used in image recognition, were used to classify different image representations such as spectrogram obtained from environmental sounds. In the study conducted on three separate environmental datasets, the GoogleNet model provided the best classification performance. Researchers stated that GoogleNet has a deeper network architecture than AlexNet, which affects performance (Boddapati, Petef, Rasmusson, & Lundberg, 2017). Tokozume et al. proposed an end-to-end environmental sound classification system, which they named EnvNet, using CNN in their study. In their study, they compared the performances for different time periods ranging from 0.5 s to 5 s for the input length. They stated that the accuracy was at a high level in the T=1.0~2.5 sec range. In their study, they achieved a 6.5% improvement in classification performance as a result of combining EnvNet with logmel-CNN (Tokozume & Harada, 2017).

VGGish (Hershey et al., 2017) is a CNN-based model trained on audio from 8 million YouTube videos to distinguish 3,000 sound classes. The VGGish model, obtained by modifying VGG16, is widely used in sound recognition. This model creates a sound classification based on log-mel spectrograms and CNN. This deep learning model has six convolutional layers and three fully connected layers, respectively. A 128-dimensional feature can be obtained for each one second segment of the input sound.

In this study, both handcrafted and deep learning methods were used for feature extraction in the classification of environmental sounds. Thus, it is aimed to compare the embeddings obtained by using a previously trained deep learning model, VGGish, with the features obtained by MFCC. The k-NN type classifier, which is mostly used in solving classification problems and has many studies on sound classification, has been used in the classification of environmental sounds with the obtained features.

Material and Methods

The study includes finding both the event that causes the environmental sound and the source group from which the environmental sound is derived. Therefore, two datasets were used. The first of these includes ten various environmental sounds. The second is categorized into five different groups according to the sounds coming from similar sources. Two methods were used to extract the features of the environmental audio signals and their performances were compared. These methods are MFCC and VGGish.

Environmental Sound Classification Dataset (ESC)

In this study, the publicly available ESC-50 dataset was used (Piczak, 2015b). The ESC-50 dataset is used to compare the environmental sound classification methods. The ESC-50 dataset was created by combining short environmental sound recordings belonging to fifty different classes, and two thousand environmental sound recordings were labeled. All sound recordings were obtained from the freesound.org project, which is an open-access database. The audio clips are 5 seconds long and have a sampling rate of 44100 Hz. There are forty audio clips in each environmental sound class.

The first dataset used in the study is the ESC-10 dataset, which is a smaller subset of ESC-50 and includes ten selected classes. This dataset represents three common sound groups. The first is transient/percussive sounds, including sneezing, dog barking, clock ticking sounds. The second is sound events with harmonic content, including crying baby and crowing rooster sounds. The third is somewhat structural sound events that include the sounds of rain, sea waves, fire crackling, helicopter, chainsaws.

Fifty different classes in the ESC-50 dataset are given in Table 1. These classes are associated with five major categories. This categorically labeled dataset, which was created according to the origin of environmental sounds, is the second dataset used in this study.

Table 1. *Categorical representation of environmental sounds*

Category	Environmental Sounds
Animals	*Crow, Sheep, Insects(flying), Hen, Cat, Frog, Rooster, Dog, Pig, Cow*
Natural soundscapes & water sounds	*Thunderstorm, Toilet flush, Rain, Pouring water, Sea waves, Wind, Crackling fire, Water drops, chirping birds, Crickets*
Human, non-speech sounds	*Crying baby, brushing teeth, Sneezing, Snoring Clapping, Breathing, Coughing, Drinking, Footsteps, Sipping, Laughing*

Interior/domestic sounds	*Washing machine, Mouse click, Door, wood creaks, Clock alarm, can opening, Clock tick, Vacuum cleaner, Glass breaking, Door knock, Keyboard typing,*
Exterior/urban noises	*Chainsaw, Hand saw, Siren, Engine, Train, Airplane, Church bells, Fireworks, Car horn, Helicopter,*

The graph of sample environmental sound signals belonging to five major categories is given in Figure 1.

