A multi-source deep learning model for music emotion recognition

Sofia Cazzaniga[1], Francesca Gasparini[1,2] and Aurora Saibene[1,2,*]

[1]University of Milano-Bicocca, Viale Sarca 336, 20126, Milano, Italy
[2]NeuroMI, Milan Center for Neuroscience, Piazza dell'Ateneo Nuovo 1, 20126, Milano, Italy

Abstract

Music has been recognized as an effective tool that could be beneficial in several applications aimed at increasing people's well-being. A personalized music recommender system can suggest playlists based on user's preferences and considering induced emotions. Being a subjective task, it is important to define a starting solid and generalizable Music Emotion Recognition (MER) model.
This model can be then refined to be adapted to the user's specific responses, ensuring a proper interaction between the recommendation system and its user. In this paper, a MER model relying on a multi-source input, composed of songs belonging to four publicly available datasets, is presented. The proposed model is based on EfficientNetB3, designed to provide high performance while being computationally efficient. Moreover, data splitting, layer modifications, and parameter setting are proposed to reduce the model overfitting.
Our proposal achieves performance comparable with those in the state of the art, providing a robust model to be adapted to a user's emotional responses in the definition of a music recommender system.

Keywords

Music Emotion Recognition (MER), perceived emotion, Mel-spetrograms, EfficientNetB3

1. Introduction

Music can positively affect health and well-being [1]. For example, music therapy is effective in improving the cognitive functions and the quality of life of people affected by dementia [2]. In fact, music is a powerful stimulus eliciting emotions and regulating mood, influencing human perception and behavior [3, 4, 5].

Emotions represent a key factor for the efficacy of applications meant to improve people's well-being and in the years researchers have strove to find a way of detecting them in music. This line of research translated into the field of *Music Emotion Recognition* (MER), mainly pertaining to the study and design of computational models to recognize emotions in songs [6].

Emotions in music can be distinguished in *perceived*, *induced*, or *intended* [7, 8]. The perceived emotion refers to the emotion that a listener identifies in a song, and is dependent from the song features, e.g., its structure, tempo, and lyrics. Instead, induced (also called felt) emotions are strictly related to the listener's own preferences and memories, and thus they are influenced by factors beyond the music itself. While emotions can be induced according to the listener's own context, a song can be composed by an artist to express a specific emotion, i.e., an *intended* emotion.
In this work we focus on listener-centric emotions, and thus only on the perceived and induced ones.

According to the reported definitions, it can be said that music playlists based on user's preferences, memories, and affective states can provide a better set of songs to be recommended to a specific listener. Ideally, a music recommendation system could automatically learn a user's emotional state and refine its recommendations over time and usage.

Italian Workshop on Artificial Intelligence for Human Machine Interaction (AIxHMI 2024), November 26, 2024, Bolzano, Italy.
*Corresponding author.
s.cazzaniga33@campus.unimib.it (S. Cazzaniga); francesca.gasparini@unimib.it (F. Gasparini); aurora.saibene@unimib.it (A. Saibene)
https://mmsp.unimib.it/ (F. Gasparini); https://mmsp.unimib.it/ (A. Saibene)
0000-0002-6279-6660 (F. Gasparini); 0000-0002-4405-8234 (A. Saibene)

However, a first step is required to provide such a custom playlist and consists in classifying a specific pool of songs by perceived emotions to have a starting point for the subsequent development of user-centered recommendations.

In this work a Deep Learning (DL) based MER model exploiting a time-frequency representation of songs is presented. These songs are selected from heterogeneous publicly available datasets, i.e., *4Q* [9, 10], *PMEmo* [11], *Emotion in Music* [12], and *Bi-Modal Emotion Dataset* [13]. Notice that we work with the emotion labels corresponding to the four quadrants of the Russel's Circumplex model of affect [14] and thus in the valence/arousal (V/A) plane.

The paper is organized as follows. Section 2 briefly presents some common MER strategies. Section 3 provides an overview of the used datasets. Section 4 describes the developed processing pipeline, highlighting the importance of a correct preparation of heterogeneous data (Section 4.1) to feed in the proposed DL-based MER model (Section 4.2). Section 5 is devoted to the presentation of the results and their discussion. Finally, conclusions are drawn, and possible developments of the work are provided (Section 6).

2. Related Work

In this section, representative literature works providing their MER strategies and using the datasets described in Section 3 are briefly reported. Up to our knowledge, no papers are using all the selected datasets at once and presenting a clear data preparation step intended to allow a correct use and comparison of data obtained from different sources as it is done in this study, that is intended to be an incremental contribution to the body of knowledge of the analyzed topic, providing a good starting point for a user-centered emotion-based playlist.

