STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXVII**, Number 1, 2022
DOI: 10.24193/subbi.2022.1.04

MUSIC RECOMMENDATIONS BASED ON USER'S MOOD USING CONVOLUTIONAL NEURAL NETWORKS

ANDREI PETRESCU

Abstract. This paper proposes a method for music recommendations using emotions, using deep learning techniques. The method is composed of two modules. The emotion detection module, which utilizes a hybrid architecture involving a Convolutional Neural Network (CNN) and a Reccurent Neural Network using Long-Short Term Memory (LSTM) Cells. We compared individual architectures of CNNs and LSTMs against our hybrid approach, outperforming them during experiments. We evaluated the modules on our own data set, created using Spotify's API and containing 2028 songs from different genres and linguistic families, labeled with valence and arousal values. The model also outperforms other related approaches, however we did not evaluate them on the same data set. The predictions are used by the second module, for which we proposed a simple method of ordering the results based on the similarity to user's input.

1. Introduction

As the popularity of streaming services grows each year, a problem is raised when it comes on which songs (besides the one he saves) should be delivered to the final user. The focus on these services is to provide as much content as possible for a large audience. The recommendations given by them are based on user's history of liked songs, genres, new songs which may be of interest for him etc.

Received by the editors: 10 October 2021.

2010 *Mathematics Subject Classification.* 68T45.

1998 *CR Categories and Descriptors.* I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – *Object recognition*; I.2.6 [**Artificial Intelligence**]: Learning – *Connectionism and neural nets*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *Intensity, color, photometry, and thresholding* .

Key words and phrases. mood, emotion, valence, energy, convolutional neural network, recurrent neural networks, long-short term memory, hybrid, regression, classification.

The task of music emotion recognition has been an area of interest for many years. This subject was tackled from different perspectives. Music may be annotated with emotion names (labels), or by expressing emotions using continuous values. Most *machine learning* (ML) methods consider features such as: pitch, beat, tempo, rhythm, melody or harmony. These were successfully utilized as inputs for Support Vector Machines and Naive Bayes models [15]. However, traditional machine learning techniques under perform when compared to deep learning methods. When working with high-dimensional data, machine learning methods are usually insufficient to learn more complex functions. For the task at hand, CNNs will be employed to extract abstract features from two-dimensional data, in the form of audio spectrograms [3]. These features are further utilized in Deep Neural Network architectures to predict the results. In most studies, deep learning approaches outperform traditional machine learning models [1].

As mentioned, the task of emotion recognition may be viewed as a classification problem. However, by using Russel's circumplex model of affect [14], we can define emotions using two continuous measures: valence and arousal (energy). This model allows us to reformulate it as a regression problem. Valence is an indicator of "happiness", which measures the positivity of an emotion. Arousal or energy measures the intensity of that certain emotion.

In this paper, we propose a hybrid deep learning architecture for a recommendation system composed of two parts (modules): an emotion detection module and a playlist generator. Using deep learning methods, we emphasised the task of emotion detection, by creating and comparing the performance of three neural network architectures. We first compared the performance of Convolutional Neural Network and Recurrent Neural Networks (RNNs). This approach was successfully used before for classification tasks in music recommendations systems based on genre [13], but, judging by our experiments, these additionally perform well on regression tasks. The proposed CNN-LSTM (Long-Short Term Memory) hybrid network achieves comparable results to state-of-the-art approaches, having the potential to outperform them. In addition, we describe an easy method to generate a playlist based on user's input, by searching for the closest items to his emotion.

We organized the rest of the paper as follows. Section 2 describes the related work utilized in creating and experimenting upon deep learning methods for solving the task at hand. The methodology for representing the data (sound and emotions) and the experimented neural network architectures are detailed

in Section 3. We also discuss about the metrics used for the evaluation and playlist generation. Section 4 contains details about the created data set for our experiments and the results and comparison to related work. Finally, Section 5 contains the conclusions and possible future work.

2. Related work

During the 20th century, music studies emerged with the work of Kate Henver [5], who succeeds to unveil a correlation between emotions and music characteristics. Taking in consideration the work done for audio-based mood detection, in the last 20 years, different approaches have been developed. Such works include Tao Li and Mitsumori Ogihara's [7], in which they use audio features as timbre, rhythm and pitch to detect emotions in music, or Geoffroy Petter's [11] Support Vector Machine. He used Mel-Frequency Cepstral Coefficients as an input for his SVM.