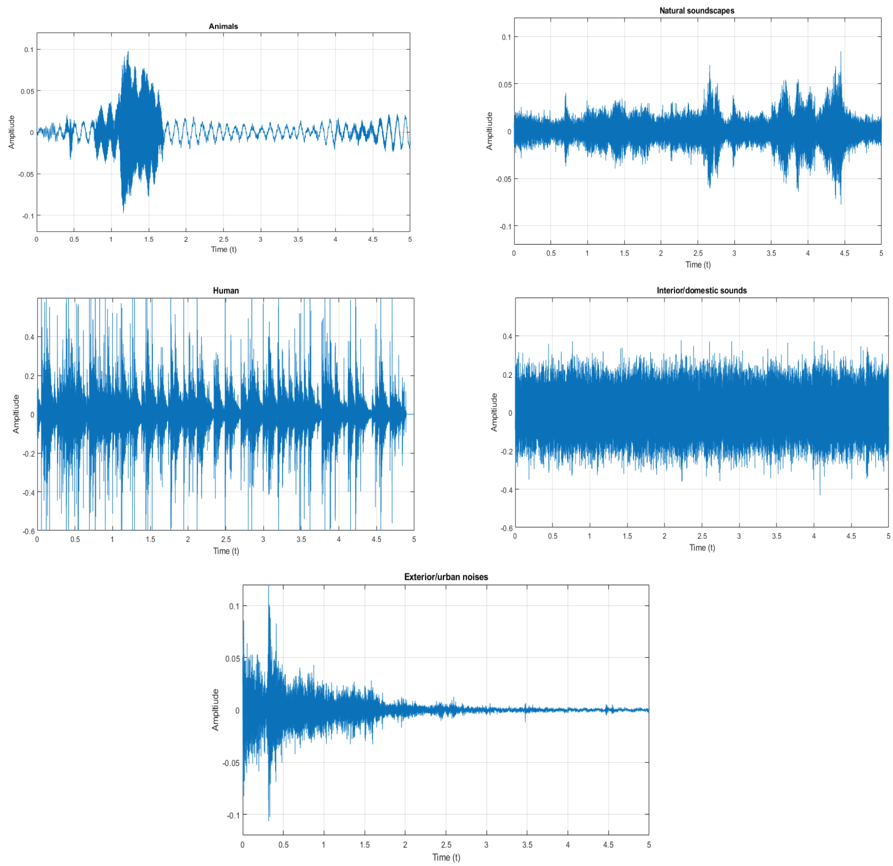


Figure 1. *Graphical representation of environmental sounds belonging to five major categories*

Feature Extraction (MFCC)

Firstly, the Mel Frequency Cepstrum Coefficients (MFCC) method was used to obtain the feature values of the ESC. MFCC is a frequently used and high-performance feature in sound recognition applications. MFCC is based on human hearing perception. Human ear sensitivity is linear up to 1 kHz and logarithmic at higher values. Accordingly, there are two types of filtering in the system with linear intervals below 1000 Hz and with logarithmic intervals above 1000 Hz. The features in the sound signal are captured as a result of this filtering. (Muda, Begam, & Elamvazuthi, 2010; Tombaloğlu & Erdem, 2016).

The conversion between the Mel scale and the frequency scale is provided by the equation in Equation 1.

$$Mel(f) = 2595x\log\left(1 + \frac{f}{100}\right) \tag{1}$$

MFCC is defined as the inverse Fourier transform of the logarithm of the time-dependent ESC(n) converted to the Mel scale Fourier transform. In this method, first, the Mel spectrum is obtained by multiplying the power spectrum of x(n) with a filter array arranged according to the Mel scale. Since the Mel spectrum does not contain complex numbers, it is sufficient to take the discrete cosine transform of the logarithm of the Mel spectrum to calculate the MFCC (Fig. 2).

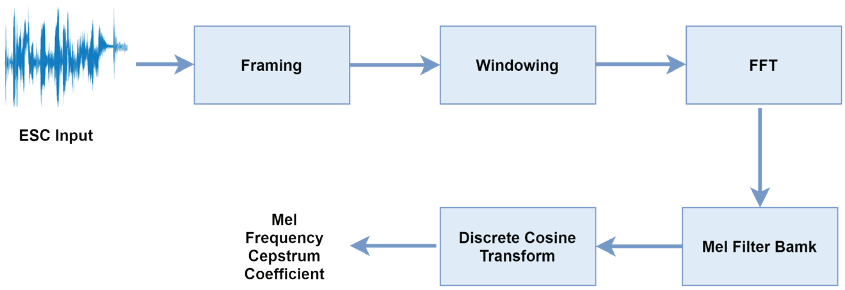


Figure 2. *MFCC Block Diagram*

As a result of the transformation, the MFCC are obtained. These coefficients express the acoustic features of the sound.

