EXPLORING DEEP LEARNING METHODOLOGIES FOR MUSIC EMOTION RECOGNITION

Pedro L. LOURO (pedrolouro@dei.uc.pt) (0000-0003-3201-6990)[1],
Hugo REDINHO (redinho@dei.uc.pt) (0009-0004-1547-2251)[1],
Ricardo MALHEIRO (rsmal@dei.uc.pt) (0000-0002-3010-2732)[1, 2],
Rui Pedro PAIVA (ruipedro@dei.uc.pt) (0000-0003-3215-3960)[1], and
Renato PANDA (panda@dei.uc.pt) (0000-0003-2539-5590)[1, 3]

[1]*CISUC, DEI*, **University of Coimbra**, Coimbra, LASI, Portugal
[2]*School of Technology and Management*, **Polytechnic Institute of Leiria**, Leiria, Portugal
[3]*Ci2 - Smart Cities Research Center*, **Polytechnic Institute of Tomar**, Tomar, Portugal

ABSTRACT

Classical machine learning techniques have dominated Music Emotion Recognition (MER). However, improvements have slowed down due to the complex and time-consuming task of handcrafting new emotionally relevant audio features. Deep Learning methods have recently gained popularity in the field because of their ability to automatically learn relevant features from spectral representations of songs, eliminating such necessity. Nonetheless, there are limitations, such as the need for large amounts of quality labeled data, a common problem in MER research. To understand the effectiveness of these techniques, a comparison study using various classical machine learning and deep learning methods was conducted. The results showed that using an ensemble of a Dense Neural Network and a Convolutional Neural Network architecture resulted in a state-of-the-art 80.20% F1-score, an improvement of around 5% considering the best baseline results, concluding that future research should take advantage of both paradigms, that is, conbining handcrafted features with feature learning.

1. INTRODUCTION

In the early stages of developing Music Emotion Recognition (MER) systems, the focus was mainly on classical machine learning (ML) techniques, which involved a significant amount of effort devoted to feature engineering [1, 2]. Music classification involves identifying gaps in dimensions such as melody, harmony, rhythm, dynamics, timbre, expressivity, texture, and form. Feature extraction algorithms capture these dimensions, and ML models are trained on them. However, current works mostly use low- and mid-level descriptors. Panda et al. [3] achieved 76% accuracy with a combination of novel emotionally relevant features based on audio analysis and a newly proposed dataset, the 4QAED dataset, surpassing the 69% accuracy plateau observed in the MIREX challenge held in 2007. The most pressing issue is the challenging and time-consuming process of designing such features, requiring expertise in signal processing, musicology, and ML to produce improvements.

As a way to avoid this process, deep learning (DL) has recently seen a rise in popularity due to its ability to automatically learn relevant features from raw input data. Recently, various DL methods have been applied to tackle MER, many of which employ convolutional (CNN) and recurrent (RNN) neural networks [4, 5]. Different approaches have been proposed for processing raw input data in audio applications. This includes end-to-end architectures [6, 7], transfer learning from larger datasets [8], and using embeddings extracted from pre-trained CNNs [9].

These techniques have limitations due to requiring large amounts of quality labeled data. Previous systems have applied audio transformations to increase the training set for chosen algorithms, but the impact of this approach for MER is not well known in comparison with other tasks such as genre recognition [10].

Neural networks lack interpretability due to their black-box nature, making it unclear which features are learned and extracted during training. In MIR, concerns have been raised about their ability to learn relevant information. However, studies such as the one by Choi et al. [11] showed that a 5-layer CNN can learn to extract features closely related to melody, harmony, percussion, and texture. In the same direction, Won et al. [12] demonstrated that a self-attention mechanism can learn relevant instrument, genre, and emotion detection information through heatmaps.

In this article, we have performed a comparative study of different classical ML and DL methodologies applied to MER. This study aims to understand these methods' effectiveness, considering the promising paths of DL-based approaches. To conduct this study, we have used the 4QAED dataset along with a recent expansion. We have explored various methodologies, including architectural improvements, audio augmentation techniques, alternative input data representations, and knowledge transfer from related tasks. Additionally, the expansion of the baseline dataset has allowed us to assess the impact of small dataset size increases on the classification accuracy of DL models.

This study produced several contributions, including a state-of-the-art F1-score of 80.20% achieved through an ensemble of Dense Neural Network (DNN) and CNN architecture while considering data augmentation. Additionally, we conducted a comprehensive comparison of various methodological enhancements for solving MER, as well as an analysis of the influence of dataset size and class balancing on classification performance. To encourage reproducible research, we provide the code used for the conducted experiments [1].

