2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) | 979-8-3503-7451-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSPW62465.2024.10669903

PERCEPTUAL MUSICAL FEATURES FOR INTERPRETABLE AUDIO TAGGING

Vassilis Lyberatos, Spyridon Kantarelis, Edmund Dervakos and Giorgos Stamou

National Technical University of Athens, Athens, Greece

ABSTRACT

In the age of music streaming platforms, the task of automatically tagging music audio has garnered significant attention, driving researchers to devise methods aimed at enhancing performance metrics on standard datasets. Most recent approaches rely on deep neural networks, which, despite their impressive performance, possess opacity, making it challenging to elucidate their output for a given input. While the issue of interpretability has been emphasized in other fields like medicine, it has not received attention in music-related tasks. In this study, we explored the relevance of interpretability in the context of automatic music tagging. We constructed a workflow that incorporates three different information extraction techniques: a) leveraging symbolic knowledge, b) utilizing auxiliary deep neural networks, and c) employing signal processing to extract perceptual features from audio files. These features were subsequently used to train an interpretable machine-learning model for tag prediction. We conducted experiments on two datasets, namely the MTG-Jamendo dataset and the GTZAN dataset. Our method surpassed the performance of baseline models in both tasks and, in certain instances, demonstrated competitiveness with the current state-of-the-art. We conclude that there are use cases where the deterioration in performance is outweighed by the value of interpretability.

Index Terms— Perceptual features, Explainable artificial intelligence, Music audio tagging

1. INTRODUCTION

Music audio tagging encompasses a range of tasks centered on assigning descriptive tags to music tracks, offering particular utility in the context of organizing extensive music collections. These tags encompass musically relevant information such as instruments [1], genres [2], mood [3], emotions [4, 5], harmony [6], rhythm [7], and additional metadata [8]. The majority of methods employed to address these tasks involve feature extraction and machine learning (ML). However, there has been a shift towards minimizing the feature extraction step, often limited to acquiring only Mel-frequency cepstrum coefficients (MFCCs), with a growing inclination towards end-to-end deep learning models [9]. This trend towards deep learning models brings with it increased opacity and challenges in providing explanations, particularly in scenarios where the ground truth itself can be subjective and open to interpretation [10, 11], as is the case with mood tags.

In these cases, the reliability of the training data dictates the reliability of the model, and without interpretability, it is difficult to detect flaws or biases in either data or models [12]. For music especially, where real-world data tend to be diverse, contrary to available datasets [13], it is essential to be able to detect and fix biases to make robust systems that can be applied in real-world settings [14]. As an extreme example, consider a mood tagging system that was trained only on pop and rock music. We cannot expect such a system to be generalized for all music genres, and it would not be possible to predict how it would behave if it were fed for instance classical music. Furthermore, we cannot expect an end-user to know the distribution of the training dataset, and the nuances of training ML models, thus such a system would not be applicable in practice, regardless of how well it performed on a test set. Contrarily, if the mood tagging system itself were interpretable, it would be easier for a user to understand why the system assigns specific tags to samples, making it more trustworthy and more likely to be applied in a real-world setting.

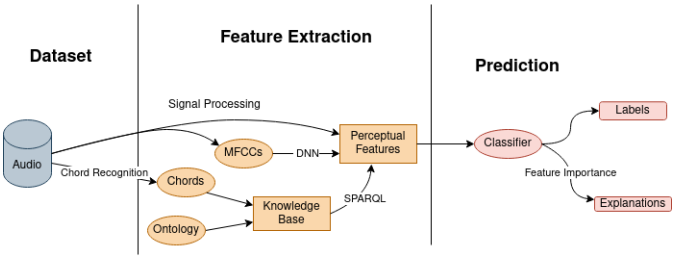


Fig. 1: Overview of the proposed pipeline

In recent years, efforts have emerged aimed at addressing the interpretability challenge within music audio tagging systems [15]. Perceptual musical features are used that are readily comprehensible to human users [16, 17, 18, 19, 20]. These features encompass characteristics imbued with both musical and acoustic significance, such as tonality, rhythmic stability, or loudness, and they exhibit a strong correlation with emotions [21, 22]. Due to this inherent correlation, these features have demonstrated their efficacy in Music Emotion Recognition (MER) tasks. These characteristics are rooted in musical concepts and theoretical knowledge, as noted in a survey [23], though quantitative studies on them remain relatively scarce. In our research, we build upon these concepts and execute a structured approach for achieving interpretable music audio tagging (Fig. 1), harnessing the power of perceptual features. Our specific contributions can be succinctly summarized as follows:

