Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 03
Issue 02

December, 2022

Research ***Article***

CREMA-D: Improving Accuracy with BPSO-Based Feature Selection for Emotion Recognition Using Speech

Kenan DONUK[1]

[1] *Cizre Vocational School, Computer Programming Department. Sirnak University, 73200, Sirnak, Türkiye*

ARTICLE INFO

Article history:
Received December 4, 2022
Revised December 15, 2022
Accepted December 21, 2022

Keywords:
BPSO Algorithm
Crema-D
CNN
SVM Algorithm
Speech Emotion

ABSTRACT

People mostly communicate through speech or facial expressions. People's feelings and thoughts are reflected in their faces and speech. This phenomenon is an important tool for people to empathize when communicating with each other. Today, human emotions can be recognized automatically with the help of artificial intelligence systems. Automatic recognition of emotions can increase productivity in all areas including virtual reality, psychology, behavior modeling, in short, human-computer interaction. In this study, we propose a method based on improving the accuracy of emotion recognition using speech data. In this method, new features are determined using convolutional neural networks from MFCC coefficient matrices of speech records in Crema-D dataset. By applying particle swarm optimization to the features obtained, the accuracy was increased by selecting the features that are important for speech emotion classification. In addition, 64 attributes used for each record were reduced to 33 attributes. In the test results, 62.86% accuracy was obtained with CNN, 63.93% accuracy with SVM and 66.01% accuracy with CNN+BPSO+SVM.

1. Introduction

Emotions are an important tool of human communication. People's cognitive and mental states manifest themselves by being reflected in their face, eyes, voice, or body with many different emotions such as surprise, disgust, fear, anger, sadness, happiness. By integrating interactive human-computer systems that continuously and automatically recognize people's emotional states with intelligent systems, these systems can be made more interactive. Emotion recognition using speech, the primary communication channel through which emotional states are conveyed, is an active area of research. Systems for speech emotion recognition (SER) are used in many different domains, e.g., call centers [1], criminal case [2]. Some studies on SER methods in the literature are as follows. Zielonka et al. They used Crema-D, RAVDESS, SAVEE, TESS and IEMOCAP datasets in their study. In their study, the performance difference of spectrogram and mel-spectrogram features used to extract emotion-representing features was compared over Resnet-18 and a custom CNN model. As a result of the comparisons, it has been shown that mel-spectrograms are more suitable for classical CNN-based education. For the Crema-D dataset, they achieved an accuracy of 46.75% in the spectrogram feature use and 53.66% in the mel-spectrogram feature test results in the classical CNN architecture test results [3]. Shankar et al. conducted a study on

[1]Corresponding author
e-mail: kenandonuk@sirnak.edu.tr
DOI: 10.55195/jscai.1214312

performance among model architectures of data augmentation. They used Gated-CNN, MLP-mixer, Bi-LSTM, Transformer architectures in their work. In the training of model architectures, the optimization hyper-parameters of the models were adjusted using the VESUS dataset. Then, the performances were evaluated by using 5-fold cross validation with IEMOCAP and Crema-D datasets, on which different data augmentation methods were applied on the models. It has been shown that speed perturbation, one of the data augmentation types, is a robust data augmentation strategy in increasing accuracy [4]. Donuk and Hanbay obtained the zcr, rmse and mfcc features of the audio signals of the Ravdess and Tess datasets. Considering the change of these features in an audio signal over time, they performed an LSTM-based classification [5]. In their study, Singh and Goel conducted a literature review on the databases used in speech emotion recognition studies between 2000-2021 and the motivations and limitations of deep learning used in speech emotion classification. Deep learning studies on speech emotion recognition were systematically summarized by examining 152 articles [6].

2. Material and Method

2.1. Crema-D Dataset

The Crowd-sourced Emotional Multi-modal Actors Dataset, Crema-D dataset contains 7442 clips created by theater directors for a total of 91 actors, 48 males and 43 females, ranging in age from 20 to 74 and of various ethnicities. Actors were asked to speak 12 specific phrases in one of six different emotions (Disgust, Fear, Sad, Anger, Neutral, Happy) and in three different intonations (Low, Medium, High). The recordings (auditory, visual, and audiovisual) were rated by 2,443 raters via crowdsourcing based on the intensity of the emotion and feeling. Human recognition of recordings made in three different modalities as auditory, visual, and audiovisual has a hit rate of 40.9%, 58.2%, and 63.6%, respectively [7]. The intensities of speech emotion labels belonging to the Crema-D data set used in our study are shown in the graph given in Figure 1.

