SELCUK UNIVERSITY JOURNAL OF ENGINEERING SCIENCES

Journal homepage: http://sujes.selcuk.edu.tr

ISSN: 2757-8828

International Open Access

Volume 23
Issue 01

April, 2024

Research Article

Multi-lingual Speech Emotion Recognition System Using Machine Learning

Emel Colakoglu [a,*], ***Serhat Hizlisoy*** [b], ***Recep Sinan Arslan*** [c]

[a] *Kayseri University, Calculated Sciences and Engineering, Kayseri 38280, Turkey*
[b] *Kayseri University, Faculty of Engineering Architecture and Design, Department of Computer Engineering, Kayseri 38280, Turkey*
[c] *Kayseri University, Faculty of Engineering Architecture and Design, Department of Computer Engineering, Kayseri 38280, Turkey*

ARTICLE INFO

Article history:
Received 20 December 2023
Accepted 29 April 2024

Keywords:
Speech Emotion Recognition,
Machine Learning,
Emo-DB,
Logistic Regression,
Text-independent

ABSTRACT

Predicting emotions from speech in different languages with high accuracy has been a challenging task for researchers in recent years. When we delve into the studies conducted in this field, it is clear that researchers generally try to recognize emotions from speech in their traditional language. However, these studies cannot be generalized for multi-lingual environments around the globe. The Turkish speech emotional dataset, which was created for use in our previous studies, was further expanded for use in this study too. Emo-db dataset was also used to benchmark the success of the proposed model. Various pre-processing stages such as standardization, sorting and resampling were applied to the data in the datasets to increase the performance of the model. OpenSMILE toolbox, which is frequently encountered in studies, was used to obtain features that provide meaningful information corresponding to the emotion in speech, and thousands of features were obtained from emobase2010 and emo_large feature sets. 8 different machine learning algorithms were used in the model to classify 4 different emotions for the Turkish dataset and 7 different emotions for the Emo-db dataset. The best recognition rates were achieved with 92.73% and 96.3%, respectively, for the Turkish dataset consisting of 1099 records and the Emo-db dataset consisting of 535 records, using the Emobase2010 as a feature set and Logistic Regression as a classifier.

This is an open access article under the CC BY-SA 4.0 license.
(https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

The reactions that an event or situation arouses in people's inner world are called emotions. Emotions are divided into subgroups such as happiness, anger, sadness, excitement, calmness, and neutral [1]. Emotion is an important topic in communication. People can convey each other's feelings through both body language and voice. However, since it is not always possible to see and interpret body language, being able to interpret emotion from sound is an area of greater importance [2].

The most common method of communication between people is speech. Speech recognition like speaker recognition [3] [4], is based on the processing of speech signal data, as well as the development of acoustic and, where necessary, language models [5]. Since emotion analysis from speech, which is a sub-branch of the speech recognition problem, is an area that can be difficult to interpret in human relations, it is a very complex process to determine human emotions using machines. While there have been investigations into this problem in recent years, particularly in the domains of machine learning and deep learning, process success is still affected by language, region, age, cultural differences, and gender [6] [7].

In SER studies visual, auditory, and physiological parameters (heartbeat, and blood pressure) can be used [1]. The first studies in emotion recognition began in the 80s [7]. While the scope of the first studies focused only on whether the emotion was positive or negative; in parallel with technological developments, new studies have come to cover many different emotions and stress levels [8]. In recent years, studies in this field have both diversified and gained more momentum. Looking at the process of artificial intelligence in general, emotion recognition studies are a condition for

* Corresponding author. E-mail address: *emelcolakoglu@gmail.com*
How to cite this article: *C. Emel, H. Serhat and A. Recep Sinan., "Multi-lingual Speech Emotion Recognition System Using Machine Learning,", Selcuk University Journal of Engineering Sciences, vol. 23, no. 1, pp. 1-11, April. 2024.*

more lively, friendly and emotional communication between machines and humans also contain essential research expectations [9].

The main uses of speech emotion recognition are healthcare, education, entertainment, driving assistance, call centers, and internet-based learning. Basically, these studies are among the study groups that are difficult to analyze. The reason is that speech signals contain different frequencies and characteristics and have an unstructured state [10] [11] [12].

Essentially, the process of recognizing emotions from speech signals relies on extraction and selection of features and the determination of classifier algorithms. The features provide insight into the emotion embedded in the speech signal, playing a crucial role in SER. In such studies, both acoustic attributes derived from speech signals and non-acoustic attributes are used. Examples of acoustic features are harmonic noise ratio (HNR), Short-term Fourier Transform (STFT), Mel Frequency Cepstral Coefficient (MFCC), Discrete Fourier transform(DFT), Perceptual Linear Prediction Coefficients, mel-spectrogram, pitch frequency, Wavelet transform (WT) etc. [13] Features such as image, text, video, and EEG signals are examples of the non-acoustic group. Much like features, classifiers are pivotal in enhancing emotion recognition accuracy through the analysis of speech signals. [14]. Choosing a good classifier and training it correctly is key to successful machine learning applications. The importance of classifiers plays a fundamental role in making sense of data, making decisions, and improving business processes in many industries and applications.

To summarize, this study makes the following key contributions:

1.In speech emotion recognition studies, emotion datasets prepared by professional speakers and in controlled environments that generally serve this purpose are used. What differentiates this study is the use of natural recordings compiled from Turkish TV series and movies.

2.Another point that differs from the studies in the literature is that the model was developed on both the speaker and the text-independent dataset.

3.In recent years, speech emotion recognition studies have focused more on the deep learning side due to high success rates. With this study, high accuracy values were achieved with machine learning algorithms.

Below is the organization of the remaining sections in the article: Section 2 discusses the literature on speech emotion recognition. Section 3 describes the architectural structure of our study. Section 4 outlines the obtained results and provides a comparison with previous studies. The conclusion and discussion are contained in Section 5.

2. Related Works

In the scope of this study, we employed machine learning algorithms to perform speech emotion recognition using the dataset we assembled. In addition, a frequently referenced dataset in academic literature, the EmoDB was also analyzed to test the model performance. Upon reviewing the literature in the domain of SER, we noted a multitude of studies. In this section, the most recent and comprehensive of these studies are summarized.