2. BACKGROUND

The connection between music and emotions has been a focus of research in music psychology. Emotion from a musical piece can be examined through expressed, perceived, and induced perspectives. Perceived emotion provides the highest level of objectivity and is the focus of most works in the literature [13].

There have been several proposals to represent the range of human emotions. These can be divided into categorical models, which cluster similar emotions together, such as Hevner's Adjective Circle [14], and dimensional models, which create a multi-dimensional plane with axes representing different biological systems theorized to process emotion in the human brain. The most widely accepted dimensional model is Russell's Circumplex Model [15], seen in Figure 1, according to the literature. Scholars have raised concerns about categorical and dimensional models. Categorical models do not reflect the continuous nature of emotions, while dimensional models are complex and require prior knowledge for accuracy [16].

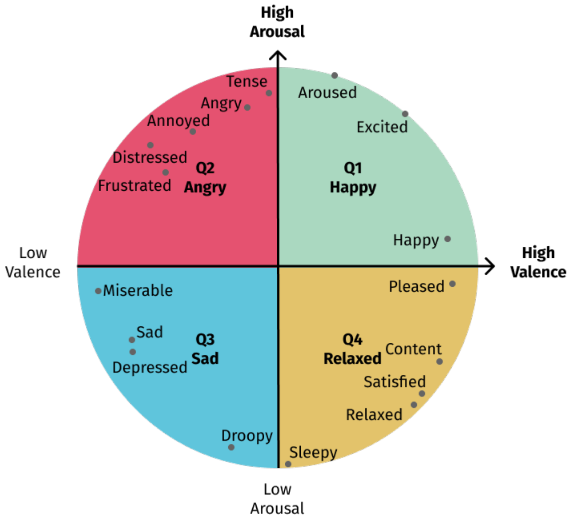

Figure 1. Russell's Circumplex Model. It is possible to represent emotions either as continuous values, which can be seen as individual points on a map, or as discrete labels that encompass a wider range of emotions.

In our team's research, Panda et al. [3] introduced the previously mentioned 4QAED dataset, which utilizes labels provided by experts from the AllMusic API. The labels were converted into A-V values [2], which correspond to the x- and y-axes of Russell's model, representing valence and arousal, respectively. The dataset takes a categorical approach by grouping all annotations into one of four quadrants rather than the continuous approach of the model. Please refer to the following section for a more detailed understanding of the dataset and its expansion.

3. METHODS

The methodologies explored for this work are presented in this section. We began by defining both ML and DL baseline methodologies in Section 3.1 and evaluating them on multiple datasets for comparison purposes.

The following section discusses the various methodologies that we explored and what motivated us to consider them. These approaches include improving the architecture by utilizing time-related information (Section 3.3), learning features from segments of entire samples (Section 3.4), obtaining alternative input representations through high-dimensional projections (Section 3.5), generating more training data through sample synthesis (Section 3.6), and utilizing learned information from related tasks (Section 3.7).

3.1 Baseline Architectures

We began our experiments by establishing a classic baseline using the state-of-the-art system by Panda et al, discussed previously, consisting on the top 100 standard and novel features, ranked using a feature selection algorithm, fed to a Support Vector Machine (SVM) classifier. For each dataset, we fine-tuned the hyperparameters of the SVM classifier using the same set of optimal features that were used in the original work.

Our team previously created a CNN architecture based on the research done by Choi et al. [17]. This architecture serves as the baseline for deep learning development. We modified the original architecture to process the extracted features on a small DNN that predicts one of the four quadrants from Russell's model, instead of outputting a binary vector. To prevent overfitting, an early stopping strategy was implemented, halting training when the training set accuracy reached a value equal to or greater than 90%, as found from previous experimentation. Unless explicitly stated otherwise, these points are the default configuration for the remaining approaches described in this section.

Regarding the optimal hyperparameters found for each methodology using a neural network as classifier, please refer to Table 1.

3.2 Explored Methodologies

We started off by examining the latest DL-based methodologies proposed for MER, focusing solely on systems that improve static emotion in music using only audio.

[1] https://github.com/g0ld3nl34f/exploring-dl-for-mer

[2] https://tivo.stoplight.io/docs/music-metadata-api/ZG9jOjQ3NjAxNTk-introduction

Table 1. Input sizes and best hyperparameters found for DL-based baseline and explored methodologies.

