Speech Classification to Recognize Emotion Using Artificial Neural Network

Siti Helmiyah[1*], Imam Riadi[2], Rusydi Umar[1], Abdullah Hanif[1]

[1]Program Studi Teknik Informatika
Universitas Ahmad Dahlan Yogyakarta
Yogyakarta, Indonesia
[2]Program Studi Sistem Informasi
Universitas Ahmad Dahlan Yogyakarta
Yogyakarta, Indonesia
*siti1708048022@webmail.uad.ac.id

Abstract-This study seeks to identify human emotions using artificial neural networks. Emotions are difficult to understand and hard to measure quantitatively. Emotions may be reflected in facial expressions and voice tone. Voice contains unique physical properties for every speaker. Everyone has different timbres, pitch, tempo, and rhythm. The geographical living area may affect how someone pronounces words and reveals certain emotions. The identification of human emotions is useful in the field of human-computer interaction. It helps develop the interface of software that is applicable in community service centers, banks, education, and others. This research proceeds in three stages, namely data collection, feature extraction, and classification. We obtain data in the form of audio files from the Berlin Emo-DB database. The files contain human voices that express five sets of emotions: angry, bored, happy, neutral, and sad. Feature extraction applies to all audio files using the method of Mel Frequency Cepstrum Coefficient (MFCC). The classification uses Multi-Layer Perceptron (MLP), which is one of the artificial neural network methods. The MLP classification proceeds in two stages, namely the training and the testing phase. MLP classification results in good emotion recognition. Classification using 100 hidden layer nodes gives an average accuracy of 72.80%, an average precision of 68.64%, an average recall of 69.40%, and an average F1-score of 67.44%.

1. Introduction

Emotions are psychological fluctuations that develop in a person to respond to internal or external stimuli. Emotion is part of the human body that comes out into expression [1. Emotions are very difficult to measure from a quantitative viewpoint [2]. According to [3], there are five basic types of emotions, namely anger, happiness, sadness, fear, and disgust, which are not easy to measure. Emotion is typically accompanied by physiological and behavioral changes in the body. Emotion may appear in facial expression and voice speech. When a person's emotion changes, his facial expression changes. Therefore, the face is a good probe to measure emotional state. Detecting emotion in speech is more complicated. Speech contains both emotional information and linguistic messages. Voice is a characteristic of a person. Voice is a form of someone's expression of a situation. Everyone's voice has different

vocals, rhythms, tempo, and stress. The difference in the characteristics of a person's voice is influenced by the language spoken and the area of residence.

Human emotion recognition or Speech Emotion Recognition (SER) is an active research topic [4]. The identification of human emotions is useful in the field of human-computer interaction. It helps develop the interface of software that is applicable in community service centers, banks, and education and others. The main component of emotion recognition is feature extraction and classification [5]. Selection of the right feature extraction algorithm is the main focus of SER activities. In turn, a suitable classification algorithm allows producing optimal emotional recognition.

Research on sound extraction of speech has recently focused on prosodic and spectral features. The observed characteristics include the pitch of the voice, length of sounds, loudness, and timbre. A study to recognize human

voice speech emotions used Hidden Markov Model (HMM) to analyze the pitch, energy, and formant [5]. Another research extracted energy feature, pitch, ZCC, and entropy and examined the extracted measure using the Mel Frequency Cepstrum Coefficients (MFCC) and the K-Nearest Neighbor (KNN) classification [6]. Research by [4] used MFCC feature extraction and Modulation Spectral Features (MSFs) and analyzed the result using Multivariate Linear Regression (MLR) and Support Vector Machine (SVM).

Several SER studies used available databases as datasets in emotion classification. [5] used HMM algorithm and a database of recorded sound film to recognize angry, happy and neutral emotion. [7] used the Berlin Emo-DB database to create an emotion recognition system employing MFCC and SVM. [8] compared the accuracy of three databases: Berlin Emo-DB, SAVEE, and TESS in recognizing emotions. Eventually, [9] developed a real-time emotion recognition system using the RAVDESS and SAVEE databases as input to the training module.

