Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 2 (No. 1), pp. 20-28, 2022
doi: 10.54569/aair.1017801
Published online: Feb, 16, 2022

Classification of Environmental Sounds With Deep Learning

Bekir Aksoy [1,*], Uygar Usta [1], Gürkan Karadağ [1], Ali Rıza Kaya [1], Melek Ömür [1]

[1] Isparta University of Applied Sciences, Department of Mechatronics Engineering

Abstract

Today, with the development of technology, environmental destruction is increasing day by day. For this reason, it is inevitable to take different measures to prevent the damage caused by environmental destruction. It is possible to prevent environmental damage by identifying the sounds that harm the environment and transferring them to the relevant units. In the study carried out, a data set of saw, rain, lightning, bark and broom sound data obtained from open access websites was created. Rain, barking and broom sounds in the data set were determined as the sounds that do not harm the environment, while saw and lightning were determined as the data set that harms the environment. The dataset was classified using VGG-13BN, ResNet-50 and DenseNet-121 deep learning architectures. When used, all three deep learning accuracy are due to over 95% study. Among these models, the VGG-13 BN model emerged as the most successful model with an accuracy rate of 99.72%.

Keywords: *Deep Learning; transfer learning; spectrogram; environmental sound detection*

1. Introduction

Environmental sustainability is essential for a high quality of life and a sustainable life. For this reason, it is of great importance that any situation that may destroy the environment can be prevented before the destruction occurs [1]. In order to prevent this destruction, it is very important to keep the damage to a minimum by classifying the sound obtained from the action taken or to inform the relevant institutions before the destruction occurs. For this purpose, it is possible to reduce the damage caused by environmental sounds by using different methods.

One of the methods used is the conversion of sound data obtained from the environment into spectrograms, which are a visual representation of the signal frequency spectrum, and classification using transfer learning [2, 3]. Spectrogram; is to calculate the frequency spectrum of an audio signal in each time slot and visually represent it on a time-frequency axis graphic, with the vertical axis frequency value and the horizontal axis time information [4]. The obtained signal is divided into certain parts and the spectrum of each part is processed to calculate. The sample image in Figure 1 is obtained by positioning these different spectra as vertical lines next to each other to create a two-dimensional image.

With the spectrogram method, it is aimed to reduce the frequency structure of audio parts to a very simple structure. To summarize; spectrogram is a visual representation of sound [5,6].

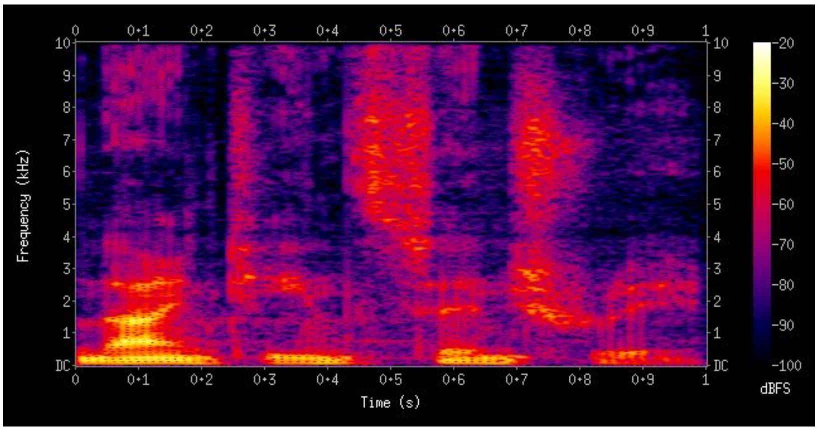

Figure 1. *Spectrogram showing frequency (Hz) and time (time) range*

One of the important usage areas of spectrogram transformation is artificial intelligence applications. When the academic studies on spectrogram transformation based on artificial intelligence are examined, the algorithm extracts the features in the image and the problem turns into a picture classification problem [7]. Thus,

*Corresponding author
E-mail address: bekiraksoy@isparta.edu.tr

Received: 02/Nov/2021; Received in revised form: 01/Feb/2022; Accepted: 01/Feb/2022.

successful results can be obtained without using a complex method. ResNet-50 algorithm is used in transfer learning [8]. As a result of the classification, an average of 90% success was achieved. Experimental results that have been tried on the prototype created data set show that the system can be used successfully in the task of classifying environmental sounds and preventing destruction. It can also be used in solving different problems in daily life by changing some components such as the data set in this proposed system. In addition, this system can perform at an equivalent level when compared to the advanced methods used in the literature.

