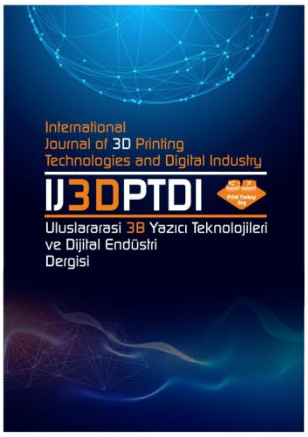
International
Journal of 3D Printing
Technologies and Digital Industry
IJ3DPTDI
Uluslararası 3B Yazıcı Teknolojileri
ve Dijital Endüstri
Dergisi

ULUSLARARASI 3B YAZICI TEKNOLOJİLERİ
VE DİJİTAL ENDÜSTRİ DERGİSİ

INTERNATIONAL JOURNAL OF 3D PRINTING
TECHNOLOGIES AND DIGITAL INDUSTRY

ISSN:2602-3350 (Online)

URL: https://dergipark.org.tr/ij3dptdi

CLASSIFICATION OF INSTRUMENT SOUNDS WITH IMAGE CLASSIFICATION ALGORITHMS

Yazarlar (Authors): Remzi GÜRFİDAN*

Bu makaleye şu şekilde atıfta bulunabilirsiniz (To cite to this article): *Gürfidan R., "Classification of Instrument Sounds with Image Classification Algorithms" Int. J. of 3D Printing Tech. Dig. Ind., 7(3): 513-519, (2023).*

DOI: 10.46519/ij3dptdi.1330052

Araştırma Makale/ Research Article

Erişim Linki: (To link to this article): https://dergipark.org.tr/en/pub/ij3dptdi/archive

CLASSIFICATION OF INSTRUMENT SOUNDS WITH IMAGE CLASSIFICATION ALGORITHMS

Remzi GÜRFİDAN[a]*

[a]Isparta University of Applied Sciences, Yalvaç Vocational School of Technical Sciences, Computer Programming, Turkey

*Corresponding Author: remzigurfidan@isparta.edu.tr

(**Received:** 02.08.23; **Revised:** 26.09.23; **Accepted:** 20.11.23)

ABSTRACT

Classification of audio files using CNN (Convolutional Neural Network) algorithm is an important application in the field of audio processing and artificial intelligence. This process aims to automatically classify audio files into different classes and can be used in speech recognition, emotional analysis, voice-based control systems and many other applications. The aim of this study is to perform spectrum transformation of instrumental sounds and classify them using image classification algorithms. The dataset contains a total of 1500 data from five different instruments. Audio files were processed, and signal and spectrogram images of each audio file were obtained. DenseNet121, ResNet and CNN algorithms were tested in experimental studies. The most successful results belong to the CNN algorithm with 99.34%.

Keywords: Instrument Sound Classification, Machine Learning, CNN Algorithm

1. INTRODUCTION

The automatic detection and classification of acoustic phenomena in audio signals is of great importance in many fields. These include security and surveillance scenarios, source recognition, communication, machine, and human interaction. The breadth of applications of voice recognition technology shows that the scope and usefulness of voice classification is not limited to human voices. Although many environmental sounds can be detected and classified with high accuracy with the partnership of a healthy human ear and mind, this work must be realized with technological infrastructures to increase human-machine interaction.

Sound classification provides a natural interface for human-machine interaction and enables users to interact more easily, quickly and in a personalized way. The ability to understand and accurately interpret environmental sounds provides users with a more intuitive and interactive experience.

Sound classification and detection can be performed in multiple digital ways. The first step in this process is data collection. The audio data to be identified or classified should be obtained from various sources or sensors and clustered. The audio data must then be pre-processed. This preprocessing usually involves noise reduction, frequency transformations or time delimitation. To classify the pre-processed audio data correctly, feature extraction is performed. This process is quite complex and challenging. Because the selection of the right feature set for the best classification directly affects the result [1]. In this process, methods such as Mel frequency Cepstral Coefficients (MFCC), spectral features, energy measurements can be used [2]. After this stage, the desired model training can be provided by choosing from machine learning algorithms and evaluation results can be obtained. The most common approach to image-based processing of audio data is to use a spectrogram image as feed data to the machine learning algorithm [3], [4].