This paper discusses the results of research on speech classification for emotion recognition. The research attempts to identify human emotions in speech using the Berlin Emo-DB emotion database. The study uses the MFCC method for feature extraction and uses the Artificial Neural Network's Multi-Layer Perceptron (MLP) method for emotional classification. MFCC method finds a good application in Emotion Recognition Systems. MFCC is superior for speaker identification and emotion recognition because its work similarly resembles how the human ear works [7]. On the other hand, the MLP method is suitable to carry out a directed learning process for pattern recognition [10].

The research utilizes the Scikit-Learn module on the Python programming language platform. Scikit-learn integrates various machine learning algorithms for supervised and unsupervised learning problems [11]. The module facilitates the study in speech processing and recognition. The research has assumed that the module can produce good speech recognition. The study has continued the previous works that have been reported elsewhere [12]. Further study is possible that examines various methods of feature extraction and classification or adding other feature extraction and classification methods [13].

2. Methods

The process of recognizing speech emotions for classification of emotions in this study follows the stages presented in Figure 1. The most important main processes are feature extraction and classification. The results of feature extraction are very influential in determining the results of the recognition of speech emotions at the classification stage. The data used in this study are speech sounds stored in the Berlin Database of Emotional Speech Berlin Emo-DB. The features of the sound samples from the database were extracted using the spectral method,

namely MFCC. The results of feature extraction are used in the classification process using Artificial Neural Networks with the Multi-Layer Perceptron model.
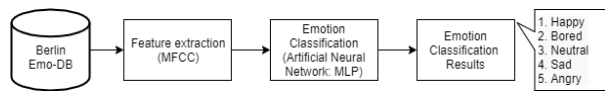
Gambar 1. Research Flow

a. Emotion Database

The data used in this study uses the Berlin Database of Emotional Speech Berlin Emo-DB is a part of a DFG SE462 / 3-1 research project in 1997 and 1999. The project director is Prof. Dr. W. Sendlmeier from the Technical University of Berlin, Institute of Speech and Communication, Department of Communication Science. The project members include Felix Burkhardt, Miriam Kienast, Astrid Paeschke, and Benjamin Weiss [14].

The emotions in this database are anger, boredom, disgust, fear, happiness, sadness and neutrality. The recording of the database creation was carried out in the anechoic room at the Technical University of Berlin, Department of Technical Acoustics. Actors were selected through a selection that resulted in 10 actors. The choice of sentences for pronunciation was also considered in the creation of this database. The spoken sentence is a normal sentence that is used everyday. This selection is useful so that the actor can pronounce it naturally without making up. This database contains about 500 samples that have a level of recognition of human emotions that is assessed according to the results of the study [14].

This study only took five types of emotions, namely anger, boredom, happiness, neutrality, and sadness, which were said by ten different actors. The study used 420 emotional data consisting of 127 data of angry emotions, 81 data of bored emotions, 71 data of happy emotions, 79 data of neutral emotions, and 62 data of sad emotions. All audio data is extracted from the Berlin Emo-DB. The audio data file provided by Berlin Emo-DB has an uneven number of emotions.

b. Feature Extraction

Determination of feature extraction is the most important stage in the recognition of speech emotions. A good extraction process can distinguish different feature patterns from one emotion class to another. The extraction of emotional features can be divided into three categories, namely prosodic features, spectral features, and sound quality features [15]. Prosidis features include frequency, duration, energy, pitch, and formant. Spectral features include LPC (Linear Predictive Coding), LPCC (Linear Predictor Ceptral Coefficients), MSF (Modulation Spectral Feature), MFCC, ZCPA (Zero Crossings with Peak Amplitudes), and others. Sound quality features include frequency and bandwidth formats, shimmer, jitter and more. The results of feature extraction are further

processed to obtain statistical values such as max, min, mean, median, kurtosis and skewness [16]. The statistical value can be used for the classification process.