Feature Extraction (VGGish Embeddings)

Due to the superior performance of deep networks in the sound classification task, the VGGish model, an 11-layer CNN pre-trained on millions of audio clips from the YouTube 8M dataset, was used in this study. This model has been modified from the VGG16 architecture to include log-Mel spectrogram-based inputs for feature extraction (Hershey et al., 2017).

In this study, firstly, all audio clips were resampled at 16 kHz. Each environmental sound signal is equally divided into five parts, and each has one second. Mel-Spectrograms are considered the input to the model. A feature matrix of 64 × 96 Mel-Spectrograms is calculated for each one-second audio segment to match the input size of the VGGish model. Here 64 is the Melband number and 96 is the spectrum number in the individual Mel spectrograms. The output of the VGGish model is a 128-dimensional feature embed. VGGish is used to extract features from a 64x96 Mel-spectrogram. The output from VGGish is feature embeddings corresponding to each 0.975 ms audio data frame.

Figures 3-7 show a comparison of Mel-spectrogram and VGGish feature embedding visualizations for the Animals, Natural, Human, Interior, and Exterior major categories.

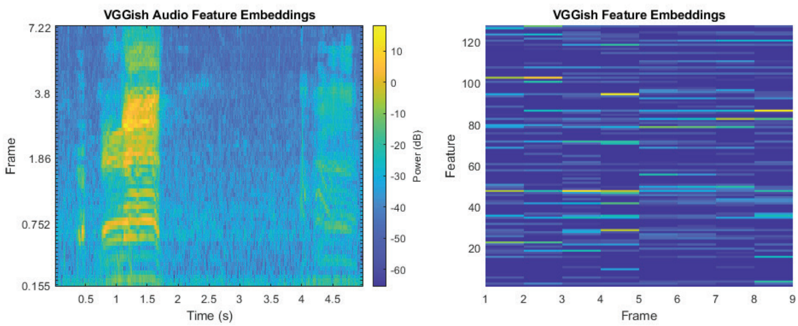


Figure 3. *Visualizing of the VGGish embeddings and the Mel spectrogram in the Animals category.*

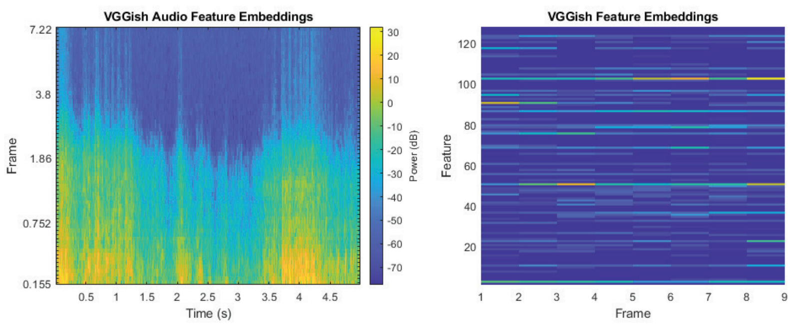


Figure 4. *Visualizing of the VGGish embeddings and the Mel spectrogram in the Natural category.*

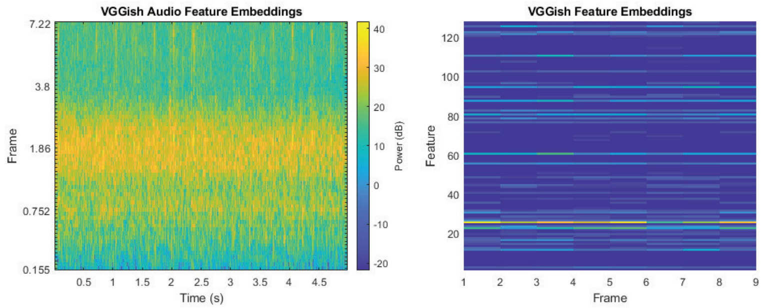


Figure 5. *Visualizing of the VGGish embeddings and the Mel spectrogram in the Human category.*

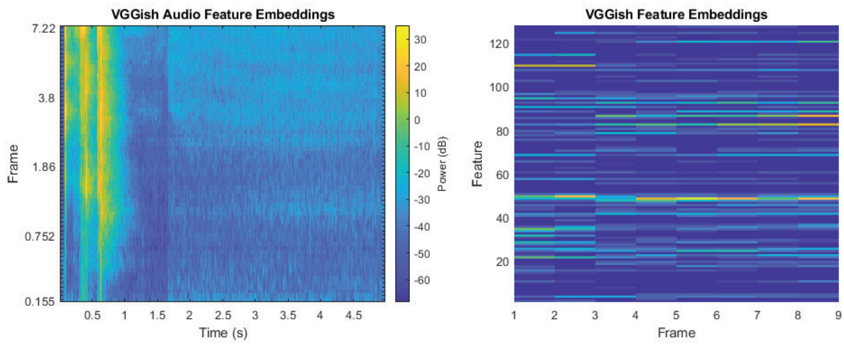


Figure 6. *Visualizing of the VGGish embeddings and the Mel spectrogram in the Interior category.*