- In our study, we integrate three distinct methods for extracting perceptual musical features: signal processing, deep neural networks, and symbolic knowledge. Particularly noteworthy is our introduction of a novel set of features directly tied to music harmony. We demonstrate that these information sources are mutually complementary, significantly enhancing classifier performance across two datasets.
- We interpret model predictions, using four different methodologies, SHAP [24], XGBoost [25] feature importance,

permutation-based feature importance, and an ablation study.

- Utilizing the interpretability methods, we discuss how we could alleviate issues with ML-based audio tagging, such as detecting biases of the training dataset which would be obscured in a deep learning setup, and better comprehend ambiguous labels.

2. EXTRACTING PERCEPTUAL FEATURES

2.1. Datasets

We employed two datasets in our experiments, the MTG-Jamendo dataset [26] and the GTZAN dataset [27]. The MTG-Jamendo dataset serves as an open resource for music audio tagging. This dataset was constructed using music accessible through the Jamendo platform, offered under Creative Commons licenses, and enriched with tags provided by content uploaders. Boasting a substantial scale, it encompasses more than 18,486 complete audio tracks, that we used, annotated with 56 tags that encompass various mood and theme categories. Our choice of this specific dataset was predicated on the anticipation that interpretability would offer heightened value in a task involving partially subjective aspects, such as mood detection [28].

The GTZAN dataset [27] is a collection comprising 1000 audio tracks of 30 seconds and features 10 genres, with 100 tracks representing each. The audio files are all 16-bit monophonic recordings in .au format, sampled at 22kHz. We opted to use this dataset to assess the generalizability of our proposed pipeline on a smaller scale and a different corpus.

2.2. Methodologies

To extract perceptual features from the datasets, we employed three different domains: Symbolic Knowledge, Deep Neural Networks, and Signal Processing.

By utilizing **symbolic knowledge**, we created a new set of interpretable features that derive from music theory. To accomplish this, we constructed a knowledge base to extract music features from each track's chords. We selected Omnizart [29] from among many chord recognition models, based on its reported performance. Our knowledge base consists of chords as instances, and classes and subclasses from the Functional Harmony Ontology (FHO) [30]. These classes and subclasses describe the harmonic function of each chord in a music track. Using SPARQL queries, we retrieved the instances of certain classes to create useful features, such as the frequency of specific harmonic patterns (e.g., *dominant-dominant-tonic*), in the form of n-grams. In total, we created a 32-dimensional feature vector $x_h \in \mathbb{R}^{32}$. First, we created *Dominants* and *Subdominants* by calculating the ratio between the sum of dominant and subdominant retrieved chords and the total number of chords, respectively. Additionally, by defining the function of a chord in any chord progression, we developed 30 more features based on bigrams and trigrams (e.g., *sub-dom*, *sub-sub-dom*). These features were defined by calculating the ratio of a specific bigram or trigram to the total number of bigrams or trigrams in the analysis, respectively. To construct these features, we also designated some chords as tonics - chords that follow a dominant without serving another function, and globs - major dominant chords that produce a great deal of harmonic tension.

