Automatic Music Mood Tagging: EMMA Dataset

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Yu Kinoshita, BSc.
Matrikelnummer 01623806

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Dipl.-Ing. Dr.techn. Peter Knees
Mitwirkung: Dipl.-Ing. Dr.techn. / Bakk.techn. Richard Vogl

Wien, 23. April 2024

Yu Kinoshita

Peter Knees

Informatics

Automatic Music Mood Tagging: EMMA Dataset

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Yu Kinoshita, BSc.
Registration Number 01623806

to the Faculty of Informatics

at the TU Wien

Advisor: Dipl.-Ing. Dr.techn. Peter Knees
Assistance: Dipl.-Ing. Dr.techn. / Bakk.techn. Richard Vogl

Vienna, 23rd April, 2024

Yu Kinoshita

Peter Knees

Erklärung zur Verfassung der Arbeit

Yu Kinoshita, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 23. April 2024

Yu Kinoshita

Acknowledgements

I would like to express my gratitude to my supervisors, Dipl.-Ing. Dr.techn. Peter Knees and Dipl.-Ing. Dr.techn. / Bakk.techn. Richard Vogl, and professors of the TU Wien for the guidance, and encouragement throughout this thesis and study. Their mentorship has been invaluable in shaping the direction of my thesis.

I also want to thank the EMMA dataset Team for providing me with the needed data and information for this thesis.

To my loving partner Lisa, your endless love and support have been my anchor through the challenges and triumphs of this journey. Without you I would not have been able to complete my studies in such a balanced and relaxed manner. I also want to thank our cats Cody and Meredith for comforting me during stressful times.

Last but not least, I am deeply grateful to my family for their unwavering support. While my father may no longer be with us, without his influence, support, and unwavering belief in me, all of this would not have been possible.

Kurzfassung

Music mood tagging Models, die vorhersagen, welche Stimmungen bestimmte Musiktitel bei Menschen auslösen können, spielen eine entscheidende Rolle bei der Anwendungen von Empfehlungsdiensten (engl. Recommender Systems) für Musik und der Erstellung von stimmungsbasierter Wiedergabelisten.

In dieser Arbeit werden Methoden und Ansätze für das Tagging von Stimmungen in Musik anhand des EMMA-Datensatzes untersucht. Dieser Datensatz beinhaltet Musiktitel und die dazugehörigen Annotationen über die Stimmungen, die sie bei Zuhörern hervorrufen sollen. Um die menschliche Wahrnehmung konsistent zu erfassen, wurden die Annotationen von einem Team von Psychologen durch kontrollierte Experimente/Umfragen erstellt.

Der EMMA-Datensatz weist jedoch auch einige Hürden auf, die das Trainieren von Stimmungs-Tagging-Modellen erschweren. Bestimmte Wertebereiche von einigen Stimmungen können schwer durch das Modell vorhergesagt werden, da bestimmte Werte durch die extreme Rechtsschiefe Verteilung im Datensatz unterrepräsentiert werden. Die Datensatzgröße ist auch sehr klein, insbesondere im Vergleich zu anderen Datensätzen im Bereich des Music Information Retrieval. Daher ist die Wirksamkeit von Oversampling und Data Augmentation als Lösung auch sehr begrenzt. Die Kombination dieser beiden Probleme verstärkt die Schwierigkeit der Aufgabe.

Um diese Herausforderungen zu bewältigen, wurden verschiedene Ansätze und neu vorgeschlagene Methoden erforscht, die durch frühere Forschungen zu ähnlichen Problemen inspiriert wurden. In dieser Arbeit werden Oversampling- und Data Augmentation Ansätze für Regressionsaufgaben mit ungleichmäßig verteilten Zielvariablen untersucht, insbesondere in Fällen, in denen Labels empfindlich auf Veränderungen der Audiomerkmale reagieren.

Der vorgeschlagene DOMR-Ansatz zeigt zwar vielversprechende Ergebnisse, doch wurden nicht immer statistisch signifikante Verbesserungen der Modellleistung erzielt. Nichtsdestotrotz wurden wertvolle Einblicke in die Grenzen und Probleme bei der Nutzung des Datensatzes für Musik-Tagging-Aufgaben und die Modellleistung gewonnen, die bei zukünftigen Entwicklungen des EMMA-Datensatzes hilfreich sein können.

Zusammenfassend lässt sich sagen, dass diese Arbeit verschiedene Herausforderungen beim Stimmungs-Tagging von Musik anhand des EMMA-Datensatzes unterstreicht und erste Einblicke in mögliche Ansätze für Oversampling- und Data Augmentation gibt. Es

sind jedoch noch weitere Forschungen in verschiedenen Bereichen erforderlich, um diese Herausforderungen zu bewältigen.

Abstract

Music mood tagging, the task of predicting mood labels to music tracks, plays a crucial role in applications such as music recommendation systems and mood-based playlist generation. This thesis investigates methodologies and approaches for music mood tagging using the EMMA dataset, a dataset comprised of music tracks annotated with the moods evoked in the listeners as score values. The annotations were created by a team of researchers in the field of psychology through controlled experiments, aiming to capture human perception consistently.

Nevertheless, the EMMA dataset also presents certain limitations that present challenges for training music mood tagging models. Certain values are difficult to predict as they are underrepresented because of the extreme right skewness of certain mood scores in the dataset. The sample size is also very small, especially compared to other datasets in the field of Music Information Retrieval. Therefore, the effectiveness of Oversampling and Data Augmentations as a solution is very limited. The combination of these two issues intensifies the difficulty of the task at hand.

To address these challenges, various approaches and newly proposed methods, inspired by past research of similar tasks were explored. In an effort to overcome these challenges, this thesis investigates Oversampling and Data Augmentation approaches for regression tasks with imbalanced target variables, particularly in cases where labels are sensitive to changes in the audio features.

While the proposed DOMR approach shows promising results, statistically significant improvements in model performance were not always achieved. Nevertheless, valuable insights into the limitation and problems when utilizing the dataset for music tagging tasks, and model performance were gained, which can help future developments of the EMMA dataset. In conclusion, this thesis has shed light on various challenges in music mood tagging using the EMMA dataset and provided initial insights into possible oversampling and data augmentation approaches. However, further research into various areas is still necessary to overcome these limitations.

Contents

CHAPTER 1

Introduction

1.1 Motivation and Problem Statement

Music is a medium that can evoke a wide range of emotions in humans. The ability to automatically generate mood tags for music could greatly improve recommender systems and the experience for listeners. Music tags in the context of Music Information Retrieval (MIR) refer to label or metadata attached to a piece of music to describe its characteristics. Predicting accurate tags could allow listeners to seamlessly discover and enjoy music that resonates with their current emotional state, enhancing their interaction with music.

As shown in the works of Kim et al. [KSM+10] and Korzeniowski et al. [KNM+20], substantial research has already been conducted in the domain of automatic music tagging. Popular datasets in the Music Information Retrieval (MIR) realm, such as MTG-Jamendo [BWT+19], Million Song Dataset [BMEWL11] and MagnaTagATune [LWM+09], include mood annotations. However, they primarily focus on music-related information retrieval tasks and do not specifically address the connection between music and emotion. Further, hardly any existing works deal with the intensity of emotions evoked by music. By doing so, mood tagging has to be transformed into a regression problem (real-valued), as opposed to the more commonly used modeling as a classification of discrete emotional states in the existing literature.

In terms of technical approaches, machine learning (ML) techniques have emerged as the predominant approach for solving automatic music tagging problems as highlighted by Won et al. [WFBS20]. ML algorithms learn patterns and relationships from data, enabling them to make predictions or classifications. Among various ML models, Convolutional Neural Networks (CNNs) have demonstrated state-of-the-art performance in music tagging tasks. CNNs, a type of neural network commonly used for image and audio processing tasks, excels at capturing spatial and temporal dependencies in data through its hierarchical architecture of convolution and pooling layers. By analyzing spectrograms of the audio data, CNNs can learn patterns and correlations that correspond to specific tags in music.

Especially in music mood prediction, state-of-the-art approaches are based on audio processing [KNM+20, PS19, WFBS20], with some approaches, such as Delbouys et al. [DHP+18], expanding to multi-modal methods using information gained from the lyrics of a music piece.

A significant hurdle in this research domain is the acquisition or creation of appropriate datasets. In particular, it can be difficult to annotate data consistently since the way humans perceive emotions are subjective. EMMA, a dataset created by psychologists from the University of Innsbruck containing information on the emotional effects of hundreds of music excerpts from different genres, aims to solve this matter. [SVJ+24]

Different from most other datasets that predominantly utilize the circumplex model of emotion developed by James Russell [Rus80], which includes emotions not specifically relevant to music, EMMA employs the Geneva Emotion Music Scale (GEMS) [ZGS08]. The GEMS notation was derived from extensive research in the domain of emotion and music, and incorporates nine distinct emotion labels, offering a refined representation of the emotions evoked by music. This dataset has the potential to be used for training a model that can accurately predict the emotion evoked in the listener.

The EMMA dataset, despite its advantage in the context of music, faces challenges due to its limited size of 364 tracks and imbalances in GEMS scores. As the variance of scores in the dataset is limited and predominantly right-skewed, a naive approach of training a CNN with the data leads to a model predicting values that gravitate towards the mean values. This situation underscores the need for targeted research into the challenges encountered when training machine learning models with the EMMA dataset and GEMS. It is important to explore advanced data augmentation, processing, and resampling techniques to effectively counteract the issues of annotation imbalances and the dataset's restricted sample size. Such approaches can be essential to minimize prediction errors and increase the robustness of the models. Variants of Transfer Learning methods using pre-trained supervised and/or unsupervised models could also improve the prediction performance. [MKO+22, LG20, WFBS20]

1.1.1 Expected Result

This thesis aims to develop a music mood regression model (automatic mood tagging) that accurately captures the emotional content of music pieces in a way that is consistent with human perception. The expected outcome is a model whose output can be used to improve recommender systems and enhance the listening experience of users by allowing them to more easily find music that matches their current mood. The model should be able to predict a mood score for each GEMS label, given a music track. This score ultimately helps to identify the associated emotion of the track.

Furthermore, this thesis aims to investigate and address the challenges encountered in training a machine learning model on the EMMA dataset, particularly focusing on issues related to dataset characteristics. The insights gained from this thesis could contribute to the further development and refinement of the EMMA dataset itself.

Given all of the above challenges, the research objective of this thesis is:

"How can a music mood regression/classification model be developed that accurately captures the emotional content of music pieces in a way that is consistent with human perception, using the EMMA dataset?"

In this thesis, the following research questions will be reflected on:

Research question 1: How much can resampling methods improve the performance of a regression model in audio processing trained on the EMMA dataset?

Given the sensitive nature of the song-emotion connection, new resampling strategies are proposed, as traditional oversampling methods like SMOTR may not be ideal due to potential synthetic changes in the features. The performance will be evaluated on RMSE/R^2.

Research question 2: How much can data augmentation methods improve the performance of a regression model in audio processing trained on the EMMA dataset?

To address the imbalanced target emotion scores data transformation approaches such as compression, pitch shifting, time shift and other criteria will be used as data augmentation methods. The performance increase will be evaluated on RMSE/R^2.

Research question 3: How much can a combination of resampling and data augmentation methods improve the performance of a regression model for audio processing trained on the EMMA dataset?

To further improve the model's performance, a resampling method including data augmentation will be proposed. The performance increase will be evaluated on RMSE/R^2.

1.2 Structure

The structure of this thesis is organized as follows:

1. Related Work

   An extensive literature research of related work is needed to get a good understanding of the relationship between music and emotions, and pre-work done to the music tagging tasks/multi-label classification tasks. Additionally, this study will explore research in data augmentation methods, delving into potential strategies to address issues related to data limitations and imbalances

2. EMMA Dataset

   The EMMA dataset and its unique differences in comparison to other existing datasets will be described in detail. An in-depth analysis will be conducted, discussing its strengths and shortcomings, in order to understand its implication and potential in the field of music mood tagging research.

3. Methodology

   The approach, dataset, and model architectures employed for both regression and classification tasks will be described. The EMMA dataset will be analyzed, specifically on its skewness and imbalances regarding the annotations. Implemented pre-processing methods, models, and metrics will be described in detail. Further, an in-depth description of the oversampling methods proposed in this thesis, addressing the dataset's challenges, will be provided. Finally, the approach to examine the dependency of the methods on particular train-test set compositions will be explained.

4. Results

   In this chapter, the results will be systematically presented, focusing on the performance metrics of each model under different conditions. This includes an analysis of how various proposed oversampling and data transformation approaches have impacted model performance. The results will be compared with each other and highlighted, providing a comprehensive understanding of the efficacy of each method employed. Next, the models will undergo validation through 10-stratified fold cross-validation to assess their performance. Additional cross-validations will be conducted using 10 different seeds for selected models.

5. Discussion

   In this chapter, the results of the experiments and approaches of the proposed methods and the dataset will be explored. Further, the limitations arising from imbalances and data scarcity in particular will also be analyzed. This analysis will investigate the implications of these constraints on the effectiveness of the methods and the reliability of the results. Finally, potential strategies to mitigate these limitations will be discussed.

   Building upon identified limitations, potential directions for future research will also be explored.

6. Conclusion

   In the final chapter, the main findings of the thesis will be summarized and their implications will be discussed. The initial research questions presented at the beginning will also be revisited.

CHAPTER 2

Related Work

This chapter gives a brief overview of existing work on automatic music tagging.

There has been a lot of research in the area of automatic music tagging over the last decade. Various methods have been investigated to classify music pieces. In the past, classification models using different features such as intensity, timber, fluctuation, and rhythm [SEL11, LLZ06] have been trained, as these tend to be features that are associated with emotions as discussed in the works of Gabrielsson and Sarkar et al. [Gab16, SCD+20] . For example, higher harmonics may awake emotions such as disgust, fear, or surprise, while lower harmonics may awaken happiness or pleasantness in the listener [Gab16]. These features, such as timbre, intensity and pitch were used to train models such as Naive Bayes, k-NN, AdaBoost [SEL11], SVM classifiers, and SVM regressors [HRDH09, YLSC08].

However, feature engineering is a difficult task and is also a major research field [SCD+20]. Therefore, as of recent, more research into neural network models has been conducted as it can avoid the task of feature engineering. Particularly, following the success in image processing and speech recognition [SKrM+13], architectures based on Convolutional Neural Networks (CNN) have been used for automatic music and speech tagging of various labels, such as genre, mood[SCD+20], and instruments. [WFBS20, PNP+18, CFS16]

2.1 Convolutional Neural Network variants

2.1.1 Fully Convolutional Network

Choi et al. [CFS16] proposed a Fully Convolutional Network (FCN) which only includes convolutional layers and subsampling layers without fully connected layers. Previously, CNNs usually include fully connected layers (also called dense layers) that often act as a post-processing of the output of the convolutional layers. These dense layers tend to make up the majority of the parameters in a network and are prone to overfitting. For data processing, raw audio data is transformed into Mel spectrograms and is taken as input features. 2D kernels are used to learn temporal and spectral structures of the respective music piece. After pilot experiments, the authors decided to trim the audio signals down to 29.1s clips and downsample it to 12kHz. Even though this architecture is rather simple, it was able to achieve an AUC of 0.894 on the MagnaTagATune dataset and an AUC of 0.851 on the Million Song Dataset.

2.1.2 Musicnn

Another approach by Pons et al.[PS19, PNP+18] is the Musicnn model. This model consists of two parts, the front-end and back-end block. The front-end network is the so-called "musically motivated CNN", which introduces musically motivated features by including vertical layers that should capture pitch-invariant timbral features and horizontal filters that should capture the temporal energy envelope from the log-mel spectrogram. These features are put into the dense layers in the mid-end, which should extract higher-level representations. The output of the mid-end is then finally put into the temporal-pooling back-end, which predicts the tags from the features(see fig. 2.1). This architecture is able to achieve an ROC-AUC of 0.9069 on the MagnaTagATune Dataset and 0.8801 on the Million Song dataset [PS19].

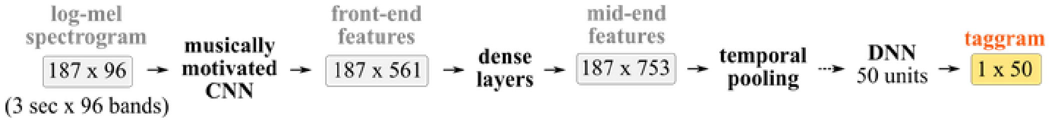


Figure 2.1: Musicnn architecture, image from [PS19]

2.1.3 Harmonic CNN model

Similar to Musicnn, Won et al. [WCNS20] tries to improve representation learning using domain knowledge. The model introduces the so-called harmonic filters, which act as a feature extractor. This turns spectrogram inputs into a 3-dimensional representation consisting of harmonic, frequency, and time ("harmonic tensor") through harmonic constant-Q transformation (HCQT) and also introduces a triangular band-pass filter to eliminate redundant convolutions in order to increase the efficiency. The HCQT output is then fed into a fully Convolutional Neural Network to classify the music tags (see fig. 2.2).

This model achieves an ROC-AUC of 0.9141 on the MagnaTagATune Dataset.

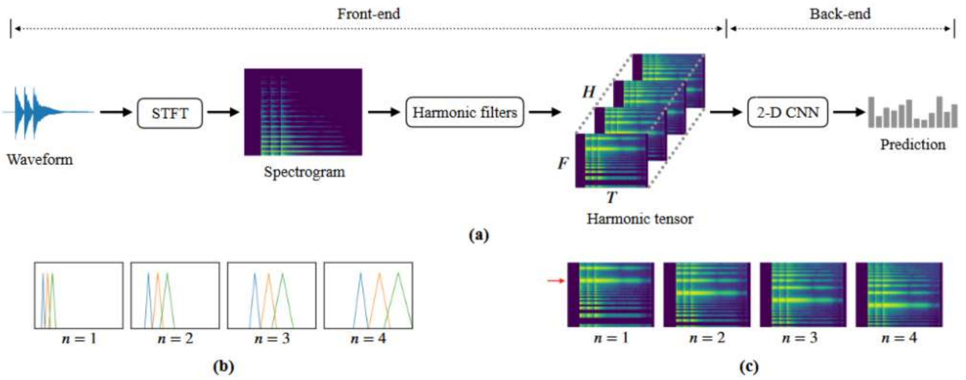


Fig. 1: (a) The proposed architecture using Harmonic filters. The proposed front-end outputs the Harmonic tensor and the back-end processes it depending on the task. The Harmonic filters and the 2-D CNN are data-driven modules that learn parameters during training. (b) Harmonic filters at each harmonic. (c) An unfolded Harmonic tensor. The red arrow indicates the fundamental frequency.

