Engineering Science and Technology, an International Journal 24 (2021) 760–767

HOSTED BY

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: www.elsevier.com/locate/jestch

Full Length Article

# Music emotion recognition using convolutional long short term memory deep neural networks

Serhat Hizlisoy [a,*], Serdar Yildirim [b], Zekeriya Tufekci [a]

[a] Department of Computer Engineering, Çukurova University, Adana, Turkey
[b] Department of Computer Engineering, Adana Alparslan Turkes Science and Technology University, Adana, Turkey

ARTICLE INFO

*Article history:*
Received 12 May 2020
Revised 23 September 2020
Accepted 30 October 2020
Available online 14 November 2020

*Keywords:*
Music emotion recognition
Convolutional long short term memory deep neural networks
Turkish emotional music database

ABSTRACT

In this paper, we propose an approach for music emotion recognition based on convolutional long short term memory deep neural network (CLDNN) architecture. In addition, we construct a new Turkish emotional music database composed of 124 Turkish traditional music excerpts with a duration of 30 s each and the performance of the proposed approach is evaluated on the constructed database. We utilize features obtained by feeding convolutional neural network (CNN) layers with log-mel filterbank energies and mel frequency cepstral coefficients (MFCCs) in addition to standard acoustic features. Classification results show that the best performance is obtained when the new feature set is combined with the standard features using the long short term memory (LSTM) + deep neural network (DNN) classifier. The overall accuracy of 99.19% is obtained using the proposed system with 10 fold cross-validation. Specifically, 6.45 points improvement is achieved. Additionally, the results also show that the LSTM + DNN classifier yields 1.61, 1.61 and 3.23 points improvements in music emotion recognition accuracies compared to k-nearest neighbor (k-NN), support vector machine (SVM), and Random Forest classifiers, respectively.


## 1. Introduction

Music has always been in our lives by serving as many social and individual purposes [1]. Music emotion recognition (MER) is a subfield of music information retrieval (MIR) that aims to determine the affective content of music applying machine learning and signal processing techniques. Music emotion recognition systems have many application areas such as music suggestion systems (Spotify), automatic playlist generation, music therapy, and so forth. However, determination of the emotional category of music is quite challenging, and several issues need to be addressed such as emotion labeling of music excerpts, feature extraction and choose of the classification algorithm.

To build a music emotion recognition system, annotated emotional music database is needed. There are two approaches, i.e., categorical and dimensional, for labeling the emotions in music. In the categorical approach, emotions are characterized by discrete labels such as sad, happy, angry and fear [2]. In the second approach, emotions are represented in dimensional space. Russell [3] proposed a model that consists of two dimensions: valence and arousal. Thayer's model [4], which is adapted from Russell's circumplex model and consists of valence and arousal dimensions, is generally used for musical definition in the dimensional approach. These dimensions affect underlying stimuli that could influence mood responses [5–8]. Arousal axis shows emotions from calm to excited, and valence shows the measure of displeasure vs. pleasure. Categorical models are problematic because there is no consensus on category numbers. There is no such problem in the case of dimensional models. The advantage of dimensional models is the reduced uncertainty when compared with the categorical approach. This model provides a reliable way for people to measure emotion into two distinct dimensions. Therefore, the two-dimensional model is applied to annotate music excerpts.

Panda et al. [9] proposed a database of 903 audio clips labeled with 5 emotion clusters by the MIREX mood classification task [10]. Y.-C. Lin et al. [11] use tags from AMG to create a database of 7922 music annotated with 183 emotional tags [12]. A database created by Y.H. Yang et al. [13] contains 1240 Chinese pop music annotated with valence and arousal utilizing rankings scheme. Another large database consists of 744 music collected by M. Soleymani et al. [14]. In this database music excerpts were selected from the free music archive (FMA). AMG1608 contains 1608

* Corresponding author.
*E-mail addresses:* shizlisoy@cu.edu.tr (S. Hizlisoy), syildirim@atu.edu.tr (S. Yildirim), ztufekci@cu.edu.tr (Z. Tufekci).
Peer review under responsibility of Karabuk University.

https://doi.org/10.1016/j.jestch.2020.10.009
2215-0986/

30-second fragments of music in various music genres annotated by 665 subjects on valence and arousal [15]. Google introduced AudioSet [16] in 2017 that includes more than two million 10-s audio excerpts extracted from YouTube videos directly. The audio excerpts are labeled with 527 categories [17]. The database for emotional analysis of music (DEAM) consists of 1802 excerpts and full music annotated with valence and arousal [18]. In this study, we construct a Turkish emotional music database composed of 124 Turkish traditional music excerpts.

Several acoustic features have been explored by the researchers of MER. The features relevant to music and emotion were employed such as timing, dynamics, articulation, pitch [19,20], melody, harmony [21,22], tonality [23], and rhythm [19,24]. Classical audio features were also used such as energy [22], Zero Crossing Rate (ZCR) [19], MFCCs [19], log-mel filterbank energies, Linear Prediction Coefficients (LPC) [19], chromagram, centroid [19], spread, skewness, kurtosis, slope, roll-off [19,24], flux [19], and contrast. There are also tools available to extract those features such as openSMILE [25], MIRToolbox [22], YAAFE [26], jAudio [23], and Marsyas [27]. In this study, some of these tools are utilized to extract the aforementioned features. In addition, deep learning ensures an alternative way to find suitable features for the MER. Compared to existing acoustic features, researchers have explored the achievement of CNN to automatically learn emotional information via a deep architecture that can provide a higher-level representation from the annotated signal. CNN [17,28–32] has demonstrated successful results in various research areas including sound classification, image classification, speech recognition, and speech emotion recognition, in the last few years. However, a few studies were conducted to extract features using CNN [28,33,34] in music emotion recognition. In previous studies, CNN was fed with the raw audio signal, spectrogram or gammatone filterbanks [17,28,32,35]. MFCCs and log-mel filterbank energies are the most widely utilized features for speech recognition [35–38] since these features convey the most relevant information for speech. Same could also be true for music. Therefore, in this study, MFCCs and log-mel filterbank energies are fed to CNN as input to extract features.