Sainath Adapa (2019) proposed a structure for environmental noise classification with a low number of data (less than 100 labeled data per class). He stated that in this structure, using transfer learning models together with the data augmentation method, higher performance is achieved compared to alternative approaches. He used this structure in Urban Voice recognition [9].

Muhammad Huzaifah (2017) proposed the visualization of an audio signal through various time-frequency representations such as Spectrograms, as these representations are a rich representation of the temporal and spectral structure of the original signal. To obtain such a representation, the waveform transforms, the constant Q transform, and the Mel measurements were compared and used two different datasets in CNN networks to evaluate their effects on classification performances [10].

Zhichao Zhang, Shugong Xu, Shan Cao, and Shunqing Zhang (2018) used convolutional and pooling layers to extract high-level feature representations from network architecture, spectrogram-like features. They observed the effects of network architectures on performance and feature distributions in different datasets [11].

Justin Salamon and Juan Pablo Bello (2016) argue that the success of deep convolutional neural networks in learning the distinctive features of spectral-temporal patterns makes these networks suitable for environmental sound classification. However, it has been argued that the relative scarcity of labeled data precludes exploitation of this family of high-capacity models. These studies have two contributions to the literature: first, a deep convolutional network architecture is proposed for environmental sound classification. Second, it is proposed to use voice data augmentation to overcome the problem of data scarcity [12].

Juncheng Li, Wei Dai, Florian Metze*, Shuhui Qu, and Samarjit Das (2017) conducted experiments on six different feature sets in their studies. These sets used Mel-frequency cepstral coefficients, binaural mel-frequency cepstral coefficients, log mel-spectrum and two different temporal pooling features. Deep neural network models using large data sets surpass traditional models in performance and stated that they have the highest performance among all the methods studied [13].

A total of 713 voice data obtained from open access websites were used in the study. It is aimed to reduce the damage by classifying the environmental sounds with the use of deep learning methods in the created data set. It is aimed to reduce the damage by classifying the environmental sounds with the use of three different deep learning methods, the data set Resnet-50, DenseNet-121 and VGG-13 BN. Among the three different deep learning architectures used, VGG-13 BN was determined as the deep learning architecture that classifies sounds with 99.72% accuracy.

2. Materials and Method

2.1. Materials

In the study, a data set created from environmental sounds found on open-source websites was used. There are 713 pieces of data in the data set. The dataset consists of 5 different classes: barking, chainsaw, thunder, rain and vacuum cleaner. The dataset was trained with transfer learning using VGG-13BN, ResNet-50 and DenseNet-121 architectures. The results obtained from the architectures were analyzed over the performance evaluation criteria of sharpness, f1 score, sensitivity and accuracy [14, 15]. It was tested using data not available in the post-analysis data set. The data set used in the study, deep learning algorithms and performance evaluation criteria are given in detail below.

2.1.1. Data Set

Each data that makes up the data set has been downloaded from various internet sites in the form of audio files with the extension "wav". Data with different extensions were taken as video clips from open source platforms on the internet and converted into audio data with wav extension. The data set consists of 5 classes: barking, chainsaw, thunder, rain and vacuum cleaner. Among these classes, the sounds of thunder, chainsaw and rain were chosen as the sounds that could cause environmental destruction. Dog barking and vacuum cleaner are included as distinctive sounds. In the dataset, there are approximately 140 voice data belonging to each class. Audio data is converted into spectrograms, which are a visual representation of the frequency spectrum of the signal. With this transformation, audio data is transformed into images and classified. Of the 140 data per class, 80% was used in training and the remaining 20% in testing. An example of a spectrogram is shown in Figure 2.

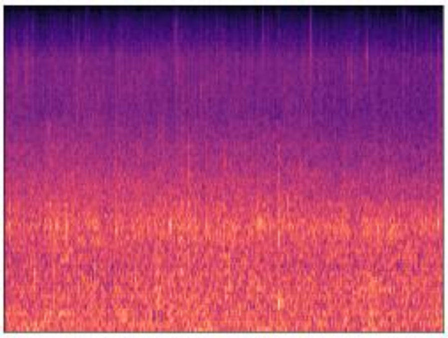

Figure 2. *Spectrogram image formed by preprocessing a sound data*

2.1.2 Deep Learning Architectures

2.1.2.1. VGG 13 BN

VGG-based deep learning methods are frequently used in many areas. One of these architectures, VGG 13 BN architecture, was used in the study. The VGG-13 BN architecture consists of 13 layers (10 convolutional layers + 3 fully connected layers). In the VGG 13 BN architecture, stack normalization layers are used after the convolutional layers [16]. 224x224 images are used as input. In convolutional layers, 3x3 kernels are used. In the pooling layers, the shift value is 2 and the kernel size is 2x2 [17].