Figure 2.2: Harmonic CNN architecture, image from [WCNS20]

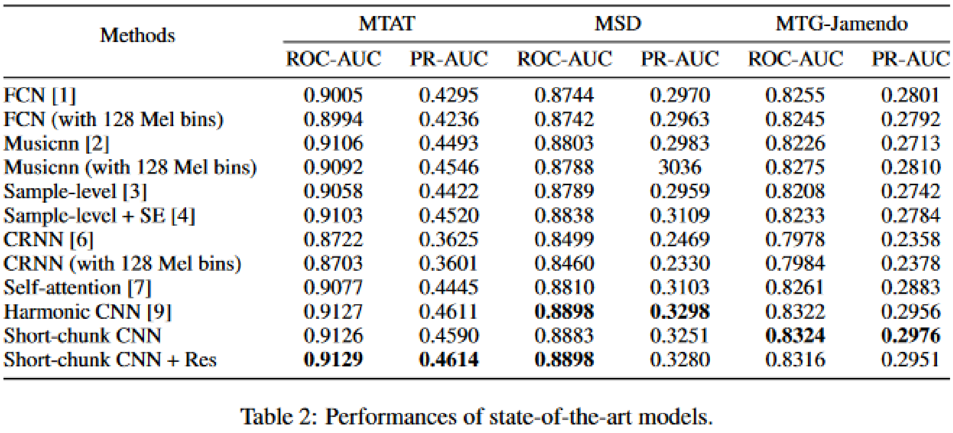
2.1.4 Short-chunk CNN model

Following the findings of the previous work of the harmonic CNN model, Won et al. [WFBS20] implements the short-chunk CNN, a simple CNN model with 3x3 filters, which is trained on small audio chunks . This architecture consists of 7 layers with residual connections and a dense output layer. This is quite similar to the FCN model, but instead of using 29.1-second snippets, it only uses 3.69s long segments of the input audio signal. Despite its rather simple architecture, it is able to achieves increased model performance, achieving an ROC-AUC of 0.9129 on the MagnaTagATune Dataset. Additionally, when Won et al. compared model performance of the short-chunk CNN, Harmonic CNN and Musicnn using three different datasets, the short-chunk CNN model was able to achieve state-of-the-art performances on the MagnaTagATune and MTG-Jamedo dataset and also was able to achieve equal performance as the harmonic CNN model on the Million Song Dataset.(see fig. 2.3)

When looking at all of the above-mentioned architectures, it seems that a complex CNN architecture may not improve the performance of the models much. Other methods, such as different data augmentation, data quality, and pre-processing methods could be a better direction to improve the model's performance.

2.2 Datasets

To develop effective models for automatic music tagging, researchers heavily rely on the availability of comprehensive and diverse datasets. There are several datasets that have been used to train and evaluate the existing machine learning algorithms.



Methods	MTAT		MSD		MTG-Jamendo	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
FCN [1]	0.9005	0.4295	0.8744	0.2970	0.8255	0.2801
FCN (with 128 Mel bins)	0.8994	0.4236	0.8742	0.2963	0.8245	0.2792
Musicnn [2]	0.9106	0.4493	0.8803	0.2983	0.8226	0.2713
Musicnn (with 128 Mel bins)	0.9092	0.4546	0.8788	3036	0.8275	0.2810
Sample-level [3]	0.9058	0.4422	0.8789	0.2959	0.8208	0.2742
Sample-level + SE [4]	0.9103	0.4520	0.8838	0.3109	0.8233	0.2784
CRNN [6]	0.8722	0.3625	0.8499	0.2469	0.7978	0.2358
CRNN (with 128 Mel bins)	0.8703	0.3601	0.8460	0.2330	0.7984	0.2378
Self-attention [7]	0.9077	0.4445	0.8810	0.3103	0.8261	0.2883
Harmonic CNN [9]	0.9127	0.4611	**0.8898**	**0.3298**	0.8322	0.2956
Short-chunk CNN	0.9126	0.4590	0.8883	0.3251	**0.8324**	**0.2976**
Short-chunk CNN + Res	**0.9129**	**0.4614**	**0.8898**	0.3280	0.8316	0.2951

Table 2: Performances of state-of-the-art models.

Figure 2.3: Model comparison, image from [WFBS20]

2.2.1 MagnaTagATune

MagnaTagATune is a dataset used in a wide range of publications related to audio classification. It was created by Law et al. [LWM$^+$09] using an online game called TagATune where two players receive either the same or different music excerpts for which they have to choose suitable tags. Then, the players have to review each other's tags and speculate on whether or not they had been listening to the same song. For a tag to be accepted into the dataset, two or more players had to have chosen the same tag in the games.

The dataset itself is comprised of 6.622 songs, 517 albums, and 270 artists. Summed up, the dataset contains 16.389 music clips, with each music clip being 29 seconds long. A wide range of genres is included, such as Classical, New Age, Electronica, Rock and Pop, among others. Each song is associated with a vector of binary tags (annotations).

2.2.2 Million Song Dataset

The Million Song Dataset (MSD) is another frequently used collection of songs and annotations in publications. It contains 1,000,000 songs as the name already implies, comprising 44,745 artists, 7,643 tags from Echo Nest API and 2,321 tags from MusicBrainz [BMEWL11]. These tags include information such as pitches, timbre, beats and other criteria of music excerpts. This dataset not only includes metadata information such as genre, year, duration and other criteria, but also similarity relationship features between artists. Further, many complementary datasets were created by the MSD community, such as lyrics, various user data, genre labels and more.

2.2.3 MTG-Jamendo Dataset

The MTG-Jamendo Dataset is an open dataset for music auto-tagging created by Bogdanov et al. [BWT+19]. It contains over 55.000 tracks with 195 tags such as genre, instruments, theme and mood. For mood, there are 6,611 annotations varying from label to label with different agreement rates. Most of them were annotated by three annotators.

2.2.4 MoodyLyrics Dataset

The MoodyLyrics dataset created by Çano et al. [cM17] differs significantly from the datasets mentioned above, as it comprises mood annotations along with song lyrics. This dataset explores mood annotations derived from the lyrical content and evaluates its performance compared to subjective music tagging methodologies. The annotations are systematically created using word lexicons, such as ANEW and WordNet, offering a potential solution to the challenge of subjective mood annotation in musical pieces.

2.3 Data Augmentation Methods In Music Classification Tasks

Data augmentation is a technique to increase the diversity and quantity of the training dataset by modifying the data in order to counteract problems such as overfitting, small volume of data and other criteria.

Two types of data augmentation methods in audio processing, sound segmentation, and sound transformation were researched by Mignot et al. [MP19] specifically in the context of musical genre classification. They focus on class-preserving data augmentation methods.

2.3.1 Sound segmentation

The audio signal of songs in the training dataset can be cut into shorter (overlapping) snippets to virtually increase the number of training examples. This is a method commonly used in the context of machine learning for different audio [PS19, PNP+18, USG14, WC05, SG15]. Depending on the window size of each excerpt, this can also cause problems. For example, songs containing multiple musical mood changes may require changing class labels. On the other hand, audio snippets that are too short may miss crucial information [MP19].

2.3.2 Sound transformations

Mignot et al. [MP19] implemented so-called elementary transformations, filtering, equalizing, noise addition, scale changes (pitch shifting and time stretching), distortions, quantization, dynamic compression, format encoding/decoding (e.g. MP3, GSM), and reverberation. These elementary transformations are then combined and applied at random to each song.

2.3.3 Results of Mignot et al.

The experiments of Mignot et al. [MP19] show that shorter song segments can be used in training and testing to increase the model performance significantly for small datasets, with 30-second segments performing the best. Shorter segments may decrease the performance. In the experiment, segments shorter than 15 seconds decreased the performance. On the contrary, Pons et al. [PNP+18] were able to achieve great performance on music classification tasks using 3-second segments. Although, it needs to be mentioned that different from classic CNNs, musicnn uses "musically motivated" filters.

Further, Mignoet et al. [MP19] were not able to answer the question of which combination of elemental transformations would improve the model. Individual transformations lead to an overfitting problem and using several transformations in a chain does not improve the model's performance on the original data, but it can increase the robustness against sound degradation.

CHAPTER 3

EMMA Dataset

The EMMA dataset [SVJ+24] was created by a group of researchers of the psychology and computer science fields from the University of Innsbruck by conducting a controlled experiment with 567 English- and German-speaking participants. The music excerpts were carefully selected by researchers with extensive knowledge of music and emotion studies, as well as students with heavy involvement with music-related activities. The set of music was selected under the consideration that it should represent a broad range of music from different genres and eras. Therefore, this dataset contains classical music that is either instrumental, vocal or both, from several major canonical periods that represent different ranges of arousal and valence. It is also comprised of pop and rap/hip-hop songs with several sub-genres from different decades.

Many works in music emotion recognition (MER), such as the 1000 Song Database from Soleymani et al. [SCS+13], AMG1608 from Chen et al. [CYWC15], MoodSwings from Speck et al. [SSMK11] and more [SVJ+24], use the so-called circumplex model of emotion, which assumes that emotions can be displayed in a two-dimensional space consisting of arousal and valence, because of its simplicity and elegance [SVJ+24]. However, it can be said that it has two significant limitations. First, it is a general model of affect and is not specifically designed for music-evoked emotions. Second, there are distinct emotions that occupy similar spaces in the model. According to Strauss et al. [SVJ+24], important differences of several positive valence and low arousal emotions evoked by music are also absent in the model. Another drawback is that it accounts for “utilitarian” emotions (e.g. anger, fear, or disgust). These emotions have limitations when describing what kind of emotion music can trigger in a human (e.g. music rarely triggers anger).

Therefore, the EMMA dataset uses the so-called Geneva Emotional Music Scale ("GEMS"). These emotion labels were derived from extensive research on emotion and music. Zentner et al. [ZGS08] started out with 515 initial emotion terms and eliminated those that were rarely evoked by music with the help of experiments. The scale includes nine first-order factors that can be assigned to one of the three second-order factors Sublimity ("Wonder", "Transcendence", "Tenderness", "Nostalgia", "Peacefulness"), Vitality ("Power", "Joyful activation"), and Unease ("Tension", and "Sadness"). For this thesis, only the first-order factors will be considered. Each factor can be rated between a score of 0 and 100. Emotions that were not selected in the experiment for the EMMA dataset are set to a scale of 0 [SVJ+24].

3.1 Advantages, Disadvantages & Differences

One major difference of this dataset in comparison to other datasets in this field (see section 1.3.2) is the experiment setup. While other datasets were created using an online game, online music platform, and/or annotators, the EMMA was created in a controlled setting while also considering psychological and musical aspects [SVJ+24]. However, it remains to be seen whether this method can be considered as an improvement. As previously mentioned, the collection of music excerpts was selected by researchers with extensive knowledge on music and human emotion, as well as students with strong connections to music [SVJ+24].

Further, the EMMA dataset sets itself apart by deploying scores to represent emotions, instead of to the binary labels commonly used in other datasets. This scoring approach allows for a more nuanced depiction of the intensity of emotions experienced, offering a richer understanding beyond a simple 'Yes' or 'No' response. The EMMA dataset also considers musically relevant emotions outside of the basic emotions theory. It deliberately omits labels that are not suited to describe the emotions evoked by music in listeners. This approach provides a more comprehensive and less constrained framework for understanding the emotional impact of music [SVJ+24].

One drawback of this dataset becomes visible when it is applied to a classification task, as the annotations consist of scores ranging from 0 to 100. The conversion of the scores into binary values for classification can be complicated, since the distribution of each emotion score varies. Selecting a threshold at which score the song will be considered to the emotion label is also not an easy task. For instance, it can be disputed whether classifying a song with a joy score of only 2 is in fact capable of evoking joy in listeners. Determining a threshold to classify each emotion label is not only difficult, but could itself be a subject of research.

Imbalanced score labels present another challenge in this dataset. In every mood, we can see that the scores are right-skewed (see figure 3.1). Especially, "Sadness", "Tenderness" and "Tension" have extreme imbalances in the scores. E.g. for "Sadness", most of the values are in the range of 1 to 5, while there are not many data excerpts for higher values (see figure 3.1).

Additionally, due to the dataset's imbalances, training a machine learning model is made much more challenging by the small dataset size of only 364 songs, especially compared to the size of other datasets in this field, such as the MagnaTagATune with 16,389 songs and MTG-Jamedo with over 55,000 songs [BWT+19, LWM+09].

Due to the limited availability of certain ranges of emotion scores, predicting those ranges is more difficult. This limitation not only complicates the training process for certain features, but also complicates evaluating the model. The reliability of the model's predictions and its performance metrics can become heavily dependent on the composition of the train-test sets. When certain emotion ranges are underrepresented in the testing set, it may lead to skewed results. Therefore, results will need to be validated using various methods, such as the cross-validation.

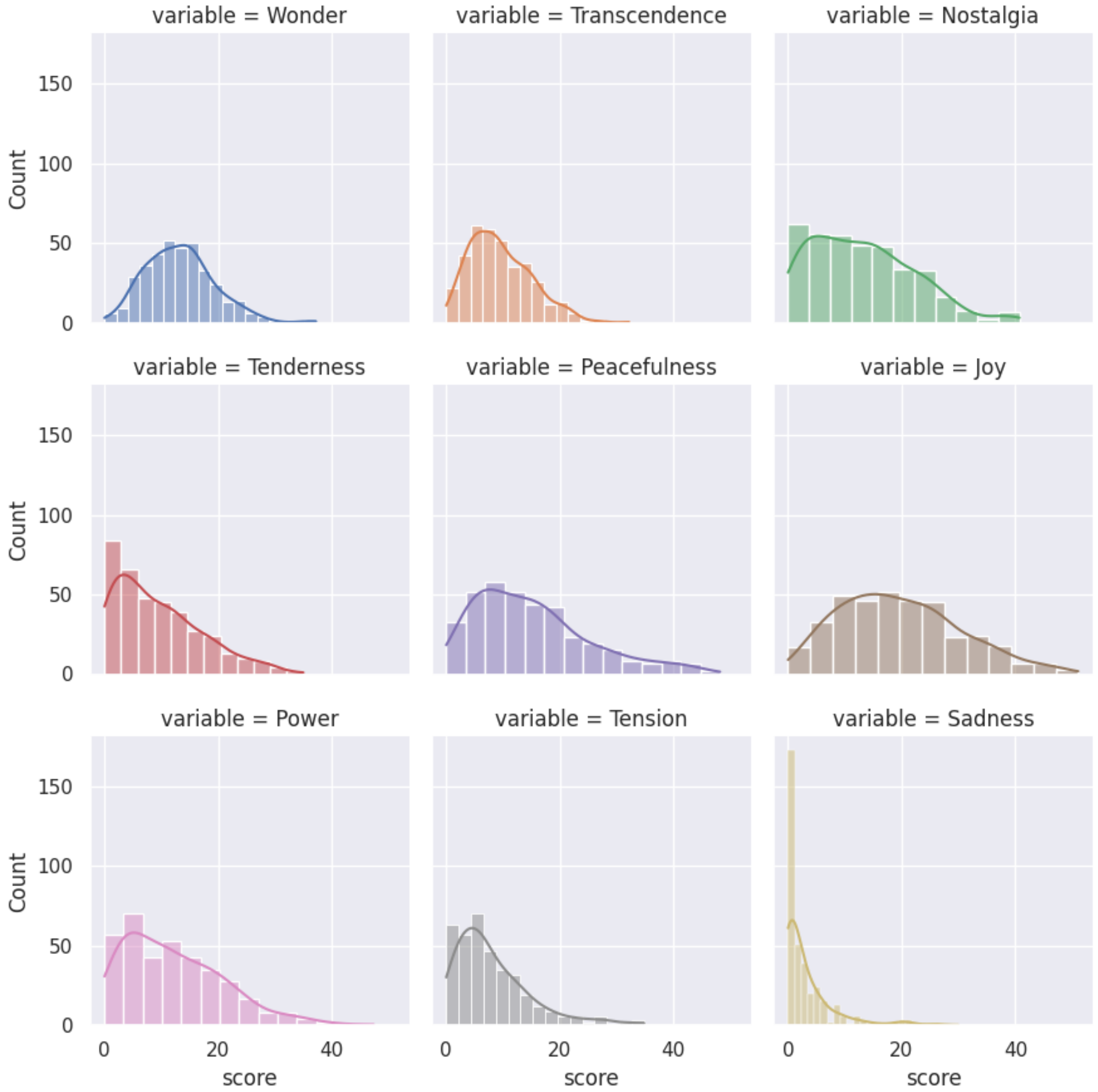


Figure 3.1: EMMA Dataset distribution [SVJ+24]

CHAPTER 4

Methodology

4.1 Approach

First, chosen state-of-the-art models were trained without any specific data augmentation methods or fine-tuning of the model to study the prediction performance. The next steps are determined through studying the prediction performance and exploratory data analysis (EDA).

4.1.1 Pre-processing

The dataset in this thesis will be divided into training, validation, and test sets in a 0.7, 0.15, and 0.15 ratio, respectively. Following this, the audio data will be segmented into smaller snippets, with the duration of each snippet varying between approximately 2 to 30 seconds, depending on the architecture being used. This thesis will tailor the snippet length to the specific architecture.

The audio snippets will be transformed into Mel-spectrograms as seen in the works of Choi et al. [CFS16], Pons et al. [PNP$^+$18], Won et al. [WCNS20] and Mccallum et al.[MKO$^+$22]. These studies have demonstrated the effectiveness of Mel-spectrograms for automatic tagging tasks. Mel-spectrograms provide an effective time-frequency representation of sound, mimicking human perception through the application of the so-called Mel scale. This scale, based on the human auditory system's response to sound, enables the representation of sound in a manner closely aligned with human perception. The raw audio data will be transformed into spectrograms with overlapping frames and each spectrogram will be transformed into the mel scale using a set of filterbanks.

Later in this thesis, various data augmentation methods, which form a part of the pre-processing stage, will be examined.

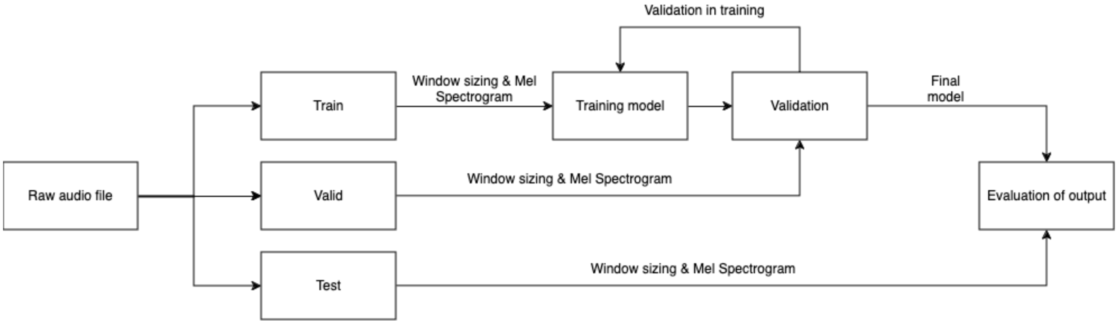


Figure 4.1: Architecture

4.1.2 Metrics

R^2 statistics and Root Mean Squared Error (RMSE) will be used to evaluate the models as these metrics have also been used in various other music emotion recognition works such as Cunningham et al.[SHJR21], Yang et al.[YLSC07], Guan et al. [GCY12], Choi et al. [CFSC17] and He et al. [HF20].

