RESEARCH-ARTICLE

Joint Learning of Emotions in Music and Generalized Sounds

FEDERICO SIMONETTA, Gran Sasso Science Institute, L'Aquila, AQ, Italy

FRANCESCA CERTO, University of Milan, Milan, MI, Italy

STAVROS NTALAMPIRAS, University of Milan, Milan, MI, Italy

Open Access Support provided by:

University of Milan

Gran Sasso Science Institute

Published: 18 September 2024

Citation in BibTeX format

AM '24: Audio Mostly 2024 - Explorations in Sonic Cultures
September 18 - 20, 2024
Milan, Italy

AM '24: Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures (September 2024)
https://doi.org/10.1145/3678299.3678328
ISBN: 9798400709685

Joint Learning of Emotions in Music and Generalized Sounds

Federico Simonetta
GSSI - Gran Sasso Science Institute
L'Aquila, Italy
federico.simonetta@gssi.it

Francesca Certo
Stavros Ntalampiras
francesca.certo@studenti.unimi.it
stavros.ntalampiras@unimi.it
University of Milan
Milan, Italy

ABSTRACT

In this study, we aim to determine if generalized sounds and music can share a common emotional space, improving predictions of emotion in terms of arousal and valence. We propose the use of multiple datasets as a multi-domain learning technique. Our approach involves creating a common space encompassing features that characterize both generalized sounds and music, as they can evoke emotions in a similar manner. To achieve this, we utilized two publicly available datasets, namely IADS-E and PMEmo, following a standardized experimental protocol. We employed a wide variety of features that capture diverse aspects of the audio structure including key parameters of spectrum, energy, and voicing. Subsequently, we performed joint learning on the common feature space, leveraging heterogeneous model architectures. Interestingly, this synergistic scheme outperforms the state-of-the-art in both sound and music emotion prediction. The code enabling full replication of the presented experimental pipeline is available at https://github.com/LIMUNIMI/MusicSoundEmotions.

CCS CONCEPTS

• **Hardware** → *Digital signal processing*; • **Applied computing** → *Sound and music computing*; • **Computing methodologies** → *Supervised learning by classification.*

KEYWORDS

music, emotions, generalized sounds, affective computing, automl

ACM Reference Format:
Federico Simonetta, Francesca Certo, and Stavros Ntalampiras. 2024. Joint Learning of Emotions in Music and Generalized Sounds. In *Audio Mostly 2024 - Explorations in Sonic Cultures (AM '24), September 18–20, 2024, Milan, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3678299.3678328

1 INTRODUCTION

Emotions have always played a fundamental role in human lives, and they are currently receiving increasing attention in the technological field [9]. Although emotions and computer science have traditionally been viewed as two distinct concepts due to the lack of consciousness in computers, which prevents them from experiencing emotions, numerous studies have been conducted over the years to demonstrate computers' ability to identify people's moods and emotions.

The present research aims to determine whether generalized sounds and music can share a common emotional space. The study proposes a novel multi-domain learning approach that harnesses the power of affective computing. Utilizing datasets representing both music and general sounds, which are capable of eliciting comparable emotional responses, the study creates a shared feature space for emotion prediction. Traditionally distinct audio domains converge in this research, suggesting that a unified model can offer enhanced performance in interpreting emotional responses to a broad spectrum of auditory stimuli.

Affective Computing utilizes two primary emotion representation frameworks: categorical (e.g., happiness, anger) and dimensional (e.g., arousal, valence). In this study, we will specifically focus on datasets that provide annotations of perceived emotions in the continuous space of valence and arousal.

The study of computational techniques for analyzing and recognizing emotions in sounds is known as Audio Emotion Recognition (AER), which is a subfield of Affective Computing [9]. While speech and music have been extensively studied in the literature, recent research has also explored general sound events that may impact human emotional states.