Methodology	Input		Best Hyperparameters			
	Type	Size	Epochs	BS	Optimizer	LR
DL Baseline	Mel-spectrogram	942×128 (≈30s)	200	150	SGD	1e-2
Baseline with GRU	Mel-spectrogram	942×128 (≈30s)	200	150	SGD	1e-2
CRNN	Mel-spectrogram	942×128 (≈30s)	200	50	SGD	1e-3
Hybrid	Mel-spectrogram and	942×128 (≈30s)	100	300	SGD	1e-2
Augmented	Handcrafted Features	1714	100	300	SGD	1e-2
ShortChunk CNN	Mel-spectrogram	116×128 (≈3.5s)	100	50	SGD	1e-3
Sample CNN	Waveform	59049 (≈3.5s)	150	50	SGD	1e-3
Baseline + TFM	Mel-spectrogram	942×128 (≈30s)	200	150	SGD	1e-2
Baseline + SB	Mel-spectrogram	942×128 (≈30s)	200	150	SGD	1e-2
Baseline + RG	Mel-spectrogram	942×128 (≈30s)	200	150	SGD	1e-2
Artists CNN	Mel-spectrogram	129×128* (≈3s)	200	100	Adam	1e-2
CRNN MTAT	Mel-spectrogram	942×128 (≈30s)	200	16	Custom**	Custom**

*A sample rate of 22.05kHz was used per the original implementation instead of 16kHz for the remaining methodologies.
**See Section 3.7 for details.

An important resource for our experiments is the work by Won et al. [18]. Here, a set of architectures from previous works on automatic music tagging are gathered and tested against each other. These comprise simple CNN-based architectures, end-to-end approaches, and even theory-motivated ones, all available in a GitHub repository [3].

3.3 Architecture Improvements

We enhanced our DL baseline's architecture by incorporating two Gated Recurrent Units (GRU) [19] to enable learning time-domain-specific features. This was done as a starting point to improve the overall architecture. We also experimented with an implementation of the CRNN architecture, which was adapted from the previously mentioned work. This way it is possible to assess the DL baseline's feature learning portion ability to preserve information related with time.

A simple ensemble of a baseline CNN and a DNN, both previously pre-trained, performed remarkably well, referred as Hybrid Augmented form herein. The DNN portion was fed with all 1714 features found to be relevant in the same work used as a basis for the classical baseline. The fused information is then post-processed by a smaller DNN. This approach combined the information extracted from both paradigms to enhance the overall classification, with the addition of synthetic samples for pre-training the CNN portion. The complete architecture can be seen in Figure 2.

3.4 Segment-level Approaches

In our previous work, we utilized the complete 30-second samples accessible on 4QAED as the model's input. However, humans can easily recognize emotions in smaller samples. Considering the small size of our datasets, breaking down these samples into smaller segments can benefit us by increasing the number of training examples, which is an indirect way of data augmentation. A straightforward approach that follows this idea is introduced in [18] as ShortChunk CNN. During the model's training phase, each segment was treated as an individual sample, while for testing, all the predictions related to the segments of a sample were combined to obtain the final prediction, which is known as a many-to-one approach.

Previous deep learning works commonly used convolutional layers to extract features from spectral representations, which require specific parameters. However, the ideal parameters are not architecture-independent, requiring optimization should any of the layers change. To avoid this issue, Lee et al. [6] propose an end-to-end architecture, the Sample CNN. They suggest working directly with the raw audio signal and using a sequence of one-dimensional convolutional blocks, similar to the two-dimensional variant, and processing the output through a dense layer. It is important to note that these models were originally designed to output one of a set of labels, which varied depending on the dataset used. They were later translated from PyTorch to TensorFlow and reworked to output categorical labels.

3.5 Data Representations

As mentioned earlier, using Mel-spectrograms might not be the best approach for training a machine learning model to classify emotions. Embeddings, popular in Natural Language Processing (NLP) due to providing a smaller, more efficient representation of the location of words in sentences in a lower-dimensional space, are considered as an alternative to spectral representations. Recently, Koh et al. [9] applied this idea to audio by using the OpenL3 deep audio embedding library [4] and training classical ML techniques classifier on its output. The embeddings are derived from a Mel-spectrogram representation, resulting in a feature matrix of size 298×512.