Agarla et al. (2023) [15], conducted a study on speech emotion recognition using 7 datasets consisting of 5 different languages. These datasets are: one in French (CaFE), three in English (TESS, RAVDESS, SAVEE), Persian (ShEMO), German (EmoDB) and Italian (EMOVO). A cross-language semi-controlled model is targeted. 3 different deep architectures were taken into account. These are EmotionCNN, BYOL-S, and HuBERT. In EmotionCNN, Cochleagrams and classifier CNN are used as features, while in BYOL-S the feature is MFCC and the classifier CNN and HuBERT are the waveforms feature and Transformer classifier. There is also an adaptation module to make use of the unlabeled expressions of the new language. The results were grouped as cross-lingual, multilingual, and pseudo-labeled. The highest accuracy value in the new language was obtained with the HuBERT architecture and in the German language (89.0%). Utilizing a Transformer architecture, the HuBERT method adjusted to the new domain through the application of a pseudo-tagging strategy on unlabeled phrases.

Langari et al. (2020) [16] used MFCC and DFrFT as features in their study. Genetic Algorithm and Cuckoo Search were used together in feature selection. The study incorporates PDREC, SAVEE, and EMO-DB as its datasets. The classifier used is SVM. Accuracy values of SAVEE (80%, 80%), EMO-DB (97.57%, 97.21%), and PDREC (91.46%, 87.31%) were achieved (WA, UA).

Issa et al. (2020) [17] incorporated EMO-DB, RAVDESS and IEMOCAP datasets into their research. The extracted features are Spectral contrast feature, MFCC, Mel scale spectrogram, Tonnetz presentation and Chromogram. The classifier is 1D CNN. The highest accuracy value was achieved in EmoDB with 96.34%.

Abdulmohsin et al. (2021) [18] utilized the datasets RAVDESS, SAVEE, and Emo-DB in their research. In the pre-processing process, "Dataset Reordering" and "removing harmless energy" were applied to the data. The features extracted from the data are Fourier function, Zero Crossing, entropy, energy, deviation of the ZC, Haar, the loudness function, MFCC, MATLAB fitness function, GTCC the pitch function and harmonic ratio. F-score was used in feature selection and artificial neural networks were used in classification. For the RAVDESS, Emo-DB and SAVEE datasets, accuracy rates in classification reached in this study are 86.1%, 96.3%, and 91.7%, respectively.

Ancilin et al. (2021) [19] used SAVEE, Berlin, eNTERFACE, Ravdess, EMOVO and Urdu datasets in this study. The features are MFMC, MFCC, LPCC, LFPC and

Mel frequency magnitude coefficient. SVM is classifier. Best accuracy values by dataset: SAVEE (75.63%), Berlin (81.50%), Ravdess (64.31%), eNTERFACE (56.41%), EMOVO (73.30%), Urdu (95.25%).

Guo et al. (2022) [20] used EmoDB and IEMOCAP datasets in this publication. Extracted features MGDCC wasa spectrogram of DRP, RP and magnitude. The classifier was KELM. The highest success rate was achieved with the multi-channel MCMA model. These values were EmoDB (94.02%; 93.66%) and IEMOCAP (57.58%; 58.07%) (WA; UA)

Liu et al. (2020) [21] used spectrograms as features and ELM (overlearning machine) as classifiers. Within the model, a local-global system wasemployed, comprising both a local representation learning module and a global representation learning module. This system was designed to analyze both global and local information in speech. The research presented a newly proposed conscious deep representation learning method. Utilizing a CNN-based approach, the local representation learning module is employed. Meanwhile, in the global representation learning module, a CapsNet model named DenseCap was developed, featuring a novel routing algorithm. The dataset used was IEMOCAP. The accuracy value obtained with the proposed model (TFCNN+DenseCap+ELM) is 70.34%(WA) and 70.78%(UA).

In the research of Wang et al. (2020) [22], both the EMODB and EESDB datasets were part of the dataset collection for the study. The extracted features were MFCC and WPC. For feature selection, the Sequential Floating Advanced Search (SFFS) method was utilized, and SVM was employed as the classifier. The highest accuracy value was obtained in the EMODB dataset with 79.5%.

Jia et al. (2021) [23] used IEMOCAP and a private dataset in their research. The extracted feature wasMFCC and the classifier was BLSTM. The highest accuracy values were (78% (SER), 70% (IEMOCAP)).

Zhao et al. (2021) [24] used the IEMOCAP dataset in these studies. The features obtained from this dataset werelog-Mel spectrograms and TCN was the classifier and LSTM for the decoder. BLSTM was used in comparison. The highest accuracy values were WA (65.0%) and UA (66.1%).

Singh et al. (2023) [25] created two models for each emotion according to female and male gender in their training methods. Using this educational methodology, a set of 16 emotion models was created by CNN. RAVDESS was the dataset used in this study. MFCC and its variants were used as features. In this context, an accuracy value of 72.07% was achieved.

In the Ntalampiras (2021) [26] study, they used EmoDB as the dataset, temporal modulation, Log-mel spectrogram features as the feature, and SNN (Siamese Neural Network) as the classifier. The highest accuracy value obtained is 82.1%.

Li et al. (2023) [27] used Mel and VTMel spectrograms as input in their study. CNN-DBM was preferred to obtain features from these spectrograms, and SVM and LSTM were preferred as classifiers. This study utilizeed the datasets EMO-DB, SAVEE, and RAVDESS. The best accuracy values in all three datasets were obtained with the LSTM classifier. EMO-DB (96.82%), SAVEE (91.56%) and RAVDESS (82.34%).

3. Proposed Method

The architectural structure of the study is shown in Figure1. Using a dataset of 1099 records collected for this study, classification tasks are carried out through machine learning algorithms. Furthermore, the results of the EmoDB dataset, a commonly utilized dataset in speech emotion recognition literature, are outlined to assess our model's performance compared to previous studies.

To obtain the features, emo_large and emobase2010 sets, which are the feature sets that come with the OpenSMILE application, are used. In this way, 2 different feature sets obtained for 2 datasets.