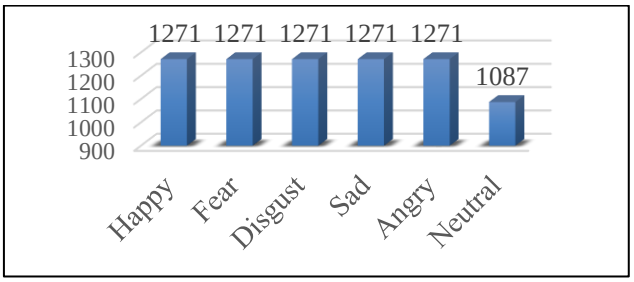

Figure 1 Crema-D dataset emotion distribution

Audio data of speech emotions are stored in digital media by analog-digital conversion. While performing this cycle, the quality of the signal is determined depending on the number of samples per second. In addition, a quantitative dimension of each sound sample of the signal is determined. In Figure 2, a digital representation of a sound with a sampling frequency of 22050 is given.

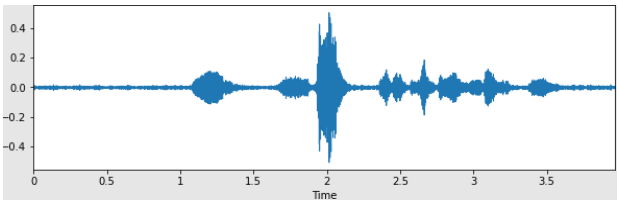

Figure 2 Audio signal

2.2. Data Preprocessing

It has been reported that the classification using the available sample values of the audio signals has little contribution to the performance of the model used [8]. For this reason, acoustic features that best represent emotions are needed. Some preprocessing steps are required to extract the appropriate features. These can be summarized as extracting the silent parts with low amplitude values from the audio signal, converting the obtained signals to the appropriate format for the model to be used in the classification, and finally extracting the features from the signals.

2.3. Data Cleaning and Editing

Before extracting the voice features from the speech recordings of the Crema-D dataset, the silent parts of the wave graphs that we think do not represent the current emotion were clipped using the librosa [9] library. With this process, emotions will be better represented while classifying from attributes. Figure 3 shows the original and clipped state of the signal to the "angry" sound sense.

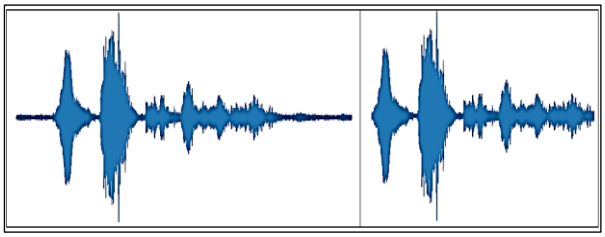

Figure 3 Cropped audio signal

To extract features with the same spatial dimensions from the audio signals obtained after clipping, all audio recordings were arranged to have the same sample size.

2.4. Feature Extraction

Acoustic sound features to represent speech emotions are needed to be used in classification from data set signals converted into a format suitable for feature extraction. There are different attributes representing sound in the literature. These can be listed as Zero Crossing Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Flux, Mel Frequency Cepstral Coefficients, Chroma Vector, Chroma Deviation. These attributes can be used together as well as alone. Mel Frequency Cepstral Coefficients (MFCCs), which are widely used among these attributes, were used in our study.

2.4.1. Mel-frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) are the most used features to represent the sound with a certain number of coefficients. To extract the MFCCs features, the audio signal needs to undergo a series of processing. First, the preprocessed audio signals are divided into frames with the same number of samples. With the division process, it is aimed to extract more consistent features from the audio signals. The number of frames to be obtained can be obtained with the formula given in Equation 1.

$$\left(\frac{Number\ of\ samples - Frame\ length}{Hop\ length}\right) + 1 \quad (1)$$

Hamming windowing is applied for each frame of the audio signals split into frames. With Hamming windowing, the frequency spectrum to be obtained from audio signals is improved. Thus, spectral leakage due to discrete frames of the audio signal is prevented. After this step, Fast Fourier Transform (FFT) is applied for each of the framed signals.