The study's baseline dataset yielded a 72% F1-score with the Random Forest (RF) classifier from the scikit-learn library [5], which is almost similar to the classical baseline. The experiment was further extended to the baseline

[3] https://github.com/minzwon/sota-music-tagging-models

[4] https://github.com/marl/openl3

[5] https://scikit-learn.org/stable/

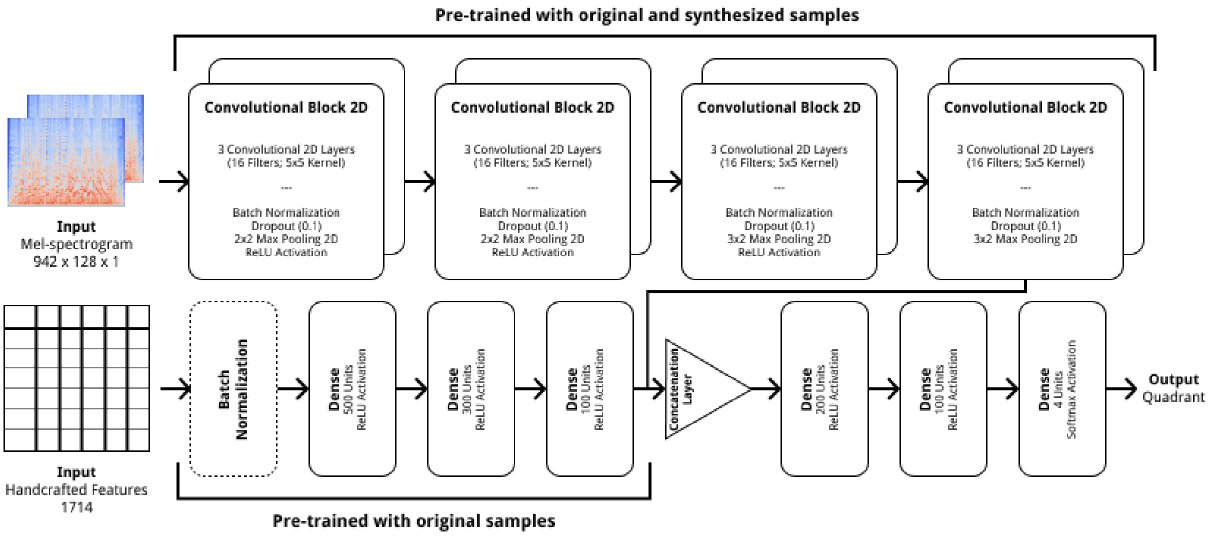

Figure 2. Hybrid Augmented architecture. Both the CNN and feature processing DNN are previously pre-trained on training data, with the addition of synthesized samples for the latter. Information from the aforementioned networks are concatenated and further processed by a smaller DNN.

dataset extension, and the embeddings generated by the autoencoder used in the DeepSMOTE approach, described in the next section, were also tested for comparison purposes.

3.6 Data Augmentation

We have delved deeper into both classical and deep learning (DL) approaches for augmenting data. In the classical approach, we have applied various audio augmentation techniques directly to the audio signal of a sample. These techniques randomly increase or decrease a factor associated with the transformation, such as time shifting (shifts start or end by five seconds), pitch shifting (increasing or decreasing pitch by two semitones), time stretching (speeding up or slowing down by 50%), and power shifting (increasing or decreasing amplitude by 10 dB), transformations used to obtain the synthesized sample for the Hybrid Augmented methodology. More of these techniques were experimented using the audiomentations library [6] , including:

- Time-Frequency Masking (TFM), popular in the field of SER, which applies a mask over a portion of the time- and frequency-domain [20];
- Seven-Band Parametric Equalization (SB), applying a seven-filter pass on the sample, changing its timbre in the process;
- Random Gain (RG), randomly increasing or decreasing the loudness of a sample.

A factor is chosen randomly from a set of predefined intervals for each transformation. For instance, the RG predefined interval lies between [-12.0, 12.0] dB. These intervals have been kept the same as the default values in the library.

The use of Generative Adversarial Networks (GANs) [21] for Deep Learning techniques has been tested by our team with less than satisfactory results. Beyond the overly complex process of training a GAN, the lack of constraints when sampling the learned space from the data can result in noisy and emotionally ambiguous samples.

We considered using SMOTE [22] to generate samples with some constraints but found that applying it to raw audio signal produced noisy samples due to the high dimensionality of the audio signal. Thus, we used the autoencoder applied for training the above-mentioned GAN to reduce the dimensionality of the sample, leading to a significant decrease in the number of values from around 482k to 60k, similar to the DeepSMOTE approach [23]. To the best of our knowledge, this was the first time this technique was applied to music samples.