2.1.2.2. DenseNet-121

One of the artificial intelligence methods used for images is the DenseNet-121 model. In the DenseNet-121 model, traditional convolutional networks with the number of layers L have L connections, while the DenseNet architecture has as many connections as L(L+1)/2. For each layer, feature maps of previous layers are used as input, and its own feature maps are used as input for all subsequent layers [18, 19]. One of the important advantages of DenseNet architectures is that it reduces the gradient disappearance problem [20, 21].

Thus, it significantly reduces the number of parameters by reusing the features while enhancing the feature dispersion. The DenseNet architecture is shown in Figure 3 [22]. DenseNet-121 architecture consists of 121 layers [23].

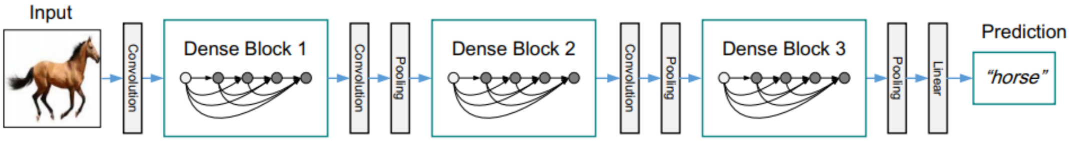

Figure 3. *DenseNet Architecture [22]*

2.1.2.3. ResNet-50

The third deep learning architecture used in the study is the Resnet-50 architecture. ResNet architecture has a high number of layers [24]. It is a deep learning architecture created to solve the problem of performance degradation in convolutional neural network architectures due to the large number of layers [25]. However, networks with a large number of layers achieve success as the depth increases, but after a while, they tend to decrease due to the disappearance of the gradients. To avoid this downward trend, ResNet architectures use hopping connections. By adding jump links, gradients can be easily moved from layer to layer. Thus, from the first layer, even the lowest layers can access the activations in the upper layers, so very deep networks can be trained with ResNet architectures [26].

2.1.3 Performance Evaluation Criteria

Since the classification process was carried out using deep learning methods in the study, sharpness, accuracy, F1 score and sensitivity methods were used to evaluate the performance of the models. The results obtained according to the precision, accuracy, F1 score and sensitivity performance evaluation criteria of the models trained with deep learning architectures are evaluated on the complexity matrix and correct and incorrect predictions are obtained for each class. In Table 1, the complexity matrix structure is shown in tabular form [27].

Table 1. *Structure of the Complexity Matrix*

		ESTIMATED VALUE	
		POSITIVE	**NEGATIVE**
REAL VALUE	**POSITIVE**	TRUE POSITIVE (TP)	FALSE POSİTİVE (FP)
	NEGATIVE	FALSE NEGATİVE (FN)	TRUE NEGATİVE (TN)

2.1.3.1 Precision

Precision answers the question of how accurate the positive predictions are in its models. Equalized as defined at Eq. (1).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

2.1.3.2 Sensivity

Sensitivity answers the question of how accurately true positives are detected in classification models. It is expressed mathematically in Eq. (2).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (2)$$

2.1.3.3 F1 Score

The F1 score shows us the harmonic mean of the precision and sensitivity values. It is expressed mathematically in Eq. (3).

$$\text{F1} = \frac{2* Precision *Sensitivity}{Precision+Sensitivity} \quad (3)$$

2.1.3.4 Accuracy

Accuracy is used to measure the success of the model, but it is not sufficient by itself. It is expressed mathematically in Eq. (4).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Mathematical expressions of the mean of accuracy, sensitivity [28-30], f1 score and precision [29-30] measurements calculated for each class when performing performance evaluations in multi-class problems (more than 2 classes) are given in Eqs. (5)-(8);

$$\text{Average Precision} = \frac{\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i}}{N} \quad (5)$$

$$\text{Average Sensitivity} = \frac{\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i}}{N} \quad (6)$$

$$\text{Average F1 Score} = \frac{\sum_{i=1}^{N}\frac{2* Precision_i * Sensitivity_i}{Precision_i+Sensitivity_i}}{N} \quad (7)$$

$$\text{Average Accuracy} = \frac{\sum_{i=1}^{N}\frac{TP_i+TN_i}{TP_i+FP_i+TN_iFN_i}}{N} \quad (8)$$

