Encoding Emotion in Music via Acoustic Features: A Weakly Supervised Machine Learning Study

Jiaming Mao
University of Chicago
jmao0220@uchicago.edu

May 2025

Abstract

This study investigates how accurately and interpretably machine learning models can classify emotional content in music using only Spotify-derived acoustic features. Over 80,000 tracks were weakly labeled with six basic emotions—*joy*, *sadness*, *anger*, *disgust*, *fear*, and *surprise*—by applying a transformer-based classifier (DistilRoBERTa) to Last.fm user-generated tags. A top-k multilabel evaluation strategy was used to reflect the non-exclusive nature of musical emotion.

Three models—Random Forest, K-Nearest Neighbors, and Multi-Layer Perceptron—were trained and compared. Random Forest achieved the highest micro-average F1 score (0.690) and exact match accuracy (38.77%), outperforming others particularly in precision on low-frequency classes like *fear* and *disgust*. However, some emotion categories remained difficult to separate due to co-occurrence and semantic ambiguity.

Feature importance comparisons revealed that `acousticness`, `energy`, `valence`, and `loudness` contributed most to emotional inference, while features like `mode` and `key` had limited value. SHAP-based interpretation showed that errors often stemmed from low signal strength or conflicting feature cues rather than random noise. Ablation studies confirmed that emotional prediction is driven by a compact yet expressive acoustic subspace.

Limitations include reliance on transformer-generated labels, absence of lyrical features, and potential genre-specific biases. Still, the findings demonstrate the feasibility of weakly supervised, audio-based multilabel emotion recognition and provide interpretable insight into how models learn affect from music.

1 Introduction

Music has long been recognized as a powerful conduit for emotional expression, and its ability to evoke affective responses is deeply embedded in both biological and cultural dimensions (Huron 2015; Perlovsky 2010). Computational models have increasingly attempted to classify emotional responses to music by analyzing audio features such as tempo, timbre, energy, and spectral properties. Although model performance (e.g. accuracy, F1 score) remains a common evaluation focus (Yang 2024; Yoo, Hong, and Hyeocheol Kim 2024), the recent literature suggests a more nuanced perspective: different machine learning models vary not only in prediction performance but also in their sensitivity to different musical features (Xia and F. Xu 2022; L. Xu et al. 2021).

This paper investigates the question: *How accurately can supervised machine learning models predict user-assigned emotional categories of songs using only audio features, and which audio features contribute most to these predictions?* Rather than simply comparing which model performs best, the objective is to examine how models such as Random Forest, MLP, and KNN use characteristics differently to make predictions. This approach is motivated by recent findings that indicate that the effectiveness of features such as MFCCs, energy, roll-off, and pitch can vary substantially between models (Juthi et al. 2020; Rosner and Kostek 2018;

Garg et al. 2022). Some models may rely more on spectral sharpness or rhythm, while others may favor dynamic or harmonic information. Understanding these patterns helps clarify both the behavior of models and the nature of musical emotion itself.

The theoretical foundation for this investigation comes from emotion models such as Russell's circumplex model and Thayer's valence arousal plane (Helmholz, Meyer, and Robra-Bissantz 2019), as well as psychological studies on how acoustic features induce affect (McCraty et al. 1998; Leubner and Hinterberger 2017). Furthermore, my study builds on the literature on domain-specific emotion recognition, particularly research integrating lyrics or domain-tuned lexicons (Bandhakavi et al. 2017; L. Xu et al. 2021). These works suggest that interpretability and modality-specific contributions are essential for building robust, generalizable emotional classifiers.

By analyzing feature importance rankings across classifiers and exploring areas of agreement and divergence, this paper offers a new perspective on model interpretability in music emotion recognition. The goal is not only to identify which features predict emotion best, but to ask *why certain features matter more to some models than others*, and what this implies for both machine learning research and affective music theory.

Specifically, this study makes the following contributions:

- It constructs a weakly labeled dataset of over 80,000 songs by mapping Last.fm tags to emotion categories using a transformer-based semantic model (DistilRoBERTa), enabling large-scale training without manual annotation.
- It systematically compares how three classifiers—Random Forest, K-Nearest Neighbors, and MLP—use acoustic features to predict emotional categories, highlighting both performance trade-offs and distinct model-specific feature importances.
- It introduces a multi-perspective interpretability framework including permutation-based feature attribution, SHAP analysis, and error-driven model interpretation, offering practical insights into classifier behavior and emotion-specific confusion.