Music can convey emotions through melody and lyrics, thus affecting the emotional state of listeners. Music Emotion Recognition (MER) systems have been developed for various applications, including medical applications to improve patients' physical and mental healthand music players to recommend songs based on the user's mood. Zhang et al. [13] introduced the PMEmo dataset, extensively studied for music emotion recognition (MER) using multimodal or audio-only approaches. Among the various works about MER, three publications used PMEmo with a strong and accurate validation pipeline. First, De Berardinis et al. [5] introduced EmoMucs, a computational model that considers the role of different musical voices in predicting emotions induced by music. Second, Chowdhury et al. [3] proposed a method to trace music emotion predictions back to sound sources and intuitive perceptual qualities. Third, Huang et al. [7] proposed an end-to-end attention-based deep feature fusion approach for MER.

Despite the importance of sound events in individuals' daily lives, research on emotion prediction of sound events has received less attention when compared to speech and music. Previous works investigated the analogies between music, speech, and sounds in the context of emotion recognition. Weninger et al. [11], Ntalampiras [8], and Coutinho et al. [4] all studied various techniques, such

AM '24, September 18–20, 2024, Milan, Italy

ACM ISBN 979-8-4007-0968-5/24/09
https://doi.org/10.1145/3678299.3678328

as analyzing analogies between speech, music, and sound events, constructing shared emotional spaces, and employing transfer learning to improve the prediction of emotion in music and speech. In a seminal work, Bradley and Lang [2] developed the widely used International Affective Digitized Sounds (IADS) dataset, which has been extended [12] and utilized in various studies. Abri et al. [1] developed machine and deep learning models to predict the emotions associated with certain sounds, and compared the accuracy of those predictions using IADS-E based on a well-defined validation strategy.

This study focuses on acoustic stimuli in the form of environmental sounds, noises, and music. We investigate the ability of regression models to predict dimensional emotions while utilizing different types of sounds. Specifically, we analyze two types of sounds: *music* and *environmental sounds*. Our experimental results demonstrate that training models with both music and generic sounds lead to more robust models and more accurate emotion predictions for both types of sounds. The expectation is that, by capturing intricate emotional elements across diverse sound types, the approach outlined in this research stands to contribute appreciably to the practical and theoretical advancements in affective computing as it pertains to the world of the Internet of Sounds.

The primary contributions of this study are as follows: (i) a novel multi-modal learning strategy for Audio Emotion Recognition (AER) models combining two different types of sounds, (ii) new models that surpass the state-of-the-art in emotion recognition for both music and environmental sounds, and (iii) an accurate analysis of the impact of the proposed augmentation strategy on the two types of sounds.

The remainder of this paper is organised as follows: section 2 explains the proposed methodology including the datasets, feature extraction and modeling process. Section 3 presents the experimental set-up and analysed the obtained resutls, while section 4 summarizes the main findings.

2 METHODOLOGY

The approach in this study analyzes the efficacy of various regression models in predicting emotions evoked by general sound events and music. The models were trained using both single and combined data sets.

Specifically, this study aims to showcase the efficacy of combining two distinct types of sounds in enhancing emotion prediction accuracy. This finding underscores the presence of certain audio characteristics that elicit comparable emotional responses in individuals, regardless of the domain of sound. Moreover, the proposed approach capitalizes on the availability of diverse datasets to offer a straightforward, yet potent augmentation strategy.

The overall pipeline involves clustering the valence-arousal labels and sub-sampling the datasets based on the distribution of the samples across the clusters. Then, 5-fold cross-validation is accurately used to compare the impact of merging different data domains on the accuracy of AutoML pipelines. The general pipeline is represented in Fig. 2.

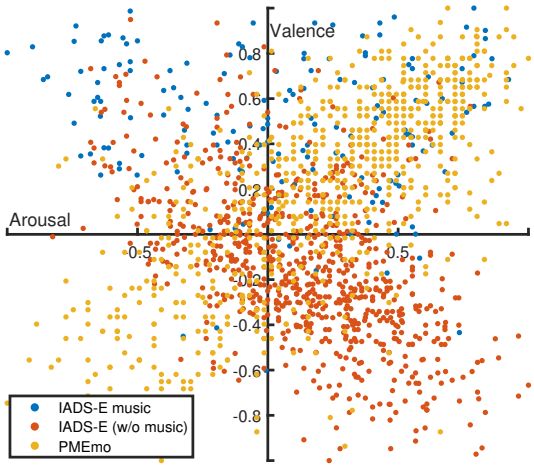

Figure 1: Distribution of the ratings in both datasets on the valence-arousal plane.

