Yuzuncu Yil University
Journal of the Institute of Natural & Applied Sciences
https://dergipark.org.tr/en/pub/yyufbed

Research Article

Classification of Emotion with Audio Analysis

Coşkucan BÜYÜKYILDIZ*[1], İsmail SARITAŞ[2], Ali YAŞAR[3]

[1]Selçuk University, Faculty of Technology, Institute of Science, 42250, Konya, Türkiye
[2]Selçuk University, Faculty of Technology, Electrical and Electronic Engineering, 42250, Konya, Türkiye
[3]Selçuk University, Faculty of Technology, Mechatronic Engineering Department, 42250, Konya, Türkiye
Coşkucan BÜYÜKYILDIZ, ORCID No: 0000-0002-8190-5914, İsmail SARITAŞ,
ORCID No: 0000-0002-5743-4593, Ali YAŞAR, ORCID No: 0000-0001-9012-7950
*Corresponding author e-mail: coskucan94@hotmail.com

Article Info

Received: 16.12.2022
Accepted: 20.03.2023
Online August 2023

DOI:10.53433/yyufbed.1219879

Keywords
Audio features,
Classification,
Emotion identifier,
Machine learning database,
Support vector machine

Abstract: Classification is an important technique used to distinguish data samples. The aim of this study is to classify according to emotions by extracting audio features. Two male and two female individuals expressed four different emotions as "fun", "angry", "neutral" and "sleepy" in the voice data. We used to “MFCC” as a Cepstral feature, “Centroid, Flatness, Skewness, Crest, Flux, Slope, Decrease, Kurtosis, Spread, Entropy, roll off point” as Spectral Feature, “Pitch, Harmonic ratio” as Periodicity Features in the sound features. After, we applied to the data that all the classification algorithms located in the classification learner toolbox in Matlab and we tried to classify the emotion with the algorithm that provides the highest accuracy. Each data in the classification study has twenty-six features inputs and one labeled output value. According to the results, support vector machine algorithm provided the highest accuracy performance. Considering the performances obtained, this study reveals that it is possible to distinguish and classify sounds using sentimental data and sound feature parameters.

Ses Analiziyle Duyguların Sınıflandırılması

Makale Bilgileri

Geliş: 16.12.2022
Kabul: 20.03.2023
Online Ağustos 2023

DOI:10.53433/yyufbed.1219879

Anahtar Kelimeler
Destek vektör makinesi,
Duygu tanımlama,
Makine öğrenmesi veritabanı,
Ses öznitelikleri,
Sınıflandırma

Öz: Sınıflandırma, veri örneklerini ayırt edebilmek için kullanılan önemli bir tekniktir. Bu çalışmada öz nitelikler çıkartılarak, duygulara göre sesin sınıflandırılması amaçlanmıştır. Neşeli, sinirli, nötr ve uykulu olmak üzere dört farklı duyguda konuşan iki erkek ve iki kadın bireyden alınan ses verileri kullanılmıştır. Sesin özniteliklerinde; Kepstral özellik olarak “Mel-Frekansı Kepstral Katsayıları”, Spektral Özellik olarak “Ağırlık Merkezi, Pürüzsüzlük, Çarpıklık, Tepe, Akış, Eğim, Azalma, Basıklık, Yayılma, Entropi, Yuvarlanma noktası”, Periyodisite Özelliği olarak “Ses perdesi, Harmonik oran” kullandık. Daha sonra, Matlab’da bulunan “sınıflandırma öğrenici” araç kutusunda yer alan tüm sınıflandırma algoritmalarını veriye uyguladık ve en yüksek doğruluğu sağlayan algoritmayla duyguyu tahmin etmeye çalıştık. Sınıflandırma çalışmasında yer alan her bir veri, yirmi altı öz nitelik girdisi ve bir etiketli çıktı değerine sahiptir. Performans sonuçlarına göre, destek vektör makine algoritması en yüksek doğruluk değerini sağlamıştır. Elde edilen performans çıktıları göz önüne alındığında, bu çalışma, duyusal veriler ve ses öznitelikleri kullanılarak sesleri ayırt etmenin ve sınıflandırmanın mümkün olduğunu ortaya koymaktadır.

1. Introduction

Speech is the most effective form of communication today. The tone of voice used during speech often gives clues about emotions and people act according to these feelings. Therefore, one of the ways to understand the other person correctly is to do an emotion analysis (Koolagudi et al., 2009). In order to do an emotional analysis, first of all, it is necessary to make a voice or speech analysis. With voice analysis, emotional changes in voice can be detected, and a data set can be created. Thus, voice or speech analysis is vital to improve communication and future interaction.

Emotion analysis models obtained by processing voice data are used to automate the business processes of companies and extract meaningful information. For this reason, we can determine the attitude of individuals or a particular group about the subject of the study. However, a large number of data to obtain positive or negative opinions makes it difficult to analyze. Therefore, classification studies using data mining and machine learning for emotion analysis have become one of the critical topics studied (Kaynar et al., 2016). Thanks to the classification studies created in this way, it is possible to reduce the possibility of making mistakes by minimizing workforce and obtaining more apparent results.

Machine learning is the development of automated techniques to learn how to make super predictions based on previous observations. In the literature, machine learning is defined under four headings; supervised, unsupervised, semi-supervised, and reinforcement learning.

Table 1. Instances with known labels (Kotsiantis, 2007)

Case	**Feature 1**	**Feature 2**	**…**	**Feature n**	**Class**
1	xxx	x	3501	xx	good
2	xxx	x	3899	xx	good
3	xxx	x	3810	xx	bad
…					…

Each sample in the dataset used in machine learning algorithms can be represented using many different data features. Features can be continuous, categorical, or binary. The learning is called supervised if the data samples are given with known labels. Examples of supervised learning are shown in Table 1 (Kotsiantis, 2007).

Classification is performed to determine the group to which the data types in a dataset belong. Thus, it provides an effective solution for parsing data. To distinguish the groups, the relevant data is labeled, and the model created with the selected algorithm learns this data during the training. Studies for classification can be created using machine learning algorithms such as Random Forest, Naive Bayes, K-Nearest Neighbors, Decision Tree, and Support Vector Machines. The data to be trained in the model are pre-labeled.