2 Literature Review

This study builds on multiple strands of prior research, drawing on theories of emotional structure, semantic projection from social tagging systems, machine learning models for text-based emotion detection, acoustic feature analysis in music information retrieval (MIR), and model interpretability techniques. Together, these literatures inform the construction of our labeling pipeline, the design of our audio-based classifiers, and our strategy for interpreting feature contributions.

2.1 Theoretical Models of Emotion Representation

Understanding emotion begins with selecting a representational framework that captures the complexity of affective states. Russell's Circumplex Model (Russell and Steiger 1982), which characterizes emotion in a two-dimensional valence-arousal space, remains foundational in both psychological and MIR research. Its strength lies in expressing subtle gradients between emotions rather than reducing them to discrete categories. Recent studies, such as Longo et al. (2024), reaffirm the model's generalizability across multimodal contexts, demonstrating that textual, auditory, and visual expressions of emotion continue to cluster meaningfully within the valence-arousal plane (Raufi, Finnegan, and Longo 2024). These findings support the choice of emotion-related features like energy and danceability in this project, which serve as valence-arousal proxies for downstream prediction.

2.2 Challenges in Music Emotion Annotation

Accurate labeling of music emotion data remains a central challenge in MIR. Manual annotation is prohibitively expensive and often inconsistent, as emphasized by Cano and Schuller (2017), who note that high-quality labeled datasets are rare (Cano and Morisio 2017). To address this, researchers increasingly turn to user-generated metadata such as Last.fm tags, which offer a noisy but scalable alternative. The challenge, however, is to translate these free-text tags into reliable emotion labels. While previous studies such as Olha et al. (2023) have proposed using word embeddings like Word2Vec for semantic clustering of tags (Olha Zalutska et al. 2023), this study adopts a more context-aware approach by applying a transformer-based zero-shot classifier (DistilRoBERTa) to map tags directly to emotion categories. This strategy better handles polysemous or stylistically nuanced tags and enables large-scale emotion labeling without relying on rigid keyword lists.

2.3 Music Information Retrieval and Acoustic Feature Relevance

Content-based MIR provides the foundation for choosing acoustic features that relate to emotional perception. Casey et al. (2008) identify a wide range of relevant features—from timbral texture to rhythm and pitch—that help structure automated MIR systems (Casey et al. 2008). Building on this, Kamenetsky et al. (1997) conducted psychological experiments revealing that musical parameters such as tempo and intensity reliably shift emotional perception (Kamenetsky, Hill, and Trehub 1997). These findings support the idea that low-level audio features can encode high-level affective meaning, making them well-suited for use as inputs in emotion-predictive models.

2.4 Semantic Labeling Using Transformer-Based Language Models

To improve the semantic labeling of noisy, user-generated text inputs, transformer-based models such as BERT, RoBERTa, and their distilled variants provide significant advantages. Unlike static embedding models, transformers can capture contextual nuance in short and ambiguous tags through self-attention mechanisms. Acheampong et al. (2021) synthesize findings across domains and show that transformer models outperform traditional approaches on short-text classification tasks, a common characteristic of Last.fm tags (Acheampong, Nunoo-Mensah, and Chen 2021). In the specific context of music tagging, Olha Zalutska et al. (2023) demonstrate that RoBERTa maintains high classification accuracy even when applied to semantically fuzzy or stylistically diverse user labels. Building on this evidence, this study implemented DistilRoBERTa as the core tagging model, allowing for fine-grained emotion classification while retaining scalability. This approach enables the system to assign emotional categories such as *fear* or *surprise* to compound or metaphorical tags like `trippy jazz` or `dark nostalgia`—cases where static models often fail. Recent work by Artemova et al. (2025) further validates the use of transformer-generated labels, showing that minimal human review of LLM outputs can rival or exceed traditional crowd-annotation, especially in subjective domains like emotion recognition. Similarly, Hannah Kim et al. (2024) show that prompt-engineered transformers can successfully classify real-world, weakly structured inputs, such as open-ended student responses in educational settings. These developments collectively justify the use of transformer-based pipelines as robust tools for weak supervision in emotion classification tasks.

