Percussion and Instrumentation in Music Emotion Recognition: a Feature Engineering Approach

Hugo Redinho, Pedro Lima Louro, André C. Santos, Ricardo Malheiro, Rui Pedro Paiva, Renato Panda

***Abstract*—We propose a new set of features for audio-based Music Emotion Recognition (MER) that are related to percussion and individual instrument information. One limitation of current feature engineering approaches in MER is that they primarily focus on melodic elements. However, the percussive elements and instrumentation are also essential for conveying and recognizing emotions in music. Our approach leverages the Demucs framework for music source separation (which enables drum channel separation) and the MT3 framework for automatic music transcription and instrument recognition. Building on the results of these frameworks, we created a new set of features that primarily capture information about musical texture, rhythm, dynamics, expressivity, tone color, and musical form. To validate our work, we utilized the MERGE dataset, which comprises over 3000 30-second audio clips annotated with Russell's emotion quadrants. To evaluate the impact of the new features, we compared classification results with those obtained using current state-of-the-art features, demonstrating statistically significant improvements in F1 score (from 71.1% to 74.2%). Moreover, the novel features helped to reduce the confusion between quadrants 3 and 4 (a common difficulty in MER models). The most significant finding of the present study is the impact of separately analyzing the drum channel, whose features proved particularly relevant.**

***Index Terms*—Music Emotion Recognition, Percussion, Instrumentation, Musical Texture, Music Information Retrieval, Feature Engineering, Machine Learning, Music Source Separation, Automatic Music Transcription**

I. Introduction

In light of the tremendous expansion of available music in recent years, there has been a significant focus on categorization and filtering strategies to manage this extensive volume of data effectively. Music Emotion Recognition (MER) research addresses this challenge by utilizing machine learning (ML) methodologies to identify the emotional content within musical compositions. Various facets within this domain have been explored, including music emotion classification [1]–[3], music emotion variation detection [4], [5], and automatic playlist generation [6], [7], among others.

Automatic emotion-based song classification is a challenging problem due to the diverse range of variables that can differ from song to song, including the vocalist, instrumentation, tonality, melody, tempo, dynamics, and other characteristics.

H. Redinho, P. L. Louro, A. C. Santos, R. Malheiro, R. P. Paiva, and R. Panda are with the University of Coimbra, CISUC, Department of Informatics Engineering, LASI, Portugal. Email: redinho@student.dei.uc.pt, {pedrolouro,andresantos,rsmal,ruipedro,panda}@dei.uc.pt.

R. Malheiro is also with the Polytechnic Institute of Leiria, School of Technology and Management, Portugal.

R. Panda is also with the Ci2 — Smart Cities Research Center, Polytechnic Institute of Tomar, Portugal.

Music Emotion Recognition has been addressed through classical feature engineering, deep learning, single-modality (audio-only or lyrics-only), and bimodal audio-lyrics approaches, among others. Regarding classical feature engineering methodologies, Panda et al. [8], [9] sustained, in a recent survey, that one of the key MER challenges is the lack of acoustic emotionally relevant features, as most approaches utilize features designed for other audio analysis tasks (e.g., speech recognition or genre classification). There, it was demonstrated that features designed explicitly for emotion recognition are needed to narrow the so-called semantic gap and break the current glass ceiling in MER (as well as in other Music Information Retrieval (MIR) problems) [10].

Besides feature engineering, several deep learning methodologies have been proposed in MER [11]–[18], mainly in Western musical traditions, but also in other cultures, e.g., [19]. This includes a recent work by our team where a hybrid methodology combining traditional feature engineering with feature learning outperformed feature engineering and feature learning methods when acting separately [18]. Moreover, the impact of combining handcrafted and learning features has also been discussed in other domains besides MER, e.g., [20], [21]. This highlights the potential of regarding feature engineering and feature learning as companions rather than competitors, and is particularly true in problems with not-so-large datasets, as is often the case in MER.

In addition, several systems have adopted bimodal approaches that combine audio and lyrics, achieving considerable improvements compared to systems using only one or the other [22], [23]. In fact, audio has been shown to be a better predictor of arousal, while lyrics are more relevant for predicting valence [24]. Audio-only MER techniques often struggle to distinguish between low-arousal quadrants in Russell's circumplex model, specifically quadrants 3 (sad) and 4 (relaxed) [9], [25]

Despite the increasing relevance of deep learning and bimodal audio-lyrics solutions, feature engineering remains a crucial component in MER. Therefore, this article focuses on the proposal of emotionally relevant features, particularly in the audio domain, following the research directions highlighted by the recently conducted survey in [8] and other limitations in the MER state of the art.

One of the limitations of current feature engineering methods in MER is that they primarily focus on the melodic elements, as in [9]. However, the percussive elements are also essential for conveying and recognizing emotions in music. Various music psychology studies have examined how factors such as the selection of percussive instruments, playing

techniques, dynamics, and timing influence emotional expression and perception [26], [27]. According to those studies, percussive elements such as the instrument and stroke type help discriminate between high and low arousal and positive and negative valence [26]. Besides percussion, the presence and blending of specific instruments (either melodic or percussive) can influence the textural aspects of music and are also associated with different emotions [27]. As pointed out in [8], musical texture features are notably underrepresented in MER. Hence, MER systems could benefit from computational models that explore the aforementioned percussive and instrumentation aspects of music.

Therefore, in this work, we aim to answer the following main research questions (RQ):

- RQ1. Can specific percussive features that target musical dimensions such as rhythm, texture, dynamics, and expressivity enhance MER systems?
- RQ2. Can specific markers of instrument presence and extent enhance MER systems?
- RQ3. Can such features reduce the confusion between the low arousal quadrants (quadrants 3 and 4)?

A possible method to address the first research question is to isolate the percussive elements of songs. Next, features that capture the aforementioned musical dimensions in the drum channel should be extracted. For the second research question, identifying the presence and extent of musical instruments in the songs is essential for finding an answer. The features resulting from the two approaches will help answer RQ3.

Therefore, we aim to help bridge the current semantic gap in MER by introducing novel feature sets that leverage automatic music transcription (with instrument identification) and music source separation methodologies, specifically Magenta MT3 [28] and Demucs [29], respectively. By integrating the outputs of these state-of-the-art frameworks and building upon our prior feature engineering contributions [9], we have created novel features[1] that primarily capture aspects of percussion and musical texture. Besides texture, we propose new features for rhythm, melody, dynamics, tone color, and musical form.

To validate our work, we used the MERGE dataset [24], which comprises 3554 song excerpts with a duration of approximately 30 seconds each. The dataset is briefly described in the upcoming sections.

To assess the effectiveness of the new features, we performed a comparative analysis of the classification results on this dataset between the proposed features and the state-of-the-art features from Panda et al. [9] (henceforth referred to as "baseline features"). This was done using Support Vector Machine (SVM) classifiers and the ReliefF feature selection algorithm, replicating the logic in [9] while varying only the feature set to ensure that differences in results arise solely from these changes (for the sake of a fair comparative analysis). In addition to SVMs, we also provide a comparison of different state-of-the-art classical and deep learning approaches. Moreover, we also report regression results for arousal-valence prediction, using Support Vector Regression (SVR).

[1]Available at: https://github.com/mir-cisuc/Percussion-and-Instrumentation-in-MER-Feature-Sets

Our experiments demonstrated that incorporating new features enhanced the F1 score for classifying music emotions from 71.0% to 74.2%, compared to using the baseline features. This model utilized the top-ranked 300 features, 108 of which were new. Moreover, the separation of the original sound waveform into drum and non-drum channels proved particularly significant. Based on this separation, rhythm, dynamics, and expressivity features from the drum stem, and texture features from MT3 the non-drum stem, were amongst the top-ranked. Notably, the novel features contributed to a reduction in the confusion between quadrants 3 and 4, addressing a common challenge in music emotion recognition classification.

The main contributions of this article are the following:

- a novel set of features for audio-based Music Emotion Recognition (MER) that capture the percussive and instrumentation content of songs;
- an analysis of the impact of the proposed features in music emotion classification;
- a comparative analysis of different classical machine learning and deep learning approaches for MER.

The paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed methods, including the datasets and features (both baseline and novel). Section 4 discusses the conducted experiments, the attained results, and their implications. Finally, Section 5 concludes with a summary of findings and future research directions.

II. Related Work

A. Music and Emotion

Understanding and defining emotions has been an enduring challenge for humans. In the field of emotion recognition, emotions are often categorized into three levels: expressed, perceived, and felt (or induced) emotions [30]. This work, like many others in the field, is specifically concerned with perceived emotions.

For quite some time, psychological researchers have discussed the representation and categorization of emotions, leading to the development of different emotion paradigms (such as categorical or dimensional) and their respective taxonomies (e.g., Hevner [31] and Russell's [25] emotion models).

In our research, we use Russell's circumplex model of emotion, which has been supported by multiple studies [32]. This model, also employed in our prior work [9], is a two-dimensional framework based on two main axes: valence (ranging from pleasure to displeasure, essentially depicting the polarity of emotion in terms of positive and negative states, or pleasantness) and arousal (representing the level of activation or energy in the emotional experience). The outcome, depicted in Figure 1, is a two-dimensional arousal-valence (AV) space, where the X-axis corresponds to valence and the Y-axis to arousal.

The framework comprises four quadrants with the following characteristics: Quadrant 1 (Q1) represents positive valence and arousal, corresponding to emotions such as excitement or enthusiasm. Quadrant 2 (Q2) reflects negative valence and positive arousal, encompassing emotions like anxiety, fear, or anger. Quadrant 3 (Q3) entails negative valence and arousal,

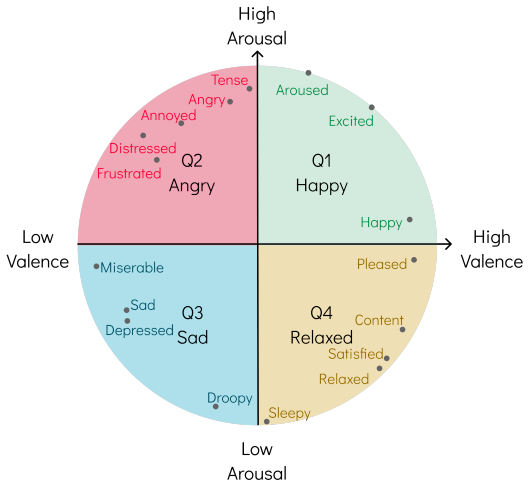


Fig. 1. Russell's Circumplex Model of Emotion (adapted from [25]).

denoting emotions like depression or sadness. Quadrant 4 (Q4) denotes positive valence and negative arousal, indicative of emotions like contentment or serenity.

B. Relations between Musical Elements and Emotions

In the study of the relations between musical elements and emotion, various features like articulation, dynamics, harmony, loudness, musical form, pitch, and rhythm have been associated with emotion in previous research (e.g., [7], [33], [34]). However, many of these relationships are not completely understood and require further investigation. Furthermore, there are certain attributes that pose challenges in terms of extraction and quantification from audio signals. To systematize the relations between musical elements and emotion, Panda et al. [8] employed the categorization proposed by Owen et al. [35], where musical attributes are classified into eight categories, each corresponding to a distinct musical concept: melody, harmony, rhythm, dynamics, tone color (or timbre), expressivity, musical texture, and musical form.

