Copenhagen Business School

M.Sc. Business Administration and Data Science

**Music Emotion Recognition with the MTG Jamendo Dataset**

**Report for Exam Project**  
**Data Mining, Machine Learning, and Deep Learning**

Authors: Marcel Wilhelm Frey (158295)

Richard Dudzak (158482)

Yile Huang (158299)

Julian Felix Thier (158304)

No of pages: xx

No of characters: xxxxx

Hand-in date: May 19, 2023

[1. Introduction 3](#_Toc134537176)

[2. Related Work and Research Questions 3](#_Toc134537177)

[2.1 MER Overview 3](#_Toc134537178)

[2.2 Challenges in MER 4](#_Toc134537179)

[2.3 Machine Learning Models in MER 4](#_Toc134537180)

[Conceptual Framework 5](#_Toc134537181)

[Methodology 5](#_Toc134537182)

[Results 5](#_Toc134537183)

[Discussion 5](#_Toc134537184)

**List of Figures**

Figure 1: xxx xx

Figure 2: xxx xx

Figure 3: xxx xx

Figure 4: xxx xx

Figure 5: xxx xx

**List of Tables**

Table 1: xxx xx

Table 2: xxx xx

Table 3: xxx xx

# Introduction

„Emotion is the essence of music“ (Yang et al., 2018, p. 365). With music, artists are able to express feelings and induce feelings in listeners. This integral element of music led to the development to a separate discipline in machine learning research, known as music emotion recognition (MER). Machine learning models for MER attempt to automatically recognize emotion labels or intensities of certain emotions based on music-related input features. The models can be used for a wide range of applications, for example as a part of music recommendation systems, automatic music composing (Ferreira & Whitehead, 2021), music visualization or even psychotherapy (Han et al., 2022, p. 1).

# 2. Related Work and Research Questions

## 2.1 MER Overview

Music Emotion Recognition emerged to a machine learning research field with several distinct sub-areas (Han et al., 2022, p. 2 ff). Because of distinct MER approaches, emotion models, datasets and types of features, the field is diverse what makes it hard to compare findings from these different sub-areas.

First of all, there are two fundamentally different approaches to music emotion recognition: song-level MER and music emotion variation detection (MEVD). Song-level MER assigns a label to a whole song, while MEVD considers emotion as a changing process throughout pieces of music (Han et al., 2022, p. 2). For the purpose of this paper, we will focus on song-level MER.

Secondly, models proposed in the past use emotion labels based on different emotion models, resulting in fundamentally different classifiers. The applied emotion models can either be based on a categorical definition of emotions or a dimensional definition (Han et al., 2022, p. 2). This is an essential difference because it will determine whether a classification or regression model will be constructed.

Furthermore, the labels are dependent on whether the chosen model reflects emotions expressed by music or the emotional reaction induced by the music in the listener (Fan et al., 2017, p. 369; Kim et al., 2010, p. 256).

The labels in our data coincide with an established emotion model, first proposed by James William in 1884 in his book „What is emotion?“. James defines emotion as "the mental state that arises spontaneously rather than through conscious effort and is often accompanied by physiological changes." (SOURCE!!!). This implies that for this paper, we are looking at the emotions induced by music. Choosing this persepective on emotions can be regarded as reasonable, e.g. applying MER in music recommendation, the similarity of induced feelings in the listener is the more relevant input factor. Former datasets in MER were rather using other emotion models with a tendency towards models based on the emotions expressed by music (Han et al., 2022, p. 3). This adds to the novelity of the dataset discussed in this paper.

In general, the emotion labels are

Lastly, MER research differs in the types of features used for training. There are many alternatives, including symbolic music scores, lyrics and pysiological data generated from listeners (Han et al., 2022, p. 3). However, prior research suggested that audio signals and especially timbre features of the audio provide the best performance in MER when used as individual features (Barthet et al., 2012a, p. 497). The most commonly used representation of timbre features of a song is the Mel Frequency Cepstral Coefficient (MFCC) (Han et al., 2022, p. 4). MFCCs show frequencies on the mel frequency scale which adapts pitches to the human hearing perceptions (Muda et al., 2010, p. 139).

MER models can also be based on a combination of feature types, e.g. audio and lyrics (Kim et al., 2010, p. 263).

## 2.2 Machine Learning Models in MER

In recent years, the focus of the reseach in MER has been on deep learning (Han et al., 2022, p. 1). Apart from that, in the sub-field of song-level MER with categorical emotions, support vector machines (SVM) have been the most frequently used tradional machine learning model (Han et al., 2022, p. 4; Kim et al., 2010, p. 261) and have been proven to be the best performing model for MER based on audio features (Barthet et al., 2012b, p. 247). However, k-nearest neighbors (KNN), decision trees, random forests and naive bayes have also been used in the past. Within deep learning, for song-level MER with categorical emotions, CNN-based models are common (Han et al., 2022, p. 6).

In this paper, we will apply CNNs and SVMs, thus the most commonly used models for this type of classification task.

## 2.3 Motivation and Research Questions

The research field of MER faces several severe challenges. Apart from the difficulty in comaring findings implied by section 2.1, it is a main issue for MER that emotions are highly subjective. The same piece of music might induce different emotional reactions in people, and even the perception of the same person is inconsistent over time (Han et al., 2022, p. 8). This implies that it is very hard or even impossible to get emotion labels for songs that every listener would agree to, meaning that one can expect noise in the labels of datasets for MER.