SVM [11,39], k-NN [40], random forest (RF) [41], DNN [17,42], LSTM [17,42,43], gaussian mixture model (GMM) [44] and deep convolutional neural network (DCNN) [28] are used as a classifier for MER. Sarkar et al. [28] showed that DCNN classifier outperforms k-NN and SVM classifiers for MER. Sainath et al. [40] compared the performance of CLDNN with LSTM, CNN + LSTM, and LSTM + DNN for speech recognition that CLDNN outperforms the other classifiers. Therefore, in this study, we propose the CLDNN architectures that use MFFCs and log-mel filterbank energies as input. This architecture was previously applied to speech recognition and music detection [17,35–38,45]. There are also lots of studies that employ deep learning for MER [28,34,43]. In this study, we propose a technique that consists of four convolutional layers, LSTM layer and fully connected (FC) layers for music emotion recognition [35].

This paper is structured as follows. The details of the Turkish emotional music database are presented in Section 2. Section 3 provides an overview of the system including feature extraction and classification. In Section 4, it is described how the experiment was done. Section 5 presents the results and discussions. Finally, Section 6 mentions possible future works and concludes the paper.

## 2. New emotional Turkish music database

Creating a database is a labour-intensive and time-consuming process. Many databases in different languages have been created for music emotion recognition recently [14,46,47]. However, Turkish music is not among the studies for MER. In this study, a new Turkish emotional music database (TEM) is created. This database is composed of 124 Turkish traditional music excerpts with a duration of 30 s, representing the most salient part of the music. Generally, two approaches are employed in MER for database creation. In the first approach, only music signals are available to the evaluator, whereas in the second approach lyrics with metadata are also presented. In this study, we follow the first approach covering various genres and styles. The music excerpts are evaluated by 21 university students. The emotional content of each music is rated on a scale of [-5, 5] for valence and arousal dimensions. The valence dimension represents a positive to negative axis whereas arousal represents excited to calm axis. During the annotation process, evaluators were allowed to listen to music excerpts many times, and they were able to change their annotations through the program called AnnoEmo [48], which is also seen in Fig. 1.

All annotators have voluntarily participated in this study. Each music excerpt was annotated by 21 students who are highly knowledgeable about Turkish music. They are not only music listeners but also able to use instruments. Each annotator annotated all the 124 music excerpts in the dataset. The gender balance of annotators is distributed as 29% female and 71% male. The mean of annotations for each excerpt was calculated by taking the average over all annotators' decision for arousal and valence.

The average of the annotator's ratings was then used as the ground truth of each music excerpt. Fig. 2 represents the distribution of the average of the annotator's ratings of music excerpts on valence and arousal dimensions.

As can be seen from Fig. 2, music excerpts are distributed into three quadrants based on the distribution of the average of the annotator's ratings of music excerpts on valence-arousal dimensions. Therefore, in this study, we focus on 3-class music emotion classification. The upper right quadrant contains high arousal and positive valence (HAPV) emotion like happiness, while the lower left quadrant contains low arousal and negative valence (LANV) emotion like sadness, and the upper left quadrant contains high arousal and negative valence (HANV) emotion like anger as presented in Fig. 2. The distributions of music excerpts into three quadrants are 75, 38, and 11, respectively.

Also, Fig. 3 explains the distributions of the standard deviations. The averages of standard deviations are 0.243 and 0.276 for arousal and valence, respectively. To evaluate the agreement among annotators in terms of categorical classes, we utilize Krippendorf's $\alpha$ and Intraclass correlation coefficients as inter-rater reliability measures. The results given in Table 1 show that, the agreement between the annotators in terms of the categorical classes is quite high.

## 3. CLDNN architecture for music emotion recognition

This section explains the proposed approach based on CLDNN architecture for Turkish music emotion recognition. This architecture uses the output of CNN as features and LSTM + DNN as the classifier. Fig. 4 represents the proposed approach. The LSTM layer consists of 200 hidden units. The output of the LSTM is connected to 2 fully-connected (FC) DNN layers. FC layers transform the features into a more discriminative space where it is easier to find output targets. Each FC layer has 100 hidden units, and they are activated by the rectified linear units (RELU) which handle a threshold of 0 for the negative values [49]. Finally, a softmax output layer is added to obtain a final decision.