In Panda et al. [8], an extensive review of audio feature engineering in MER was conducted. The survey revealed that several musical dimensions, including texture, melody, expressivity, and rhythm, are underrepresented in MER and MIR. Furthermore, it was found that these underrepresented dimensions, particularly texture, are valuable for distinguishing songs with positive from negative valence [9]. For example, music psychology studies have shown that thinner textures are associated with positive emotions; in contrast, thicker textures are associated with negative emotions [36].

As a result, the creation of new features that effectively capture these musical dimensions is crucial.

As previously mentioned, current feature engineering methods in MER mainly focus on melodic elements, as in [9]. However, percussive elements are also vital for expressing and perceiving emotions in music. In general terms, as pointed out by Laukka and Gabrielsson [27], "[t]he communicative code used in emotional expression with drums appears to be similar to that used with other instruments". Therefore, musical dimensions such as rhythm, dynamics, texture, expressivity, tone color, and form are also employed in the emotional expression of drumming performances. For example, different drum patterns, i.e., grooves, play a relevant role in emotion expression [27].

Moreover, research shows that factors like the choice of percussive instruments and stroke types influence emotional expression and perception [26], [27]. For example, Huang et al. [26] concluded that the bass drum is strongly correlated with negative valence due to its deep sound (associated with somber emotions). In contrast, snare drums are perceived as more arousing due to their brighter sounds, which convey a sense of positivity and energy. A similar emotional connotation can be drawn to other percussion instruments, such as cymbals. Regarding stroke types, low-energy positive emotions (i.e., quadrant 4) are primarily linked to single-stroke sounds, as they provide a sense of steadiness. Conversely, drum rolls (a technique that allows the creation of tremolo with percussion instruments) add a sense of disturbance and tension, which is associated with negative valence. These rumbling sounds are likely to evoke terror and sadness alike, depending on the dynamics and instrument used; for example, playing louder can increase arousal, while playing softer can decrease it. In general, louder drum sounds lead to positive arousal (Q1 and Q2).

Therefore, MER systems could benefit from the design and creation of features that specifically capture the previously mentioned percussive elements of music. To this end, a possible approach is to isolate the percussive elements of songs, as discussed in the following subsection.

Besides percussion, instrumentation in general also has an impact on the expression and perception of emotion. The presence and blending of specific instruments (either melodic or percussive) can influence the textural aspects of music and is also associated with different emotions [27]. For example, in a study conducted by Chan et al. [37], 60% of general listeners perceived the clarinet as calm across various pitches and dynamics. Additionally, the instruments with the lowest pitch ranges from the bowed string, brass, and woodwind families (specifically the double bass, tuba, and bassoon) were generally rated as less happy and more sad or angry compared to other instruments within the same family.

Hence, MER systems could benefit from computational features that specifically capture the mentioned instrumentation and texture elements of music. One potential approach is to identify the presence and extent of various musical instruments in songs, as discussed in the following paragraphs.

C. Music Source Separation and Automatic Music Transcription

In our current research, we explore the development of new emotionally relevant musical features using two frameworks: Magenta MT3 and Demucs.

Magenta MT3 [28] is an automatic music transcription system that also identifies the notes played by different instruments (up to 128 instruments, as specified in the General MIDI Level 1 standard [38]). The output of MT3 is a MIDI

file containing the transcribed music. MT3 was evaluated using the usual metrics, namely Frame F1 (a binary metric on whether the predicted and final notes match), Onset F1 (a metric that considers a prediction correct if it has the same pitch and is within 50ms of the referenced onset), and finally Onset-Offset F1 (as the name suggests, this metric combines the aforementioned metrics, but now notes must also have matching endings). A final metric, multi-instrument F1, is also presented in [28], which combines Onset-Offset F1 with the requirement that the instrument predicted to play a certain note must match the original instrument of the reference note. The comparative analysis in [28] showed that MT3 outperforms all the compared approaches across six state-of-the-art datasets. However, the attained F1 scores vary considerably across datasets, ranging from 0.68 to 0.88, 0.5 to 0.96, 0.33 to 0.84, and 0.34 to 0.82 for the four metrics mentioned previously. This variability indicates that, despite notable advancements, further improvement is needed in both note and instrument recognition. This limitation affects other tasks, as in our study. Despite its inaccuracies, we follow a best-effort approach and employ the results from MT3 to create texture, rhythm, and melody features. Moreover, the method previously introduced by Panda et al. [9] was based on Dressler's multi-pitch estimator [39]. A limitation of this method is that it does not identify the musical instruments used, which requires frameworks such as MT3.

Regarding Demucs [29], this is a system for separating different music stems, such as vocals and drums, from an original audio track. This framework has been found to have superior performance compared to other state-of-the-art music source separation methods, achieving a final Signal-to-Distortion-Ratio (SDR) of 9.20 dB on the test set of the MUSDB dataset. In addition, human evaluations showed that Demucs also offers advantages in terms of audio naturalness. However, some artefacts are present in the separated sources, mainly due to bleeding (i.e., leakage of one source into another), especially between the vocals and other sources. As with MT3, despite its inaccuracies, we follow a best-effort approach and utilize the output from Demucs to generate features related to rhythm, particularly those associated with percussive elements, as well as dynamics, form, and tone color.

D. MER Datasets

Our approach is validated using the MERGE dataset [40] for the reasons that follow.

A set of preliminary audio MER datasets was previously suggested in the context of MER challenges. For example, the MIREX AMC dataset (for static MER) and later the DEAM dataset [41] (focused more on MEVD) were developed as a result of the consecutive benchmarks for the 2013, 2014, and 2015 MediaEval Emotion in Music tasks. However, these datasets, along with other MER datasets created over the years, have limitations, including inadequate emotion taxonomies, emotion classes with both acoustic and semantic overlap, noisy annotations, limited size, and noise.

The Million Song Dataset (MSD) [42] is a widely known large-scale dataset in the field of Music Information Retrieval (MIR). Data annotation in this dataset relies on user-provided tags from platforms like Last.fm[2]. However, it suffers from the limitations of approaches based on social tags, namely sparsity due to the cold-start problem and popularity bias, multiple spellings of tags, malicious tagging, or ad-hoc labeling techniques [43]. For example, when a subject uses the tag "hate" on Last.fm, this might either mean that the song is about "hate" or that the person hates the song.

Another alternative is, for example, the MuSe dataset [44]. The dataset includes annotations for 90,408 songs. However, it is important to note that the audio itself is not directly available. Instead, the tracks are identified by metadata, including Spotify IDs for 61,630 tracks. Nevertheless, according to Bogdanov et al. [45], only 41,021 30-second audio previews are currently accessible via the Spotify API. Similar to the MERGE dataset, emotion values are derived from social tags associated with music tracks on Last.fm by using Warriner's dictionary of emotional ratings of words [46]. However, contrary to MERGE, no human manual validation is performed in the MuSe dataset. Therefore, the previously mentioned problems of annotations based on social tagging are not handled. This may lead to noisy annotations, as confirmed by the low R^2 scores obtained for arousal and valence (0.143 and 0.089, respectively), as discussed in [45].

Among more controlled audio MER datasets, alternatives include EMOPIA [47] and TROMPA-MER [48]. Regarding EMOPIA, it contains 1087 pop piano recordings (and corresponding MIDI files) annotated into the four Russell quadrants. EMOPIA is focused on a specific instrument (piano) and genre (pop), lacking diversity. As for TROMPA-MER, this dataset comprises 1161 audio clips representing diverse cultures, including popular music from Spain and Portugal, music from the Griot tradition in West Africa, traditional and Popular Music from Latin America, the Middle East, and choir music. The song excerpts were annotated into 11 emotion tags related to the four Russell quadrants, aiming at personalized MER. Despite its high diversity and annotation quality control, its size is relatively small, which is why we favored the MERGE dataset.

Despite the potential of various datasets, MERGE is, to the best of our knowledge, the largest and most diverse controlled audio MER dataset, comprising 3554 audio clips that span a wide range of musical genres. This study selected MERGE for analysis, as outlined in the following section.

III. Materials and Methods

The overall architecture of our approach is summarized in Figure 2. The main building blocks are described in the following paragraphs. We begin by providing a brief introduction to the dataset used in this work. Next, we briefly describe the employed baseline (state-of-the-art) features. We then detail the novel features proposed in this work, organized into the different musical dimensions. Finally, we detail the evaluated emotion classification strategies.

[2]https://www.last.fm/.

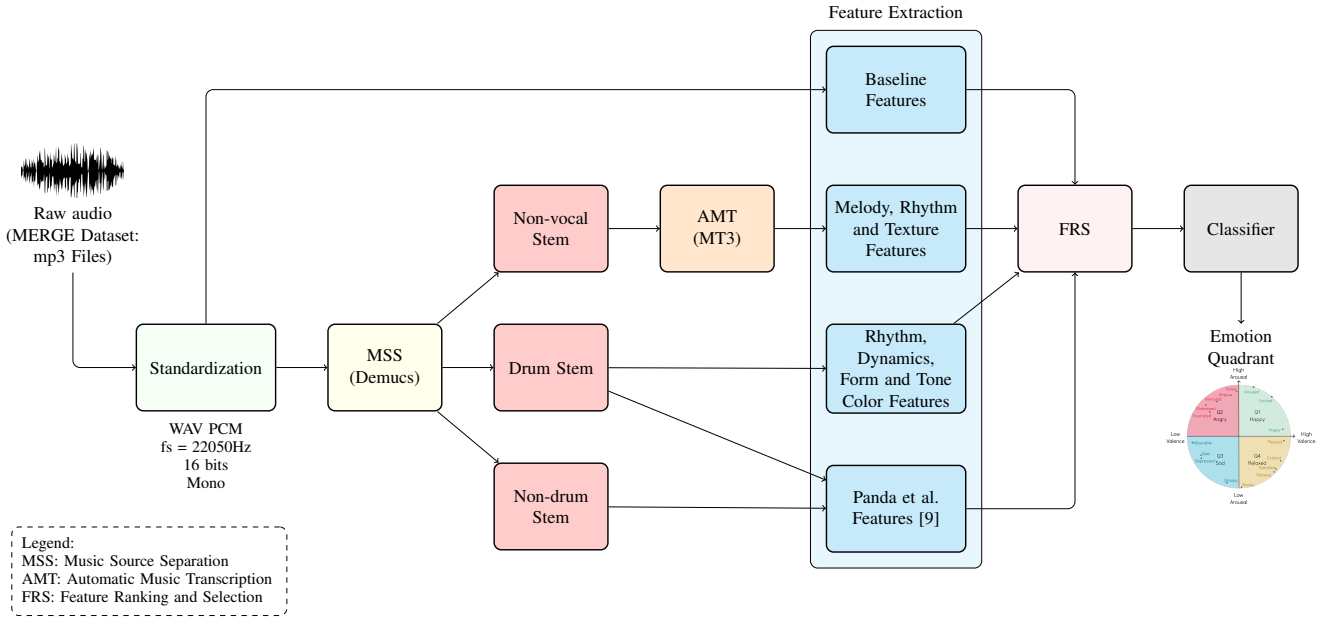


Fig. 2. Overall architecture of our approach.