3 different filters were applied as a group on the obtained feature groups and the results were compared. These filtering groups are sorting-standardization; sorting-standardization-resample. In the Weka application, sorting and resampling filters for feature sets and standardization were applied in the Python environment.

Feature sets were divided into groups as 90-10%, 80-20%, and 10-fold cross-validation as training and testing.

Python language was used in the development process. Python has been frequently used in analysis with machine learning algorithms in recent years. Python is a language with rich libraries in many areas such as data analysis, classification, natural language processing, data visualization, machine learning and deep learning. Within the scope of this study, Pandas, OS, Matplotlib, Sklearn, Numpy, Imblearn and Random libraries were used. The Sklearn library contains many ML algorithms and has a very detailed documentation.

Within the scope of the study, 2 datasets were used. These are the dataset we created ourselves and EmoDB. While creating our dataset, TV series and movies in Turkish were examined. 68 unique TV series and movies are included in the scope. In this number, different episodes of the same series are counted as one. As a result of this examination, records were grouped according to 4 different emotions (happy, angry, sad and neutral). The samples were recorded with the Camtasia Recorder application, and then all operations such as editing, cutting, separating image files from speech signals, noise cleaning, amplifying the volume, cleaning the gaps were also done with Camtasia Studio. All recordings are in wav format, with a sampling frequency of 22050 and are single-channel. Table 1 lists the emotion distributions of our 1.099 record dataset.

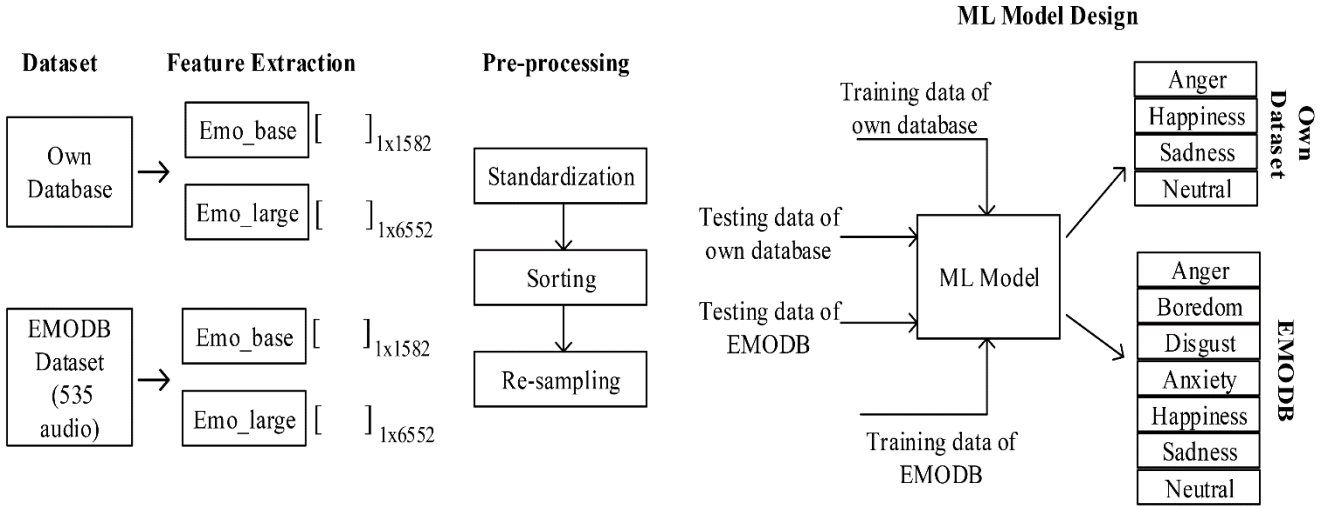

Figure 1: Architectural Structure

Table 1: Distribution of Emotion in a Data Set of 1.099 Records

Angry	Happy	Sad	Neutral
276	274	278	272

Table 2: Emotion Counts in EmoDB(F: Fear, D: Disgust , S: Sad, H: Happy, B: Bored , N: Neutral ,A: Angry)

	F	D	S	H	B	N	A
Male	36	11	25	27	35	39	60
Female	33	35	37	44	46	40	67
Total=535	69	46	62	71	81	79	127

EmoDB, on the other hand, is an emotion dataset prepared in German. It contains a total of 535 records and 7 emotions. These emotions are happiness, neutral, anger, sadness, boredom, fear, and disgust.

Table 2 lists the emotion distribution of the EmoDB dataset. The entire dataset consists of 5 short and 5 long sentences, meaningful and meaningless. It has voiced 5 female and 5 male professional speakers.

When creating the dataset, we paid attention that there is no music in the TV series and movies examined during the preparation phase, the speeches that will reflect a single emotion are selected throughout the recording, there is only one speaker throughout the entire recording, that the sound is quality, that the text of the speech is different, there are no gaps or it does not interfere with the perception of the emotion, and Attention was paid to the use of recordings of different actors and actresses in order to support the speaker independent structure. In order to achieve this, many TV series and movies were watched and eliminated.

The OpenSMILE application used to obtain the feature has a flexible and modular structure for ML and signal processing. It is basically aimed at obtaining features from audio signals. However, it can also be used to analyze visual signals, physical signals, and signals from other physical sensors. C++ has also been developed. It can run on a variety of systems such as Windows, Linux, and MacOS. OpenSMILE includes a powerful C API and allows for seamless integration with other applications [28]. Within the scope of the study, OpenSMILE was integrated through the Python interface and the emobase2010 (1,582) and emo_large (6,552) feature sets that came with OpenSMILE were used. Emo_large feature set contains 6.552 feature values [29]. In the Emobase2010 feature set, there are 34 LLD features. With these 34 features and their deltas, 21 different functions were applied in a total of 68 feature sets obtained and 1.428 features were obtained. There are also 4 pitch-based features. 19 different functions were applied to the 8 features obtained with the delta of these and 152 features were obtained. By adding the number of most recent Pitch starts and sum of input time, the feature value is 1.582 [28].