For this reason, classification studies are included in the subject of supervised learning in the literature. The model learned from the training data to determine the category of the input data is called a classifier. The classifier can be implemented as a binary classifier or a multi-class classifier. The input data in the binary classifier is divided into one of two categories at the output. Input data in a multi-class classifier can be determined as one of more than two output categories (Rebala et al., 2019).

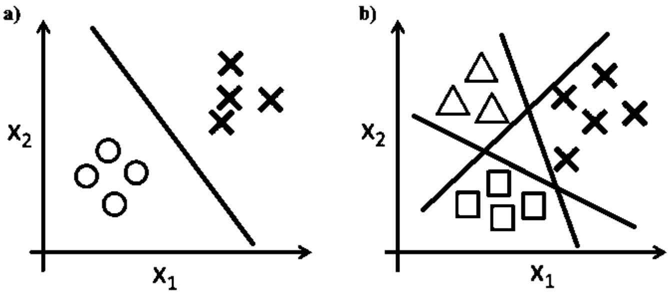


Figure 1. a) Binary classification b) Multi-class classification.

Features can be extracted from raw data such as audio, picture, and video used in machine learning. A model can be trained with machine learning algorithms using these features, and desired outputs can be obtained. In order to increase the performance and accuracy of the classification study to be carried out, it is crucial to extract the correct number of features and to ensure that these features are decisive (Lech et al., 2020). In order to analyze the sounds with machine learning, it is necessary to create a data pool by taking sound samples of many different emotions.

In this paper, we made a classification study by using supervised learning. It is a multi-class classifier because we use more than one label in the classification. Additionally, the audio dataset, which is publicly available in OpenSLR was obtained from (https://coe.northeastern.edu/Research/AClab/Speech%20Data/). In order to analyze the sounds, 26 sound features were extracted, and a classification study was conducted by comparing machine learning algorithms with each other.

We have extensively trained and evaluated classification algorithms. We performed a detailed analysis on performance metrics together with methods such as Confusion Matrix and ROC curve. Also, we used different features such as spectral and periodicity with MFCC in this study. As a result of these comprehensive analyzes and the use of features, we have demonstrated the success of the Cubic SVM algorithm. In looking the literature, we see that there are many different studies on the classification of emotions with sound analysis. These studies have shown the superiority of SVMs in classification by using different methods. In this context, we presented information about the different studies in below.

Jain et al. (2018) aimed to classify due to gender by using prosodic and spectral features of sound for the Hindi language. Mel-frequency Cepstrum Coefficients (MFCC) and Linear Prediction Cepstral Coefficient (LPCC) used. The study carried out with the Cubic SVM classification algorithm. According to the results, 95% accuracy rate for women and 98.75% for men was obtained.

Chatterjee et al. (2018) were used two different techniques to classify Angry, Happy or Neutral emotions. First of all, maximum cross correlation was calculated between the voice data for labeling the data to be used. In the second technique, Energy, Volume, MFCC, Zero Crossing Rate, Formants and Spectral Centroid are used as Sound features. Then, these features were used in the Cubic SVM classification algorithm. According to the results, 91.3% accuracy rate obtained.

Tuncer et al. (2021) is presented a sound classification system using twine shufpat, INCA, and TQWT techniques. Using this system, accuracy values of 87.43%, 90.09%, 84.79%, and 79.08% were obtained for RAVDESS Speech, Berlin Emo-DB, SAVEE, and EMOVO databases, respectively. In addition, another classification study was carried out by mixing these databases. In this classification study, which includes nine different emotions, an accuracy value of 80.05% was obtained. Cubic SVM provided the best classification performance in the developed system.

Mohamad Nezami et al. (2019) was used Sharif Emotional Speech Database (ShEMO) for Persian language. For the emotions of Anger, fear, happiness, sadness, neutral and surprise, a classification study was carried out with samples taken from 87 native Persian speakers. According to the results, SVM's achieves the best results for both gender-independent (58.2%) and gender-dependent models (female=59.4%, male=57.6%).

Milton et al. (2013) was used a 3-stage SVM classifier for classification. MFCC features were extracted from 535 different audio data. Linear and RBF kernels were used as hierarchical SVMs with RBF sigma equal to 1. The training and testing process of the data was carried out in the form of cross-evaluation 10 times. According to the performance results, the accuracy rate is 68%.

Aouani & Ayed (2018) were carried out two different studies, namely sound feature extraction and classification. First of all, 39 and 65 different MFCC coefficient features are extracted. Secondly, the SVM algorithm was used as a classifier. According to the results performed in the SAVEE audio database, the DSVM method using 39 MFCC coefficients showed better results than standard SVMs with a classification rate of 69.84% and 68.25%. In addition, the auto-encoder method performed better with a classification rate of 73.01%.

Sonawane et al. (2017) support vector machines are used as the classifier algorithm. In addition, MFCC audio features are used. Classification was carried out on happy, angry, sad, disgusted, surprised and neutral emotions. The results revealed that nonlinear kernel SVM achieves higher accuracy than linear SVM.

2. Material and Methods

In this section, the materials and methods used in the study are presented and examined under three sub-titles. Figure 2 shows the processes of the work performed.

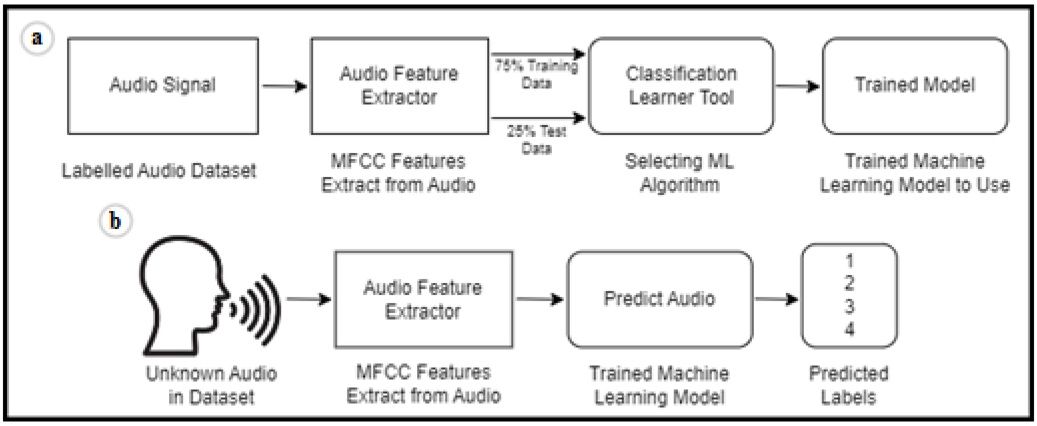


Figure 2. a) Train process of machine learning model b) Predict process of unknown audio data.