A. Dataset

In this work, we utilized the bimodal (audio and lyrics) MERGE dataset (v1.1) [24]. This dataset comprises 3554 audio clips annotated into Russell's emotion quadrants. The median clip duration is 30.06 s. Most samples are stereo recordings (99.8%), encoded at 32 kHz (69.7%), followed by 22.05 kHz (24.0%) and 44.1 kHz (6.2%), with a median bitrate of 96 kbps and a maximum of 128 kbps, ensuring consistent audio quality across the collection. In this work, the MP3 audio clips from the MERGE dataset are converted into a standardized audio format using the FFmpeg framework, with the following specifications: WAV PCM Format, 22050 Hz sampling rate, 16-bit quantization, mono-aural.

For a comprehensive understanding of the dataset, please refer to [24].

The audio clips in the dataset originate from AllMusic[3], where they were annotated with emotional tags (e.g., happy, sad) by experts [49].

As in [9], our aim is to classify songs into Russell's quadrants. Therefore, we converted the AllMusic tags into arousal and valence (AV) values using the Warriner [46] dictionary. From these AV values, the corresponding quadrants were directly obtained. Since each song was originally annotated with multiple emotion labels, we selected the quadrant where the majority of tags were mapped.

Next, the proposed quadrants were manually validated (to see if the human annotations matched the quadrants from AllMusic), and only the songs where the quadrants matched were kept.

Six subsets were generated from the complete set of annotated songs, as detailed in Table I. These subsets are denoted as MERGE followed by the modality ("audio", "lyrics", or "bimodal") and the balance status ("complete", if it is the complete and unbalanced data, or "balanced" if it is a balanced subset of the complete data, where each quadrant has the same number of songs).

Table I shows the total number of songs for each of the six aforementioned datasets and the number of songs per quadrant.

TABLE I
NUMBER OF SONGS IN THE MERGE DATASETS.

Dataset Name	Q1	Q2	Q3	Q4	Total
MERGE Audio Complete	875	915	808	956	3554
MERGE Audio Balanced	808	808	808	808	3232
MERGE Lyrics Complete	600	710	621	637	2568
MERGE Lyrics Balanced	600	600	600	600	2400
MERGE Bimodal Complete	525	673	500	518	2216
MERGE Bimodal Balanced	500	500	500	500	2000

The dataset includes a lyrics component, which is not considered in this work, as it is focused on the analysis of the acoustic counterpart. Therefore, the only datasets utilized in this article are those containing audio clips, specifically the audio and bimodal datasets.

In addition to audio clips and lyrics, each dataset includes metadata and is divided into train-validation-test (TVT) sets. Two configurations are available for the TVT splits: 70-15-15 and 40-30-30 for training, validation, and testing, respectively.

B. Baseline Features

In the context of our work, we leverage the features previously proposed by Panda et al. [9], referred to as "baseline features". These features are the current state of the art of

[3]AllMusic is a music platform that provides, along with the audio clips, metadata and annotations for the songs in the platform. https://www.allmusic.com

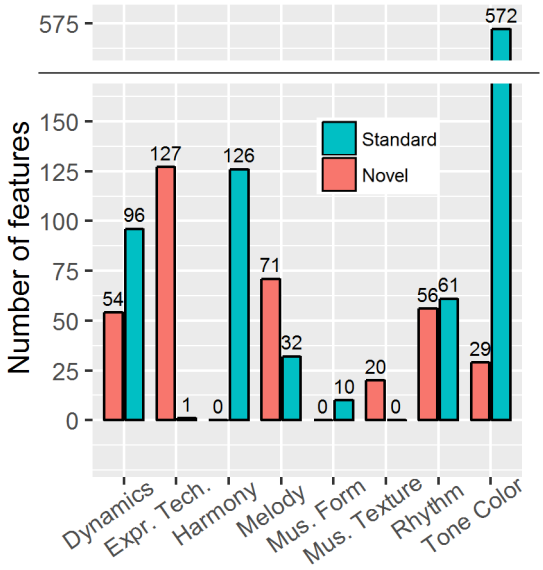


Fig. 3. Distribution of features for each musical dimension, separated by feature type (from [9]).

feature engineering MER approaches. They encompass the standard eight musical dimensions and are categorized into two groups: 1) standard features, extracted from audio frameworks such as Marsyas [50], MIR Toolbox [51], and PsySound3 [52], which have been widely used in previous studies prior to Panda et al.; and 2) new features introduced by Panda et al., which are focused on higher-level musical concepts, specifically expressivity and texture.

After initial analysis, the original feature set contained 1702 standard features, whereas Panda's new feature set contained 1086 features. Out of these, 543 were extracted from the original sound waveform, and the remaining 543 were generated from the vocal-separated audio clip. Following redundancy analysis [9], the final feature set consisted of 1255 acoustic features, with 898 from the original standard set and 357 from the new set. The distribution of these baseline features across the eight musical dimensions is illustrated in Figure 3, from [9]. Further information on these baseline features can be found in [9].

C. Novel Features

As mentioned earlier, this work proposes a set of novel percussive and instrumentation features by leveraging automatic music transcription and music source separation. In this section, we describe the proposed features, which are organized into the respective musical dimensions.

1) Music Source Separation and Automatic Music Transcription:

We use Demucs to perform two audio separations: 1) separate the vocal track and non-vocal track, and 2) separate the drum part and the non-drum (melodic) part of the track.

Using MT3, MIDI files with the notes for each instrument were extracted. As MT3 was trained on files with no vocals, the track with no vocals (resulting from the Demucs output) was used as input for MT3. This complements the analysis of the main melodic line, conducted mostly in the vocals, as performed in Panda et al. [9].

2) MT3: Novel Melody, Rhythm and Texture Features:

We employ MT3 [28] as a basis to create novel melody, rhythm, and texture features.

In Panda et al. [9], the extraction of melodic features (e.g., notes, register distribution, and melodic contour statistics) relied on the melody detection algorithm proposed by Dressler [39]. As such, only features from the main melody were created.

To extend this approach, in this work, we exploit MT3 to perform full music transcription (using the non-vocal stem, as aforementioned). As mentioned above, up to 128 instruments are transcribed according to the General MIDI Level 1 standard. From this transcription, note duration statistics are computed for the set of all melodic instruments. Six common statistics are employed: mean, standard deviation, skewness, kurtosis, maximum, and minimum. This was performed for the sake of consistency with [9].

Regarding rhythm features, note duration statistics from percussion instruments are computed (six features).

As for texture features, Panda et al. [9] previously extracted features mostly related to the statistics of musical layers, i.e., the number of simultaneous notes at each moment. These relied on the multi-pitch estimator by Dressler [39] (an extended version of Dressler's melody detection algorithm, from which multiple pitches in each short-time frame were estimated).

Here, we take advantage of the transcription performed by MT3 and compute the same texture features, but now within specific musical layers. To this end, it was important to categorize the 128 MIDI instruments into distinct groups. The classification most commonly used in the literature is the one proposed by Sachs et al. [53]. Therefore, that was the one chosen for our work. The authors organize musical instruments into five categories: idiophones, chordophones, membranophones, aerophones, and electrophones. Moreover, the MIDI standard also performs a categorization by grouping instruments into "melodic" or "percussive". We also employ this binary organization in this work. Hence, texture features specific to idiophones, chordophones, membranophones, aerophones, and electrophones, as well as melodic and percussive instruments, are computed for a total of 154 features.

Based on the information on individual instruments (as well as their respective categories), we also propose the following novel texture features:

1) **Instrument presence** (*Texture*)**.** In [9], measuring the presence or absence of a certain instrument in an audio track was impossible. With the transcription performed by MT3, this is now possible. To encapsulate this, for each of the 128 MIDI instruments, a one-hot-encoding feature was created with the value of 1 if any note of that instrument is played throughout the audio track and 0 otherwise. Furthermore, we also check the presence of percussion instruments, resulting in 47 features that correspond to the 47 percussive sounds

from MIDI channel 10 (this channel is exclusively reserved for percussion instruments) [54]). Hence, the same one-hot encoding reasoning was applied for the 47 percussion instruments. Furthermore, the total number of instruments of each specific group present in the song is calculated for each of the five instrument groups, as well as for the binary instrument grouping (melodic versus percussive).

2) **Instrument notes** (*Texture*)**.** This feature represents the number of notes of a specific instrument in a song. The total amount of melodic, percussion, and notes for each of the five instrument groups is also calculated. The same reasoning was applied to the 47 percussion instruments mentioned above.

3) **Instrument duration percentage** (*Texture*)**.** Based on note duration from MT3 transcription, we calculate the percentage of time an instrument plays throughout each song. For this, the sum of the durations of all notes of a particular instrument is calculated. As before, the same approach was followed for the 47 percussion instruments. Then, that amount is divided by the length of the song. Once again, this process was repeated for each of the five instrument categories and the binary groups.

In total, 700 new texture features were created. This high number of texture features results mostly from the one-hot encoding of the 128 MIDI instruments.

3) Demucs: Novel Rhythm, Dynamics, Form and Tonal Color:

Relying on the music source separation provided by Demucs [29], we extract features pertaining to rhythm, dynamics, form, and tone color. As mentioned earlier, one of Demucs's operating modes involves separating a song into its drum stem and non-drum (melodic) stem. We employed the drum stem as a basis for our novel features, as follows.

First, the audio waveform resulting from the drum stem is frame-wise analyzed using a 1024-sample frame length, with a hop size of 128 samples (to maintain consistency with the previously extracted features in Panda et al. [9]). Then, novel rhythm, dynamics, and form features, as well as typical spectral (tone color) feature statistics (see the paragraphs below), are extracted from each frame. The sequences resulting from frame-wise feature extraction are then summarized into the six previously mentioned statistics: mean, standard deviation, skewness, kurtosis, maximum, and minimum.

The following paragraphs describe the proposed percussion features:

1) **Drum extent percentage** (*Rhythm*)**.** This feature represents the amount of drums present in the percussion track. First, the Root-Mean-Square (RMS) energy for each frame is computed. Then, the number of frames above a certain threshold is calculated. The ratio between this number of frames and the total number of frames in the song gives the drum extent of the audio track. This threshold was defined experimentally as 0.025. More specifically, a sensitivity analysis was performed in the range 0.02-0.05, with a step of 0.005, leading to the obtained 0.025 threshold.
To determine this threshold, a separate small dataset containing 40 songs was used. This dataset consisted of 20 excerpts (each 30 seconds in duration) with low amplitude and percussive content (five excerpts from each quadrant) and 20 excerpts with a high amount of percussion (also five excerpts from each quadrant). Then, different threshold values were tested until the calculated drum extent accurately represented the audible amount of drum in the tracks.

2) **Drum amplitude and intensity information** (*Dynamics*)**.** Using the audio waveform, amplitude and intensity information were calculated.
Amplitude information relies on the Full-Wave-Rectified (FWR) waveform of the drum stem. We start by taking the absolute value of the waveform. Then, using the FWR waveform, we computed the six aforementioned statistics.
On the other hand, intensity information is based on the RMS energy of each frame, as described above. The usual six statistics were also computed.

