JOIV : Int. J. Inform. Visualization, 6(1-2): Data Visualization, Modeling, and Representation - May 2022 250-255

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Development on Deaf Support Application Based on Daily Sound Classification Using Image-based Deep Learning

Ji-Hee An [a], Na-Kyoung Koo [a], Ju-Hye Son [a], Hye-Min Joo [a], Seungdo Jeong [a,*]

[a] *Department of Smart Information and Telecommunication Engineering, Sangmyung University, Cheonan, Chungnam, Republic of Korea*

Corresponding author: [*]sdjeong@smu.ac.kr

***Abstract*— According to statistics, the number of hearing-impaired persons among the disabled in Korea accounts for 27% of all persons with disabilities. However, there is insufficient support for the deaf and hard of hearing's protective devices and life aids compared to the large number. In particular, the hearing impaired misses much information obtained through sound and causes inconvenience in daily life. Therefore, in this paper, we propose a method to relieve the discomfort in the daily life of the hearing impaired. It analyzes sounds that can occur frequently and must be recognized in daily life and guide them to the hearing impaired through applications and vibration bracelets. Sound analysis was learned by using deep learning by converting sounds that often occur in daily life into the Mel-Spectrogram. The sound that actually occurs is recorded through the application, and then it is identified based on the learning result. According to the identification result, predefined alarms and vibrations are provided differently so that the hearing impaired can easily recognize it. As a result of the recognition of the four major sounds occurring in real life in the experiment, the performance showed an average of 85% and an average of 80% of the classification rate for mixed sounds. It was confirmed that the proposed method can be applied to real-life through experiments. Through the proposed method, the quality of life can be improved by allowing the hearing impaired to recognize and respond to sounds that are essential in daily life.**

***Keywords*— Sound analysis; Mel-Spectrogram; YOLO; deep learning; hearing impaired.**

Manuscript received 25 Nov. 2021; revised 19 Jan. 2022; accepted 28 Mar. 2022. Date of publication 31 May 2022.

I. INTRODUCTION

According to the statistics published on the Ministry of Health and Welfare website in Korea, there were a total of 2.63 million people with disabilities, and about 396,000 people with hearing impairments accounted for the largest proportion, excluding the physically challenged [1]. However, according to the '2017 White Paper for the Disabled', only 18.7% of the respondents answered that visual alarms and signal alerts for the hearing impaired were installed [2]. Most things can be easily responded to and solved by accurately recognizing the sound that occurs in daily life, but deaf people feel much inconvenience in their daily lives because they cannot hear the sound. Therefore, if a sense that can replace hearing can be used, it will be possible to provide more convenience in daily life.

According to an article on the deaf, hearing impaired people have a high rate of not recognizing sounds and vibrations, so the risk of experiencing risks in daily life is relatively high, and smartphones, vibration digital alarm clocks, and vibration wrist watches are frequently used as auxiliary devices to recognize and notify risk situations [3-8]. This article determined that a smartphone application that uses sight instead of hearing and a vibration bracelet that uses tactile feel are highly likely to be easily encountered and recognized by the hearing impaired. We intend to propose a method to provide auditory information using this.

In this paper, among the important but easy to miss in daily life, four sounds were selected: baby cry, washing machine shut down sound, doorbell sound, and fire alarm sound. A deep learning framework using the YOLO model [9], [10] was used to recognize the sound and judge it according to the learning result, and it was implemented to enable bracelet vibration control using Bluetooth communication.

In the study of risk detection aids for the hearing impaired, the device is designed to operate if the decibel is above a certain decibel by measuring the decibel with Arduino [11]. However, regardless of what sound it is, it distinguishes whether it is a dangerous situation or not only by detecting a specified decibel abnormality. Even if there are no threats such as laughter or applause around the device, there is a disadvantage of notification regardless of the risk if the decibel above the threshold is measured. The study on the