As advancements in deep learning technology occurred, new models emerged based on fewer feature engineering. Music Information Retrieval Evaluation eXchange (MIREX) competition has unveiled the evolution of state of the art. Thomas Lidy and Alexander Schindler's work [8] shows the potential of audio-based models, using CNNs. However, these approaches are based on labeling music by their emotions. Two state-of-the-art methods using CNNs, that use the same data representation as our work, consist in the works of Delbouys et al. [3] and Bhattarai et al. [1]. These methods implement a fully-connected network as the final layer, which will output the prediction.

Recent music recommendation systems utilize deep learning in order to identify musical content [16]. Raju et al. utilize a hybrid network using a CNN and an LSTM module for music genre classification [13]. This method inspired our approach, which was further compared with other similar works that implement this type of network for emotion recognition. A very similar method was implemented by Malik et al. [10], who uses a bidirectional-GRU module, instead of consecutive LSTM layers.

3. Methodology

This section will approach in multiple subsections the methodology used in elaborating emotion detection method. First, it is discussed how the data is represented in particularly the audio signal and the emotion model utilized. In the following subsections, we will present the deep learning methods, which

form the hybrid architecture. These methods will also be individually implemented for comparison. The final model consists in a combination of a CNN module and a RNN using LSTM cells. The combination of these models was created setting the goal to achieve better performance than state-of-the art and to outperform the individual models.

3.1. **Data Representation.**

3.1.1. *Emotion Representation.* Russell's valence-arousal (V-A) model [14] is one of the most widely used models for describing emotions. Because people experience interactions differently, this approach seeks to express emotions objectively in a way that mere labels can not. Emotions are represented on a two-dimensional space in this model. The positive effect of an emotion is represented by valence. Valence levels can be interpreted as being negative or positive, or low or high, depending on the scale employed. Happiness, for example, might be classified as a positive or high-valence emotion. The intensity of an emotion is represented by arousal. Anger, for example, is a powerful emotion with a high arousal value. Therefore, the scale utilized determines the representation. Valence and arousal are given values between 0 and 1 in our experiments. The V-A model describes happiness as having high values, near to 1, for both components.

3.1.2. *Mel-Spectrograms.* The audio signal may be represented in the form of an image in the form of a spectogram. Using spectrograms, we can take advantage of CNN's performance on multidimensional data [9]. Applying the Short-Time Fourier Transform [9], we obtain the spectrogram.

Inspired by Delbouys et al [3], we took into consideration that the human ear cannot differentiate sounds of low or high frequencies. Beginning from a frequency f, we can re-scale the values to Mel Scale. The converted value m is obtained applying the following formula [6]:

$$m = 2595 * \log(1 + \frac{f}{500}) \tag{1}$$

After converting all the frequencies using Formula. 1, we will obtain the final mel-spectrogram. The output will be saved as a 128 x 128 px grayscale image as shwon in Figure. 1.

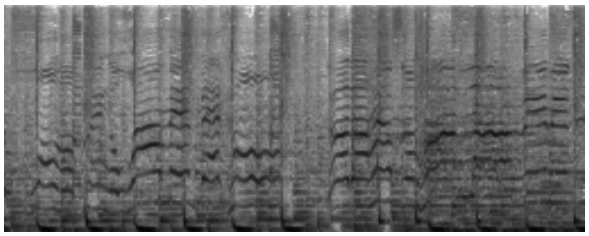

FIGURE 1. Example of a mel-spectogram

3.2. **Proposed Architecture.** In order to create a recommendation system based on emotions for music, we divided it into two parts. The first part will detect an emotion from a musical piece and the second will deliver the recommendations based on a user's input (his emotion using the valence-arousal model described in the Section. 3.1.1). This paper focuses on the emotion detection, for which we created a hybrid neural network, that may outperform state-of-the-art approaches. In the following sections we will present the architecture, the performance metrics used for the evaluation and a proposal for generating recommendations.