With FFT, the audio signal is passed from the time domain to the frequency domain. Thus, the amplitude and frequency values that make up the audio signal are extracted. FFT formula is given in Equation 2. In the equation, N represents the number of samples in the frame, *n* represents the relevant sample, *x(n)* represents the value of the signal in the n sample, k the current frequency, *X(k)* represents the amplitude and phase values of the *k* frequency in the audio signal [5].

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-i\frac{2\pi kn}{N}} \quad (2)$$

With the formula given in Equation 3, the powers of the frequencies are calculated by squaring the values obtained by FFT.

$$P_n = \frac{(HFD(x_n))^2}{N_{HFD}} \quad (3)$$

It has been reported that the human ear perceives sound frequencies linearly up to 1000 Hz and logarithmically for values after 1000 Hz [10]. With an empirical frequency scale called the Mel scale, sound frequencies are converted to frequency values suitable for the nature of the human ear. Calculation of sound frequency (f) from Mel scale type is given in Equation 4. After this stage, Mel spectrogram is obtained by applying Mel triangle filters (usually 12 filters) to the frequency spectrum. The powers of the Mel bands are obtained by multiplying the Mel filters in each different frequency range with the frequency values in the FFT applied frames [5].

Discrete Cosine Transform (DCT) is applied as the last step in MFCC feature extraction. In this step, commonly 13 MFCC coefficients are obtained for each signal frame. The number of coefficients in our study is 40. MFCC are feature coefficients that are often used in audio classification tasks. The coefficients obtained in each frame are added along the column to obtain the MFCC feature matrix. The MFCC coefficient extraction formula is given in Equation 5. $Ct(n)$ in the formula represents the *n*th MFCC coefficient of the t-frame. M indicates the number of MFCCs. The value of $X'n(m)$ shows the logarithmic energy of the *m*th mel filter [11].

$$Mel(f) = 2595 * (1 + \frac{f}{700}) \quad (4)$$

$$C_t(n) = \sum_{m=0}^{M-1} X'_n(m)\, cos(\frac{\pi n(m-0.5)}{M}) \quad (5)$$

2.5. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a heuristic optimization technique developed by Kennedy and Eberhart [12] in 1995. This optimization was inspired by the movements of birds and fish in search of food, which move in flocks. For the herd to reach its goal, the communication of individuals in the herd with each other has been mathematically revealed. PSO is an algorithm used to solve nonlinear problems. In this algorithm, a population of candidate particles is created to find the best solution in the problem space. Each particle in the population tries to find the best solution. The particle that gives the best solution in the population becomes the leader (GBest) of the population, that is, it refers to the global best particle. The best solution obtained by the particles in the search space individually represents the individual best particle. The individual best particle is expressed by PBest. While the particles are searching for solutions in the search space, they update their velocities with every change of position by referencing the GBest and PBest solutions.

Velocity and position updates are given in Equations 6 and 7, respectively [13].

$$vn[t+1] = w[t]vn[t] + c1r1(xL,n[t]- xn[t]) + c2r2(xG,n[t] - xn[t]) \quad (6)$$

$$xn[t+1] = xn[t] + vn[t+1] \quad (7)$$

The expressions of the formulas given in Equations 6 and 7 are explained below [13].

vn[t+1] :The updated velocity of the particle
w[t] :Inertia coefficient
vn[t] :The previous velocity of the particle
c1 :Coefficient of remembering own best position
r1 :Random number between 0-1
xL,n[t] :The particle's best position ever
xn[t] :The previous position of the particle
xG,n[t] :Swarm leader's position
r2 :Random number between 0-1
c2 :Coefficient of remembering the position of the swarm leader

2.6. Proposed System

The proposed system is shown in Figure 4. The MFCCs feature coefficients obtained from the audio recordings in the Crema-D dataset were used in the training and testing stages of the CNN network.