2.1 Data sets

This work investigates the behavior of models trained on specific sound types when applied to different ones. We aim to determine if two classes of sounds can share a common emotional space, improving predictions of emotion in terms of arousal and valence. We used the IADS-E and PMEmo datasets to analyze the emotional space of sound events and music, respectively.

IADS-E [12] expands the existing auditory affective sample database. It includes ratings from 207 participants using the SAM and basic-emotion rating scales on 935 sounds, including those from IADS-2 [2]. The results showed that emotions in sounds can be distinguished on affective rating scales, providing a larger corpus of natural, emotionally evocative auditory stimuli covering a wide range of categories. IADS-E also provides a semantic categorization based on labels assigned by 10 additional users, comprising 10 classes. While the IADS-E dataset constitutes a large set of sounds comprehensive of several categories that are rare in other existing datasets, it is limited by the asbence of speech sounds and utterances. Nevertheless, in our study, these categories are partially covered by music recordings.

The PMEmo dataset [13] consists of 794 popular music choruses, with 767 annotated with static emotion labels in terms of arousal and valence. It is designed for research on Music Emotion Recognition and Music Information Retrieval. Similar to the IADS-E dataset, the SAM technique was used to annotate emotional experiences along arousal and valence dimensions. In this study, we used static arousal and valence ratings to train regression models.

To standardize ratings and aid regression model learning, labels in IADS-E and PMEmo datasets were rescaled to $[-1, 1]$. Fig. 1 shows ratings of IADS-E music, IADS-E without music and PMEmo, illustrating their complementarity. No single dataset covers all quadrants of the valence-arousal plane, suggesting a holistic approach when learning from a combined dataset.

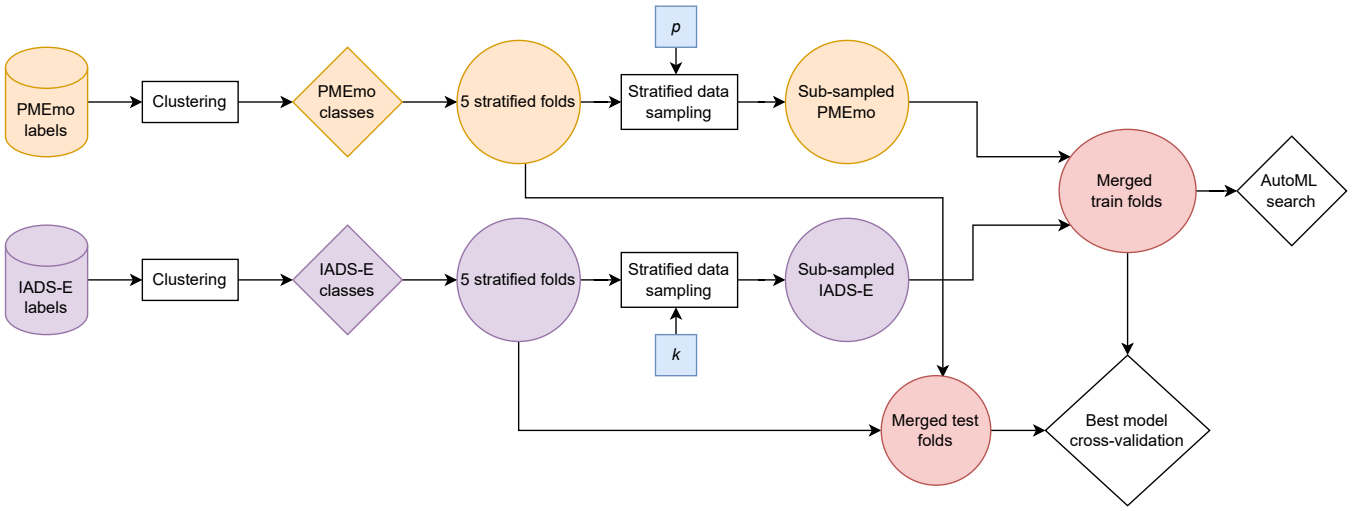

Figure 2: The overall pipeline: first, clustering is suitably used for applying stratified sampling; then, datasets are sub-sampled according to the parameters k and p; finally, the model learns the merged sub-populations and is tested on the original test folds.