Starting from the less recent dataset, i.e., *Emotion in Music*, a MER strategy is devised to predict the four emotions in the V/A plane using different classifiers receiving in input (i) L^3-Net or (ii) VGGNet-based deep audio embeddings [15]. The approach consists of two steps. Firstly, deep audio embeddings are extracted from each song with one of the two approaches. Secondly, the corresponding emotion category is classified. The authors evaluate the performances using accuracy, highlighting that the use of L^3-Net provides better performances compared to the VGGNet-based model. The best performance on *Emotion in Music* is achieved with L^3-Net combined with a Multi-Layer Perceptron (MLP) classifier (72% accuracy). Notice that similar results (71% accuracy) are obtained by combining L^3-Net with a Support Vector Machine (SVM) or a random forest classifier.

Instead, *Malheiro et al.* [13], authors of the *Bi-Modal Emotion Dataset*, exploit audio features such as rhythm, melody, and timbre, as inputs to an SVM. Using a stratified 10-fold cross validation approach, they achieve 72.60% accuracy on the four classes corresponding to the four quadrants of the V/A plane.

In recent years, the advent of Convolutional Neural Networks (CNNs) encouraged researchers to exploit the capabilities of these DL models to capture audio features effectively by treating audio signals as images. For example, *Sarkar et al.* [16] use a VGGNet-based architecture, with log-magnitude Mel-scale spectrograms of 5 s segments as input. This method achieved a performance of 77.82%, marking a 6.10% improvement over the earlier results reported for the *Bi-Modal Emotion Dataset* by *Malheiro et al.* [13].

Considering an approach similar to the one proposed by us, *Sung et al.* [17] combine the *4Q*, *Bi-Modal Emotion Dataset*, and *PMEmo* datasets in a unique dataset. They employ two CNNs with six (CNN-6) and ten (CNN-10) layers, respectively, taking log Mel-Spectrograms of 60 s audio segments as input. The songs are processed to ensure they fit within the 60 s constraint by truncating signals lasting more than 60 s, and zero-padding the shorter ones. The models were evaluated using a stratified k-fold cross validation approach with k = 5. The CNN-6 model achieved on the four V/A plane quadrants classification a best micro F1-Score of 60.42%, while the CNN-10 model reached 62.92% on the first fold.

3. Datasets

Considering that the proposed learning model is based on a DL strategy requiring a large number of data, four datasets, i.e., *4Q* [9, 10], *PMEmo* [11], *Emotion in Music* [12], and *Bi-Modal Emotion Dataset* [13], have been chosen from the literature.

The selection criteria consisted of the online availability of (i) the original audio files (ii) with their metadata, such as song title and artist, and (iii) the presence of emotional labels according to human annotators for each audio.

Table 1 summarizes the names of the datasets, the dataset year of publication, and the link to the available online resources.

Table 1
Summary of the used datasets

Original Name	Year	Link
4Q	2018	https://mir.dei.uc.pt/downloads.html
PMEmo	2018	https://github.com/HuiZhangDB/PMEmo
Emotion in Music	2013	https://cvml.unige.ch/databases/emoMusic/
Bi-Modal Emotion Dataset	2016	https://mir.dei.uc.pt/downloads.html

Table 2
Conversions of emotional labels

Dataset	Original Labels	A-V-	A-V+	A+V-	A+V+
PMEmo	[0,1]	$A \leq 0.5, V \leq 0.5$	$A \leq 0.5, V > 0.5$	$A > 0.5, V \leq 0.5$	$A > 0.5, V > 0.5$
Emotion in Music	[1,9]	$A \leq 5, V \leq 5$	$A \leq 5, V > 5$	$A > 5, V \leq 5$	$A > 5, V > 5$
Bi-Modal Emotion Dataset	[-4,4]	$A \leq 0, V \leq 0$	$A \leq 0, V > 0$	$A > 0, V \leq 0$	$A > 0, V > 0$

Remind that in this paper the four quadrants of Russel's Circumplex model of affect [14] are considered. Thus, besides briefly describing the datasets, the conversions of the provided labels according to the affect model of interest are reported in Table 2. The letters A and V stand for arousal and valence, respectively. The minus (-) and plus (+) symbols are used to mark the values as low or high.

4Q is composed by songs collected from *AllMusic API*. The authors removed duplicate songs as well as files with missing metadata information. The resulting dataset presents 900 songs (lasting around 30 s), balanced in the four V/A plane quadrants.

The *PMEmo* has been devised to support MER-based studies requiring large music content libraries. The authors define an initial pool of songs by accessing the 2016-2017 songs of *Bilboard Hot 1000*, *iTunes Top 100*, and *UK Top 40 Singles*, resulting in 487, 616, and 226 songs, respectively. Duplicates were removed, obtaining the final 794 songs (lasting 10-90 s), annotated by at least 10 people with values between 0 and 1 for both valence and arousal.

The *Emotion in Music* dataset contains 744 audio signals. These signals last 45 s, having that those 45 random seconds were extracted from the original songs. These songs were selected from an initial set of 1000 songs taken from the *Free Music Archive* (https://freemusicarchive.org/). For each audio clip, metadata and both continuous and static annotations are available. In this study, only the latter annotations on the whole song are considered with values between 1 and 9 for both valence and arousal.