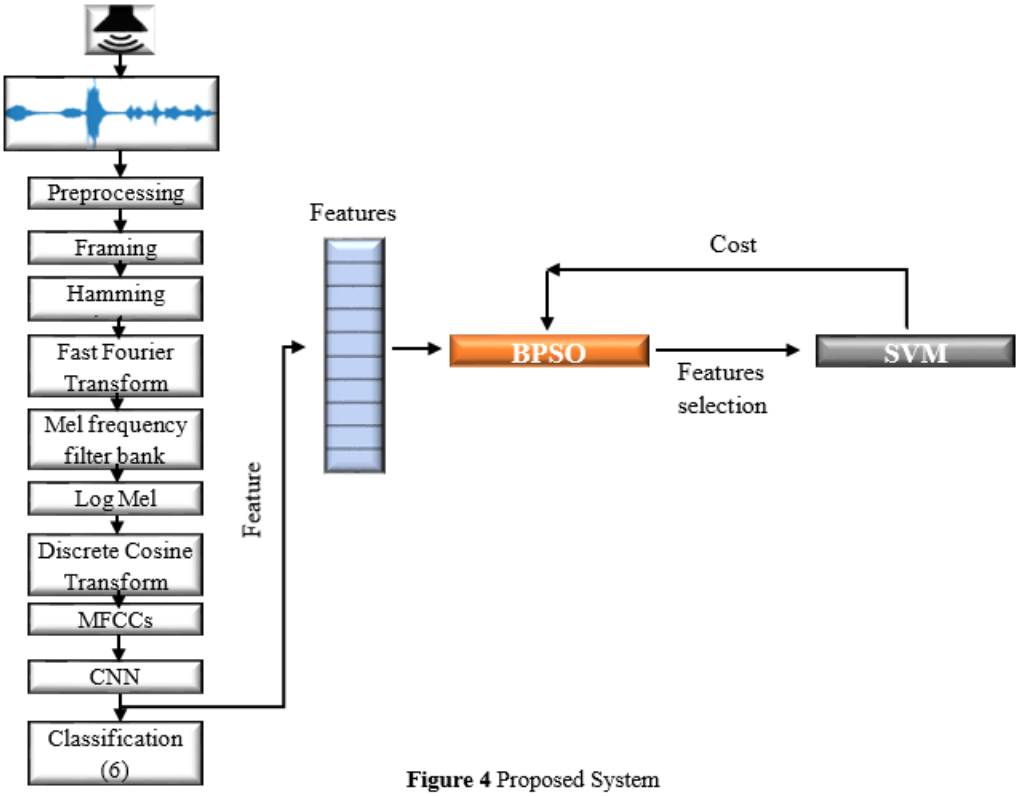

Figure 4 Proposed System

The MFCCs features obtained from the data set are divided into two as 80% training and 20% testing. Training and testing data are shown in Table 1.

Table 1 Crema-D dataset features

Number of clips	Frame feature count (MFCCs)	Number of frames
Training :5953	40	216
Test :1489	40	216

When Table 1 is examined, when the audio recording signals belonging to the equal sample number are divided into frames, 216 frames are obtained. For each frame, 40 MFCCs feature coefficients were extracted. The CNN structure of the proposed system is given in Table 2.

Table 2 CNN structure

Layer	Output Shape	Parameter
input_1 (InputLayer)	[(?, 40, 216, 1)]	0
conv2d (Conv2D)	(?, 40, 216, 64)	640
batch_normalization (BatchNormalization)	(?, 40, 216, 64)	256
activation (Activation)	(?, 40, 216, 64)	0
max_pooling2d (MaxPooling2D)	(?, 20, 108, 64)	0
dropout (Dropout)	(?, 20, 108, 64)	0
conv2d_1 (Conv2D)	(?, 20, 108, 64)	36928
batch_normalization_1 (BatchNormalization)	(?, 20, 108, 64)	256
activation_1 (Activation)	(?, 20, 108, 64)	0
max_pooling2d_1 (MaxPooling2D)	(?, 10, 54, 64)	0
dropout_1 (Dropout)	(?, 10, 54, 64)	0
conv2d_2 (Conv2D)	(?, 10, 54, 128)	73856
batch_normalization_2 (BatchNormalization)	(?, 10, 54, 128)	512
activation_2 (Activation)	(?, 10, 54, 128)	0
max_pooling2d_2 (MaxPooling2D)	(?, 5, 27, 128)	0
dropout_2 (Dropout)	(?, 5, 27, 128)	0
conv2d_3 (Conv2D)	(?, 5, 27, 128)	147584
batch_normalization_3 (BatchNormalization)	(?, 5, 27, 128)	512
activation_3 (Activation)	(?, 5, 27, 128)	0
max_pooling2d_3 (MaxPooling2D)	(?, 2, 13, 128)	0
dropout_3 (Dropout)	(?, 2, 13, 128)	0
flatten (Flatten)	(?, 3328)	0
dense (Dense)	(?, 64)	213056
batch_normalization_4 (BatchNormalization)	(?, 64)	256
activation_4 (Activation)	(?, 64)	0
dropout_4 (Dropout)	(?, 64)	0
dense_1 (Dense)	(?, 6)	390