2.2 Feature Extraction

Feature extraction plays a crucial role in Affective Computing as it transforms audio samples into numerical features that can be processed by machine learning algorithms without compromising the original information. This stage is paramount in determining the relevant features to be used as input in predictive models, ultimately influencing the accuracy of emotion prediction.

In this study, we utilized the `openSMILE` toolkit [10] to extract audio features from the IADS-E and PMEmo datasets. The `openSMILE` toolkit integrates features from both speech and music domains, providing a versatile software for feature extraction that is domain-independent.

For this purpose, we employed the ComParE 2013 configuration [10] packaged with version 3.0.1 of the `openSMILE` toolkit. The extracted feature set captures essential parameters of spectrum, energy, and voicing. Furthermore, ComPaRe applies various statistical functions to these low-level descriptors to capture diverse aspects of their temporal evolution. In total, 6375 static features are extracted for each sound sample using ComPaRe.

2.3 Model Selection and Validation on Combined Data Sets

To evaluate the impact of the proposed augmentation strategy on different models, we employed three approaches. Firstly, we utilized a linear model based on ElasticNet and a Support Vector Regression (SVR) model as implemented in the `sklearn` library. Secondly, we adopted a state-of-the-art AutoML method to evaluate the impact of various pre-processing and feature selection strategies on model performance. The AutoML also enabled us to compare a variety of regressors, both linear and non-linear, and to create an ensemble of the best performing models [6].

Both ElasticNet and SVR were trained using the principal components (PCs) of the features extracted from the IADS-E and PMEmo datasets. The number of PCs was determined by the cumulative explained variance ratio, which was itself optimized as a hyper-parameter of the model. The optimization of ElasticNet and SVR was performed using a successive halving grid search. Both the AutoML and the successive halving grid search were executed on a single machine with 32 GB of RAM and 12 cores. For the sake of space, the code documentation provides detailed information on the hyperparameter space for ElasticNet and SVR. Both AutoML and grid-search optimized the mean squared error (MSE).

To construct a mixed dataset, we merged the PMEmo and IADS-E datasets in proportions governed by two parameters, which dictate the respective dataset contributions to the mix. we define the combined dataset size using the equation:

$$k \times |IADS\text{-}E| + p \times |PMEmo|, \tag{1}$$

where $k, p \in [0, 1]$. Here, $|IADS\text{-}E|$ and $|PMEmo|$ represent the total number of samples in the PMEmo and IADS-E datasets, respectively. We constrain the parameters such that either k or p is set to 1, ensuring that one dataset is fully included while the inclusion of the other is adjustable. For a given pair (k, p), both AutoML and successive halving were utilized with 5-fold cross-validation. The hyper-parameter optimization was conducted to search for models that performed well *on average* for both PMEmo and IADS-E when trained on the augmented dataset.

We then selected the best-performing model from the hyper-parameter optimization and the best-performing ensemble from AutoML, and evaluated them using a similar 5-fold cross-validation procedure. Interestingly, this time we observed the performance on the validation fold of IADS-E and PMEmo separately.

The cross-validations were performed using a stratified sampling in order to address the potential issue of overfitting due to non-equitable representation of the dataset across the folds. Since, the target labels of the datasets at hand are continuous, we employed the Ward hierarchical method for clustering and then associated a

Table 1: Prediction results in terms of RMSE achieved by the considered model types when using a) PMEmo, b) IADS-E, and c) the fully augmented dataset.