3.2.1. *Convolutional Neural Network.* These types of networks are known for the capability of extracting abstract features from multidimensional data. Beginning from a 128 x 128 matrix, representing the mel-spectrogram extract the features, further processed. Each convolution layer multiplies parts of the last layer's output. The result may, or may not outline the important information contained by the input matrix. Another type of layer, that CNNs use, is the pooling layer. This study utilizes the Max-Pooling layers to down-sample the the output of convolutions. Max-Pooling layers downsizes the input using the maximum values from a stride. A stride is a portion from a matrix having a fixed size [1].

The CNN module (Figure 2), utilized for the hybrid and individual architectures, will use multiple convolution and pooling layers. This architecture was inspired from the works of Liu et al. [9], that used a CNN architecture for music emotion recognition. Our CNN alternates between a convolution layer and a max-pooling layer as shown in Figure 2. Between them, we use a batch normalization layer to keep values under control and avoid over fitting. For individual analysis during the experiments, we connected the module to

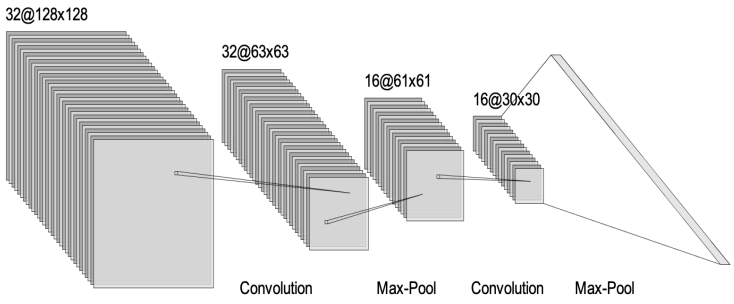

FIGURE 2. CNN module

a fully connected Deep Neural Network (DNN) (Figure 3). Its final layer has only two neurons computing the values of valence and energy.

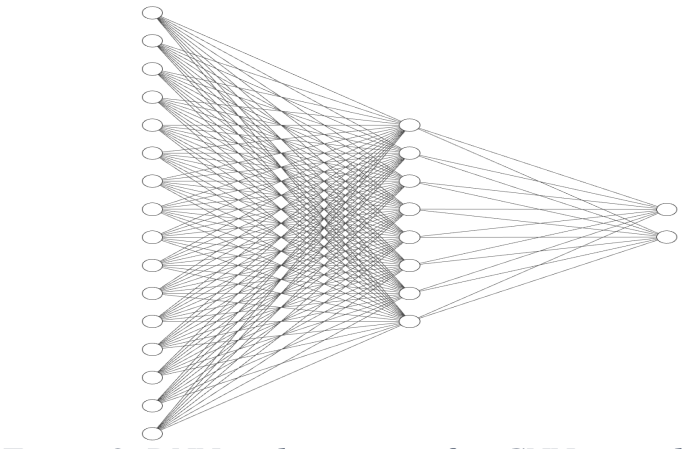

FIGURE 3. DNN module

3.2.2. *LSTM-RNNs.* The second method studied involves RNNs. This type of architecture has proven its performance on solving emotion recognition problems. However, traditional RNNs suffer from what is known as "the vanishing/exploding gradient" problem. This prevents RNNs from further learning. To avoid this problem, we use LSTM Cells, described by formulas (2) and (3) [4]:

$$c_t = f_t \odot \ c_{t-1} + i_t \odot \widetilde{c}_t \tag{2}$$

$$\widetilde{c}_t = g(W_c * x_t + U_c * \ h_{t-1} + b_c) \tag{3}$$

$$h_t = o_t \odot g(c_t) \tag{4}$$

In Formulas (2) and (3), t is the current timestamp, c_t is the current cell value, $\widetilde{c}_t$ is the proposed cell, h_t is the hidden state, g is the activation function. W, U, b are the prior computed parameters, which are updated during backpropagation. During the computations, LSTMs use three types of signal gates: f_t forget gate, o_t output gate, i_t input gate. The formulas for computing the gates are [4]:

$$i_t = \sigma(W_i * x_t + U_i * \ h_{t-1} + b_i) \tag{5}$$

$$f_t = \sigma\left(W_f * x_t + U_f * h_{t-1} + b_f\right) \tag{6}$$

$$o_t = \sigma(W_o * x_t + U_o * \ h_{t-1} + b_o) \tag{7}$$

In Formulas (5), (6) and (7), the parameters $W_{i,f,o}$, $U_{i,f,o}, b_{i,f,o}$ correspond to each gate for signal computations and are updated using backpropagation during the training stage. Our two-dimensional input, may be divided in fixed time-steps lengths. The module uses two LSTM layers with 40 and 2 units. For individual comparison with the hybrid module, the module is connected to a Dense layer, that will compute the output.