2.2. Method

Work flow diagram of the classification of environmental sounds with deep learning is given in Figure 4. In the first stage, a data set was created with the audio data found on open source websites. The data set consists of 5 classes. These classes are bark, thunder, rain, vacuum cleaner and chainsaw sound. There are 140 pieces of data for each class. As a preprocessing step, the audio data were converted into spectrograms and converted into images. Images were trained with DenseNet-121, ResNet-50 and VGG-13 BN deep learning algorithms. 80% of the data in the data set was used for training and 20% for testing. A learning rate of 0.003 was used for each algorithm. Then, the test results of the algorithms were obtained by making predictions with the data that is not in the data set.

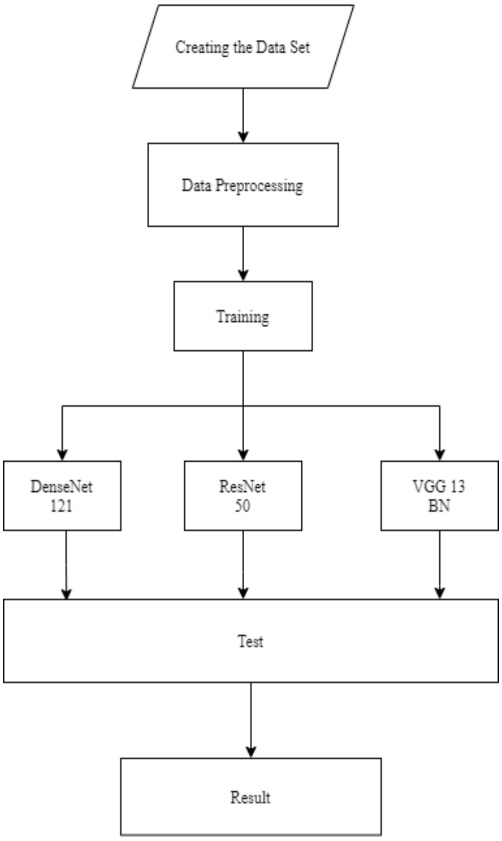

Figure 4. *Work flow diagram*

3. Research Findings

In the study, three different models of sound data were trained using ResNet-50, DenseNet-121 and VGG-13 BN deep learning models. After the training, the model was evaluated according to the performance evaluation criteria by examining the complexity matrices. After the evaluation, the models were tested using an image that was not in the data set for each class, and the following results were obtained. In Figure 5, the test and complexity matrix results of the ResNet-50 model are given.

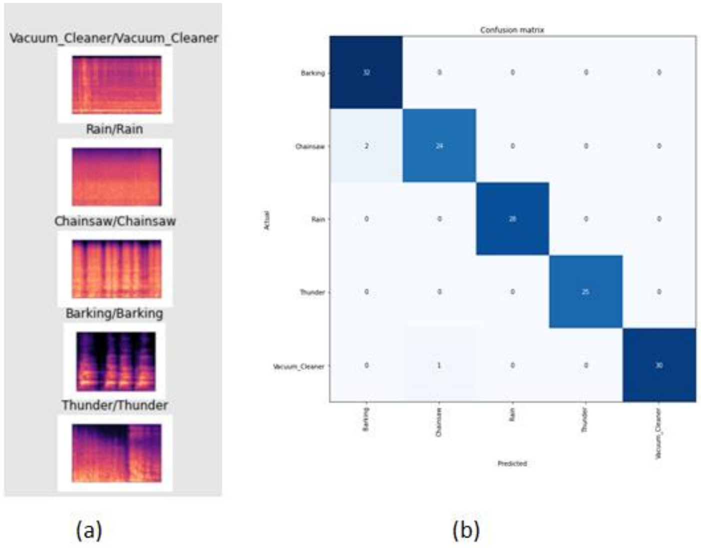

Figure 5. *ResNet-50 (a) test result, (b) complexity matrix*

When the complexity matrix is examined, 142 test data according to the ResNet-50 model are examined; He classified 32 of the 34 test data belonging to the Barking class as correct and 2 as incorrect (Chainsaw). It classified 24 of 25 test data belonging to Chainsaw class as correct and 1 as incorrect (Vacuum_Cleaner). It correctly classified all 28 test data belonging to the Rain class. It correctly classified all 25 test data belonging to the Thunder class. It correctly classified all 30 test data belonging to the Vacuum_Cleaner class. The performance evaluation criteria of the ResNet-50 model are shown in Table 2.