development of a risk prevention application for the hearing impaired through automata synthesis was designed to give notifications according to the surrounding decibels in each case of daily life and driving using automata synthesis [12]. In this study, there is a disadvantage that the information provided to the hearing impaired is provided only through a smartphone, so there may be situations in which it may not be recognized quickly. In research on the development of sound information delivery aids for the hearing impaired, a transmitter is attached to a home appliance or electronic device that wants to receive a notification, recognizes a sound through a voice recognition module, and transmits a signal to the receiver to notify the receiver with an LCD screen and a vibration module gives [13]. This method has a disadvantage in that multiple transmitters and receivers are needed to detect various sounds because the transmitter and receiver communicate one-to-one. As a study related to sound analysis, a study on an application that classifies and informs a baby's cry [14], [15] and a study on city noise [16], [17] and specific sounds [18], [19] were conducted. Most of the above studies have limitations in being applied to a limited situation according to a specific sound or environment. The study of Boddapati et al. can be said to be the same study as the sound visualization method in this paper [20].

In this paper, we aim to overcome the limitations of the related studies discussed above and proceed with the classification of various sounds in the indoor environment. In addition, we propose a method to increase the efficiency in providing the classification results to the visually impaired. That is, in order to provide visual information using an application and to directly feel it without being limited to providing visual information, a vibration bracelet is produced so that it can be intuitively recognized and responded.

II. Materials and Methods

This section presents the overall system configuration and a detailed description of the proposals.

A. Proposed System Configuration

The overall system configuration diagram proposed in this paper is presented in Fig. 1. As presented in the system configuration diagram, the pre-collected learning sound data is imaged, and the learned model is stored in the server.

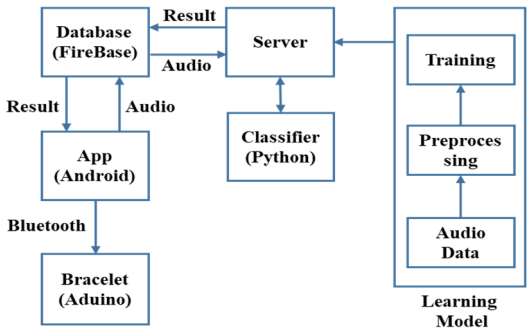

Fig. 1 System configuration

In the application, if a sound exceeding the reference value is detected by detecting the decibel, the recording proceeds, and the recorded file is uploaded to the FireBase [21]. The server downloads the file, executes a preprocessing process of imaging the recorded file and uploads the result value to the FireBase as text. When the result value is uploaded, the application displays a notification window and sends a signal to the bracelet to execute a vibration matching the result value.

The application runs in the background and measures the decibel of sound entering the microphone in real-time. The types of sounds to be classified consist of four types: washing machine end sound, baby cry, doorbell sound, and fire alarm sound.

TABLE I
Decibel of Type of Sound

Type of sound	Decibel(dB)
Daily life noise	About 25 ~ 30 dB
Doorbell	About 68dB
Washing machine end sound	About 63db

As can be seen in Table 1, daily life noise was measured to be about 25 to 30 dB, doorbell sound to be about 68 dB, and washing machine end sound to be about 63 dB. Thus the decibel at which recording starts was set to 60dB, and sound recording was set to be performed for 10 seconds from the moment 60dB was exceeded. Conventionally, sound analysis requires analysis of sound data itself, so it is not possible to utilize an image-based deep learning framework or machine learning network. The video-based deep learning framework is relatively well configured and can be easily used in mobile applications, so this study used the video-based deep learning framework. Therefore, since an image-based discrimination network was used, there is an advantage that it is easy to use by additionally expanding it to various existing applications [20].

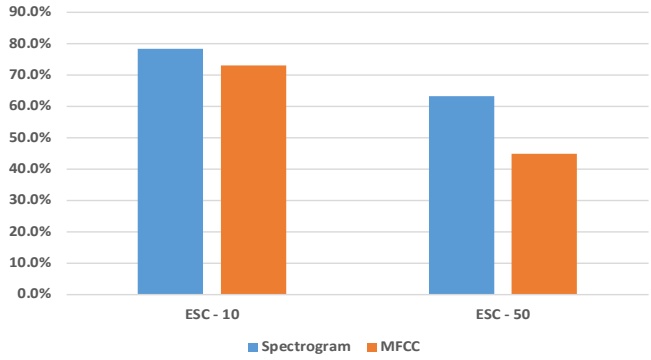

Fig. 2 Comparison of sound analysis performance of AlexNet

Figure 2 is constructed based on the contents described in the previous study [20], and shows the results when the ESC-10 dataset and ESC-50 are trained with AlexNet by extracting features using Spectrogram and MFCC, respectively [22], [23]. As shown in the results, it can be seen that the results of learning with the spectrogram are better. Therefore, in this study, spectrogram features were extracted and used for sound analysis.