We leveraged the power of **deep neural networks** (DNNs) to efficiently capture perceptual features from a dataset curated by music experts [17]. To build our DNN model, we used a vgg-ish architecture trained with multiple linear regression. We first extracted MFCCs with 40 coefficients, utilizing a window and FFT size of 2048, from the initial 15 seconds of each audio file. This procedure resulted in an input MFCC $K \in \mathbb{R}^{S \times P}$, where S is the number of segments and P is the number of MFCCs. The model's output is a feature vector $x_m \in \mathbb{R}^p$, where $p \in \mathbb{N}$ corresponds to the number of extracted features. We trained our model on the "Mid Level Perceptual Music Features" dataset [17], comprising 5000 songs, each lasting 15 seconds in audio format. Each song is annotated by music experts with seven scalar perceptual features: *Melodiousness*, *Articulation*, *Rhythmic Stability*, *Rhythmic Complexity*, *Dissonance*, *Tonal Stability*, and *Minorness*.

Utilizing **signal processing** we extracted acoustically meaningful features. In order to choose features appropriate for our task a filtering process is essential having as a criterion their interpretability. If x is the input audio file, we define our process as a function E, where $E(x) = x_s$, $x_s \in \mathbb{R}^q$ and $q \in \mathbb{N}$ (q the number of the extracted features). For this approach, we used features extracted with a signal processing tool called Essentia [31]. The features were filtered out from three music experts, retaining only those that hold acoustical and musical significance and are easily understandable by humans. They ended up choosing 23 scalar features[1]: *Danceability*, *Loudness*, *Chords Changes Rate*, *Dynamic Complexity*, *Zero Crossing Rate*, *Chords Number Rate*, *Pitch Salience*, *Spectral Centroid*, *Spectral Complexity*, *Spectral Decrease*, *Spectral Energyband High*, *Spectral Energyband-Low*, *Spectral Energyband Middle High*, *Spectral Energyband Middle Low*, *Spectral Entropy*, *Spectral Flux*, *Spectral Rolloff*, *Spectral Spread*, *Onset Rate*, *Length*, *BPM*, *Beats Loud* and *Vocal-Instrumental*.

The description of the features and the code for the implementation of the experiments can be found in the GitHub repository [2]. By concatenating x_h, x_m, x_s, and normalizing the values with a standard scaler, we end up with the feature vector used to train the models.

3. EXPERIMENTS

When dealing with a set of perceptual features, there exist two strategies for achieving explainable predictions of audio tags. The first approach involves employing explainable-by-design ML models , like Decision Trees. The second approach revolves around the utilization of post hoc black-box explanation techniques capable of elucidating the functioning of any ML model [32]. In our research, we adopt a combination of both approaches. We incorporate explainable-by-design models due to their capacity to provide readily understandable outputs when fed with human-interpretable features as input. Additionally, we leverage post hoc black-box explanation methods, such as SHAP values, to enhance our interpretability efforts.

3.1. Classification

After experimenting with a wide array of traditional ML classifiers, we ended up choosing XGBoost as the most effective classifier. XG-

[1] https://essentia.upf.edu/streaming_extractor_music.html#music-descriptors

[2] https://github.com/vaslyb/PerceptibleMusicTagging

Boost is an ensemble ML classifier based on decision trees that employs a gradient-boosting framework. Our model achieved a ROC-AUC score of 0.729, in the multi-label mood recognition task of the MTG Jamendo dataset, slightly surpassing the baseline of 0.725 [33] while falling slightly below the performance of state-of-the-art intelligent systems 0.769 [34]. Similar results were obtained when applying XGBoost to the multi-class classification task using the GTZAN dataset, where we achieved an accuracy of 0.79, showing a slight deviation from the state-of-the-art performance 0.84 [35, 36]. While our findings may not outperform the state-of-the-art deep learning models, the presence of interpretable model outputs serves the valuable purpose of identifying potential biases within the model. Furthermore, we achieve commendable results using models characterized by a limited number of trainable parameters.

3.2. Explanations

Interpretability played a pivotal role in the selection of XGBoost as the predictive modeling technique. It enabled us to gain insights into the significance of input features in the overall predictions. This characteristic has rendered it a valuable tool for improving our understanding of the classification process. Our study utilized four methods to extract feature importance from our XGBoost model, including the built-in feature importance, permutation method, SHAP values, and ablation studies.