The classification of musical instrument sounds is a subject of study that can be found in a wide range of fields. Classification of musical instrument sounds plays an important role in music recognition and labeling applications. Classification or clustering operations vary depending on the use of the information in the music recordings [5]. For instance, a music library or streaming service can group instrument sounds into playlists for consumers or propose songs based on their tastes. It can also be applied to the teaching and learning of music. It can be used to teach pupils how to identify the instrument being played correctly or to hone abilities like note identification. Students will be better able to identify instruments and develop their musical skills as a result. In addition, music creation and editing software heavily relies on the classification of musical instruments. For example, an audio editing software can automatically detect and categorize different instrument sounds, allowing users to edit faster and more accurately. For those working in this field, it is used in musical research and analysis. Classification of instrument sounds is very suitable for studying the distribution of instruments or characteristic sounds of a particular musical genre. This provides an important tool for understanding musical patterns and structures.

Our motivation for conducting this study is to classify audio files using machine learning techniques and image classification algorithms. Although this is already covered in the existing literature, this study enriches the existing literature by comparing the results of more than one algorithm. In addition, the dataset used in the training of machine learning algorithms in the study is a unique dataset created by the researcher of this study. It is aimed to share the dataset with the scientific world by further enriching it in future studies.

The second part of the study includes the review of existing studies in the literature, the third part includes how the voice classification process is performed and the methods used, the fourth part includes the findings and interpretations obtained, and the last part includes the conclusions obtained from the study.

2. RELATED WORKS

Two distinct but related approaches—a perceptual approach and a taxonomy approach—have been studied in research on the automatic classification of musical instrument sound. The former seeks to produce perceptual similarity functions for timbre grouping, sound retrieval, and search and retrieval based on timbral similarity. The second tries to produce sound labeling indices based on user- or culture-side taxonomies. Different methods for similarity-based grouping and categorization of sounds into preset instrumental categories are described after evaluating the pertinent features employed in these two domains [6]. Sound classification studies are also very important in the field of healthcare. Researchers have developed an electronic stethoscope that can store and classify respiratory sounds. They used two types of machine learning algorithms to classify the sounds. Spectrogram pictures in convolutional neural networks (CNN) and support vector machines (SVM) both use mel frequency cepstral coefficient (MFCC) features as one of them [7]. The objective is to develop classification models and methods to recognize aberrant respiratory sounds (cracking, wheezing) for the automatic diagnosis of respiratory and pulmonary disorders. Another study suggests a deep CNN-RNN model that uses Mel-spectrograms to categorize respiratory sounds. Additionally, a patient-specific model tuning technique was put into place, which screens respiratory patients first, then uses the scant amount of patient data to create patient-specific classification models for accurate anomaly identification [8]. Again, artificial intelligence algorithms have been utilized for the classification of heart rhythm sound in the health field. In this study, four different pathological classifications were studied using SVM [9]. Since CNN algorithm is popular in voice classification, different CNN topologies were prepared to classify lung sounds with this method. In order to make the results more successful, tests were performed by combining with SVM [10].