The R^2 statistic, also known as the coefficient of determination, indicates the percentage of variance of the dependent variable (y) that can be explained by the independent variables (x) in a regression model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.1}$$

The RMSE is a widely used metric in regression tasks, that shows the accuracy of a regression model. It takes the rooted mean value of the squared errors between the predicted and actual values:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{n}} \tag{4.2}$$

4.2 Models

Through a literature review of previous work, including studies by Won et al. [[WFBS20] and Choi et al. [CFS16, CFSC17], a simple baseline model has been selected to compare the performance of the developed model. The following Convolutional Neural Network models were explored in this thesis. They were derived from various existing papers on music audio classification tasks.

4.2.1 Convolutional Neural Networks (CNN)

CNNs, originally motivated by the way biological vision systems work, try to capture spatial information (low-level information) and use this information to capture high-level information [LB98]. Therefore, CNNs have traditionally been used for visual analysis tasks. Interestingly, many state-of-the-art audio classification models also rely on CNNs as shown in the evaluation work of state-of-the-art approaches of Won et al. [WFBS20]. This can be traced back to the assumption that acoustic events can be learned by seeing their time-frequency domain, the Spectrograms which visualize the spectrum of frequencies of a signal. Choi et al. [CFS16] justifies the use of CNNs in automatic tagging as follows, music tags are considered high-level features that represent an information level above the intermediate features such as beats, cords, tonality and other criteria. This suits the CNN well as it learns hierarchical features over multilayered information structures.

As already mentioned in section 4.1.1, the input data is transformed into the time-frequency domain, specifically the Mel-spectrograms are widely adopted for tasks such as tagging [DS13] or onset detection [SB14]. The Mel scale is used to mimic the way humans perceive the difference of pitch or frequency [Moo12]. This scale is therefore suited for tasks that capture the content of music pieces in a way that is consistent with human perception.

Fully Convolutional Network

As already mentioned in chapter 2, Choi et al. [CFS16] introduced the fully convolutional network (FCN), which is essentially a CNN without fully connected layers. A dense layer is used to predict the class labels as output. This will be left out in this regression task as we do not have any class labels to predict. Therefore this model will be converted from a multiclass classification model to a multivariate regression model. Further, an average-pooling layer is implemented in the FCN as the last layer instead of a sigmoid activation function. The rest of the architecture follows generic CNN structures. 1D 3x3 convolutions are used with the numbers of layers varying from four to seven, depending on setup. Further, each layer has batch normalization layers, max pooling layers to reduce the size of the feature map and dropout layers for regularization. Based on the Choi et al. approach, the chunk size for this model is set to 30 seconds, amounting to a total of 625 training samples.

The FCN architecture is rather simple, but effective in audio classification tasks, achieving state of the art performance in certain tasks as seen in the work of Won et al. [WFBS20] where they conduct comparisons between state-of-the-art models. Therefore, the 6-layer FCN that performed the best on the EMMA dataset with an R^2 of 0.29 and RMSE of 6.32 will act as base model for this thesis. (See table 4.1)

Model	MAE	RMSE	R^2
7-Layer FCN	4.900	6.446	0.276
6-Layer FCN	4.777	6.318	0.287
5-Layer FCN	4.998	6.595	0.237
4-Layer FCN	4.914	6.528	0.230

Table 4.1: Performance metrics of various Fully convolutional networks (based on works from Choi et al.[CFS16])

Short-Chunk CNN Model

Furthermore, the Short-Chunk CNN model from Won et al. [WCNS20, WFBS20] mentioned in chapter 2 which was able to produce promising results, is implemented. For this model a very short audio chunk size of 3.6 seconds is used. Additionally, the sigmoid layer in the last layer is removed to obtain scores as output instead of binary class labels.

Table 4.2 illustrates that the short-chunk CNN model performed relatively poorly compared to the 6-layer FCN model, with an RMSE of 7.14 and R^2 of 0.104.

Model	RMSE	R^2
Short-chunk CNN	7.140	0.104

Table 4.2: Performance metrics of Short-chunk CNN

4.2.2 Transfer Learning - MULE (Musicset Unsupervised Large Embedding)

In recent years research has been conducted in so-called audio understanding models which can help in solving diverse problems in audio related tasks such as speech, music, environment and more [MKO$^+$22, WLW$^+$21]. These models produce audio representations that can be used in various models and settings. Choi et al. [CFSC17] were also able to produce promising performances on various different audio tasks using a pre-trained Convolutional Neural Network that was trained to predict music tags from the Million Song Dataset.

Another research of McCallum et al. [MKO$^+$22] compares the performance of audio representations produced by supervised and unsupervised learning models. The models were trained using 3-second snippets of the raw audio file. The supervised model, a Short-Fast Normalizer-Free Net F0, was trained on mel-spectrograms and existing binary labels. For unsupervised learning, the so-called SimCLR objective, which is a contrastive self-supervised learning algorithm from Chen at al. [CKNH20], is deployed. Using the features produced by these models, simple multi-layer perceptron (MLP) classifiers were trained. These models have been able to deliver promising results in audio and visual understanding.

McCallum et al. concluded that supervised learning on large-scale annotated datasets can achieve state of the art performances, but also that unsupervised learning methods can achieve promising results, especially on labels or information that are not much present in the supervised dataset.

In this thesis a simple feed forward neural network model with four fully connected layers with ReLU activation functions, batch normalization and dropout layers is trained using the audio representation of the above mentioned unsupervised model, in particular MULE (Musicset Unsupervised Large Embedding) [MKO+22]. This will help determine the direction of research to improve the prediction performance of the annotations of the EMMA dataset.

Table 4.3 depicts that the model has only a slightly improved RMSE of 6.04 compared to the FCN model and a similar R^2 of 0.29.

Model	**RMSE**	$\mathbf{R^2}$
Transfer learning unsupervised audio embeddings	6.0427	0.2936

Table 4.3: Transfer learning unsupervised audio embeddings

4.3 Exploratory Data Analysis

In this section, a comprehensive exploratory data analysis is performed to highlight the characteristics of the dataset. Further, the imbalances present within the dataset will be examined.

The skewness is interpreted by using the rule of thumb for skewness interpretation by Bulmer [Bul79]:

- Highly skewed: Skewness is less than 1 or greater than +1
- Moderately skewed: Skewness is between 1 and $\frac{-1}{2}$ or between $\frac{+1}{2}$ and +1
- Approximately symmetric: Skewness is between $\frac{-1}{2}$ and $\frac{+1}{2}$

Analyzing the skewness presented in table 4.4 using the interpretation by Bulmer, we can conclude that Tension and Sadness exhibit highly skewed distributions. Conversely, Wonder, Transcendence, Nostalgia, Tenderness, Peacefulness, and Power demonstrate moderately skewed distributions. Joy is the only emotion that has an approximately symmetric distribution.

Upon closer examination of table 4.5, which shows the percentiles and maximum values, we can see that, although the differences between the Q1, Q2, and Q3 percentile values are relatively small, a pronounced discrepancy can be observed between the 75th percentile and the maximum values across all emotion labels, indicating a rather high concentration of the data on the lower end (left side) of the distribution, suggesting a right-skewed distribution. Additionally, it is evident that the maximum values for each emotion label are notably distant from the highest attainable GEMS score of 100.

	Skewness
Wonder	0.630
Transcendence	0.746
Nostalgia	0.633
Tenderness	0.850
Peacefulness	0.944
Joy	0.484
Power	0.853
Tension	1.468
Sadness	2.579

Table 4.4: Skewness of emotions scores (EMMA dataset)

	Q1 (25th)	Q2 (50th)	Q3 (75th)	Maximum
Wonder	8.89	13.22	16.71	37.21
Transcendence	5.28	8.71	13.24	32.16
Nostalgia	5.24	12.47	18.95	40.71
Tenderness	3.30	7.74	13.88	34.970
Peacefulness	7.13	13.18	20.32	48.16
Joy	11.56	18.60	26.69	51.04
Power	5.07	10.61	18.06	47.34
Tension	3.32	6.18	11.25	34.79
Sadness	0.0	1.46	4.23	29.94

Table 4.5: Percentile of emotions scores (EMMA dataset)

In chapter 4 we were able to observe model output that can be contributed to the right-skewness of the data. Comparing the predictions of the baseline model (FCN model) with the original annotations of the test set, we can see that certain values are not predicted consistently. For example, we can observe that the model does not predict any scores above around 21 for "Power" as illustrated in fig.4.2, fig.4.4 and fig.4.3 Similarly, the model makes no predictions above 17 for "Wonder". Similar behavior can be observed across all emotion scores.

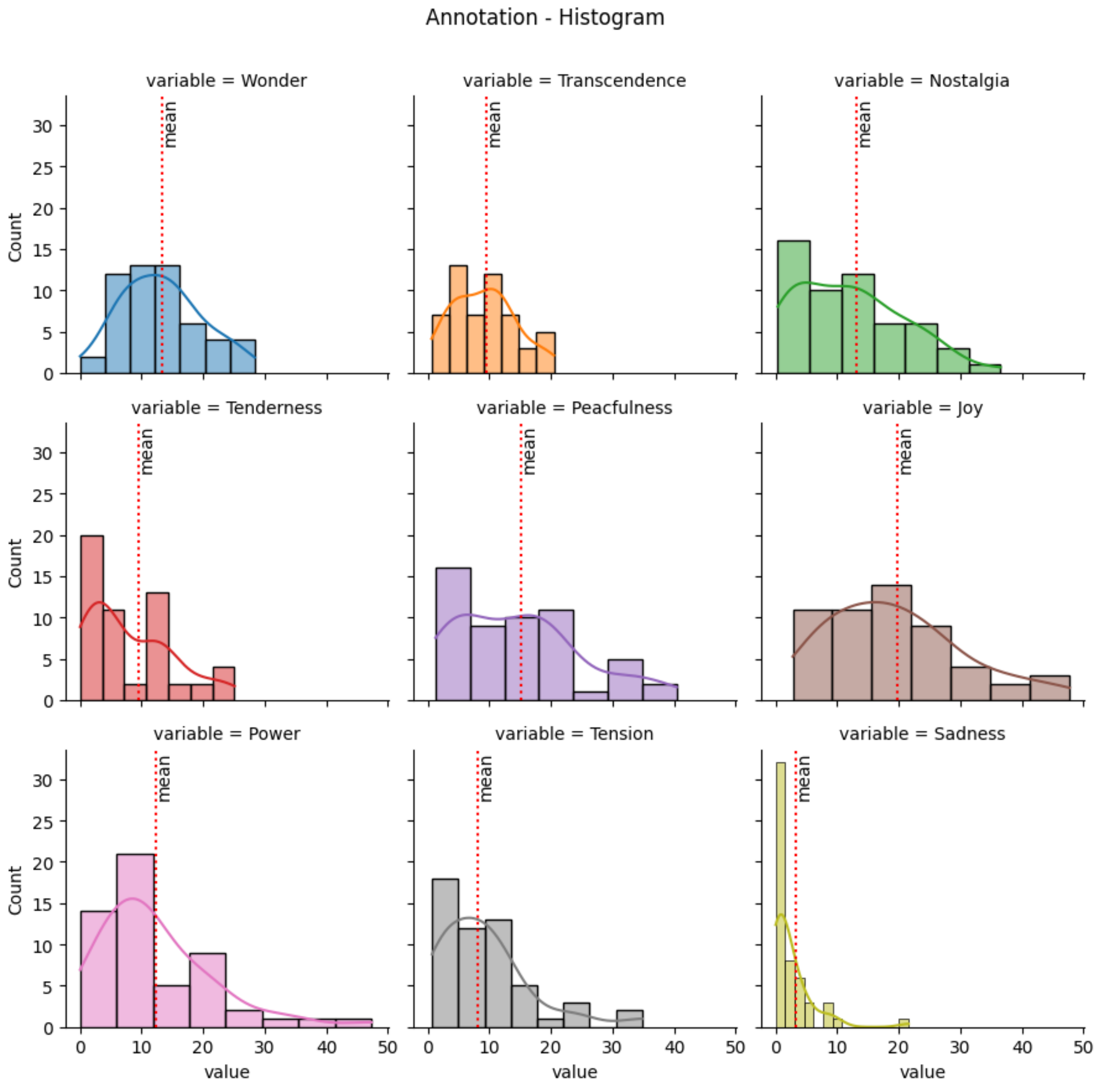


Figure 4.2: Test set Annotation

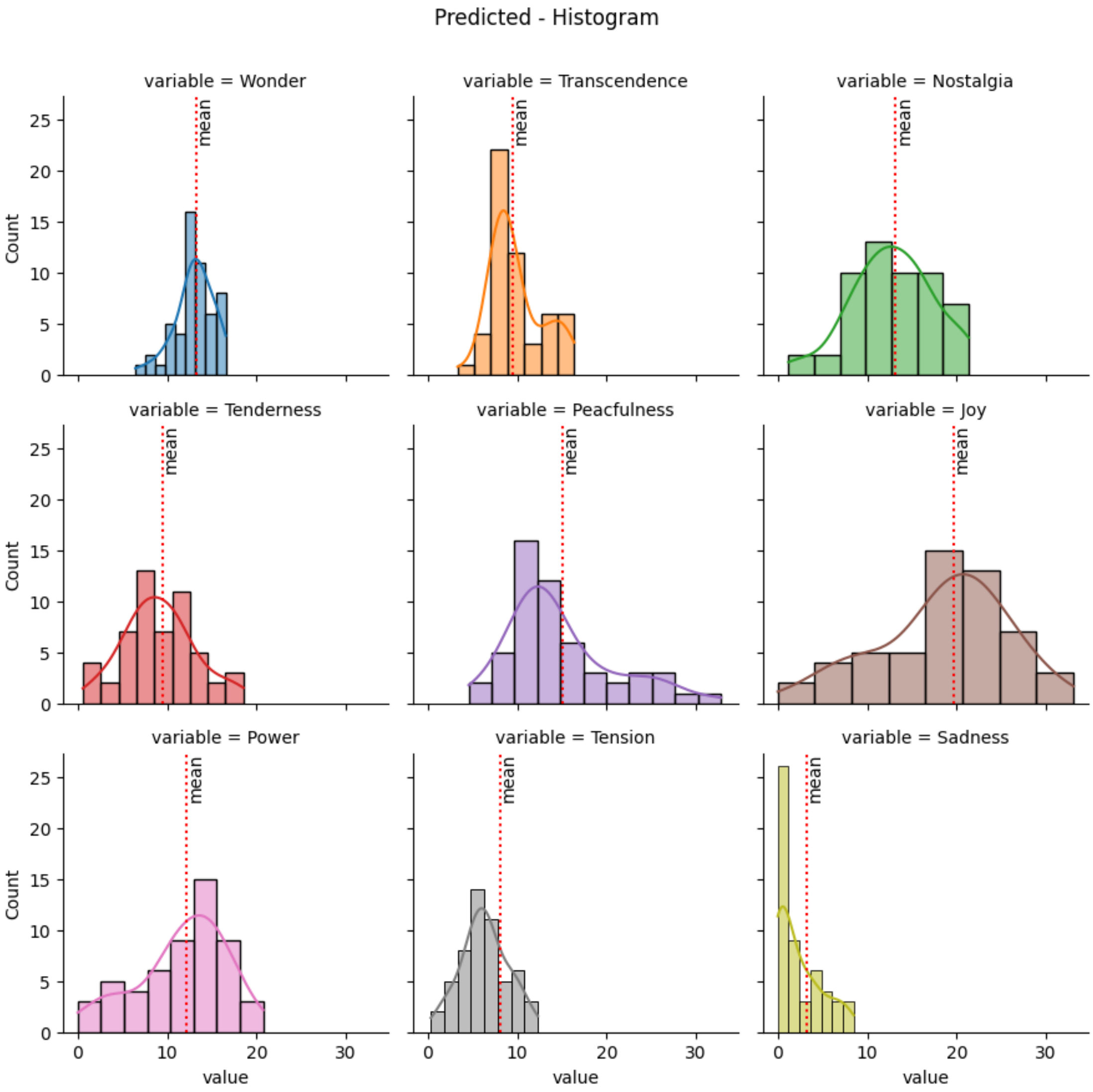


Figure 4.3: Baseline Test set Prediction

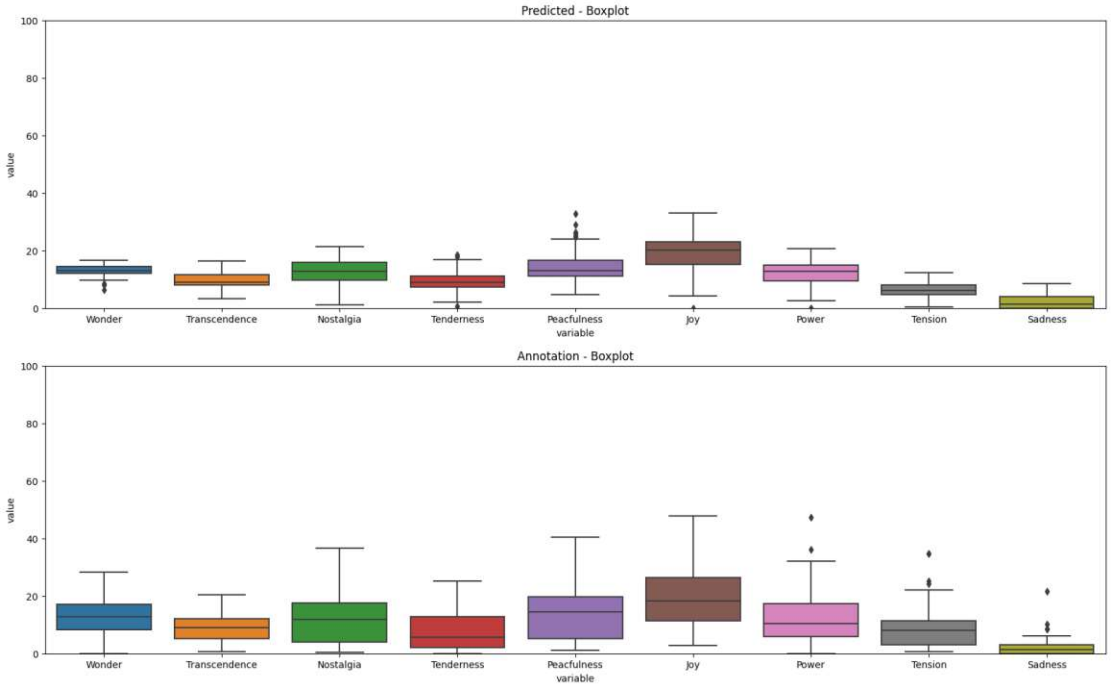


Figure 4.4: Baseline Boxplot Prediction vs Annotation (Test set)

All of this suggests that the performance problem may not lie in the model's architecture but rather in the structure of the dataset, as already discussed in the Chapter 3.