2.5 Weak Supervision and Label Quality Considerations

Weak supervision techniques are essential in large-scale emotion classification tasks where human-annotated ground truth is impractical to obtain. In affective computing and educational research, label proxies—such as semantic projections or model-generated tags—are increasingly accepted, provided they are transparently validated. Kim et al. (2024) present a powerful example of this approach through a human-centered LLM-integrated dashboard for writing education, where ChatGPT-based feedback and behavior logs are automatically analyzed using fine-tuned models (Hannah Kim et al. 2024). Their system detects misuse

patterns, aligns student inputs with curricular learning objectives, and enables teachers to provide adaptive feedback based on weakly supervised signals derived from chat data and model predictions. This reinforces the feasibility of using LLM-generated labels, especially when paired with targeted human oversight. Inspired by their methodology, this study treats Last.fm tag-derived emotion labels as a form of distant supervision, acknowledging their noise but leveraging semantic embeddings to minimize inconsistency. Kim et al.'s design also underscores the importance of involving end users (in their case, teachers) in refining model interpretations—paralleling our inclusion of interpretable model diagnostics to mitigate label-induced bias. Their iterative process, combining NLP expertise with qualitative insights, sets a methodological precedent for managing noisy data pipelines in emotion-related applications. While risks such as semantic drift and confirmation bias remain, their work provides a practical blueprint for transparent and human-aligned weak supervision systems, validating the broader use of model-based labeling in complex, subjective domains like musical emotion classification. Importantly, Uplabdhee et al. (2025) demonstrate that even in the absence of gold-standard annotations, weakly supervised pipelines using rule-based emotion taggers and acoustic feature heuristics can achieve strong multi-label classification results. Their hybrid model, evaluated on genre-diverse music corpora, supports the viability of scalable labeling strategies when human annotation is infeasible. These findings align with the approach taken in this study: leveraging transformer-based emotion projections on Last.fm tags as a proxy for listener affect across a broad corpus of songs.

2.6 Multi-Label Emotion Classification in Music

Affective responses to music are inherently multidimensional: a single track may evoke joy and nostalgia simultaneously, or anger interlaced with excitement. Capturing this complexity requires moving beyond traditional single-label classification. Multi-label frameworks allow each musical input to be associated with a set of emotional categories, better reflecting real-world listening experiences. Ahsan, Kumar, and Jawahar (2015) frame music annotation as a multi-label classification task and compare algorithms such as binary relevance, classifier chains, and ensemble methods, finding that multi-label models outperform one-hot emotion classifiers in expressive range and precision. More recently, Uplabdhee et al. (2025) provide a comprehensive evaluation of multi-label music emotion recognition using deep neural networks and rule-based heuristics, emphasizing that multi-label output not only improves classification metrics but also aligns more closely with psychological models of emotion. Their study reveals that incorporating co-occurrence information and feature-level fusion enhances both recall and emotional plausibility. This study adopts the same perspective, implementing a top-k dynamic matching strategy to allow each track to receive multiple emotional predictions. This design choice reflects both empirical success in recent literature and theoretical alignment with the non-exclusive nature of emotional expression in music.

2.7 Interpretability of Machine Learning Models in Emotion Prediction

Interpretability methods are essential for understanding the behavior of black-box models in emotion classification. As Parthasarathy et al. (2017) argue, accurate predictions alone are insufficient without insight into which features drive these predictions (Parthasarathy, Lotfian, and Busso 2017). Tools like SHAP values and permutation importance help quantify individual feature contributions, improving transparency. Though initially developed in contexts like fraud detection, these methods translate effectively to MIR applications. This study applys permutation feature importance to assess which acoustic features most strongly influence model outputs, ensuring that emotion predictions remain interpretable and aligned with psychological theories of affect. Kim et al. (2024) similarly advocate for interpretability in educational AI tools, showing that teachers rely on NLP-enhanced dashboards to gain contextual and semantic insights into student behavior, rather than relying on statistical summaries alone (Hannah Kim et al. 2024). Their approach strengthens the case for embedding interpretability at every level of the modeling pipeline in subjective domains like emotion.

Together, these strands of literature provide both the conceptual and methodological foundations for the current study. The following section describes how these theories were operationalized into a data labeling pipeline and feature-driven classification framework.

3 Data and Methods

3.1 Dataset Construction and Emotion Labeling

This study investigates how machine learning models classify songs into emotional categories based on audio features extracted from Spotify. The emotion labeling pipeline begins with user-generated tags from the Last.fm subset of the Million Song Dataset (Thierry et al. 2011), which includes two primary SQLite files: `tags.db`, containing user-assigned tags for tracks and their IDs, and `track_metadata.db`, containing track metadata including track ID, name, and artist. The final dataset includes over 80,000 songs labeled with one or more of six emotions and enriched with Spotify-derived audio features. The pipeline proceeds as follows:

Tag-to-Emotion Mapping. To generate emotion labels from Last.fm's user-generated tags, this study adopted a transformer-based model: DistilRoBERTa fine-tuned for multi-class emotion classification (Hartmann 2022). Compared to static embeddings, DistilRoBERTa captures contextual meaning by leveraging attention-based token interactions. Each tag was passed through the model, which returned a ranked list of predicted emotion labels with associated confidence scores. This study retained only predictions that exceeded a confidence threshold of 0.8 and excluded neutral outputs to focus on emotion-rich tags. For example, when applying the classifier to the tag `"happy"`, the model returned *joy* with a confidence score of 0.993—indicating strong affective alignment and semantic clarity.