B. Mel-Spectrogram

In this paper, the sound was converted into an image to classify the sound. Mel-Spectrogram shown in Fig. 3 was introduced as a method of imaging sound data [24]-[26].

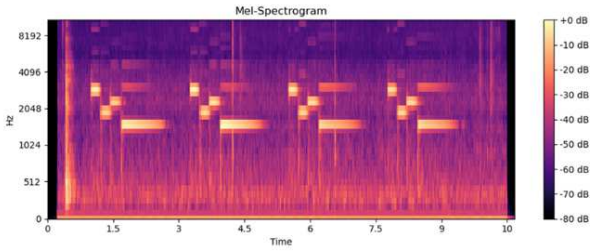

Fig. 3 An example of Mel-Spectrogram

The human ear perceives sound non-linearly. Mel-Scale is the expression of the relationship between the physical frequency and the frequency perceived by a person by reflecting the characteristics of the human hearing organ. Mel-Spectrogram is to obtain frequency components of sound using Mel-Scale.

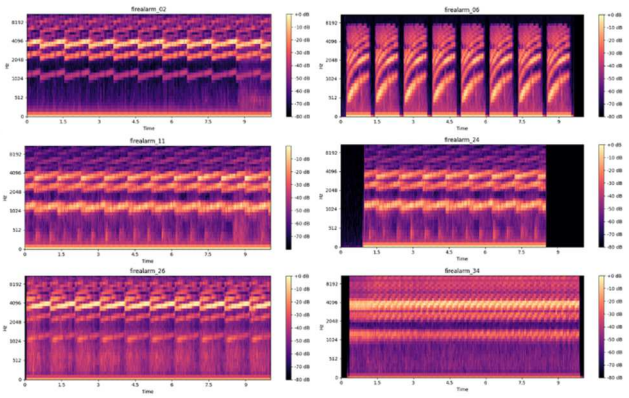

Fig. 4 Mel-Spectrogram of fire alarm sound

Figure 4 is an image converted from a fire alarm sound into Mel-Spectrogram. As shown in the figure, it can be seen that the shape of the image of the same type of sound is quite similar. Mel-Spectrogram can be thought of as a kind of numerical value because it is created based on frequency components. Therefore, sound classification is possible through an image-based discrimination framework through imaging of frequency components.

1) Mel-Spectrogram generation: Figure 5 shows the process of generating Mel-Spectrogram from input sound data. The intensity of the y-axis signal according to the x-axis time is stored in wav format through the microphone. The wav file performs the resampling and normalization processes using python's librosa library. Next, when FFT is performed, the time domain is changed to the frequency domain. In order to consider time together, a three-dimensional heat map data called spectrogram is used by adding a time dimension (frame). Finally, apply a log scale to create a Mel-spectrogram. In the log scale process, the Mel-Scale process proceeds [11].

Fig. 5 Mel-Spectrogram generation process

2) Unifying data length: Even if the sound is of the same type, if the length of the recorded sound is different, the generated Mel-Spectrogram is converted as if it were a different image. The use of these data can greatly affect the sound classification results. Figures 6 and Fig. 7 show that the generated images are different when the sound is the same, but the length of the recorded sound is different.