To address this problem, different methods will be explored. For typical classification problems, there exist various methods to address the imbalances, but for regression problems things get more complicated. Especially, re-sampling methods such as over-sampling, under-sampling [KKP+06], synthetic minority over-sampling (SMOTE) [FGHC18] and other criteria cannot be simply converted for regression task use, as they make use of minority/majority classes, which do not exist in a regression problem.

4.4 Addressing The Imbalance

4.4.1 DenseLoss

In typical cost-sensitve learning methods, the cost of missclassification is modified in a way that certain missclassified classes have higher error costs than other classes [KKP+06]. Steininger et al. [SKD+21] introduced the so-called DenseLoss, a cost-sensitive learning approach for neural network regression tasks. This method makes use of a sample weighting method called DenseWeight, also introduced by Steininger et al., which assigns weights to each data point according to its rarity in the dataset using the kernel density estimation (KDE):

$$p(y) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{y - y_i}{h}) \tag{4.3}$$

The weighting is calculated using the normalized density function p' and a hyper-parameter α which controls the strength of density-based weighting:

$$f'(\alpha, y) = 1 - \alpha p'(y) \tag{4.4}$$

The Gaussian kernel is used as the kernel function and the Silverman's rule is used for the bandwidth selection.

The DenseLoss is an approach implemented on the algorithm-level. Thus, making it flexible in comparison to approaches on the data-level (i.e. re-sampling methods). It integrates the DenseWeight into a loss function to obtain a cost-sensitive approach. In training, the calculation of the gradient will take the weights into account. Meaning, rarer data points will output a larger gradient than common data points based on DenseWeight. It also can be adapted to any gradient descent optimization algorithm.

4.4.2 Oversampling Approach And Single-Class Ensemble Models

Under certain conditions a one-class model can outperform a mutli-class model [KKP+06]. Raskutti and Kowalczyk [RK04] show in their experiments that in extremely unbalanced datasets, the one-class models can indeed outperform multi-class models.

Further, since each label score is differently distributed, we can also observe the effect of different architectures, loss function and data augmentation methods on each individual label more closely using single-task models. Single-class models also have the advantage that classic re-sampling methods can be applied more easily as each individual label can be assigned its own re-sampled dataset.

If results are promising, an ensemble model of single-class regression models could be considered.

Density Based Oversampling For Single Variable Regression (DOSR)

Inspired by the weighting method of DenseLoss from Steininger et al. [SKD+21] and KDE based class oversampling from Kamalov et al. [Kam20], this thesis proposes an oversampling technique that relies on kernel density estimation. The hyperparameter 'threshold (t)' plays a crucial role, determining which kernel density values are classified as a minority data point and, consequently, included in the pool of candidates for oversampling. This threshold effectively divides the dataset into segments akin to 'rare' - minority and 'common' - majority data points, paralleling the minority-majority class dichotomy in classification. Two specific strategies are considered:

DOSR: Each data point within the candidate pool is oversampled once.

DOSR Percentile: Data points are oversampled with a frequency proportional to their rarity, as determined by their percentile ranking. Data in the third quartile (Q3) will undergo oversampling once, the second quartile (Q2) twice, and, finally, the first quartile (Q1) will be oversampled three times.

However, this method has a limitation that it cannot be used for multivariate models. Since each emotion score yields distinct kernel density estimation values, this leads to varying candidate pools for each emotion label, complicating the application of this oversampling method across multiple variables. The architecture of the proposed approach is illustrated in Figure 4.5.

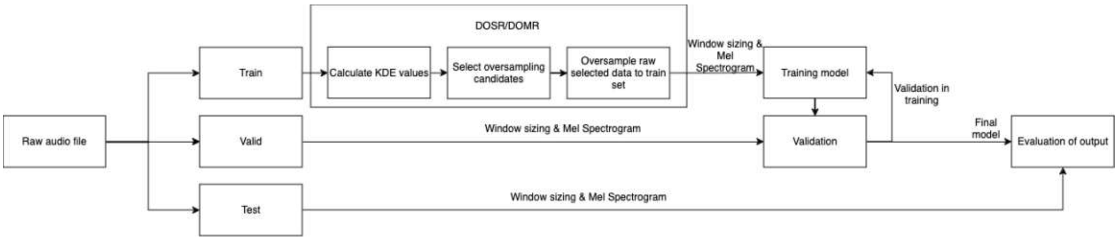

DOSR/DOMR
Raw audio file
Train
Valid
Test
Calculate KDE values
Select oversampling candidates
Oversample raw selected data to train set
Window sizing & Mel Spectrogram
Training model
Validation in training
Window sizing & Mel Spectrogram
Validation
Final model
Evaluation of output
Window sizing & Mel Spectrogram

Figure 4.5: DOSR & DOMR architecture

Density Based Oversampling For Multivariate Regression (DOMR)

To apply the Density Based Oversampling For Single Variable Regression (DOSR) approach on a multivariate model, the kernel density estimation of each emotion label will be aggregated by calculating the average values for each song. This way we end up with one single kernel density estimation value which will then be used to determine if a sample should be oversampled. Similar to DOSR, the hyperparameter threshold (t) will determine which values will be considered for oversampling.

In addition, a modified approach, akin to the percentile method used in DOSR, is proposed (termed as the DOMR Percentile). Here, data points with lower average KDE values will be subject to more frequent oversampling, rather than a singular instance. Data in the first quartile (Q1, 25th percentile) will be oversampled three times, data in the second quartile (Q2, 50th percentile) will be oversampled twice and data in the third quartile (Q3, 75th percentile) only once.

Finally, the DOMR No-Pool is introduced as an alternative strategy. This method retains individual KDE values for each emotion score, without averaging. Under this model, if any emotion score's KDE value falls below the threshold (t), that data point is included in the oversampling candidate pool.

4.4.3 Data transformation

Inspired by Mignot et al. [MP19], the following class-preserving elementary transformations are implemented to improve the models performance. The changes from the transformations will be clearly noticeable, but in a way that it preserves the class labels.

The strength of all of the changes will be based on the work of Mignot et al.. The strength of the transformation (γ) is set to 1. The transformations are implemented using the audiomentations python package [JTC+23].

Pitch shifting and Time stretching

The strength of the pitch shifting factor (s_p) in semitones and time stretching scaling (s_t) in percentage are randomly chosen using the formula from Mignot et al. [MP19]:
$\left(\frac{s_p}{4}\right)^2 + \left(\frac{s_t}{5}\right)^2 = \gamma^2$

a_p and a_t set to 4 and 5 respectively, represent the maximum changes.

Mp3 compression

The audio file is compressed into an Mp3 file using the lame encoder with a bitrate of 40 kbps [MP19].

Background noise

One random background noise is added to the audio with a signal-to-noise ratio of 18 dB [MP19]. The background noise are taken from the ESC dataset, a labeled collection of environmental audio recordings for benchmarking [Pic].

Time shift

An imperceptible time shift is applied to the audio file. It is shifted by $\pm 50\gamma$ in milliseconds [MP19].

Low pass filter

A low pass filter is applied to the audio. The starting frequency is chosen between 20 and 150 hertz and the slope is given by $a = 6\gamma$ in dB/decade [MP19].

Gain

A gain transition is applied to the audio varying between -4γ and +4γ in dB, simulating a fade in and fade out [MP19].

Compression

The audio is compressed with a release time of 100 ms and a threshold of –60dB [MP19].

Application

Inspired by Mignot et al. [MP19], the transformations will be applied in a chain of transformations. Four configurations will be explored:

- Raw audio data + 1 transformation
- Raw audio data + 2 transformations
- Raw audio data + 4 transformations
- Raw audio data + 8 transformations

The applied transformation will be randomly chosen, the combination will be without repetition. Meaning, if a transformation was already applied to the audio, it will not be chosen again. Other than obtaining doubling the sample size, this could improve the robustness of the model against noise.

4.4.4 Density Based Oversampling For Multivariate Regression With Data Transformation (DOMR+)

In light of the observed performance gains from the data transformation method and the Density-Based Oversampling for Multivariate Regression (DOMR) method, I propose an integrated approach that incorporates data transformation into DOMR. This approach will retain the kernel density estimation calculation for each emotion score from DOMR. In contrast to DOMR, where raw data is oversampled, this method will oversample using the transformed data, employing the same transformation techniques inspired by Mignot et al. [MP19]. The hyperparameters will be the same as for DOMR, the "threshold (t)" and "number of transformations" for the number of chained transformations without repetition.

Further, similar to DOMR, this approach proposes two additional variants:

DOMR+ Percentile: This variant will focus on oversampling data points that fall within the lower percentiles of kernel density estimation values with greater frequency. (see section 4.4.2)

DOMR+ no-pool: This version will maintain individual kernel density estimation values for each emotion label.

The proposed approach's architecture is visually depicted in Figure 4.6.

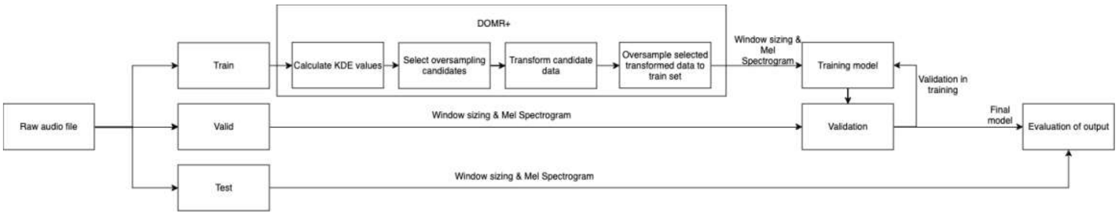

DOMR+
Raw audio file
Train
Valid
Test
Calculate KDE values
Select oversampling candidates
Transform candidate data
Oversample selected transformed data to train set
Window sizing & Mel Spectrogram
Training model
Validation in training
Validation
Window sizing & Mel Spectrogram
Final model
Evaluation of output
Window sizing & Mel Spectrogram

Figure 4.6: DOMR+ architecture

4.5 Dependency Of Training Set Composition

Due to the imbalances in the EMMA dataset, the models will be examined of its dependency on a particular train set composition and the performance will be validated. A 10-fold Cross-Validation will be used where the dataset will be split into ten equal sized subsamples. One subsample will be used as the test set and the rest will be used as the train-validation set for training. Each iteration the model trained on the train-validation set (9 folds) will be evaluated on the test set (1 fold). This is repeated ten times and each of the 10 subsamples will be used once as a test set for evaluation. The evaluation metrics from each iteration will be collected and inspected for further implication. If the metrics are close to each other, it implies that the model is not dependent on the training set composition. On the other hand, if only a few fold variations perform well, it is implied that the model is indeed strongly influenced by the training set composition. Figure 4.7 presents a graphical illustration of the process.

In order to maintain similar data distribution throughout the folds, a stratified CV-fold using the KDE values is proposed. First, the kernel density estimation values for each data point will be calculated. Based on these values, the data will be categorized in 10 equal sized splits according to its "rareness". The assembly of equal-sized folds will then be conducted according to this categorization.

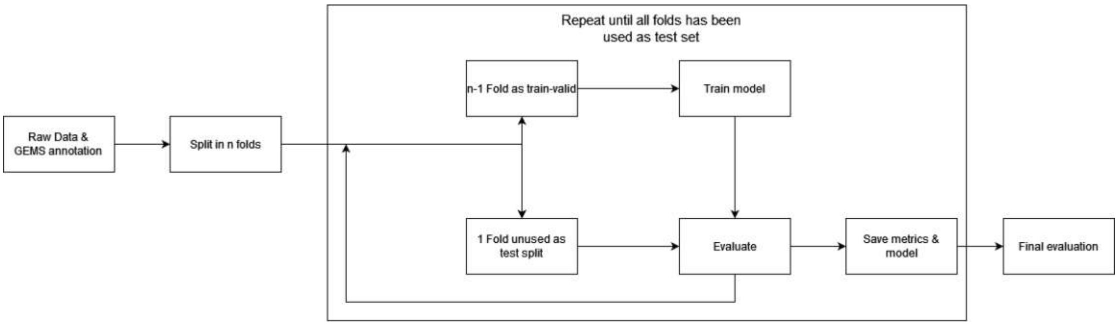

Repeat until all folds has been used as test set
Raw Data & GEMS annotation
Split in n folds
n-1 Fold as train-valid
Train model
1 Fold unused as test split
Evaluate
Save metrics & model
Final evaluation

Figure 4.7: Dependency of training set composition

CHAPTER 5

Results

5.1 DenseLoss as Loss function

FCN models are trained using the DenseLoss function from Steininger et al. [SKD$^+$21]. Kernel density estimations are calculated for each emotion score using the training set and based on that the weight for the loss function will be calculated.

A multivariate FCN 6 layer regression model using the DenseLoss function [SKD$^+$21] is trained using audio snippets of 30 seconds, the sample size amounts to 625 samples. Looking at the metrics (see table 5.1), the models trained with DenseLoss decreased in performance in comparison to the baseline model. But this can be expected as every emotion score has different distributions and applying the DenseLoss on each score may be practical. But looking at the RMSE and R^2 of each mood score in table 5.2 of the Multivariate FCN6 Denseloss model, Denseloss does not improve the performance at all. Another explanation is that the Denseloss was introduced by Steiniger et al. for a single variable regression model instead of a multivariate regression model.

Therefore, an ensemble of single variable FCN 6 layer regression models deploying the DenseLoss function is introduced to analyze the performance difference when applying the DenseLoss on only a single emotion score. The parameter α was tuned in the range between 1 and 3. Table 5.2 shows that a marginal performance increase can only be observed for Power and Tension while all the other mood scores decreased in performance.

When examined more closely, we can see that the performance decrease from the DenseLoss model is mainly caused by the performance of the emotions Wonder, Joy and Peacefulness as seen in table 5.3. Implying that the score distribution of these scores may not be suitable for the use of the DenseLoss function. Subsequently, we change the ensemble model by incorporating the more effective single-variable model, chosen between the baseline and DenseLoss models. This integration yields a new mixed model termed Ensemble FCN6 L1 & DenseLoss. While the performance of this mixed model improves, it still falls short of outperforming the multivariate regression model (refer to Table 5.4). Therefore, based on the analysis of the ensemble FCN models, it could be concluded that multivariate regression models tend to exhibit superior performance across this task overall.

Model	RMSE	R^2
FCN6 Baseline	**6.272**	**0.292**
FCN6 DenseLoss	7.409	0.037

Table 5.1: DenseLoss performance on FCN

Model	Wonder	Trascendence	Nostalgia	Tenderness	Peacfulness	Joy	Power	Tension	Sadness
Multivariate FCN Baseline	6.295803	4.117	6.609	5.372	6.631	8.601	7.950	7.757	3.333
Multivariate FCN6 Denseloss	6.630	4.134	9.342	7.405	8.625	10.017	8.645	8.001	3.883
Ensemble FCN6	6.376	4.052	8.627	5.574	6.738	9.458	8.414	7.216	3.449
Ensemble FCN6 DenseLoss	7.512	4.458	6.713	5.736	10.28	14.44	7.838	6.868	4.342

Table 5.2: Multivariate FCN6 vs Ensemble FCN6 vs Ensemble FCN6 Denseloss RMSE

Model	Wonder	Trascendence	Nostalgia	Tenderness	Peacfulness	Joy	Power	Tension	Sadness
Ensemble FCN6 DenseLoss	-0.329	0.236	0.401	0.309	-0.023	-0.767	0.293	0.201	3.333

Table 5.3: R^2 of DenseLoss on single variable models

Models	**Overall RMSE**	$\mathbf{R^2}$
Multivariate FCN6 Baseline	6.272	0.292
Ensemble FCN6	6.672	0.213
Ensemble FCN6 DenseLoss	7.576	-0.012
Ensemble FCN6 L1 & DenseLoss	6.356	0.277

Table 5.4: Overall RMSE and R^2

5.2 Density Based Oversampling For Single Variable Regression (DOSR)

Next, we examine the performance of the Density Based Oversampling For Single Variable Regression proposed in section 4.4.2 through single variable models and Ensemble models. Two proposed models are implemented, "Ensemble FCN6 DOSR" and "Ensemble FCN6 DOSR percentile". For each single variable model the oversampling kernel density estimation value threshold (t) is finetuned in the space between 0.3 and 0.7. For each model in the ensemble, the threshold with the best performance is chosen.

Model	Wonder	Trascendence	Nostalgia	Tenderness	Peacfulness	Joy	Power	Tension	Sadness
Multivariate FCN6 Baseline	6.295	4.117	6.609	5.372	6.631	8.601	7.950	7.757	3.333
Ensemble FCN6	6.376	4.1586	8.627	5.574	6.738	9.458	8.414	7.216	3.485
Ensemble FCN6 DOSR	6.161	3.850	6.156	5.117	7.057	7.569	7.681	6.712	3.528
Ensemble FCN6 DOSR percentile	6.509	3.832	6.161	5.804	7.200	8.559	7.833	8.120	3.040

Table 5.5: Multivariate FCN vs Ensemble FCN vs Ensemble SV FCN oversampling vs Ensemble SV FCN oversampling percentile

Model	RMSE	R^2
Multivariate FCN6 Baseline	6.272	0.292
Ensemble FCN6	6.672	0.213
Ensemble FCN6 DOSR	5.981	0.348
Ensemble FCN6 DOSR percentile	6.223	0.286

Table 5.6: Overall RMSE and R^2 for FCN6 DOSR

In table 5.5 we are able to observe that the Ensemble FCN6 DOSR, which oversamples each data point under the threshold equally, outperforms both Multivariate FCN6 Baseline and Ensemble FCN6 across the board, except for the emotion scores Peacefulness and Sadness, which only performs slightly inferior. Even the overall RMSE and R^2 statistic show improved values with a RMSE of 5.98 and R^2 statistic of 0.348. (Table 5.6)

The ensemble model, Ensemble FCN6 DOSR percentile, consisting of single-variate models deploying the DOSR percentile approach, where the oversampling amount is determined based on the percentile range of the kernel density estimation values, shows inferior performance compared to the Ensemble FCN6 DOSR model. However, it still outperforms the Ensemble FCN6 model across most emotion scores. Despite this improvement, it remains insufficient to surpass the performance of the Multivariate FCN Baseline model.