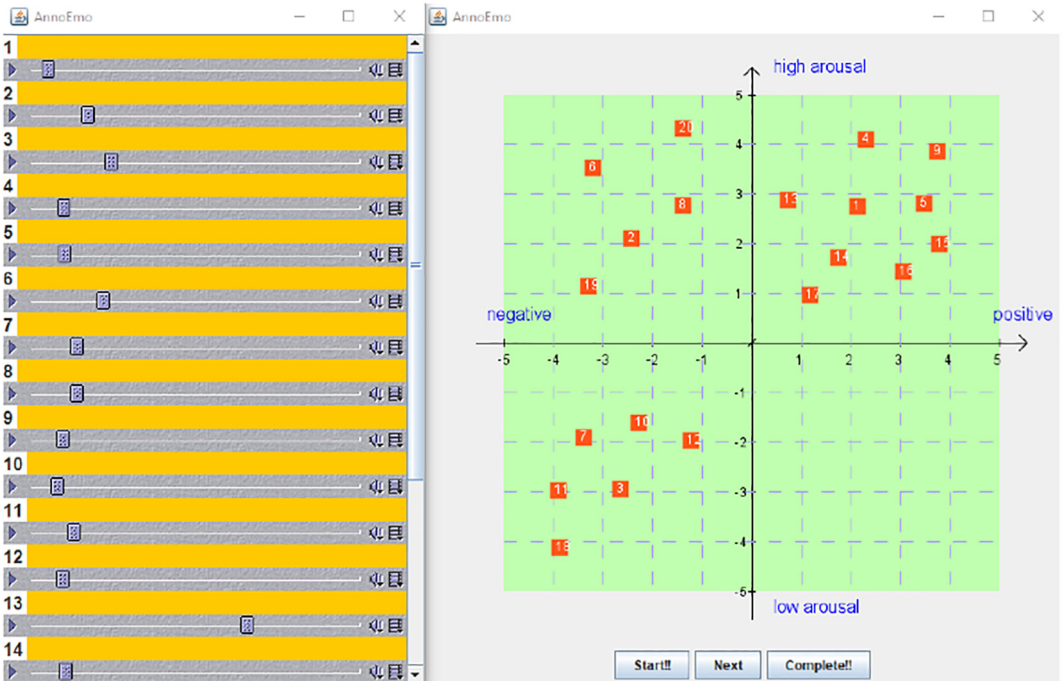


**Fig. 1.** A snapshot of AnnoEmo for annotating the arousal and valence values.

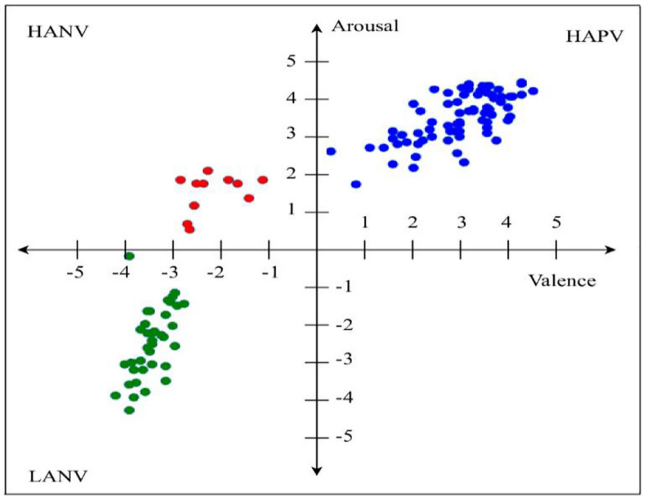


**Fig. 2.** The distribution of the average of the annotator's ratings of music excerpts on valence-arousal dimensions.

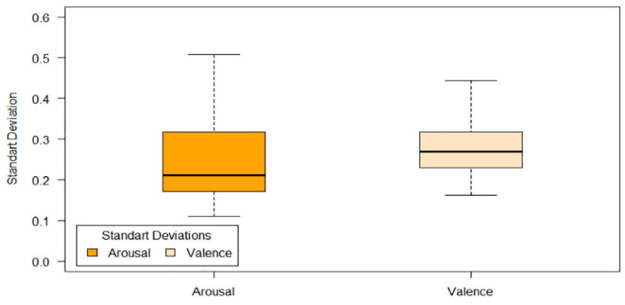


**Fig. 3.** The distributions of the standard deviations for each emotion primitives.

**Table 1**
Annotation agreement between annotators for the categorical classes.

| | Arousal | Valence |
|---|---|---|
| Krippendorff's α | 0.750 | 0.747 |
| Intraclass Correlation Co. | 0.808 | 0.806 |

### 3.1. One dimensional CNN for feature extraction

In this paper, we use a one-dimensional CNN to extract features for music emotion recognition. CNN's consist of one or more convolutional layers as a class of feed-forward neural networks. Each convolution layer has a number of filters to produce new feature maps. The depth of the network is very substantial to achieve better accuracy. However, as the depth of network increases, accuracy can reach the saturation point and then decrease. After convolution layers, pooling layers are usually added to decrease the number of parameters in the pattern by taking the average or maximum values for every subsection of the matrix. Finally, the flatten layer transforms the features into a $1 \times 768$ feature vector for every 30 s music excerpts. Fig. 5 and Table 2 illustrate the detailed structure of the proposed CNN, which is implemented utilizing Keras [50] with tensorflow backend.

In [28] raw audio signal and mel-scaled spectrogram are fed as input to the CNN to obtain features. Their results demonstrate that there is no significant improvement on MER performance using raw audio signal based CNN features [28]. Therefore, we do not use raw audio based CNN features for MER. Log-mel filterbank energies can be considered as a smooth version of the mel-scaled spectrogram. Therefore, log-mel filterbank energy based feature may give better performance than mel-scaled spectrogram based features. Log-mel filterbank energies and MFCCs are the most widely used features because they are considered to convey the most relevant information for speech recognition and MER. Therefore, we propose to use log-mel filterbank energies and MFCCs as input to CNN to obtain CNN based features.