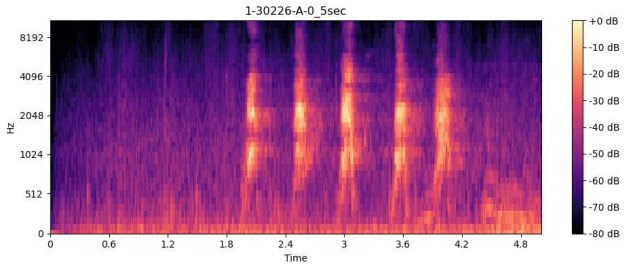

Fig. 6 Conversion result of 5 seconds sound data

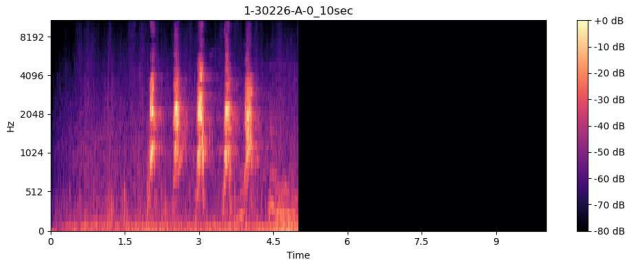

Fig. 7 Conversion result of 10 seconds of sound data

If the initial sound data length was ignored and learned, the accuracy was very low. In order to solve this problem, in this study, a pre-work that equalizes the length of the recorded sound data was applied. Figure 8 is the result of editing a baby's cry into 10 seconds and converting it into Mel-Spectrogram. In order to improve performance, the accuracy could be improved by editing all data to 10 seconds and training.

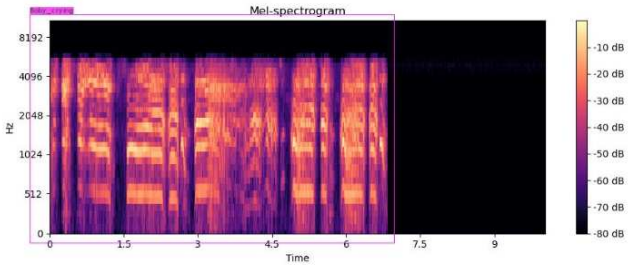

Fig. 8 Normalization result of baby crying sound data

C. Deep-learning

1) YOLO learning model: YOLO is a representative single-step object detection algorithm [28], [29]. YOLO has several advantages. First, it's very fast. It does not require complex pipelines and can execute neural networks with new images at test time. Second, we learn the general features of the object. With the generalization of objects, learning is better performed compared to powerful detection methods such as R-CNN. Since YOLO is powerful in generalization, it is less likely to misclassify or malfunction even if a new domain or unexpected input comes in. The proposed system requires a fast processing speed in the process of deriving a result value after recording for 10 seconds. The YOLO model was introduced to take advantage of these advantages.

2) Training data composition: Table 2 shows the training data configuration for the learning model of this study. In Table 2, time (s) is the sum of the total time of sound sources recorded for each class, Size(MB) is the total file size of these sound sources, and number of image files means the number of preprocessed image files.

TABLE II
COMPOSITION OF TRAINING DATA

Class	Time(s)	Size(MB)	Numbers of image files
Baby_crying	1130	17.25	113
Door_bell	910	129	91
Washer_Alarm	680	121	68
Fire_Alarm	430	74.9	43
Breathing	690	89.4	69
Clapping	680	103	68
Coughing	700	95.4	70
Class_breaking	660	103	66
Snoring	700	86	70
Dog	1220	180	122
Door_Lock	720	84.4	72
Washing_Machine	760	67.9	76
Vaccum_Cleaner	730	99.7	73
Total	10010	1250.95	1001

3) Test data composition: In order to evaluate the performance of the proposed sound analysis method, evaluation data were separately generated for each class. In addition, since the classification of sounds that can occur in daily life must be made, ambient noise will affect it. In order to check how the performance of the proposed system is affected in such a situation, two sounds corresponding to each class were randomly selected, and 100 composite sound files played at the same time were created and applied to the experiment.

TABLE III
COMPOSITION OF TEST DATA

Class	Numbers of image files
Baby_crying	42
Door_bell	31
Washer_Alarm	30
Fire_Alarm	20
Breathing	2
Clapping	6
Coughing	8
Class_breaking	11
Snoring	8
Dog	15
Door_Lock	5
Washing_Machine	4
Vaccum_Cleaner	9
Total	191

D. User Interface

1) Smartphone application: In this study, sound recording and primary alarm provision are made through a smartphone. Figure 9 shows the main screen of the smartphone application implemented in this paper. The circular shape on the left side of the logo of the application shown in the picture is a double expression of the shape of the bracelet and the ear. It also has the meaning of helping enough to replace the function of the ear through the bracelet. The zigzag shape on the right represents the vibration of the vibrating bracelet, showing that it takes the place of sound through other senses. There is a button to connect Bluetooth in the upper right part of the main screen, and through this, it is linked with the vibration bracelet. It is set to continuously measure decibels using a thread and record for 10 seconds when a certain decibel is exceeded. The recorded sound is stored in the database, and the result of sound classification is sent back from the server.