3) **Self-Similarity Matrix (SSM) features** (*Form*)**.** Musical form, also known as musical structure, describes the layout or structure of a composition, typically divided into various sections. Self-similarity matrix (SSM) representations of music can help identify song structures [55], which we exploit as follows. Using Mel-Frequency Cepstral Coefficients (MFCC) derived from the audio signal, a set of features that aim to capture information regarding the similarity between each pair of frames in the audio signal is computed. First, the MFCCs are calculated (13 coefficients per frame). Then, the SSM [55] is computed from the MFCC matrix using cosine similarity as the distance metric. Finally, the SSM is thresholded to create a binary matrix, as in (1):

$$\text{Thresholded\_SSM} = \begin{cases} 1 & \text{if SSM} > 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Here, we denote cells with a value of one as positive cells, corresponding to frames for which a similarity was found. A sequence of positive cells forms a pattern. This SSM serves as a base from which the following features are proposed and extracted:

a) **Total number of positive cells.** Corresponds to the total number of positive cells where there are similarities.
b) **Total number of positive cells per second.** The total number of positive cells is divided by the total number of seconds in the song to get the total

number of positive cells per second.

c) **Pattern duration.** An array of pattern durations is computed, where all the repeated patterns' durations are kept. From this array, the six normal statistics are computed.

4) **Spectral features** (*Tone Color*)**.** A set of spectral features was extracted from the signal in each frame, namely: 1) spectral centroid, 2) spectral bandwidth, 3) spectral contrast, 4) spectral flatness, 5) spectral rolloff, and 6) spectral entropy. The same six statistics were computed.

4) Panda's Features Extracted from Demucs Stems:

Besides the previous features, we also extract Panda's new features, previously proposed in [9], here directly from the drum and non-drum stems after Demucs separation and referred to as "reused" features (in contrast to the other "strictly novel" features). This set comprises 543 features from each stem (as mentioned earlier, before feature redundancy analysis, Panda's new feature set consisted of 543 features).

In particular, the drum channel allows us to capture percussive features from the rhythmic, texture, dynamics, and expressivity dimensions. One example of the former is tremolo features (presence, rate, coverage, and salience, as proposed in [9]), which can capture the characteristics of drum rolls. In fact, in drum performance, a tremolo is typically performed as a roll. As previously stated, drum rolls create a sense of disturbance and tension, which is linked to negative valence.

5) Summary of Feature Extraction:

Table II summarizes the number of strictly novel features for each musical dimension (across MT3 and Demucs). In total, 769 features are proposed, comprising 712 features derived from the transcription performed by MT3 and 57 features derived from the source separation carried out by Demucs.

As can be seen, out of the 769 features, 653 belong to the texture dimension (84.9%). As previously mentioned, this high proportion of texture features stems largely from the one-hot encoding of the 128 MIDI instruments.

TABLE II
NUMBER OF NOVEL FEATURES FROM EACH NEWLY PROPOSED FEATURE SET (EXCLUDING REUSED FEATURES FROM PANDA ET AL. [9]).

Dimension	#Feat Total	#Feat MT3	#Feat Demucs
Melody	6	6	0
Rhythm	7	6	1
Dynamics	12	0	12
Tone Color	36	0	36
Texture	700	700	0
Form	8	0	8

In addition to these "strictly novel" features, we compute what we designated as "reused" features (further discussed in Table IX, Section IV-C). These are Panda's features extracted from the drum and non-drum stems (543 features each).

The feature extraction process can be summarized as in Algorithm 1.

Algorithm 1 Extraction of Novel Features

1. Separate the vocal and percussion tracks using Demucs, creating vocal, non-vocal, drum, and non-drum stems.
2. Standardize each of the four stems.
3. Extract the MIDI files using MT3 on the non-vocal stem.
4. Extract all proposed features related to the MIDI file (melody, rhythm, and texture features).
5. Extract the proposed features from the drum stem (rhythm, dynamics form, and tone color features).
6. Extract Panda's features on Demucs stems (drum and non-drum).

D. Feature Ranking and Dimensionality Reduction

Due to the high number of features used, performing feature reduction is paramount. Thus, feature redundancy analysis is performed as in [9].

We employed a two-stage pruning strategy to reduce redundancy and multicollinearity while retaining the most informative descriptors (full details are provided in [9]). First, features with zero variance are excluded. Next, we ranked the remaining descriptors using the ReliefF feature ranking algorithm [56]. We then iteratively removed redundant features, i.e., the lowest-ranked features (least significant) whose Pearson correlation with higher-ranked features exceeded 0.90. This resulted in a smaller, less redundant, and ranked feature set that is used in the following sections.

E. Emotion Classification and Regression

Regarding classification, SVMs [57] were selected as they were employed in our previous work [9]. To tune the hyperparameters, Bayesian search [58] was chosen instead of the grid search approach used in [9], as it achieved comparable results while taking less time to find the optimal set of parameters.

Additionally, we perform a comparative analysis employing several state-of-the-art methods:

- a Random Forest classifier [59], which employs the same handcrafted features used with the SVM;
- a Dense Neural Network (DNN), which also receives as inputs the same handcrafted features;
- a Convolutional Neural Network (CNN), which learns the most relevant patterns from Mel-spectrogram representations of the raw audio; here, the training data is increased using data augmentation techniques, as in [18];
- a hybrid CNN+DNN ensemble architecture, adapted from [18].

The hybrid DL architecture consists of an ensemble of the CNN and DNN models above, used as separate input branches, which combines handcrafted features and feature learning (Figure 4). Then, the outputs from the two branches (specifically, the class probabilities from each) are fed into a smaller DNN (a Convolutional 1D layer) that performs the final classification. The first two models are pre-trained before

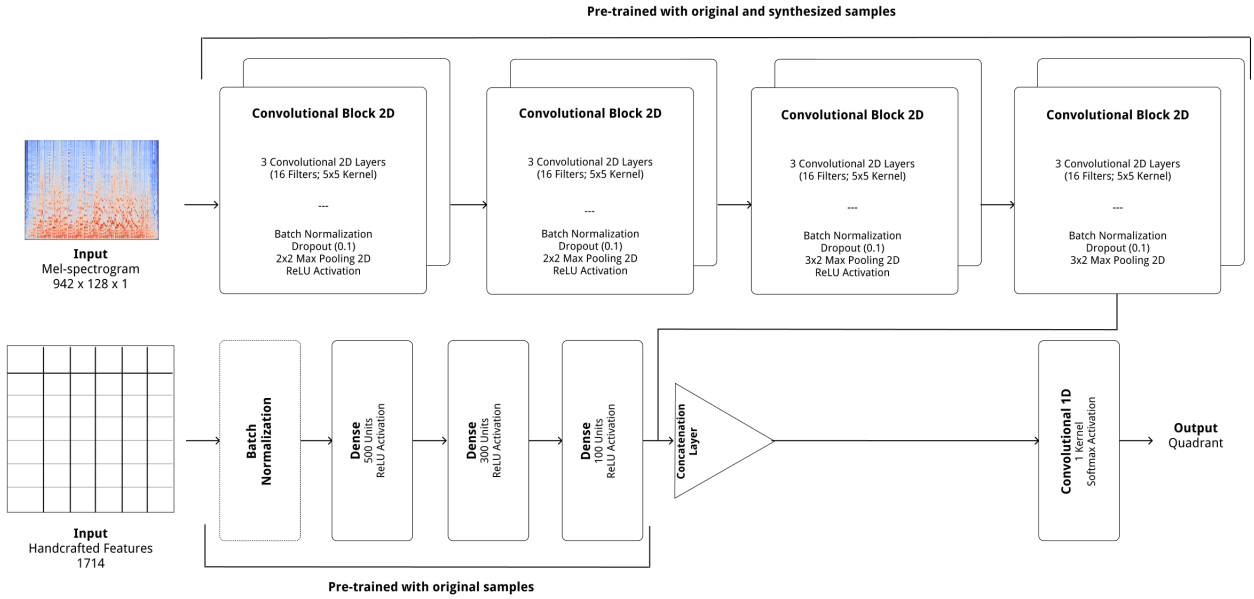


Fig. 4. The adapted hybrid deep learning architecture used in the present study. The feature processing and learning portion is as seen in [18], while the classification portion is substituted by a late fusion layer.

training the classification layer. In this study, instead of using the early fusion approach proposed in [18], we adopted a late fusion strategy due to its superior performance.

The experiments were then validated using two approaches: 1) 10 repetitions of 10-fold cross-validation (hereafter abbreviated to 10x10CV) [59], where we report the average (macro-weighted) results (e.g., F1 score); 2) the two aforementioned TVT splits (40-30-30 and 70-15-15), as proposed in the MERGE dataset [24].

Our study was mainly focused on the impact of the novel features on emotion classification, replicating the approach in [9]. For completeness, since the MERGE dataset also provides annotations for arousal and valence, we conducted a brief emotion regression analysis. Here, we trained Support Vector Regressors (SVR) on the 70-15-15 TVT split (the preferred data split recommended in the MERGE dataset [24]) and report R^2 scores.

In all experiments, we conduct statistical significance tests. For each subset, since we use the same test split for all comparisons, we perform paired tests. If the distributions are Gaussian (as determined by the Kolmogorov-Smirnov test), we use the paired t-test. Otherwise, we apply the Wilcoxon signed-rank test. We define a p-value of 0.05 as the threshold for statistical significance ($p < 0.05$). In all comparisons, we test the best-performing model against all the others.

IV. Results and Discussion

In this section, we discuss our classification results and assess the impact of our novel features on the audio MERGE dataset.

A. Classification Results

Table III summarizes the results obtained using only baseline features, only novel features, and the combination of both in the four MERGE data subsets, across three data-splitting scenarios (10x10CV, TVT 40-30-30, and TVT 70-15-15). Additionally, the average results for the four MERGE subsets are reported.

As can be observed, the combination of baseline and novel features improves the F1 score in all subsets and data-splitting alternatives (with statistical significance).

In terms of data splitting, the average F1 scores of the four datasets improved from 71.0% to 73.2%, 67.6% to 68.8%, and 68.1% to 71.3% statistically significant improvements of 2.2%, 1.2%, and 3.2%) for 10x10CV, TVT 40-30-30, and TVT 70-15-15, respectively.

Moreover, the 10x10CV experiment generally gave the highest and most stable results across the four subsets (standard deviation of 0.0 and 0.7 for the baseline and combined features, respectively), whereas the 70-15-15 TVT splitting was the one with the largest variability (standard deviation of 1.7 and 2.8 for the baseline and combined features, respectively). This is justified by the fact that the repetition process conducted in 10-fold cross-validation was less sensitive to the specific aspects of a particular test split. In TVT 70-15-15, the fact that only one (and smaller) test set was employed might justify its larger variability across data subsets. On the other hand, the use of TVT 70-15-15 generally leads to better results in comparison to TVT 40-30-30. This might be justified by its larger training set. Hence, we observe a trade-off between classification performance and variability in TVT: larger training sets lead to higher F1 scores at the expense of a larger variability across data subsets.