The *Bi-Modal Emotion Dataset* collects 200 songs (lasting 30 s). The annotation of the dataset was performed by 39 people assigning values between -4 and 4 to valence and arousal. To improve the consistency of the ground truth, the songs with a standard deviation above 1.2 were excluded. As a result, the final audio dataset contains 162 audio clips.

4. Proposed Processing Pipeline

In this section the proposed processing pipeline depicted in Figure 1 is described. Data preparation is required to provide a correct comparison of audio signals coming from different datasets. Moreover, data are converted into 300x300x3 pixels time-frequency images (i.e., Mel-Spectrograms) to provide a correct input for our *EfficientNetB3* [18] based DL MER model. The model is then introduced, reporting details on the architecture and the training process.

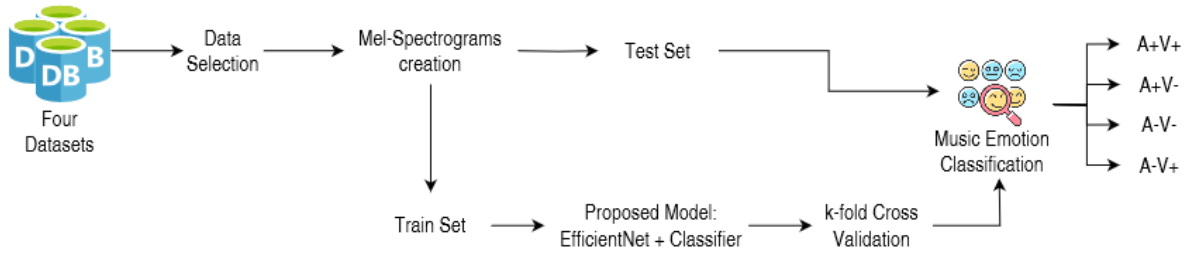

Figure 1: Proposed processing pipeline.

4.1. Data Preparation

The 2600 songs resulting from the previously described dataset selection, present different characteristics in terms of duration and signal acquisition.

Audio clips lasting 24 s are extracted from each song, wanting to maintain as much data as possible, while ensuring a sufficient time span to elicit emotions. In fact, a listener's emotion seems to stabilize in around 15 s from the song start [11]. Given this observation, the central part of the signal is extracted for those songs lasting more than 24 s to include part of the stabilization phase. Songs lasting less than 24 s are removed.

Secondly, a song selection is made to ensure a balanced distribution of the audios of each dataset in the four quadrants of the V/A plane. In the case of unbalanced distributions, the data placed in the extreme corners are selected from each of the quadrants of the V/A plane, i.e., as far as possible from the origin and axes. A graphical representation of this selection process is shown in Figure 2 for *PMEmo*, where the non-selected data include also those that were eliminated based on the criteria of the audio length.

Therefore, the final merged dataset of 1637 audio signals is composed of all the *4Q* and *Bi-Modal Emotion Dataset*, and 232 *PMEmo* and 346 *Emotion in Music* data.

All the audio signals are then downsampled to 22050 Hz, which is the lowest sampling rate among the four datasets. An anti-aliasing filter is introduced to avoid distortions. Volume normalization is not performed, considering that the volume influences the subjective perception of the song. Mel-Spectrograms are generated from the entire 24 s segments and converted to decibel units as a form of normalization. In fact, this conversion is performed by considering the maximum value among the spectrograms of each dataset and using this value as the reference maximum. This method is chosen to preserve the unique characteristics of individual songs across different datasets. Finally, the Mel-Spectrograms are saved with an image size of 300x300x3 pixels to be correctly used as inputs to our EfficientNetB3-based DL MER model. The use of time-frequency images is intended to understand if the proposed model can learn morphological characteristics of the songs bounded to the annotated emotion. Figure 3 provides examples of the generated Mel-Spectrograms for each V/A quadrant using the songs of *4Q*, i.e., *Little Saint Nick* by The Beach Boys (A+V+), *Only Two Can Play* by High Contrast (A+V-), *The Christmas Song* by Nat King Cole (A-V-), and *The Garden* by Vern Gosdin (A-V+).

The obtained images are divided into train (90%) and test (10%) sets. Notice that the division is performed by balancing the data in terms of class and dataset. The resulting distribution in terms of classes and datasets is depicted in Figure 4.