Fig. 9 Main screen of application

In the application that received the result, a notification appears on the top bar as shown in Fig. 10, and when the notification is pressed, an icon image of the corresponding sound is displayed in a large size. At the same time, the bracelet generates vibrations with different frequencies and vibration lengths corresponding to sound. In addition, a recording window to record important sounds and an additional setting window are created so that it can be used conveniently within the application.

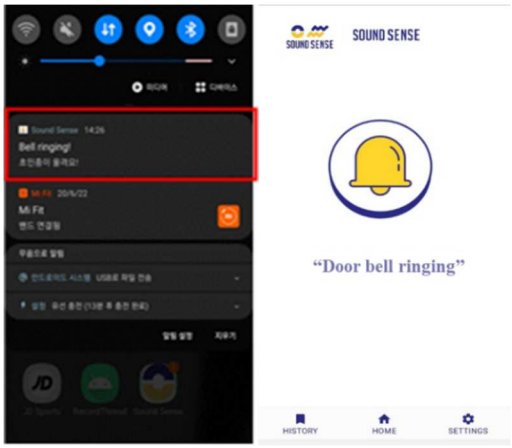

Fig. 10 Information screens of alarm

2) Vibration bracelet: The vibration bracelet can communicate with the application using the Arduino Nano and Bluetooth module. A vibration motor was used to transmit vibration to the skin sense. The vibration bracelet was implemented to vibrate for a total of four sounds that can frequently occur in daily life, and a frame was manufactured using a 3D printer, as shown in Fig. 11.

Fig. 11 Implemented vibration bracelet

III. Results and Discussion

To evaluate the performance of the proposed sound analysis network, an experiment was conducted on a computer equipped with Windows 10 operating system, i7-8700K 3.70GHz CPU, 23GB RAM, and GTX 1070 Ti GPU.

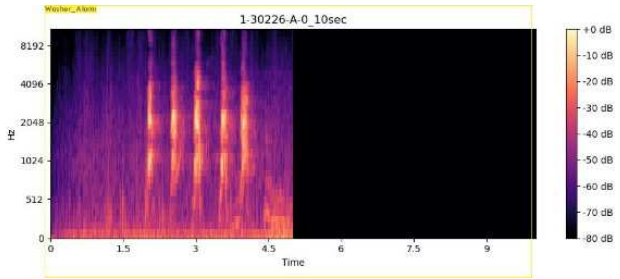

Fig. 12 Information screens of alarm

A recognition experiment was conducted after learning a deep learning model using only four target sounds in the first experiment. However, as shown in Fig. 12, when external noise was recorded together, there was a problem in that the false recognition rate was increased. The sound analysis results in this study finally aim to accurately classify the four types of sounds that occur in life. Therefore, if external noise is included, it is highly likely to affect learning and evaluation results adversely. Therefore, for the final experiment, a class was added by selecting sounds generated in the indoor environment among the ESC (Environmental Sound Classification)-50 dataset [30], and it was integrated with four classes corresponding to the final goal. A training dataset was created by selecting 60 pieces of data per class and trained 30000 times. A test dataset was created that was randomly selected among data excluding the training dataset, and classification accuracy was measured for all test datasets. As a result of the experiment, by including a countermeasure against external noise, it was possible to reduce the false classification rate. In addition, as shown in Fig. 13, even when two sounds are mixed, both sounds were recognized separately.