Table 2. *Performance Evaluation Criteria*

Precision	Sensitivity	F1 Score	Accuracy	Test
0.98	0.976	0.975	0.9912	5/5

When Table 2 is examined, it is seen that the ResNet-50 architecture has been successfully classified over 97% according to all performance evaluation criteria. The test result (a) and complexity matrix (b) of the DenseNet-121 model are given in Figure 6.

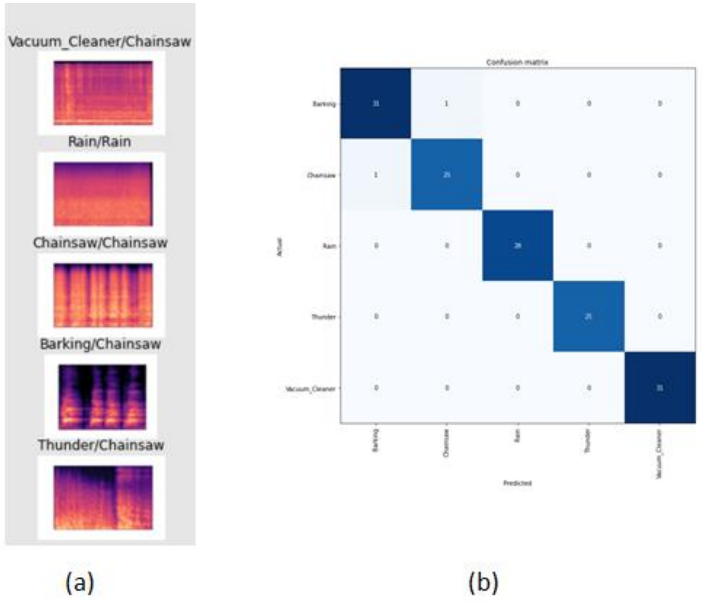

Figure 6. *DenseNet-121 (a) test result, (b) complexity matrix*

When the complexity matrix is examined, 142 test data according to the DenseNet-121 model are examined; He classified 31 of the 32 test data belonging to the Barking class as correct and 1 as incorrect (Chainsaw). It classified 25 of the 26 test data belonging to the Chainsaw class as correct and 1 as incorrect (Barking). It correctly classified all 28 test data belonging to the Rain class. It correctly classified all 25 test data belonging to the Thunder class.

It correctly classified all 31 test data belonging to the Vacuum_Cleaner class. The performance evaluation criteria of the DenseNet-121 model are shown in Table 3.

Table 3. Performance *Evaluation* Criteria.

Precision	Sensitivity	F1 Score	Accuracy	Test
0.9858	0.9858	0.9858	0.99438	2/5

The test result (a) and complexity matrix (b) of the VGG-13 BN model are given in Figure 7.

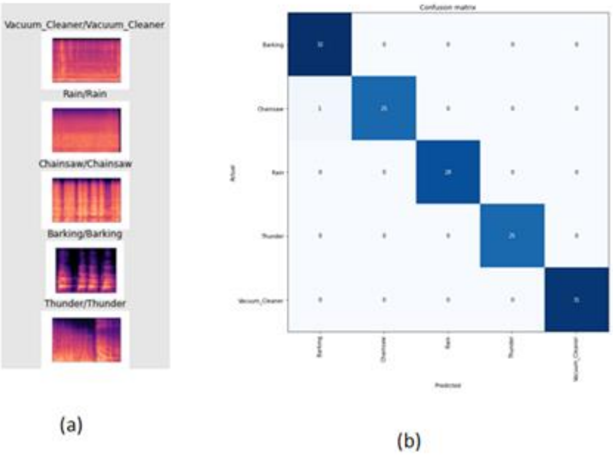

Figure 7. *VGG-13 BN (a) test result, (b) complexity matrix*

When the complexity matrix is examined, 142 test data according to the VGG-13 BN model are examined; He classified 32 of the 33 test data belonging to the Barking class as correct and 1 as incorrect (Chainsaw). It correctly classified all 25 test data belonging to the Chainsaw class. It correctly classified all 28 test data belonging to the Rain class.