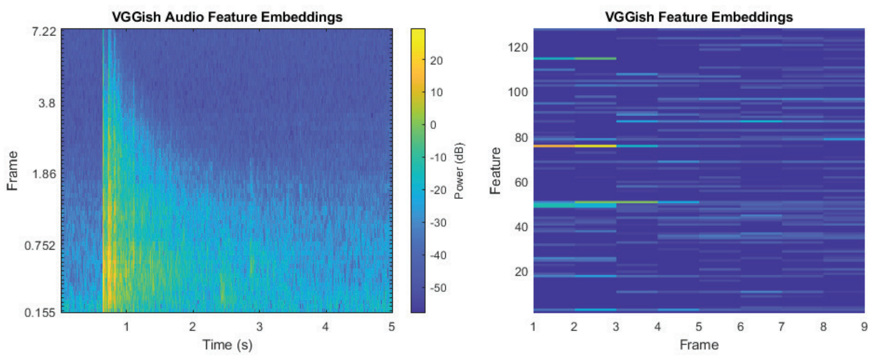


Figure 7. *Visualizing of the VGGish embeddings and the Mel spectrogram in the Exterior category.*

k-NN Algorithm

The k-nearest neighbor algorithm (k-NN) is one of the most well-known and used algorithms among machine learning algorithms. The output of K-NN classification is class membership. An item is classified by a majority vote of its neighbors. The item is assigned to the class that is most common among its nearest neighbors (Sutton, 2012). When an unknown item is encountered, the k closest samples are determined from the training set and the class label of the new sample is assigned according to the majority vote of the class labels of its k nearest neighbors (Bhatia, 2010). If k = 1, the item is simply assigned to the class of its nearest neighbor. The distance between the item to be classified and each training items is calculated with the Euclidean equation given in Equation 2 (Wasule & Sonar, 2017).

$$D(Y,X_i) = \sqrt{\sum_{j=1}^{n} (y_j - x_{ij})^2} \tag{2}$$

In Figure 8, the item u is classified as class C since three of its neighbors are of class C.

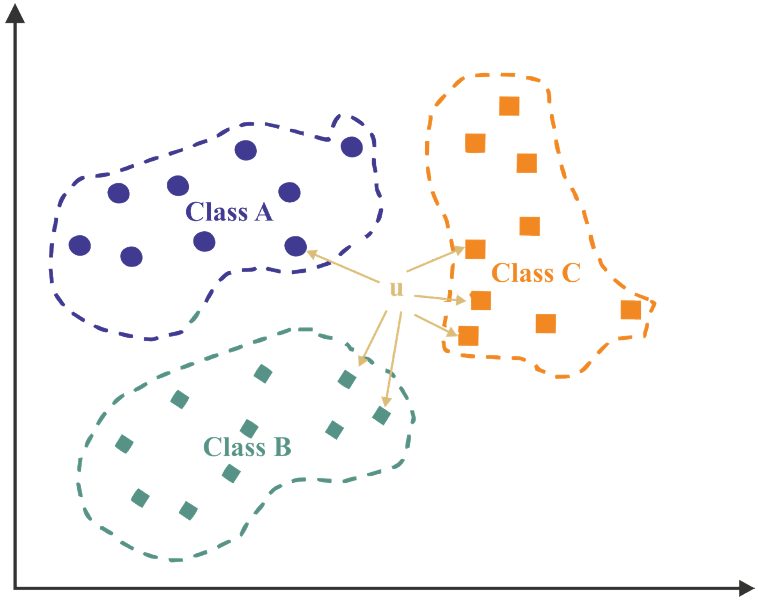


Figure 8. *Example of a three-class k-NN classifier*

In this study, trial-and-error method was applied to select the best k value and the k value was chosen as five.

Results

In this study, the ESC-10 dataset containing audio data from ten different categories on the ESC-50 dataset and the environmental dataset belonging to five major categories were used in the classification of environmental sounds. MFCC feature extraction and the VGGish feature extraction process were applied on both datasets. 5-fold cross-validation was applied to the obtained features. The accuracy rate of the model was used to measure the model performance. The k-NN classifier was used to classify the obtained features.