A closer look at specific data subsets reveals that the best results are almost always attained in the Bimodal Complete subset (an exception is TVT 70-15-15 using only baseline features, where the best results of 70.1% were achieved in the audio complete subset). In particular, in the Bimodal Complete subset, results improved from 71.0% to 74.2%, 69.2% to

TABLE III
SUMMARY OF RESULTS USING BASELINE FEATURES, NOVEL FEATURES, AND THEIR COMBINATION IN THE DIFFERENT MERGE DATA SUBSETS ACROSS THREE SCENARIOS: 10X10CV, TVT 40-30-30, AND TVT 70-15-15. AVERAGE RESULTS OF THE FOUR DATA SUBSETS ARE HIGHLIGHTED IN GRAY. BEST OVERALL RESULTS ARE HIGHLIGHTED IN GREEN.
(#FEAT: NUMBER OF FEATURES USED)

Features	Dataset	10x10CV		TVT 40-30-30		TVT 70-15-15	
		F1 score	#Feat	F1 score	#Feat	F1 score	#Feat
Baseline only	Audio Complete	71.0% ± 2.3	200	66.4%	400	70.1%	250
	Audio Balanced	70.9% ± 2.3	200	68.2%	250	68.5%	250
	Bimodal Complete	71.0% ± 2.6	300	69.2%	400	67.6%	400
	Bimodal Balanced	71.0% ± 2.8	200	66.7%	300	66.1%	200
	Average	71.0% ± 0.0	225	67.6% ± 1.3	337.5	68.1% ± 1.7	275
Novel only	Audio Complete	63.9% ± 2.2	300	60.6%	200	61.9%	250
	Audio Balanced	63.6% ± 2.5	300	61.2%	250	61.5%	300
	Bimodal Complete	65.8% ± 2.8	300	63.9%	250	68.0%	300
	Bimodal Balanced	62.8% ± 3.4	250	61.5%	400	66.7%	250
	Average	64.0% ± 1.3	287.5	61.8% ± 1.4	275	64.5% ± 3.3	275
Baseline+Novel	Audio Complete	72.9% ± 2.2	200	67.5%	300	72.2%	400
	Audio Balanced	72.8% ± 2.4	200	69.2%	400	69.8%	250
	Bimodal Complete	**74.2% ± 2.7**	300	**70.0%**	400	**74.7%**	300
	Bimodal Balanced	72.8% ± 3.1	200	68.4%	300	68.3%	250
	Average	73.2% ± 0.7	225	68.8% ± 1.1	350	71.3% ± 2.8	300

70.0%, and 67.6% to 74.7% (improvement of 3.2%, 0.8% and 7.1%) for 10x10CV, TVT 40-30-30, and TVT 70-15-15, respectively. Here, the impact of the novel features was highest in the TVT 70-15-15 split, with a rise of 7.1%, whereas, in TVT 40-30-30, the scores only improved by 0.8%. These results indicate that the employed dataset influences the impact of the novel features, even though, as previously mentioned, their use improved the results in all the performed experiments.

Regarding the results obtained by employing only the novel features, we achieved 64.0%, 61.8%, and 64.5% F1 scores for 10x10CV, TVT 40-30-30, and TVT 70-15-15, respectively. These results are clearly above the chance level, which suggests the potential of the novel features for emotion recognition (as will be discussed later).

In addition to presenting the overall classification results, we also examine the results achieved in each quadrant. This analysis enables us to identify which emotions are more challenging to classify and the impact of the baseline and novel features on this process. Tables IV and V show the confusion matrices for the MERGE Bimodal Complete dataset using only the baseline features and using the combined feature set, respectively, for the 10x10CV experiment (which attained an average F1 score of 74.2%). Table VI highlights the differences[4] in the F1 score from the baseline and the baseline plus novel features for each quadrant.

As can be observed, a significantly higher proportion of songs from Q1 and Q2 were accurately categorized compared to Q3 and Q4. This confirms previous findings in the literature indicating that emotions with higher arousal are easier to recognize than emotions with lower arousal, e.g., [9]. In the higher arousal quadrants, Q2 obtained the highest F1 score.

[4]This matrix represents the element-wise subtraction of Table IV from Table V. Improvements are indicated by positive values along the diagonal (increased correct classifications) and negative values in the off-diagonal cells (reduced misclassifications).

TABLE IV
CONFUSION MATRIX FOR THE MERGE BIMODAL COMPLETE DATASET USING BASELINE FEATURES (PERCENT VALUES IN 10X10CV).

Actual \ Predicted	Q1	Q2	Q3	Q4
Q1	76.3	10.5	4.7	8.5
Q2	8.7	89.5	1.2	0.6
Q3	7.9	2.8	56.8	32.6
Q4	13.2	0.7	30.0	56.2

TABLE V
CONFUSION MATRIX FOR THE MERGE BIMODAL COMPLETE DATASET USING ONLY BASELINE PLUS NOVEL FEATURES (PERCENT VALUES IN 10X10CV).

Actual \ Predicted	Q1	Q2	Q3	Q4
Q1	78.6	8.3	4.2	8.8
Q2	7.0	91.5	1.2	0.3
Q3	7.9	3.3	60.1	28.7
Q4	11.4	0.5	27.9	60.2

The reason for this could be that several excerpts from Q2 are

TABLE VI
DIFFERENCES IN THE CONFUSION MATRICES FROM TABLES IV AND V. (PERCENT VALUES IN 10X10CV)

Actual \ Predicted	Q1	Q2	Q3	Q4
Q1	2.3	−2.2	−0.5	0.3
Q2	−1.7	2.0	0.0	−0.3
Q3	0.0	0.5	3.3	−3.9
Q4	−1.8	−0.2	−2.1	4.0

categorized under the heavy metal genre, which is known for its distinct, noise-like acoustic features. In fact, as discussed in Panda et al. [9], the best features to discriminate Q2 are those that capture dissonance, timbre (e.g., MFCCs), and spectral flatness (which indicates how noise-like the sound is). Our analysis confirmed the relevance of these features, as discussed in the following subsection.

This is consistent with other studies in the literature, which demonstrate the challenge of discerning valence in low-arousal quadrants [9]. In fact, some songs from these quadrants exhibit significant acoustic overlap and share musical characteristics related to contrasting emotional elements. For example, a song might have a happy accompaniment or melody with a sad voice or lyric. This aligns with the conclusions presented in Hong's work [60].

However, using only audio information, we observe an increase in the F1 score in all quadrants. Moreover, adding the new features only led to worse results in two cases: a slightly increased confusion between Q1 and Q4 (0.3%) and between Q3 and Q2 (0.5%). This demonstrates the relevance of the newly proposed features for MER studies.

Additionally, the quadrants with the highest increases are Q3 and Q4, with an increase of 3.3% and 4%, respectively. Therefore, the proposed features (namely, percussion features related to rhythm, dynamics, and expressivity, as discussed below) make a positive contribution to helping discriminate between Q3 and Q4, despite the potential for further improvement. The better separation between Q3 and Q4 is particularly important: these quadrants are opposite in valence but similar in arousal, and their confusion often leads to poor recommendations (e.g., recommending a sad song in a calm playlist). Improvements in this area have direct implications for the quality of music classification, playlist generation, and music recommendation systems. It is well known that lyrics play a significant role in distinguishing between low-valence and high-valence songs [40]. However, it is worth noting that by relying solely on acoustic information, this confusion can be reduced, as our results show.

We also performed a comparative analysis of the methods described in Section III-E, using the best-performing TVT split (70-15-15), summarized in Table VII.

TABLE VII
SUMMARY OF BEST RESULTS OBTAINED USING A COMBINATION OF BASELINE AND NOVEL FEATURES IN THE DIFFERENT MERGE DATA SUBSETS USING TVT 70-15-15 AS THE EVALUATION STRATEGY. BEST OVERALL RESULTS ARE HIGHLIGHTED IN GREEN. SOME OF THE PRESENTED DL-BASED METHODOLOGIES DO NOT REQUIRE PREVIOUS FEATURE EXTRACTION.
(#FEAT: NUMBER OF FEATURES USED)

Approach	Dataset	F1 score	#Feat
SVM	Audio Complete	72.2%	400
	Audio Balanced	69.8%	250
	Bimodal Complete	**74.7%**	300
	Bimodal Balanced	68.3%	250
Random Forest	Audio Complete	71.0%	400
	Audio Balanced	69.5%	250
	Bimodal Complete	**71.1%**	400
	Bimodal Balanced	66.6%	200
DNN	Audio Complete	**70.6%**	250
	Audio Balanced	69.5%	200
	Bimodal Complete	69.3%	200
	Bimodal Balanced	65.5%	250
CNN with Data Augmentation	Audio Complete	**75.9%**	—
	Audio Balanced	71.3%	—
	Bimodal Complete	69.8%	—
	Bimodal Balanced	69.4%	—
Hybrid CNN + DNN Ensemble	Audio Complete	75.0%	250
	Audio Balanced	**76.3%**	200
	Bimodal Complete	70.7%	200
	Bimodal Balanced	68.4%	250

When comparing purely feature engineering approaches (SVM, Random Forest, and DNN), we observe that the SVM model performs the best, achieving an F1 score of 74.7%. In contrast, the Random Forest model reaches 71.1%, while the DNN scores 70.6%. This finding is consistent with previous research [9]. Additionally, we observed that the size of the dataset affects the performance of the DNN approach, with larger datasets yielding higher accuracy scores, as expected.

As for purely deep learning approaches, the CNN model outperformed all classical approaches, including the SVM, on the largest dataset (MERGE Audio Complete), achieving a 75.9% F1 score. However, as expected, its scores decreased in the smaller datasets.

Finally, the best overall results were obtained with the hybrid CNN+DNN ensemble, which combines handcrafted features and feature learning, yielding a 76.3% F1 score using the MERGE Audio Balanced dataset. However, using the Audio Complete dataset, a score of 75.0% was obtained, which is lower than the result of the CNN model. Our hypothesis is that the DNN classification layer was overly simple. In the future, a more complex classification layer will be designed.

B. Regression Results

In the following, we conducted a brief emotion regression analysis. As previously mentioned, our focus was on emotion classification, replicating the approach in [9], for which a more comprehensive analysis was conducted.

Table VIII summarizes the results obtained using only baseline features and the combination of baseline and novel

features in the four MERGE data subsets, for the TVT 70-15-15 data split.

TABLE VIII
SUMMARY OF BEST RESULTS OBTAINED USING A COMBINATION OF BASELINE AND NOVEL FEATURES IN THE DIFFERENT MERGE DATA SUBSETS USING TVT 70-15-15 AS THE EVALUATION STRATEGY AND AN SVM AS A REGRESSION TASK. BEST OVERALL RESULTS ARE HIGHLIGHTED IN GREEN.
(#FEAT: NUMBER OF FEATURES USED)

Dataset	Emotion Axis	R2 Score (Baseline)	R2 Score (Baseline + Novel)
Audio Complete	Arousal	0.479	0.525
	Valence	0.364	0.362
Audio Balanced	Arousal	0.527	**0.541**
	Valence	0.306	**0.377**
Bimodal Complete	Arousal	0.498	0.514
	Valence	0.284	0.367
Bimodal Balanced	Arousal	0.522	0.519
	Valence	0.360	0.334

As shown, the combination of baseline and novel features increased the R^2 score for arousal in all cases, except for a slight decrease in the Bimodal Balanced subset. Moreover, the improvement was particularly noticeable in the Audio Complete subset (from 0.479 to 0.525). As for valence, substantial improvements were achieved in the Bimodal Complete subset (from 0.284 to 0.377) and in the Audio Balanced subset (from 0.306 to 0.377). As with arousal, the results for the Bimodal Balanced subset decreased slightly. The reason for this is unclear and should be further studied in the future.