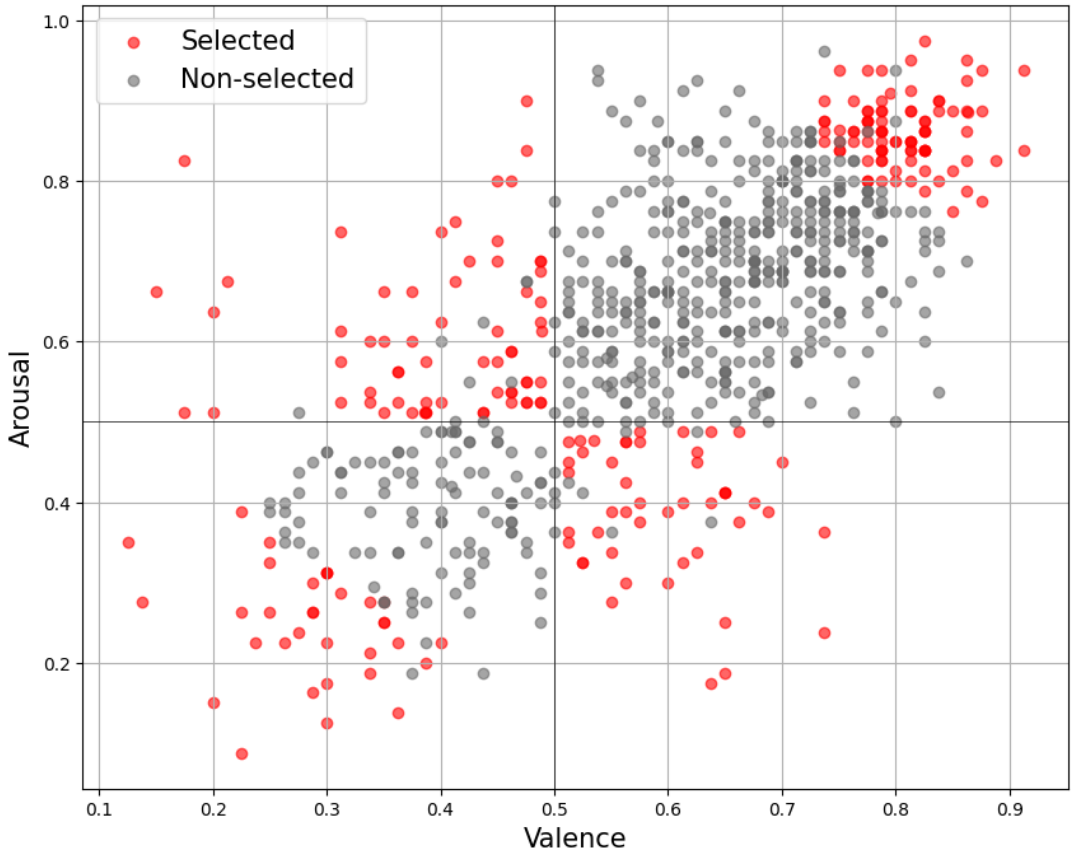

Figure 2: Distribution of songs in the V/A plane for *PMEmo*.

Moreover, one of the possible variations in the signal morphology can be due to the difference in genre. In fact, songs in the same genre tend to be composed with some recurrent structures. Therefore, a balancing in terms of genre in the train and test set division is also introduced for *Emotion in Music*, which is the only dataset presenting a one-on-one association of songs and genres.

4.2. Proposed MER Model

The proposed MER model is based on the CNN *EfficientNetB3* [19]. This choice is due to different factors.

The network works with images, which we wanted to use to understand if an initial DL model could correctly predict emotions by exploiting time-frequency information characterizing the audio signals. Moreover, *EfficientNetB3*, as well as other EfficientNet variations, is designed to provide high performance while being computationally efficient, i.e., requiring fewer parameters and computational resources compared to other architectures such as ResNet. This is especially true considering that this architecture exploits a compound scaling. In fact, the network depth (i.e., number of layers), width (i.e., number of channels in each layer), and input image resolution are scaled uniformly.

Another choice-driving characteristic is represented by the fact that the network is scaled from the baseline *EfficientNetB0*, which is optimized using neural architecture search.

Moreover, the network processes 300x300x3 pixels input images through 24 layers, with a structure comprising an initial stem, seven Mobile Inverted Bottleneck (MBConv) blocks, and a final fully connected layer. The MBConv blocks are crucial components of the network, characterized by (i) a depthwise convolution, reducing computational costs by processing channels independently, (ii) a pointwise convolution, increasing model capacity while maintaining computational efficiency, (iii) a squeeze-and-excitation module capturing channel-wise dependencies by computing statistics and