2.1. Dataset

Audio recordings in English and French are available in the Dataset. In these recordings, five different senses, such as "Amused, Neutral, Sleepy, Disgusted and Angry" were used. Two female and two male English and one male French native speaker were asked to read the sentences while expressing one of these feelings. English and French sentences are sourced from an open-source database. English data were recorded in two insulated rooms on the Northeastern University campus. All data was recorded at 44.1k and down sampled at 16k and recorded in 16-bit PCM WAV format (Adigwe et al., 2018).

In this study, only the English language was used together with the emotional expressions "Amused, Neutral, Sleepy and Angry" in the related data set. For each emotion used, 100 voice recordings were evaluated. Samples of 3 seconds were used for each sound recording. The count of audio records belonging to Neutral, Amused, Angry and Sleepy emotions is shown in Table 2. In addition, the names and gender information of the speakers are also presented in Table 2 below.

Table 2. Used dataset content

Speaker	Gender	Neutral	Amused	Angry	Sleepy
Jennie	Female	417	222	523	466
Bea	Female	373	309	317	520
Sam	Male	493	501	468	495
Josh	Male	302	298	-	263

2.2. Feature extraction

When feature extraction is performed, the original data in the dataset is preserved. In addition, numerical values that can be processed are obtained from the raw data. A model trained with feature extraction gives better results than a model trained with raw data. The features of the sounds were obtained with the “audio feature extractor” function in Matlab. The features used in the study are explained below.

Mel-frequency cepstral coefficients (MFCCs): Mel-frequency cepstrum (MFC) can be described as the short-term power spectrum of a sound in signal processing. Also, this cepstrum uses a linear cosine transform of a log power spectrum on a nonlinear Mel-frequency scale. Mel-frequency cepstral coefficients (MFCCs) make up the MFC (Vyas & Kumari, 2013). While obtaining the cepstrum coefficients used as the feature vector, MFCC is generally used in speaker recognition applications. Because MFCC imitates the frequency selectivity of the human ear, it achieves discriminating values for a good speaker. In addition, the MFCC coefficients are much less affected by the changes and the sound wave structure (Eskidere & Ertaş, 2009).

Spectral Centroid: Spectral centroid defines the location of the center of mass of the spectrum. This term is associated with the brightness of the sound (Grey & Gordon, 1978).

Spectral Flatness: Spectral flatness or tonality coefficient is also known as Wiener entropy. It is used to measure how pure a tone the sound is. It can be measured from the decibel of sound (Dubnov, 2004).

Spectral Skewness: The spectral skewness measures the symmetry of the spectrum around its centroid. The skewness can take positive, zero, negative, or undefined values (Lerch, 2012).

Spectral Kurtosis: A Spectral kurtosis is a statistical tool that determines the presence of a series of transients in the signal and their position in the frequency domain (Antoni, 2006).

Spectral Crest: Spectral crest is the ratio of the peak values in the audio signal to the arithmetic mean. It shows the peak in the sound spectrum. The higher the peak value, the higher the tone. However, the lower the peak value, the higher the audio noise (Peeters et al., 2011).

Spectral Flux: Spectral flux is the rate of change in the sound's power spectrum. The difference between the power spectrum of the frame in the sound and the previous frame calculates it. Each difference is squared, and the result is the sum of the squares (Giannoulis et al., 2013).

Spectral Slope: In digital signal processing, it measures the rate of change of the spectrum of an acoustic sound, which decreases towards higher frequencies, calculated using linear regression and the central wavelet of the signal (Peeters, 2004).

Spectral Decrease: Spectral decrease refers to the amount of reduction in the spectrum. It shows the lower frequency slopes of the signal. It is an important criterion for instrument recognition (Peeters et al., 2011).

Spectral Spread: The spectral spread can be expressed as the distribution around the weighted center of the spectrum. The signals concentrated in this center are the points where the spectral spread is low (Giannakopoulos & Pikrakis, 2014).

Spectral Entropy: The spectral entropy can be expressed as the spectral power distribution of the signal obtained using the normalized Fourier transform. This entropy is based on Shannon's entropy in information theory. It is a widely used measurement in speech recognition applications (Misra et al., 2004).

Spectral Roll-off: The spectral roll-off point calculates the rolling frequency of a spectrum. The frame at this point contains 85% of the spectrum power at low frequencies. It is used to distinguish between loud and silent speech and music (Mitrović et al., 2010).

Pitch: Indicates that the perceived sound is high and low. High-pitch values are associated with thin sounds, while low-pitch values are associated with deep sounds.

Harmonic Ratio: The harmonic ratio is obtained by dividing the frequency power by the total power in an audio signal frame (Mitrović et al., 2010).

2.3. Classification algorithms

This study used all classification algorithms in the classification learner in the Matlab tool for comparison purposes. 75% of the dataset was set as training data. 25% was used as test data. The four different emotions in the dataset are labeled as follows. This labeled the data is added to data table as output. Labeled values and classes are shown following in Table 3.

Table 3. Labeled features

Label	**1**	**2**	**3**	**4**
Class	Amused	Angry	Neutral	Sleepy

The classification algorithms used in this study are as follows; Cubic SVM, Bagged Trees, Medium Gaussian SVM, Wide Neural Network, Fine Gaussian SVM, Quadratic SVM, Medium Neural Network, Weighted KNN, Cosine KNN, Medium KNN, Fine KNN, Trilayered Neural Network, Cubic KNN, Bilayered Neural Network, Narrow Neural Network, Coarse KNN, Coarse Gaussian SVM, Linear SVM, Subspace Discriminant, Fine Tree, Subspace KNN, Quadratic

Discriminant, Boosted Trees, Medium Tree, RUSBoosted Trees, Kernel Naive Bayes, Coarse Tree, Gaussian Naive Bayes.