5.3 Density Based Oversampling For Multivariate Regression (DOMR)

As evident from Table 5.2, it becomes apparent that multivariate regression models may outperform the ensemble models. Therefore, a multivariate approach of the DOSR, the Density Based Oversampling For Multivariate Regression (DOMR) proposed in section 4.4.2, is implemented.

Model	Wonder	Trascendence	Nostalgia	Tenderness	Peacefulness	Joy	Power	Tension	Sadness
Multivariate Baseline	6.295	4.117	6.609	5.372	6.631	8.601	7.95	7.757	3.333
DOMR	6.022	3.848	6.227	**4.891**	7.092	**7.821**	**6.865**	7.365	3.433
DOMR percentile	**5.698**	3.725	**6.057**	5.134	6.651	8.229	7.385	6.985	3.455
DOMR no-pool	6.362	**3.684**	6.795	5.225	**6.541**	8.549	7.871	**6.856**	**3.259**

Table 5.7: R^2 of FCN6 Baseline/FCN6 DOMR/FCN6 DOMR percentile/FCN6 DOMR no-pool

Model	RMSE	R^2
Baseline	6.272	0.292
DOMR	5.952	0.355
DOMR percentile	**5.924**	**0.366**
DOMR no-pool	6.126	0.332

Table 5.8: Average RMSE & R^2 over all emotion scores: FCN6 Baseline/FCN6 DOMR /FCN6 DOMR percentile/FCN6 DOMR no-pool

The optimal "threshold (t)" parameter for the DOMR model has been identified as 0.6, meaning that data points with average kernel density estimation values below 0.6 are selected for oversampling. This method increases the training sample size from 625 to 710. According to Table 5.7, the RMSE performance for each individual emotion score in the DOMR model demonstrates improvement compared to the baseline, except for 'Sadness', which shows a slight decrease in performance. The average metrics of the DOMR model also demonstrate improvement over the Multivariate Baseline, with an R^2 statistic of 0.355 and an RMSE of 5.952.

For the DOMR percentile model, with a threshold of 0.6 the total training sample size is increased to 751. Further, we can observe in table 5.8, that the DOMR percentile model outperforms the DOMR model with a R^2 of 0.366 and RMSE of 5.924.

Finally, the DOMR no-pool approach is implemented. Instead of pooling (averaging) the kernel density estimation value of the scores, this method includes every data point as soon as any of the KDE values falls below the threshold. The training set with oversampled data, with the parameter "threshold" set to 0.3, yields a sample size of 1020. The optimal number of data transformation chains is found to be 1. While this model does not surpass the performance of the other two approaches, it still demonstrates an improvement over the baseline, as detailed in Table 5.8.

5.4 Data Transformation

Class-preserving elementary transformations, introduced in section 4.4.3, are applied to each data point (song) in a chain at random to create new samples. The transformed data is subsequently added to the dataset, effectively doubling the available data to 1250 samples. Randomly selected from a pool of audio transformation techniques—including pitch shifting, time stretching, MP3 compression, background noise addition, time shifting, low-pass filtering, gain adjustment, and compression (see section 4.4.3)—the transformations are applied sequentially, without repeating any previously applied technique. Four configurations are implemented each differentiating on the number of transformations applied to the data.

Model	**RMSE**	**R^2**
Baseline	6.272	0.292
DT 1x	6.224	0.295
DT 2x	6.151	0.313
DT 4x	**6.030**	**0.366**
DT 8x	7.294	0.047

Table 5.9: Average RMSE & R^2 over all emotion scores: FCN6 model with chained Data transformations

The average metrics in table 5.9 reveal that all data transformation configurations, except 8x lead to an improvement of the model. DT 1x outputs similar metrics as the baseline model, but if we take a look at the model with higher transformation chains, the performance improves. The model with four data transformation chains performs the best with a RMSE of 6.03 and R^2 statistic of 0.366. A decrease in RMSE can also be observed across all emotions scores, except Nostalgia, Tenderness and Peacefulness, which only perform slightly inferior than the baseline model (see table 5.10). Finally, the DT 8x model, which includes songs that were transformed eight times, has decreased performance metrics. This decline could be attributed to excessive alteration of the audio data resulting from chaining eight transformations consecutively.

Model	**Wonder**	**Transcendence**	**Nostalgia**	**Tenderness**	**Peacefulness**	**Joy**	**Power**	**Tension**	**Sadness**
Baseline	6.295	4.117	6.609	5.372	6.631	8.601	7.95	7.757	3.333
DT 1x	6.266	4.041	6.420	4.883	6.857	8.924	8.116	6.741	3.773
DT 2x	6.258	4.120	5.990	5.392	7.440	8.526	7.854	6.109	3.671
DT 4x	5.488	3.971	6.173	5.488	7.824	8.524	7.610	6.038	3.155
DT 8x	7.139	5.077	8.168	6.242	9.138	9.244	8.066	8.793	3.779

Table 5.10: R^2: FCN6 model with chained data transformation per emotion scale

5.5 Density Based Oversampling For Multivariate Regression With Data Transformation (DOMR+)

The newly proposed Density Based Oversampling For Multivariate Regression With Data Transformation (DOMR+, section 4.4.4), combines the previously implemented DOMR and Data Transformation approaches. Instead of oversampling raw data, the transformed data will be oversampled. Same as for the DOMR model, three approaches introduced in section 4.4.4 are implemented: DOMR+, DOMR+ Percentile and DOMR+ No-Pool.

The parameters threshold (t) and number of transformation chains are fine tuned for each model. They are optimized in the parameter space between 0.3 and 0.7 for tolerance (t) and 1, 2, 4 and 8 chains of transformations.

For the DOMR+ model with a 'threshold (t)' set at 0.6, the training set's sample size increases to 710. The most effective number of data transformation chains for this model is identified as 2, leading to an improved performance with an RMSE of 5.906 and an R^2 of 0.378, thus surpassing the baseline model.

Regarding the DOMR+ Percentile model, the optimal parameters are a 'threshold (t)' of 0.7 and a single data transformation chain. This configuration results in a training sample size of 853. The model achieves superior results compared to the baseline, with a lower RMSE of 5.75 and a higher R^2 of 0.395.

Finally, the DOMR+ No-Pool model, with its 'threshold (t)' set to 0.3 and using two data transformation chains, also shows promising performance. It achieves an RMSE of 5.907 and an R^2 of 0.375. While its performance is comparable to the DOMR+ model, it does not outperform the DOMR+ Percentile approach.

Model	**RMSE**	**R^2**
FCN6 Baseline	6.272	0.292
DOMR+	5.906	0.378
DOMR+ percentile	**5.75**	**0.395**
DOMR+ no-pool	5.907	0.375

Table 5.11: Average RMSE & R^2 over all emotion scores: FCN6 model with DOMR+

Model	**Wonder**	**Transcendence**	**Nostalgia**	**Tenderness**	**Peacefulness**	**Joy**	**Power**	**Tension**	**Sadness**
Baseline	6.296	4.118	6.610	5.373	6.632	8.602	7.951	7.758	3.333
DOMR+	**5.946**	**3.637**	**5.659**	4.771	6.656	8.702	7.906	6.570	3.312
DOMR+ percentile	5.999	**3.438**	5.676	4.792	**6.419**	**7.918**	**7.097**	7.112	3.363
DOMR+ no-pool	6.194	4.073	6.054	**4.499**	7.057	8.475	7.135	**6.459**	**3.224**

Table 5.12: R^2 per emotion scale : FCN6 model with DOMR +

5.6 Comparing The Performance

The results in table 5.13 reveal several insights into the performance of the methods. The Ensemble FCN6 L1 & DenseLoss model stands out as the only approach that fails to surpass the baseline model. However, it can be observed in table 5.2 that single variable models performance can be improved by DenseLoss, but not enough to surpass the multivariate baseline model.

Moreover, both proposed oversampling methods, DOSR and DOMR, demonstrate superior performance over the baseline model in terms of RMSE and R^2 statistic. Mirroring the way the multivariate baseline model surpasses the ensemble single-variable model, the multivariate DOMR model outperforms the single-variable DOSR model in R^2 and across all emotion labels, except for 'Nostalgia'.

Interestingly, the data transformation methods yield results comparable to the DOMR method, with both achieving an R^2 of 0.366. However, their slightly higher RMSE may be attributed to the addition of every piece of data in the training set.

Most notably, the data transformation-integrated DOMR, or DOMR+, outshines all other approaches. It beats the baseline with an R^2 of 0.395, an increase of 0.1, and an RMSE of 5.75, an improvement of 0.877.

Models	**RMSE**	**R^2**
FCN6 Baseline	6.272	0.292
Ensemble FCN6 L1 & DenseLoss	6.356	0.277
Ensemble FCN6 DOSR	5.981	0.348
FCN6 DT 4x	6.030	0.366
FCN6 DOMR percentile	5.924	0.366
FCN6 DOMR+ percentile	**5.75**	**0.395**

Table 5.13: Model comparison

5.7 Dependency Of Training Set Composition

Although we are able to observe that model performance can be improved by addressing the data imbalance through various combinations of oversampling and data transformation methods, we need to validate whether or not the models are dependent on the train-test composition due to the imbalances in the dataset. Similarly, the proposed methods to overcome the imbalances are expected to be dependent on the train-test composition, especially regarding the set hyperparameters. For this, the 10-fold Cross-Validation will be used, which is described more in detail in the section 4.5 in the chapter methodology.

5.7.1 Cross-Validation

The dataset has been randomly split into 10 equal sized folds. One fold will be used to test the performance of the model while the rest will be used to train the models. Meaning each configuration introduced in the earlier chapters will be trained ten times with different fold compositions. Then the RMSE and R^2 statistic of each model will be evaluated.

First, we can see in table 5.14, that the performance of the base model (Mutlivariate FCN6) achieves a mean R^2 of 0.2706, a maximum value of 0.3662 and a minimum value of 0.1665. First implications that the models performance is dependent on the train-test composition can be seen through the rather high gap between the maximum and minimum values.

Data Transformation

FCN6 models are trained with additional transformed data with different transformation chain configurations, as already mentioned in the previous chapter. Looking at the performance in table 5.14 and 5.15, it can be observed that the models using data transformations do not improve the models' average R^2 statistics. Table 5.14 shows the models' R^2 statistics, indicating that while data transformations do not enhance the average R^2, but there is a noticeable improvement in the maximum R^2 values, especially the 1x and 2x chained transformations. Contrary to the results in 5.4, it can be seen that the model trained with the additional 4x data transformation does not improve the performance in terms of R^2 or RMSE metrics. As expected, the model trained with 8x data transformations shows the poorest performance, similar to the results in 5.4.

	Baseline	**DT 1x**	**DT 2x**	**DT 4x**	**DT 8x**
Mean	0.2706	0.2563	0.2680	0.2217	-0.0716
Median	0.2746	0.2778	0.2661	0.2293	-0.0619
Max	0.3662	0.3784	0.4395	0.3434	0.1463
Min	0.1665	0.1080	0.1537	0.0383	-0.3643

Table 5.14: R^2 statistics results from validation: FCN6 Baseline and FCN6 with chained data transformations

	Baseline	**DT 1x**	**DT 2x**	**DT 4x**	**DT 8x**
Mean	6.2930	6.3397	6.2706	6.4730	7.6604
Median	6.3663	6.3848	6.3118	6.5952	7.5253
Max	6.6026	6.9203	6.7541	7.1684	8.5693
Min	5.6987	5.7303	5.5754	5.6205	6.4647

Table 5.15: RMSE results from validation: FCN6 Baseline and FCN6 with chained data transformations

Density Based Oversampling For Multivariate Regression (DOMR)

Different to the findings in section 5.3, the models utilizing DOMR performed poorly, as can be seen in table 5.16 and 5.17. Mean R^2 and RMSE do not show any improvements regardless of the threshold. The only noteworthy observation is that while the maximum R^2 value for DOMR 0.6 saw some improvement, the minimum value deteriorated.

	Baseline	**DOMR t=0.5**	**DOMR t=0.6**	**DOMR t=0.7**
Mean	0.2706	0.2181	0.2502	0.2332
Median	0.2746	0.2266	0.2592	0.2245
Max	0.3662	0.3531	0.4094	0.3537
Min	0.1665	0.0545	-0.0016	0.1075

Table 5.16: R^2 statistics results from validation: FCN6 DOMR with different thresholds

	Baseline	DOMR t=0.5	DOMR t=0.6	DOMR t=0.7
Mean	6.2930	6.5293	6.3227	6.4191
Median	6.3663	6.5534	6.2739	6.4358
Max	6.6026	7.2463	7.2730	6.8530
Min	5.6987	5.6752	5.7033	5.8260

Table 5.17: RMSE results from validation: FCN6 DOMR with different thresholds

DOMR Percentile

Models utilizing the DOMR percentile approach show an even less favorable performance, with metrics declining across the board compared to the baseline, as illustrated in Tables 5.18 and 5.19.

	Baseline	DOMR perc. t=0.5	DOMR perc. t=0.6	DOMR perc. t=0.7
Mean	0.2706	0.2305	0.2186	0.2068
Median	0.2746	0.2669	0.2200	0.2044
Max	0.3662	0.3587	0.3562	0.3518
Min	0.1665	0.0953	0.0750	0.0434

Table 5.18: R^2 statistics results from validation: FCN6 DOMR percentile with different thresholds

	Baseline	DOMR perc. t=0.5	DOMR perc. t=0.6	DOMR perc. t=0.7
Mean	6.2930	6.4913	6.4564	6.5067
Median	6.3663	6.5805	6.5251	6.5589
Max	6.6026	6.9928	6.8758	6.9432
Min	5.6987	5.9884	5.8921	5.8498

Table 5.19: RMSE results from validation: FCN6 DOMR percentile with different thresholds

Density Based Oversampling For Multivariate Regression With Data Transformation (DOMR+)

For the DOMR+ approach, different thresholds with data transformation chains of 1x, 2x and 4x have been trained. 8x chained transformations have been omitted as they did not deliver any improvements in any of the previously trained models.

In general, across all configurations, the average performance of the DOMR+ model does not surpass that of the baseline model in terms of both R^2 and RMSE. Again, similar to the DOMR 0.6 in table 5.16, for the DOMR+ model with the threshold set to 0.6, the maximum value for the R^2 increased, regardless of data transformation configuration (see table 5.20 and 5.21). In terms of the RMSE no improvement can be observed as indicated in tables 5.22 and 5.23.

	Baseline	DOMR+ t=0.5 dt=1	DOMR+ t=0.5 dt=2	DOMR+ t=0.5 dt=4	DOMR+ t=0.6 dt=1	DOMR+ t=0.6 dt=2	DOMR+ t=0.6 dt=4
Mean	0.2706	0.2581	0.2112	0.2450	0.2258	0.2390	0.2194
Median	0.2746	0.2660	0.2197	0.2408	0.2488	0.2370	0.2143
Max	0.3662	0.3730	0.3553	0.4066	0.3684	0.4042	0.4293
Min	0.1665	0.1356	0.0490	0.1528	0.0380	0.1112	0.0956

Table 5.20: R^2 statistics from validation: FCN6 DOMR+ with different thresholds

	Baseline	DOMR+ t=0.7 dt=1	DOMR+ t=0.7 dt=2	DOMR+ t=0.7 dt=4
Mean	0.2706	0.2438	0.1953	0.2078
Median	0.2746	0.2503	0.1859	0.2355
Max	0.3662	0.3699	0.3472	0.3590
Min	0.1665	0.0464	0.0893	0.0505

Table 5.21: R^2 statistics from validation FCN6 DOMR+ with threshold 0.7 and data transformation chains

	Baseline	DOMR+ t=0.5 dt=1	DOMR+ t=0.5 dt=2	DOMR+ t=0.5 dt=4	DOMR+ t=0.6 dt=1	DOMR+ t=0.6 dt=2	DOMR+ t=0.6 dt=4
Mean	6.2930	6.3236	6.5454	6.3778	**6.4474**	6.3797	6.5039
Median	6.3663	6.4620	6.6573	6.5362	**6.4817**	6.4304	6.5590
Max	6.6026	6.6258	7.1189	6.7982	**7.2910**	6.7949	6.9556
Min	5.6987	5.6449	5.7798	5.5838	**5.6939**	5.5589	6.0086

Table 5.22: RMSE results from validation: FCN6 DOMR+ with different thresholds

	Baseline	DOMR+ t=0.7 dt=1x	DOMR+ t=0.7 dt=2x	DOMR+ t=0.7 dt=4x
Mean	6.2930	6.3216	6.5884	6.5684
Median	6.3663	6.3027	6.5807	6.4906
Max	6.6026	6.7765	7.1616	7.4793
Min	5.6987	5.7748	5.8036	5.6396

Table 5.23: RMSE results from validation: FCN6 DOMR + with threshold 0.7 and data transformation chains

DOMR+ Percentile

Similar to DOMR+, models using the DOMR+ percentile approach (introduced in subsection 4.4.4), are trained with different thresholds and data transformation chains of 1x, 2x and 4x. The trained models also fail to show improvements in mean R^2 and RMSE as depicted in tables 5.24 and 5.25, with the exception of the model with a 0.8 threshold and 1x data transformation, though this improvement is only minimal. Another notable observation is that the maximum value for R^2 using the DOMR+ percentile approach is higher than the baseline model across the board, except for the model with the threshold of 0.8 and 4x data transformation.