Only labels with a classification confidence above 0.8 were retained, and any predictions labeled as `neutral` were excluded to ensure the dataset contained only emotionally salient tags. This thresholding step was critical for improving the affective clarity of the weakly labeled dataset. For instance, the tag `"happy"` was confidently mapped to *joy* with a probability score of 0.993, whereas low-confidence or semantically ambiguous tags were discarded. See Appendix A for a detailed justification of this threshold selection.

Tag	Predicted Emotion
happy	joy
melancholy	sadness
ragecore	anger
trippy jazz	surprise

Table 1: Sample entries generated by DistilRoBERTa classifier.

This process yielded emotion labels for a large number of Last.fm tags, which were then linked to individual tracks via the tags' associated `track_id`s. The resulting file recorded 297,730 tracks associated with at least one emotion-labeled tag. The emotion label distribution across these tags was as follows: 123,053 labeled as *joy*, 85,313 as *sadness*, 30,165 as *anger*, 22,692 as *disgust*, 19,569 as *surprise*, and 16,938 as *fear*. Importantly, no additional semantic or genre-based tag filtering was performed at this stage-a choice that is revisited in the *Limitations* section.

Justification for Labeling Strategy In emotion recognition tasks where large-scale human annotation is impractical, transformer models offer a scalable proxy for subjective judgments. Prior studies such as Hartmann (2022) show that fine-tuned emotion classifiers outperform crowd-sourced baselines in both semantic coherence and label consistency. Recent work also supports the use of model-generated annotations as reliable weak labels when paired with light filtering (Artemova et al. 2025; Hannah Kim et al. 2024). While such predictions inevitably carry noise, filtering by model confidence—e.g., retaining only outputs with scores above 0.8 and excluding neutral labels—enhances alignment with affect-rich input.

This approach enabled efficient multi-label annotation of over 80,000 music tags and produced a usable training corpus without manual review. It also better reflects the distributed and ambiguous nature of listener

sentiment, a critical consideration in music informatics. By leveraging confidence-aware predictions and filtering heuristics, the pipeline provides a scalable yet interpretable form of weak supervision for emotion modeling.

Track Metadata Retrieval. Emotion-labeled tags were mapped to corresponding Last.fm `track_id`s using `tags.db`. Track-level metadata such as titles and artist names was retrieved from the `track_metadata.db` SQLite database, which contains over one million records in the `songs` table. To address SQLite's parameter limit, track IDs were grouped into batches of 500 and queried using parameterized SQL. Retrieved metadata was organized into a tag-to-track mapping, with duplicates removed via Python's `set()` to avoid inflating the dataset.

Track ID	**Title**	**Artist Name**
TRMMMYQ128F932D901	Silent Night	Faster Pussy cat
TRMMMKD128F425225D	Tanssi vaan	Karkkiautomaatti
TRMMMRX128F93187D9	No One Could Ever	Hudson Mohawke
TRMMMCH128F425532C	Si Vos Querés	Yerba Brava
TRMMMWA128F426B589	Tangle Of Aspens	Der Mystic

Table 2: Sample entries from `track_metadata.db`.

This yielded a set of emotion-labeled tracks with valid metadata, forming the basis for downstream Spotify matching and model training. The final dataset retained the multi-label structure, each track could be associated with multiple emotions, which better capturing the complexity of musical affect.

The resulting dataset contains 82,950 matched tracks, each labeled with one or more emotions. Figure 1 shows the distribution across categories.