3.2.3. *The proposed Hybrid Network.* In the previous subsections we described the individual modules, which combined will form the hybrid architecture. The network begins with the CNN, reshaping the input by dividing it into time steps. We used the following formula for shaping the input:

$$inputShape_{Hybrid} = (no.\ of\ images,\ no.\ of\ t.\ steps,\ 128, \frac{128}{no.\ of\ t.\ steps},\ 1) \tag{8}$$

The first parameter from Formula (8) represents the number of images processed by the network. For each image the network outputs one prediction.

The second parameter is the number of time steps used. Controlling this parameter will affect the performance of the model. Last three parameters refer the image's sizes. The output of the CNN module is flattened and directed to the LSTM module, which will predict the emotions.

3.2.4. *Performance Evaluation.* We compared our model with state-of-the-art approaches that utilize CNNs or RNNs to detect emotions from music. These approaches use different performance metrics specific for a regression model. Therefore, we evaluated our models using all performance metrics met in the compared papers. The performance metrics are:

- *Mean Squared Error* (MSE)

$$MSE = \frac{1}{n} * \sum_{i=1}^{n} \left(Y_i - \widetilde{Y_i}\right)^2 \tag{9}$$

- *Root Mean Squared Error* (RMSE)

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} \left(Y_i - \widetilde{Y_i}\right)^2} \tag{10}$$

- *Mean Absolute Error* (MAE)

$$MAE = \frac{1}{n} * \sum_{i=1}^{n} |Y_i - \widetilde{Y_i}| \tag{11}$$

- R^2 score (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - \widetilde{Y_i}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} \tag{12}$$

In Formulas (9), (10), (11) and (12) the expected value for the i-th input is denoted by Y_i. $\widetilde{Y_i}$ represents the predicted value and $\overline{Y_i}$ is the mean of the expected values. To obtain a better performance, our models require, on one hand to minimize MSE, RMSE and MAE values, on the other hand to maximize the R^2 score.

3.2.5. *Recommendations.* After using the model for predicting the valence and arousal values for all input images, we can generate a list (playlist) of music based on an user's mood. The goal is to recommend music as close as possible to user's emotions. A user may characterize his current mood using the

valence-arousal model described in Section. 3.1.1. Each song is now characterized by these values too, therefore we need to choose the most similar songs for an user's mood. The values range from 0 to 1 and the data can be represented in a two-dimensional space. For computing the similarity we can use a distance based metric. An example of such a metric is the Euclidian distance computed using the following formula:

$$D_i = \sqrt{(v_i - v_u)^2 + (a_i - a_u)^2} \tag{13}$$

In Formula (13), v_i and a_i are the valence and arousal values corresponding to a song i. v_u, a_u are the valence and arousal values from an user's input. For each song we will compute this distance D_i and sort them from the lowest to highest score. From the obtained list, we can extract first n songs, which will form the playlist for a user.

4. Experimental Evaluation

This section presents the conducted experiments and the obtained results. For testing the proposed methods, we created and utilized our own data set containing the generated mel-spectrograms and corresponding values for valence and arousal. In the following subsections we discuss the results for the individual architectures (CNN and LSTM) and the novel hybrid approach for solving this task.

4.1. **Data Set.** Using Spotify's API [2], we collected the necessary audio files. This API provides a 30 seconds mpeg-3 audio clip as a preview for the entire song. This audio clip is the most meaningful and listened part of the song, according to API documentation [2]. For creating this data set we collected audio previews from the following genres: Jazz, Rock, Classical, Hip-Hop, Folk and Electronic. The data set may include songs from one or more sub genres from the ones mentioned above [12]. The music belongs to the following linguistic families: Latin, Slavic, Germanic, Indo-Iranian and Japanic. Another characteristic is that the data set contains instrumental and non instrumental songs, because we require that our methods may focus on melody and not on verses. The API also provides the valence and arousal (referred as energy) values for each song.

The final data set [12] to be used during the training and testing phases is composed of the generated mel-spectrograms (in the form of a gray scale

image) from the collected audio clips and contains 2028 entries. To each image, we associate the values for emotion. The mean values for valence and arousal are 0.476 and 0.526. The variance values for both features are 0.064 and 0.071.