Researchers have developed a variety of sound classification studies and have presented findings in different concepts. Researchers have carried out a study in acoustic scene classification, which is the task of classifying environments from the sounds they produce [11]. To address the environmental sound categorization (ESC) issue, deep characteristics were favored. Using fully connected layers of a recently created CNN model that was trained end-to-end on spectrogram images, deep features are recovered. The suggested CNN model's fully linked layers are combined to create the feature vector [12]. It has been looked at how to take representative characteristics out of a new deep neural network model for a music recommendation engine and categorization system. The categorization and recommendation of musical genres were performed on a dataset using the auditory features gathered from these networks [13]. The learning and combining of multimodal data representations for music genre classification is offered as a method. Audio tracks, written reviews, and cover art pictures are used to train intermediate representations of deep neural networks, which are then integrated for categorization. Following that, single- and multi-label research on genre classification look at the effects of different learned representations and their combinations. The findings from these two experiments show how mixing learned representations from different modalities improves classification accuracy and shows that different modalities include complementary information [14]. Another study eliminates the need for manual feature selection by converting the audio signal of music into an audio spectrum as a consistent representation. According to the audio spectrum's characteristics, the research combined 1D convolution, gating, residual connectivity, and attention mechanisms, and it proposed a music feature extraction and classification model based on convolutional neural networks that can extract more useful audio spectrum features for the music category [15]. Convolutional neural networks have been studied in various studies to see if they can be successfully used for ambient sound categorization tasks, especially given the small number of datasets available in this area. The results of the tests indicate that a convolutional model performs better than conventional strategies based on manually created features and is on par with other feature learning techniques. Given the substantially longer training timeframes, the outcome is by no means revolutionary, but it demonstrates that convolutional neural networks may be used to classify environmental sounds even with small datasets and straightforward data augmentation [16]. Another study looked into the combination of CNN-based models. Three different methodologies were utilized in the study to categorize 43 different species of birds. Bird species were categorized using a VGG-style network. To further characterize bird cries, another SubSpectralNet was added. An attempt was made at class-based fusion to further enhance classification performance. It selectively blends four different streams from CNN [17]. In study, we investigated the sound categorization procedure for a honeybee colony in each scenario using spectrogram image features. Six groups were investigated in order to train the categorization models. The model's accuracy was 99.82%, which was the highest in Logistic Regression. The study's findings demonstrate the high effectiveness of applying spectrogram picture features to comprehend honeybee sound classifications [18]. The purpose of this research is to extract spectrogram images of features from audio samples. The 7-layer or 9-layer CNN architectures chosen at random and used in this model training were created from scratch. This research suggests a way for meaningful data augmentation by taking modifications applied directly to audio samples into account rather than using existing data augmentation schemes for graphics. The outcomes show that the suggested approach is efficient, reliable, and highly accurate. ResNet-152, one of the models employed, achieved 99.04% for the ESC-10 and 99.49% for the Us8k datasets. For ESC-50, DenseNet-161 achieved 97.57% [19]. This study proposes an auditory classification approach that can discriminate between typical and unusual noises made during concrete pouring. The researchers describe an experiment in which data on background noise, main noise from construction, and symptoms that can impair structural quality are recorded throughout concrete pouring. A deep learning-based categorization algorithm was created to foresee occurrences that could compromise the quality and safety of structures by analyzing the acoustic data collected from real construction sites and experiments. Convolutional neural networks (CNN) and recurrent neural networks (RNN) both displayed excellent performance in the classification model, with respective scores of 94.38% and 93.26% [20].

3. METHOD AND DISCUSSION

Recordings were made in a sound-isolated environment during the solo performance of the instruments to create the data set. The recordings were then divided into periods of 6 seconds each. There are 300 recordings for each data class. The fragmented audio files were converted to wav format. Spectrum analysis of the wav format audio files was performed.

Figure 1 shows the fragmented audio files after the analysis. The blue-coloured ones are the signal graphs of the audio files, while the yellow and red coloured ones are the spectrum graphs. The analysis results were saved as image files with jpg extension. Each instrument sound group was labelled, and the dataset was made ready for training. The model to be created was trained with deep learning algorithms.

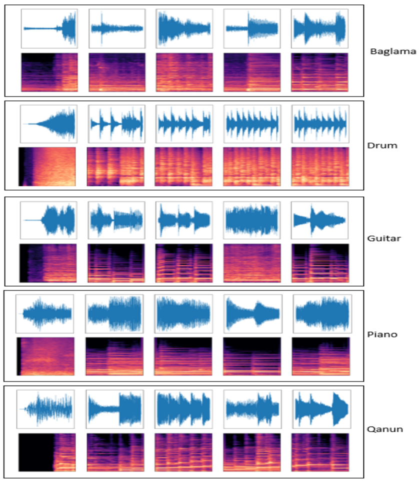

Figure 1. Signal and spectrogram graphs of audio files

Figure 2 shows the architecture of the CNN model proposed for this study. Convolution processes and pooling processes are shown in detail.