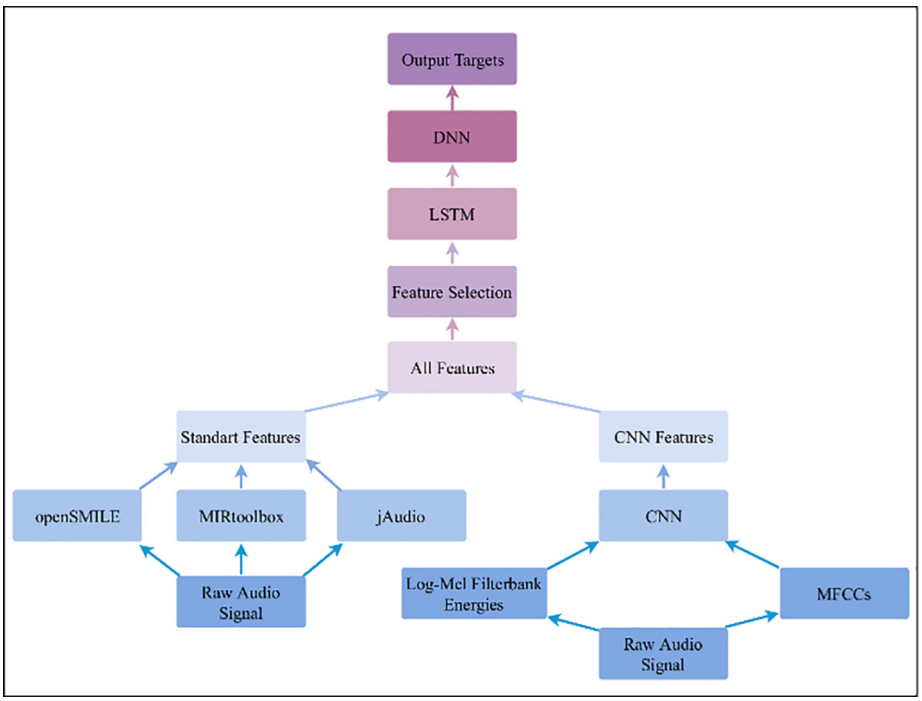


**Fig. 4.** The architectures of a CLDNN.

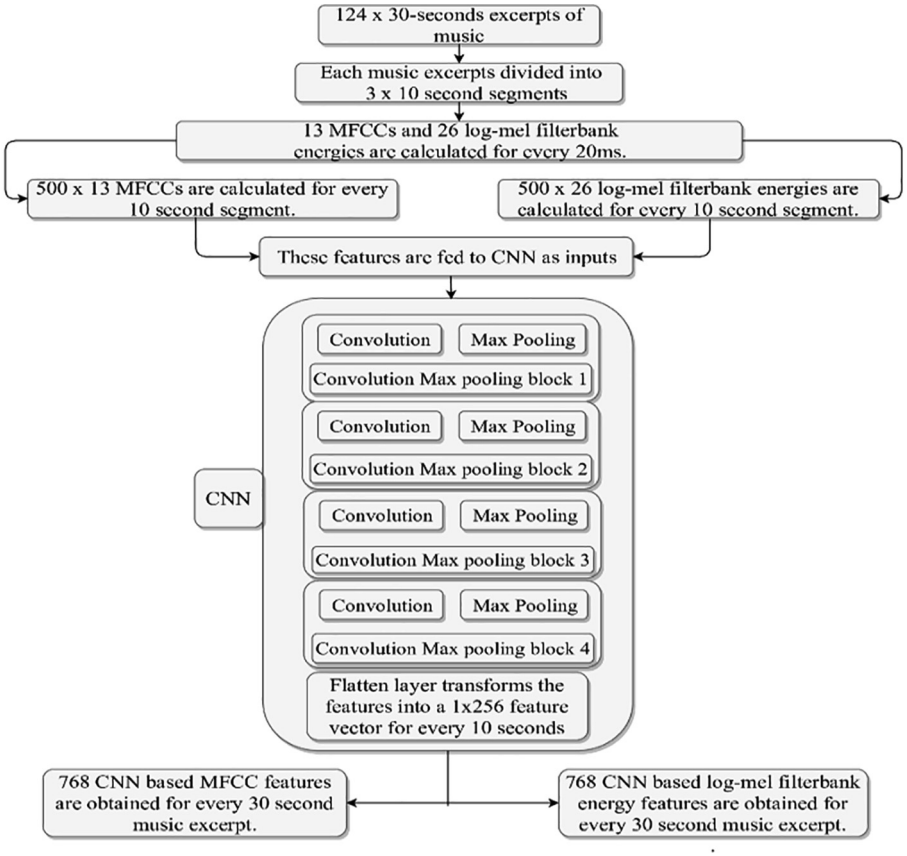


**Fig. 5.** The structure of a 1D CNN, consisting of convolutional layers, max-pooling layers, and flatten layer.

In this paper, log-mel filterbank energies and MFCCs are fed as input to CNN to obtain features. The number of samples increased from 124 to 372 by dividing into three 10-second of music to achieve sufficient data size for feature extraction applying deep learning, and then 13 MFCCs and 26 log-mel filterbank energies are calculated using 30 ms Hamming window every 20 ms. As a

**Table 2**
Configuration of the convolutional layers (C), pooling layers (P) and flatten layer (F) for the CNN considering different type inputs and different input sizes.

| Description | | Log-Mel Filterbank Energies | | | | MFCCs | | | |
|---|---|---|---|---|---|---|---|---|---|
| Input Shape | | 500 × 26 | | | | 500 × 13 | | | |
| | | Dim | Filter | Filter Size | Stride | Dim | Filter | Filter Size | Stride |
| Layers | C1 | 491 | 64 | 10 | 1 | 491 | 64 | 10 | 1 |
| | P1 | 122 | 64 | 4 | 4 | 122 | 64 | 4 | 4 |
| | C2 | 120 | 128 | 3 | 1 | 120 | 128 | 3 | 1 |
| | P2 | 30 | 128 | 4 | 4 | 30 | 128 | 4 | 4 |
| | C3 | 28 | 128 | 3 | 1 | 28 | 128 | 3 | 1 |
| | P3 | 7 | 128 | 4 | 4 | 7 | 128 | 4 | 4 |
| | C4 | 6 | 128 | 2 | 1 | 6 | 128 | 2 | 1 |
| | P4 | 2 | 128 | 3 | 3 | 2 | 128 | 3 | |
| | F | 1 | 256 | | | 1 | 256 | | |

result, 500 × 13 MFCCs and 500 × 26 log-mel filterbank energies are calculated for every 10-second of music. These features are fed to CNN as shown in Table 2 to generate 256 CNN based features for every 10-second of music. As a result of all these, we obtain 768 CNN based MFCC features and 768 CNN based log-mel filterbank energy features for every 30-second music. The reason for combining these features is to prevent overfitting of the 10-second of music, some of which are reserved as test and some as train. Therefore, 124 Turkish music excerpts with a duration of 30 s are used during the classification.

Fig. 6 displays the steps for computing log-mel filterbank energies and MFCCs. In the first step, music excerpts are divided into overlapping frames utilizing the hamming window. In the second step, the discrete time Fourier transform (DTFT) is calculated for each frame. Then the square of magnitude spectrum is computed. The fourth step gives filterbank energies which are computed by summing all energies in each mel-scaled triangular filterbank. The outputs of the fifth step are the log-mel filterbank energies which are computed by taking the logarithm of mel filterbank energies. In the last step, MFCCs are calculated by taking the discrete cosine transform (DCT) of log-mel filterbank energies.

