Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

GU J Sci, Part C, 13(3): 849-858 (2025)

Using of Deep Learning Models In Acoustic Scene Classification

Zehra BOZDAĞ KARAKEÇİ[1*] Harun ÇİĞ[2]

[1]*Harran University, Faculty of Engineering, Department of Software Engineering, Şanlıurfa, Turkey*

[2]*Harran University, Faculty of Engineering, Department of Software Engineering, Şanlıurfa, Turkey*

Article Info

Research article
Received: 15/11/2024
Revision: 14/03/2025
Accepted: 04/06/2025

Keywords

Signal processing
Deep learning
Acoustic scene classification
Audio processing
Model performance

Makale Bilgisi

Araştırma makalesi
Başvuru: 15/11/2024
Düzeltme: 14/03/2025
Kabul: 04/06/2025

Anahtar Kelimeler

Sinyal işleme
Derin öğrenme
Akustik sahne sınıflandırması
Ses işleme
Model performansı

Graphical/Tabular Abstract (Grafik Özet)

This study shows that a simple CNN outperforms more complex models in classifying stage sounds using mel-spectrograms, achieving 59% accuracy on the TAU Acoustic Scene 2023 dataset, emphasizing the efficiency and effectiveness of lightweight models in ambient sound analysis. / Bu çalışma, mel-spektrogramlar kullanarak sahne seslerini sınıflandırmada basit bir CNN'in daha karmaşık modellerden üstün olduğunu göstermektedir. TAU Acoustic Scene 2023 veri setinde %59 doğruluk elde etmiş olup, hafif modellerin çevresel ses analizinde hem verimli hem de etkili olduğunu vurgulamaktadır.

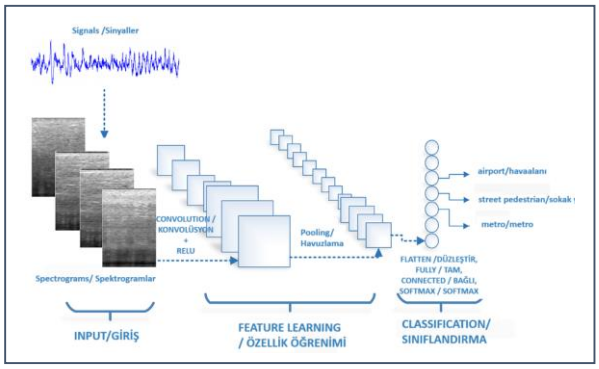

***Figure A**: Template of the Study /**Şekil A:** Çalışmanın Şablonu*

Highlights (Önemli noktalar)

- *Ambient sound analysis provides environmental context through surrounding audio. / Ortam sesi analizi, çevreleyen ses aracılığıyla çevresel bağlam sağlar.*
- *Deep learning methods are increasingly applied to this field, surpassing traditional techniques. / Derin öğrenme yöntemleri bu alanda giderek daha fazla uygulanmakta ve geleneksel tekniklerin önüne geçmektedir.*
- *Mel-spectrograms from the TAU Acoustic Scene 2023 dataset were used for classification. / Sınıflandırma için TAU Akustik Sahne 2023 veri setinden elde edilen Mel-spektrogramları kullanıldı.*

***Aim (Amaç):** To evaluate and compare the performance of various deep learning models for acoustic scene classification. / Akustik sahne sınıflandırması için çeşitli derin öğrenme modellerinin performansını değerlendirmek ve karşılaştırmak.*

***Originality (Özgünlük):** The study focuses on using mel-spectrogram representations of stage sounds. A comparison of simple and complex deep learning models is presented in terms of classification performance and efficiency. / Çalışma, sahne seslerinin mel-spektrogram gösterimlerinin kullanılmasına odaklanmaktadır. Basit ve karmaşık derin öğrenme modellerinin sınıflandırma performansı ve verimliliği açısından bir karşılaştırması sunulmaktadır.*

***Results (Bulgular):** A simple CNN model outperformed more complex models with a 59% accuracy rate. The CNN achieved the highest performance despite having the fewest parameters. / Basit bir CNN modeli, %59'luk bir doğruluk oranıyla daha karmaşık modellerden daha iyi performans gösterdi. CNN, en az parametreye sahip olmasına rağmen en yüksek performansı elde etti.*

***Conclusion (Sonuç):** Simpler deep learning models, like CNNs can be both effective and computationally efficient for acoustic scene classification tasks. / CNN'ler gibi daha basit derin öğrenme modelleri, akustik sahne sınıflandırma görevleri için hem etkili hem de hesaplama açısından verimli olabilir.*

*Corresponding author, e-mail: zbozdag@harran.edu.tr
DOI: 10.29109/gujsc.1585401

GU J Sci, Part C, 13(3): 849-858 (2025)

Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

Using of Deep Learning Models In Acoustic Scene Classification

Zehra BOZDAĞ KARAKEÇİ[2*] Harun ÇİĞ[2]

[1]*Harran University, Faculty of Engineering, Department of Software Engineering, Şanlıurfa, Turkey*

[2]*Harran University, Faculty of Engineering, Department of Software Engineering, Şanlıurfa, Turkey*

Article Info

Research article
Received: 15/11/2024
Revision: 14/03/2025
Accepted: 04/06/2025

Keywords

Signal processing
Deep learning
Acoustic scene classification
Audio processing
Model performance

Abstract

Ambient sound analysis has become more prominent with the rise of portable and wearable devices. It provides valuable insights into a person's environment by analyzing surrounding sounds. Recently, deep learning methods, frequently used in image and text processing, have been applied to this field and are proving more effective than traditional machine learning techniques.