It can be challenging to determine the best SMOTE implementation to use due to the many alternatives available. Kovács' article on SMOTE variants [24], along with the accompanying repository [7] , provides a comprehensive resource for making a decision. The article compares over 80 variants, but we focused on the most widely used ones, namely SMOTE, BorderlineSMOTE, and Adasyn. After conducting preliminary tests, BorderlineSMOTE, particularly the Borderline_SMOTE2 implementation, was found to be the most suitable option.

3.7 Transfer Learning

One way to address the challenge of dealing with a small dataset is to leverage the knowledge learned from a domain with a larger data corpus. This involves transferring the learned weights from a pre-trained network to a new network with a different task. Our team has previously experimented with this approach by utilizing the learned weights of a network trained for genre recognition to improve emotion recognition. Instead of using a larger dataset, we took

[6] https://github.com/iver56/audiomentations

[7] https://github.com/analyticalmindsltd/smote_variants

Table 2. Datasets used for evaluation with respective sample distribution.

Dataset	Q1	Q2	Q3	Q4	Total
Original-4QAED	225	225	225	225	900
New-4QAED C	434	440	397	358	1629
New-4QAED B	343	343	343	343	1372

advantage of the learned information about genres, which are closely linked to specific emotion quadrants. For instance, heavy metal is typically associated with Q2, while reggae is connected to Q4 [25].

In the same direction, we conducted experiments to transfer the knowledge from the models introduced by Park et al. [8] in their article on artist classification. We opted for the simpler model in our work, which encompasses a series of 5 one-dimensional convolutional blocks, a global average pooling layer, and a dense layer that produces a 256-value vector. To carry out the experiment, we retrieved the model's weights from the article's accompanying repository [8] , loaded and froze them, and substituted the last layer with a dense layer that outputs to one of the quadrants.

We also evaluated the impact of utilizing data from larger music datasets using the available weights for the CRNN model on Won's repository trained on three datasets, but for this article we focus on MagnaTagATune (MTAT) [26] as it provided the best results. It should be noted that the CRNN model was trained for multi-label classification. The optimization process for the model was adaptive, which means that it changed during certain epochs of training. It started with Adam optimizer and a learning rate of 1e-3 until epoch 80, then shifted to the SGD optimizer with a learning rate of 1e-4, which decreased to 1e-5 at epoch 100, and finally to 1e-6 at epoch 120. According to the authors, this approach leads to a more stable training process and better results at 200 epochs. The model was optimized with a batch size of 16, which minimized the necessary computational resources for the training process.

4. EVALUATION DETAILS

As previously stated in Section 2, our team utilized the 4QAED dataset [9] for experimentation. The dataset, which consists of 900 samples, was evenly distributed across Russell's four quadrants. Each quadrant corresponds to a specific set of emotions: Q1 for happiness and excitement; Q2 for anger and frustration; Q3 for sadness and melancholy; and Q4 for serenity and contentment. The dataset provides 30-second excerpts of complete songs and two sets of emotionally relevant handcrafted features as data sources. The first set contains 1714 features found to be relevant for emotion recognition, while the second set contains the top 100 features obtained after feature selection. The dataset also provides categorical labels for one of the four quadrants as targets.

An expanded version of the dataset is also considered, increasing the number of available samples from 900 to 1629. The datasets are referred as Original-4QAED and New-4QAED, respectively. Table 2 shows the quadrant distribution of the datasets, including two variations of the New-4QAED dataset, a complete (C) version and a balanced subset (B) that has 1372 samples. The balanced subset takes into account the distribution of genre in each quadrant to prevent any potential bias.

4.1 Data Preprocessing

The input data for these methodologies were obtained through Mel-spectrogram representations generated using the Python library librosa [10] with default parameters. However, the sample rate was set to 16 kHz after experimenting with different values. This was done to reduce the computational complexity of the model and reduce the necessary resources for training and inference.

It is worth noting that DL-based architectures are robust to a lack of information related to lower sample rates when compared with higher rates, as has been observed in other studies [27].