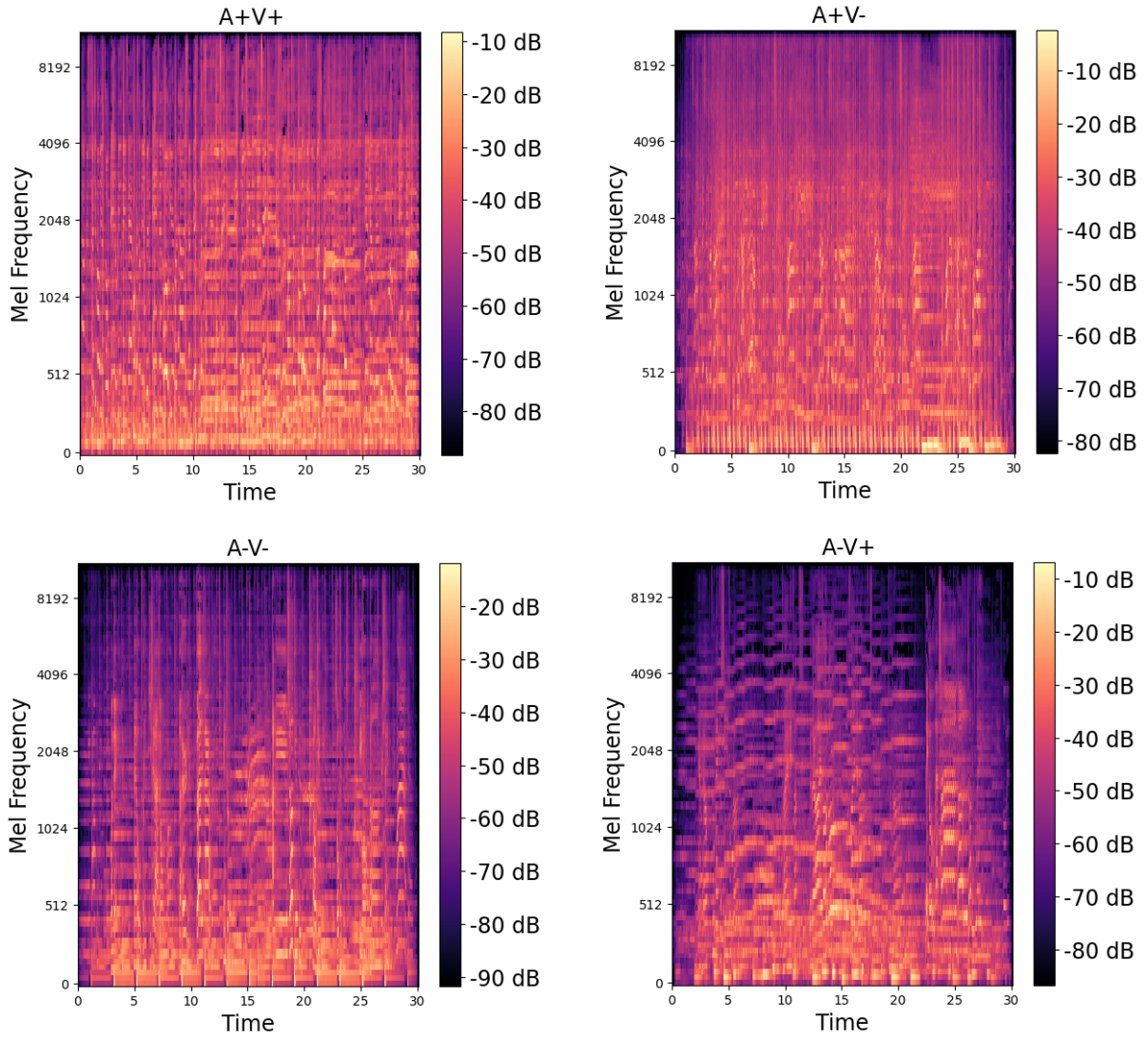

Figure 3: Example of four Mel-Spectrograms in dB units on songs covering the four quadrants from *4Q*.

learning feature re-weighting, enhancing model adaptability and feature representation.

Considering the limited number of available data of this study, *EfficientNetB3* is pre-trained using *ImageNet* [20].

The following layers are added to the last layer of the original network, proposing a modification to reduce data dimensionality with dense layers and prevent overfitting through dropout layers:

- Flatten Layer: flattens the output from the previous layer (in this case, the last layer of *EfficientNetB3*) into a 1D array, i.e., a flat vector.
- Dense layer: consists of 512 neurons, applies a linear transformation to the flat vector and the ReLU activation function.
- Dropout layer (0.5): randomly sets 50% of the input units to 0 at each network update during training. This helps prevent overfitting by reducing the co-dependency between neurons.
- Dense layer of (128 neurons) with ReLU, dropout layer (0.3), dense layer (64 neurons) with ReLU, and dropout layer (0.1).
- Output layer as a dense layer consisting of four neurons corresponding to the number of classes in the classification task (i.e., the V/A plane quadrants). It applies the softmax activation function, which converts the raw output into probability scores for each class, ensuring that the sum of the probabilities for all classes is equal to 1.