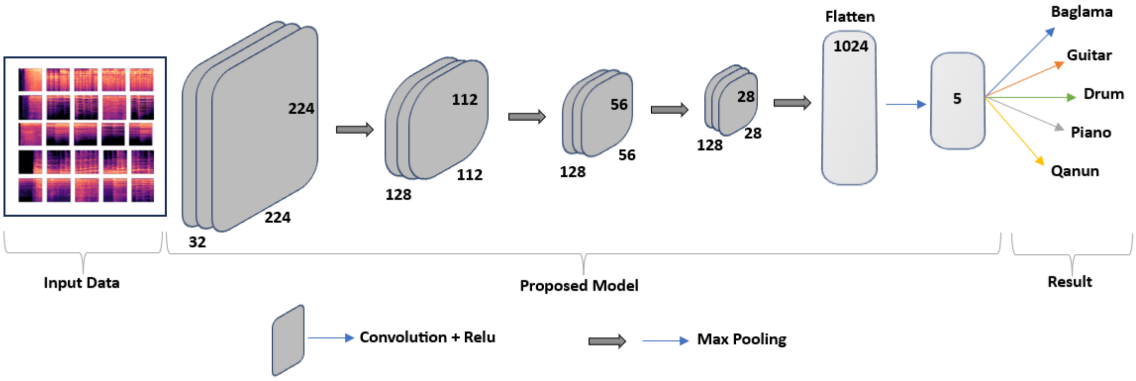

Figure 2. CNN model architecture

The CNN model implemented in this study includes convolution, pooling, smoothing and density layers. The layers and their values used in the architecture created for this dataset are shown in Table 1.

Table 1. Model summary of the selected CNN algorithm

Model Layer	Output Shape Form	Params.
Conv-2d (Conv2D)	(None, 222, 222, 32)	896
maxpooling2D(MaxPooling-2D)	(None, 111, 111, 32)	0
Conv-2d_1 (Conv-2D)	(None, 109, 109, 128)	36992
maxpooling2D_1(MaxPooling-2D)	(None, 54, 54, 128)	0
Conv-2d_2 (Conv-2D)	(None, 52, 52, 128)	147584
maxpooling2D_2(MaxPooling-2D)	(None, 26, 26, 128)	0
Conv-2d_3 (Conv-2D)	(None, 24, 24, 128)	147584
maxpooling2D_3(MaxPooling-2D)	(None, 12, 12, 128)	0
Flatten (Flatten)	(None, 18432)	0
Dense (Dense)	(None, 1024)	18875392
Dense-1 (Dense)	(None, 5)	5125

Figure 3 shows the correct and incorrect predictions of the model on the confusion matrix during the testing process at the end of training. Figure 3 shows the confusion matrix for the three algorithms trained and tested.