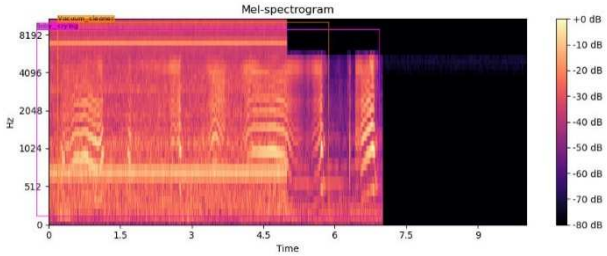

Fig. 13 Information screens of alarm

As a result of the classification accuracy measurement, when the number of weights to be trained in the YOLO learning model was tested in units of 1000 from 20000 to 30000, the accuracy was about 85% on average. When the training weights were set to 29000, the performance was the best, and the classification accuracy was 88.18%. Table 4 shows some of the classification results in the experiment conducted while changing the learning weight.

TABLE IV
CLASSIFICATION ACCURACY ACCORDING TO VARYING WEIGHTS

Weights	Accuracy(%)
2100	80.33
2700	85.43
2900	88.18

In order to evaluate the performance when various sounds are mixed and recorded in daily life, a complex sound source file was created and evaluated separately. A total of 1001 training data were learned, and the two were randomly mixed in the evaluation dataset to generate 100 composite sound files, and the performance was evaluated, showing an accuracy of about 80%. When compared with single sound source data, the accuracy is somewhat lower, but it can be meaningful as a result showing the possibility of distinguishing both mixed sounds.

IV. Conclusion

Many people with hearing impairments have difficulty even in their daily lives because they cannot hear the sound. This paper has dealt with ways to solve these difficulties and provide help. In other words, it was proposed to provide information so that the hearing impaired can intuitively recognize by classifying four sounds, including crying baby sounds, doorbell sounds, washing machine end sounds, and fire alarm sounds, which are important but easy to miss for a normal life.

In the proposed method, the application first records the sound over 60 decibels, and then transmits it to the server. The server that receives the recorded sound classifies the sound and delivers the classification result to the application. The application notifies the hearing impaired to recognize information through visual information and a vibration bracelet intuitively. The vibration was easily recognized by varying the number of vibrations and vibration periods for each sound. Since the sound to be classified in the paper is likely to deteriorate classification performance due to the

influence of other external noises, a learning dataset including everyday sounds was created in addition to the classification target of four sounds, and through learning, sounds containing external noise were also classified. In learning, sound data was visualized as Mel-Spectrogram and learned using the YOLO classification model.

The test was conducted using separate image data that was not included in the learning. The appropriate number of weights was derived by comparing the performance according to the number of learning weights in this process. The experiment confirmed that the accuracy of the total data was about 85% on average, which was sufficient to apply to the auxiliary system for the hearing impaired proposed in this paper. In addition, it was confirmed that the classification accuracy was about 80% in the experiment conducted by separately generating data mixed with two sounds. Additional performance improvement can be expected if the basic learning dataset is strengthened through future research and learning by adding sound files and complex sound files for living noise.

REFERENCES

[1] Registration Status for Diabled, Website of the Ministry of Health and Welfare in Korea, viewed at May 10, 2021. Available: http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&CONT_SEQ=365339.

[2] "2017 White Paper for the Disabled," Korea Disabled People's Development Institute, 2017.

[3] Rodriguez-Villarreal, Kevin, et al., "Development of Warning Device in Risk Situations for Children with Hearing Impairment at Low Cost," Development, vol.10, no.11, 2019.

[4] Hu, Menghan, et al., "An overview of assistive devices for blind and visually impaired people," International Journal of Robotics and Automation, vol.34, no.5, pp. 580-598, 2019.

[5] Saleem, Muhammad Imran, et al., "Full Duplex Smart System for Deaf & Dumb and Normal People," 2020 Global Conference on Wireless and Optical Technologies (GCWOT). IEEE, 2020.

[6] Tapu, Ruxandra, Bogdan Mocanu, and Titus Zaharia., "Wearable assistive devices for visually impaired: A state of the art survey," Pattern Recognition Letters, vol.137, pp. 37-52, 2020.

[7] Abdelmagid, Fatima, Hamda Fasla, and Mourad Elhadef, "Jusoor: A Wearable Communication Device for the Deaf-Blind: An Ideation-Themed Capstone Project," The 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research. 2021.