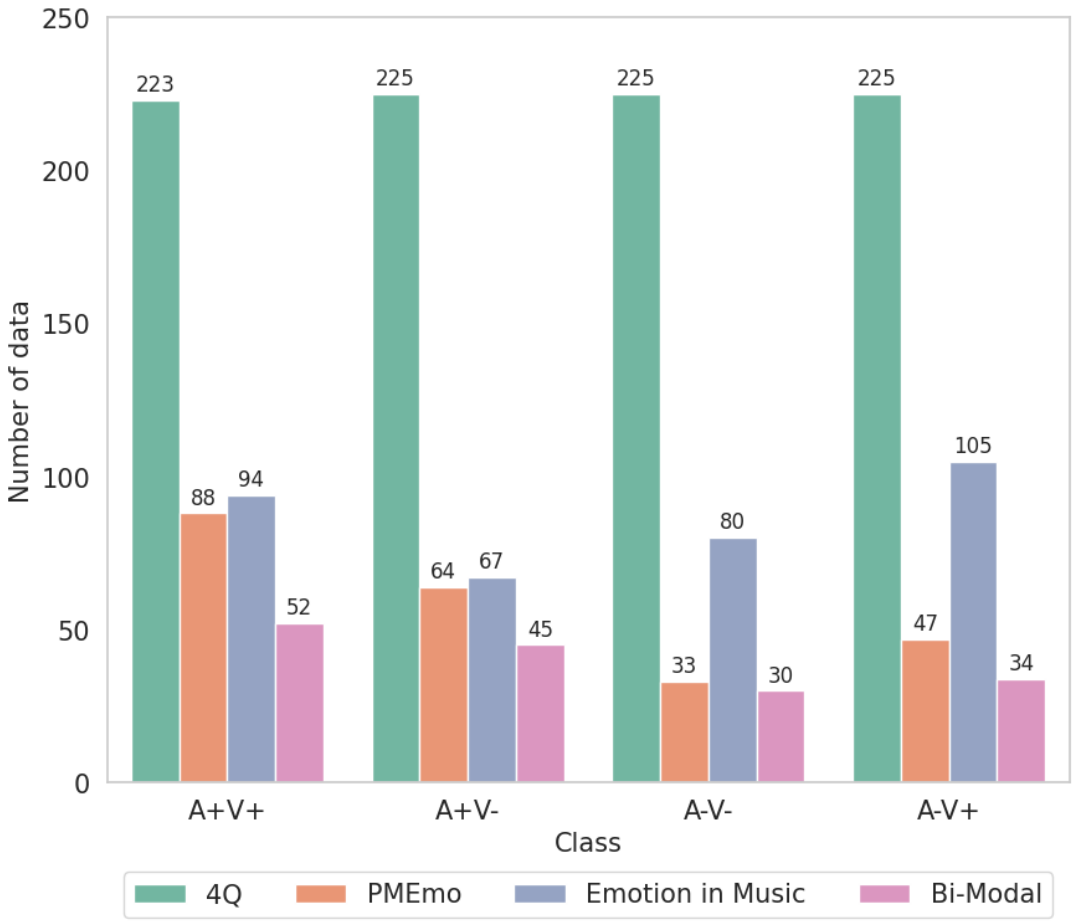

Figure 4: Data distribution per class and dataset.

The code is implemented in *Python* and executed on the *Kaggle* platform, with the following hardware specifications: *Intel(R) Xeon(R) CPU @ 2.00 GHz, 29 GB RAM, and NVIDIA TESLA T4(x2) 15GB GPU.*

The training is performed using the following parameters:

- Loss function: *sparse categorical cross-entropy loss*, used in multi-class classification tasks.
- Optimizer: *Adam* [21] with learning rate equal to 10^{-5}.
- Epochs: 50.
- Callbacks: the *EarlyStopping* callback is added to stop the training if the validation loss fails to decrease, restoring the model to its best weights. This tool is added to prevent overfitting, ensuring that the model generalizes well on unseen data.

A stratified k-fold cross-validation approach (with k = 5) is used to mitigate bias or dependencies introduced by arbitrary partitioning of data into train and validation sets, ensuring that each fold is representative of the overall class distribution.

5. Results and Discussion

The test set is used to evaluate each model produced by the stratified k-fold cross-validation procedure. The results from each of the five training sessions are analyzed and combined to provide a robust estimate of the model expected performance. This aggregated performance serves as a final evaluation metric, reflecting the model overall effectiveness and reliability.

Table 3 summarizes the obtained results in terms of accuracy and F1-score for each fold, while Table 4 reports the average precision, recall, and F1-score for each of the four classes. Remind that A and

V appearing in Table 4 correspond to the arousal and valence dimensions, respectively. The + and - symbols represent the high or low valence and arousal.

Table 3
Training results with stratified 5-fold cross validation

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	All Folds
Accuracy	0.64	0.65	0.62	0.64	0.65	0.64
F1-Score	0.64	0.65	0.62	0.63	0.65	0.64

Table 4
Average performance results

Class	Precision	Recall	F1-score
A+V+	0.68	0.68	0.68
A+V-	0.71	0.78	0.74
A-V-	0.62	0.58	0.60
A-V+	0.54	0.53	0.53

Before presenting the results, notice that the model tends to overfit on the validation set. Figure 5 depicts an example of this trend. The overfitting is likely due to the complexity of the network and the limited training data. However, early stopping is employed to prevent excessive overfitting, and the weights corresponding to the best performance before the onset of overfitting are saved. These weights represent the point at which the model had the best balance between learning from the training data and maintaining its ability to generalize on unseen data.