The sort operation used in the preprocessing process is the ascending or descending ordering of certain features. In this way, it is possible to identify missing or distorted data and analyze the dataset. The "resample" process is a method of data transformation. It is used in the fields of data analysis and machine learning. It allows to change the size or sample of the existing dataset. It is generally used to organize unbalanced data sets [30]. Standardization

(also known as Z-Score Normalization) is a statistical process used to eliminate differences between different variables or characteristics of data and to make these variables comparable on the same scale. This concept is basically based on standard deviation and variance values. It is calculated as in formula 1 [31][36].

$$Z = \frac{(x-\mu)}{\sigma} \quad (1)$$

Within the scope of this study, sorting, resampling and standardization filters are used with different grouping and the results compared.

In feature selection, the methods can be classified into three primary groups. These are filters, embedded methods, and wrappers [32]. Within the scope of the study, Information Gain and Correlation filters from the filter group; In embedded methods, the effects on performance were examined by using Decision Trees, Lasso, Ridge, LDA, PCA, Random Forest, Gradient Enhancer, and they were not included in the scope of the study because no additional success increase was observed in the result of ML algorithms with high recognition rate.

In the classification process, the results of 8 different ML algorithms were listed and compared. Support Vector Machines, Linear Discriminant Analysis, K-Nearest Neighbour, , Decision Tree, Naive Bayes (NB), Logistic Regression, Extra Tree and Random Forest classifiers were used in the analysis part made with ML algorithms. In Table 3, you can find the parameter values employed for these 8 classifiers, which differ from their default configurations.

Table 3: Differences in Parameter Assignments from Defaults in Classifiers

Classifier	Parameters Assigned Different from Default Value
DVM	Random_state=0, tol=0.001
KNN	N_neighbors=5
LDA	-
LR	Solver='lbfgs', max_iter=1,000, random_state=0
KA	Random_state=0
NB	-
EA	N_estimators=100, random_state=0
RO	Max_depth=2, random_state=0

4. Experimental Results

Speech emotion analysis was conducted in this study using both our prepared dataset and EmoDB. Our dataset consists of 1.099 records, while EmoDB contains 535 records. With OpenSMILE, the features have been extracted. Two sets of features were used, Emo_large and Emobase2010. Results were obtained with 8 different classifiers on these two feature sets and two datasets. Dataset division is conducted with 80-20%, 90-10%, and 10-fold cross-validation, creating training and testing components. The results of all analyses are listed in Table 4.

Table 4: Results of All Algorithms

		Sort+Std			Sort+Std+Resample		
Dataset	**CLS**	**90-10 %**	**80-20 %**	**10 Floor**	**90-10 %**	**80-20 %**	**10 Floor**
Our Data Set emo_large (6552)	DVM	80.00	81.36	79.53	**88.18**	**86.36**	90.53
	KNN	70.91	69.55	71.70	70.00	72.27	74.70
	LDA	**86.36**	79.09	73.53	81.82	81.36	86.72
	LR	85.45	**86.82**	**80.44**	**88.18**	**86.36**	**92.63**
	KA	53.64	55.45	51.05	75.45	75.00	79.43
	NB	56.36	57.27	52.96	50.91	52.27	56.96
	EA	77.27	78.64	74.43	87.27	87.27	87.99
	RO	60.91	64.55	53.96	59.09	63.64	59.24
Our Data Set Emobase 2010(1582)	DVM	**81.82**	**79.09**	**78.07**	90.00	**90.45**	89.35
	KNN	72.73	65.45	66.06	68.18	70.45	71.61
	LDA	49.09	56.82	46.67	84.55	84.55	84.26
	LR	72.73	74.09	74.98	**92.73**	**90.45**	**89.72**
	KA	54.55	51.36	48.77	77.27	79.55	81.89
	NB	69.09	60.45	55.69	60.00	60.91	61.69
	EA	75.45	73.64	70.88	87.27	88.64	86.99
	RO	65.45	62.73	53.05	60.91	61.82	59.70
EmoDB emo_large (6552)	DVM	81.48	84.11	82.99	87.04	92.52	91.60
	KNN	75.93	71.03	70.47	74.07	81.31	73.63
	LDA	**92.59**	**85.98**	**89.72**	88.89	88.79	89.90
	LR	83.33	82.24	88.99	**92.59**	**93.46**	**94.21**
	KA	66.67	59.81	58.69	79.63	83.18	80.39
	NB	68.52	71.03	67.26	75.93	77.57	74.60
	EA	83.33	78.50	81.11	87.04	90.65	88.79
	RO	68.52	63.55	59.06	66.67	65.42	62.99
EmoDB Emobase 2010(1582)	DVM	87.04	82.24	85.22	90.74	91.59	92.16
	KNN	81.48	77.57	73.63	75.93	83.18	74.96
	LDA	87.04	83.18	81.13	92.59	91.59	90.48
	LR	**88.89**	**87.85**	**87.83**	**96.30**	**95.33**	**94.02**
	KA	66.67	51.40	61.64	74.07	77.57	80.39
	NB	72.22	74.77	74.56	87.04	85.05	79.83
	EA	87.04	84.11	81.32	87.04	92.52	90.29
	RO	74.07	70.09	58.87	74.07	76.64	59.64

When we examine the results according to Table 4 in our dataset, the highest recognition rate after sort and standardization was obtained in emo_large feature set.