Initially, we utilized the **XGBoost feature importance**, which involved assessing the weight of features based on their frequency in trees. By aggregating this information across all tags, we obtained a holistic overview of the features crucial to the task. Additionally, examining the weight of each specific tag allowed us to discern the importance of features on a tag-by-tag basis.

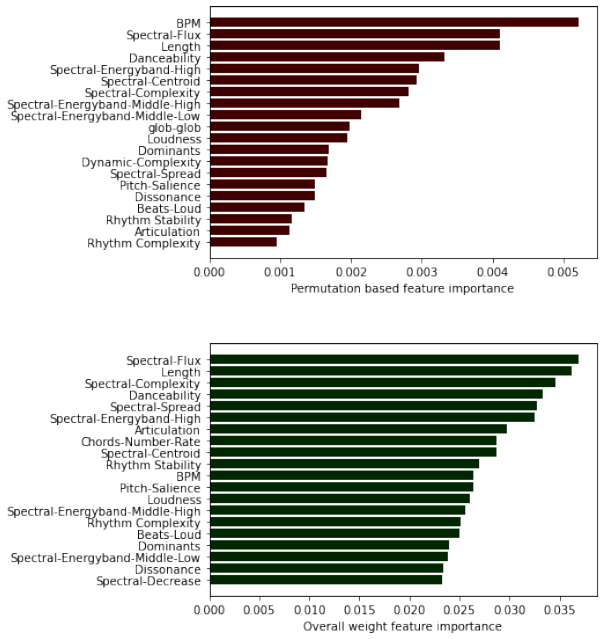
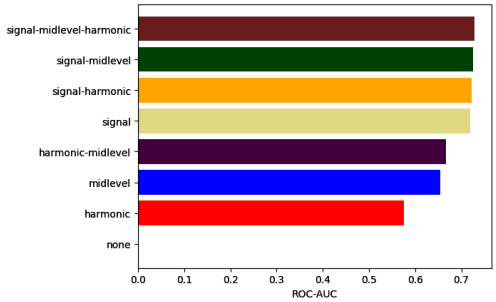
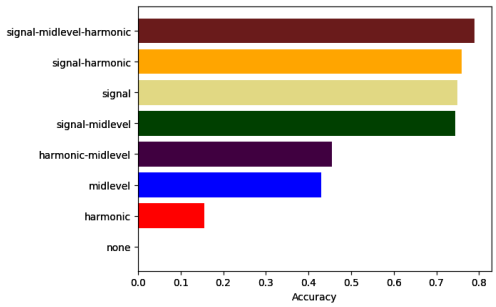


Fig. 2: Overall and permutation-based feature importance of the model trained on the MTG Jamendo dataset



(a) MTG Jamendo



(b) GTZAN

Fig. 3: Ablation studies with different three groups of features: harmonic, signal, and midlevel on both datasets

To further enhance our insights, we employed the **permutation method** to compute a second set of feature importance explanations. This involved iteratively discarding each feature from the training process and assessing the resulting change in performance. This approach provided an overall perspective on the importance of features across all mood tags.

Subsequently, we incorporated **SHAP values** to extract feature importance, leveraging game theory concepts to estimate the contribution of each feature to predictions. Unlike the weight and permutation methods, SHAP values indicated whether features impacted the predictions positively, or negatively. The results were visualized through plots, showcasing the distribution of impacts each feature had on the model output, sorted by the sum of SHAP value magnitudes over all samples.

In line with recommended practices from the literature, we conducted an **ablation study** with different feature groups, categorized as Mid-level, Harmonic, and Signal based on their origin. By experimenting with all possible combinations of these groups, we gained valuable insights into the distinct impacts each set of features had on the model's performance.

4. DISCUSSION

The results of permutation-based and the overall weight XGBoost feature importance for the Jamendo dataset are shown in Fig. 2. The two methods produce similar results, which indicates that the displayed features are likely indeed important. Signal processing features were the most important for predictions, notably *Spectral-Flux*,

Spectral-Complexity and *Spectral-Centroid*.