4.2 Experimental Setup

The experiments were conducted on a server that was shared among the team. The server had two Intel Xeon Silver 4214 CPUs, which had a total of 48 cores and ran at a clock speed of 2.20GHz, and three NVIDIA Quadro P500 GPUs with 16GB of dedicated memory. The latter were necessary to develop and evaluate each network within a reasonable time frame. However, due to high demand, we also used Google Collaborator [11] during the evaluation process. Depending on availability, this platform offered a similar GPU and either an NVIDIA PCIE or an NVIDIA T4, both with 16GB of dedicated memory.

Almost all DL-based approaches were developed using the TensorFlow Python library [12] . This library enables the creation and optimization of intricate models in a straightforward and efficient manner. Additionally, in Section 3.7, we discussed how pre-trained CRNN models' provided weights were utilized with the PyTorch library [13] .

5. RESULTS AND DISCUSSION

In this section, we begin by describing the metrics considered and the evaluation strategy for the conducted experiments. The outcomes for each approach and datasets taken into consideration are presented in accordance with the broad categories discussed in Section 3.2. Refer to Table 3 for the comprehensive summary of the results.

In order to evaluate the performance of a classification model, three common metrics are Precision, Recall, and F1-score. Precision measures the proportion of true positive predictions within all positive predictions made by the model. Recall measures the proportion of true positive predictions within all actual positive samples. F1-score is a combined metric that takes into account both Precision

[8] https://github.com/jongpillee/ismir2018-artist

[9] Available at: http://mir.dei.uc.pt/resources/MER_audio_taffc_dataset.zip

[10] https://github.com/librosa/librosa

[11] https://colab.research.google.com/

[12] https://github.com/tensorflow/tensorflow

[13] https://pytorch.org/

and Recall, and is defined as the harmonic mean between the two. These metrics can be easily calculated using the scikit-learn library [14], which is widely used for machine learning tasks.

To obtain the metrics, we first optimized the relevant hyperparameters on Original-4QAED. We used a grid search strategy to experiment with a set of possible values to serve as a baseline for performance on New-4QAED. To ensure a fair comparison, we utilized the same parameters.

To ensure reliable results and handle small dataset sizes, a 10-fold and 10-repetition stratified cross-validation strategy is employed. This results in a total of 100 different train-test splits. In each repetition, the original dataset is randomly divided into ten portions while maintaining an equal distribution of quadrants. Nine of these portions are used for training, while the remaining one is used for testing. The portion held-out for testing changes for each train-test split, resulting in ten different combinations for each repetition.

The values of the hyperparameters that were tested using this approach varied depending on the methodology employed. In cases where the methodology was based on the baseline CNN, adjacent values were tested to accommodate any potential variations in the data. When the reported results did not have accompanying optimal hyperparameters values, they were used, and if not, the baseline CNN values were utilized as a fallback. However, regardless of the approach used, it is possible to draw meaningful conclusions concerning the effects of different dataset sizes and quadrant distributions.

Regarding improvements, increasing the dataset size had a positive impact on the performance of the baseline CNN with GRU and CRNN methodologies. The F1-score improved from 60.07% to 61.99% and 60.35% to 63.33%, respectively, when comparing the Original- to New-4QAED C datasets. This improvement was better compared to the DL Baseline. Additionally, the optimization phase was more stable with the increased dataset size. However, there was a slight decrease in performance when the balanced variations of the latter were applied, highlighting the importance of dataset size for optimal results.

Still, in terms of architecture improvements, it was found that the Hybrid Augmented methodology produced the best results. This particular methodology achieved an F1-score of 80.20% on the balanced subset of New-4QAED. The outcome was heavily influenced by both the size and quadrant distribution, with the latter potentially being linked to the biased nature of the DNN, much like traditional ML techniques.

It was also observed that using Time-Frequency Masking, Seven-Band Parametric Equalization, and Random Gain improved the results. These techniques achieved the best results with an increase of around 1.5% F1-score compared to the DL Baseline on Original-4QAED. They consistently performed better on New-4QAED as well. These findings suggest that more research is needed on data augmentation in MER, as most of the existing techniques are borrowed from other fields. As discussed in Section 1, the emotional impact of such techniques on the resulting samples is not yet fully understood.

All of the segment-level methodologies were found to perform similarly or worse than the DL Baseline. The poor performance may be attributed to the smaller size of the datasets used for training, compared to the ones used in the original proposal of the architectures. This means that the amount of available training data was limited, which may have hindered the performance of the models. Additionally, splitting the samples into smaller segments could have introduced more variability in the data, making it harder for the models to learn relevant features for distinguishing each quadrant. Further investigation is needed to verify this hypothesis.