4.2. **Experimental Setup.** As mentioned before, we trained and compared the performance of the individual modules and the hybrid network in solving the same task. We used k-fold cross-validation in order to detecting overfitting. We chose $k = 4$ for our implementation and each model is trained for 5000 epochs. Multiple activation function were used in our models, however the last layer of neurons is activated using the sigmoid activation. In case of the CNN module, we used ReLU activation functions. For the LSTM module we utilized hyperbolic tangent activation and sigmoid activation for recurrent activation. The rest of hyperparameter values are presented in Table. 1. For the hybrid network we took into consideration the possible effect of sequence lengths. We tested the performance on lengths of 16 and 32 time steps.

Hyperparameter	**Value**
Dropout β	0.2
Learning Rate	0.001
Data Set Split Ratio	80/20 (training/test)
Batch Size	50
No. of Epochs	3000

TABLE 1. Hyperparmeter values

4.3. **Results and Discussion.** The first experiment involves testing different time sequence lengths to achieve a better performance for the hybrid model. We will present the average results for the train and test data sets, but, in order to compare our approach with the related work, we considered computed the results for the entire data set (Table 4). The obtained results (Table 2) point out that a smaller time sequence length improves the performance, when applied on the testing data set. The performance increases for the training data as well, but, for the most cases, it is similar. The length's decrease has the greatest impact on the coefficient of determination (R^2 Score), which increases for the testing data. If we compare the results to the individual modules, the hybrid network obtains the best performance (Table 3). However, the CNN module achieves close results and obtains a better RMSE score for the testing data. The LSTM module performs the worst out of all three and

has a tendency of over fitting. Therefore, the most impact on the overall performance of the hybrid network it is due to the CNN module. Although, the extracted abstract features are better utilized by the LSTM module in order to predict the emotions, than the Deep Neural Network utilized for the individual CNN. A graphical illustration of the performances of the considered ML models on the testing data is illustrated in Figure 4.

Length	**MSE**		**RMSE**		**MAE**		**R^2 Score**	
	Train	Test	Train	Test	Train	Test	Train	Test
16	**0.001**	**0.031**	**0.027**	**0.178**	**0.022**	**0.139**	**0.996**	**0.413**
32	0.001	0.038	0.028	0.196	0.022	0.152	0.995	0.292

TABLE 2. Performance results of the hybrid network for different time sequence lengths

Architecture	**MSE**		**RMSE**		**MAE**		**R^2 Score**	
	Train	Test	Train	Test	Train	Test	Train	Test
CNN	0.011	0.035	0.105	0.189	0.081	0.153	0.931	0.323
LSTM	0.007	0.061	0.029	0.247	**0.020**	0.195	0.854	0.234
CNN+LSTM	**0.001**	**0.031**	**0.027**	**0.178**	0.022	**0.139**	**0.996**	**0.413**

TABLE 3. Performance results for all the studied methods

Architecture	**R^2 Score**
CNN	0.849
LSTM	0.798
CNN+LSTM	**0.901**

TABLE 4. Performance results for the entire data set

4.4. **Comparison to Related Work.** Throughout the literature, the task of predicting the mood from music was considered a classification problem. Data was labeled with different emotion names, which, in our opinion, simplifies the spectrum of existing emotions. By abstracting an emotion and quantifying it, we used the valence/energy model, which can accommodate a larger scale of emotions, without diminishing their complexity.

This approach was utilized before by other authors such as Delbouys et al. [3], for predicting emotions using CNNs. Tables 5, 6 and 7 compare our results