In this study, we evaluated the performance of different deep learning models using mel-spectrograms of 3 classes of stage sounds based on TAU Acoustic Scene 2023 dataset. Our results indicate that a simple Convolutional Neural Network (CNN) model gives better classification results compared to other more complex models in classification tasks. Despite having the fewest parameters, the CNN model achieved the highest success with 59% accuracy. This suggests that simpler models can be highly effective for acoustic scene classification, highlighting the value of more efficient and computationally feasible approaches in this domain.

Akustik Sahne Sınıflandırmasında Derin Öğrenme Modellerinin Kullanımı

Makale Bilgisi

Araştırma makalesi
Başvuru: 15/11/2024
Düzeltme: 14/03/2025
Kabul: 04/06/2025

Anahtar Kelimeler

Sinyal işleme
Derin öğrenme
Akustik sahne sınıflandırması
Ses işleme
Model performansı

Öz

Taşınabilir ve giyilebilir cihazların yükselişiyle ortam sesi analizi daha da belirgin hale geldi. Çevresel sesleri analiz ederek bir kişinin ortamına dair değerli içgörüler sağlar. Son zamanlarda, görüntü ve metin işlemede sıklıkla kullanılan derin öğrenme yöntemleri bu alana uygulandı ve geleneksel makine öğrenme tekniklerinden daha etkili olduğu kanıtlandı.

Bu çalışmada, TAU Akustik Sahne 2023 veri setine dayalı 3 sınıf sahne sesinin mel-spektrogramlarını kullanarak farklı derin öğrenme modellerinin performansını değerlendirdik. Sonuçlarımız, basit bir Evrişimli Sinir Ağı (ESA) modelinin sınıflandırma görevlerinde diğer daha karmaşık modellere kıyasla daha iyi sınıflandırma sonuçları verdiğini göstermektedir. En az parametreye sahip olmasına rağmen, ESA modeli %59 doğrulukla en yüksek başarıyı elde etti. Bu, daha basit modellerin akustik sahne sınıflandırması için oldukça etkili olabileceğini ve bu alanda daha verimli ve hesaplama açısından uygulanabilir yaklaşımların değerini vurgulamaktadır.

1. INTRODUCTION (GİRİŞ)

Voice analysis covers a variety of voice-related problems such as automatic speech recognition [1], speech-to-text translation [2], automatic audio captioning [3], speech emotion recognition [4] [5], voice-based question-and-answer generation [6], vocal sound classification, and music score analysis [7,8].

With the advancing technology and the increase in the amount of data accessed, the interest in voice analysis has increased day by day. Table 1 presents the total number of publications in voice analysis by years [9], illustrating the distribution across different periods. The data suggests a projected increase in research activity in this field in the coming years.

Acoustic scene classification (ASC) is one of the problems in the field of audio analysis. ASC aims to

*Corresponding author, e-mail: zbozdag@harran.edu.tr
DOI: 10.29109/gujsc.1585401

classify digital audio signals into mutually exclusive scene categories, which can be, for example, an indoor environment (such as a house) or an outdoor environment (such as a park) [10]. In everyday life, people often use the ability to perceive and react to the scene they are in by listening to the sounds around them. Technological devices utilize ASC processing in various fields and applications. For instance, wearable smartwatches can classify the user's environment and adjust notifications accordingly. Similarly, smartphones and other devices equipped with digital sensors detect and classify environmental sounds. By definition, ASC seeks to identify the type of scene in which a given audio signal is recorded.

To utilize an acoustic audio signal for deep learning, the signal typically needs to be transformed into an appropriate format. Common methods for transforming audio signals include spectrograms, mel-frequency cepstral coefficients, linear prediction coding, and wavelet decomposition [11–14]. The resulting feature matrices or images are then used as input data for deep learning algorithms. In audio signal processing studies, the choice of transformation methods and the applied deep learning models significantly impact processing performance [14–16].