According to the results, the best classification outputs were obtained from the support vector machine algorithms. For this reason, only support vector machines are examined in this section. Support Vector Machine is a supervised learning method generally used in classification problems. One of the benefits of SVM is that it can build machine-learning models with intermittent data and few data samples. SVMs are binary linear classifiers and parametric learners. Parametric approaches use parameters to train the model. Each data item in the used dataset can have more than one feature, and each feature has a place in the coordinate plane. Therefore, It is drawn as a point in a multidimensional space with as many features as it has (Kishore et al., 2022). These algorithms create a hyperplane boundary to separate the data points into two categories. At least one hyperplane is required to specify data boundaries. Data samples that this hyperplane can separate are called linearly separable. SVM algorithms can parse nonlinear data samples using kernel functions (Rebala et al., 2019).

Cubic SVM: Support vector machine operations in multidimensional space can take a long time. For this reason, we can perform operations in a shorter time by using kernel functions. Cubic SVM is realized using a Polynomial kernel function. These functions are a non-stationary kernel. The polynomial kernel is generally suitable for problems where all the training data are normalized (Metlek & Kayaalp, 2020). 3rd-degree polynomials are expressed as cubic polynomials. Therefore, in the equation below, the cubic kernel function is 3rd degree. For an SVM type classifier, if the kernel function used is cubic, it is expressed as in Equation (1) (Jain et al., 2018). Here, the product of the vectors x_i and x_j is related to the similarity between them, that is, the angle between them.

$$\kappa(x_i, x_j) = (x_i^T, x_J)^3 \tag{1}$$

3. Results

According to the results, the Cubic SVM algorithm provided the best prediction response with 84.2%. Looking at the top five best scores in Figure 3, it is seen that SVM algorithms are in the majority. Looking at the last five classification algorithms in Figure 3, it is seen that the lowest score belongs to the Tree and Naive Bayes algorithm.

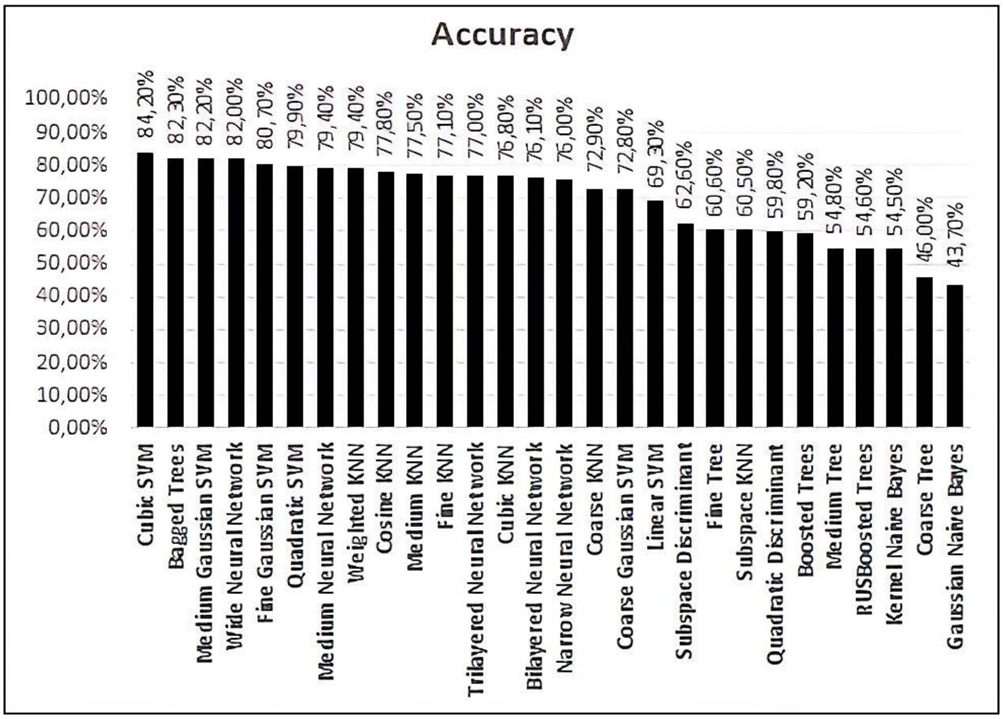


Figure 3. Accuracy table of classification algorithms for emotion analysis.

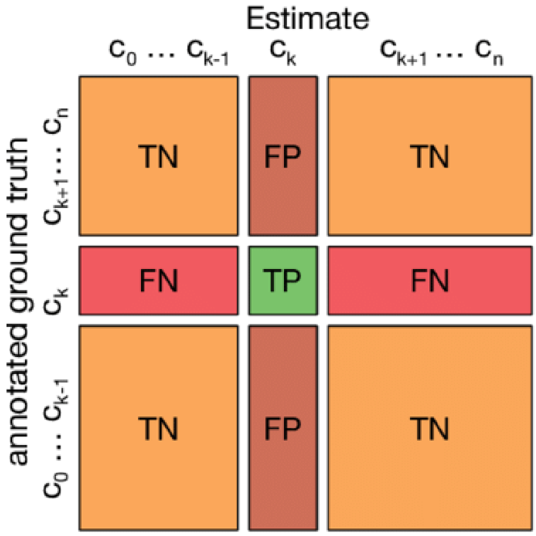


Figure 4. Confusion matrix for multi-class classification (Krüger, 2016).