Test set \ Train set			***IADS-E (no music)***	***PMEmo***	***IADS-E (no music) + PMEmo***
IADS-E (no music)	***Linear***	*Arousal*	**2.14e-01 ± 2.06e-02**	2.80e+04 ± 3.54e+04	1.69e+00 ± 4.08e+00
		Valence	3.06e-01 ± 1.98e-02	3.60e+04 ± 2.80e+04	**2.61e-01 ± 1.25e-02**
	SVM	*Arousal*	1.92e-01 ± 1.03e-02	2.66e-01 ± 1.96e-03	**1.25e-01 ± 1.15e-02**
		Valence	3.58e-01 ± 9.42e-03	3.62e-01 ± 4.82e-03	**2.41e-01 ± 2.87e-02**
	AutoML	*Arousal*	1.85e-01 ± 1.38e-02	2.45e-01 ± 1.17e-02	**1.04e-01 ± 5.93e-03**
		Valence	2.71e-01 ± 1.15e-02	3.24e-01 ± 2.25e-02	**2.53e-01 ± 1.31e-02**
PMEmo	***Linear***	*Arousal*	6.17e-01 ± 1.87e-01	2.92e-01 ± 8.19e-02	**2.29e-01 ± 2.37e-02**
		Valence	1.32e+00 ± 4.11e-01	4.45e-01 ± 4.72e-01	**2.41e-01 ± 2.87e-02**
	SVM	*Arousal*	3.62e-01 ± 1.46e-02	2.10e-01 ± 4.88e-03	**1.93e-01 ± 9.23e-03**
		Valence	4.14e-01 ± 1.41e-02	2.55e-01 ± 2.09e-02	**1.68e-01 ± 2.10e-02**
	AutoML	*Arousal*	3.20e-01 ± 9.28e-03	1.93e-01 ± 7.22e-03	**1.80e-01 ± 1.53e-02**
		Valence	3.89e-01 ± 3.16e-02	2.23e-01 ± 2.12e-02	**1.53e-01 ± 2.62e-02**

Table 2: Ablation study demonstrating how the proposed learning scheme compares with the state of the art.

Test set	**Method**	***RMSE***		R^2	
		Valence	*Arousal*	*Valence*	*Arousal*
IADS-E (with music)	*Abri et al., 2021*	.289 *	.195 *	.370	.563
	AutoML on IADS-E (with music)	.279	.193	.379	.566
	AutoML on IADS-E (with music) + PMEmo	**.249**	**.163**	**.508**	**.692**
PMEmo	*de Berardinis et al., 2020*	.232	.223	.481	.610
	Chowdhury et al., 2021	.310	.250	.400	.600
	Huang et al., 2022	.231	.216	.508	.655
	AutoML on PMEmo	.223	.193	.525	.727
	AutoML on IADS (no music) + PMEmo	.153	.180	.775	.762
	AutoML on IADS-E (with music) + PMEmo	**.152**	**.137**	**.780**	**.861**

* These values were re-normalized so that they are referred to labels in $[-1, 1]$

Note: all the results are referred to 5-fold cross-validation and labels normalized in $[-1, 1]$; except for the proposed model, for papers showing multiple configurations of the proposed models, the best results were picked, even if produced by different model configurations.

class to each cluster. The same approach was used for sub-sampling the datasets, i.e. when k and p are not set to 1. To ensure an adequate number of samples in each fold, we set the minimum number of samples in a cluster to 25 as our stopping criterion.

3 EXPERIMENTAL SET-UP AND RESULTS

We conducted experiments on ElasticNet, SVR, and AutoML, utilizing p and k values from the set $[0, 1]$. This approach only considered augmentation sets generated by the sum of PMEmo and IADS-E. The results of the best-performing models for each approach are presented in Table 1. Additionally, we evaluated the AutoML for various p and k values, as shown in Fig. 4.

Given that IADS-E includes 170 samples categorized as music, we have excluded them from the experiments, except when comparing with the state-of-the-art, so as to avoid confounding effects when analyzing the inter-relationships between generic sounds and music. This decision was made to ensure the validity and reliability of the obtained results.

We conducted an initial experiment to assess the performance of ElasticNet, SVR, and AutoML on the PMEmo and IADS-E datasets when using either one or both of the datasets. Hence, we executed the AutoML for 8 hours, while the SVM hyper-parameter optimization lasted approximately 4.5 hours and the ElasticNet model 3-5 minutes depending on the size of the train set. The results are presented in Fig. 3 and Table 1. Our findings indicate that all models demonstrate improved performance with the augmentation strategy, particularly AutoML and SVM models. However, the linear ElasticNet model may not fully capture the complexity of the shared space. Therefore, non-linear models might be more appropriate for leveraging this type of augmentation. In fact, ElasticNet does not benefit from the augmentation when predicting emotions conveyed by generalized sounds.