MFCC is a parametric representation of speech signals applied to speech speech recognition and is becoming popularly used for voice identification and emotion recognition [7] [17] [18] [19]. By applying cepstral analysis, MFCC tries to mimic the workings of the human hearing organ [19] [20]. MFCC is a coefficient representing perceptual sound with logarithmic frequency bands that mimic human vocals [21]. According to [13] the most commonly used number of coefficients is the coefficient of 20, the use of a coefficient of 10-12 is considered sufficient depending on the spectral shape. Research [22] on speech recognition uses a factor of the number of coefficients between 9 and 13, and research [19] uses a number of coefficients of 13. The results of feature extraction are then searched for statistical values such as mean, STD, max, min, kurtosis, skewness, and median. 16]. The aim is to reduce the value obtained from the MFCC extraction process [21].

c. Emotion Classification Using Artificial Neural Networks: Multi-Layer Perceptron (MLP)

Emotion classification is an important step in recognizing emotions according to each emotion class. The Multi-Layer Perceptron is a feedforward network model consisting of several neurons connected by neuron connecting weights [23]. The neurons are arranged in a layer consisting of an input layer, one or more hidden layers, and an output layer [24]. The MLP learning process updates the return weight (backpropagation). This weight update is carried out to find the most optimal value to produce the correct classification results.

The MLP classification process is divided into two stages, namely the training stage and the testing phase. These two stages are carried out on different data, namely dividing the research data into two, namely training data and testing data. The training stage is carried out for network initialization and formation, namely to determine the number of input layers, hidden layers, and output layers. During the training phase, parameters such as learning rate, number of iterations, and error threshold were determined. The training stage uses a lot of data to produce a good classification pattern. The testing phase is carried out to find out the classification results that have been made at the training stage whether it provides correct and accurate classification results.

3. Result

a. MFCC Feature Extraction

Figure 2 shows an example of a voice signal which is the raw data in this study, namely data that has not gone through feature extraction. The signal is a typical wave of angry emotional signals uttered by an actor. The horizontal axis on the graph is time and the vertical axis is the power of sound (amplitude).
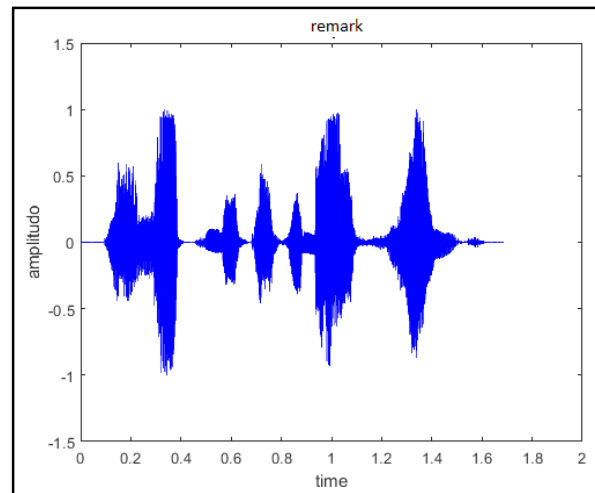
Gambar 2. An example of data is in the form of a voice signal

Data as in Figure 2 is processed using MFCC with a coefficient of 13. The process of feature extraction is carried out using modules from the Python programming language [25]. The results of feature extraction are processed to determine the mean value which in turn is used for the classification process of angry, bored, happy, neutral, and sad emotions using the MLP method. The value of the MFCC feature extraction results for some voice data can be seen in Figure 3. The extraction results for each data signal are 13 numbers, which are the MFCC coefficient.

Gambar 3. MFCC Feature Extraction for 5 voice data samples

b. Emotion Classification

This study uses 420 emotional data, with 127 data on angry emotions, 81 data for bored emotions, 71 data for happy emotions, 79 data for neutral emotions, and 62 data for sad emotions. Each emotional speech data is processed using MFCC with a coefficient of 13 to get the feature extraction value. From the results of feature extraction, the mean value is determined which is used as input in the artificial neural network using the Multi-Layer Perceptron model. The output of the artificial neural network is five classes of emotions, namely anger, boredom, happiness, neutrality, and sadness. The parameters used in the MLP model can be seen in Table 1. There are three layers used, namely the input layer, one hidden layer, and the output layer. The input node is connected to the MFCC feature

extraction result coefficient 13 and the output node is five emotion classes.