Furthermore, the MER field lacks authoritative large-scale diversified datasets (Han et al., 2022, p. 8). Datasets used by researchers are often not published because of copyright restrictions (Bogdanov, Won, et al., 2019, p. 1; Han et al., 2022, p. 3). For auto-tagging, recent studies commonly use three datasets which are known for their limitations: They only offer short audio segments instead of whole tracks, unique tracks and artists are not covered well, and many recordings suffer from low audio quality (Bogdanov, Won, et al., 2019, p. 1). This incentivized Bogdanov et al. to develop a new dataset for different Music Information Retrieval tasks. The dataset is based on music available on Jamendo – a platform for royalty-free music – under creative common licenses and tags provided by the content uploaders. (Bogdanov, Won, et al., 2019, p. 1). One main novely of the dataset is the availability of songs of high-quality full length audio instead of only short song segments (Bogdanov, Won, et al., 2019, p. 1). This allows for new variations of input features for the training of the model. We will train the models MFCCs based on 30s extracts from the beginning, the middle and the end of the songs and will compare the resulting models. The MFCC approach will be benchmarked against using the mere raw audio input. Additionally, we might construct a FCN using the whole songs with varying length, or looping the songs to the same length and construct a CNN based on this. We will be testing alterations of previously proven algorithms on the MTG-Jamendo dataset. This leads us to the following research questions:

1. How do different CNN structures perform in comparison to support vector machines on the MTG-Jamendo Dataset?
2. Which of the suggested data preprocessing measures is the most adequate for these techniques?

# Methodology

## 3.1 Dataset description

* Summary of EDA results

## 3.2 Data Preprocessing

## Machine Learning Models

* CNN:
  + VGGNet used as benchmark for MTG-Jamendo (Bogdanov, Porter, et al., 2019, p. 2; Choi et al., 2016)
  + Shallow CNN better than deep CNN (Liu et al., 2017, p. 3)
* SVM

## 3.4 Evaluation Metrics

* Micro-/macro accuracy
* Others? Macro ROC-AUC and PR-AUC? (Bogdanov, Won, et al., 2019, p. 2)

# 4. Results

4.1 Result Summary

4.2 Model Complexity analysis

# 5. Discussion

5.1 Model Comparisons

5.2 (Error Analysis)

5.3 Future Work & Conclusion

# References

Barthet, M., Fazekas, G., & Sandler, M. (2012a). *Multidisciplinary Perspectives on Music Emotion Recognition: Implications for Content and Context-Based Models*.

Barthet, M., Fazekas, G., & Sandler, M. (2012b). Music Emotion Recognition: From Content to Context-Based Models. In *From Sounds to Music and Emotions: 9th International Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers* (pp. 228–252). Springer.

Bogdanov, D., Porter, A., Tovstogan, P., & Won, M. (2019). *MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo*.

Bogdanov, D., Won, M., Philip Tovstogan, Alastair Porter, & Xavier Serra. (2019). *The MTG-Jamendo Dataset for Automatic Music Tagging*.

Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016*, 805–811.

Fan, J., Tatar, K., Thorogood, M., & Pasquier, P. (2017). RANKING-BASED EMOTION RECOGNITION FOR EXPERIMENTAL MUSIC. *International Symposium on Music Information Retrieval (ISMIR)*, 368–375. http://metacreation.net/project/emusic/.

Ferreira, L. N., & Whitehead, J. (2021). *Learning to Generate Music With Sentiment*.

Han, D., Kong, Y., Han, J., & Wang, G. (2022). A survey of music emotion recognition. *Frontiers of Computer Science*, *16*(6).

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., & Turnbull, D. (2010). MUSIC EMOTION RECOGNITION: A STATE OF THE ART REVIEW. *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, 255–266.

Liu, X., Chen, Q., Wu, X., Liu, Y., & Liu, Y. (2017). *CNN based music emotion classification*.

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal of Computing*, *2*(3), 138–143.

Yang, X., Dong, Y., & Li, J. (2018). Review of data features-based music emotion recognition methods. *Multimedia Systems*, *24*(4), 365–389.

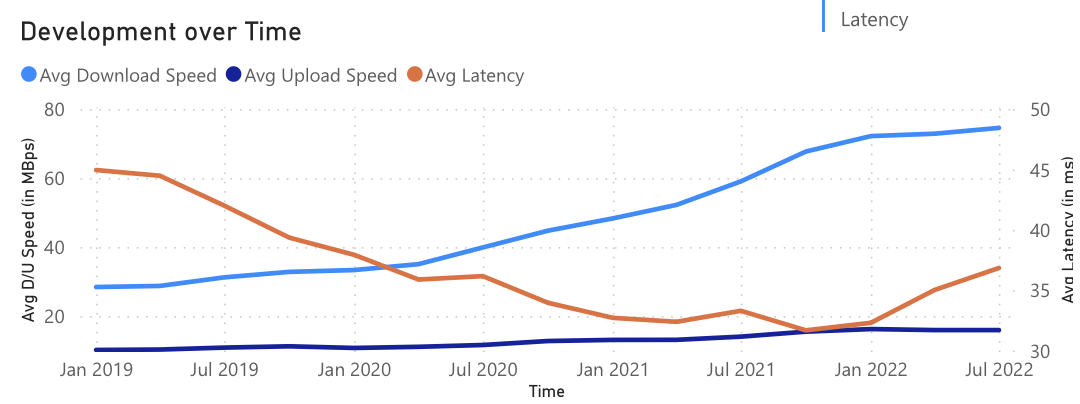


Figure 4: Layout for visualisations