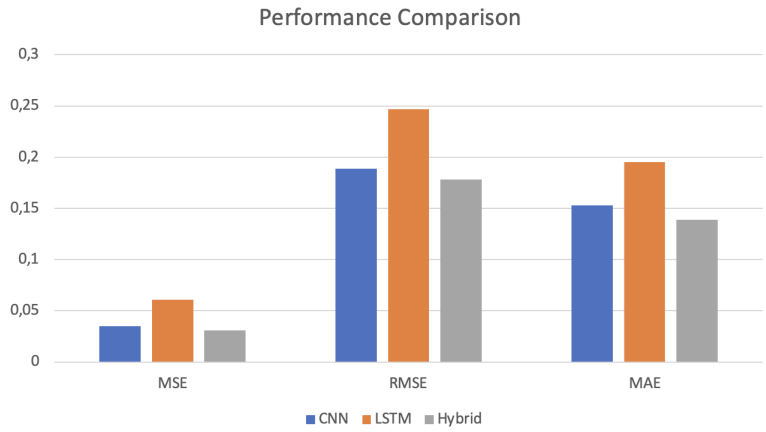

FIGURE 4. Comparison between the performance of studied methods (Table 3) on the testing data

to three similar approaches, those of Bhattarai et al. [1], aforementioned Delbouys et al. [3] and Malik et al. [10]. The last approach is the most similar to our hybrid network. Instead of two stacked LSTM layers, the authors use a bidirectional GRU network. We wanted to compare the impact on performance of a simpler recurrent network to a more complex one, in the form of Malik et al.'s approach [10]. The works compared in our study evaluate their models using parts (train/test) or entire data sets.

Approach	**Dataset size**	**R^2 Score**
Bhattarai et al. [1]	1000 x 96 x 1360	85.61%
CNN	2028 x 128 x 128	84.90%
LSTM		79.80%
CNN + LSTM		**90.10%**

TABLE 5. Comparison between approaches using R^2 for the entire dataset

As shown in Tables 5 and 6, our hybrid approach achieves an R^2 score of 90.10% applied on the entire data set an a score of 41.30% for testing data. These results outperform the the proposed methods of Bhattarai et al. [1] and Delbouys et al. [3], that utilize CNNs to detect emotions. In our individual experiments using only CNNs, we achieved close results to

Approach	**Dataset size**	**R^2 Score**
Delbouys et al. [3]	18644 x 40 x 1292	24.3%
CNN	2028 x 128 x 128	**31.70%**
LSTM		23.40%
CNN + LSTM		**41.30%**

TABLE 6. Comparison between approaches using R^2 for the test data

Approach	**Dataset size**	**RMSE**
Malik et al. [10]	431 x 60 x 260	0.255
CNN	2028 x 128 x 128	0.189
LSTM		0.247
CNN + LSTM		**0.178**

TABLE 7. Comparison between approaches using RMSE for the test data

those approaches, even outperforming Delbouy's method. When comparing the two similar hybrid networks, our method obtained a RMSE value of 0.178, outperforming the approach proposed by Malik et al. [10]. However, even if we obtained better results than the compared related works, we cannot state the superiority of our method, because we did not experiment upon the same data sets. The data sets used by the authors are privately owned and cannot be accessed without permission. Figure 5 visually represents the comparison between our proposed hybrid method (CNN+LSTM) and the approaches of Bhattarai et al. [1], Delbouys et al. [3] and Malik et al. [10], as represented in Table 5, Table 6 and Table 7.

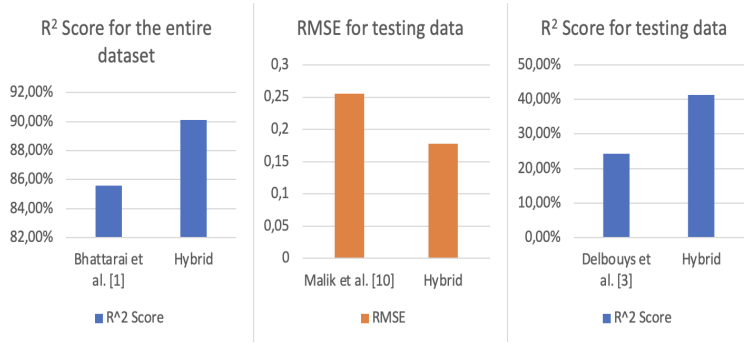

FIGURE 5. Performance comparison between the proposed hybrid method and the approaches of Bhattarai et al. [1], Delbouys et al. [3] and Malik et al. [10] (Table 5, Table 6, Table 7)

5. Conclusions and future work

In this study, we proposed a method for generating music recommendations based on emotions. Our approach consists of two modules: an emotion recognition module and a recommendation generator based on an user's input. First, we introduced the valence and arousal method created by Russell [14] for annotating data. Then, we studied different deep learning methods for emotion detection, beginning from the mel-spectrogram of a song. These methods are using CNNs, RNNs with LSTM cells and a hybrid network combining a CNN and an LSTM module. Even if the paper focuses on emotion recognition, we detailed a method to generate a personalized playlist, having the music labeled with valence and arousal values. This method searches for k-closest songs to a user's emotion.