The objective of the second experiment was to assess the performance of AutoML, which was found to be the best-performing approach, when adding a portion of one dataset to another. The experiment ran for 4 hours, during which AutoML searched for the optimal hyper-parameters. The results are presented in Figure 4, where we observe that the most effective augmentation is achieved when both k and p are close to 1. Notably, adding a small amount of music to the training set significantly improves valence prediction for general sounds. Moreover, even without utilizing any sounds from the target dataset, arousal prediction achieves an R^2 value

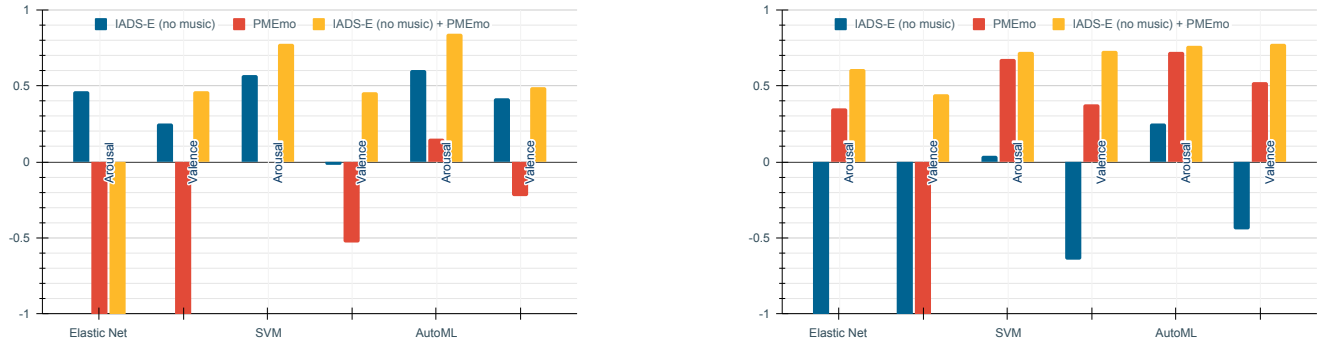

Figure 3: R^2 values according to the various training sets. The test set of the left plot was IADS-E (no music), while for right plot PMEmo. Negative values are truncated at -1.

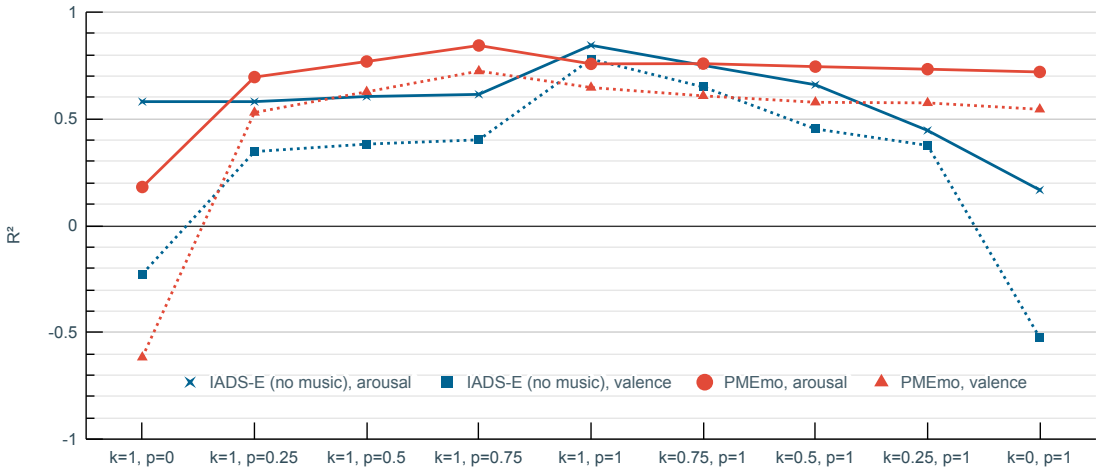

Figure 4: R^2 of the AutoML optimization when different augmentation ratios are used in the train set, i.e. for different values of k and p in the formula $k \times$ *IADS-E* $+ p \times$ *PMEmo*. Each line represents a different test set, while IADS-E dataset was used without the music samples.

greater than 0.15 demonstrating stronger feature sharing across music and general sounds for arousal than for valence.