C. Feature Analysis

As previously mentioned, the results obtained using only the new features were significantly better than random chance, indicating the potential of these features for emotion recognition (see Table III). This was further supported by the results obtained when combining the novel and baseline features, which outperformed using only the baseline features.

In the following paragraphs, we discuss the impact of specific features on the results attained in the Bimodal Complete dataset in the 10x10CV experiment (F1 score of 74.2%). As shown in Table III, the best score in this experiment was obtained with the top-ranked 300 features.

In total, out of 300 features, 108 are novel (either strictly novel or reused), i.e., 36%. Table IX shows the distribution of features across musical dimensions, framework (MT3 versus Demucs), and novelty level (strictly novel versus reused).

Regarding musical dimensions, most of the selected novel features capture texture information (60 of 108, i.e., 55.6%).

Regarding the employed music source separation and automatic music transcription platforms, 62 features are based on Demucs (25 from the drum stem and 37 from the non-drum stem), and 46 features are based on MT3.

In terms of the novelty of the selected features, 57 are strictly novel (i.e., 52.8% of the novel features and 19% of the total number of selected features). The other 51 features are reused features from Panda, namely melody (one feature), rhythm (two), dynamics (six), tone color (three), expressivity (25), and texture (14) features, all from the Demucs stems.

TABLE IX
NUMBER OF NOVEL FEATURES ACROSS MUSICAL DIMENSIONS, FRAMEWORKS (MT3 VS. DEMUCS), AND NOVELTY LEVELS (STRICTLY NOVEL VS. REUSED) IN THE TOP 300.
(MERGE BIMODAL COMPLETE DATASET AND 10X10CV)

Dimension	#Feat	Framework	#Feat	Novelty	#Feat
Melody	1	MT3	0	Strictly Novel	0
				Reused	0
		Demucs	1	Strictly Novel	0
				Reused	1
Rhythm	3	MT3	0	Strictly Novel	0
				Reused	0
		Demucs	3	Strictly Novel	1
				Reused	2
Dynamics	12	MT3	0	Strictly Novel	0
				Reused	0
		Demucs	12	Strictly Novel	6
				Reused	6
Tone Color	7	MT3	0	Strictly Novel	0
				Reused	0
		Demucs	7	Strictly Novel	4
				Reused	3
Expressivity	25	MT3	0	Strictly Novel	0
				Reused	0
		Demucs	25	Strictly Novel	0
				Reused	25
Texture	60	MT3	46	Strictly Novel	46
				Reused	0
		Demucs	14	Strictly Novel	0
				Reused	14

Finally, the distribution of strictly novel features is the following: rhythm (1 feature: drum extent percentage, based on Demucs drum stem), dynamics (6 features: drum amplitude and intensity information, from the Demucs drum stem), tone color (4 features: spectral features statistics, namely bandwidth, centroid, and contrast, also from the Demucs drum stem), and texture (46 features related to instrument extent, notes, and duration percentage, from the transcription conducted by MT3). Regarding the proposed musical form features, these did not prove relevant. The top-ranked form feature was the average pattern duration, which appeared only in position 493, i.e., below the top 300. A possible reason for this might be that the employed song excerpts have a duration of 30 seconds, which might be insufficient to capture structural variations in the song. This feature should be evaluated in scenarios where complete songs are employed.

Table X lists the top 20 features from the 300-feature set, where it can be observed that eight features are novel (i.e., 40%). The top 20 contains one rhythm, five dynamics, seven tone color, one expressivity, and six texture features. This confirms the findings in [9], where tone color (e.g., statistics from the magnitude spectrum and MFCC coefficients) and texture features (e.g., statistics about the number of musical layers) proved to be particularly relevant. Here, in addition to the original baseline texture features, three novel features of

this dimension ranked in the top 20 (obtained from the Demucs non-drum stem). Unlike Panda et al's work [9], where several expressivity features reached the top, in the current work, only one feature of this kind appeared in the top 20, notably tremolo salience, related to the presence of drum rolls. Nevertheless, 25 expressivity features (obtained from Demucs drum and non-drum stems) appear in the top 300.

The high rank of percussion features related to rhythm, dynamics, and expressivity (drum extent, intensity, and tremolo, respectively) confirms the conclusions of the studies discussed in Section II, particularly their ability to improve the discrimination between positive and negative valence (as well as positive and negative arousal). In addition to these features, the presence and extent of specific drum instruments also had a positive impact on the results. In particular, the instrument duration percentage of bass drums (linked to negative emotions) ranked 101st in the feature set.

Regarding the presence and extent of melodic instruments, these features were not as relevant as the percussive ones. Nevertheless, the presence of string instruments ranked 170th.

Moreover, as mentioned previously, the presence of features such as MFCCs, spectral flatness measure, and spectral dissonance in the top-20 confirmed the findings from [9] regarding their relevance for Q2.

The most significant finding of the present study is the impact of separately analyzing the drum channel, whose features proved particularly relevant. The separate analysis of the non-drum (melodic) stem was also relevant. In fact, all eight novel features result from the analysis in these two channels. The separated analysis of the drum channel led to the increased relevance of dynamics features, namely regarding drum amplitude and intensity, which are novel contributions. Moreover, rhythm (drum extent percentage from the drum stem) and texture features (from the non-drum stem, i.e., containing only the melodic component of the songs) also proved relevant.

Overall, from the set of eight new features, three are strictly novel and pertain to rhythm (drum extent percentage) and dynamics (drum intensity and amplitude information mean statistics), all from the source separation conducted by the Demucs framework. The other five are Panda's features extracted from the Demucs stems, related to musical texture (mean and std statistics of the number of layers, and the percentage of thicker texture), dynamics (mean note intensity), and expressivity (mean of tremolo salience).

Therefore, by focusing on the top 20 features, it is evident that all the new features are based on the source separation performed by Demucs, highlighting the impact of features derived from the separation of drum and non-drum stems.

Besides the improvement in classification performance, the fact that a significant percentage of novel features are ranked in the top 300 and top 20 demonstrates their relevance for Music Emotion Recognition.

D. Analysis of Classification Errors

To gain a clearer understanding of the classification errors, we conducted an analysis of the misclassified songs based on both audio listening and the songs' metadata.

1) *Listening-based Acoustic Analysis:*

We conducted an in-depth listening analysis of selected examples that were consistently wrong across the four dataset variants, observing recurring patterns, especially confusion between low-arousal quadrants Q3 (sad) and Q4 (calm). These findings reinforce previous literature highlighting the challenge in distinguishing subtle acoustic differences when arousal is low but valence differs. These misclassifications often occurred when musical components, i.e., vocals and instrumental (as well as lyrics), convey opposed or ambiguous emotional cues. Some examples are:

- MT0002903697[5] (Jimmy Buffett – Blue Heaven Rendezvous): Annotated as calm (Q4) but predicted as sad (Q3). Although the accompaniment is serene and relaxing, featuring a gentle guitar and laid-back rhythm, Jimmy Buffett's vocal expression and timbre, along with subtle vibrato, might introduce a bit of melancholy, creating emotional ambiguity.
- MT0028335228 (Jim Brickman – Love of My Life): Annotated as calm (Q4) but predicted as sad (Q3). While predominantly calm, specific piano passages with strongly marked chords and subtle vocal tension introduce emotional complexity, which may contribute to the model's confusion.
- MT0004910983 (The Moody Blues – Melancholy Man): Annotated as sad (Q3) but predicted as calm (Q4). The song's slow tempo, long and soft background chords, and smooth harmonies create a gentle, calming atmosphere. Although the lyrics and singing clearly express sadness, the soft dynamics and relaxed arrangement make the song sound calm. This mix of sad emotion with peaceful music likely caused the misclassification.

We also observed specific errors related to moderate arousal (especially between Q1 and Q4). Such cases involved songs with moderate rhythmic energy that span perceptual boundaries between calm and happy emotional states, sometimes containing internal dynamic variations (e.g., slightly energetic parts). Notable examples include:

- MT0008131425 (Duke Ellington – Danke Schoen): Annotated as happy/excited (Q1) but predicted as calm (Q4). Although featuring the typical jazz swing rhythm and related elements suggestive of moderate happiness, the smooth timbre, soft dynamics, and articulation create a relaxed atmosphere, making emotional categorization ambiguous.
- MT0000202045 (Bob Marley & the Wailers – Exodus) and MT0002803746 (Lena Horne – Tomorrow Mountain): Both annotated as calm (Q4) but predicted as happy/excited (Q1). Both tracks feature lively rhythms and dynamics that introduce a sense of movement and may be perceived as excitement. However, the relaxed vocals of Bob Marley, in the first example, or slightly softer dynamics near the end might have influenced both the classifier and the annotators.

Additionally, annotation ambiguity itself may cause misclassification:

[5]This code is the song ID in the dataset.

TABLE X
LIST OF THE BEST (TOP) 20 FEATURES

Rank	Feature	Type	Musical Dimension
1	Musical Layers Distribution (MLD3) - Thicker Texture (percentage)	Baseline	Texture
2	FFT Spectrum - Average Power Spectrum (median)	Baseline	Tone Color
3	Musical Layers (ML) statistics (mean)	Baseline	Texture
4	Mel Frequency Cepstral Coefficients - MFCC1 (mean)	Baseline	Tone Color
5	Musical Layers (ML) statistics (mean), extracted from the Demucs / non-drum stem	Novel	Texture (Demucs)
6	Drum Extent Percentage	Novel	Rhythm (Demucs)
7	Musical Layers Distribution (MLD3) - Thicker Texture (percentage), from Demucs / non-drum	Novel	Texture (Demucs)
8	FFT Spectrum - Spectral 2nd Moment (median)	Baseline	Tone Color
9	FFT Spectrum - Average Power Spectrum (mean)	Baseline	Tone Color
10	Drum Intensity Information (mean)	Novel	Dynamics (Demucs)
11	Musical Layers (ML) statistics (std)	Baseline	Texture
12	Drum Amplitude Information (mean)	Novel	Dynamics (Demucs)
13	Note Intensity (NI) statistics (mean), extracted from the Demucs / drum stem	Novel	Dynamics (Demucs)
14	Musical Layers (ML) statistics (std), extracted from the Demucs / non-drum stem	Novel	Texture (Demucs)
15	Spectral Flatness Measure - SFM15 (mean)	Baseline	Tone Color
16	Loudness (skewness)	Baseline	Dynamics
17	Spectral Flatness Measure - SFM12 (mean)	Baseline	Tone Color
18	Spectral Dissonance (S) (skewness)	Baseline	Dynamics
19	Spectral Crest Factor - SCF15 (mean)	Baseline	Tone Color
20	Tremolo Salience (mean), extracted from the Demucs / drum stem	Novel	Expressivity (Demucs)

- MT0002962258 (Beastie Boys – Body Movin'): Annotated as happy/excited (Q1) but predicted as tense/agitated (Q2). Despite rhythmic and danceable elements indicative of happiness, aggressive vocals and harsher sounds could justify tension, suggesting annotation difficulties in distinguishing between tense and excited expressions (near-neutral valence).
- MT0003649209 (The Ramones – Rock N' Roll High School): Annotated as happy/excited (Q1), predicted as Q2. The fast tempo and danceable punk rhythm are consistent with excitement and high arousal. On the other hand, the clip features distorted guitars, aggressive drumming, and raw, shouted vocal delivery, which may indicate annotation uncertainty rather than model error.