We obtained similar results in our ablation study (Fig. 3), where we found that signal processing features were the most useful, while harmonic features were the least useful for both datasets. Considering that harmonic features are extracted from the chord progression, this observation could be attributed to the quality of the chord recognition tool. Harmonic features had a substantial impact only on predicting the *Blues* labels, a genre where chord progressions are distinct and easy to detect throughout a track. In terms of mid-level features, they appeared to be significant in improving the ROC-AUC metric for the Jamendo dataset. For both datasets, the best results were achieved by the combination of all features.

We show examples of interpreting feature importance computed with SHAP values and XGBoost for predicting mood labels in the MTG Jamendo dataset. Specifically, we analyze one of the best-performing labels (ROC-AUC > 0.8) as shown in Fig. 5. For the highly-predictable *Trailer* mood predictions rely on low values of *Length* and high values of *Spectral-Flux*, a feature that is used for voice detection. This is aligned with the fact that trailers are usually shorter than songs and have voice-overs.

For GTZAN, we show examples of the interpretation for one predicted genre, *Disco* that reached 0.72 F1-score (see Fig. 4). For this label, it is noticeable that high values of *Beat-Loud* and middle values of *BPM* are the most prevalent features. These observations align with the common perception of Disco music as a genre characterized by distinguishable beats, easy to dance to.

The above examples are meant to illustrate the value of interpretability for the audio tagging task. We can gain insights on a case-by-case basis which would not be possible in an end-to-end setup. Whether or not these insights are worth the additional work needed for feature extraction, and the slight deterioration in performance, depends on each application's priorities. We argue that interpretability should be a priority for music audio tagging systems, as they suffer from label ambiguity [37], in addition to being at risk of the typical biases of MIR datasets [38].

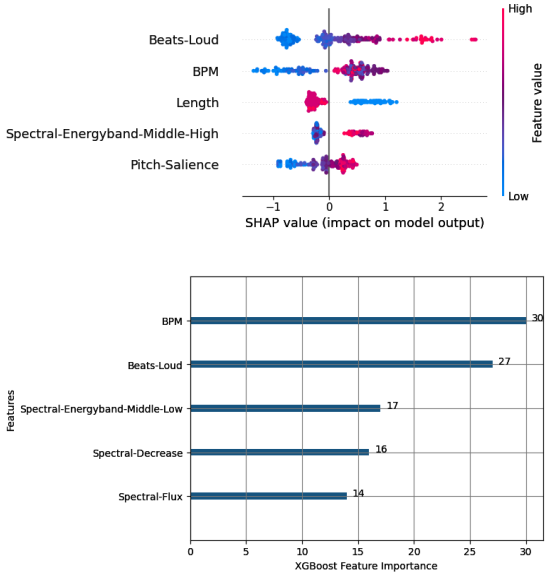


Fig. 4: Interpretations of the prediction for the label *Disco* in the GTZAN dataset based on SHAP values and model's feature importance.

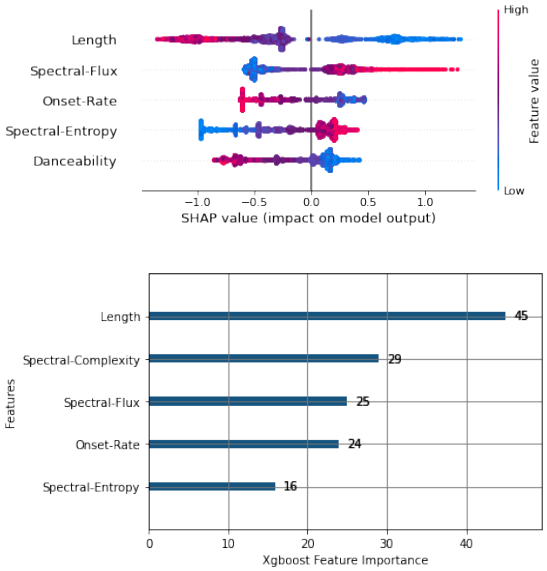


Fig. 5: Interpretations of the prediction for the label *Trailer* in the MTG Jamendo dataset based on SHAP values and model's feature importance