The MFCC coefficients obtained from the ESC-10 dataset were trained with the k-NN model, and the overall accuracy rate of the obtained model was obtained as 71.60%. The normalized complexity matrix of this model is given in Figure 9. In addition, Precision, Recall, and F1-Score values for each class of this MFCC feature-based model are given in Table 2.

Table 2. *Statistical values for each class in the MFCC-based ESC-10 dataset*

Class	Precision	Recall	F1-Score
Dog	0.51	0.81	0.63
Rain	0.64	0.90	0.75
Sea waves	0.60	0.91	0.72
Baby cry	0.90	0.61	0.73
Clock tick	0.86	0.83	0.85
Person sneeze	0.67	0.21	0.32
Helicopter	0.87	0.90	0.88
Chainsaw	0.71	0.74	0.73
Rooster	0.83	0.36	0.50
Fire crackling	0.84	0.89	0.87

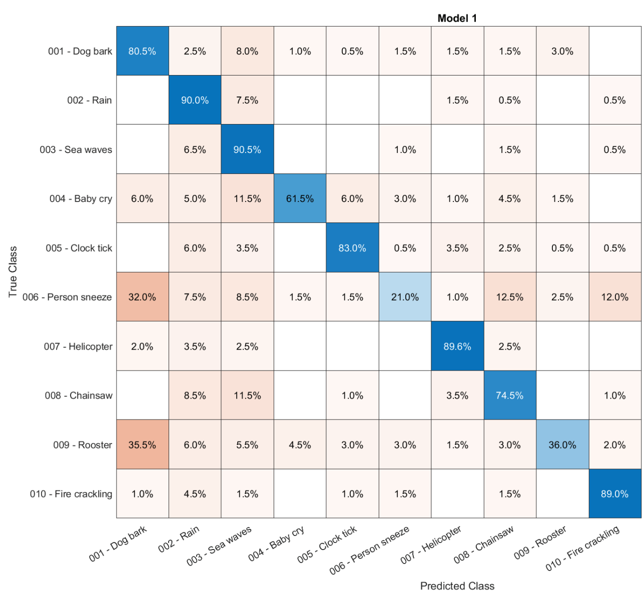


Figure 9. *The complexity matrix obtained for the MFCC-based ESC-10 dataset*

The VGGish features obtained from the ESC-10 dataset were trained with the k-NN model and the overall accuracy rate of the obtained model was obtained as 93.20%. The normalized complexity matrix of this model is given in Figure 10. The Precision, Recall, and F1-Score values of this VGGish feature-based model are also given in Table 3 for each class.

Table 3. *Statistical values for each class in VGGish based ESC-10 dataset*

Class	Precision	**Recall**	**F1-Score**
Dog	0.90	0.88	0.89
Rain	0.96	0.98	0.97
Sea waves	0.98	0.96	0.97
Baby cry	0.88	0.91	0.89
Clock tick	0.90	0.98	0.94
Person sneeze	0.79	0.71	0.75
Helicopter	0.98	0.99	0.98
Chainsaw	0.99	0.97	0.98
Rooster	0.86	0.75	0.80
Fire crackling	0.997	0.99	0.98

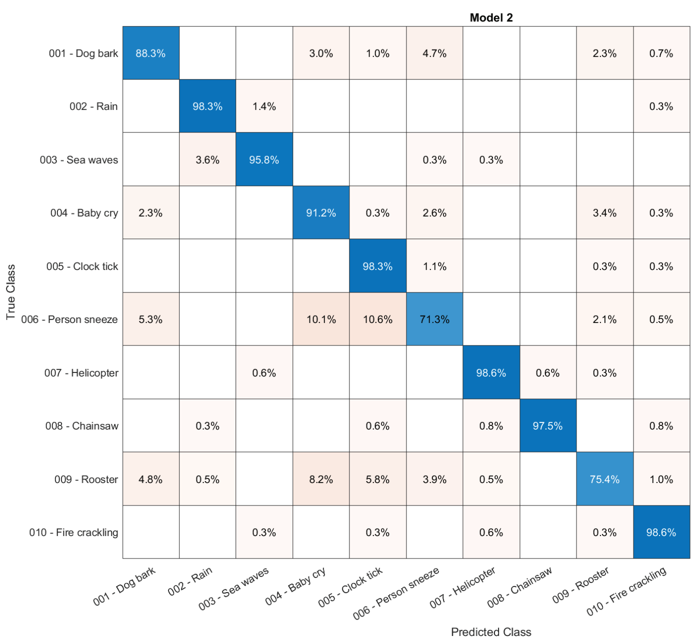


Figure 10. *The obtained complexity matrix for the VGGish-based ESC-10 dataset*

The MFCC coefficients obtained from the categorized ESC dataset were trained with the k-NN model and the overall accuracy rate of the obtained model was obtained as 68.20%. The normalized complexity matrix of this model is given in Figure 11. In addition, the Precision, Recall, and F1-Score values for each class of this MFCC feature-based model are given in Table 4.