The remaining methodologies related to knowledge transfer and data representation did not perform well compared to the baseline. The former significantly underperformed compared with the original implementations. This may imply that such approaches are not useful for emotion recognition, especially in the case of methodologies initially developed for multi-label classification, where larger datasets are used. The inadequate performance of these methodologies may be due to their significant differences from the learned features required for the task. As a result, crucial information for emotion recognition may be lost due to the abundance of irrelevant features. In the future, it may be worth experimenting with an ensemble of models trained for emotion recognition and other related tasks.

Finally, regarding embedding-based methodologies, we were not able to replicate the results presented for the OpenL3 embeddings on Original-4QAED, which was a 72% F1-score, reaching at most 55.70%. It seems that the unclear data splitting, where the authors followed an 80/10/10 train-validation-test data splitting instead of 10-fold cross-validation, might have contributed to this. Additionally, the original approach did not disclose the parameters used for creating the RF classifier, so we assumed default parameters from the scikit-learn implementation. To maintain consistency, we applied the usual method for cross-validation. As for the autoencoder embeddings, these exhibited better performance on New-4QAED overall than OpenL3 embeddings. This observation suggests that OpenL3 embeddings may not be the most appropriate choice for MER.

Furthermore, the DeepSMOTE-based augmentation did not show any significant improvement over the DL baseline. The reason behind this lack of improvement could be the high dimensional embedding space, which provides little variability when sampled as compared to the original samples. To overcome this, it might be helpful to reduce the input data size by using the segments of the whole samples. This should decrease the embedding space dimension and provide more relevant synthesized samples.

6. CONCLUSION AND FUTURE DIRECTIONS

The study aimed to compare the effectiveness of ML and DL methodologies for static emotion recognition in music with audio with datasets of different sizes. The fo-

[14] https://scikit-learn.org/stable/

Table 3. Datasets used for evaluation with respective sample distribution.

Methods	**Original-4QAED**			**New-4QAED C**			**New-4QAED B**		
	P	R	F1	P	R	F1	P	R	F1
SVM Baseline	75.63%	76.03%	75.59%	69.92%	70.26%	69.79%	70.03%	70.05%	69.82%
DL Baseline	61.60%	61.21%	60.62%	62.46%	63.99%	61.66%	61.39%	63.42%	60.28%
Baseline with GRU	61.58%	61.01%	60.07%	62.29%	62.46%	61.99%	60.69%	60.01%	58.85%
CRNN	**65.14%**	**65.07%**	**64.63%**	**64.20%**	**64.03%**	**64.09%**	**63.31%**	**63.34%**	**62.54%**
Hybrid Augmented	**67.81%**	**68.08%**	**68.04%**	**68.15%**	**68.14%**	**67.85%**	***80.56%***	***80.50%***	***80.24%***
ShortChunk CNN	**64.66%**	61.48%	60.61%	**64.07%**	62.13%	61.84%	60.23%	59.19%	57.07%
Sample CNN	**62.64%**	61.26%	60.92%	**65.17%**	62.62%	60.78%	**62.43%**	56.70%	54.46%
OpenL3	55.67%	56.75%	55.70%	53.92%	54.49%	53.62%	53.03%	53.18%	52.85%
Autoencoder	50.63%	50.40%	50.18%	53.78%	55.45%	53.56%	53.56%	54.76%	53.69%
Baseline + TFM	**63.05%**	**62.75%**	**62.03%**	62.51%	62.17%	61.82%	62.33%	61.85%	61.39%
Baseline + SB	**63.38%**	**62.79%**	**62.12%**	62.54%	62.16%	61.73%	62.13%	61.71%	61.01%
Baseline + RG	**63.37%**	**63.13%**	**62.24%**	63.02%	62.80%	62.08%	62.35%	62.10%	61.36%
Baseline + DeepSMOTE	61.91%	61.61%	60.70%	62.40%	62.02%	61.47%	61.62%	61.62%	60.48%
Artists CNN	51.95%	53.93%	50.85%	51.81%	53.29%	50.27%	51.56%	52.43%	50.22%
CRNN Pre-trained MTAT	51.93%	51.71%	50.21%	52.97%	53.72%	51.70%	52.16%	52.50%	51.44%

cus was more on DL to address the semantic gap found in traditional ML approaches by exploring various approaches, such as improving existing architectures, using segment-level models, applying data augmentation, performing knowledge transfer, and using different data representations as input.