We experimented and evaluated all the studied methods and concluded that combining CNN's ability to extract abstract features from multidimensional inputs and LSTM's performance on sentiment analysis tasks, we can obtain a model that outperforms both of them, when used individually. However, from our experiments, it was indicated that the CNN module has the most impact on the performance. Our work was compared to other similar approaches that utilize CNN for music emotion recognition. Based on the results, our method achieves better performance than the compared models. The experiments were not conducted on the same data sets, therefore we are not able to state the superiority of our approach.

Even if the hybrid network achieves better performance, we need to take int account the added computational expense. Using only a CNN module for this task, achieves very close performance and, in future work, we need to further experiment on different architectures, which may be able to achieve better results. However, considering the work of Malik et al. [10], we may further improve the performance using bidirectional RNN layers instead of sequential LSTM layers. We shall further analyze other hybrid architectures involving CNNs and forms of RNNs, that may be able to improve the prediction results.

References

[1] Bhattarai, B., and Lee, J. Automatic music mood detection using transfer learning and multilayer perceptron. *International Journal of Fuzzy Logic and Intelligent Systems 19*, 2 (2019), 88–96.

[2] Clifton, A., Pappu, A., Reddy, S., Yu, Y., Karlgren, J., Carterette, B., and Jones, R. The spotify podcast dataset. *arXiv preprint arXiv:2004.04270* (2020), 1–4.

[3] Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., and Moussallam, M. Music mood detection based on audio and lyrics with deep neural net. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018* (2018), pp. 370–375.

[4] Dey, R., and Salem, F. M. Gate-variants of gated recurrent unit (GRU) neural networks. In *IEEE 60th International Midwest Symposium on Circuits and Systems, MWSCAS 2017, Boston, MA, USA, August 6-9, 2017* (2017), IEEE, pp. 1597–1600.

[5] Hevner, K. Experimental studies of the elements of expression in music. *The American Journal of Psychology 48*, 2 (1936), 246–268.

[6] Kamm, T., Hermansky, H., and Andreou, A. G. Learning the mel-scale and optimal vtn mapping. In *Center for Language and Speech Processing, Workshop* (1997), pp. 1–8.

[7] Li, T., and Ogihara, M. Detecting emotion in music. *CiteSeer* (2003), 1–3.

[8] Lidy, T., and Schindler, A. Parallel convolutional neural networks for music genre and mood classification. *MIREX2016* (2016), 1–4.

[9] Liu, T., Han, L., Ma, L., and Guo, D. Audio-based deep music emotion recognition. *AIP Conference Proceedings 1967*, 1 (2018), 040021.

[10] Malik, M., Adavanne, S., Drossos, K., Virtanen, T., Ticha, D., and Jarina, R. Stacked convolutional and recurrent neural networks for music emotion recognition. *CoRR abs/1706.02292* (2017).

[11] Peeters, G. A generic training and classification system for mirex08 classification tasks: audio music mood, audio genre, audio artist and audio tag. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR'08)* (2008), Citeseer.

[12] Petrescu, A. Spotify dataset. https://github.com/AndreiPetrescu99/SpotifyDataset.git/, 2022.

[13] Raju, A., R.S, D., Gurang, D., Kirthika, R., and Rubeena, S. Ai based music recommendation system using deep learning algorithms. *IOP Conference Series: Earth and Environmental Science 785* (06 2021), 012013.

[14] Russell, J. A. A circumplex model of affect. *Journal of personality and social psychology 39*, 6 (1980), 1161.

[15] Tan, K., Villarino, M., and Maderazo, C. Automatic music mood recognition using russell's twodimensional valence-arousal space from audio and lyrical data as classified using svm and naïve bayes. *IOP Conference Series: Materials Science and Engineering 482* (03 2019), 012019.

[16] Yang, G. Research on music content recognition and recommendation technology based on deep learning. *Security and Communication Networks 2022* (03 2022), Article ID 7696840.

Department of Computer Science, Babeş-Bolyai University, 1, M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania

Email address: andrei.petrescu@stud.ubbcluj.ro