Tabel 1. MLP Method Emotion Classification Parameters

Parameter	Value
Learning rate	0.0001
Error threshold value (epsilon)	1e-08
Epoch	1000
Cross Validation	5 folds
Layer	3 layer
Node input	13
Node Hidden Layer	10
Node output	5

Observations regarding classification performance require training data and test data. Training and testing data were obtained by dividing 420 existing data with a comparison of 80% of data for training and 20% of data for testing. Validation of the performance of the classification process was carried out by using five-fold cross validation. The data presented next is the average of the performance for each fold. Tests were carried out for various numbers of hidden layer nodes. The test results for the number of different hidden layers are presented in Tables 2-6.

Tabel 2. Test Results with Hidden Layer Node = 10

Emotion Class	Accuracy	Precision	Recall	F1-Score	Total
Angry	61.00	84.20	88.40	86.20	32
Bored	68.00	46.20	44.60	45.00	17
Happy	64.00	44.20	36.00	39.60	10
Neutral	62.00	43.00	48.00	45.00	15
Sad	65.00	77.20	72.00	74.40	10
Macro averaged	64.00	58.96	57.80	58.04	84

Table 2 shows that the average accuracy of emotion recognition is 64%, the average precision is 58.96%, the average recall is 57.80% and the average F1-score is 58.04%. The addition of a hidden layer to 30 results in a better classification. Table 3 shows a significant increase in accuracy value to 68%. The same increase occurred for precision and recall so that the F1 score was bigger at 62.84%.

Tabel 3. Test Results with Hidden Layer Node = 30

Emotion Class	Accuracy	Precision	Recall	F1-Score	Total
Angry	74.00	85.20	91.00	88.00	32
Bored	67.00	50.40	43.40	46.20	17
Happy	68.00	50.20	40.00	44.40	10
Neutral	67.00	51.80	57.20	54.20	15
Sad	67.00	79.20	84.00	81.40	10
Macro averaged	68.60	63.36	63.12	62.84	84

Test Results with Hidden Layer Node = 50

Emotion Class	Accuracy	Precision	Recall	F1-Score	Total
Angry	75.00	86.00	85.80	86.00	32
Bored	74.00	58.80	51.80	55.00	17
Happy	64.00	42.40	40.00	40.60	10
Neutral	73.00	58.80	63.80	61.00	15
Sad	64.00	80.60	90.00	84.80	10
Macro averaged	70.00	65.32	66.28	65.48	84

An increase in classification performance was also seen when the number of hidden layer nodes was increased to 50 (see Table 4). Accuracy increased by 1.4 points compared to the classification with 40 hidden layers (Table 3). Meanwhile, the F1-score increased by more than 2.5 points. This shows that the use of the number of hidden layers of 50 significantly results in a better classification than the number of hidden layers of 40. When the number of hidden layers is increased to 80, there is no increase in the performance of the previous classification (see Table 5). The addition of a hidden layer of 30 new nodes only increased accuracy by 1.8% and increased the F1-score by 1.6%.

Tabel 4. Test Results with Node Hidden Layer = 80

Emotion Class	Accuracy	Precision	Recall	F1-Score	Total
Angry	76.00	86.80	89.00	87.80	32
Bored	75.00	59.60	47.00	52.20	17
Happy	71.00	48.60	44.00	45.60	10
Neutral	69.00	60.40	70.60	65.00	15
Sad	68.00	80.40	90.00	84.80	10
Macro averaged	71.80	67.16	68.12	67.08	84

Test Results with Hidden Layer Node = 100

Emotion Class	Accuracy	Precision	Recall	F1-Score	Total
Angry	77.00	85.80	91.00	88.20	32
Bored	67.00	64.40	48.00	52.20	17
Happy	68.00	56.20	42.00	47.20	10
Neutral	77.00	62.20	72.00	66.40	15
Sad	75.00	74.60	94.00	83.20	10
Macro averaged	72.80	68.64	69.40	67.44	84