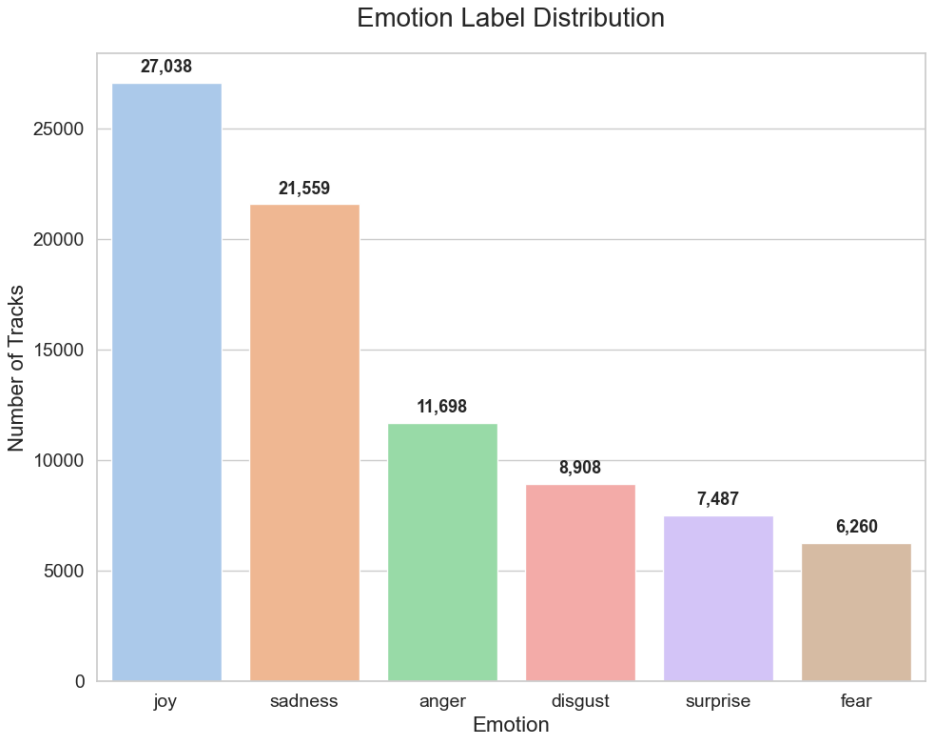

Figure 1: Multi-label emotion distribution across matched tracks.

As shown, *joy* and *sadness* dominate the label space, while emotions like *surprise* and *fear* are underrepresented. To address this imbalance during training, this study employed stratified sampling, undersampling of dominant classes, and synthetic oversampling for minority ones. While these techniques improved class representation, residual imbalance remains a source of modeling bias—further discussed in the Limitations section.

3.2 Track Matching and Audio Feature Retrieval

After generating emotion labels and retrieving track metadata from the Last.fm dataset, I aligned each track with acoustic features provided by Spotify. This involved two main steps: Spotify ID retrieval using the Spotify Web API, and audio feature matching via a public Kaggle dataset.

Spotify Track Matching. Spotify track IDs were retrieved by querying the Spotify Web API (*Spotify Web API* n.d.) using a combination of track titles and artist names, implemented via the `spotipy` Python package. To ensure precision, results were filtered based on exact string matches in both the `name` and `artist` fields. Ambiguous, duplicated, or timed-out matches were discarded. Due to API limitations and occasional metadata mismatches, not all tracks could be enriched with Spotify audio features. These constraints and their implications for potential model bias are further discussed in the Limitations section.

Audio Feature Retrieval and Final Dataset Composition. Once Spotify track IDs were obtained, they were joined with the Kaggle dataset *"8+ Million Spotify Tracks, Genre, Audio Features"* (Grosse 2022), accessed on April 15, 2025. This dataset contains over 8.7 million entries and provides 13 Spotify-computed audio features per track, including `acousticness`, `danceability`, `energy`, `loudness`, `valence`, and `tempo`, among others. The join was performed using Spotify track IDs to ensure accurate alignment across sources.

Spotify ID	Acousticness	Danceability	Energy	Loudness (dB)	Valence	Tempo (BPM)
2jKoVlU7VAmExKJ1Jh3w9P	0.180	0.893	0.514	-5.08	0.787	95.85
4JYUDRtPZuVNi7FAnbHyux	0.272	0.520	0.847	-5.30	0.799	177.37
6YjKAkDYmlasMqYw73iB0w	0.078	0.918	0.586	-2.89	0.779	95.52
2YlvHjDb4Tyxl4A1IcDhAe	0.584	0.877	0.681	-6.28	0.839	94.83
3UOuBNEin5peSRqdzvlnWM	0.170	0.814	0.781	-3.33	0.536	93.44

Table 3: Sample entries from the Kaggle `audio_features` table.