Table 4. *Statistical values for each class in the MFCC-based categorized ESC dataset*

Class	Precision	Recall	F1-Score
Animals	0.72	0.60	0.65
Exterior	0.80	0.71	0.75
Human	0.53	0.68	0.60
Interior	0.60	0.76	0.67
Natural	0.75	0.69	0.72

Figure 11. *The complexity matrix obtained for the MFCC-based categorized ESC dataset.*

The VGGish features obtained from the categorized ESC dataset were trained with the k-NN model and the overall accuracy rate of the obtained model was obtained 87.30%. The normalized complexity matrix of this model is given in Figure 12. In addition, Precision, Recall, and F1-Score values for each class of this MFCC feature-based model are given in Table 5.

Table 5. *Statistical values for each class in the VGGish-based categorized ESC dataset*

Class	Precision	Recall	F1-Score
Animals	0.90	0.73	0.81
Exterior	0.93	0.97	0.95
Human	0.80	0.87	0.83
Interior	0.80	0.89	0.84
Natural	0.94	0.95	0.94

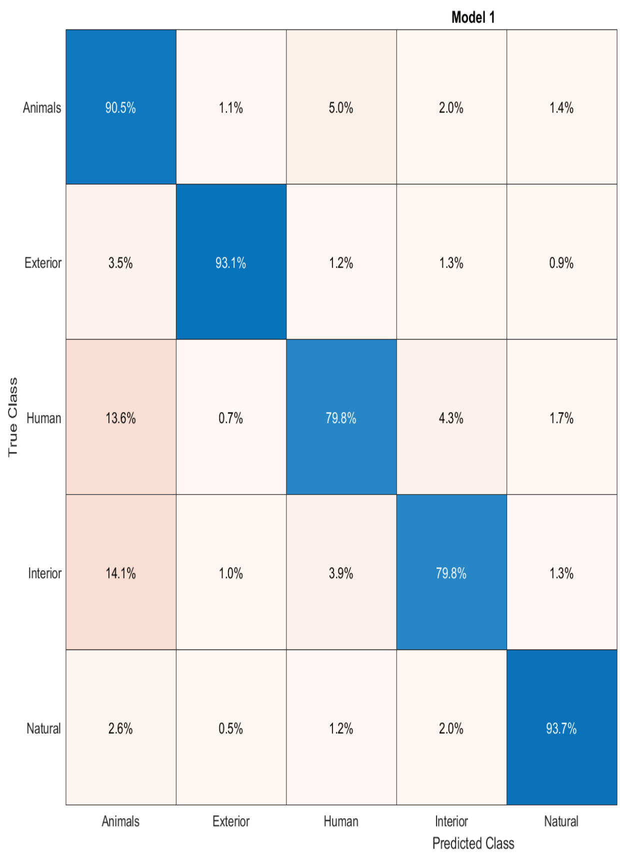


Figure 12. *The complexity matrix obtained for the VGGish-based categorized ESC dataset.*

With VGGish embeddings, a 21.60% better classification accuracy was achieved on the ESC-10 dataset and 19.10% better classification accuracy on the categorized ESC dataset. In environmental sound classification, VGGish embeddings outperforms MFCC feature extraction.

References

Bansal, R., Shukla, N., Goyal, M., & Kumar, D. (2021). Enhancement and Comparative Analysis of Environmental Sound Classification Using MFCC and Empirical Mode Decomposition. In T. Senjyu, P. N. Mahalle, T. Perumal, & A. Joshi (Eds.), *Smart Innovation, Systems and Technologies* (Vol. 195, pp. 227–235). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-7078-0_21

Bhatia, N. (2010). Survey of nearest neighbor techniques. *ArXiv Preprint ArXiv:1007.0085*.