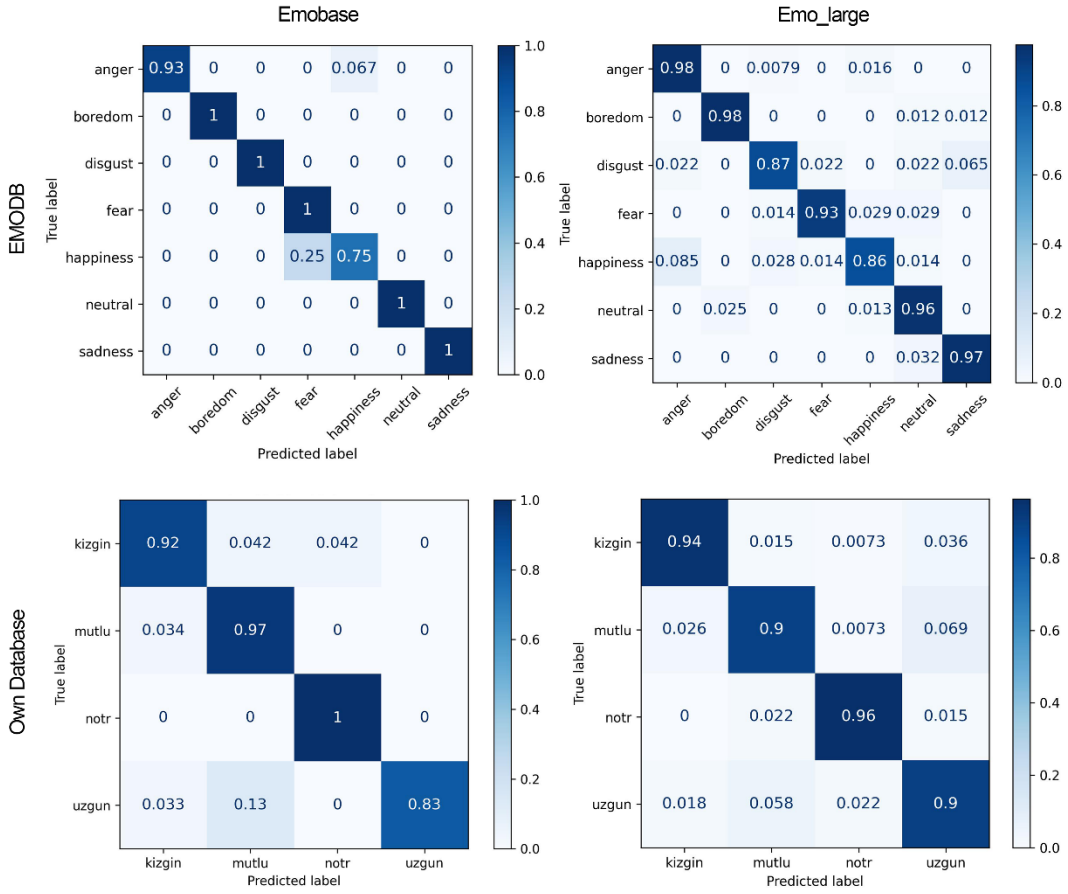

Figure 2: Confusion Matrices (Our Dataset and EmoDB)

The best value was obtained with the Logistic Regression classifier and 80-20% training-test data segmentation and an accuracy value of 86.82%. In the EmoDB dataset, the highest accuracy value after sorting and standardization was obtained in the emo_large dataset and its value was 92.59%. The classifier from which this value is obtained is Linear Discriminant Analysis and 90-10% training-test segmentation.

After sort, standardization and resample pre-processing, the highest accuracy value in our dataset is in the emobase2010 feature set, at 90-10% data breakdown it was also obtained by Logistic Regression (92.73%). In the EmoDB dataset, the highest accuracy value was obtained in the emobase2010 dataset, 90-10% data breakdown and Logistic Regression after 3 filters. The accuracy value is also 96.30%.

Considering the result as a whole, the highest accuracy rate in our dataset is 92.73% and the highest accuracy rate in the EmoDB dataset is 96.30% using the same model.

In both datasets, the highest accuracy value was obtained by Logistic Regression. Table 5 lists the parameters used in the Logistic Regression classifier.

When Table 5 is examined, the parameters used differently from the default value are random_state, solver and max_iter. When dividing data into training and testing, a different splitting process occurs each time. With the constant value given to Random_state, it is ensured that this division is always done in the same order. In the Solver parameter, the optimization method to be used is defined. Max_iter determines the maximum number of times the algorithm will repeat.

Table 5: Logistic Regression Parameters

Parameter List	**Value Used**	**Parameter Range**	**Default Value**
penalty	L2	'L1', 'elasticnet', 'L2','none	L2
dual	False	True, False	False
Tol	1e-4	Float	1e-4
C	1.0	Float	1.0
fit_intercept	True	False, True	True
warm_start	False	False, True	False
random_state	0	Int RandomState instance, none	none
Solver	lbfgs	'Newton-CG', 'lbfgs', 'Saag', 'Saga', 'Liblinear',	Liblinear
max_iter	1000	Int	100
class_weight	none	'balanced', none,'dict'	none
multi_class	ovr	'ovr', 'multinomial', 'auto'	ovr
Verbose	0	Int	0
n_jobs	None	Int, None	None
l1_ratio	None	Float, None	None
intercept_scaling	1	Float	1

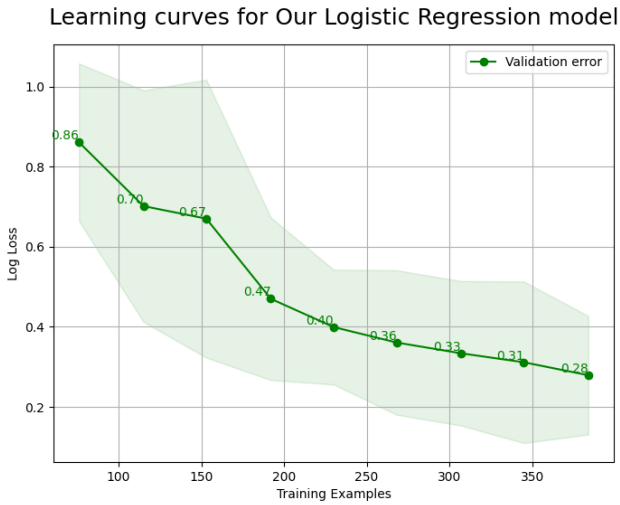

Figure 3: Learning Curve (Model Loss for EmoDB)