Following the join, 297,730 tracks with at least one emotion label were identified. Due to Spotify API constraints and metadata inconsistencies, only 82,950 records could be reliably enriched with audio features. To reduce over-representation of frequently tagged songs, a deduplicated version was created by grouping tracks by unique `title + artist` combinations and aggregating their associated emotion labels into multilabel annotations. This process yielded 39,635 unique tracks.

Two analysis-ready views were maintained:

- **Raw dataset** (`spotify_emotion_feature_multi.csv`): 82,950 records, each representing one emotion-tag match for a track.
- **Deduplicated dataset**: 39,635 unique tracks with consolidated multilabel emotion annotations.

All tracks include 12 Spotify-computed audio features capturing rhythmic, acoustic, dynamic, and tonal characteristics. Table 4 summarizes the schema, and Table 5 provides example records.

Column	Description
`track_id`	Spotify track identifier
`emotion`	Emotion label(s): joy, sadness, anger, fear, surprise, disgust
`acousticness`	Confidence the track is acoustic (0.0–1.0)
`danceability`	Suitability for dancing (0.0–1.0)
`energy`	Intensity and activity (0.0–1.0)
`instrumentalness`	Probability of instrumental nature (0.0–1.0)
`key`	Musical key (0 = C, ..., 11 = B)
`liveness`	Likelihood of live performance (0.0–1.0)
`loudness`	Track loudness in decibels (dB)
`mode`	Modality: major (1) or minor (0)
`speechiness`	Presence of spoken words (0.0–1.0)
`tempo`	Tempo in beats per minute (BPM)
`time_signature`	Estimated beats per bar (e.g., 4)
`valence`	Positivity and musical cheerfulness (0.0–1.0)

Table 4: Final dataset schema with 12 Spotify audio features and multi-label emotion tags.

track_id	emotion	acousticness	danceability	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence
61ajg7tHtzucOvK4WhCGBG	joy	0.0668	0.7490	0.6660	6.39E-06	10	0.0864	-4.8090	0	0.1580	186.07	4	0.2680
7dzUZec5MnWMyQnk5klnKR	joy	0.0194	0.6710	0.8200	0.0133	0	0.3610	-5.7960	1	0.0480	122.36	4	0.6850
1QQgmN383kUqjioRoTSfF3	joy	0.0001	0.2890	0.9380	0.0615	1	0.1060	-6.0830	1	0.1630	173.88	4	0.3490
6atVS7UZBxoyJkkteM62u5	joy	0.1720	0.6660	0.7130	0.0320	2	0.1770	-3.5510	1	0.0384	94.71	4	0.4910

Table 5: Sample records from the final dataset with audio features and emotion labels.

3.3 Descriptive Statistics of Audio Features

To characterize the acoustic profile of the music, I computed summary statistics across the deduplicated dataset. Table 6 reports the distributional properties of each audio feature, while Table 7 shows example songs and their associated multilabel annotations.

Deduplicated Dataset (39,635 unique tracks). Since each track on Last.fm may include multiple user-generated tags—often linked to different emotions—the initial mapping process produced multiple entries per song, each representing a distinct tag-emotion pair. To avoid disproportionate representation of frequently tagged tracks, I deduplicated the dataset by grouping entries based on unique `title + artist` pairs. All associated emotion labels were then aggregated into a single multilabel annotation per track. This process resulted in a final dataset of 39,635 unique tracks, each assigned one or more emotion categories.

Table 7 provides a sample from the final dataset, showing the multilabel format and associated metadata after deduplication. See Appendix B for descriptive statistics of the raw dataset (82,950 records) prior to deduplication.

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Acousticness	39,635	0.2486	0.3080	0.0000	0.0048	0.0869	0.4340	0.9960
Danceability	39,635	0.5231	0.1773	0.0000	0.3980	0.5290	0.6530	0.9800
Energy	39,635	0.6503	0.2508	0.0000	0.4710	0.6910	0.8670	1.0000
Instrumentalness	39,635	0.1643	0.2993	0.0000	0.0000	0.0007	0.1410	0.9930
Key	39,635	5.3163	3.5597	0.0000	2.0000	5.0000	9.0000	11.0000
Liveness	39,635	0.1984	0.1636	0.0088	0.0953	0.1310	0.2630	1.0000
Loudness (dB)	39,635	-8.8165	4.4138	-50.0140	-11.1400	-7.9290	-5.6310	3.7440
Mode	39,635	0.6593	0.4740	0.0000	0.0000	1.0000	1.0000	1.0000
Speechiness	39,635	0.0767	0.0843	0.0000	0.0340	0.0458	0.0790	0.9600
Tempo (BPM)	39,635	122.08	29.58	0.0000	99.34	120.01	140.04	219.93
Time Signature	39,635	3.9041	0.3986	0.0000	4.0000	4.0000	4.0000	5.0000
Valence	39,635	0.4868	0.2596	0.0000	0.2720	0.4810	0.6990	0.9890

Table 6: Descriptive statistics of Spotify audio features across 39,635 unique tracks.