Table 1. Total number of audio analyzed publications by year (Yıllara göre ses analizli yayınların toplam sayısı)

	2015	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
Publications (total)	23,287	26,105	30,883	37,592	37,155	48,972	55,151	58,498	65,618

In previous ASC studies, traditional machine learning techniques such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Random Forest (RF), Decision Tree (DT) and Hidden Markov Models (HMM) have been widely used. In these traditional approaches, classification usually consists of two stages: first, the features of the audio signal are extracted and then these features are classified by traditional machine learning methods [17–19]. Recently, deep learning methods have been successfully applied to ASC, yielding excellent results with models such as Time Delay Neural Networks, Bidirectional Long Short-Term Memory, Feed Forward Neural Networks, EfficientNet-based architectures (e.g., VGG, ResNet, Inception, Xception, and MobileNetV2), Fully Convolutional Neural Networks (FCNN), and Convolutional Recurrent Neural Networks (CRNN) [20–23].

CNNs reveal complex spatial and temporal dependencies more successfully than traditional methods. Acoustic Scene Classification (ASC) is the problem of characterizing the complex patterns and contextual information of sound. One of the methods to address this problem is to obtain and process spectrograms of signals. The Mel-spectrogram feature visually represents how the frequency content of sounds changes over time. For example, when an audio signal contains audible sound events such as "dog barking," "speech," or "applause," these events can be identified in the corresponding Mel-spectrogram feature due to their unique visual patterns. From this perspective, the Mel-spectrogram feature can be considered as an image [24].

For the last 20 years, CNNs have been the most widely used and highest-performing method in image processing fields such as classification, object detection, and segmentation [25–27]. The highest success rates are achieved with CNNs in studies conducted on Mel-spectrogram images [28–30]. However, it is still not fully understood what features CNNs learn, whether they generalize well across different datasets or overfit to a specific dataset, and how they encode information [31] [32].

Deep learning methods automatically extract the features required for classification from the given input by convolution operations. To successfully utilize these models, large datasets and server computers equipped with powerful graphics processing units (GPUs) are typically required. Advancements in technological devices and increasingly powerful computing systems have enabled the wider application of deep learning models. Research findings indicate that deep learning methods are being used more frequently in voice anawlysis, yielding more successful results [33]. These models are capable of learning complex patterns and recognizing subtle differences in audio data. Additionally, they learn faster and more

accurately than traditional models, making them particularly suitable for real-time audio classification and analysis.

In our study, we utilized the dataset released for the 1st task of the DCASE Challenge 2023. Mel-spectrograms were generated from the audio signals to be used in image classification models, which were then analyzed using contemporary models and a simple Convolutional Neural Network (CNN) architecture.

The contributions of our work to literature are as follows:

- We analyzed the performance of contemporary deep learning models on the dataset, with MobileNetV3-Small outperforming the other models in the benchmark.

- The simple CNN model, despite having fewer parameters, achieved the highest accuracy in the benchmark.

The remainder of this paper is organized as follows: Section 2 provides details about the dataset used in the study, while Section 3 outlines the experiments conducted using the dataset. Section 4 describes the methods applied, Section 5 presents the evaluation results, and finally, Section 6 offers the conclusions and suggestions.

2. DATASET (VERİ SETİ)

The DCASE Challenge is an annual competition for researchers focused on the computational analysis of audio events and acoustic scene analysis, providing a platform for presenting and discussing their results [34]. Each year, different tasks are defined, and datasets are released for the analysis of audio in order to accomplish these tasks. In this study, we utilized the dataset published for the first task of the competition, which involves voice classification. The general schematic of the voice classification process is presented in Figure 1.

The dataset consists of recordings from 12 European cities, captured in 10 distinct acoustic scenes (Airport, Indoor shopping mall, Metro station, Pedestrian street, Public square, Street with medium level of traffic, Traveling by a tram, Traveling by a bus, Traveling by an underground metro and Urban park) using 4 different devices. Additionally, it includes synthetic data from 11 mobile devices derived from the original recordings. The dataset is identical to the TAU Urban Acoustic Scene 2022 Mobile development dataset, with each audio file having a duration of 1 second. In our study, we utilized sound recordings from three classes: airport (0), street pedestrian (1), and metro (2). Table 2 provides the distribution of audio recordings across these classes [35].

The sound characteristics of ASC can be significantly affected by local infrastructure, population density, recording interval, cultural factors and changes in the recording device resolution [36,37]. These effects can cause the performance of the sound classification system to degrade. Adapting a model trained on specific cities and devices to other cities or devices creates a domain adaptation problem [38]. As a solution to this problem, collecting data from various data domains is an important method among the solutions to the domain adaptation problem. The dataset used in our study includes sound data from a total of 12 European cities: Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm and Vienna. Also, different device types were used for sound recordings.

Mel-spectrograms of audio signals are frequently utilized for audio classification tasks [39–42]. In fact, the finalists of the DCASE Challenge employed mel-spectrogram features in their models [43]. To generate mel-spectrogram images from the audio files, a Hamming window with a sample length of 2048, an overlap of 1024, and 128 mel bins was used. Sample mel-spectrograms for the different classes are shown in Figure 2.