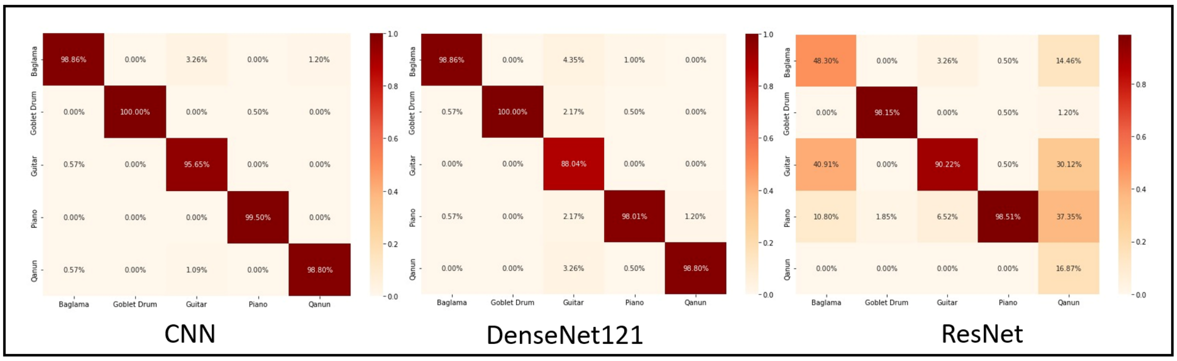

Figure 3. Complexity matrix of algorithms

The graphs of the success and loss values of the algorithms tested as the decision mechanism of the proposed classification model during training are shown in Figure 4. Comparative results of other machine learning algorithms tested will also be discussed in this section. The dataset is divided into two parts in two stages: 80% training data and 20% test data and 90% training data and 10% test data. With the resulting dataset, 10 epochs of training were repeated. Based on the val_accuracy metric, the CNN algorithm achieved 99.34% accuracy, denseNet121 96.70%, ResNet 85.97%. The loss value decreased to 0.0269 in the CNN model, which gave the most successful result.

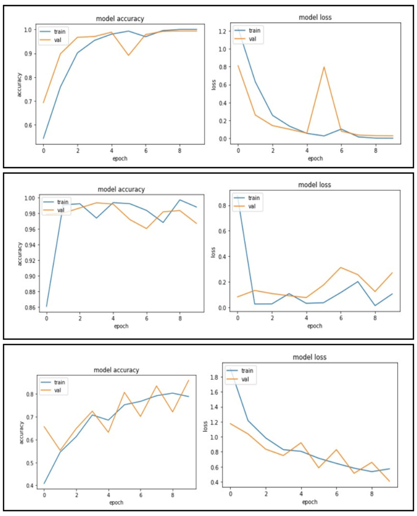

Figure 4. Variation graph of success and loss values of different algorithms according to the number of repetitive trainings

Table 2 shows the values of the metrics obtained according to the training results of different image-based learning algorithms. The most successful results on the dataset in the current study belong to the CNN model preferred in this study. Based on the values in this table, it is seen that the trained model is successfully trained, and the dataset is suitable for training.

Table 2. Metric values in training results of different algorithms

Data Split	Models Metrics	CNN	Dense Net121	ResNet
80% - 20%	Loss	3.58e-04	0.1044	0.5730
	accuracy:	1.0000	0.9880	0.7885
	f1_m	1.0000	0.9880	0.7310
	precision_m	1.0000	0.9880	0.8330
	recall_m	1.0000	0.9880	0.7754
	val_loss	0.0269	0.2701	0.4082
	val_accur.	0.9934	0.9670	0.8597
	val_f1_m	0.9934	0.9672	0.8279
	val_prec_m	0.9934	0.9672	0.8884
	val_recall_m	0.9934	0.9672	0.8557
90% - 10%	Loss	0.0348	0.1821	0.8730
	accuracy:	0.9879	0.9696	0.7983
	f1_m	0.9879	0.9687	0.7410
	precision_m	0.9879	0.9687	0.8430
	recall_m	0.9879	0.9687	0.7954
	val_loss	0.0029	0.7683	0.4280
	val_accur.	1.0000	0.9290	0.8490
	val_f1_m	1.0000	0.9310	0.8375
	val_prec_m	1.0000	0.9326	0.8685
	val_recall_m	1.0000	0.9295	0.8755

The ROC curve reveals the performance of a trained artificial intelligence model in its classification capability. The ROC Curve is a straight line starting from the top left corner and represents the performance of a correctly working classification model. The area under the curve is called the AUC (Area Under the Curve) and is used to measure the overall performance of a model. The AUC value can vary between 0 and 1. The closer the reading is to 1, the better the performance of the model. ROC curves are interpreted to select the appropriate threshold value. The AUC metric represents the area under the ROC curve plot. Figure 5 shows the plot of the algorithms tested against the AUC value of the ROC curve and the average AUC value.