We also compared the effects of using genuine data versus randomized data in AutoML training. Random labels for the IADS-E dataset were used to train AutoML models for 90 minutes. The performance of these models was contrasted with those trained on the actual IADS-E data, as shown in Figure 5. The models trained with random data struggled to find optimal configurations, especially with $k = 1$ and $p < 1$. In contrast, adding even a small amount of PMEmo data to the real IADS-E dataset improved the models' inference on PMEmo. This observation, underscored in Figure 5, illuminates the informational exchange between the IADS-E and PMEmo datasets, suggesting that leveraging commonalities across domains may facilitate the development of superior models.

The third experiment compared the proposed approach with the state-of-the-art. We used AutoML trained on three datasets: PMEmo, IADS-E, and the fully augmented dataset. We also repeated the experiment by including music samples in the IADS-E dataset. For the literature study, we considered all existing works with PMEmo and IADS-E. Standard cross-validation strategies and continuous valence-arousal labels were used. The results are presented in Table 2. We rescaled the RMSE values from [1] for a fair comparison. Our proposed method surpasses the state-of-the-art due to the model's effectiveness and the augmentation strategy. The ensemble discovered by AutoML is a combination of HistGradientBoostingRegressor models. However, the resulting solution is slow due to its complexity, resulting in more than 1 hour for the 5-fold cross-validation on a i9-9820X machine with 64GB of RAM. A smaller, more direct model like Support Vector Regressor (SVR) provides improved performance with faster cross-validation time on the same machine (about 30 seconds). Table 1 compares the SVM performance on the PMEmo dataset with the state-of-the-art results in Table 2.

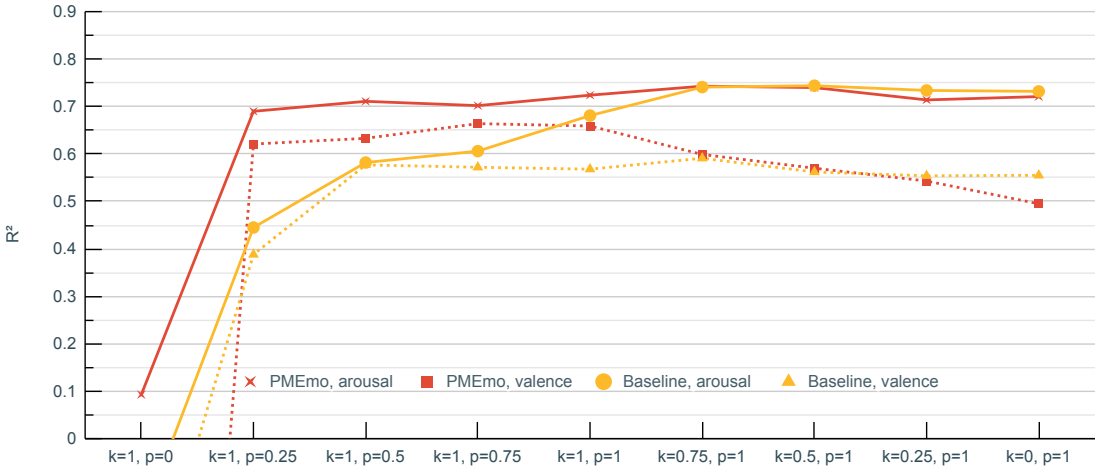

Figure 5: R^2 of the AutoML optimization when different augmentation ratios are used in the train set, i.e. for different values of k and p in the formula $k \times$ *IADS-E* $+ p \times$ *PMEmo*. Both lines represent R^2 scores obtained on the PMEmo validation folds. The baseline is obtained by adding a randomized version of IADS-E to the train set in which the labels were synthesized by uniform random sampling.

Overall, we argue that the proposed model provides more than satisfactory AER results, where learning from the augmented feature space formed by both music and generalized sounds offers a considerable improvement.

4 CONCLUSION

This paper proposes a new method to improve Audio and Music Emotion Recognition (AER and MER) models. Our approach creates a shared feature space for both types of sounds, allowing for more accurate emotion recognition models in both music and generalized sounds. We extensively validated our proposed strategy and found that non-linear models are necessary for proper modeling of the shared space. Interestingly, our experimental results show that arousal prediction benefits more than valence when learning is done in the shared feature space.