5. CONCLUSIONS AND FUTURE WORK

We have created a specialized pipeline for music audio tagging with an emphasis on delivering interpretable results. Our approach incorporates established feature extraction methods alongside the introduction of a fresh set of features derived from the realm of music harmony concepts. Our experimental findings underscore the pivotal role of feature engineering within our proposed pipeline, as it exerts an influence on both performance and interpretability. It is imperative that each feature's meaning remains comprehensible to humans, emphasizing the significance of this step.

For future work, we plan to further investigate ways for improving performance given the explanations of predictions. Furthermore, we plan to study explanations regarding not only their performance improvement capabilities but also their usefulness and informativeness for end-users of music audio tagging systems. Finally, we plan to extend the feature extraction procedure with more features that are based on notions of music theory.

Acknowledgments

We express our sincere gratitude to Christos Garoufis for his useful insights and thoughtful comments on our work. The research project is implemented in the framework of H.F.R.I call "Basic research Financing (Horizontal support of all Sciences)" under the National Recovery and Resilience Plan "Greece 2.0" funded by the European Union –NextGenerationEU(H.F.R.I. Project Number: 15111 - Emotional Artificial Intelligence in Music Expression)

6. REFERENCES

[1] Antti Eronen, "Comparison of features for musical instrument recognition," in *WASPAA*. IEEE, 2001.

[2] Edmund Dervakos, Natalia Kotsani, and Giorgos Stamou, "Genre recognition from symbolic music with cnns," in *EvoMUSART*. Springer, 2021.

[3] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis Vlahavas, "Multi-label classification of music by emotion," *EURASIP*, vol. 2011, no. 1, pp. 1–9, 2011.

[4] Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang, "1000 songs for emotional analysis of music," in *Proc. of the 2nd ACM international workshop on Crowdsourcing for multimedia*, 2013.

[5] Vassilis Lyberatos, Spyridon Kantarelis, Eirini Kaldeli, Spyros Bekiaris, Panagiotis Tzortzis, Orfeas Menis-Mastromichalakis, and Giorgos Stamou, "Employing crowdsourcing for enriching a music knowledge base in higher education," in *AIET*. Springer, 2023, pp. 224–240.

[6] Tsung-Ping Chen and Li Su, "Harmony transformer: Incorporating chord segmentation into harmony recognition," *neural networks*, 2019.

[7] Haruto Takeda, Takuya Nishimoto, and Shigeki Sagayama, "Rhythm and tempo recognition of music performance from a probabilistic approach," in *ISMIR*, 2004.

[8] Pasquale Lisena, Albert Meroño-Peñuela, and Raphaël Troncy, "Midi2vec: Learning midi embeddings for reliable prediction of symbolic music metadata," *Semantic Web*, vol. 13, no. 3, pp. 357–377, 2022.

[9] Jordi Pons Puig, Oriol Nieto Caballero, Matthew Prockup, Erik M Schmidt, Andreas F Ehmann, and Xavier Serra, "End-to-end learning for music audio tagging at scale," in *ISMIR*, 2018.

[10] Morgan Buisson, Pablo Alonso-Jiménez, and Dmitry Bogdanov, "Ambiguity modelling with label distribution learning for music classification," in *ICASSP*, 2022.

[11] Hendrik Vincent Koops, W Bas De Haas, John Ashley Burgoyne, Jeroen Bransen, Anna Kent-Muller, and Anja Volk, "Annotator subjectivity in harmony annotations of popular music," *Journal of New Music Research*, vol. 48, no. 3, pp. 232–252, 2019.

[12] Cynthia Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019.

[13] Andre Holzapfel, Bob Sturm, and Mark Coeckelbergh, "Ethical dimensions of music information retrieval technology," *TISMIR*, vol. 1, no. 1, pp. 44–55, 2018.