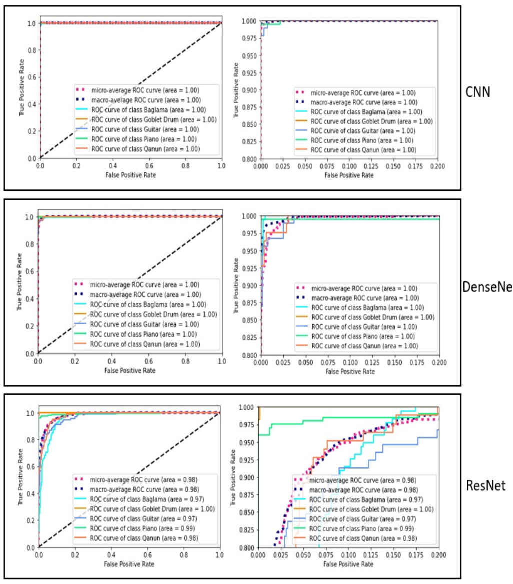

Figure 5. ROC Curve Plots according to training results of different algorithms

When the ROC curve graphs prepared for CNN, DenseNet121 and ResNet are examined, it is seen that the CNN algorithm has the most successful results. The x-axis shows the specificity values, and the y-axis shows the sensitivity values. the red dashed diagonal line is the representation of a ROC model making a random prediction. therefore, the trained model should stay on this ROC curve. otherwise, it means that it produces worse results than a random prediction. We see that this is not the case with our models. The ROC curve shows the change in sensitivity and specificity of the model at different threshold values. The threshold value determines the point at which the model will classify as positive or negative. we expect this point to be close to the upper left corner. The closest value is seen in the CNN algorithm. The best point on the ROC curve is usually in the upper left corner and is the point of highest sensitivity and highest specificity. This point represents the point where the model achieves the best balance and performs best. This is again best represented by the CNN algorithm. Table 3 shows how the current study compares with similar studies in terms of year, technique, success metric and success value.

Table 3. Comparison of the current study with similar studies

Ref.	Technique	Acc.	Metrics	Year
[12]	CNN,VGG16, VGG19,AlexNet,ResNet50	%86,7	Avg_acc.	2020
[18]	Logistic Regression	%99,82	Accuracy	2022
[19]	CNN	%99,49	Accuracy	2021
[20]	CNN, RNN	%94,38	Accuracy	2023
[21]	CNN	%91	Accuracy	2021
[22]	CNN	%95	Accuracy	2018
This Work	CNN, DenseNet121, ResNet	%99,34	Val_acc.	2023

4. CONCLUSION

Automatic detection and classification of acoustic phenomena in audio signals is of great importance in many fields. The aim of this study is to classify audio files using machine learning techniques and image classification algorithms. The dataset for training the machine learning models is unique as it was created by the author of the study. In the study, learning processes were performed with CNN, DenseNet121 and ResNet algorithms and comparisons were made. When the success rates obtained are analysed, CNN was the most successful algorithm with 99.34%. Unlike the off-the-shelf DenseNet and ResNet architectures, the CNN architecture we created was a more suitable training algorithm for this dataset. In future studies, we aim to enrich the dataset and perform instrument extraction in files containing complex audio nodes. In this way, an exam tool that can be used in the aptitude exams of gifted students will be realized.

REFERENCES

1. Antti E., ‘Automatic musical instrument recognition’, Master Thesis, TAMPERE UNIVERSITY OF TECHNOLOGY, Finland, 2001.

2. Cotton C. V. and Ellis D. P. W., ‘Spectral vs. spectro-temporal features for acoustic event detection’, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Pages. 69–72, 2011,

3. Lee H., Largman Y., Pham P., and Ng A. Y., ‘Unsupervised feature learning for audio classification using convolutional deep belief networks’, Adv Neural Inf Process Syst, Vol. 22, 2009.