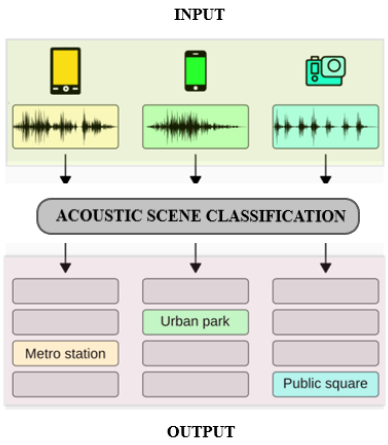

Figure 1. General schematic of the acoustic scene classification system (Akustik sahne sınıflandırma sisteminin genel şeması).

Table 2. Dataset information (Veri seti bilgisi)

Classes	Training Data	Test Data	Total
airport (0)	8000	2000	8000
pedestrian street (1)	8000	2000	8000
metro (2)	8000	2000	8000
Total	24000	6000	30000

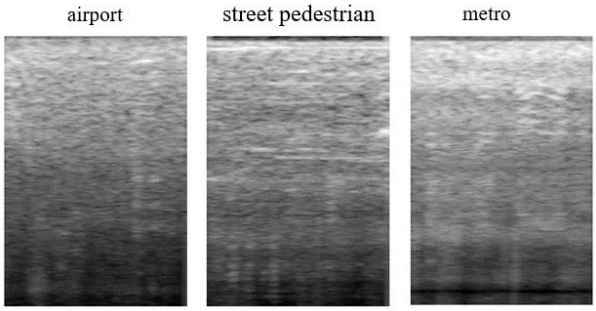

Figure 2. Examples of mel-spectrograms of sound classes (Ses sınıflarının mel-spektrogramlarına örnekler).

3. STUDIES ON THE DCASE DATASET
(DCASE VERİ SETİ ÜZERİNDEKİ ÇALIŞMALAR)

The high performance achieved through the use of deep learning in image classification problems has led to its widespread use in other problems as well [44]. In recent studies, it has been observed that deep learning techniques are predominantly applied to the ASC problem. When we examine the studies conducted over the years according to the first task description published by DCASE Challenge, deep learning models have been commonly used.

In 2017, Duppada and Hiray enhanced ASC performance by employing ensemble methods with various deep neural network models (LeNet, SqueezeNet, 1D CNN) using mel-spectrogram features [45]. In 2018, Eghbal-zadeh et al. incorporated an intraclass covariance analysis layer into the VGG model to improve performance [46]. A review of the DCASE competition by Gharib et al. in 2018 highlighted that data augmentation techniques, such as data replication, were widely used to enhance ASC performance [47]. In 2019, Wu and Lee's work on improving voice texture significantly boosted voice classification accuracy using CNN-based approaches [24]. In 2020, Zhang et al. explored the reuse of pre-trained models for different tasks [48]. Additionally, Tonami et al. demonstrated that multi-task learning, which combines the analysis of audio events and acoustic scenes, improves classification success[49]. In 2021, studies focusing on audio event detection with a curriculum learning approach achieved notable success by adopting training strategies based on the complexity of event learning [50]. Utilizing the DCASE 2017 dataset, in 2022, methods were developed that combined both scene and instantaneous sounds for event detection, while the Qwen-Audio application, introduced in 2023, provided more comprehensive solutions for universal audio understanding with its large-scale audio-language model [51,52].

The classification results for 15 classes, as announced on the official DCASE 2023 Challenge web page, ranked Schmid and his team in the top two positions, with accuracies of 62.7% and 61.4%, respectively [53]. The team developed a novel deep learning architecture featuring a regularized receptive field and residual inverted bottleneck blocks. Jiaxin et al. secured 3rd place with an accuracy of 60.8%, having designed a tutor model based on blueprint separable convolution (BSConv) [54].

These studies demonstrate that methodological diversity and innovation in the field of ASC contribute significantly to performance improvements, with variations in model designs and data processing methods playing a critical role in achieving success.

4. METHOD (METOT)

In our study, we utilized current image classification algorithms with varying parameter counts and technical architecture. ResNet, MobileNet, and EfficientNet architectures, which are frequently employed in the literature for image classification tasks, were included. Additionally, a CNN model with a simplified layer structure and a smaller number of parameters was developed. Comparisons of these models were conducted based on their parameter counts and classification accuracy.

ResNet is an architecture based on a residual network model, with variants such as ResNet101, ResNet50, and ResNet18, distinguished by the number of layers. In our study, we used the ResNet18 model, which has a depth of 18 layers and 11.7 million parameters. This variant has the fewest layers and parameters compared to the other ResNet models [55].

In 2017, Google introduced MobileNet, specifically designed for mobile and embedded applications to minimize resource requirements [56]. In this

network architecture, both the number of parameters (disk space) and computational complexity (power consumption and latency) are significantly reduced. MobileNet decreases the number of parameters by a factor of 7 compared to traditional full convolutional models, while sacrificing only 1% in accuracy. The MobileNet model achieves this efficiency by utilizing depth wise separable convolutions and pointwise convolution techniques.