Track ID	Emotion Labels	Title	Artist
0sRRjH8LDDnDl60lqXjhIT	anger, sadness, joy	Higher Than the Stars	The Pains Of Being Pure At Heart
7e0qA4bZ1LZqZzdmG7jZ6F	joy, surprise	Walking on a Dream	Empire of the Sun
3yfqSUWxFvZELEM4PmlwIR	sadness	Skinny Love	Bon Iver
1kPpge9JDLpcj15qgrPbYX	fear, anger	Closer	Nine Inch Nails

Table 7: Sample of matched emotion labels with track metadata. Each row represents a unique song associated with one or more emotion labels.

Interpretation. Both datasets exhibit comparable central tendencies; however, the deduplicated version presents marginally higher average values in features such as `acousticness`, `speechiness`, and `valence`. This pattern suggests that songs with a greater number of associated tags—often filtered out during deduplication—are more likely to belong to high-energy, low-acoustic genres such as pop or electronic music. While outliers and feature skewness remain—evident in attributes like near-zero `tempo` or unusually low `loudness`—these values have been preserved to maintain the integrity of the original distribution. Future normalization or error analysis may warrant additional handling of these anomalies.

At a broader level, the descriptive statistics reflect several structural characteristics of the dataset. Features including `acousticness`, `instrumentalness`, and `speechiness` are strongly right-skewed toward zero, indicating a dominance of vocal, non-acoustic tracks. In contrast, attributes such as `danceability` and `energy` show approximately symmetric distributions centered around mid-range values, with a slight skew toward higher energy. Approximately 65% of tracks are composed in a major key, as indicated by the distribution of the `mode` variable.

The `tempo` feature spans an extensive range, from near zero to over 200 BPM, with a dataset-wide mean of approximately 122 BPM. Values close to zero likely reflect metadata inconsistencies or silence-dominated recordings. Similarly, extreme values in `loudness`—such as entries around -50 dB—may result from parsing errors or low-quality source material. While these anomalies have been retained to avoid artificial smoothing, they may influence model sensitivity to tempo and amplitude-related features.

3.4 Feature Distribution by Emotion Category

To examine how acoustic features vary across emotion categories, I generated boxplots of all 12 Spotify-derived features grouped by emotion label. Figures 2–4 highlight three of the most informative features: `acousticness`, `valence`, and `energy`. Additional plots are provided in Appendix C.