	Baseline	DOMR+ perc. t=0.6 dt=1	DOMR+ perc. t=0.6 dt=2	DOMR+ perc. t=0.6 dt=4	DOMR+ perc. t=0.7 dt=1	DOMR+ perc. t=0.7 dt=2	DOMR+ perc. t=0.7 dt=4
Mean	0.2706	0.2319	0.2615	0.2032	0.2505	0.2167	0.2136
Median	0.2746	0.2398	0.2543	0.2180	0.2803	0.2259	0.2030
Max	0.3662	0.4003	0.3984	0.3758	0.3791	0.3717	0.4042
Min	0.1665	0.0464	0.1737	0.0472	0.0616	0.0441	0.0411

Table 5.24: R^2 statistics results from validation: FCN6 DOMR+ percentile with different thresholds

	Baseline	DOMR+ perc. t=0.8 dt=1	DOMR+ perc. t=0.8 dt=2	DOMR+ perc. t=0.8 dt=4
Mean	0.2706	0.2721	0.2448	0.1622
Median	0.2746	0.2788	0.2552	0.1312
Max	0.3662	0.3877	0.4605	0.3007
Min	0.1665	0.1505	0.0546	0.0437

Table 5.25: R^2 statistics results from validation: FCN6 DOMR + percentile with threshold 0.8 and data transformation chains

	Baseline	DOMR+ perc. 0.6_1	DOMR+ perc. 0.6_2	DOMR+ perc. 0.6_4	DOMR+ perc. 0.7_1	DOMR+ perc. 0.7_2	DOMR+ perc. 0.7_4
Mean	6.2930	6.3893	6.3287	6.5469	6.3225	6.4781	6.5153
Median	6.3663	6.4233	6.3569	6.5425	6.4381	6.5353	6.5390
Max	6.6026	7.1020	7.0352	7.0165	6.6781	7.0935	6.8693
Min	5.6987	5.6660	5.7353	6.0521	5.7223	5.8719	5.9142

Table 5.26: RMSE from validation FCN6 DOMR+ percentile with different threshold

	Baseline	DOMR+ perc. 0.8_1	DOMR+ perc. 0.8_2	DOMR+ perc. 0.8_4
Mean	6.2930	6.2843	6.3483	6.7315
Median	6.3663	6.3901	6.2785	6.7283
Max	6.6026	6.7214	6.9918	7.4032
Min	5.6987	5.5662	5.8874	5.9791

Table 5.27: RMSE after CV fold DOMR + percentile with threshold 0.8 and data transformation chains

	Baseline	DT 2x	DOMR	DOMR perc.	DOMR+ DT 1x	DOMR+ DT 2x	DOMR+ DT 4x	DOMR+ perc. DT 1x	DOMR+ perc. DT 2x	DOMR+ perc. DT 4x
Mean	6.2930	6.2706	6.3227	6.4564	6.3216	6.3797	6.3778	6.2843	6.3483	6.5469
Median	6.3663	6.3118	6.2739	6.5251	6.3027	6.4304	6.5362	6.3901	6.2785	6.5425
Max	6.6026	6.7541	7.2730	6.8758	6.7765	6.7949	6.7982	6.7214	6.9918	7.0165
Min	5.6987	5.5754	5.7033	5.8921	5.7748	5.5589	5.5838	5.5662	5.8874	6.0521

Table 5.28: Best performing models per RMSE for CV-fold

Implications and Limitations

The results of the models above suggest that the implemented approaches do not improve model performance on average. Generally, there is a huge discrepancy between the maximum and minimum value with a difference of 0.3086 on average throughout the models. We can see this more clearly in table 5.29, which depcits the RMSE and R ^{2} of each fold of the best performing model, FCN6 DOMR+ percentile with a threshold of 0.7 and 1x data transformation. Notably, both the R^2 and RMSE values exhibit significant fluctuations across different folds. All of this could suggest that the performance of the model and the effectiveness of the given approach is highly dependent on the train-test composition.

The results may also reflect the data scarcity issue of the data set. As the data is extremely right-skewed for several emotion labels, the chosen training folds may not represent all emotion ranges very well. As certain emotion ranges are rarer, they may have not been selected for some of the folds. Figure 5.1 depicts the kernel density estimation values of each of the 10 folds. Especially for the labels "Wonder", "Transcendence", "Nostalgia", "Peacefulness" and "Joy", we can observe some differences between the folds distribution. This issue will be further addressed in the next section.

Fold	R^2	RMSE
1	0.1686	6.7214
2	0.2870	6.5118
3	0.3116	5.8255
4	0.2607	5.5662
5	0.1505	5.9534
6	0.2418	6.4214
7	0.2707	6.5698
8	0.3389	6.3589
9	0.3877	6.3516
10	0.3034	6.5627

Table 5.29: FCN6 DOMR+ percentile, threshold: 0.7, 1x DT

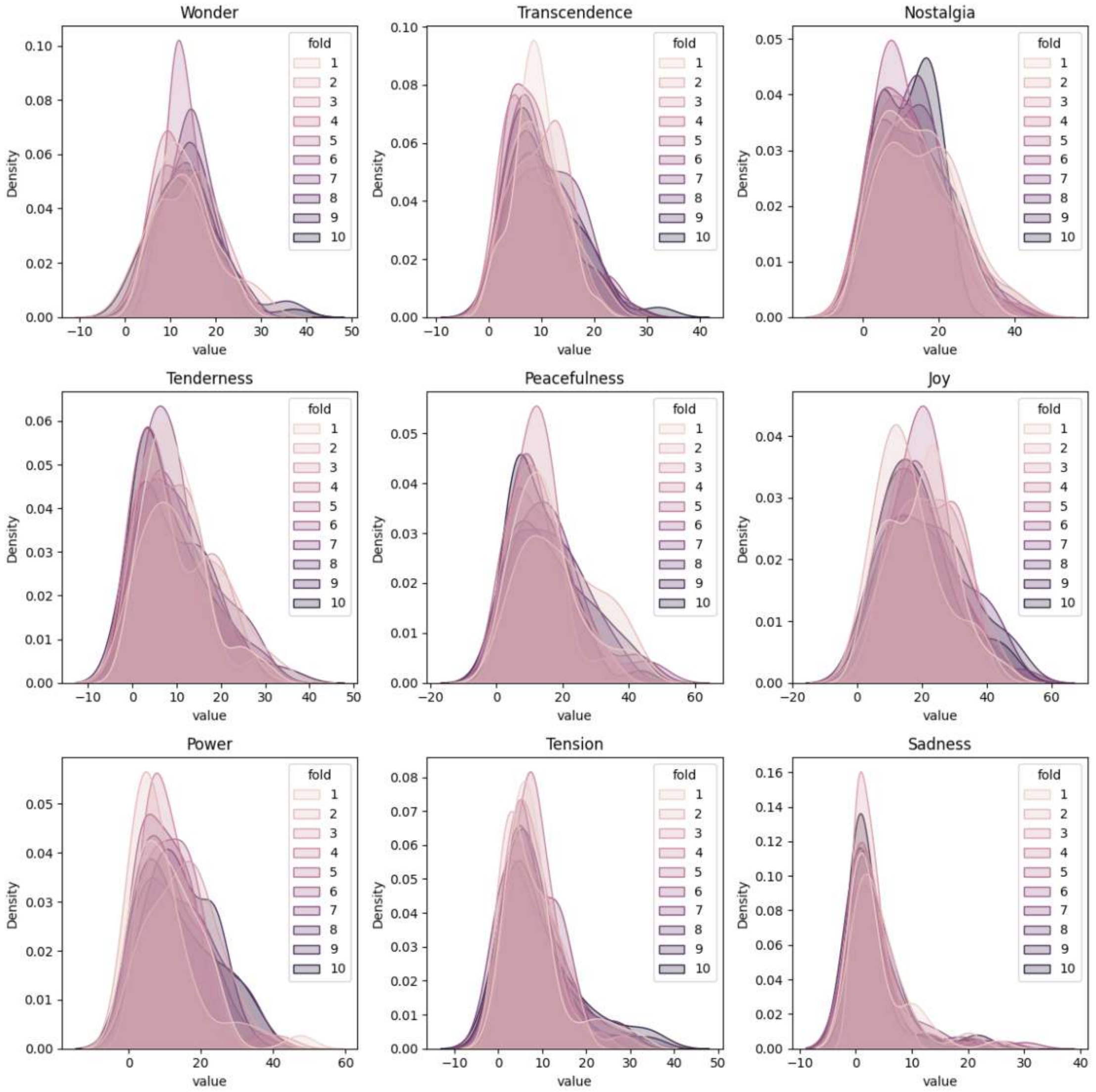


Figure 5.1: CV: Kernel density estimation plots for each fold and emotion label

5.7.2 Stratified Cross-Validation using KDE-values

To counteract the issue of sampling folds of unequal distributions, a stratified cross-validation inspired approach is proposed, described in section 4.5 in chapter Methodology.

Data transformation

Models with different data transformation configurations have been trained under stratified fold sampling. Different to the non-stratified fold sampling, the results depicted in table 5.30 show that the DT 1x and DT 2x models exhibit superior performance compared to the baseline models on average. Further, an increase for the maximum R^2 value can be seen for all data transformation configurations, except for DT 8x — consistent with earlier findings in this thesis. This pattern is also reflected in RMSE performance, although with a marginal improvement (see table 5.31).

	Baseline	DT 1x	DT 2x	DT 4x	DT 8x
Mean	0.2397	0.2508	0.2527	0.2146	-0.1521
Median	0.2378	0.2659	0.2530	0.2149	-0.1468
Max	0.2947	0.3188	0.3024	0.3413	0.1784
Min	0.1656	0.1729	0.1587	0.1111	-0.4461

Table 5.30: R^2 statistics results from stratified validation: FCN6 Baseline and Data transformation

	Baseline	DT 1x	DT 2x	DT 4x	DT 8x
Mean	6.4346	6.3608	6.3557	6.4836	7.8432
Median	6.4187	6.3585	6.3703	6.6637	7.9232
Max	6.9655	7.0083	6.6499	6.9137	8.7804
Min	5.9962	5.9130	5.9660	5.7676	6.5186

Table 5.31: RMSE results from stratified validation: FCN6 Baseline and Data transformation

Density Based Oversampling For Multivariate Regression (DOMR)

The tables 5.32 and 5.33 show, different to the results of the non-stratified fold sampling and consistent with the finding in section 5.3, that the models utilising DOMR have improved performance in regards to the R^2 and RMSE. The thresholds 0.5 and 0.7 show an improved performance in the average performance and also in the maximum value of R^2. Again, the RMSE improvement is only minimal.

	Baseline	DOMR t=0.5	DOMR t=0.6	DOMR t=0.7
Mean	0.2397	0.2430	0.2122	0.2539
Median	0.2378	0.2550	0.2134	0.2436
Max	0.2947	0.2989	0.2930	0.3500
Min	0.1656	0.1499	0.0753	0.2035

Table 5.32: R^2 statistics results from stratified validation: FCN6 Baseline and DOMR

	Baseline	DOMR t=0.5	DOMR t=0.6	DOMR t=0.7
Mean	6.4346	6.4343	6.5440	6.3459
Median	6.4187	6.3746	6.4996	6.3291
Max	6.9655	6.8729	7.1522	6.8176
Min	5.9962	5.9813	6.0754	5.8168

Table 5.33: RMSE results from stratified validation: FCN6 Baseline and DOMR

DOMR Percentile

Similarly to the results obtained from the non-stratified folds method, models trained using the DOMR percentile approach perform poorly and exhibit inferior performance compared to the baseline model in terms of R^2 and RMSE, as depicted in Tables 5.34 and 5.35.

	Baseline	DOMR perc. t=0.5	DOMR perc. t=0.6	DOMR perc. t=0.7
Mean	0.2397	0.2020	0.1742	0.1815
Median	0.2378	0.2073	0.1912	0.1708
Max	0.2947	0.2733	0.2804	0.3098
Min	0.1656	0.1191	0.0311	0.0829

Table 5.34: R^2 statistics from stratified validation: FCN6 Baseline and DOMR percentile

	Baseline	DOMR perc. t=0.5	DOMR perc. t=0.6	DOMR perc. t=0.7
Mean	6.4346	6.5903	6.6559	6.5888
Median	6.4187	6.5756	6.5840	6.6256
Max	6.9655	7.2895	7.3794	7.3330
Min	5.9962	6.2462	6.0705	5.7659

Table 5.35: RMSE from stratified validation: FCN6 Baseline and DOMR percentile

Density Based Oversampling For Multivariate Regression With Data Transformation (DOMR+)

Again, for the DOMR+ approach models with varying thresholds and data transformation chain configurations have been trained. Different to the results of the non-stratified fold sampling method, models with a threshold of 0.5 and data transformation chain of 1x and 4x exhibits superior performance compared to the baseline model on average, as illustrated in tables 5.36, 5.37, 5.38 and 5.39. Specifically, the DOMR+ model with a threshold of 0.5 and 4x data transformation (DOMR+ t=0.5 dt=4) outperforms the baseline model in terms of mean, median, maximum, and minimum values for both R^2 and RMSE. This result is also consistent with the findings in section 5.5 .

	Baseline	DOMR+ t=0.5 dt=1	DOMR+ t=0.5 dt=2	DOMR+ t=0.5 dt=4	DOMR+ t=0.6 dt=1	DOMR+ t=0.6 dt=2	DOMR+ t=0.6 dt=4
Mean	0.2397	0.2579	0.2388	0.2459	0.2365	0.1878	0.1863
Median	0.2378	0.2658	0.2416	0.2525	0.2297	0.2162	0.2090
Max	0.2947	0.3712	0.3156	0.3047	0.3238	0.2671	0.2862
Min	0.1656	0.1543	0.1510	0.1956	0.1512	0.0738	0.0456

Table 5.36: R^2 statistics results from stratified validation: FCN6 DOMR+ with different threshold

	Baseline	DOMR+ t=0.7 dt=1	DOMR+ t=0.7 dt=2	DOMR+ t=0.7 dt=4
Mean	0.2397	0.2149	0.2060	0.1475
Median	0.2378	0.2044	0.2201	0.1729
Max	0.2947	0.3546	0.3304	0.2489
Min	0.1656	0.1046	0.0513	-0.1246

Table 5.37: R^2 statistics results from stratified validation: FCN6 DOMR + with threshold 0.7 and data transformation chains

	Baseline	DOMR+ t=0.5 dt=1	DOMR+ t=0.5 dt=2	DOMR+ t=0.5 dt=4	DOMR+ t=0.6 dt=1	DOMR+ t=0.6 dt=2	DOMR+ t=0.6 dt=4
Mean	6.4346	6.3477	6.4441	6.3920	6.4301	6.6108	6.5905
Median	6.4187	6.2211	6.4150	6.3045	6.4663	6.5889	6.5254
Max	6.9655	7.1883	6.8034	6.7810	7.0846	7.2848	7.0922
Min	5.9962	5.9043	6.0616	5.9919	5.9058	6.0689	6.0968

Table 5.38: RMSE results from stratified validation: FCN6 DOMR+ with different threshold

	Baseline	DOMR+ t=0.7 dt=1	DOMR+ t=0.7 dt=2	DOMR+ t=0.7 dt=4
Mean	6.4346	6.4680	6.5372	6.7933
Median	6.4187	6.4834	6.5380	6.7121
Max	6.9655	7.0861	7.4439	7.5620
Min	5.9962	5.7142	5.9264	6.2182

Table 5.39: RMSE esults from stratified validation: FCN6 DOMR + with threshold 0.7 and data transformation chains

DOMR+ Percentile

We observe similar trends in the performance of models trained using the DOMR+ percentile approach compared to those trained using DOMR+. Consistent with the results in section 4.4.4, tables 5.24, 5.41, 5.42 and 5.43 show that the models with various DOMR+ configurations beat the performance of the baseline models. Models with the thresholds 0.6 and 0.8 with 1x, 2x and 1x data transformation chains respectively perform the best in terms of R^2 and RMSE.

	Baseline	DOMR+ perc. t=0.6 dt=1	DOMR+ perc. t=0.6 dt=2	DOMR+ perc. t=0.6 dt=4	DOMR+ perc. t=0.7 dt=1	DOMR+ perc. t=0.7 dt=2	DOMR+ perc. t=0.7 dt=4
Mean	0.2397	0.2413	0.2459	0.2279	0.2159	0.2136	0.1608
Median	0.2378	0.2605	0.2544	0.2171	0.2098	0.2134	0.1716
Max	0.2947	0.3098	0.3333	0.2677	0.3174	0.3212	0.3215
Min	0.1656	0.1440	0.1507	0.1984	0.1125	0.1328	0.0219

Table 5.40: R^2 statistics aesults from stratified validation: FCN6 DOMR+ percentile with different threshold

	Baseline	DOMR+ perc. t=0.8 dt=1	DOMR+ perc. t=0.8 dt=2	DOMR+ perc. t=0.8 dt=4
Mean	0.2397	0.2532	0.2344	0.2149
Median	0.2378	0.2455	0.2310	0.2011
Max	0.2947	0.3251	0.3559	0.3299
Min	0.1656	0.1795	0.1340	0.0868

Table 5.41: R^2 statistics results from stratified validation: FCN6 DOMR + percentile with threshold 0.8 and data transformation chains

	Baseline	DOMR+ perc. t=0.6 dt=1	DOMR+ perc. t=0.6 dt=2	DOMR+ perc. t=0.6 dt=4	DOMR+ perc. t=0.7 dt=1	DOMR+ perc. t=0.7 dt=2	DOMR+ perc. t=0.7 dt=4
Mean	6.4346	6.4224	6.3873	6.4385	6.4957	6.5301	6.6861
Median	6.4187	6.4466	6.4821	6.4145	.4684	6.5154	6.7812
Max	6.9655	6.7519	6.8402	6.8087	6.9971	7.0099	7.1474
Min	5.9962	5.9919	5.9182	6.1264	5.9245	6.0200	6.0199

Table 5.42: RMSE esults from stratified validation: FCN6 DOMR+ percentile with different threshold

	Baseline	DOMR+ perc. t=0.8 dt=1	DOMR+ perc. t=0.8 dt=2	DOMR+ perc. t=0.8 dt=4
Mean	6.4346	6.3539	6.4486	6.5026
Median	6.4187	6.3728	6.3823	6.5090
Max	6.9655	6.8595	7.2862	7.4882
Min	5.9962	5.8793	6.0666	5.8323

Table 5.43: RMSE esults from stratified validation: FCN6 DOMR + percentile with threshold 0.8 and data transformation chains

Implications and Limitations

The stratified fold approach is implemented in order to create a more equal distribution throughout the folds, thereby mitigating the issue of certain emotion ranges not being present in some folds due to their rarity. As intended, in comparison to the non-stratified folds (see graph 5.1), the implementation of the stratified cross-validation inspired approach is able to achieve a more uniform distribution across all folds, as the graph 5.2, which illustrates the kernel density estimation values of the folds, suggest. Therefore, it can be argued that the results of this experiment could provide stronger evidence for the effectiveness of the proposed approaches.

While the models evaluated with the non-stratified fold approach do not deliver an improvement in performance in terms of average R^2, except the FCN6 DOMR+ percentile with threshold 0.8 + DT 1x (table 5.25), multiple models evaluated with the stratified fold approach demonstrate improved performance.

Every proposed approach appears to improve performance, except for the DOMR percentile model, which resulted in a inferior performance regardless of the set parameters.