Our strategy is a simple and effective way to enhance the performance of existing models. Lightweight non-linear models like Support Vector Machines can outperform complex neural networks by utilizing the shared feature space. We believe this method has great potential for various tasks and should be further investigated. In the future, we aim to create a feature space that includes a wide range of data classes to facilitate diverse recognition tasks.

REFERENCES

[1] Faranak Abri, Luis Felipe Gutiérrez, Prerit Datta, David R. W. Sears, Akbar Siami Namin, and Keith S. Jones. 2021. A Comparative Analysis of Modeling and Predicting Perceived and Induced Emotions in Sonification. *Electronics* 10, 20 (Oct. 2021), 2519. https://doi.org/10.3390/electronics10202519

[2] Margaret M. Bradley and Peter J. Lang. 2007. *The International Affective Digitized Sounds (2nd Edition; IADS-2): Affective ratings of sounds and instruction manual.* Technical report B-3. University of Florida, Gainesville, Fl.

[3] Shreyan Chowdhury, Verena Praher, and Gerhard Widmer. 2021. Tracing Back Music Emotion Predictions to Sound Sources and Intuitive Perceptual Qualities. In *18th Sound and Music Computing Conference*. Zenodo, Virtual. https://doi.org/10.5281/ZENODO.5045121

[4] Eduardo Coutinho, Jun Deng, and Björn Schuller. 2014. Transfer learning emotion manifestation across music and speech. In *IJCNN*. IEEE, 3592–3598.

[5] Jacopo de Berardinis, Angelo Cangelosi, and Eduardo Coutinho. 2020. The Multiple Voices of Musical Emotions: Source Separation for Improving Music Emotion Recognition Models and Their Interpretability. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*. Online, 310–317.

[6] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2022. Auto-Sklearn 2.0: Hands-Free AutoML via Meta-Learning. *The Journal of Machine Learning Research* 23, 1 (Jan. 2022), 261:11936–261:11996.

[7] Zi Huang, Shulei Ji, Zhilan Hu, Chuangjian Cai, Jing Luo, and Xinyu Yang. 2022. ADFF: Attention Based Deep Feature Fusion Approach for Music Emotion Recognition. In *Proc. Interspeech 2022*. 4152–4156. https://doi.org/10.21437/Interspeech.2022-726

[8] Stavros Ntalampiras. 2017. A transfer learning framework for predicting the emotional content of generalized sound events. *The Journal of the Acoustical Society of America* 141, 3 (2017), 1694–1701.

[9] Stavros Ntalampiras, Federico Avanzini, and Luca Andrea Ludovico. 2019. Fusing Acoustic and Electroencephalographic Modalities for User-Independent Emotion Prediction. In *2019 IEEE ICCC*. 36–41. https://doi.org/10.1109/ICCC.2019.00018

[10] Bjorn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. 2014. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In *INTERSPEECH-2014*. 427–431. https://doi.org/10.21437/Interspeech.2014-104

[11] Felix Weninger, Florian Eyben, Björn W. Schuller, Marcello Mortillaro, and Klaus R. Scherer. 2013. On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology* 4 (2013), 292.

[12] Wanlu Yang, Kai Makita, Takashi Nakao, Noriaki Kanayama, Maro G. Machizawa, Takafumi Sasaoka, Ayako Sugata, Ryota Kobayashi, Ryosuke Hiramoto, Shigeto Yamawaki, Makoto Iwanaga, and Makoto Miyatani. 2018. Affective Auditory Stimulus Database: An Expanded Version of the International Affective Digitized Sounds (IADS-E). *Behavior Research Methods* 50, 4 (Aug. 2018), 1415–1429. https://doi.org/10.3758/s13428-018-1027-6

[13] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. 2018. The PMEmo Dataset for Music Emotion Recognition. In *ACM Conference on International Conference on Multimedia Retrieval* (Yokohama, Japan) *(ICMR '18)*. ACM, New York, NY, USA, 135–142. https://doi.org/10.1145/3206025.3206037