3. Experimental Results

The CNN network is trained with training data. The error value of the network reached its minimum value at 45 epochs. Data normalization, maxpooling and dropout (0.3) were applied after all convolution operations except the last classification layer. “Adam” optimization algorithm is used to learn the network better. For the loss calculation, categorical_crossentropy was considered suitable as the loss function. While the classification layer is activated with the “softmax” activation function, the “ReLU” activation function is used in all other layers. The batch size of the network is 32.

After the training process, the performance of the model was tested in the testing phase. The test results were realized with an accuracy rate of 62.86%. Precision, recall, f1-score results of emotions were obtained via the sklearn library [14] in Table 3.

Table 3 Performance measurement metrics

Emotion	Precision	Recall	F1-score	Support
Angry	0.64	0.80	0.71	244
Digust	0.52	0.68	0.59	254
Fear	0.65	0.46	0.54	279
Happy	0.75	0.53	0.62	243
Neutral	0.71	0.77	0.73	235
Sad	0.59	0.56	0.57	234

CNN network trained with MFCCs features of speech sound recordings performs classification with 62.86% accuracy. At this stage, feature vectors of 5953x64 dimensions belonging to the fully connected layer of the trained CNN model before the classification layer are sent to the BPSO space for feature selection. In feature selection with BPSO, it creates swarm of solution snippets to find which features represent emotion better in 64-unit feature vectors. In BPSO, each particle, unlike PSO, consists of a 64-unit binary vector containing random "0" and "1" values. The new positions of the particles in motion in the search space must be determined at each iteration. For this, it is necessary to update the velocities of the particles. By applying the Sigmoidal function to the velocities of the particles, values between 0-1 are obtained. Then, the obtained values are compared with a randomly determined number between 0-1 and the new positions of the units of the

particle are updated as "0" or "1" [15]. Position calculation formulas in BPSO optimization are given in Equation 8 and Equation 9.

$$Sig(vn[t+1]) = \frac{1}{1+e^{-(vn[t+1])}} \quad (8)$$

$$xn[t+1] = \begin{cases} 0, if\ rand() \geq Sig(vn[t+1]) \\ 1, if\ rand() < Sig\ (vn[t+1]) \end{cases} \quad (9)$$

Filtering is performed by comparing the particle swarm with the feature vectors obtained from the CNN. The index values of the units with the value "1" of the particle are marked and the features with these index values in the feature vector form the new vector representing the emotion. The resulting new feature vectors are classified by SVM and an error value is obtained [15]. This process ends with finding the solution candidate with the lowest error rate as a result of SVM classification. As a result of applying BPSO to 64 feature layers of each record obtained from the CNN network, features that better represent emotions were obtained from 64 features. Thus, the number of attributes has been reduced from 64 to 33. The accuracy obtained because of the SVM classification made with 33 feature vectors of each record increased by 3.15% and reached 66.01% accuracy. BPSO was performed with 40 iterations and 100 particles. The optimization graph of BPSO is given in Figure 5. In addition, the classification accuracy rates of CNN, SVM and CNN+BPSO+SVM, respectively, using the MFCCs attributes of the data set are given in Table 4.