In 2018, MobileNet-v2 was introduced as an evolution of its predecessor, maintaining accuracy levels while significantly reducing memory usage. This improvement is achieved through an inverted residual structure and shortcut connections between linear bottlenecks [57]. MobileNet-v2 processes 224x224x3 images, has a depth of 53 layers, and contains 3.5 million parameters. In 2019, the third generation of MobileNets, MobileNet-v3, was developed using a combination of Network Architecture Search (NAS) and the NetAdapt algorithm. This led to the creation of two models, MobileNet-v3 Large and MobileNet-v3 Small, designed to accommodate different resource constraints [58].

EfficientNets are a family of models developed by Google AI researchers, designed to scale network dimensions at a constant rate. EfficientNet employs a novel approach that scales the network's depth, width, and resolution uniformly with a fixed ratio. The EfficientNet-B0 model, for example, has a depth of 82 layers, processes input images at a resolution of 224x224x3, and contains 5.3 million parameters [59]. Compared to earlier models such as ResNet, DenseNet, and Inception, EfficientNet models are significantly more computationally efficient.

The simple CNN model consists of two blocks followed by a final output layer. Each block contains a Conv2d (3x3) layer, a ReLU activation function, and a Dropout layer. The Conv2d (3x3) layer is a 2D convolutional layer, ReLU introduces non-linearity, and Dropout is used to prevent overfitting by randomly omitting some neurons during training. Table 3 outlines the architecture of the developed simple CNN model. The MaxPool2d function reduces the size of the feature maps, while the Flatten and Linear layers prepare the feature maps for output. The implementation of the model was done using the PyTorch library in the Python programming language.

PyTorch and torchaudio libraries were used to process and classify audio signals. The audio signals were downsampled from 16 kHz to 8 kHz by applying a resampling process. After resampling, spectrogram and mel spectrogram transformations were performed. Data augmentation techniques were also applied to improve the model's overall accuracy. These techniques include TimeStretch, Frequency Mask, and Time Mask.

The mel spectrograms obtained from the processed audio data were resized to 224x224 pixels, and their values were normalized between 0 and 1 using min-max normalization. These processes enable the audio data to be converted into a format that can be effectively used in deep learning models.

All network models were trained for 10 epochs, with 80% of the dataset allocated for training and 20% for testing. Cross-Entropy was employed as the loss function, and the Adam optimizer was used as the optimization algorithm [60]. Adam is a variant of the stochastic gradient descent algorithm. A learning rate of 0.001 and a weight decay of 0.01 were applied during the training process.

Table 3. Architecture of the simple CNN model
(Basit CNN modelinin mimarisi)

Model Layers
Block
Block
MaxPool2d
Block
Block
MaxPool2d
Flatten
Linear

Evaluation metrics (Değerlendirme ölçütleri): The accuracy calculation formula used to compare the performance of the models is given in Equation (1). TP represents the number of True Positives, TN represents the number of True Negatives, FP represents the number of False Positives, and FN represents the number of False Negatives [61].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

The formula for calculating Precision is provided in Equation 2. Precision measures the proportion of correct predictions made by the model out of all positive predictions.

Recall, given in Equation 3, measures how many of the actual positive instances the model correctly identifies.

The F1-Score, calculated using Equation 4, represents the harmonic mean of Precision and Recall, providing a balanced measure of a model's accuracy.

5. METHOD RESULTS (YÖNTEM SONUÇLARI)

In this study, we utilized several state-of-the-art image classification network models with different parameter counts and architectures (ResNet18, MobileNetV3-Small, EfficientNet-B0, and MobileNetV2) to classify mel-spectrogram images of audio files. Additionally, the developed simple CNN model was employed for classification. Table 4 presents the Accuracy, Recall, Precision, and F1-Score of the models on the test dataset. As shown in the table, the simple CNN architecture achieves higher accuracy compared to the other models. Given that the audio files in the dataset are only 1 second in duration, they lack many of the distinguishing features necessary for effective class differentiation. As a result, the models generally show lower accuracy rates. However, as seen in Table 4, the simple CNN model demonstrates the highest accuracy.

Table 5 presents the number of parameters and accuracy values for the models. The simple CNN model, which has the lowest number of parameters, achieves the highest performance. Similarly, the MobileNetV3-Small model, which also has a relatively small number of parameters, ranks second in terms of accuracy.