In order to better understand a classification problem and measure its accuracy, the results obtained should be interpreted and evaluated after the relevant model is trained. In this case, the confusion matrix can be used as a widely used tool in machine learning, and we can evaluate the outputs visually. The columns of the Confusion Matrix show the results of the prediction classes, and the rows show the results of the actual class. According to this matrix, all possible states of the classification problem are shown. In addition, it also displays the total number of misclassified data for each class included in the matrix. True Positive, True Negative, False Positive, and False Negative metrics can be calculated using the Confusion Matrix (Yasar, 2022). An example of a multiclass confusion matrix with n classes is shown in Figure 4. The data prediction points in the confusion matrix are expressed in Equation (2) (Krüger, 2016).

$$C = (c_{ij}) \tag{2}$$

Where c_{ij} is the number of time-steps where the class was actually i and class j has been estimated. When considering the class k ($0 \leq k \leq n$), the confusion matrix provides four types of classification results due to one classification target k (Tharwat, 2020):

True Positive (TP) is the number correctly determined in the classes true and predicted. The equation is expressed in the following Equation (3).

$$C = (c_{kk}) \tag{3}$$

False Negative (FN) is the number of points where the true class is true and the predicted class is false. The equation is expressed in the following Equation (4).

$$\Sigma_{i \in N \setminus \{k\}} c_{ki} \tag{4}$$

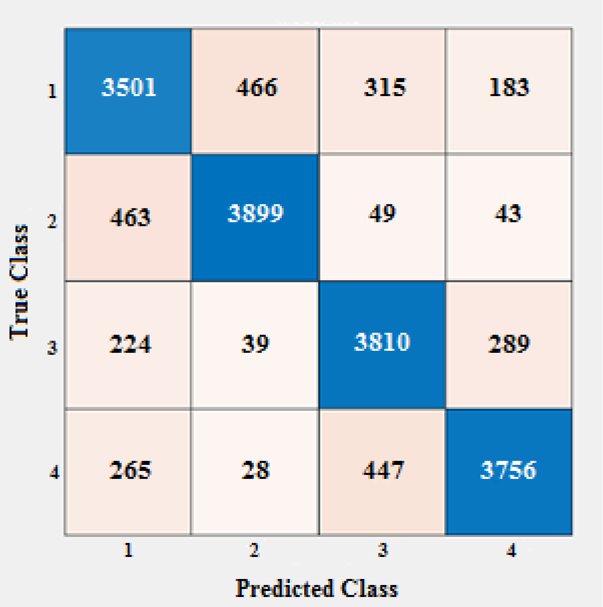
False positive (FP) is the number of points correctly predicted in the predicted class and incorrectly predicted in the true class. The equation is expressed in the following Equation (5).

$$\Sigma_{i \in N \setminus \{k\}} c_{ik} \tag{5}$$

True Negative (TN) is the number of points where the true and predicted classes are false. The equation is expressed in the following Equation (6).

$$\sum_{ij\in N\setminus\{k\}} c_{ij} \tag{6}$$

In this study, the dimension of a Confusion Matrix is 4 × 4. The multi-class confusion matrix results realized with the Cubic SVM algorithm are shown in Table 4. Classification parameters were calculated separately for each classifier.



	1	2	3	4
1	3501	466	315	183
2	463	3899	49	43
3	224	39	3810	289
4	265	28	447	3756

Figure 5. Confusion matrix of validation data for cubic SVM algorithm.

Table 4. Analysis of the confusion matrix

Class	False Positive	False Negative	True Positive	True Negative
1	952	964	3 501	12 360
2	533	555	3 899	12 790
3	811	552	3 810	12 604
4	515	740	3 756	12 766

3 501 of the "amused" class was correctly classified and this corresponds to 19.7% of all 17 777 audio features in the data. 3 899 of the "angry" class are correctly classified and this corresponds to 21.9% of all data. 3 810 of the "neutral" class are correctly classified and this corresponds to 21.4% of all data. 3 756 of the "sleepy" class are correctly classified and this corresponds to 21.2% of all data. 84.2% of the predictions are correct, and 15.8% are wrong. According to these results, the most used performance metrics are presented below.

The Accuracy (Acc) performance metric is one of the frequently used criteria in classification problems. It is expressed as the ratio of correctly classified samples to the total number (Vyas & Kumari, 2013). The equation is expressed in the following Equation (7).

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{7}$$

where P and N indicate the number of positive and negative samples, respectively.

Error Rate (Err) is the inverse of the accuracy metric. It is the ratio of the values that incorrect predictions to the total number (Tharwat, 2020). The equation is expressed in the following Equation (8).

$$\mathrm{Err} = 1 - \mathrm{Acc} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{8}$$

Precision is the ratio of the true positive values to the sum of the predicted positive classes. This metric is more insensitive to unbalanced data. Therefore, it can be preferred instead of the accuracy metric for this data type. This value is high in a classification problem with good results. The equation is expressed in the following Equation (9).

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

Sensitivity is also referred to as True Positive Rate or Recall. It is a measure of positive examples labeled as positive by a classifier. It should be higher. The equation is expressed in the following Equation (10).

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

False positive rate (FPR) is a ratio of the total number of incorrectly classified positive examples to the total number of ground truth negatives. The equation is expressed in the following Equation (11).

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{11}$$

Using the metrics obtained in the confusion matrix analysis, Table 5 was created with the above measurement methods. The values in the table were calculated separately according to the classes.

Table 5. Performance metrics that used for classification

Class	Accuracy	Error Rate	Precision	FPR	Sensitivity
1	89.2%	10.8%	78.6%	7.15%	78.4%
2	93.8%	6.2%	78.9%	4.00%	87.5%
3	92.3%	7.7%	75.7%	6.04%	87.3%
4	92.9%	7.1%	81.9%	3.87%	83.5%

The ROC curve can be created with the performance metrics in the table above. This curve is created by comparing the true positive rate vertically and the false positive rate horizontally. Although ROC is a probability curve, it is one of the commonly used metrics to evaluate the performance of machine learning algorithms in the presence of inconsistent datasets. The area under the ROC curve is briefly expressed as AUC. The AUC shows how well the model can distinguish classes. The higher the AUC value, the easier it is to distinguish classes (Yasar et al., 2018).