### 3.2. Standard audio features

In this work, standard audio features are extracted utilizing publicly available tools such as MIRtoolbox [20], OpenSMILE [23], and jAudio [21]. The MIRtoolbox offers a set of functions to extract musical features from music excerpts such as timbre, tonality, and pitch [1]. Similarly, jAudio is also a tool to extract music-related features from the music excerpts, including harmonic change detection function (hcdf), mode, inharmonicity, tonal, chromogram, key clarity, tempo, and fluctuation. OpenSMILE is able to extract numerous low-level descriptors (LLDs) which are usually obtained from the short-time spectrum of the audio signal (e.g., energy, MFCCs, Mel-based features, pitch and spectral features such as roll-off, variance, centroid, flux, skewness, slope, kurtosis, spread, decrease, contrast). In this work, we extracted low-level features related to timbre and energy with openSMILE toolkit. In total, 7368 standard features (openSMILE = 6553 features, jAudio = 468 features, MIRtoolbox = 348 features) have been extracted for each music excerpt.

### 3.3. Feature selection

Feature selection is a method to decrease the feature size by choosing the most salient features. In this study, we apply the Correlation-based Feature Selection (CFS) [51] method for feature selection. CFS is an algorithm that aims to find a subset of features that are unrelated to each other and highly correlated with the class by evaluating each feature subset using an objective function. CFS calculates a heuristic measure of the feature subset $F$ as:

$$Merit_F = \frac{nt_{cf}}{\sqrt{n + n(n-1)t_{ff}}} \quad (1)$$

where $n$ is the number of features in sub feature space $F$, $t_{ff}$ is the average feature-feature inter-correlation, and $t_{cf}$ is the average class-feature correlation. CFS calculates $t_{ff}$ and $t_{cf}$ using a symmetric information gain.

## 4. Experimental setup

In this study we carried out 3-class music emotion classification since music excerpts are distributed into three quadrants of arousal-valence plane based on the distribution of the average of the annotator's ratings. Different information sources were utilized for feature extraction. In addition to standard audio features, CNN was employed to extract features using log-mel filterbank energies and MFCCs.

In this study, we used four convolutional layers each of which consist of 64, 128, 128, and 128 filters, respectively to produce feature maps. Basic modules such as; RELU and batch normalization were implemented following each convolutional layer. In the pooling layers, there are 4 and 3 max pooling layers to reduce dimensionality without padding. Besides, a dropout rate of 0.05 was applied to reduce overfitting. Adaptive moment estimation (ADAM) [52] optimizer with a 0.0001 learning rate was picked to find the best features. A batch size of 10 samples was chosen for training the CNN, and trained up to 100 epochs with early stopping. Also, categorical cross-entropy was chosen as a loss function to optimize neural networks.

To determine the most relevant features we applied CFS method. All the classification experiments were carried out utilizing 10-fold cross validation. In 10-fold cross-validation, the dataset

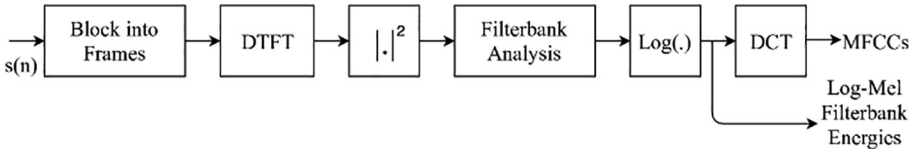


**Fig. 6.** Extraction of the MFCCs and Log-Mel filterbank energies.

is randomly divided into 10 equal parts. After that, data is stratified to get approximately the same class distribution for all the training sets. Then, 9 parts are selected as training data, the remaining part is for testing to calculate an error rate. This process is repeated 10 times until each subset is used as both training data and test data. The results are presented in terms of average accuracy (Eq. (2)), recall (Eq. (3)) where $R_i$ is a ratio of the number of music excerpts that are truly classified as class $C_i$ ($C_1$ = HAPV, $C_2$ = HANV, and $C_3$ = LANV) to the number of music excerpts belongs to class Ci in the data, precision $P_i$ (Eq.4), where $P_i$ refers to a ratio of the number of the truly classified music excerpts to the total number of music excerpts classified to class $C_i$ by the classifier and f-measure (Eq.5) for each class for multi-class classification where $l$ , $FN_i$ , $FP_i$ , $TN_i$ , and $TP_i$ refer to the number of class, false negatives, false positives, true negatives and true positives. The all evaluation metrics except accuracy are then calculated by averaging results from the dataset.

$$AverageAccuracy = \frac{\sum_{i=1}^{l} \frac{TP_i+TN_i}{TP_i+FN_i+FP_i+TN_i}}{l} \times 100 \quad (2)$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (3)$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (4)$$

$$F_i = \frac{2P_iR_i}{P_i + R_i} \quad (5)$$

## 5. Results and discussion

First, we evaluated the classification performance of LSTM + DNN classifier with various feature sets. Features are obtained by feeding different information sources into CNN layer. These information sources are log-mel filterbank energies and MFCCs that are labelled as fea_set1 and fea_set2, respectively. The classification performances of proposed feature sets are compared to that of standard audio features (fea_set3) extracted from music excerpts. Feature level combinations of feature sets and the effect of feature selection on classification performances are also evaluated. The performances of the proposed classifier are given in Table 4 in terms of classification accuracy. The best result is achieved with a combined feature set. Specifically, by combining new feature sets (fea_set1, and fea_set2) and standard features set (fea_set3), the performance increased from 87.09% to 91.93% over only standard feature set without applying CFS as shown in Table 4. This result suggests that new features provide additional discriminative information for music emotion recognition.

Table 3 displays the number of features before and after feature selection process for each feature set. Results indicated that the sizes of feature sets are reduced substantially by applying correlation-based feature selection. The numbers of selected features from each feature sets are given in Fig. 7. It can be observed from the figure that there are features selected from each feature sets. Also, as can be seen from Table 4, improved performances are achieved by employing feature selection. The performance increased from 92.74% to 99.19% after applying CFS for the combined features set over the standard feature set as seen in Table 4.