Boddapati, V., Petef, A., Rasmusson, J., & Lundberg, L. (2017). Classifying environmental sounds using image recognition networks. *Procedia Computer Science*, *112*, 2048–2056. https://doi.org/10.1016/j.procs.2017.08.250

Chu, S., Narayanan, S., & Kuo, C. C. J. (2009). Environmental sound recognition with timeFrequency audio features. *IEEE Transactions on Audio, Speech and Language Processing*, *17*(6), 1142–1158. https://doi.org/10.1109/TASL.2009.2017438

Dogan, S., Akbal, E., & Tuncer, T. (2020). A novel ternary and signum kernelled linear hexadecimal pattern and hybrid feature selection based environmental sound classification method. *Measurement: Journal of the International Measurement Confederation*, *166*. https://doi.org/10.1016/j.measurement.2020.108151

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., … Wilson, K. (2017). CNN architectures for large-scale audio classification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (pp. 131–135). https://doi.org/10.1109/ICASSP.2017.7952132

Jia-Ching Wang, Jhing-Fa Wang, Kuok Wai He, & Cheng-Shu Hsu. (2006). Environmental Sound Classification using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings* (pp. 1731–1735). IEEE. https://doi.org/10.1109/IJCNN.2006.246644

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *ArXiv Preprint ArXiv:1003.4083*. Retrieved from http://arxiv.org/abs/1003.4083

Piczak, K. J. (2015a). Environmental sound classification with convolutional neural networks. In *IEEE International Workshop on Machine Learning for Signal Processing, MLSP* (Vol. 2015-Novem, pp. 1–6). https://doi.org/10.1109/MLSP.2015.7324337

Piczak, K. J. (2015b). ESC: Dataset for environmental sound classification. In *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference*. https://doi.org/10.1145/2733373.2806390

Salamon, J., & Bello, J. P. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, *24*(3), 279–283. https://doi.org/10.1109/LSP.2017.2657381

Sutton, O. (2012). Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction. *Introduction to k Nearest Neighbour Classification*, 1–10.

Tokozume, Y., & Harada, T. (2017). Learning environmental sounds with end-to-end convolutional neural network. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (pp. 2721–2725). https://doi.org/10.1109/ICASSP.2017.7952651

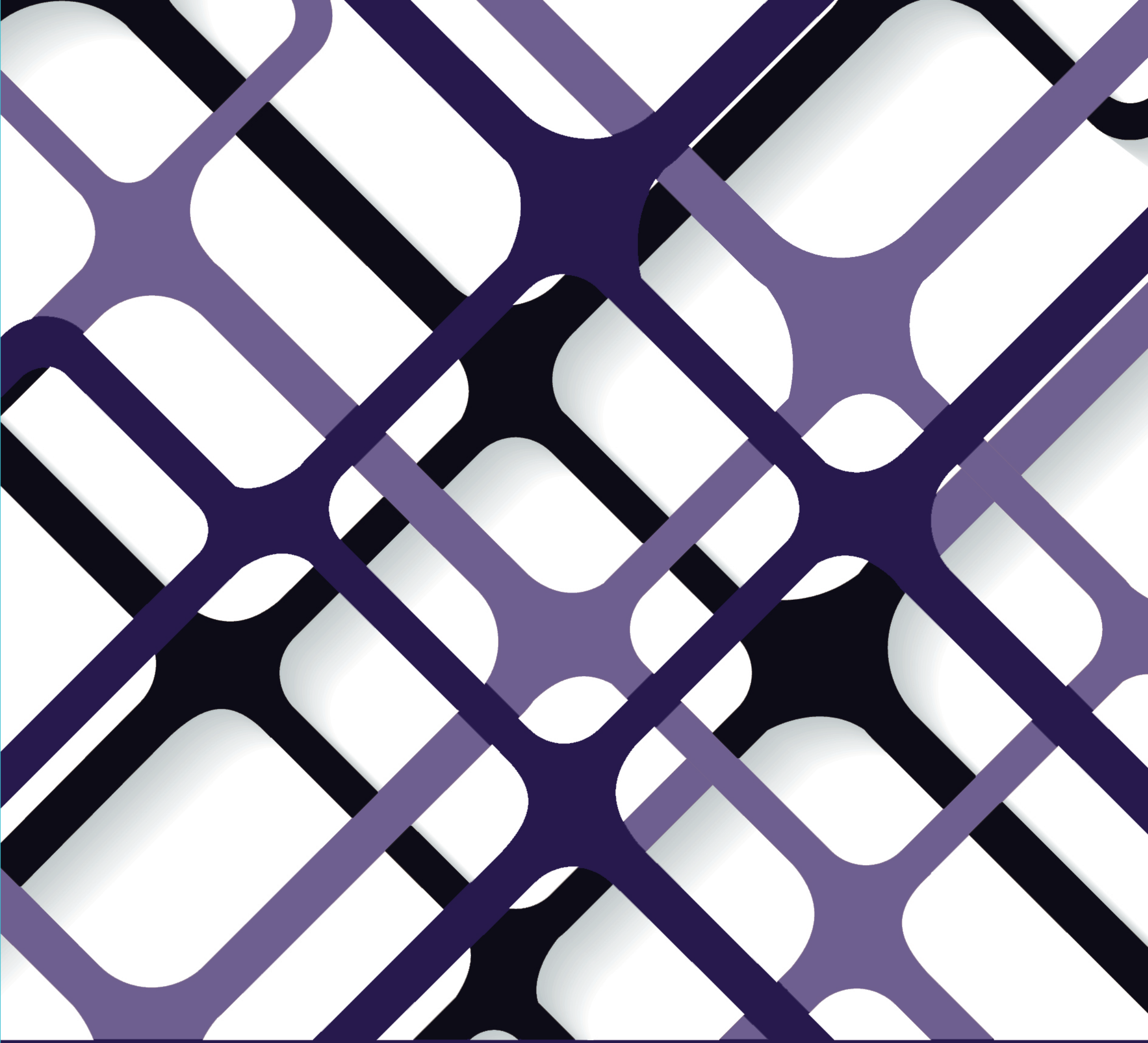
Tombaloğlu, B., & Erdem, H. (2016). Development of a MFCC-SVM Based Turkish Speech Recognition system. *2016 24th Signal Processing and Communication Application Conference, SIU 2016 - Proceedings*, 929–932. https://doi.org/10.1109/SIU.2016.7495893