Table 4. Evaluation results of the models (Modellerin değerlendirme sonuçları)

Model	Data Replication	Pre-trained	Number of Epochs	Accuracy	F1-Score	Recall	Precision
Simple CNN	√		10	0.5908	0.5490	0.6007	0.5587
MobileNetV3-Small	√	√	10	0.5633	0.5333	0.5633	0.5368
MobileNetV2	√	√	10	0.5068	0.4901	0.5021	0.5045
EfficientNet-B0	√	√	10	0.5407	0.5126	0.5431	0.5303
ResNet-18	√	√	10	0.4455	0.3813	0.4778	0.3805

Table 5. Parameter numbers of the models (Modellerin parametre sayıları)

Model	Parameter Number	Accuracy
Simple CNN	**157 939**	**0.5908**
MobileNetV3-Small	2 542 856	0.5633
MobileNetV2	3 504 872	0.5068
EfficientNet-B0	5 288 548	0.5413
ResNet-18	11 689 512	0.4455

Among the models selected for comparison, MobileNetV3-Small, which has the lowest number of parameters, is optimized for devices with lower computational requirements. It uses Squeeze-and-Excitation (SE) blocks and the hardswish activation function. Similarly, MobileNetV2 provides efficient computation with Depthwise Separable Convolution and offers low latency on mobile devices using Inverted Residual Blocks.

EfficientNet-B0 is a balanced model in terms of both the number of parameters and accuracy, optimized through Neural Architecture Search (NAS). ResNet18 improves the trainability of deeper networks by using residual connections (skip connections). It is a convolutional neural network model that is not very deep but still powerful.

The reason why the developed Simple CNN model achieves higher accuracy compared to advanced deep learning architectures can be attributed to several fundamental factors. These include dataset size and the compatibility of feature distribution. While highly optimized models (such as EfficientNet and ResNet) use techniques like parameter reduction, quantization, and narrow filters to minimize computational costs, these optimizations can sometimes lead to a loss of important visual information. Furthermore, pre-trained models, which are usually trained on large and diverse datasets (such as ImageNet), may have limited generalization ability when applied to specific datasets. In contrast, the Simple CNN model generalized better without overfitting because it was trained on a smaller dataset that was more aligned with the target task than those used for pre-trained models.

6. CONCLUSIONS AND SUGGESTIONS
(SONUÇLAR VE ÖNERİLER)

In our study, acoustic scene sounds from the TAU Urban Acoustic Scene 2023 Mobile development dataset were classified using state-of-the-art deep learning models and simple CNN architecture. The analysis reveals that among the deep networks tested, the MobileNetV3-Small model, despite having fewer parameters, outperforms other models with an accuracy rate of 56.33%. In contrast, the ResNet18 model, which has the most parameters, exhibited the lowest performance. Interestingly, the simple CNN model, designed with only 157,939 parameters, achieved the highest accuracy of 59.08%, surpassing deeper and more complex networks.

These findings indicate that highly complex and deep architectures are not necessarily required for effective acoustic scene classification, and that compact models with fewer layers and parameters can perform competitively.

Furthermore, it was observed that mel-spectrogram images tend to lose discriminative features as they pass through deeper convolutional layers in complex networks. This may explain the superior performance of simpler models, which better preserve these features throughout the learning process.

Another factor influencing model performance is the nature of the dataset itself — since it contains 1-second audio clips, many distinctive scene-specific features may not be fully captured. This inherent limitation makes classification more challenging and is reflected in the observed accuracy rates.

The findings of this study have several real-world applications, especially in areas where efficient and accurate acoustic scene classification is important, such as wearable devices, smart home systems, and autonomous vehicles. Moreover, simpler models are well-suited for resource-constrained environments, such as low-power devices. These models can maintain competitive performance while reducing computational load and energy consumption, making them ideal for deployment in embedded systems and mobile platforms. For example, previous studies have shown that low-complexity models can effectively perform acoustic scene classification in resource-constrained environments [62].

In the future, additional features can be incorporated into the audio analysis process to enhance model performance. For instance, alongside mel-spectrograms, feature extraction methods commonly used in literature such as wavelet transforms and scalograms—could be explored. These methods can provide a more detailed analysis of the time-frequency characteristics of signals and help strengthen the discriminative features of the classes. By integrating such techniques, the accuracy of the models can be improved through the enhanced representation of the distinguishing characteristics of each class.

DECLARATION OF ETHICAL STANDARDS
(ETİK STANDARTLARIN BEYANI)

The author of this article declares that the materials and methods they use in their work do not require ethical committee approval and/or legal-specific permission.

Bu makalenin yazarı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

AUTHORS' CONTRIBUTIONS (YAZARLARIN KATKILARI)

Zehra BOZDAĞ KARAKEÇİ: She conducted the experiments, analyzed the results and performed the writing process.

Deneyleri yapmış, sonuçlarını analiz etmiş ve makalenin yazım işlemini gerçekleştirmiştir.

Harun ÇİĞ: He conducted the experiments and analyzed the results.

Deneyleri yapmış ve sonuçlarını analiz etmiştir.

CONFLICT OF INTEREST (ÇIKAR ÇATIŞMASI)

There is no conflict of interest in this study.

Bu çalışmada herhangi bir çıkar çatışması yoktur.

REFERENCES (KAYNAKLAR)

[1] Kriman S, Beliaev S, Ginsburg B, Huang J, Kuchaiev O, Lavrukhin V, Leary R, Li J, Zhang Y, Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 2020; arXiv:1910.10261.