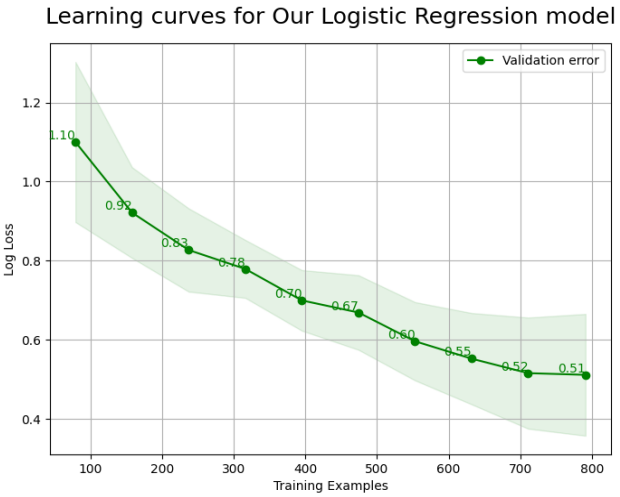

Figure 4: Learning Curve (Model Loss for Our DB)

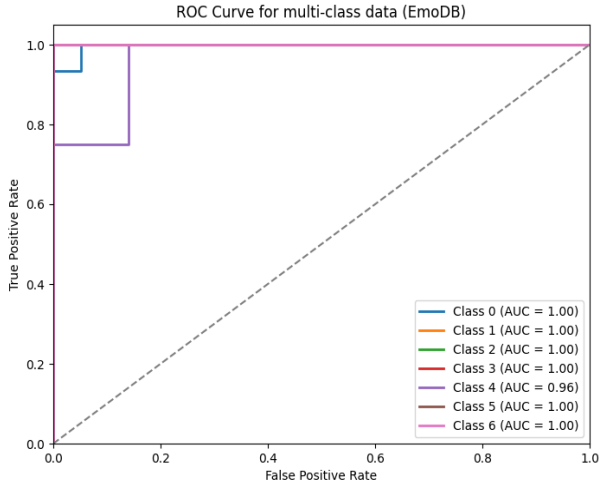

Figure 5: ROC Curve (EmoDB)

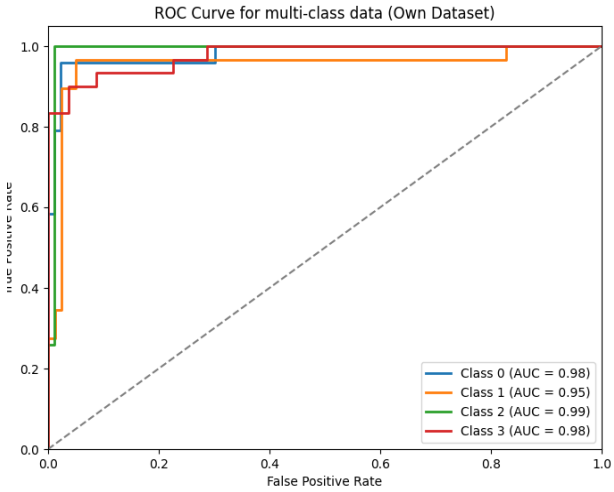

Figure 6: ROC Curve (Our Dataset)

When both datasets are examined on the basis of binary feature sets, the confusion matrices for the highest accuracy values obtained are shown in Figure 2.

In the emobase2010 feature set for the EmoDB dataset, the highest recognition rate was obtained by logistic regression (96.30%). On the basis of emotion, the lowest recognition rate is happiness with 75%, and the highest recognition rates are boredom, disgust, fear, neutral and sadness with 100%. Again, in the EmoDB dataset, 94.21% accuracy was obtained by logistic regression for emo_large feature set. Happiness again had the lowest accuracy value (86%), while anger and boredom had become the emotion with the highest accuracy at 98%.

In our dataset, the highest recognition rate in logistic regression was obtained with 92.73% for the emobase2010 feature set. On the basis of emotion, neutral emotion was the highest value with 100%; The feeling of sadness has the lowest accuracy value at 83%. In our dataset, 92.63% accuracy was achieved for emo_large feature set. In the emotion breakdown, neutral again had the highest value with 96%; Happy and sad received the lowest value at 90%.

Figure 3 and Figure 4 show the test loss chart for EmoDB and our dataset. These curves are used to evaluate how models will perform with a varying number of test samples. Ideally, test loss should be low and steadily decreasing. In both of these graphs, there is a decrease in the values over the course of the sample. It is observed that

Table 6: Comparison with Previous Studies

Work	Dataset	Features	Cls	Best Acc (%)
Agarla et al. [15]	RAVDESS, SAVEE, TESS, CaFE, EmoDB, EMOVO, ShEMO	Cochleagrams, MFCC, Waveforms	CNN Transformer	89
Langari et al.[16]	EmoDB SAVEE PDREC	DFrFT, Kepstral Features	SVM	97.57
Gan et al [33]	SAVEE, IEMOCAP, EMO-DB, MSP-IMPROV, eNTERFACE05	Spectrograms	Bi-GRU CNN Multiple Fusion	89.50
Issa et al [17]	IEMOCAP, RAVDESS EMO-DB	Tonnetz Notation, MFCCs, Spectral Contrast Feature, Chromagram, Mel Scale Spectrogram	1D CNN	96.34
Abdulmohsin et al. [18]	RAVDESS SAVEE EmoDB	Harmonic Ratio, Energy, Entropy, Zero Transition Deviation, MFCC, Pitch, Loudness, Conformity, Fourier, ZC, Energy Deviation, GTCC	Neural Network	96.30
Chen et al. [34]	RAVDESS and IEMOCAP	INTERSPEECH 2010 Paralinguistic Challenge feature set	CNN BiLSTM	88.88
Ancilin et al.[19]	SAVEE, Ravdess, eNTERFACE Berlin, Urdu, EMOVO,	MFCC, MFMC, LFPC, LPCC	SVM	95.25
Guo et al. [20]	EmoDB IEMOCAP	DRP, Magnitude Spectrogram MGDCC, RP,	KELM	94,02
Liu et al. [21]	IEMOCAP	Spectrograms	ELM	70.78
Wang et al. [22]	EmoDB	MFCC, WPC, SFFS	SVM	79.50
Shahin et al. [35]	SAVEE, The Arabic Emirati, RAVDESS	Mel spectrogram, MFCC, Chroma, and Root Mean Square (RMS), ZCR	KNN, MLP, DT, LR, NB, SVM-RBF	89.17
Jia et al. [23]	Private SER	MFCC	BLSTM	78.00
Zhao et al. [24]	IEMOCAP	Three-dimensional Log-Mel Spectrogram	LSTM, A TCN WITH 1D convolutional modules,BLSTM	66.1
Singh et al. [25]	RAVDESS	MFCC	CNN	72.07
Ntalampiras[26]	EmoDB	Log-mel Spectrogram Temporal Modulation	Siamese Neural Network	82,10
Li et al [27]	EMO-DB, SAVEE, RAVDESS	Mel and VTMel Spectrograms	SVM LSTM	96.82
Recommended Method	**Our Dataset EmoDB**	**Emo_large Emobase2010**	**LR**	**92.73 (OwnDB) 96.3 (EmoDB)**