[8] Yağanoğlu, M., "Real time wearable speech recognition system for deaf persons," Computers & Electrical Engineering, vol.91, p.107026, 2021.

[9] Fang, Wei, Lin Wang, and Peiming Ren, "Tinier-YOLO: A real-time object detection method for constrained environments," IEEE Access, vol.8 pp. 1935-1944, 2019.

[10] Wu, Dihua, et al., "Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments," Computers and Electronics in Agriculture, vol.178, pp.105742, 2020.

[11] Rajagukguk, Juniastel, and Nurdieni Eka Sari, "Detection system of sound noise level (SNL) based on condenser microphone sensor," Journal of Physics, vol.970, no.1, IOP Publishing, 2018.

[12] H. Y. Oh, H. Kwon, G. Kwon, "Developing Application for The Hearing Impaired Using the Synthesized Automaton," in Proceeding of Korea Computer Congress, pp.2031-2033, 2015.

[13] D. K. Heo, B. K. Lee, S. J. Lee, Y. J. Nam, J. H. Kwon, B. S. Song, "Development of An Assistive Device for Sound Information Transmission for Hearing Impaired People," in Proceedings of the Korean Society of Rehabilitation and Welfare Engineering Conference, pp. 18-20, 2015.

[14] Ji, Chunyan, et al., "A review of infant cry analysis and classification," EURASIP Journal on Audio, Speech, and Music Proc., pp. 1-17, 2021.

[15] Burileanu, C., "Recent Experiments and Findings in Baby Cry Classification," Future Access Enablers for Ubiquitous and Intelligent Infrastructures: Third International Conference, FABULOUS 2017, Bucharest, Romania, October 12-14, 2017, Proceedings. vol. 241, Springer, 2018.

[16] Alsouda, Y., Pllana, S., Kurti, A., "Iot-based urban noise identification using machine learning: performance of SVM, KNN, bagging, and random forest," in Proceedings of the international conference on omni-layer intelligent systems, 2019, p. 62-67.

[17] Alsouda, Y., Pllana, S., Kurti, A., "A machine learning driven IoT solution for noise classification in smart cities," arXiv preprint arXiv:1809.00238, 2018.

[18] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Processing Letters, vol.24, no.3, pp. 279-283, 2017.

[19] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep Convolutional Neural Networks with Mixup for Environmental Sound Classification," in Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision, pp. 356-367, 2018.

[20] V. Boddapati, A. Petef, and L. Lundberg, "Classifying Environmental Sounds Using Image Recognition networks," Procedia Computer Science, vol.112, pp.2048-2056, 2017.

[21] Khawas, Chunnu, and Pritam Shah, "Application of firebase in android app development-a study," International Journal of Computer Applications. vol.179, no.46, pp. 49-53, 2018.

[22] V. Bisot, S. Essid and G. Richard, "HOG and subband power distribution image features for acoustic scene classification," in Proceeding of EUSIPCO, pp. 719-723, 2015.

[23] A. Rakotomamonjy and G. Gasso, "Histogram of Gradients of Time-Frequency Representations for Audio Scene Classification," IEEE/ACM Trans. Audio, Speech, and Language Process, vol.23, no.1, pp. 142-153, 2015.

[24] Dörfler, Monika, Roswitha Bammer, and Thomas Grill, "Inside the spectrogram: Convolutional Neural Networks in audio processing," in Proceeding of IEEE International Conference on Sampling Theory and Applications, 2017.

[25] Dong, Mingwen. "Convolutional neural network achieves human-level accuracy in music genre classification," arXiv preprint arXiv:1802.09697, 2018.

[26] Zhou, Quan, et al. "Cough recognition based on mel-spectrogram and convolutional neural network," Frontiers in Robotics and AI, vol.8, 2021.

[27] Mushtaq, Z., and Su, S. F., "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," Applied Acoustics, vol.167, 2020.

[28] Redmon, Joseph, and A. Farhadi, "YOLO9000: better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263-7271.

[29] Wu, Xifang, et al, "Real-time vehicle color recognition based on yolo9000," in International Conference in Communications, Signal Processing, and Systems, Springer, Singapore, 2018. p. 82-89.

[30] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1015-1018, 2015.