[2] Shivakumar KM, Aravind KG, Anoop TV, Gupta D. Kannada speech to text conversion using CMU Sphinx. Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016; 3-1:6.

[3] Mathur A, Saxena T, Krishnamurthi R. Generating subtitles automatically using audio extraction and speech recognition. Proceedings of the 2015 IEEE International Conference on Computational Intelligence and Communication Technology (CICT). 2015; 621–626.

[4] Sakurai M, Kosaka T. Emotion recognition combining acoustic and linguistic features based on speech recognition results. Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics. 2021; 824-827.

[5] Yağcı M, Aygül ME. Derin öğrenme tabanlı gerçek zamanlı vücut hareketlerinden duygu analizi modeli. Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji. 2022; 12: 664–674.

[6] Fayek HM, Johnson J. Temporal reasoning via audio question answering. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020; 1-1.

[7] Ewert S, Müller M. Estimating note intensities in music recordings. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing. 2011; 385-388.

[8] Türkmen MC, Ergin AA. Music note detection using matrix pencil method. In: Proceedings of the 31st IEEE Conference on Signal Processing Communications Applications, SIU 2023, 2023; 1-4.

[9] Audio Analyzing in Publications - Dimensions, (n.d.). https://app.dimensions.ai/discover/publication?search_mode=content&search_text=Audio Analyzing&search_type=kws&search_field=full_search (accessed September 3, 2024).

[10] Barchiesi D, Giannoulis DD, Stowell D, Plumbley MD. Acoustic scene classification: Classifying environments from the sounds they produce. IEEE Signal Processing Magazine. 2015; 32-16:34.

[11] Zaman K, Sah M, Direkoglu C, Unoki M. A survey of audio classification using deep learning. IEEE Access. 2023; 11-1:1.

[12] Kumari RSS, Sugumar D, Sadasivam V. Audio signal classification based on optimal wavelet and support vector machine. In: Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA). 2007; 2-544:548.

[13] Mahanta SK, Basisth NJ, Halder E, Khilji AFUR, Pakray P. Exploiting cepstral coefficients and CNN for efficient musical instrument classification. Evolving Systems. 2024; 15(3): 1–13.

[14] Chu HC, Zhang YL, Chiang HC. A CNN sound classification mechanism using data augmentation. Sensors. 2023; 23(15): 6972.

[15] Dong S, Xia Z, Pan X, Yu T. Environmental sound classification based on improved compact bilinear attention network. Digital Signal Processing. 2023; 141.

[16] Mu W, Yin B, Huang X, Xu J, Du Z. Environmental sound classification using temporal-frequency attention based convolutional neural network. Scientific Reports. 2021; 11(1): 21552.

[17] Piczak KJ. ESC: Dataset for environmental sound classification. In: Proceedings of the 2015 ACM Multimedia Conference (MM). 2015; 1015–1018.

[18] Aytar Y, Vondrick C, Torralba A. SoundNet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems. 2016; 29: 892–900.

[19] Baelde M, Biernacki C, Greff R. A mixture model-based real-time audio sources classification method. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing. 2017; 371–375.

[20] Guzhov A, Raue F, Hees J, Dengel A. Esresnet: Environmental sound classification based on visual domain models. In: Proceedings