4. Abdel-Hamid O., Mohamed A. R., Jiang H., Deng L., Penn G., and Yu D., ‘Convolutional neural networks for speech recognition’, IEEE Trans Audio Speech Lang Process, Vol. 22, Issue 10, Pages 1533–1545, 2014.

5. Özbek, M. E., Savacı, F. A., Genelleştirilmiş Gauss yoğunluk modellemesi ile müzik aletlerinin sınıflandırılması. 2007 IEEE 15th Signal Processing and Communications Applications, 2007.

6. Perfecto Herrera-Boyer G. P. S. D., ‘Automatic Classification of Musical Instrument Sounds’, J New Music Res, Vol. 32, Issue 1, Pages 3–21, 2003.

7. Aykanat M., Kılıç Ö., Kurt B., and Saryal S., ‘Classification of lung sounds using convolutional neural networks’, EURASIP J Image Video Process, Vol. 2017, Issue 1, Pages 1–9, 2017.

8. Acharya J. and Basu A., ‘Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning’, IEEE Trans Biomed Circuits Syst, Vol. 14, Issue 3, Pages 535–544, 2020.

9. Redlarski G., Gradolewski D., and Palkowski A., ‘A System for Heart Sounds Classification’, PLoS One, Vol. 9, Issue 11, Pages e112673, 2014.

10. Bardou D., Zhang K., and Ahmad S. M., ‘Lung sounds classification using convolutional neural networks’, Artif Intell Med, Vol. 88, Pages 58–69, 2018.

11. Barchiesi D., Giannoulis D. D., Stowell D., and Plumbley M. D., ‘Acoustic Scene Classification: Classifying environments from the sounds they produce’, IEEE Signal Process Mag, Vol. 32, Issue 3, Pages 16–34, 2015.

12. Demir F., Abdullah D. A., and Sengur A., ‘A New Deep CNN Model for Environmental Sound Classification’, IEEE Access, Vol. 8, Pages 66529–66537, 2020.

13. Elbir A. and Aydin N., ‘Music genre classification and music recommendation by using deep learning’, Electron Lett, Vol. 56, Issue 12, Pages 627–629, 2020.

14. Oramas S., Barbieri F., Nieto Caballero O., and Serra X., ‘Multimodal deep learning for music genre classification’, Transactions of the International Society for Music Information Retrieval, Vol. 1, Issue 1, Pages 4–21, 2018.

15. Zhang J., ‘Music Feature Extraction and Classification Algorithm Based on Deep Learning’, Sci Program, Vol. 2021, 2021,

16. Piczak K. J., ‘Environmental sound classification with convolutional neural networks’, IEEE International Workshop on Machine Learning for Signal Processing, MLSP, Vol. 2015-November, 2015.

17. Xie J., Hu K., Zhu M., Yu J., and Zhu Q., ‘Investigation of Different CNN-Based Models for Improved Bird Sound Classification’, IEEE Access, Vol. 7, Pages 175353–175361, 2019.

18. Mekha P., Teeyasuksaet N., Sompowloy T. and Osathanunkul K., "Honey Bee Sound Classification Using Spectrogram Image Features," 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Pages 205-209, Chiang Rai, 2022.

19. Mushtaq, Z., Su, S. F., & Tran, Q. V. Spectral images based environmental sound classification using CNN with meaningful data augmentation. Applied Acoustics, Vol. 172, Issue 107581, 2021.

20. Kim, I., Kim, Y., & Chin, S. (2023). Deep-Learning-Based Sound Classification Model for Concrete Pouring Work Monitoring at a Construction Site. Applied Sciences, Vol 13, Issue 8, Pages 4789.

21. Massoudi M., Verma S. and Jain R., "Urban Sound Classification using CNN," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Pages 583-589, Coimbatore, 2021.

22. Jaiswal K. and Kalpeshbhai Patel D., "Sound Classification Using Convolutional Neural Networks," 2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Pages 81-84, Bangalore, 2018.