[14] Saumitra Mishra, Bob L Sturm, and Simon Dixon, "Local interpretable model-agnostic explanations for music content analysis.," in *ISMIR*, 2017, vol. 53, pp. 537–543.

[15] Minz Won, Sanghyuk Chun, and Xavier Serra, "Toward interpretable music tagging with self-attention," *arXiv preprint arXiv:1906.04972*, 2019.

[16] Anders Friberg, Erwin Schoonderwaldt, Anton Hedblad, Marco Fabiani, and Anders Elowsson, "Using listener-based perceptual features as intermediate representations in music information retrieval," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1951–1963, 2014.

[17] Anna Aljanaki and M. Soleymani, "A data-driven approach to mid-level perceptual musical feature modeling," in *ISMIR*, 2018.

[18] Shreyan Chowdhury, Andreu Vall Portabella, Verena Haunschmid, and Gerhard Widmer, "Towards Explainable Music Emotion Recognition: The Route via Mid-level Features," 2019, ISMIR.

[19] Shreyan Chowdhury and Gerhard Widmer, "Towards explaining expressive qualities in piano recordings: Transfer of explanatory features via acoustic domain adaptation," in *ICASSP*. IEEE, 2021, pp. 561–565.

[20] Shreyan Chowdhury, Verena Praher, and Gerhard Widmer, "Tracing back music emotion predictions to sound sources and intuitive perceptual qualities," *CoRR*, 2021.

[21] Alf Gabrielsson and Erik Lindström, "The role of structure in the musical expression of emotions," *Handbook of music and emotion: Theory, research, applications*, vol. 367400, pp. 367–44, 2010.

[22] Lage Wedin, "A multidimensional study of perceptual-emotional qualities in music," *Scandinavian journal of psychology*, 1972.

[23] Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang, "A survey of music emotion recognition," *Frontiers of Computer Science*, vol. 16, no. 6, pp. 1–11, 2022.

[24] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017.

[25] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *SIGKDD*. 2016, KDD '16, ACM.

[26] 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, ICML*, 2019.

[27] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[28] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull, "Music emotion recognition: A state of the art review," in *ISMIR*, 2010.

[29] Yu-Te Wu, Yin-Jyun Luo, Tsung-Ping Chen, I-Chieh Wei, Jui-Yang Hsu, Yi-Chin Chuang, and Li Su, "Omnizart: A general toolbox for automatic music transcription," *Journal of Open Source Software*, vol. 6, no. 68, pp. 3391, 2021.

[30] Spyridon Kantarelis, Edmund Dervakos, Natalia Kotsani, and Giorgos Stamou, "Functional harmony ontology: Musical harmony analysis with description logics," *Journal of Web Semantics*, vol. 75, pp. 100754, 2023.

[31] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Herrera Boyer, Oscar Mayor, Gerard Roma Trepat, Justin Salamon, José Ricardo Zapata González, Xavier Serra, et al., "Essentia: An audio analysis library for music information retrieval," ISMIR, 2013.

[32] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[33] Philip Tovstogan, Dmitry Bogdanov, and Alastair Porter, "Media-eval 2021: Emotion and theme recognition in music using jamendo," in *Proc. of the MediaEval 2021 Workshop, Online*, 2021, pp. 13–15.

[34] Vincent Bour, "Frequency dependent convolutions for music tagging," in *Proc. of the MediaEval 2021 Workshop, Online*, 2021.

[35] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino, "Masked modeling duo: Learning representations by encouraging both networks to model the input," in *ICASSP*, 2023.

[36] Rodrigo Castellon, Chris Donahue, and Percy Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *ISMIR*, 2021.

[37] Juan Sebastián Gómez Cañón, Estefanía Cano, Herrera Boyer, Emilia Gómez Gutiérrez, et al., "Joyful for you and tender for us: The influence of individual characteristics and language on emotion labeling and classification," ISMIR, 2020.

[38] Cory McKay and Ichiro Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?," in *ISMIR*, 2006, pp. 101–106.