**Table 3**
The number of features and the number of features selected by CFS for each feature set.

| Feature Type | # of features | #of selected features |
|---|---|---|
| Fea_set1 | 768 | 61 |
| Fea_set2 | 768 | 50 |
| Fea_set3 | 7368 | 80 |
| Fea_set1 + Fea_set3 | 8136 | 103 |
| Fea_set2 + Fea_set3 | 8136 | 112 |
| Fea_set1 + Fea_set2 + Fea_set3 | 8904 | 110 |

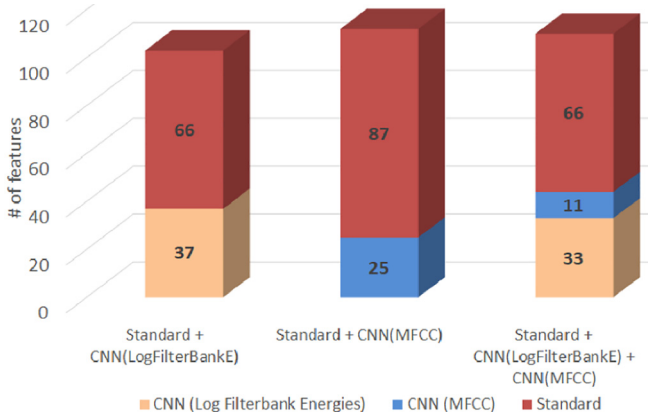


**Fig. 7.** The number of features selected by CFS from each feature types in the combined feature set.

**Table 4**
The classification performance of the proposed LSTM + DNN architecture using different feature sets.

| Features | Full feature Set | CFS applied |
|---|---|---|
| Fea_set1 | 87.09 | 93.54 |
| Fea_set2 | 90.32 | 96.77 |
| Fea_set3 | 87.09 | 92.74 |
| Fea_set1 + Fea_set3 | 88.70 | 98.38 |
| Fea_set2 + Fea_set3 | 88.70 | 96.77 |
| Fea_set1 + Fea_set2 + Fea_set3 | 91.93 | 99.19 |

These results demonstrate that proposed features based on CNN provide additional information to music emotion classification. Performance of the LSTM + DNN classifier is compared with the SVM, k-NN and Random Forest classifiers for the standard feature set in Table 5 and combined feature set in Table 6. Results in Table 5 display that, for the standard feature set, after applying CFS, LSTM + DNN yields improvements of 3.23, 2.42 and 0.81 points in music emotion recognition accuracies compared to that of k-NN, SVM and Random Forest classifiers, respectively. For the combined feature set 1.61, 1.61 and 3.23 points improvements are achieved. As can be seen from Table 5 and 6, the LSTM + DNN classifier gives better results than SVM, k-NN and Random Forest classifiers in terms of accuracy, recall, precision and f-measure. The detailed classification results for each emotion class (HAPV = high arousal positive valence, HANV = high arousal negative valence, LANV = low arousal negative valence) for each classifier are given in Table 7.

## 6. Conclusion

In this paper, we have realized deep learning based architecture on emotion recognition from Turkish music. New Turkish emotional music database composed of 124 Turkish traditional music excerpts with a duration of 30 s is constructed to evaluate the performance of the approach. The music excerpts were annotated on valence and arousal dimensions. We used 3-class music emotion since classification music excerpts are distributed into three quadrants based on the distribution of the average of the annotator's

**Table 5**
Comparison of performance of proposed methods with other classifiers before and after feature selection in terms of accuracy using standard features (fea_set3).

| | K-NN | | Random Forest | | SVM | | LSTM + DNN | |
|---|---|---|---|---|---|---|---|---|
| | Full Feature Set | CFS | Full Feature Set | CFS | Full Feature Set | CFS | Full Feature Set | CFS |
| Accuracy | 83.87 | 89.51 | 87.09 | 91.93 | 87.09 | 90.32 | 87.09 | 92.74 |
| F-Measure | 0.839 | 0.904 | 0.843 | 0.918 | 0.866 | 0.905 | 0.864 | 0.926 |
| Precision | 0.840 | 0.919 | 0.844 | 0.917 | 0.862 | 0.907 | 0.861 | 0.925 |
| Recall | 0.839 | 0.895 | 0.871 | 0.919 | 0.871 | 0.903 | 0.871 | 0.927 |

**Table 6**
Comparison of performance of proposed methods with other classifiers before and after feature selection in terms of accuracy using combined feature set (fea_set1 + fea_set2 + fea_set3).

| K-NN | | | Random Forest | | SVM | | LSTM + DNN | |
|---|---|---|---|---|---|---|---|---|
| | Full Feature Set | CFS | Full Feature Set | CFS | Full Feature Set | CFS | Full Feature Set | CFS |
| Accuracy | 88.70 | 97.58 | 87.90 | 95.96 | 89.51 | 97.58 | 91.93 | 99.19 |
| F-Measure | 0.878 | 0.976 | 0.852 | 0.957 | 0.904 | 0.974 | 0.915 | 0.992 |
| Precision | 0.874 | 0.978 | 0.854 | 0.961 | 0.919 | 0.976 | 0.915 | 0.992 |
| Recall | 0.887 | 0.976 | 0.879 | 0.960 | 0.895 | 0.977 | 0.919 | 0.992 |

**Table 7**
Performances of classifiers in terms of recall, precision, and f-measure for each class before and after feature selection using a combined feature set (fea_set1 + fea_set2 + fea_set3).