Uzkent, B., Barkana, B. D., & Cevikalp, H. (2012). Non-speech environmental sound classification using SVMs with a new set of features. *International Journal of Innovative Computing, Information and Control*, *8*(5 B).

Wasule, V., & Sonar, P. (2017). Classification of brain MRI using SVM and KNN classifier. In *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)* (pp. 218–223). https://doi.org/10.1109/SSPS.2017.8071594

Zhang, X., Zou, Y., & Shi, W. (2017). Dilated convolution neural network with LeakyReLU for environmental sound classification. In *International Conference on Digital Signal Processing, DSP* (Vol. 2017-Augus). https://doi.org/10.1109/ICDSP.2017.8096153

Zhang, Z., Xu, S., Zhang, S., Qiao, T., & Cao, S. (2019). Learning Attentive Representations for Environmental Sound Classification. *IEEE Access*, *7*, 130327–130339. https://doi.org/10.1109/ACCESS.2019.2939495



RESEARCH & REVIEWS IN ENGINEERING - II

SEPTEMBER/2021

EDITORS
PROF. DR. BANU NERGİS
ASSOC. PROF. DR. SELAHATTİN BARDAK
ASSOC. PROF. DR. MAHMUT KAYAR
DR. ARİF FURKAN MENDİ

gece kitaplığı

İmtiyaz Sahibi / Publisher • Yaşar Hız
Genel Yayın Yönetmeni / Editor in Chief • Eda Altunel
Kapak & İç T asarım / Cover & Interior Design • Gece Kitaplığı
Editörler / Editors • Prof. Dr. Banu NERGİS
Assoc. Prof. Dr. Selahattin BARDAK
Assoc. Prof. Dr. Mahmut KAYAR
Dr. Arif Furkan MENDİ
Birinci Basım / First Edition • © Eylül 2021
ISBN • 978-625-8002-45-4

Gece Kitaplığı / Gece Publishing
Türkiye Adres / Turkey Address: Kızılay Mah. Fevzi Çakmak 1. Sokak Ümit Apt. No: 22/A Çankaya / Ankara / TR
Telefon / Phone: +90 312 384 80 40
web: www.gecekitapligi.com
e-mail: gecekitapligi@gmail.com

Baskı & Cilt / Printing & Volume
Sertifika / Certificate No: 47083

RESEARCH & REVIEWS IN ENGINEERING - II

September / 2021

EDITORS
PROF. DR. BANU NERGİS
ASSOC. PROF. DR. SELAHATTİN BARDAK
ASSOC. PROF. DR. MAHMUT KAYAR
DR. ARİF FURKAN MENDİ

gece kitaplığı

CONTENTS