of the International Conference on Pattern Recognition. 2020; 3504–3511.
[21] Jin G, Zhai J, Wei J. CAA-Net: End-to-End Two-Branch Feature Attention Network for Single Image Dehazing. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. 2022; E106-A(1): 1–9.
[22] Spoorthy V, Mulimani M, Koolagudi SG. Acoustic scene classification using deep learning architectures. In: Proceedings of the 6th International Conference for Convergence in Technology (I2CT). 2021; 1–6.
[23] Yang L, Tao L, Chen X, Gu X. Multi-scale semantic feature fusion and data augmentation for acoustic scene classification. Applied Acoustics. 2020; 163: 107238.
[24] Wu Y, Lee T. Enhancing sound texture in CNN-based acoustic scene classification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019; 815–819.
[25] Bozdağ Karakeçi Z, Talu MF. Multi-scale residual segmentation network for histopathological image. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi. 2024; 15(3): 623–632.
[26] Spanhol FA, Oliveira LS, Petitjean C, Heutte L. Breast cancer histopathological image classification using Convolutional Neural Networks. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). 2016; 2560–2567.
[27] Zhao Z-Q, Zheng P, Xu S-T, Wu X. Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems. 2019; 30(11): 3212–3232.
[28] Tchinda BS, Tchiotsop D, Djoufack Nkengfack LC, Tchinda R. Diagnosis of epileptic seizures from electroencephalogram signals using log-Mel spectrogram and a deep learning CNN model. Heliyon. 2025;11(2): e42993
[29] Sharan RV, Mascolo C, Schuller BW. Emotion recognition from speech signals by Mel-spectrogram and a CNN-RNN. In: Proceedings of the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2024; 1–4.
[30] Radhakrishnan BL, Kirubakaran E, Jebadurai IJ, Selvakumar IA, Andrew J. A CNN-based deep learning model for the in-home sleep stage detection system using Mel-spectrogram. In: Thomas KV, editor. Manipal Interdisciplinary Health Science and Technical Reports-2023. 2024; 98–103.
[31] Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digital Signal Processing: A Review Journal. 2018; 73: 1–15.
[32] Dosilović FK, Brčić M, Hlupić N. Explainable artificial intelligence: A survey. In: Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2018; 210–215.
[33] Ding B, Zhang T, Wang C, Liu G, Liang J, Hu R, Wu Y, Guo D. Acoustic scene classification: A comprehensive survey. Expert Systems with Applications. 2024; 238:121902.
[34] DCASE2023 Challenge - DCASE, (n.d.). https://dcase.community/challenge2023/index (accessed September 3, 2024).
[35] Heittola T, Mesaros A, Virtanen T. Acoustic scene classification in DCASE 2020 Challenge: generalization across devices and low complexity solutions. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop. 2020; 1–5.
[36] Mesaros A, Heittola T, Virtanen T. A multi-device dataset for urban acoustic scene classification. arXiv preprint arXiv:1807.09840. 2018.
[37] Bear HL, Heittola T, Mesaros A, Benetos E, Virtanen T. City classification from multiple real-world sound scenes. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). 2019; 1–5.
[38] Wang M, Deng W. Deep visual domain adaptation: A survey. Neurocomputing. 2018; 312: 135–153.
[39] Abdoli S, Cardinal P, Koerich AL. End-to-end environmental sound classification using a 1D convolutional neural network. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019; 255–259.
[40] Seo S, Kim C, Kim JH. Convolutional neural networks using log Mel-spectrogram separation for audio event classification with unknown devices. Journal of Web Engineering. 2022; 21(5): 1115–1133.
[41] Zhang T, Feng G, Liang J, An T. Acoustic scene classification based on Mel spectrogram decomposition and model merging. Applied Acoustics. 2021; 178: 107956.
[42] Nguyen MT, Lin WW, Huang JH. Heart sound classification using deep learning techniques based on log-mel spectrogram. Circuits, Systems, and Signal Processing. 2023; 42: 1039–1058.

[43] Abeßer J. A review of deep learning-based methods for acoustic scene classification. Applied Sciences (Switzerland). 2020; 10(15): 5251.

[44] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM. 2017; 60(6): 84–90.

[45] Duppada V, Hiray S. Ensemble of deep neural networks for acoustic scene classification. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop. 2017; 1–5.

[46] Eghbal-Zadeh H, Dorfer M, Widmer G. Deep within-class covariance analysis for robust audio representation learning. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop. 2017; 1–5.

[47] Gharib S, Derrar H, Niizumi D, Senttula T, Tommola J, Heittola T, Virtanen T, Huttunen H. Acoustic scene classification: A competition review. In: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP). 2018; 1–6.

[48] Zhang R, Zou W, Li X. Cross-task pre-training for on-device acoustic scene classification. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop. 2019; 1–5.

[49] Tonami N, Imoto K, Yamanishi R, Yamashita Y. Joint analysis of sound events and acoustic scenes using multitask learning. IEICE Transactions on Information and Systems. 2021; E104-D (9): 1017–1025.

[50] Tonami N, Imoto K, Okamoto Y, Fukumori T, Yamashita Y. Sound event detection based on curriculum learning considering learning difficulty of events. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021; 841–845.

[51] Tonami N, Imoto K, Nagase R, Okamoto Y, Fukumori T, Yamashita Y. Sound event detection guided by semantic contexts of scenes. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022; 851–855.

[52] Chu Y, Xu J, Zhou X, Yang Q, Zhang S, Yan Z, Zhou C, Zhou J. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. arXiv preprint arXiv:2307.04765. 2023.

[53] Schmid F, Morocutti T, Masoudian S, Koutini K, Widmer G. CP-JKU submission to DCASE23: Efficient acoustic scene classification with CP-Mobile. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop. 2023; 1–5.

[54] Tan J, Li Y. Low-complexity acoustic scene classification using blueprint separable convolution and knowledge distillation. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop. 2023; 1–5.

[55] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 770–778.

[56] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. 2017.

[57] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018; 4510–4520.

[58] Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV, Adam H. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019; 1314–1324.

[59] Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). 2019; 97: 6105–6114.

[60] Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR). 2015.

[61] Baratloo A, Hosseini M, Negida A, El Ashal G. Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. Emergency (Tehran, Iran). 2015; 3(2): 48–49.

[62] Koutini K, Henkel F, Eghbal-zadeh H, Widmer G. Low-complexity models for acoustic scene classification based on receptive field regularization and frequency damping. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop. 2020; 1–5.