test loss tends to increase in models with overfitting. In our model, it can be seen from the graph that there is no such problem. If the test loss value is high or an overfitting condition is observed, it would be appropriate to try correction methods to improve the model. However instead of commenting only by looking at the loss graph, other metrics (accuracy, sensitivity, specificity, ROC Curve, etc.) should also be used jointly in model evaluation.

Figure 5 and Figure 6 show the ROC Curve graphs over the Logistic Regression classifier, where the highest accuracy values are obtained for our dataset and EmoDB. These graphs are useful for evaluating class-based performances and observing differences between classes. In both graphs, AUC values are close to 1. This indicates that the model has a successful classification ability. The fact that the curves in the graphs look similar explains that the classes are closely predicted. The fact that the curves

are located above the graph diagonal indicates that the classifier does not make a random prediction.

Recent studies of emotion recognition from speech are also listed in Table 6.

According to Table 6, if we evaluate the previous studies in the literature in general, high accuracy rates have been achieved in the datasets prepared for this purpose. Again, higher recognition rates were obtained in DL algorithms. MFCC and spectrograms were generally used as features. SVM is generally preferred as a classifier on the ML side and CNN is generally preferred as a classifier on the DL side Our own model has reached a high accuracy value by using ML algorithms. Performing a study dedicated to the Turkish language is immensely valuable, with the goal of achieving better performance.

5. Conclusions

In this study, the Turkish emotion speech dataset (4 emotions) that we prepared was used. In addition, EmoDB (7 emotions), which is the most preferred dataset in the literature in general, was included as a dataset in the analysis in order to evaluate the success of the model.

Opensmile tool was used to obtain features from both datasets. 8 different feature sets that come with Opensmile have been analyzed and the working process has been continued using emobase2010 and emo_large. The model was built entirely on machine learning algorithms and 8 different machine learning algorithms were used and compared.

In the process, the sequencing, standardization and re-sample pre-processing stages were applied to the dataset. A positive effect on performance has been obtained. Many different methods and algorithms have been tried in the feature selection phase, but no additional recognition increase rate has been observed on the results obtained with a high recognition rate. Therefore, the study was carried out but not included in the model. In addition, parameter optimization has been a situation that has a positive effect on performance.

The fact that the study was carried out on a Turkish dataset, the development of a model independent of both text and speaker, and the high recognition rates were achieved with machine learning algorithms was a situation that differentiated and strengthened the study.

As a result of all the analyzes, emotion recognition rates of 92.73% in our own dataset and 96.3% in the EmoDB dataset were reached. Both results were obtained by Logistic Regression.

From now on, we plan to develop a model with deep learning algorithms on the same dataset. However, the number of samples that we think is sufficient for machine learning needs to be increased in the process of deep learning study. Again, in subsequent studies, an increase in the number of emotions analyzed can be achieved in order to make a full comparison with the EmoDB dataset on the basis of emotion.

References

[1] S. Hizlisoy and Z. Tufekci, "Türkçe Müzikten Duygu Tanıma," *Avrupa Bilim ve Teknoloji Dergisi*, Vols. 6-12, 2020.
DOI: https://doi.org/10.31590/ejosat.802169

[2] A. Tepecik and E. Demir, "Türkçe Ses Kayıt Verilerinin CountVectorizer ve TF-IDFVectorizer Yöntemleri ile BERT Modelleri Olarak Google Colab Platformunda ve RapidMiner'da Makine Öğrenmesi Algoritmalarıyla Analizi," *Fırat Üniversitesi Fen Bil. Dergisi*, vol. 34(1), pp. 19-29, 2022.

[3] R. S. Arslan, N. Barisci, N. Arici and S. Kocer, "Detecting and correcting automatic speech recognition errors with a new model," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 29, no. 5, pp. 2298-2311, 2021.

[4] S. Hızlısoy And R. S. Arslan, "Text Independent Speaker Recognition Based On Mfcc And Machine Learning," *Selcuk University Journal Of Engineering Sciences*, Vol.20, No.3, Pp.73-78, 2021.

[5] R. S. Arslan and N. Barışçı, "The effect of different optimization techniques on end-to-end Turkish speech recognition systems that use connectionist temporal classification," *in 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, 2018.

[6] K. B. Bhangale and M. Kothandaraman, "Speech emotion recognition using the novel PEmoNet (Parallel Emotion Network)," *Applied Acoustics*, https://doi.org/10.1016/j.apacoust.2023.109613, Volume 212, September 2023, 109613 2023.

[7] E. Çolakoğlu, "Speech Emotion Recognition In A New Turkish Emotional Speech Dataset," Kayseri University, Institute Of Graduate Education, Msc Thesis., Kayseri, 2023.

[8] C. Parlak and B. Diri, "Farklı Veri Setleri Arasında Duygu Tanıma Çalışması," *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik dergisi*, vol. 16, no. 48, pp. 21-29, 2014.

[9] Z.T. Liu, B.H. Wu, M.T. Han, W.H. Cao and M. Wu, "Speech emotion recognition based on meta-transfer learning with domain adaption,"*Applied Soft Computing,* Volume 147, November 2023, 110766 2023.