These examples highlight some of the main issues identified. In the future, a more thorough review of all misclassified clips, especially with blind input from multiple annotators on the more ambiguous cases, could enhance our understanding and help improve the dataset.

*2) **Metadata Analysis**:*

Complementing the acoustic analysis, we examined dataset metadata to identify if models struggled with specific genres, artists, or temporal periods. Among genres, Holiday music consistently showed the highest error rate (34.2%), significantly above the overall average (25.8%). This likely results from the characteristic acoustic profile of Holiday music, which often features emotionally nuanced lyrics that convey nostalgia or melancholy. Although the provided annotations pertain to the audio part only, annotators' familiarity and cultural associations with these songs may further complicate objective assessments. This complexity is even greater when artists known for a melancholic style perform Holiday covers.

Reggae and New Age genres also exhibited higher error rates (28.7% and 31.2%, respectively), although for different reasons. Reggae's distinctive rhythmic patterns and moderate tempos can blur perceived arousal levels, leading to songs being misclassified as calmer than intended. New Age music, with its ambient textures and minimalist arrangements, often creates ambiguity between calm and sad states, contributing to frequent Q3 versus Q4 confusions.

At the artist level, most misclassified ones over the four dataset variants are Bob Dylan and Willie Nelson. Bob Dylan's calm (Q4) tracks were frequently misclassified as sad (Q3), likely due to his distinctive vocal delivery and characteristic instrumental arrangements, particularly the prominent use of harmonica. Interestingly, most of these include Holiday genre tracks, such as "O' Little Town of Bethlehem" (MT0010290788) and "O' Come All Ye Faithful (Adeste Fideles)" (MT0010305995). Willie Nelson's case is similar, with calm (Q4) tracks, such as "Pretty Paper" (MT0006037085), misclassified as sad (Q3), likely due to his unique vocal timbre. These findings illustrate how both contradicting acoustic cues and strong cultural associations can challenge the annotation of perceived emotions. They also raise the question of whether MER classifiers sometimes are learning to identify specific artists rather than genuine emotional content, especially when distinctive vocal or instrumental signatures are present.

Finally, we observed a temporal trend in misclassification, with accuracy declining for songs from the 1970s and earlier. This pattern may result from the dataset being composed mostly of contemporary recordings, whereas older songs differ in style, instrumentation, and recording quality. For instance, Duke Ellington's "Danke Schoen" (MT0008131425), from the 1960s, illustrates how jazz-influenced arrangements and subtle rhythmic nuances can create emotional ambiguity.

V. Conclusions and Future Work

This work proposed several novel percussion and instrumentation features based on the outcomes of music source separation and automatic music transcription frameworks. The newly proposed features helped increase the obtained F1 score, achieving 74.2% with 300 features on the MERGE Bimodal Complete dataset, using 10x10-fold cross-validation. Compared to using only baseline features, a statistically significant improvement of 3.2% was observed.

Revisiting the three research questions posed in the introduction, regarding RQ1 (*Can specific percussive features that target musical dimensions such as rhythm, texture, dynamics, and expressivity enhance MER systems?*), we have demonstrated the answer to be yes. In fact, features such as drum extent percentage, drum amplitude information, drum intensity information, note intensity and tremolo salience (both from the drum stem) were among the top-ranked features.

As for RQ2 (*Can specific markers of instrument presence and extent enhance MER systems?*), the answer is also yes. As we have shown, of the 108 selected novel features, 46 are related to instrument extent, notes, and duration percentage, extracted from the transcription conducted by MT3.

Finally, regarding RQ3 (*Can such features reduce the confusion between the low arousal quadrants (quadrants 3 and 4)?*), the answer is, once again, yes. In fact, the newly proposed features primarily helped to decrease the confusion between the third and fourth quadrants, a well-known issue in MER. The better separation between Q3 and Q4 is particularly important: these quadrants are opposite in valence but similar in arousal, and their confusion often leads to poor recommendations (e.g., recommending a sad song in a calm playlist). Improvements in this area have direct implications for the quality of music classification, playlist generation, and recommendation systems.

Besides applications in Music Emotion Recognition, the proposed features might also be helpful in other Music Information Retrieval Problems, e.g., music genre classification.

Additionally, a significant proportion of novel features were present in the set of 300 selected features (36%). This percentage increased to 40% when considering the top 20 features, which demonstrates the relevance of the proposed features for MER studies. Moreover, the separation of the original sound waveform into drum and non-drum channels proved particularly relevant. Based on this separation, rhythm, dynamics and expressivity features from the drum stem and texture features from MT3 and the non-drum stem were amongst the top-ranked.

One limitation of this study is the absence of lyrics. As previously mentioned, audio-only MER systems often struggle to distinguish between low-arousal quadrants in Russell's circumplex model. Although our approach helped reduce the confusion between Q3 and Q4, practical MER systems likely require a bimodal approach that combines audio and lyrics. Therefore, in the future, we will explore bimodal approaches by combining audio and lyrics, as this is a promising approach to reducing confusion between Q3 and Q4, since valence information is mainly captured by the lyrics [61]. To further study the impact of bimodal approaches, both congruent (emotion matching) and incongruent (emotion mismatch) combinations of audio and lyrics should be addressed, e.g., analyzing songs with sad lyrics and melancholic musical features (congruent) or upbeat songs with sad lyrics (incongruent).

Moreover, despite the care to create a quality-controlled dataset, the combined acoustic and metadata analyses highlight the complexity of music emotion recognition and annotation. Besides incorporating lyrics, future ideas include genre-specific features or utilizing the genre itself as a feature, and improving annotation reliability through additional listening tests and blind re-annotations of challenging examples to strengthen the dataset.

Another limitation is the relatively small size of the MERGE dataset used. Although this is not a large-scale dataset, to our knowledge, it is the largest publicly available and quality-controlled bimodal static MER dataset. Nevertheless, despite its size limitations, the hybrid deep learning approach outperformed the purely feature engineering approach. This result confirms the need for larger, quality-controlled MER datasets to better exploit deep learning methodologies. In fact, the lack of quality and sizeable data is one of the most significant drawbacks of these approaches.

Still another limitation is the propagation of errors from MT3 and Demucs. As previously discussed, MT3, despite notable advancements, needs further improvements in both note and instrument recognition. As for Demucs, artifacts caused by bleeding are a current limitation.

Moreover, additional features could still be developed from the aforementioned tools to help achieve a more representative portrayal of the underrepresented musical dimensions.

Finally, this work, as most MER works, focuses on the Western musical culture. In the future, we plan to conduct MER studies with datasets such as TROMPA-MER [48] and JUMusEmoDB [19] (comprising sitar and sarod clips from Indian Classical Music) to assess whether the conclusions drawn from this work generalize to other musical traditions.

Acknowledgments

This work is funded by FCT - Foundation for Science and Technology, I.P., within the scope of the projects: MERGE - DOI: 10.54499/PTDC/CCI-COM/3171/2021 financed with national funds (PIDDAC) via the Portuguese State Budget; and project CISUC - UID/CEC/00326/2020 with funds from the European Social Fund, through the Regional Operational Program Centro 2020. Renato Panda was supported by Ci2 - FCT UIDP/05567/2020.

References

[1] Y. Feng, Y. Zhuang, and Y. Pan, "Popular Music Retrieval by Detecting Mood," in *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR 2003*, vol. 2, no. 2. Toronto, Canada: ACM Press, 2003, pp. 375–376. [Online]. Available: http://portal.acm.org/citation.cfm?doid=860435.860508

[2] C. Laurier and P. Herrera, "Audio Music Mood Classification Using Support Vector Machine," in *8th International Society for Music Information Retrieval Conference – ISMIR 2007*, Vienna, Austria, 2007, pp. 2–4.

[3] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A Regression Approach to Music Emotion Recognition," *IEEE Transactions on Audio, Speech, and Language Processing – TASLP*, vol. 16, no. 2, pp. 448–457, 2008. [Online]. Available: https://ieeexplore.ieee.org/document/4432654

[4] L. Lu, D. Liu, and H.-J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *IEEE Transactions on Audio, Speech, and Language Processing – TASLP*, vol. 14, no. 1, pp. 5–18, jan 2006. [Online]. Available: https://ieeexplore.ieee.org/document/1561259

[5] R. Panda and R. P. Paiva, "Using Support Vector Machines for Automatic Mood Tracking in Audio Music," in *130th Audio Engineering Society Convention 2011 (AES 130)*. London, UK: Audio Engineering Society, 2011, pp. 579–586.

[6] A. Flexer, D. Schnitzer, M. Gasser, and G. Widmer, "Playlist Generation Using Start and End Songs," in *9th International Society of Music Information Retrieval Conference – ISMIR 2008*, Philadelphia, Pennsylvania, USA, 2008, pp. 173–178.

[7] O. C. Meyers, "A Mood-Based Music Classification and Exploration System," MSc, Massachusetts Institute of Technology, 2007. [Online]. Available: https://dspace.mit.edu/handle/1721.1/39337

[8] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio Features for Music Emotion Recognition: a Survey," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 68–88, oct 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9229494/

[9] R. Panda, R. Malheiro, and R. P. Paiva, "Novel Audio Features for Music Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, oct 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8327886/

[10] Ò. Celma, P. Herrera, and X. Serra, "Bridging the Music Semantic Gap," in *Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, held in conjunction with the European Semantic Web Conference*, vol. 187, no. 2, Budva, Montenegro, 2006, pp. 177–190.

[11] M. Bilal Er and I. B. Aydilek, "Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features," *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, p. 1622, 2019. [Online]. Available: https://www.atlantis-press.com/article/125927469

[12] M. Russo, L. Kraljević, M. Stella, and M. Sikora, "Cochleogram-based approach for detecting perceived emotions in music," *Information Processing and Management*, vol. 57, no. 5, p. 102270, sep 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306457319310635

[13] J. S. Gomez Canon, E. Cano, P. Herrera, and E. Gomez, "Transfer Learning from Speech to Music: Towards Language-Sensitive Emotion Recognition Models," in *2020 28th European Signal Processing Conference (EUSIPCO)*. Amsterdam, Netherlands: IEEE, jan 2021, pp. 136–140. [Online]. Available: https://ieeexplore.ieee.org/document/9287548/

[14] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology, an International Journal*, vol. 24, no. 3, pp. 760–767, jun 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2215098620342385

[15] N. He and S. Ferguson, "Music emotion recognition based on segment-level two-stage learning," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 3, pp. 383–394, sep 2022. [Online]. Available: https://link.springer.com/10.1007/s13735-022-00230-z

[16] X. Han, F. Chen, and J. Ban, "Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure," *Electronics*, vol. 12, no. 4, p. 978, feb 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/4/978

[17] Y. R. Pandeya and J. Lee, "GlocalEmoNet: An optimized neural network for music emotion classification and segmentation using timbre and chroma features," *Multimedia Tools and Applications*, vol. 83, no. 30, pp. 74 141–74 158, feb 2024. [Online]. Available: https://link.springer.com/10.1007/s11042-024-18246-4

[18] P. L. Louro, H. Redinho, R. Malheiro, R. P. Paiva, and R. Panda, "A Comparison Study of Deep Learning Methodologies for Music Emotion Recognition," *Sensors*, vol. 24, no. 7, p. 2201, mar 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/7/2201

[19] S. Nag, M. Basu, S. Sanyal, A. Banerjee, and D. Ghosh, "On the application of deep learning and multifractal techniques to classify emotions and instruments using indian classical music," *Physica A: Statistical Mechanics and its Applications*, vol. 597, p. 127261, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437122002291

[20] J. Li and Y. Zhang, "The Death of Feature Engineering? BERT with Linguistic Features on SQuAD 2.0," apr 2024, preprint. [Online]. Available: http://arxiv.org/abs/2404.03184

[21] D. Gibert, J. Planes, C. Mateu, and Q. Le, "Fusing feature engineering and deep learning: A case study for malware classification," *Expert Systems with Applications*, vol. 207, p. 117957, nov 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417422011927

[22] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music Mood Detection Based on Audio and Lyrics with Deep Neural Net," in *19th International Society for Music Information Retrieval Conference – ISMIR 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., Paris, France, 2018, pp. 370–375.