It correctly classified all 25 test data belonging to the Thunder class. It correctly classified all 31 test data belonging to the Vacuum_Cleaner class. The performance evaluation criteria of the VGG -13 BN model are shown in Table 4.

Table 4. *Performance Evaluation Criteria.*

Precision	Sensitivity	F1 Score	Accuracy	Test
0.9938	0.9922	0.9928	0.9972	5/5

In Table 5, the results of the performance evaluation criteria of three different deep learning architectures used in the study are given.

Table 5. *ResNet-50, DenseNet-121 and VGG-13 BN Performance Evaluation Criteria*

Model	Precision	Sensitivity	F1 Score	Accuracy	Test
ResNet-50	0.98	0.976	0.975	0.9912	5/5
DenseNet-121	0.9858	0.9858	0.9858	0.99438	2/5
VGG-13 BN	**0.9938**	**0.9922**	**0.9928**	**0.9972**	**5/5**

When Table 5 is examined, it is seen that all three architectures have an accuracy rate of over 97%. Among these three architectures, according to the VGG 13 BN accuracy performance evaluation criterion, approximately 1.78% more successful performance was obtained compared to the ResNet-50 architecture and approximately 0.7% more successful than DenseNet-121.

4. Result

Today, the damages caused by environmental sounds have become more important with the advancement of technology. In the study carried out, five different environmental sounds were classified as Barking, Chainsaw, Rain, Thunder, VacuumCleaner with ResNet-50, DenseNet-121 and VGG-13 BN architectures.
The results obtained from the architectures were evaluated according to four different performance evaluation criteria, and deep learning models were tested for each class by using an image that was not included in the data set. According to the performance evaluation criteria, the most successful model was found to be VGG-13 BN with 99.38% acuity, 99.22% sensitivity, 99.72% accuracy and 99.28% F1 score. In addition, the images that are not in the data set were tested on the models and it was determined that the deep learning models predicted all the images correctly. The ResNet-50 model performs very close to the VGG-13 BN model when the performance evaluation criteria and test results are examined. Although the performance evaluation criteria scores of the ResNet-50 model are lower than DenseNet-121, it has been determined that it performs better when looking at the test results. Thus, it was determined that the ResNet-50 model had a better test result than the DenseNet-121 model.

According to the results obtained from deep learning architectures in the study, it was seen that the best models that can be used in environmental sound detection are VGG 13-BN and ResNet-50. It is thought that performance can be increased with different deep learning architectures by using transfer learning on the data set used in further studies. In future academic studies, it is planned to increase the number of environmental sounds and to carry out different studies by using different deep learning architectures.

Acknowledgement

We would like to thank everyone who made the data sets consisting of environmental sounds used in the study available as open source on different websites.

References

[1] Önder S, Gülgün B. "Gürültü Kirliliği Ve Alınması Gereken Önlemler: Bitkisel Gürültü Perdeleri". *Ziraat Mühendisliği*, 35, 54-64, 2010.

[2] Felipe G Z, Maldonado Y, da Costa G, Helal L G. "Acoustic scene classification using spectrograms". *In 2017 36th International Conference of the Chilean Computer Science Society (SCCC)*, 1-7, 2017.

[3] Olah C. (2020). Understanding lstm networks, August 2015. *URL https://colah. github. io/posts/2015-08-Understanding-LSTMs.* Accessed on, 10.

[4] Nwe T L, Dat T H, Ma B. "Convolutional neural network with multi-task learning scheme for acoustic scene classification". *In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* 1347-1350, 2017.

[5] Kodaloğlu G. Segmentation of snore sounds and detection of sleep apnea with statistical change detection algorithms. *MSc Thesis*, Ankara University, Ankara, Turkey, 2019.

[6] Başbuğ A M. Sound event recognition and acoustic scenes retrieval. *MSc Thesis*, Başkent University, Ankara, Turkey, 2019.

[7] Toraman S, Arslan Tuncer S, Balgetir F. "Is it possible to detect cerebral dominance via EEG signals by using deep learning?" *Medical Hypotheses*, Elazığ: Fırat University, 131, 2019.

[8] Talo M. "Meme Kanseri Histopatolojik Görüntülerinin Konvolüsyonal Sinir Ağları ile Sınıflandırılması". *Fırat University Journal of Engineering Science,* 31(2), 391-398, 2019.