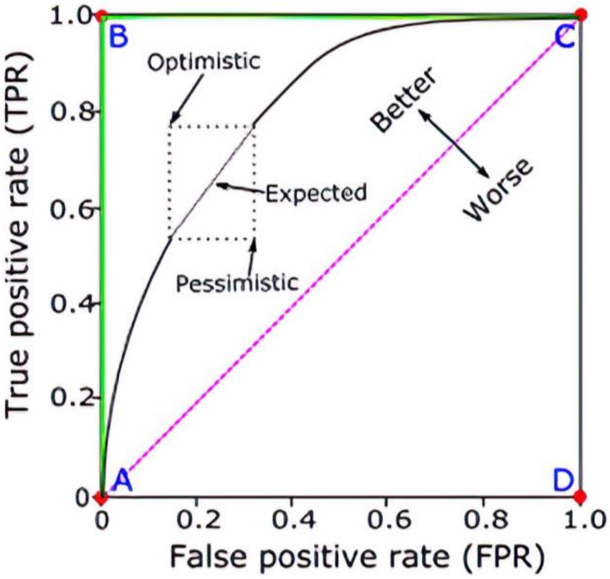


Figure 6. Characteristic and measuring points of the ROC curve (Kishore et al., 2022).

Figure 6 shows the possible behavior of the ROC curve. Four important coordinate points, A, B, C, and D, show the ROC curve's performance. Here, at point B (0,1), the best results of the curve are obtained. Therefore, the green curve shown in Figure 6 passes through point B. The area under the green curve (AUC) is equal to one (Tharwat, 2020).

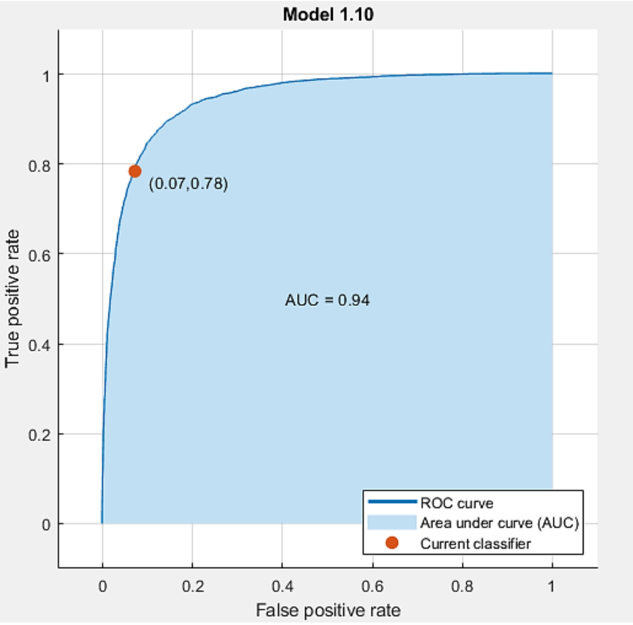


Figure 7. ROC curve for class 1 (Amused).

As shown in Figure 7, the AUC value for class 1 (Amused) in validation data was calculated as AUC = 0.94. According to the result, the classifier prediction of the Cubic SVM algorithm for class 1 is high.

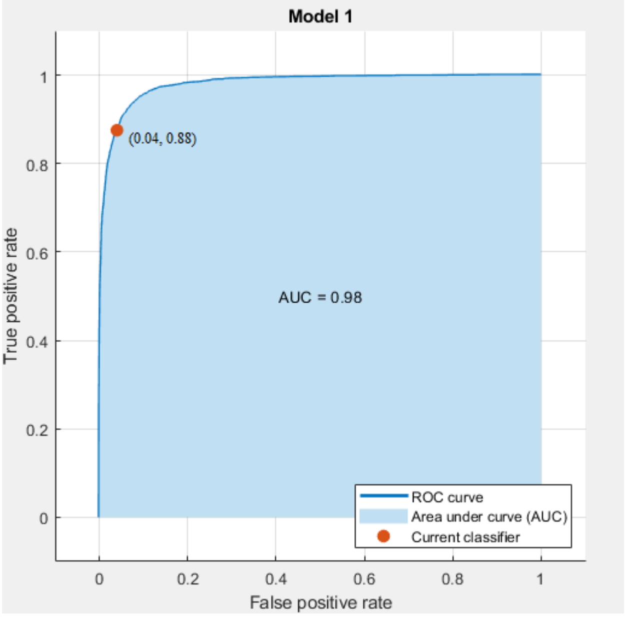


Figure 8. ROC curve for class 2 (Angry).

As shown in Figure 8, the AUC value for class 2 (Angry) in validation data was calculated as AUC = 0.98. According to the result, the classifier prediction of the Cubic SVM algorithm for class 2 is high.

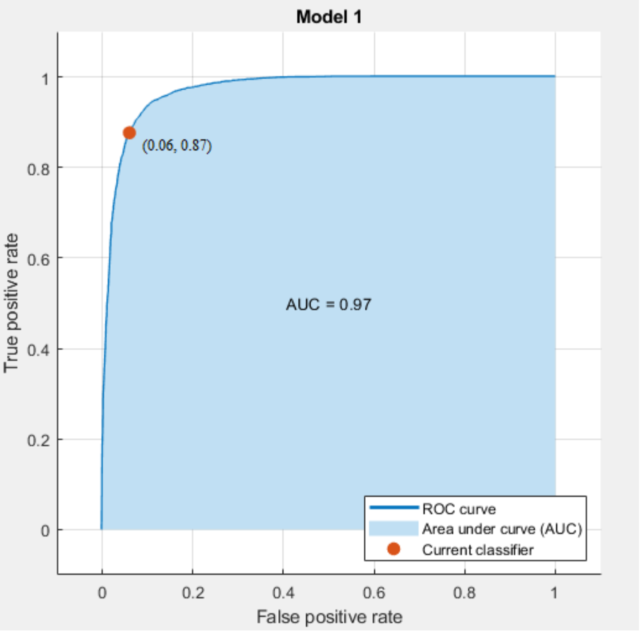


Figure 9. ROC curve for class 3 (Neutral).

As shown in Figure 9, the AUC value for class 3 (Neutral) in validation data was calculated as AUC = 0.97. According to the result, the classifier prediction of the Cubic SVM algorithm for class 3 is high.