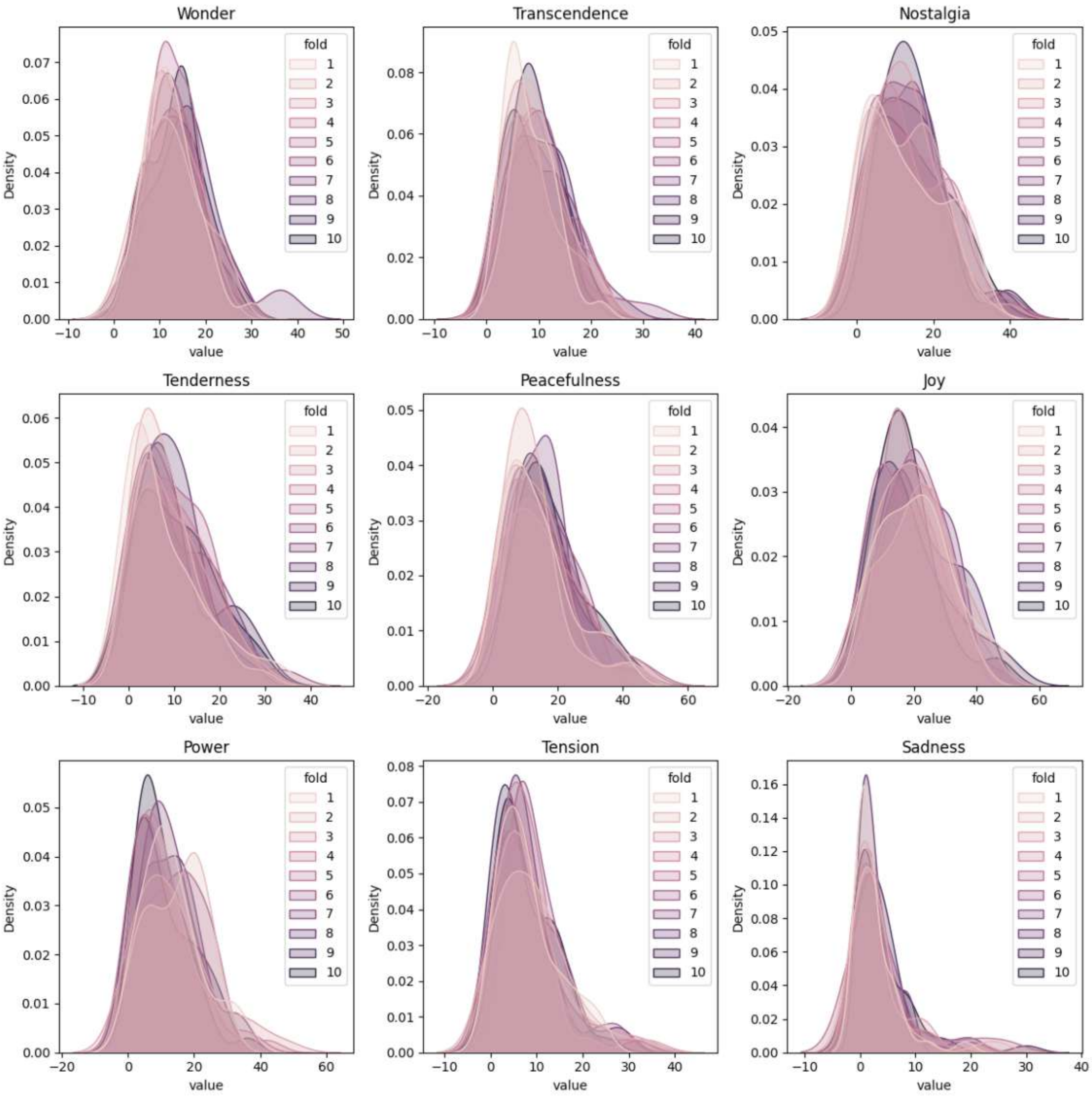


Figure 5.2: Stratified CV: Kernel density estimation plots for each fold and emotion label

5.8 T-Test

A dependent pairwise t-test is conducted to statistically compare the performance of the proposed approaches to the baseline model and to evaluate whether the mean difference in metrics between the two models is statistically significant. For this analysis, the R^2 statistics from the stratified validations are used. Table 5.44 and 5.45 depict the t-statistics and p-value of the pairwise t-tests for R^2 and RMSE of the promising models. T statistic is calculated as:

$$t = \frac{\bar{X}_D - \mu_0}{s_D/\sqrt{n}}$$

R^2

The results of table 5.44 suggest that there is no significant variability (p-value > 0.05) in the R^2 statistic across most of the proposed models. We can examine the t-statistic to find out which of the two groups' mean is higher or lower. A negative t-statistics value indicates that the mean R^2 of the proposed model is higher than the R^2 of the baseline model (see t-test formula). Negative t-statistics can be seen in table 5.44 for the Models deploying 1x and 2x chained data transformation (DT 1x & DT 2x), DOMR with threshold set to 0.5 and 0.7 (DOMR 0.5 & DOMR 0.7), DOMR+ with a threshold set to 0.5 and employing 1x and 4x chained data transformations (DOMR+ 0.5 1x & DOMR 0.5 4x), as well as the DOMR+ percentile with a threshold set to 0.6 and employing 1x and 2x chained data transformations (DOMR+ perc. 0.6 1x & DOMR perc. 0.6 2x), and with a threshold set to 0.8 and employing 1x chained data transformation (DOMR+ perc. 0.8 1x). These findings align with the results presented in Section 5.7.2, but it is important to note that despite the negative t-statistics, these values are not statistically significant (p-value < 0.05).

Model	t-stat	p-value	significance
DT 1x	-0.498	0.623	FALSE
DT 2x	-0.694	0.4961	FALSE
DT 4x	0.909	0.375	FALSE
DOMR t=0.5	-0.163	0.871	FALSE
DOMR t=0.6	1.087	0.291	FALSE
DOMR t=0.7	-0.770	0.451	FALSE
DOMR perc. t=0.5	1.753	0.096	FALSE
DOMR perc. t=0.6	2.244	0.037	TRUE
DOMR perc t=0.7	2.122	0.047	TRUE
DOMR+ t=0.5 dt=1	-0.694	0.496	FALSE
DOMR+ t=0.5 dt=2	0.048	0.961	FALSE
DOMR+ t=0.5 dt=4	-0.359	0.723	FALSE
DOMR+ perc. t=0.6 dt=1	-0.069	0.945	FALSE
DOMR+ perc. t=0.6 dt=2	-0.274	0.787	FALSE
DOMR+ perc. t=0.6 dt=4	0.779	0.445	FALSE
DOMR+ perc. t=0.8 dt=1	-0.659	0.517	FALSE
DOMR+ perc. t=0.8 dt=2	0.227	0.822	FALSE
DOMR+ perc. t=0.8 dt=4	0.827	0.4189	FALSE

Table 5.44: Paired T-test: R^2 stratified validation

RMSE

Table 5.45 shows that for each negative t-statistics value in table 5.44, we can observe a positive t-statistics value, indicating that the mean RMSE of the proposed model is higher than the RMSE of the baseline model. However, similar to the R^2 analysis, these differences in RMSE are not statistically significant.

Model	t-stat	p-value	significance
DT 1x	0.5420	0.5945	FALSE
DT 2x	0.6688	0.5121	FALSE
DT 4x	-0.3052	0.7637	FALSE
DOMR t=0.5	0.0029	0.9977	FALSE
DOMR t=0.6	-0.7277	0.4762	FALSE
DOMR t=0.7	0.6531	0.5219	FALSE
DOMR perc. t=0.5	-1.0734	0.2973	FALSE
DOMR perc. t=0.6	-1.3211	0.2030	FALSE
DOMR perc. t=0.7	-0.8704	0.3955	FALSE
DOMR+ t=0.5 dt=1	0.5417	0.5947	FALSE
DOMR+ t=0.5 dt=2	-0.0779	0.9388	FALSE
DOMR+ t=0.5 dt=4	0.3321	0.7437	FALSE
DOMR+ perc. t=0.6 dt=1	0.0982	0.9228	FALSE
DOMR+ perc. t=0.6 dt=2	0.3465	0.7330	FALSE
DOMR+ perc. t=0.6 dt=4	-0.0328	0.9742	FALSE
DOMR+ perc. t=0.8 dt=1	0.5582	0.5836	FALSE
DOMR+ perc. t=0.8 dt=2	-0.0891	0.9300	FALSE
DOMR+ perc. t=0.8 dt=4	-0.3699	0.7158	FALSE

Table 5.45: Paired T-test: RMSE stratified validation

Model	R2	RMSE
Baseline	0.239	6.434
DT 1x	0.250	6.361
DT 2x	0.252	6.356
DT4x	0.214	6.484
DOMR t=0.5	0.243	6.434
DOMR t=0.6	0.174	6.655
DOMR t=0.7	0.253	6.346
DOMR perc. t=0.5	0.201	6.590
DOMR perc. t=0.6	0.174	6.656
DOMR perc. t=0.7	0.181	6.589
DOMR+ t=0.5 dt=1	0.257	6.348
DOMR+ t=0.5 dt=2	0.238	6.444
DOMR+ t=0.5 dt=4	0.245	6.392
DOMR+ perc. t=0.6 dt=1	0.241	6.422
DOMR+ perc. t=0.6 dt=2	0.245	6.387
DOMR+ perc. t=0.6 dt=4	0.227	6.439
DOMR+ perc. t=0.8 dt=1	0.253	6.354
DOMR+ perc. t=0.8 dt=2	0.234	6.449
DOMR+ perc. t=0.8 dt=4	0.214	6.503

Table 5.46: Stratified fold mean R2 and RMSE

5.8.1 Seeding

To further empirically assess the efficacy of the proposed approaches in enhancing model performance, chosen models are trained 10 times using the stratified fold methods, respectively outlined in section 5.7.1 and 5.7.2. Each training cycle (Cross-Validation) employs a different seed for random sampling. This process will be repeated 10 times. Doing so results in a total of 100 trained models (10 seeds × 10 folds). For comparative analysis, we employ the baseline FCN6 model alongside the best-performing models identified in sections 5.7.1 and 5.7.2: the FCN6 model with 1x chained data transformation (DT 1x), the FCN6 model with 2x chained data transformation (DT 2x), the FCN6 model applying DOMR with a threshold set to 0.7 (DOMR t=0.7), the FCN6 DOMR+ with a threshold of 0.5 and 4x chained data transformation (DOMR+ perc. t=0.5 dt=4), the FCN6 DOMR+ percentile with a threshold of 0.8 and 1x chained data transformation (DOMR+ perc. t=0.8 dt=1).

Results

When we only look at the mean values of the R^2 and RMSE in table 5.47, we can see that only the DT 1x model improved in performance.

Further, the t-test results on R^2 in table 5.48 show, that the differences between the baseline and each model, except DOMR+ t=0.5 dt=4, are statistically not significant. Although significant, the DOMR+ t=0.5 dt=4 model's t-statistic is positive, indicating that these models have a decrease in performance against the baseline model on average regarding the R^2. Similar results can be seen for the DT 2x and DOMR t=0.7 models, with positive t-statistics. Similarly, it can be said for these models for the RMSE, as seen in table 5.49, that the values are statistically not significant and decreased in performance, except for DT 2x, but only with pretty much negligible t-statistics value of 0.057. For the DT 1x model a negative t-statistic value of -1.145 for R^2 and positive t-statistics of 1.515 for RMSE can be seen in table 5.48 and 5.49, but again these values are not statistically significant.

The DOMR+ perc. t=0.8 dt=1 model by far delivers the most promising results, with a negative t-statistic of -1.726 for R^2. While the p-value of 0.087 is not below the significance threshold value (0.05), it is quite close to it. Similarly, we can see an increase in RMSE performance with a t-statistic of 3.046, which is also statistically significant with a p-value of 0.002.

Mean metrics (10 seeds, 10-stratified fold CV)		
Model	**R2**	**RMSE**
Baseline	0.258	6.377
DT 1x	0.267	6.315
DT 2x	0.254	6.376
DOMR t=0.7	0.251	6.382
DOMR+ t=0.5 dt=4	0.244	6.412
DOMR+ t=0.8 DT=1x	0.251	6.312

Table 5.47: Mean R^2 and RMSE results of 10-stratified fold CV (10 seeds)

Model	t-stat	p-value	significant
DT 1x	-1.145	0.254	FALSE
DT 2x	0.828	0.409	FALSE
DOMR t=0.7	1.297	0.197	FALSE
DOMR+ t=0.5 dt=4	2.459	0.015	FALSE
DOMR+ perc. t=0.8 dt=1	-1.726	0.087	FALSE

Table 5.48: 10 seeds x 10 folds, Paired T-test for R^2 (FCN6 model, stratified validation)

Model	t-stat	p-value	significant
DT 1x	1.515	0.132	FALSE
DT 2x	0.057	0.954	FALSE
DOMR t=0.7	-0.201	0.840	FALSE
DOMR+ t=0.5 dt=4	-1.40	0.163	FALSE
DOMR+ perc. t=0.8 dt=1	3.046	0.002	TRUE

Table 5.49: 10 seeds x 10 folds, Paired T-test for RMSE (FCN6 model, stratified validation)

CHAPTER 6

Discussion

In this chapter, we will discuss the implications of the results obtained from the experiments. We will examine what these findings suggest and identify any inherent limitations and future research directions.

6.1 Implications

This thesis contributes to the ongoing research that aims to improve the accuracy and robustness of music mood tagging models through explorations of different oversampling techniques for imbalanced audio dataset, specifically for discrete or continuous variables, which could pave the way for future research in this area.

Furthermore, this thesis can help further develop the EMMA dataset by highlighting areas where additional annotations or data collection efforts may be needed. By identifying limitations in the current dataset, researchers can prioritize future efforts to ensure the usability of the dataset. This includes improving the diversity and coverage of emotion scores within the dataset, as well as refining the annotation process to better align with the requirements of machine learning tasks. This in turn could also help the usability for other use cases of the dataset.

Upon reviewing the results presented in Chapter 5, it becomes apparent that none of the proposed approaches were able to significantly improve the performance of the models. However, certain approaches demonstrate promising potential, as evidenced by the findings in tables 5.48 and 5.49. Particularly noteworthy is the DOMR+ percentile model with a threshold of 0.8 and a data transformation chain of 1, which exhibit an increased average R^2 performance compared to the baseline model. Its p-value of 0.08, which is close to the significance threshold, suggests potential significance. It was able to achieve an increase in performance in regards to the average RMSE, which was also statistically significant with a p-value of 0.002. Further investigation with adjustments to the model parameters and performance measurements on different datasets could validate its effectiveness.

6.2 Limitations

6.2.1 Dataset size & Imbalance

Arguably, the biggest drawback of the EMMA dataset is its extremely small sample size compared to other datasets in the field. Already mentioned in section 2.2, existing datasets can reach a sample size up to a million music samples (e.g. Million Song Dataset [BMEWL11]). While the EMMA dataset arguably offers superior data quality compared to its counterparts, its sample size significantly diminishes its potential. Furthermore, in addition to its small sample size, the right skewness observed in certain emotion scores, such as Sadness and Tension as seen in chapter 3, further exacerbates the issue. This skewness results in certain data ranges being even less represented in an already data-scarce dataset. Similarly, due to the small sample size and right skewness, the performance of the model can get very dependent on the train-test composition. This dependency is evident in the contrasting results obtained from the non-stratified and stratified fold methods, as discussed in sections 5.7.1 and 5.7.2.

Consequently, the DOMR approaches, which oversample according to the rarity of the data point, have limited data to work with. For instance, the smaller the pool of data for a certain range, the less data will be oversampled from it, making it difficult to balance out the skewness. Additionally, the diversity of the oversampled data will also be limited. This can be also observed in the metrics, as the R^2 improves, the RMSE hardly improves. Generally, oversampling, regardless of the method, will most likely not improve the model significantly due to this constraint. Further, even if effective oversampling methods exist, the samples size cannot be enlarged enough if the original sample amount is too low to begin with. We can see in section 5.3, how a rather tolerant threshold point of 0.6 increases the sample size from 625 to only 710.

6.2.2 Annotation

The spread of the emotion scores appears to have low variance (see figure 4.4), which may present challenges for effectively capturing the nuances of music mood tagging. Therefore, the current annotation methodology may not be optimally suited for this kind of task. Further refinement of the annotation/scoring process, while considering machine learning tasks, could make the dataset more suitable for music mood tagging tasks.

6.2.3 Model selection & Evaluation

The choice of model architecture and hyperparameters in this thesis may not have been optimal. While a variety of models was explored, there may exist other architectures or configurations that could yield improved performance. The various models implemented in this thesis only had marginal or no differences to the performance. This could partly be attributed to the limited availability of data, which may have hindered the models' ability to learn the underlying patterns in the audio data effectively.

Moreover, the evaluation metrics used in this thesis, such as R^2 and RMSE, provide insights into model performance, but may not capture all aspects of music mood tagging accuracy. Further, future research could consider additional evaluation metrics or qualitative assessments to provide a more comprehensive understanding of model performance. R^2 and RMSE cannot capture whether or not predicted scores reflect the emotions perceived by human listeners.

6.3 Future Work

While this thesis provides first insights into the performance, problems and approaches for music mood tagging using the EMMA dataset, there are still numerous areas open to be researched. In this section, we outline potential directions for further investigation.

EMMA dataset

As discussed in the limitations section, the small sample size of the EMMA dataset poses a significant challenge in training a music mood tagging model. Therefore, one direction for future work could be the expansion of the EMMA dataset. Especially, rare score ranges need to be covered more by the data. Further, adjusting the dataset's coverage of a broader range of score variations could substantially improve model performance.

Combination of Datasets

Another direction for future research could be to find an appropriate taxonomy to merge the EMMA dataset with other similar datasets introduced in the related works to overcome the data scarcity problem. Datasets such as MTG-Jamendo, the Million Song Dataset, and MagnaTagATune have been extensively used in the field of Music Information Retrieval (MIR). By leveraging the strengths of multiple datasets and exploring multimodal features, it may be possible to improve mood based recommender systems and the understanding of the emotional content of music.

Oversampling, Data transformation & DOMR

The DOMR approaches implemented in this thesis have shown limitations, but also have only been applied to a rather small dataset. Therefore, training models using the DOMR+ perc. t=0.8 dt=1 approach from section 5.8.1, which yielded optimistic results to larger datasets, could provide a more comprehensive understanding of their effectiveness in enhancing model performance. Furthermore, the development of more refined techniques other than the kernel density estimation for selecting data to be used in oversampling has the potential to significantly improve the efficacy of DOMR approaches.

Models & transfer learning

While this thesis prioritized addressing the challenge of data scarcity over exploring model architecture, it is worth noting that alternative architectures may increase model performance. Additionally, while transfer learning did not enhance model performance in this thesis, it is possible that the limited data size played a role in this outcome. Further exploration of methods to improve model performance through transfer learning remains a direction for future research.

Further, as already mentioned, to capture whether the predictions are consistent with the emotions perceived by listeners, qualitative assessments could be considered in the future. For instance, this could involve conducting questionnaires with listeners or designing interactive games where participants judge the predictions made by the model based on their perceived emotions. These qualitative approaches could help ascertain whether the model's predictions align with the emotions perceived by listeners and provide valuable insights into the real-world applicability of the proposed approaches. The information gained from these assessments could also be embedded into the training process.

Multimodel data

Another promising area for future research is the integration of multimodal data for music mood tagging. While this thesis focused primarily on audio features, incorporating additional modalities such as lyrics, user metadata, or audiovisual information could provide richer context and improve the accuracy of mood prediction. Pyrovolakis et al. [PTS22] were able to achieve promising results on the MoodyLyrics dataset [cM17] by combining lyrics and audio data by constructing a uniform system. They utilized BERT for extracting features from lyrics and a CNN for processing audio data. The output from both models was then combined, and a neural network with two fully-connected layers was employed to predict the labels.