Testing the MLP model using 100 hidden layer node further improves the emotional classification results (see Table 6). However, the improvement is not sufficiently significant compared to the effort to add such a large number of hidden layer nodes. The accuracy score is 72.8% and it is just 2.8% higher than the accuracy for the 50 hidden layer nodes (see Table 4). The F1-score has increased to 67.44, which means that it is 1.96% higher than that for the 50 hidden layer nodes. The increase in the accuracy, precision, recall and F1-score for all observations with various numbers of hidden layers is presented in Figure 4. The figure confirms a significant increase for various metrics when the hidden layer nodes were increased from 10 to 50. Subsequently, the increase becomes sloping and insignificant when the hidden layer is increased to 80 and to 100. The figure probably suggests that the use of the MLP model is optimal when the number of hidden layers is around 50.
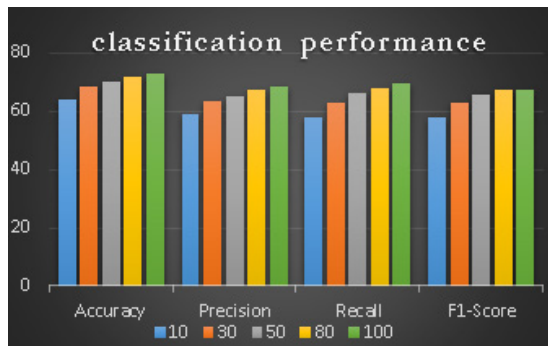
Gambar 4. Testing Comparison Chart

If we look more detail into the recognition results for each type of emotion, the results show different numbers in the accuracy value and F1-score. In the F1-score measurement, the score tends to be high for the recognition of angry and sad emotions. Trends are consistent for MLP models with varying numbers of hidden layers (see again Table 2-6). The F1-score for the recognition of angry emotions was always above 86%, while the recognition for sad emotions ranged from 74% to 84.8%. The recognition of happy emotions shows the lowest F1-score, which is below 50%. Emotion recognition performance measurement using accuracy shows variable metric values and is less consistent when the number of hidden layers is increased. When the number of hidden layers is 10, the highest accuracy value is reached for the recognition of bored emotions. However, when the number of hidden layers was increased to 30, the highest accuracy value was reached for recognition of angry and happy emotions. Furthermore, when the number of hidden layers is between 50 - 80, the highest accuracy value is reached for the introduction of angry and bored emotions. Meanwhile, the model with the number of hidden layers of 100 produces emotion recognition with the highest accuracy for angry and neutral emotions.

The recognition of emotions in this study resulted in better accuracy than the previous study [12] which was only 31.67%. This research proves that the greater the number of hidden layer nodes, the better emotion recognition is. However, the optimal number of hidden layers still needs to be examined in terms of the computation time associated with the acquisition of metric values. This study also noted that the accuracy of emotion recognition varies for the types of emotions identified.

4. Conclusion

In this research, the speech classification process for recognizing emotions starts with feature extraction, which significantly affects the result of emotion recognition. We use Mel Frequency Cepstrum Coefficients (MFCC) for the feature extraction. The next step is emotion classification that uses an artificial neural network called Multilayer Perceptron (MLP). The method produces good speech emotion recognition as indicated by the classification performance measure. In general, the performance measure steadily improves with the increase of the number of hidden layers from 10, 30, 50, 80 to 100. The accuracy attains the highest score of 72.8 when the number of hidden layers is 100. Likewise, the F1-score is at its highest value of 67.44 for the same number of hidden layers. The accuracy of the emotion recognition varies with the varying number of nodes. The F1-score is more consistent in measuring the performance of emotion recognition. Referring to the F1-score value, the model performs better in recognizing angry and sad, but it performs poorly in recognizing happy emotions.

Reference

[1]   A. B. Gumelar et al., "Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks," in 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), 2019, pp. 1–8.

[2]   S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech Emotion Recognition," 2014 Int. Conf. Adv. Electron. Comput. Commun. ICAECC 2014, vol. 7, 2015.

[3]   M. D. Pell and S. A. Kotz, "On the Time Course of Vocal Emotion Recognition," PLoS One, vol. 6, no. 11, pp. 1–16, 2011.

[4]   L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "Speech Emotion Recognition: Methods and Cases Study," Proc. 10th Int. Conf. Agents Artif. Intell., vol. 2, no. Icaart, pp. 175–

182, 2018.

[5]   B. H. Prasetio, W. Kurniawan, and M. H. H. Ichsan, "Pengenalan Emosi Berdasarkan Suara Menggunakan Algoritma HMM," J. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 3, pp. 168–172, 2017.