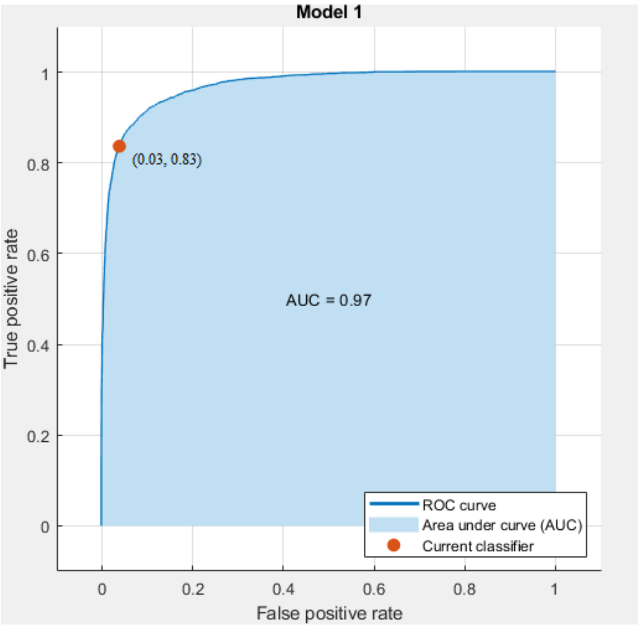


Figure 10. ROC curve for class 4 (Sleepy).

As shown in Figure 10, the AUC value for class 4 (Sleepy) in validation data was calculated as AUC = 0.97. According to the result, the classifier prediction of the Cubic SVM algorithm for class 4 is high.

Figure 11 shows the Confusion Matrix results for the voice data containing sleepy emotional expressions, which is not included in the data set. For test data, 60 different features were extracted from the 3-second audio data, and these features were classified using the previously trained cubic SVM algorithm. According to the results obtained from the model, 13 features were classified as 1st class, 17 as 3rd class, and 30 as 4th class. Looking at the results, the most data is in the 4th class. Therefore, it can be said that the estimated test data belongs to the sleepy class.

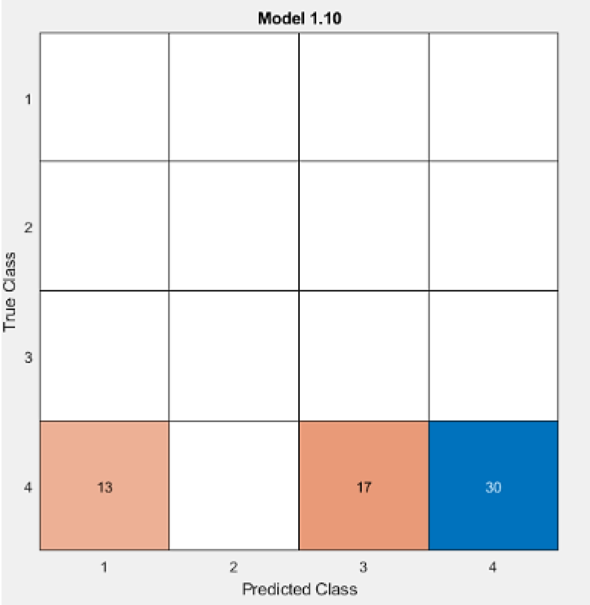


Figure 11. Confusion matrix of test data for sleepy emotion.

The Confusion Matrix in Figure 12 shows the results of audio data that include amused emotional expressions not included in the dataset. In this example, 60 different features are extracted from 3-second audio data. The outputs obtained from the model are classified as 27 features 1st class, 15 features 2nd class, 6 features 3rd class, and 12 features 4th class. As a result, the most of the data are in the 1st class. Therefore, it can be said that the predicted test data belong to the amused class. The Cubic SVM model was seen to analyze the "amused" input audio data accurately.

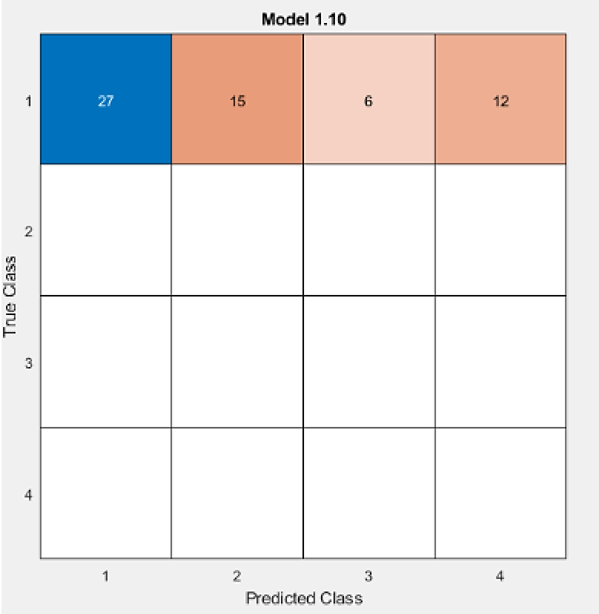


Figure 12. Confusion matrix of test data for amused emotion.

4. Conclusion

This article discusses a classification study using supervised learning to determine emotions in audio recordings. The study uses the publicly available OpenSLR audio dataset and extracts 26 sound features to be used in the machine learning algorithms. The article explains the feature extraction process and the different features used, such as Mel-frequency cepstral coefficients (MFCCs), Spectral Centroid, Spectral Flatness, Spectral Skewness, Spectral Kurtosis, Spectral Crest, Spectral Flux, Spectral Slope, Spectral Decrease, Spectral Spread, Spectral Entropy, Pitch, and Harmonic Ratio. The study uses all classification algorithms in the classification learner in Matlab.

Unlike existing studies, we used to various classification algorithms in this study and we worked with spectral and periodic features as well as cepstral coefficients. After this comprehensive training and evaluation process, we obtained accuracy values for each classification algorithm. According to results, the Cubic SVM algorithm provided the best response with 84.2% accuracy in multi-featured and multidimensional chaotic audio data. Then, the results performed with the test data on the Trained Cubic SVM model were examined, and the accuracy of the trained model was checked.

When we look at the literature in voice sentiment analysis studies, the success of SVM algorithms is clearly seen. In this study, Cubic SVM algorithm is presented as the best sound emotion classifier by comparing to other classification algorithms and by using extensive sound features. For this reason, Cubic SVM algorithm is recommended because of its high accuracy value and high performance metric outputs in order to classify emotions in English language conversations.