However, this approach may encounter potential limitations if applied to the EMMA dataset. For instance, the dataset includes classical music without lyrics, as well as songs with non-English lyrics. These factors could pose challenges in effectively integrating multimodal data and may require further research.

Data transformations

Similar to DOMR approaches, chained Data transformation methods yielded results that could potentially improve model performances. However, as observed in Table 5.7.1 and 5.7.2, transformations with more than 2x tend to lead to inferior results. This suggests a potential area for further investigation. Exploring the optimal number and types of transformations that enhance model performance without diminishing it could be a valuable research direction. Data transformation approaches, while not statistically significant in this thesis, have shown potential to have a positive effect on model performance as seen in table 5.44 and 5.48 and warrant further investigation as a potential research direction.

CHAPTER 7

Conclusion

In conclusion, this thesis has explored through extensive experimentation and analysis various methodologies and approaches for music mood tagging tasks using the EMMA dataset and several key findings have emerged.

Firstly, the limitations of the EMMA dataset, including its small sample size and the right skewness of certain emotion scores, present significant challenges for effectively training music mood tagging models. Additionally, there is not a lot of research done in oversampling or data augmentation methods tailored for audio regression tasks. While certain approaches, such as Density Oversampling for Multivariate Regression (DOMR/DOMR+) and data transformation techniques, have shown promise, none have yielded statistically significant improvements in model performance when trained and tested on the EMMA dataset. That being said, DOMR+ showed signs of potential and therefore warrants further exploration. Particularly, applying DOMR+ to other larger datasets could provide valuable insights into its effectiveness. Similarly, data transformation methods also demonstrated glimpses of improvement in model performance, suggesting their relevance for enhancing the effectiveness of the EMMA dataset in music mood tagging tasks.

Particularly for training machine learning models using the EMMA dataset, the mentioned shortcomings need to be addressed. Researchers could explore approaches to enhance the EMMA dataset, such as increasing its sample size and refining its annotation process. By addressing these weaknesses, the dataset could be more suitable for model training and be able to predict moods that are potentially consistent with human perception.

In summary, while this thesis has shed light on various challenges in music mood tagging using the EMMA dataset, further research into various areas is still needed to overcome the limitations.

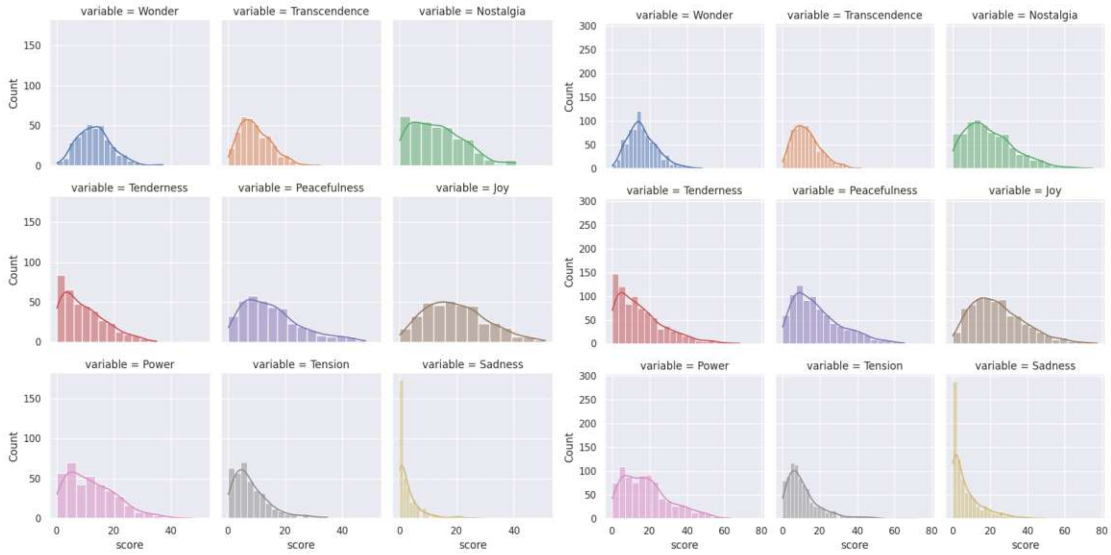
CHAPTER 8

Update: Performance on the Extended EMMA Dataset

Near the completion of this thesis, an extended version of the EMMA dataset has been released, incorporating additional data and possibly addressing some of the limitations identified in the previous version.

8.1 New Dataset

The dataset size increased from 364 to 823 songs, covering more score ranges than the previous version as depicted in figure 8.1. However, the problem of right skewness still exists. When examining the skewness values presented in Table 8.1, and referring to Bulmer's interpretation of skewness [Bul79], already introduced in section 4.3, none of the emotions are approximately symmetric, Tenderness, Tension and Sadness are highly skewed.



(a) Original

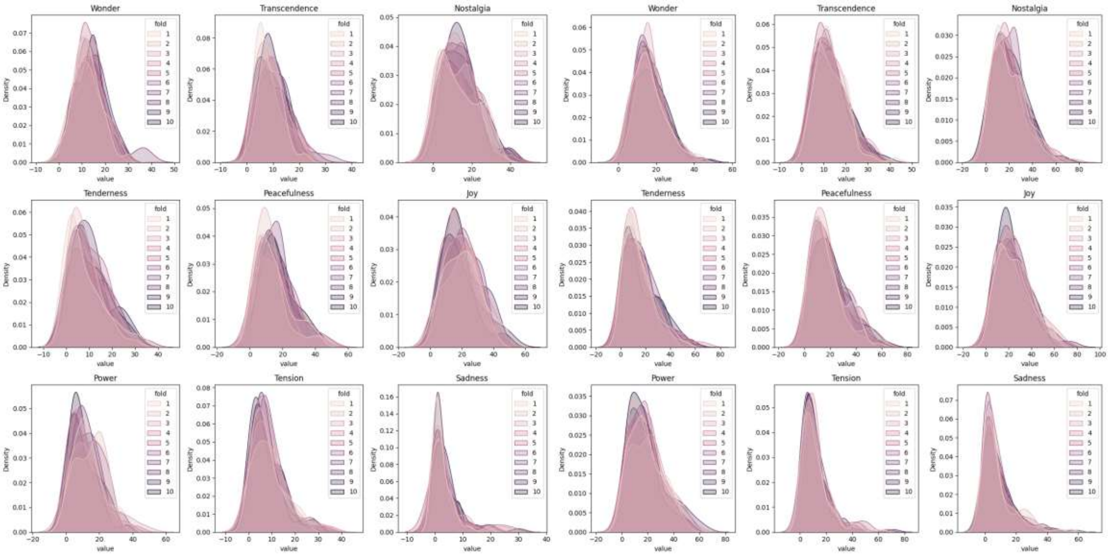
(b) Extended

Figure 8.1: Emma dataset distribution

	Skewness
Wonder	0.715
Transcendence	0.757
Nostalgia	0.839
Tenderness	1.096
Peacefulness	0.949
Joy	0.757
Power	0.804
Tension	1.992
Sadness	1.880

Table 8.1: Skewness of emotions scores (Extended EMMA dataset)

Due to the higher number of samples, the KDE based stratified folds of the extended dataset exhibit higher similarity to each other, as it can be seen by the comparisons shown in figure 8.2.



(a) Original

(b) Extended

Figure 8.2: Emma dataset fold distribution

8.2 Model Performance with the Extended EMMA Dataset

Similar to section 5.8.1, models were trained using 10-stratified fold cross-validation with 10 different seeds, resulting in a total of 100 models. For this, both the baseline FCN6 model and the best performing model from Section 5.8.1, DOMR+ percentile with parameters t=0.8 and DT=1x, were trained. The mean, median, minimum and maximum metrics are compared and a dependent pairwise t-test between the baseline model and DOMR+ percentile will be conducted.

Baseline		DOMR+ perc. t=0.8 DT=1x	
R^2	**RMSE**	**R^2**	**RMSE**
0.258	9.374	0.297	9.102
0.262	9.392	0.275	9.392
0.248	9.433	0.288	9.433
0.264	9.370	0.298	9.370
0.275	9.300	0.289	9.300
0.246	9.466	0.291	9.466
0.237	9.513	0.274	9.513
0.248	9.446	0.276	9.446
0.238	9.528	0.286	9.528
0.273	9.302	0.299	9.302

Table 8.2: Baseline FCN6 vs DOMR+ perc. t = 0.8 DT = 1x FCN6 model, Stratified fold mean R^2 and RMSE for extended EMMA dataset

Model	**Baseline**		**DOMR+ perc. t=0.8 DT=1x**	
Metric	**R^2**	**RMSE**	**R^2**	**RMSE**
Mean	0.255	9.412	0.287	9.186
Median	0.2592	9.371	0.288	9.182
Min	0.021	8.786	0.1676	8.373
Max	0.374	10.467	0.396	10.105

Table 8.3: Baseline vs DOMR+ perc. t=0.8 DT = 1x, Statistics over 100 models (10 seeds x 10 stratified folds)

R^2			
Model	**t-stat**	**p-value**	**significant**
DOMR+ perc. t=0.8 DT = 1x	-5.536	0.00000025	TRUE

Table 8.4: T-test results for R^2 against the Baseline model (extended Dataset)

RMSE			
Model	**t-stat**	**p-value**	**significant**
DOMR+ perc. t=0.8 DT = 1x	5.691	0.00000013	TRUE

Table 8.5: T-test results for RMSE against the Baseline model (extended Dataset)

Looking at the average R^2 and RMSE values of the Baseline and the DOMR+ in table 8.2, we cannot see a similar performance in R^2 and an notable increase in RMSE. This increase in RMSE could be attributed to the dataset more than doubling in size. With a broader range of score values covered by the data, the test dataset naturally includes score ranges that are more challenging to predict, thereby leading to higher prediction errors.

When comparing the DOMR+ t=0.8 DT=1x to the baseline model, as seen in table 8.3, we can see an increase in mean and median performance for the RMSE and R^2. This was not the case for the models trained on the original dataset as seen in table 5.47. Additionally, the DOMR+ model shows a higher minimum performance for both RMSE and R^2, as well as a higher maximum performance for R^2. Similarly, an increase in performance can be observed for both the minimum and maximum RMSE values.

For the t-test results shown in table 8.4, we observe a t-statistic of -5.536. The negative sign indicates that the sample mean of the DOMR+ percentile is higher than the sample mean of the baseline models, suggesting an increase in R^2 performance. Furthermore, the p-value is below the threshold of 0.05, indicating statistical significance.

Similarly, in table 8.5, the t-statistic has a positive sign with 5.691. This suggests that the sample mean of the DOMR+ percentile is lower than the sample mean of the baseline models, indicating an improvement in RMSE performance. Again, the p-value is below the threshold of 0.05, confirming statistical significance.

This contrasts with the results obtained from the original dataset, as depicted in table 5.48 and table 5.48, where the difference in R^2 was not statistically significant.

8.3 Implication, Future Work and Conclusion

With an increased volume of annotated music excerpts, the dataset offers a more comprehensive representation of emotional responses to music. This opens up research for even deeper analysis and exploration of approaches, including those already examined in this thesis, such as oversampling and data augmentation. With more data available, these methods can be more viable and effective.

Already, the DOMR+ percentile model, though not specifically fine-tuned for the extended dataset, has exhibited promising performance improvements, positioning it as a potential candidate for further investigation and application in mood tagging tasks.

In conclusion, the extension of the EMMA dataset is already a promising improvement in score distribution and sample size. While there is still room for improvement in these areas, the extended dataset already indicates progress in the right direction.

While this thesis did not delve deeply into the extended EMMA dataset, its preliminary findings already reflect the potential of the DOMR+ approach to enhance model performance in music mood tagging tasks.

List of Figures

List of Tables

Bibliography

[BMEWL11] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[Bul79] M.G. Bulmer. *Principles of Statistics*. Dover Books on Mathematics Series. Dover Publications, 1979.

[BWT+19] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019.

[CFS16] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks, 2016.

[CFSC17] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*, 2017.

[CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[cM17] Erion Çano and Maurizio Morisio. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, ISMSI '17, page 118–124, New York, NY, USA, 2017. Association for Computing Machinery.

[CYWC15] Yu-An Chen, Yi-Hsuan Yang, Ju-Chiang Wang, and Homer H. Chen. The amg1608 dataset for music emotion recognition. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 693–697, 2015.

[DHP+18] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music mood detection based on audio and lyrics with deep neural net, 2018.

[DS13] Sander Dieleman and Benjamin Schrauwen. Multiscale approaches to music audio feature learning. In *International Society for Music Information Retrieval Conference*, 2013.

[FGHC18] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[Gab16] Alf Gabrielsson. 215The Relationship between Musical Structure and Perceived Expression. In *The Oxford Handbook of Music Psychology*. Oxford University Press, 01 2016.

[GCY12] Di Guan, Xiaoou Chen, and Deshun Yang. Music emotion regression based on multi-modal features. In *Proc. International Symposium on Computer Music Modeling and Retrieval*, pages 70–77, 2012.

[HF20] NA HE and Sam Ferguson. Multi-view neural networks for raw audio-based music emotion recognition. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 168–172, 2020.

[HRDH09] Byeong-jun Han, Seungmin Rho, Roger B Dannenberg, and Eenjun Hwang. Smers: Music emotion recognition using support vector regression. In *ISMIR*, pages 651–656, 2009.

[JTC+23] Iver Jordal, Araik Tamazian, Emmanouil Theofanis Chourdakis, Céline Angonin, Tushar Dhyani, askskro, Nikolay Karpov, Omer Sarioglu, BakerBunker, kvilouras, Enis Berk Çoban, Florian Mirus, Jeong-Yoon Lee, Kwanghee Choi, MarvinLvn, SolomidHero, and Tanel Alumäe. iver56/audiomentations: v0.33.0, August 2023.

[Kam20] Firuz Kamalov. Kernel density estimation based sampling for imbalanced class distribution. *Information Sciences*, 512:1192–1201, 2020.

[KKP+06] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*, 30(1):25–36, 2006.

[KNM+20] Filip Korzeniowski, Oriol Nieto, Matthew McCallum, Minz Won, Sergio Oramas, and Erik Schmidt. Mood classification using listening data, 2020.

[KSM+10] Youngmoo Kim, Erik Schmidt, Raymond Migneco, Brandon Morton, Patrick Richardson, Jeffrey Scott, Jacquelin Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, 01 2010.

[LB98] Yann LeCun and Yoshua Bengio. *Convolutional networks for images, speech, and time series*, page 255–258. MIT Press, Cambridge, MA, USA, 1998.

[LG20] Beici Liang and Minwei Gu. Music genre classification using transfer learning. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 392–393, 2020.

[LLZ06] Lie Lu, D. Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):5–18, 2006.

[LWM+09] Edith Law, Kris West, Michael Mandel, Mert Bay, and J. Downie. Evaluation of algorithms using games: The case of music tagging. pages 387–392, 01 2009.

[MKO+22] Matthew C. McCallum, Filip Korzeniowski, Sergio Oramas, Fabien Gouyon, and Andreas F. Ehmann. Supervised and unsupervised learning of audio representations for music understanding, 2022.

[Moo12] B.C.J. Moore. *An Introduction to the Psychology of Hearing*. Emerald, 2012.

[MP19] Rémi Mignot and Geoffroy 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*, Dec 2019.

[Pic] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

[PNP+18] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale, 2018.

[PS19] Jordi Pons and Xavier Serra. musicnn: Pre-trained convolutional neural networks for music audio tagging, 2019.

[PTS22] Konstantinos Pyrovolakis, Paraskevi Tzouveli, and Giorgos Stamou. Multimodal song mood detection with deep learning. *Sensors*, 22(3), 2022.

[RK04] Bhavani Raskutti and Adam Kowalczyk. Extreme re-balancing for svms: A case study. *SIGKDD Explor. Newsl.*, 6(1):60–69, jun 2004.

[Rus80] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[SB14] Jan Schlüter and Sebastian Böck. Improved musical onset detection with convolutional neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6979–6983, 2014.

[SCD+20] Rajib Sarkar, Sombuddha Choudhury, Saikat Dutta, Aneek Roy, and Sanjoy Kumar Saha. Recognition of emotion in music based on deep convolutional neural network. *Multimedia Tools and Applications*, 79:765–783, 2020.

[SCS+13] Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, page 1–6, New York, NY, USA, 2013. Association for Computing Machinery.

[SEL11] Pasi Saari, Tuomas Eerola, and Olivier Lartillot. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1802–1812, 2011.

[SG15] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*, pages 121–126, 2015.

[SHJR21] Cunningham Stuart, Ridley Harrison, Weinel Jonathan, and Picking Richard. Supervised machine learning for audio emotion recognition. *Personal and Ubiquitous Computing*, 2021.

[SKD+21] Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110:1–25, 08 2021.

[SKrM+13] Tara N. Sainath, Brian Kingsbury, Abdel rahman Mohamed, George E. Dahl, George Saon, Hagen Soltau, Tomas Beran, Aleksandr Y. Aravkin, and Bhuvana Ramabhadran. Improvements to deep convolutional neural networks for lvcsr, 2013.

[SSMK11] Jacquelin Speck, Erik Schmidt, Brandon Morton, and Youngmoo Kim. A comparative study of collaborative vs. traditional musical mood annotation. pages 549–554, 01 2011.

[SVJ+24] Hannah Strauss, Julia Vigl, Peer-Ole Jacobsen, Martin Bayer, Francesca Talamini, Wolfgang Vigl, Eva Zangerle, and Marcel Zentner. The emotion-to-music mapping atlas (emma): A systematically organized online database of emotionally evocative music excerpts. 2024.

[USG14] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary detection in music structure analysis using convolutional neural networks. In *ISMIR*, pages 417–422, 2014.

[WC05] Kris West and Stephen Cox. Finding an optimal segmentation for audio genre classification. In *ISMIR*, pages 680–685, 2005.

[WCNS20] Minz Won, Sanghyuk Chun, Oriol Nieto, and Xavier Serrc. Data-driven harmonic filters for audio representation learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 536–540, 2020.

[WFBS20] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. Evaluation of cnn-based automatic music tagging models, 2020.

[WLW$^+$21] Luyu Wang, Pauline Luc, Yan Wu, Adrià Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, João Carreira, and Aäron van den Oord. Towards learning universal audio representations. *CoRR*, abs/2111.12124, 2021.

[YLSC07] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen. Music emotion classification: A regression approach. In *2007 IEEE International Conference on Multimedia and Expo*, pages 208–211, 2007.

[YLSC08] yi-hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer Chen. A regression approach to music emotion recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16:448 – 457, 03 2008.

[ZGS08] Marcel Zentner, Didier Grandjean, and Klaus Scherer. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion (Washington, D.C.)*, 8:494–521, 09 2008.