| | | Full Feature Set | | | CFS | | |
|---|---|---|---|---|---|---|---|
| | | Recall | Precision | F-Measure | Recall | Precision | F-Measure |
| | HAPV | 0.960 | 0.889 | 0.923 | 0.973 | 1.000 | 0.986 |
| K-NN | HANV | 0.273 | 0.429 | 0.333 | 0.909 | 1.000 | 0.952 |
| | LANV | 0.921 | 0.972 | 0.946 | 1.000 | 0.927 | 0.962 |
| | HAPV | 0.973 | 0.859 | 0.913 | 0.987 | 0.949 | 0.967 |
| RF | HANV | 0.091 | 0.500 | 0.154 | 0.636 | 1.000 | 0.778 |
| | LANV | 0.921 | 0.946 | 0.933 | 1.000 | 0.974 | 0.987 |
| | HAPV | 0.974 | 0.925 | 0.949 | 1.000 | 1.000 | 1.000 |
| SVM | HANV | 0.303 | 0.550 | 0.445 | 0.727 | 1.000 | 0.842 |
| | LANV | 0.987 | 0.881 | 0.931 | 1.000 | 0.962 | 0.980 |
| LSTM + DNN | HAPV | 0.973 | 0.924 | 0.948 | 1.000 | 0.987 | 0.993 |
| | HANV | 0.545 | 0.750 | 0.632 | 0.909 | 1.000 | 0.952 |
| | LANV | 0.921 | 0.946 | 0.933 | 1.000 | 1.000 | 1.000 |

assessments. 1D - CNN architecture is employed for feature extraction using log -mel filterbank

energies, and MFCCs. Results demonstrate that adding new features to standard audio features improves the classification performance. The effects of feature selection were also investigated. Improved performances were achieved by reducing the feature size applying correlation-based feature selection method. In the paper, we also conducted experiments to compare LSTM + DNN classifier architecture performance with other classifiers which resulted in better performances. In future, we will increase the database size to include samples from the low arousal positive valence (LAPV) quadrant. We will also evaluate the system with different databases and explore the cross-database performance of the approach.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] R. Panda, R.P. Paiva, Using support vector machines for automatic mood tracking in audio music, in: 130th Audio Eng. Soc. Conv. 2011, 2011.
[2] Y. Feng, Y. Zhuang, Y. Pan, Popular Music Retrieval by Detecting Mood, in: SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval), 2003: pp. 375–376. https://doi.org/10.1145/860500.860508.
[3] J.A. Russell, A circumplex model of affect, J. Pers. Soc. Psychol. (1980), https://doi.org/10.1037/h0077714.
[4] R.E. Thayer, Modern Perspectives on Mood, in, Biopsychology Mood Arousal (1989).
[5] S. Mo, J. Niu, A Novel Method based on OMPGW Method for Feature Extraction in Automatic Music Mood Classification, IEEE Trans. Affect. Comput. 3045 (2017), https://doi.org/10.1109/TAFFC.2017.2724515.
[6] L. Lu, D. Liu, H.J. Zhang, Automatic mood detection and tracking of music audio signals, IEEE Trans. Audio, Speech Lang. Process. 14 (2006) 5–18, https://doi.org/10.1109/TSA.2005.860344.
[7] Y.H. Yang, Y.C. Lin, Y.F. Su, H.H. Chen, A regression approach to music emotion recognition, IEEE Trans. Audio, Speech Lang. Process. 16 (2008) 448–457, https://doi.org/10.1109/TASL.2007.911513.
[8] R. Malheiro, R. Panda, P. Gomes, R.P. Paiva, Emotionally-Relevant Features for Classification and Regression of Music Lyrics, IEEE Trans. Affect. Comput. 9 (2018) 240–254, https://doi.org/10.1109/TAFFC.2016.2598569.
[9] R. Panda, R.P. Paiva, Music emotion classification: Dataset acquisition and comparative analysis, 15th Int, Conf. Digit. Audio Eff. DAFx 2012 Proc. (2012).
[10] R. Panda, B. Rocha, R.P. Paiva, Music emotion recognition with standard and melodic audio features, Appl. Artif. Intell. 29 (2015) 313–334, https://doi.org/10.1080/08839514.2015.1016389.
[11] Y.C. Lin, Y.H. Yang, H.H. Chen, Exploiting online music tags for music emotion classification, ACM Trans, Multimed. Comput. Commun. Appl. 7 S (2011), https://doi.org/10.1145/2037676.2037683.
[12] A. Aljanaki Emotion in Music: representation and computational modeling 2016 149
[13] H.H. Yang, Yi-Hsuan and Chen, CRC Press Inc, USA, Music Emotion Recognition, 2011.
[14] M. Soleymani, M.N. Caro, E.M. Schmidt, C.Y. Sha, Y.H. Yang, 1000 Songs for Emotional Analysis of Music, CrowdMM, Proc. 2nd ACM Int Work. Crowdsourcing Multimed. 2013 (2013) 1–6, https://doi.org/10.1145/2506364.2506365.
[15] Y.A. Chen Y.H. Yang J.C. Wang H. Chen The AMG1608 dataset for music emotion recognition, ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2015-Augus (2015) 693 697 10.1109/ICASSP.2015.7178058
[16] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter, Audio Set: An ontology and human-labeled dataset for audio events, ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. (2017) 776–780, https://doi.org/10.1109/ICASSP.2017.7952261.