5. Acknowledgements

This article was presented orally at the 10th International Conference on Advanced Technologies (ICAT'22) and selected to be published by ICAT'22 Editor Committee. Paper ID: 0274, Paper Name: Classification of Emotions with Audio Analysis. An expanded version of the article has been uploaded to this journal.

References

Adigwe, A., Tits, N., Haddad, K. E., Ostadabbas, S., & Dutoit, T. (2018). The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*. doi:10.48550/arXiv.1806.09514

Antoni, J. (2006). The spectral kurtosis: A useful tool for characterising non-stationary signals. *Mechanical Systems and Signal Processing*, 20(2), 282-307. doi:10.1016/j.ymssp.2004.09.001

Aouani, H., & Ayed, Y. B. (2018, March). *Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder*. 2018 4th International conference on advanced technologies for signal and image processing (ATSIP), Sousse, Tunisia. doi:10.1109/ATSIP.2018.8364518

Chatterjee, J., Mukesh, V., Hsu, H.-H., Vyas, G., & Liu, Z. (2018, August). *Speech emotion recognition using cross-correlation and acoustic features*. 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/ PiCom/ DataCom/ CyberSciTech), Athens, Greece. doi:10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00050

Dubnov, S. (2004). Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters*, 11(8), 698-701. doi:10.1109/LSP.2004.831663

Eskidere, Ö., & Ertaş, F. (2009). Mel frekansı kepstrum katsayılarındaki değişimlerin konuşmacı tanımaya etkisi. *Uludağ University Journal of The Faculty of Engineering*, 14(2), 93-110.

Giannakopoulos, T. & Pikrakis, A. (2014). Introduction to audio analysis: A MATLAB® approach. Orlando, FL, USA: Academic Press Inc.

Giannoulis, D., Massberg, M. & Reiss, J. D. (2013). Parameter automation in a dynamic range compressor. *Journal of the Audio Engineering Society*, 61(10), 716-726.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5), 1493-1500. doi:10.1121/1.381843

Jain, U., Nathani, K., Ruban, N., Raj, A. N. J., Zhuang, Z., & Mahesh, V. G. V. (2018, October). *Cubic SVM classifier based feature extraction and emotion detection from speech signals*. 2018 International Conference on Sensor Networks and Signal Processing (SNSP), Xi'an, China. doi:10.1109/SNSP.2018.00081

Kaynar, O., Görmez, Y., Yıldız, M., & Albayrak, A. (2016, September). *Makine öğrenmesi yöntemleri ile duygu analizi*. International Artificial Intelligence and Data Processing Symposium (IDAP'16), Malatya, Türkiye.

Kishore, B., Yasar, A., Taspinar, Y. S., Kursun, R., Cinar, I., Shankar, V. G., … & Ofori, I. (2022). Computer-aided multiclass classification of corn from corn images integrating deep feature extraction. *Computational Intelligence and Neuroscience*, 2022, 2062944. doi:10.1155/2022/2062944

Koolagudi, S. G., Maity, S., Kumar, V. A., Chakrabarti, S., & Rao, K. S. (2009). IITKGP-SESC: Speech Database for Emotion Analysis. In S. Ranka et al. (Eds). *Contemporary Computing: Second International Conference* (pp. 485-492). Noida, India: Springer Berlin Heidelberg. doi:10.1007/978-3-642-03547-0_46

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica (Slovenia)*, 31(3), 249-268.

Krüger, F. (2016). *Activity, context, and plan recognition with computational causal behaviour models*. (PhD), University of Rostock, Institute of Communications Engineering, Rostock, Germany.

Lech, M., Stolar, M., Best, C., & Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers in Computer Science*, 2, 14. doi:10.3389/fcomp.2020.00014

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. New Jersey, USA: Wiley-IEEE Press.

Metlek, S., & Kayaalp, K., 2020. *Makine Öğrenmesinde, Teoriden Örnek MATLAB Uygulamalarına Kadar Destek Vektör Makineleri*. Ankara, Türkiye: İksad Yayınevi.

Milton, A., Roy, S. S., & Selvi, S. T. (2013). SVM scheme for speech emotion recognition using MFCC feature. *International Journal of Computer Applications*, 69(9), 34-39. doi:10.5120/11872-7667

Misra, H., Ikbal, S., Bourlard, H., & Hermansky, H. (2004, May). *Spectral entropy based feature for robust ASR*. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada. doi:10.1109/ICASSP.2004.1325955

Mitrović, D., Zeppelzauer, M., & Breiteneder, C. (2010). Chapter 3- Features for content-based audio retrieval. In M. V. Zelkowitz (Ed.), *Advances in Computers, Vol. 78* (pp. 71-150). Burlington, USA: Elsevier. doi:10.1016/S0065-2458(10)78003-7

Mohamad Nezami, O., Jamshid Lou, P., & Karami, M. (2019). ShEMO: a large-scale validated database for Persian speech emotion detection. *Language Resources and Evaluation*, 53, 1-16. doi:10.1007/s10579-018-9427-x

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO Ist Project Report* (pp. 1-25). Paris, France: Icram.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5), 2902-2916. doi:10.1121/1.3642604

Rebala, G., Ravi, A., & Churiwala, S. (2019). *An Introduction to Machine Learning*. Cham, Switzerland: Springer.

Sonawane, A., Inamdar, M. U., & Bhangale, K. B. (2017, August). *Sound based human emotion recognition using MFCC & multiple SVM*. 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), Indore, India. doi:10.1109/ICOMICON.2017.8279046

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168-192. doi:10.1016/j.aci.2018.08.003

Tuncer, T., Dogan, S., & Acharya, U. R. (2021). Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. *Knowledge-Based Systems*, 211, 106547. doi:10.1016/j.knosys.2020.106547

Vyas, G., & Kumari, B. (2013). Speaker recognition system based on mfcc and dct. *International Journal of Engineering and Advanced Technology (IJEAT)*, 2(5), 167-169.

Yasar, A., Saritas, I., & Korkmaz, H. (2018). Determination of intestinal mass by region growing method. *Preprints*, 2018, 2018050449. doi:10.20944/preprints201805.0449.v1

Yasar, A. (2022). Benchmarking analysis of CNN models for bread wheat varieties. *European Food Research and Technology*, 249, 749-758. doi:10.1007/s00217-022-04172-y