[6]   A. Bombatkar, G. Bhoyar, K. Morjani, S. Gautam, and V. Gupta, "Emotion Recognition Using Speech Processing Using K- Nearest Neighbor Algorithm," Int. J. Eng. Res. Appl., pp. 2248–9622, 2014.

[7]   A. Watile, V. Alagdeve, and S. Jain, "Emotion Recognition in Speech by MFCC and SVM," Int. J. Sci. Eng. Technol. Res., vol. 6, no. 3, pp. 404–407, 2017.

[8]   A. R. Choudhury, A. Ghosh, R. Pandey, and S. Barman, "Emotion Recognition from Speech Signals using Excitation Source and Spectral Features," in Proceedings of 2018 IEEE Applied Signal Processing Conference, ASPCON 2018, 2018, pp. 257–261.

[9]   A. Iqbal and K. Barua, "A Real-time Emotion Recognition from Speech using Gradient Boosting," 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019, pp. 1–5, 2019.

[10]   M. Mentari, E. K. R. Sari, and S. Mutrofin, "Klasifikasi Menggunakan Kombinasi Multilayer Perceptron dan Alignment Particle Swarm Optimization," in Seminar Nasional Teknologi Informasi & Komputasi (SENASTIK), 2014, no. October 2014, pp. 47–54.

[11]   F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, no. January, pp. 2825–2830, 2011.

[12]   S. Helmiyah, I. Riadi, R. Umar, A. Hanif, A. Yudhana, and A. Fadlil, "Identifikasi Emosi Manusia Berdasarkan Ucapan Menggunakan Metode Ekstraksi Ciri LPC dan Metode Euclidean Distance," J. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 6, pp. 1177–1186, Dec. 2020.

[13]   Y. Kumar and M. Mahajan, "Machine Learning Based Speech Emotions Recognition System," Int. J. Sci. Technol. Res., vol. 8, no. 7, pp. 722–729, 2019.

[14]   F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in Ninth European Conference on Speech Communication and Technology, 2005, pp. 1517–1520.

[15]   G. Liu, W. He, and B. Jin, "Feature Fusion of Speech Emotion Recognition Based on Deep Learning," in 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), 2018, pp. 193–197.

[16]   Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local Hu moments for speech emotion recognition," Biomed. Signal Process. Control, vol. 18, pp. 80–90, 2015.

[17]   K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech Emotion Recognition Using Fourier Parameters," IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 69–75, 2015.

[18]   S. S. S and V. N. Nitnaware, "Emotion Speech Recognition using MFCC and SVM," Int. J. Eng. Res., vol. 4, no. 06, pp. 1067–1070, 2015.

[19]   K. V. Krishna Kishore and P. Krishna Satish, "Emotion Recognition in Speech using MFCC and Wavelet Features," in Proceedings of the 2013 3rd IEEE International Advance Computing Conference, IACC 2013, 2013, pp. 842–847.

[20]   G. Liu, W. He, and B. Jin, "Feature Fusion of Speech Emotion Recognition Based on Deep Learning," in 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), 2018, pp. 193–197.

[21]   M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech Based Human Emotion Recognition Using MFCC," in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 2257–2260.

[22]   P. R. Chaudhari and J. S. R. Alex, "Selection of Features for Emotion Recognition from Speech," Indian J. Sci. Technol., vol. 9, no. 39, pp. 1–5, 2016.

[23]   M. Irfan, B. A. Ardi Sumbodo, and I. Candradewi, "Sistem Klasifikasi Kendaraan Berbasis Pengolahan Citra Digital dengan Metode Multilayer Perceptron," IJEIS (Indonesian J. Electron. Instrum. Syst., vol. 7, no. 2, p. 139, 2017.

[24]   S. Motamed, S. Setayeshi, and A. Rabiee, "Speech Emotion Recognition Based on a Modified Brain Emotional Learning Model," Biol. Inspired Cogn. Archit., vol. 19, pp. 32–38, Jan. 2017.

[25]   B. Mcfee et al., "librosa : Audio and Music Signal Analysis in Python," in Proceedings of the 14th python in science conference, 2015, pp. 18–25.