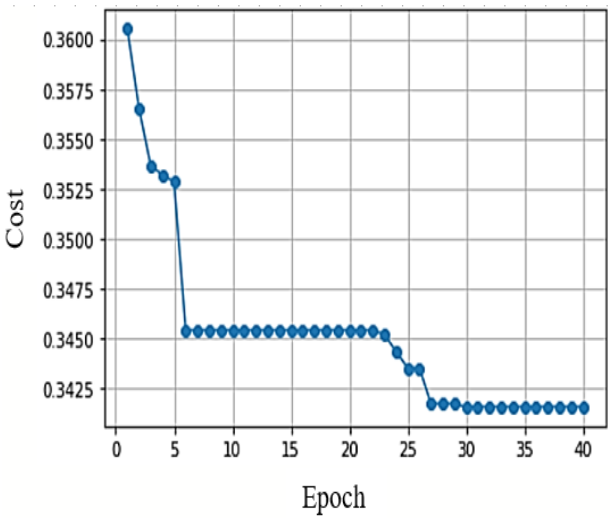

Figure 5 BPSO optimization graph

Table 4 Comparative accuracy rates

Method	**Accuracy (%)**
CNN	62.86
SVM	63.93
CNN+BPSO+SVM (Proposed method)	66.01

4. Conclusion

In our study, the effect of feature selection on classification accuracy was investigated. For this purpose, MFCCs features were extracted on the Crema-D speech dataset. The CNN architecture trained with these features achieved a test accuracy of 62.86%. By using the features in the last full connected layer of the CNN network, feature selection was performed with the help of BPSO. In the feature selection process, the error calculation was carried out with the help of the SVM algorithm. The number of features has been reduced by feature selection. In addition, the accuracy has been increased by 3.15%, reaching an accuracy rate of 66.01%.

References

[1] M. Bojanić, V. Delić, and A. Karpov, "Call Redistribution for a Call Center Based on Speech Emotion Recognition" *Applied Sciences 2020, Vol. 10, Page 4653*, vol. 10, no. 13, p. 4653, Jul. 2020, doi: 10.3390/APP10134653.

[2] A. S. S. Kyi and K. Z. Lin, "Detecting Voice Features for Criminal Case" *2019 International Conference on Advanced Information Technologies, ICAIT 2019*, pp. 212–216, Nov. 2019, doi: 10.1109/AITC.2019.8921212.

[3] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, "Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets" *Electronics (Switzerland)*, vol. 11, no. 22, Nov. 2022.

[4] R. Shankar, A. H. Kenfack, A. Somayazulu, and A. Venkataraman, "A Comparative Study of Data Augmentation Techniques for Deep Learning Based Emotion Recognition" Nov. 2022, doi: 10.48550/arxiv.2211.05047.

[5] K. Donuk and D. Hanbay, "Konuşma Duygu Tanıma için Akustik Özelliklere Dayalı LSTM Tabanlı Bir Yaklaşım" *Computer Science*, vol. 7, no. 2, pp. 54–67, 2022, doi: 10.53070/bbd.1113379.

[6] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches" *Neurocomputing*, vol. 492, pp. 245–263, Jul. 2022, doi: 10.1016/J.NEUCOM.2022.04.028.

[7] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset" *IEEE Trans Affect Comput*, vol. 5, no. 4, p. 377, Oct. 2014, doi: 10.1109/TAFFC.2014.2336244.

[8] Ö. F. ÖZTÜRK and E. PASHAEİ, "Konuşmalardaki duygunun evrişimsel LSTM modeli ile tespiti" *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, vol. 12, no. 4, pp. 581–589, Sep. 2021, doi: 10.24012/DUMF.1001914.

[9] "librosa — librosa 0.9.2 documentation." https://librosa.org/doc/latest/index.html (accessed Dec. 01, 2022).

[10] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch" *J Acoust Soc Am*, vol. 8, no. 3, pp. 185–190, Jan. 1937, doi: 10.1121/1.1915893.

[11] Q. Chen and G. Huang, "A novel dual attention-based BLSTM with hybrid features in speech emotion recognition" *Eng Appl Artif Intell*, vol. 102, p. 104277, Jun. 2021, doi: 10.1016/J.ENGAPPAI.2021.104277.

[12] J. Kennedy and R. Eberhart, "Particle swarm optimization" *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, pp. 1942–1948, doi: 10.1109/ICNN.1995.488968.

[13] K. Donuk, N. Özbey, M. Inan, C. Yeroğlu, and D. Hanbay, "Investigation of PIDA Controller Parameters via PSO Algorithm" *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*, Jan. 2019, doi: 10.1109/IDAP.2018.8620871.

[14] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python" *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Dec. 02, 2022. http://jmlr.org/papers/v12/pedregosa11a.html

[15] K. Donuk *et al.*, "Deep Feature Selection for Facial Emotion Recognition Based on BPSO and SVM" *Politeknik Dergisi*, Dec. 2022, doi: 10.2339/POLITEKNIK.992720.