The proposed Hybrid Augmented achieved the best result overall on the New-4QAED balanced dataset with an 80.20% F1 Score. This methodology is an ensemble of a CNN trained with additional synthesized samples and a DNN, which use Mel-spectrogram representations and previously extracted features from each song as input, respectively. Another notable improvement was observed by using CRNN on the increased size of the New-4QAED datasets. This approach outperformed the DL Baseline by around 2% on the complete set, outperforming methodologies that applied classical data augmentation techniques.

When comparing different methodologies, it was found that classical audio augmentation techniques and architectural improvements effectively improved performance. However, segment-level architectures, knowledge transfer from related tasks, and embedding-based input representations did not show as much improvement. Nevertheless, as discussed earlier, there is room for improvement in these areas. The results also revealed that, in most cases, the size of the dataset has a greater impact on classification performance than class balance.

The findings suggest that there is a need for continued research aimed at developing new classical features and enhancing DL architectures for improved performance. Additionally, exploring data augmentation techniques specifically for MER could be a promising approach to fully leverage DL models' ability to automatically extract relevant features. As more training data becomes available, future DL architectures should include an RNN component to capture time-domain-specific features. Finally, using various spectral representations as inputs is an exciting area for further research, as demonstrated by early experimental studies. However, it is important to address the unstable nature of these approaches before they can be fully utilized.

Acknowledgments

This work is funded by FCT - Foundation for Science and Technology, I.P., within the scope of the projects: MERGE - 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.

7. REFERENCES

[1] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 375–376.

[2] 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*, vol. 14, pp. 5–18, 2006.

[3] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, pp. 614–626, 2020.

[4] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proceedings of the 2017 International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2392–2396.

[5] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and

recurrent neural networks for music emotion recognition," in *Proceedings of the 14th Sound and Music Computing Conference*, vol. 14, 2017, pp. 208–213.

[6] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in *Proceedings of the 14th Sound and Music Computing Conference*, vol. 14, 2017, pp. 220–226.

[7] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018, pp. 637–644.

[8] J. Park, J. Lee, J. Park, J.-W. Ha, and J. Nam, "Representation learning of music using artist labels," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018, pp. 717–724.

[9] E. Koh and S. Dubnov, "Comparison and analysis of deep audio embeddings for music emotion recognition," 2021. [Online]. Available: https://arxiv.org/abs/2104.06517

[10] R. Mignot and G. Peeters, "An analysis of the effect of data augmentation methods: Experiments for a musical genre classification task," *Transactions of the International Society for Music Information Retrieval*, vol. 2, pp. 97–110, 2019.

[11] K. Choi, G. Fazekas, and M. Sandler, "Explaining deep convolutional neural networks on music classification," 2016. [Online]. Available: https://arxiv.org/abs/1607.02444

[12] M. Won, S. Chun, and X. Serra, "Toward interpretable music tagging with self-attention," 2019. [Online]. Available: https://arxiv.org/abs/1906.04972

[13] A. Pannese, M.-A. Rappaz, and D. Grandjean, "Metaphor and music emotion: Ancient views and future directions," *Consciousness and Cognition*, vol. 44, pp. 61–71, 2016.

[14] K. Hevner, "Experimental studies of the elements of expression in music," *The American Journal of Psychology*, vol. 48, pp. 246–268, 1936.

[15] 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, pp. 715–734, 2005.

[16] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, pp. 18–49, 2011.

[17] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, 2016, pp. 805–811.

[18] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of cnn-based automatic music tagging models," in *Proceedings of the 17th Sound and Music Computing Conference*, 2020, pp. 331–337.

[19] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014. [Online]. Available: https://arxiv.org/abs/1409.1259

[20] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, 2019, pp. 2613–2617.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[23] D. Dablain, B. Krawczyk, and N. V. Chawla, "Deepsmote: Fusing deep learning and smote for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 16, pp. 1–15, 2022.

[24] G. Kovacs, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," *Applied Soft Computing*, vol. 83, p. 105662, 2019.

[25] D. Griffiths, S. Cunningham, J. Weinel, and R. Picking, "A multi-genre model for music emotion recognition using linear regressors," *Journal of New Music Research*, vol. 50, pp. 355–372, 2021.

[26] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, vol. 50, 2009, pp. 387–392.

[27] K. Pyrovolakis, P. Tzouveli, and G. Stamou, "Multimodal song mood detection with deep learning," *Sensors*, vol. 22, pp. 387–392, 2009.