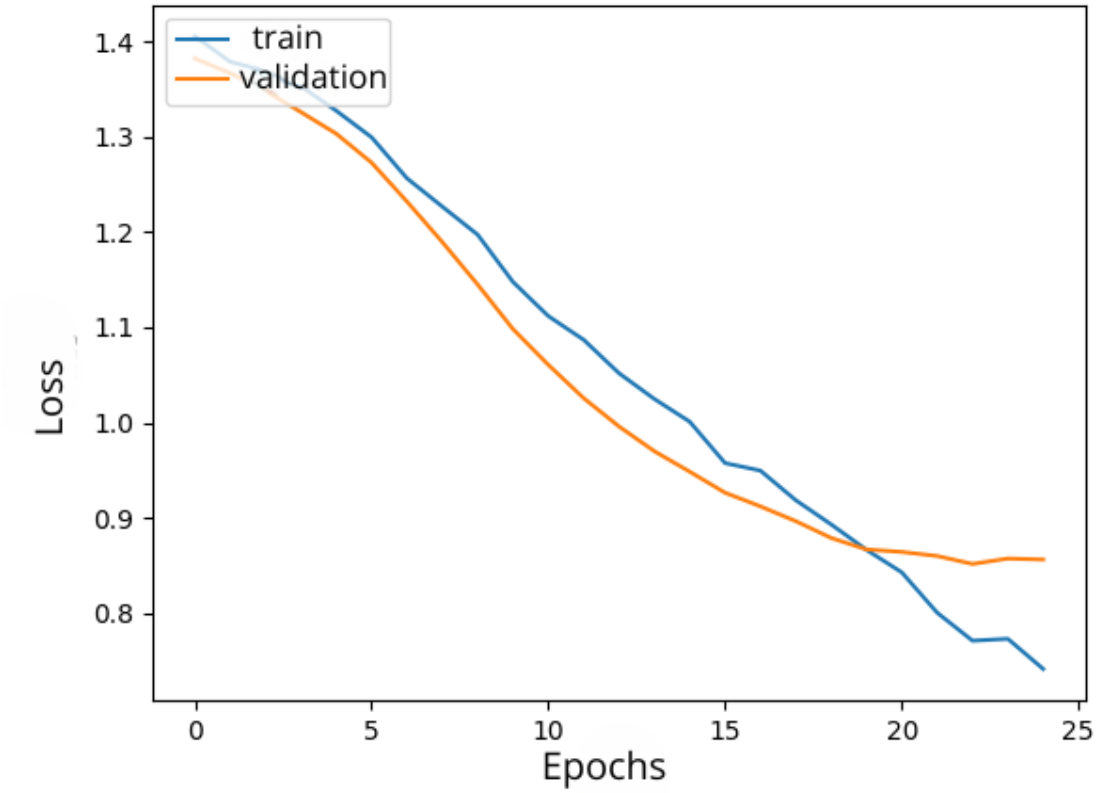

Figure 5: Plot of the model train and validation losses for fold 3.

Performance appears fairly similar and consistent among all the folds, suggesting the absence of bias due to training data selection. The model achieves 64% average accuracy and 64% F1-score. The model

Table 5
Summary table of different MER tasks on different datasets

Strategy	Dataset	Performance
L^3-Net + MLP [15]	Emotion in Music	72% accuracy
SVM [13]	Bi-Modal Emotion Dataset	72.60% accuracy
Mel-Spectrograms + VGGNet [16]	Bi-Modal Emotion Dataset	77.82% accuracy
Mel-Spectrograms + CNN-10 [17]	4Q, Bi-Modal Emotion Dataset, PMEmo	62.92% micro F1-Score
Our proposal	4Q, PMEmo, Emotion in Music, Bi-Modal Emotion Dataset	64% accuracy and F1-Score

is valued as sufficiently solid as a starting point to provide an emotion assessment of songs, considering that it significantly outperforms random guessing (which would correspond to 25% accuracy for a four-class learning task). Concerning the performance results on the four classes (Table 3), it can be noticed that the images related to the high arousal quadrants are classified better (A+V+ 68% and A+V- 74% F1-score), while the low arousal ones have a significantly lower performance (60% and 53% F1-score for the A-V- and A-V+ classes).

This can be due to the usual annotators' perceived difficulty in selecting a specific arousal value during song labeling.

Notice that no direct comparisons with the literature works are provided, considering the different classification tasks. However, a summary table (Table 5) is reported to provide a brief overview of the results obtained in different MER tasks exploiting different datasets. All the works provide a four-class emotion classification based on the V/A plane.

6. Conclusion and Future Work

In this paper, we described a robust multi-source DL MER strategy intended to provide an initial pool of songs falling into specific *perceived emotion* categories. The division of songs in the four V/A plane quadrants is intended to be used in further studies to provide a *user-centred induced-emotion-based music recommender system*.

Considering the difficulty of the classification task and the subjectivity of emotion evaluation, we value the obtained initial results (i.e., 64% average accuracy and F1-Score above chance level for a four-class task) satisfactory to start further analyses and developments of the proposed DL-based model. The results are also in line with the only literature study working on multiple datasets at a time [17], described in Section 2.

Error analysis will be performed to better understand why the classifier has lower performances for the low arousal labeled songs. Particular attention will be given to the misclassified songs by observing (i) the initial dataset from which a song is extracted, (ii) the song genre, (iii) its volume, and (iv) frequency features.

An in-depth analysis will be also performed to better assess the reason why the model performs better for certain quadrants of the V/A plane by using a cognitive appraisal and attention-based perspective. The resulting observations will be exploited to consider a modification of the model and/or its evaluation strategy.