[10] E. Çolakoğlu, S. Hızlısoy and R. S. Aslan, "T-ser: An Efficient Speech Emotional Recognition Model For Turkish Language Based On Machine Learning Algorithms," in *Innovations And Technologies In Engineering*, Eğitim Yayınevi, 2022, pp. 106-127.

[11] R. S. Arslan and N. Barışçı, "Development of Output Correction Methodology for Long Short

Term Memory-Based Speech Recognition," *Sustainability,* vol. 11, no. 15, pp. 4250-4266, 2019

[12] S. Hızlısoy, E. Çolakoğlu and R. S. Arslan, "Speech-to-Gender Recognition Based on Machine Learning Algorithms," *International Journal Of Applied Mathematics Electronics And Computers,* no. 10(4), pp. 084-092, 2022.

[13] R. S. Arslan and N. Barışçı, "A detailed survey of Turkish automatic speech recognition," *Turkish journal of electrical engineering and computer sciences,* vol. 28, no. 6, pp. 3253-3269, 2020.

[14] S. P. Mishra, P. Warule and S. Deb, "Variational mode decomposition based acoustic and entropy features for speech emotion recognition," *Applied Acoustics,* Volume 212, September 2023, 109578 2023 https://doi.org/10.1016/j.apacoust.2023.109578.

[15] M. Agarla, S. Bianco, L. Celona, P. Napoletano, A. Petrovsky, F. Piccoli, . R. Schettini and I. Shanin, "Semi-supervised cross-lingual speech emotion recognition," *Expert Systems With Applications,* no. https://doi.org/10.1016/j.eswa.2023.121368. Volume 237, Part A, 1 March 2024, 121368, 2023.

[16] S. Langari, H. Marvi and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," *Informatics in Medicine Unlocked,* no. 20, 2020.

[17] D. Issa, M. F. Demirci and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control,* no. 59, 2020.

[18] H. A. Abdulmohsin, H. B. Abdul Wahab and A. M. J. Abdul hossen, "A new proposed statistical feature extraction method in speech emotion recognition," *Computers and Electrical Engineering,* no. 93, 2021.

[19] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Applied Acoustics,* no. 179, 2021.

[20] L. Guo, L. Wang, J. Dang, E. S. Chng and S. Nakagawa, "Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition.," *Speech Communication,* no. 136, pp. 118-127, 2022.

[21] J. Liu, Z. Liu, L. Wang, L. Guo and J. Dang, "Speech Emotion Recognition with Local-Global Aware Deep Representation Learning," in *ICASSP 2020 - 2020 IEEE Uluslararası Akustik, Konuşma ve Sinyal İşleme Konferansı (ICASSP)*, Barselona İspanya, 2020.

[22] K. Wang, G. Su, L. Liu and S. Wang, "Wavelet packet analysis for speaker-independent emotion recognition," *Neurocomputing,* no. 398, pp. 257-264, 2020.

[23] N. Jia and C. Zheng, "Two-level discriminative speech emotion recognition model with wave field dynamics: A personalized speech emotion recognition method," *Computer Communications,* no. 180, pp. 161-170, 2021.

[24] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, S. Sun, H. Wang, J. Tao and B. W. Schuller, "Self-attention transfer networks for speech emotion recognition," *Virtual Reality & Intelligent Hardware,* no. 3, pp. 43-54, 2021.

[25] V. Singh and S. Prasad, "Speech emotion recognition system using gender dependent convolution neural network," *Procedia Computer Science,* Volume 218, 2023, Pages 2533-2540, 2023. https://doi.org/10.1016/j.procs.2023.01.227

[26] S. Ntalampiras, "Speech emotion recognition via learning analogies," *Pattern Recognition Letters,* no. 144, pp. 21-26, 2021.

[27] J. Li, X. Zhang, F. Li and L. Huang, "Speech emotion recognition based on optimized deep features of dual-channel complementary spectrogram," *Information Sciences*. https://doi.org/10.1016/j.ins.2023.119649, Volume 649, November 2023, 119649 2023.

[28] Eyben, F., & Schuller, B. (2015). openSMILE:) The Munich open-source large-scale multimedia feature extractor. ACM SIGMultimedia Records, 6(4), 4-13.

[29] H. Perez and A. A. Torres, "Evaluation of quantitative and qualitative features for the acoustic analysis of domestic dogs' vocalizations," *Journal of Intelligent & Fuzzy Systems,* p. 5051–5061, 2019.

[30] S. Mozaffari, The power of resampling: Techniques for boosting your ML model performance. LinkedIn. https://www.linkedin.com/pulse/power-resampling-techniques-boosting-your-ml-model-sadaf-mozaffari/. (Accessed:30 October 2023)

[31] Editör Ekibi, "Matematik ile Makine Öğrenimi Standardizasyonu (Z-Score Normalization)," *Yapay Zekaya Yönelik — Multidisipliner Bilim Dergisi,* 2020.

[32] V. Bolón, N. Sánchez and A. Alonso, "A review of feature selection methods on synthetic data," 2012, p. 483–519.

[33] C. Gan, K. Wang, Q. Zhu, Y. Xiang, D. K. Jain and S. García, "Speech emotion recognition via multiple fusion under spatial–temporal," *Neurocomputing,* no. https://doi.org/10.1016/j.neucom.2023.126623, pp. Volume 555, 28 October 2023, 126623, 2023.

[34] Z. Chen, M. Lin, Z. Wang, . Q. Zheng and C. Liu, "Spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms," *Knowledge-Based Systems,* no. https://doi.org/10.1016/j.knosys.2023.111077, pp. Volume 281, 3 December 2023, 111077, 2023.

[35] I. Shahin, O. A. Alomari, A. B. Nassif, I. Afyouni , . I. A. Hashem and A. Elnagar , "An efficient feature selection method for arabic and english speech emotion recognition using Grey Wolf Optimizer," *Applied Acoustics,* Volume 205, 30 March 2023, https://doi.org/10.1016/j.apacoust.2023.109279.

[36] S. Hızlısoy, R. S. Arslan and E. Çolakoğlu "Singer identification model using data augmentation and enhanced feature conversion with hybrid feature vector and machine learning" EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:14 https://doi.org/10.1186/s13636-024-00336-8