[23] K. Pyrovolakis, P. Tzouveli, and G. Stamou, "Multi-Modal Song Mood Detection with Deep Learning," *Sensors*, vol. 22, no. 3, p. 1065, jan 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/3/1065

[24] P. L. Louro, H. Redinho, R. Santos, R. Malheiro, R. Panda, and R. P. Paiva, "MERGE - A Bimodal Dataset for Static Music Emotion Recognition," jul 2024, preprint. [Online]. Available: http://arxiv.org/abs/2407.06060

[25] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980. [Online]. Available: http://content.apa.org/journals/psp/39/6/1161

[26] Z. Huang, W. Song, X. Ma, and A. B. Horner, "The emotional characteristics of bass drums, snare drums, and disengaged snare drums with different strokes and dynamics," *Proceedings of Meetings on Acoustics*, vol. 52, no. 1, p. 035005, 04 2024. [Online]. Available: https://doi.org/10.1121/2.0001834

[27] P. Laukka and A. Gabrielsson, "Emotional expression in drumming performance," *Psychology of Music*, vol. 28, no. 2, pp. 181–189, 2000. [Online]. Available: https://doi.org/10.1177/0305735600282007

[28] J. P. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "MT3: Multi-Task Multitrack Music Transcription," in *International Conference on Learning Representations (ICLR 2022)*, nov 2022. [Online]. Available: https://openreview.net/forum?id=iMSjopcOn0p

[29] S. Rouard, F. Massa, and A. Défossez, "Hybrid Transformers for Music Source Separation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, jun 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10096956/

[30] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Scientiae*, vol. 5, pp. 123–147, sep 2001. [Online]. Available: http://journals.sagepub.com/doi/10.1177/10298649020050S105

[31] K. Hevner, "Experimental Studies of the Elements of Expression in Music," *The American Journal of Psychology*, vol. 48, no. 2, p. 246, apr 1936. [Online]. Available: http://www.jstor.org/stable/1415746

[32] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–34, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16262989

[33] C. Laurier, O. Lartillot, T. Eerola, and P. Toiviainen, "Exploring relationships between audio features and emotion in music," in *7th Triennial Conference of European Society for the Cognitive Sciences of Music – ESCOM 2009*, vol. 3. Jyväskylä, Finland: European Society for Cognitive Sciences of Music, 2009, pp. 260–264.

[34] A. Friberg, "Digital Audio Emotions - An Overview of Computer Analysis and Synthesis of Emotional Expression in Music," in *11th International Conference on Digital Audio Effects – DAFx 2008*, Espoo, Finland, 2008, pp. 1–6.

[35] H. Owen, *Music theory resource book*. Oxford University Press, 2000.

[36] G. D. Webster and C. G. Weir, "Emotional responses to music: Interactive effects of mode, texture, and tempo," *Motivation and Emotion*, vol. 29, no. 1, pp. 19–39, Mar. 2005.

[37] H. T. Chan, B. Y. Chang, A. B. Horner, and M. H. Law, "Comparison of the emotional characteristics of western orchestral sustaining musical instrument families with different pitch and dynamics," *Proceedings of Meetings on Acoustics*, vol. 50, no. 1, p. 035001, 03 2023. [Online]. Available: https://doi.org/10.1121/2.0001712

[38] MIDI Manufacturers Association, *The Complete MIDI 1.0 Detailed Specification*, Los Angeles, CA, 1996. [Online]. Available: https://midi.org/midi-1-0-detailed-specification

[39] K. Dressler, "Automatic Transcription of the Melody from Polyphonic Music," Ph.D. dissertation, Ilmenau University of Technology, 2016.

[40] P. L. Louro, H. Redinho, R. Santos, R. Malheiro, R. Panda, and R. P. Paiva, "Merge dataset," Oct. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13904708

[41] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS ONE*, vol. 12, no. 3, mar 2017. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0173392

[42] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *12th International Society for Music Information Retrieval Conference – ISMIR 2011*, Miami, Florida, USA, 2011, pp. 591–596.

[43] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music Emotion Recognition: A State of the Art Review," in *11th International Society for Music Information Retrieval Conference – ISMIR 2010*, Utrecht, Netherlands, 2010, pp. 255–266.

[44] C. Akiki and M. Burghardt, "MuSe: The Musical Sentiment Dataset," *Journal of Open Humanities Data*, vol. 7, jul 2021. [Online]. Available: http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.33/

[45] D. Bogdanov, X. Lizarraga-Seijas, P. Alonso-Jiménez, and X. Serra, "MusAV: A dataset of relative arousal-valence annotations for validation of audio models," in *23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, Bengaluru, India, dec 2022. [Online]. Available: https://ismir2022program.ismir.net/poster_286.html

[46] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, dec 2013. [Online]. Available: http://link.springer.com/10.3758/s13428-012-0314-x

[47] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, Online, nov 2021, pp. 318–325.

[48] J. S. Gómez-Cañón, N. Gutiérrez-Páez, L. Porcaro, A. Porter, E. Cano, P. Herrera-Boyer, A. Gkiokas, P. Santos, D. Hernández-Leo, C. Karreman, and E. Gómez, "TROMPA-MER: an open dataset for personalized music emotion recognition," *Journal of Intelligent Information Systems*, vol. 60, no. 2, pp. 549–570, apr 2023. [Online]. Available: https://link.springer.com/10.1007/s10844-022-00746-0

[49] X. Hu and J. S. Downie, "Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata," in *8th International Society for Music Information Retrieval Conference – ISMIR 2007*. Vienna, Austria: Österreichische Computer Gesellschaft, 2007, pp. 67–72.

[50] G. Tzanetakis and P. Cook, "MARSYAS: a framework for audio analysis," *Organised Sound*, vol. 4, no. 3, pp. 169–175, 2000. [Online]. Available: https://www.cambridge.org/core/journals/organised-sound/article/marsyas-a-framework-for-audio-analysis/43A5D9BCB0F7BB439E1D4D1FF4B563C2

[51] O. Lartillot and P. Toiviainen, "A Matlab Toolbox for Musical Feature Extraction from Audio," in *10th International Conference on Digital Audio Effects – DAFx 2007*, Bordeaux, France, 2007, pp. 237–244.

[52] D. Cabrera, S. Ferguson, and E. Schubert, "'Psysound3': Software for Acoustical and Psychoacoustical Analysis of Sound Recordings," in *13th International Conference on Auditory Display – ICAD2007*, G. P. Scavone, Ed. Schulich School of Music, McGill University, 2007, pp. 356–363.

[53] C. Sachs, "Die Hornbostel-Sachs'sche Klassifikation der Musikinstrumente," *Die Naturwissenschaften*, vol. 2, no. 51, pp. 1056–1059, dec 1914. [Online]. Available: http://link.springer.com/10.1007/BF01495319

[54] MIDI Manufacturers Association, *General MIDI System Level 1 Specification*, Los Angeles, CA, 1996, page 6: General MIDI Percussion Map (Channel 10). [Online]. Available: https://midi.org/general-midi-level-1

[55] M. Müller and M. Clausen, "Transposition-Invariant Self-Similarity Matrices," in *8th International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[56] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.

[57] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology – TIST*, vol. 2, no. 3, pp. 1–27, apr 2011. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1961189.1961199

[58] J. Brownlee, *Probability for Machine Learning: Discover How To Harness Uncertainty With Python*. Machine Learning Mastery, 2019.

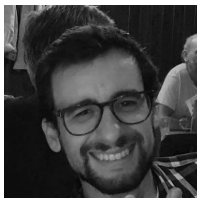
[59] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. Wiley, 2000.

[60] Y. Hong, C.-J. Chau, and A. Horner, "An Analysis of Low-Arousal Piano Music Ratings to Uncover What Makes Calm and Sad Music So Difficult to Distinguish in Music Emotion Recognition," *Journal of the Audio Engineering Society – JAES*, vol. 65, no. 4, pp. 304–320, 2017. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=18563

[61] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, "Emotionally-Relevant Features for Classification and Regression of Music Lyrics," *IEEE Transactions on Affective Computing – TAFFC*, vol. 9, no. 2, pp. 240–254, apr 2018. [Online]. Available: http://ieeexplore.ieee.org/document/7536113/

Hugo Redinho is an MSc from the University of Coimbra, specializing in Intelligent Systems, where he also concluded his Bachelor's degree in Informatics Engineering. He is a member of the Cognitive and Media Systems (CMS) research group at the Center for Informatics and Systems of the University of Coimbra (CISUC). His main research interests are related to Music Emotion Recognition (MER) and Music Information Retrieval (MIR).

Pedro Louro is a PhD Research Student at the University of Coimbra, where he also concluded his Masters degree, specializing in Intelligent Systems. He is a member of the CMS research group at CISUC. His main research interests include Music Information Retrieval (MIR), Music Emotion Recognition (MER), and Deep Learning.



André Carvalho dos Santos is a PhD Student in Informatics Engineering at the University of Coimbra (UC). He holds a BSc and MSc in Electrical and Computer Engineering from Instituto Superior Técnico (IST). He was awarded a Fulbright Scholarship and was a visiting researcher at Carnegie Mellon University (CMU). His research is focused on MIR and Computational Creativity (CC).

Ricardo Malheiro is a PhD from the University of Coimbra, where he also concluded his Master and Bachelor (Licenciatura - 5 years) degrees, respectively in Informatics Engineering and Mathematics. He is a Professor at the Polytechnic Institute of Leiria - School of Technology and Management. He is also a member of the CMS research group at CISUC. His main research interests are Natural Language Processing, Detection of Emotions in Music Lyrics and Text, and Text/Data Mining.

Rui Pedro Paiva is a Professor at the Department of Informatics Engineering of the University of Coimbra, where he concluded his Doctoral, Master and Bachelor degrees in 2007, 1999 and 1996, respectively. He is also a member of the CMS group at CISUC. His main research interests are in MIR and Health Informatics. The common research hat is the study of feature engineering, machine learning, and signal processing to analyze musical and bio signals.

Renato Panda is an Auxiliary Researcher at Ci2, Polytechnic Institute of Tomar, Portugal. His main research interests are Music Emotion Recognition (MER) and Music Information Retrieval (MIR), as well as Applied Machine Learning and Software Engineering. He earned his PhD in Informatics Engineering from the University of Coimbra in 2019. Since then, he has been a member of the CMS group at the CISUC, where he remains actively involved.