Additional aspects will be considered in future works, particularly using a larger data pool (e.g., integrating more datasets such as the Moodo [22] and the AMG1608 [23] dataset), providing a more in-depth assessment of overfitting and hyper-parameter tuning, introducing further pre-processing steps, and different feature extraction strategies. On the latter note, besides audio signal-related features, cognitive features such as expectation, familiarity, and music complexity will be introduced to enrich the understanding of the emotional content of the songs.

A deeper analysis on the influence of lyrics in the model understanding of emotions will be also performed.
Further machine and deep learning models will be considered, especially to understand the efficacy of handcrafted features in case of a limited amount of data, and to provide direct comparisons between our proposal and literature DL solutions using the same dataset.

Starting from the final perceived-emotion model trained on the literature datasets, a user-tuning will be performed to provide an induced-emotion music recommendation. An experimental run involving controlled participants will be considered to collect further data that can influence the effective outcome of the user-based music recommender system besides the evaluation of valence and arousal, i.e., song liking and whether it is known or not. Particular attention will be given to the participants' agreement on the emotional dimensions, which seem to be never provided as information in the available labeled datasets.

References

[1] G. A. Dingle, L. S. Sharman, Z. Bauer, E. Beckman, M. Broughton, E. Bunzli, R. Davidson, G. Draper, S. Fairley, C. Farrell, et al., How do music activities affect health and well-being? A scoping review of studies examining psychosocial mechanisms, Frontiers in psychology 12 (2021) 713818.

[2] C. Moreno-Morales, R. Calero, P. Moreno-Morales, C. Pintado, Music therapy in the treatment of dementia: A systematic review and meta-analysis, Frontiers in medicine 7 (2020) 160.

[3] B. P. Gold, M. T. Pearce, E. Mas-Herrero, A. Dagher, R. J. Zatorre, Predictability and uncertainty in the pleasure of music: a reward for learning?, Journal of Neuroscience 39 (2019) 9397–9409.

[4] M. Baltazar, S. Saarikallio, Strategies and mechanisms in musical affect self-regulation: A new model, Musicae Scientiae 23 (2019) 177–195.

[5] M. B. Er, H. Çiğ, I. B. Aydilek, A new approach to recognition of human emotions using brain signals and music stimuli, Applied Acoustics 175 (2021) 107840.

[6] X. Yang, Y. Dong, J. Li, Review of data features-based music emotion recognition methods, Multimedia systems 24 (2018) 365–389.

[7] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, E. Gómez, Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications, IEEE Signal Processing Magazine 38 (2021) 106–114.

[8] L. Turchet, J. Pauwels, Music emotion recognition: intention of composers-performers versus perception of musicians, non-musicians, and listening machines, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021) 305–316.

[9] R. Panda, R. Malheiro, R. P. Paiva, Novel audio features for music emotion recognition, IEEE Transactions on Affective Computing 11 (2018) 614–626.

[10] R. Panda, R. Malheiro, R. P. Paiva, Musical texture and expressivity features for music emotion recognition, in: 19th International Society for Music Information Retrieval Conference (ISMIR 2018), 2018, pp. 383–391.

[11] K. Zhang, H. Zhang, S. Li, C. Yang, L. Sun, The PMEmo dataset for music emotion recognition, in: Proceedings of the 2018 acm on international conference on multimedia retrieval, 2018, pp. 135–142.

[12] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, Y.-H. Yang, 1000 songs for emotional analysis of music, in: Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia, 2013, pp. 1–6.

[13] R. Malheiro, R. Panda, P. Gomes, R. P. Paiva, Bi-Modal Music Emotion Recognition: Novel Lyrical Features and Dataset, 2016.

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

[15] E. Koh, S. Dubnov, comparison and analysis of deep audio embeddings for music emotion recognition, in: CEUR Workshop Proceedings, volume 2897, 2021, p. 15 – 22.

URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85109687006&partnerID=40&md5=7c7420825de1dac08abe6f9ffca7cd0c, cited by: 3.

[16] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, S. K. Saha, Recognition of emotion in music based on deep convolutional neural network, Multimedia Tools and Applications 79 (2020) 765–783.

[17] B.-H. Sung, S.-C. Wei, BECMER: A Fusion Model Using BERT and CNN for Music Emotion Recognition, in: 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI), 2021, pp. 437–444. doi:10.1109/IRI51335.2021.00068.

[18] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[19] M. Tan, Q. V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2020. URL: https://arxiv.org/abs/1905.11946. arXiv:1905.11946.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, 2015. URL: https://arxiv.org/abs/1409.0575. arXiv:1409.0575.

[21] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[22] M. Pesek, G. Strle, A. Kavčič, M. Marolt, The Moodo dataset: Integrating user context with emotional and color perception of music for affective music information retrieval, Journal of New Music Research 46 (2017) 246–260.

[23] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, H. Chen, The AMG1608 dataset for music emotion recognition, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 693–697.