[9] Adapa S. "Urban sound tagging using convolutional neural networks". *arXiv preprint arXiv:1909.12699v1*, 2019.

[10] Huzaifah M. "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks". *arXiv preprint arXiv:1706.07156*, 2017.

[11] Zhang Z, Xu S, Cao S, Zhang S. "Deep convolutional neural network with mixup for environmental sound classification". *In Chinese conference on pattern recognition and computer vision (prcv)*, 356-367, 2018.

[12] Salamon J, Bello J P. "Deep convolutional neural networks and data augmentation for environmental sound classification". *IEEE Signal processing letters*, 24(3), 279-283, 2017.

[13] Li J, Dai W, Metze F, Qu S, Das S. "A comparison of deep learning methods for environmental sound detection". *In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 126-130, 2017.

[14] Özkaya U, Seyfi L. "Yere Nüfuz Eden Radar B Tarama Görüntülerinin Az Parametreye Sahip Konvolüsyonel Sinir Ağı İle Değerlendirilmesi". *Geomatik*, 6(2), 84-92, 2021.

[15] Bozkurt F, Yağanoğlu M. “Derin Evrişimli Sinir Ağları Kullanarak Akciğer X-Ray Görüntülerinden COVID-19 Tespiti”. *Veri Bilimi*, 4(2), 1-8, 2021.

[16] Turhan C G, Bilge H Ş. “Çekişmeli üretici ağ ile ölçeklenebilir görüntü oluşturma ve süper çözünürlük”. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 35(2), 953-966, 2020.

[17] Akılotu B N, Kadiroğlu Z, Şengür A, Kayaoğlu M. “Evrişimsel Sinir Ağları ve Transfer Öğrenme Yöntemi Kullanılarak Sıtma Tespiti”. *International Engineering and Science Symposium, Siirt*, 2019.

[18] Bozkurt F. “Derin Öğrenme Tekniklerini Kullanarak Akciğer X-Ray Görüntülerinden COVID-19 Tespiti”. *Avrupa Bilim ve Teknoloji Dergisi*, 24, 149-156, 2021.

[19] Kumar R. “Adding binary search connections to improve densenet performance”. In *5th International Conference on Next Generation Computing Technologies (NGCT-2019)*, 2020.

[20] Korfiatis P, Kline T L, Lachance D H, Parney I F, Buckner J C, Erickson B J. “Residual deep convolutional neural network predicts MGMT methylation status”. *Journal of digital imaging*, 30(5), 622-628, 2017.

[21] Fu Y, Aldrich C. “Flotation froth image recognition with convolutional neural networks’’. *Minerals Engineering*, 132, 183-190, 2019.

[22] Huang G, Liu Z, van der Maaten L, Weinberger K Q. “Densely connected convolutional networks”. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700-4708, 2017.

[23] Li X, Shen X, Zhou Y, Wang X, Li T Q. “Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet)’’, *PloS one*, Hangzhou: China Jiliang University, 15(5), e0232127, 2020.

[24] Duman E, Akın F. “Yüz Tanima Sürecinde Farkli Cnn Modellerinin Performans Karşilaştirmasi’’. *Uluslararası Mardin Artuklu Multidisipliner Çalışmalar Kongresi*, 35-42, 2019.

[25] Narin A. “Meme Kanserinin Evrişimsel Sinir Ağı Modelleriyle Tespitinde Farklı Görüntü Büyütme Oranlarının Etkisi’’. *Karaelmas Fen ve Mühendislik Dergisi*, 10(2), 186-194, 2020.

[26] Tan Z. Vehicle classification with deep learning. *MSc Thesis*, Fırat University, Elazığ, Turkey, 2019.

[27] Deng X, Liu Q, Deng Y, Mahadevan S. “An improved method to construct basic probability assignment based on the confusion matrix for classification problem”. *Information Sciences*, 340, 250-261, 2016.

[28] Orman A, Köse U, Yiğit T. “Açıklanabilir Evrişimsel Sinir Ağları ile Beyin Tümörü Tespiti”. *El-Cezeri*, 8(3), 1323-1337, 2021.

[29] Sokolova M, Lapalme G. “A systematic analysis of performance measures for classification tasks”. *Information processing & management*, 45(4), 427-437, 2009.

[30] Ballabio D, Grisoni F, Todeschini R. “Multivariate comparison of classification performance measures”. *Chemometrics and Intelligent Laboratory Systems*, 174, 33-44, 2018.