[17] D. de Benito-Gorron, A. Lozano-Diez, D.T. Toledano, J. Gonzalez-Rodriguez, Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset, Eurasip, J. Audio, Speech, Music Process. 2019 (2019) 1–18, https://doi.org/10.1186/s13636-019-0152-1.
[18] A. Aljanaki, Y.H. Yang, M. Soleymani, Developing a benchmark for emotional analysis of music, PLoS One. 12 (2017) 1–22, https://doi.org/10.1371/journal.pone.0173392.
[19] G. Tzanetakis, G. Tzanetakis, Manipulation, analysis and retrieval systems for audio signals, Princet. Univ. Princeton, NJ, 2002, p. 198.
[20] A. de Cheveigné, H. Kawahara, YIN, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Am. (2002), https://doi.org/10.1121/1.1458024.
[21] C. Harte, M. Sandler, M. Gasser, Detecting harmonic change in musical audio, Proc. ACM Int. Multimed. Conf. Exhib. (2006) 21–26, https://doi.org/10.1145/1178723.1178727.
[22] O. Lartillot P. Toiviainen Mir in matlab (II): A toolbox for musical feature extraction from audio, Proc. 8th Int Conf. Music Inf. Retrieval, ISMIR 2007, 2007, 127 130
[23] C. McKay jAudio: Towards a standardized extensible audio music feature extraction system 2005 Course Pap. McGill Univ Canada http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.148.5866&rep=rep1&type=pdf
[24] E. Pampalk, A. Rauber, D. Merkl, Content-based organization and visualization of music archives, Proc. ACM Int. Multimed. Conf. Exhib. (2002) 570–579, https://doi.org/10.1145/641118.641121.
[25] F. Eyben, B. Schuller, openSMILE:), ACM SIGMultimedia Rec. 6 (2015) 4–13, https://doi.org/10.1145/2729095.2729097.
[26] B. Mathieu, S. Essid, T. Fillon, J. Prado, G. Richard, Yaafe, an easy to use and efficient audio feature extraction software, Proc. 11th Int, Soc. Music Inf. Retr. Conf. ISMIR (2010, 2010,) 441–446.
[27] G. Tzanetakis, P. Cook, MARSYAS: A framework for audio analysis, Organised Sound. 4 (2000) 169–175, https://doi.org/10.1017/S1355771800003071.
[28] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, S.K. Saha, Recognition of emotion in music based on deep convolutional neural network, Multimed. Tools Appl. (2019), https://doi.org/10.1007/s11042-019-08192-x.
[29] A.L. Maas, P. Qi, Z. Xie, A.Y. Hannun, C.T. Lengerich, D. Jurafsky, A.Y. Ng, Building DNN acoustic models for large vocabulary speech recognition, Comput. Speech Lang. 41 (2017) 195–213, https://doi.org/10.1016/j.csl.2016.06.007.
[30] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, N. Arunkumar, Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS), IEEE Access. 7 (2019) 57–67, https://doi.org/10.1109/ACCESS.2018.2883213.
[31] K. Simonyan A. Zisserman Very Deep Convolutional Networks for Large-Scale Image Recognition 2014 1 14 http://arxiv.org/abs/1409.1556
[32] S. Abdoli, P. Cardinal, A. Lameiras Koerich, End-to-end environmental sound classification using a 1D convolutional neural network, Expert Syst. Appl. (2019), https://doi.org/10.1016/j.eswa.2019.06.040.
[33] X. Liu, Q. Chen, X. Wu, Y. Liu, Y. Liu, CNN based music emotion classification, (2017). http://arxiv.org/abs/1704.05665.
[34] T. Liu, L. Han, L. Ma, D. Guo, Audio-based deep music emotion recognition, AIP Conf. Proc. 1967 (2018), https://doi.org/10.1063/1.5039095.
[35] C.-W. Huang S.S. Narayanan Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition 2017 1 20 http://arxiv.org/abs/1706.02901
[36] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-End Multimodal Emotion Recognition Using Deep Neural Networks, IEEE J. Sel. Top. Signal Process. (2017), https://doi.org/10.1109/JSTSP.2017.2764438.
[37] D. Bukhari, Y. Wang, H. Wang, Multilingual Convolutional, Long Short-Term Memory, Deep Neural Networks for Low Resource Speech Recognition, in, Procedia Comput. Sci. (2017), https://doi.org/10.1016/j.procs.2017.03.179.
[38] T.N. Sainath R.J. Weiss A. Senior K.W. Wilson O. Vinyals Learning the speech front-end with raw waveform CLDNNs Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. 2015-Janua (2015) 1 5
[39] Y. Song, S. Dixon, M. Pearce, Evaluation of musical features for emotion classification, Proc. 13th Int, Soc. Music Inf. Retr. Conf. ISMIR (2012, 2012,) 523–528.
[40] B. Rocha, R. Panda, R.P. Paiva, Music Emotion Recognition: The Importance of Melodic Features, 6th Int. Work. Music Mach. Learn. – MML 2013 – Conjunction with Eur, Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases – ECML/PKDD (2013. (2013).).
[41] F. Zhang, H. Meng, M. Li, Emotion extraction and recognition from music, 2016 12th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov. ICNC-FSKD 2016. (2016) 1728–1733. https://doi.org/10.1109/FSKD.2016.7603438.
[42] R. Hennequin, J. Royo-letelier, Music Mood Detection based on Audio and Lyrics, Proc. 19th Int, Soc. Music Inf. Retr. Conf. (2018) 370–375.
[43] H. Liu, Y. Fang, Q. Huang, Music Emotion Recognition Using a Variant of Recurrent Neural, Network 164 (2019) 15–18, https://doi.org/10.2991/mmssa-18.2019.4.
[44] B.J. Han S. Rho R.B. Dannenberg E. Hwang SMERS, Music emotion recognition using support vector regression, Proc. 10th Int Soc. Music Inf. Retr. Conf. ISMIR 2009, 2009, 651 656
[45] T.N. Sainath O. Vinyals A. Senior H. Sak Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks, ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2015-Augus (2015) 4580 4584 10.1109/ICASSP.2015.7178838
[46] C.L. dos Santos, C.N. Silla, The Latin Music Mood Database, Eurasip J. Audio, Speech, Music Process. (2015), https://doi.org/10.1186/s13636-015-0065-6.
[47] D. Makris, I. Karydis, S. Sioutas, The Greek Music Dataset, in (2015), https://doi.org/10.1145/2797143.2797175.
[48] Y.H. Yang, Y.F. Su, Y.C. Lin, H.H. Chen, Music emotion recognition: The role of individuality, in: Proc. ACM Int. Multimed. Conf. Exhib., 2007. https://doi.org/10.1145/1290128.1290132.
[49] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: J. Mach. Learn. Res., 2011.
[50] F. Chollet, Keras: The Python Deep Learning library, Keras, Io, 2015.
[51] M. Hall, L.A. Smith, Feature Selection for Machine Learning : Comparing a Correlation-based Filter Approach to the Wrapper CFS : Correlation-based Feature, Int. FLAIRS Conf. (1999).
[52] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015.