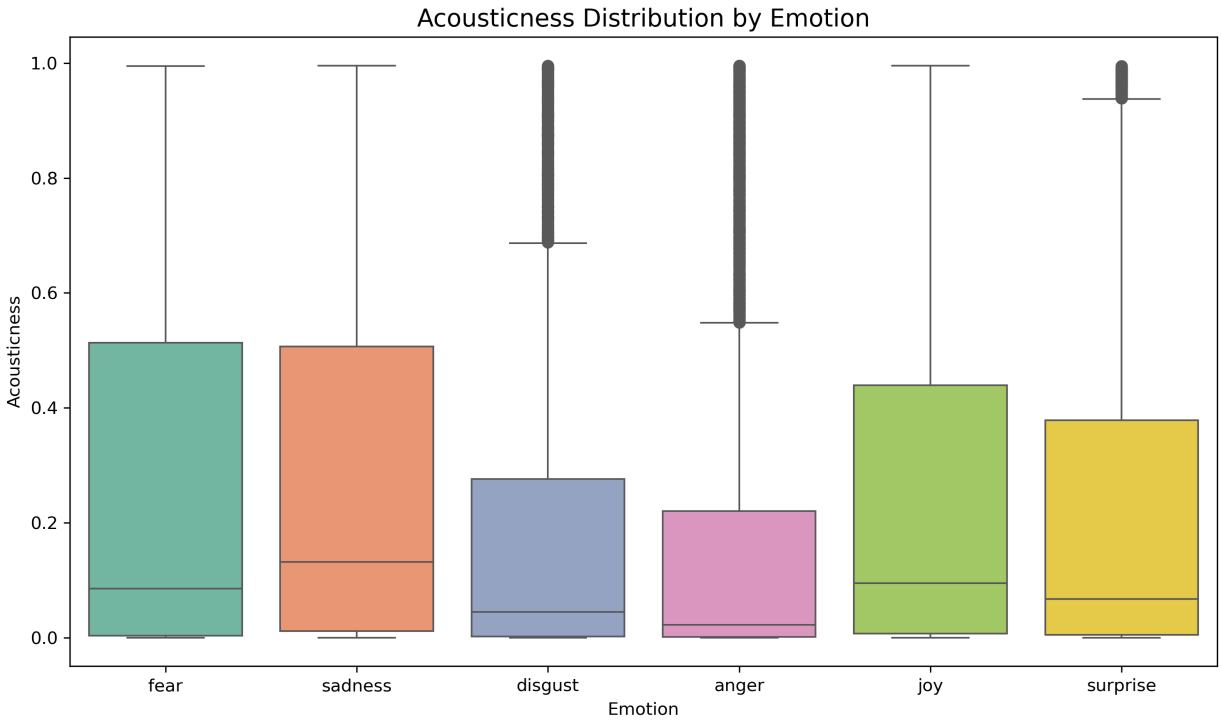

Figure 2: Distribution of `acousticness` across emotion categories.

`Acousticness` shows marked variation: tracks labeled with *fear* and *sadness* tend to be more acoustic, while *anger* and *disgust* skew toward lower values.

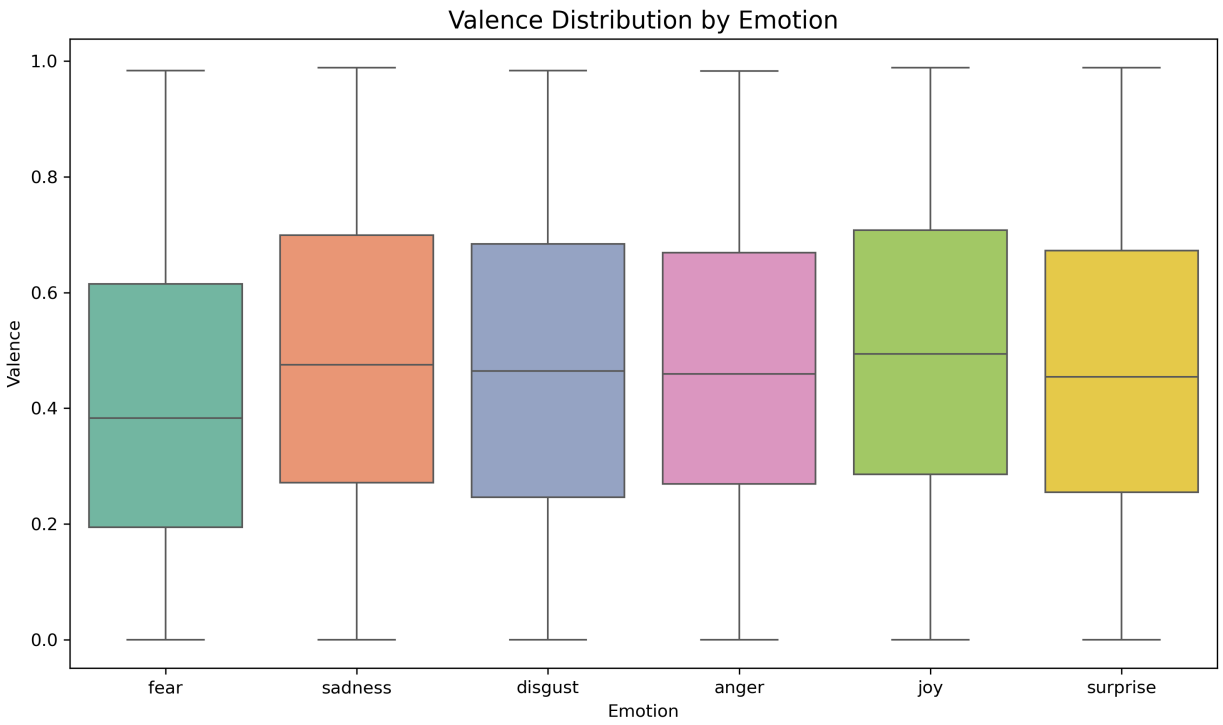

Figure 3: Distribution of `valence` across emotion categories.

`Valence` aligns well with emotional polarity—*joy* and *surprise* are associated with high valence, whereas *fear*, *sadness*, and *disgust* cluster near neutral or negative values.

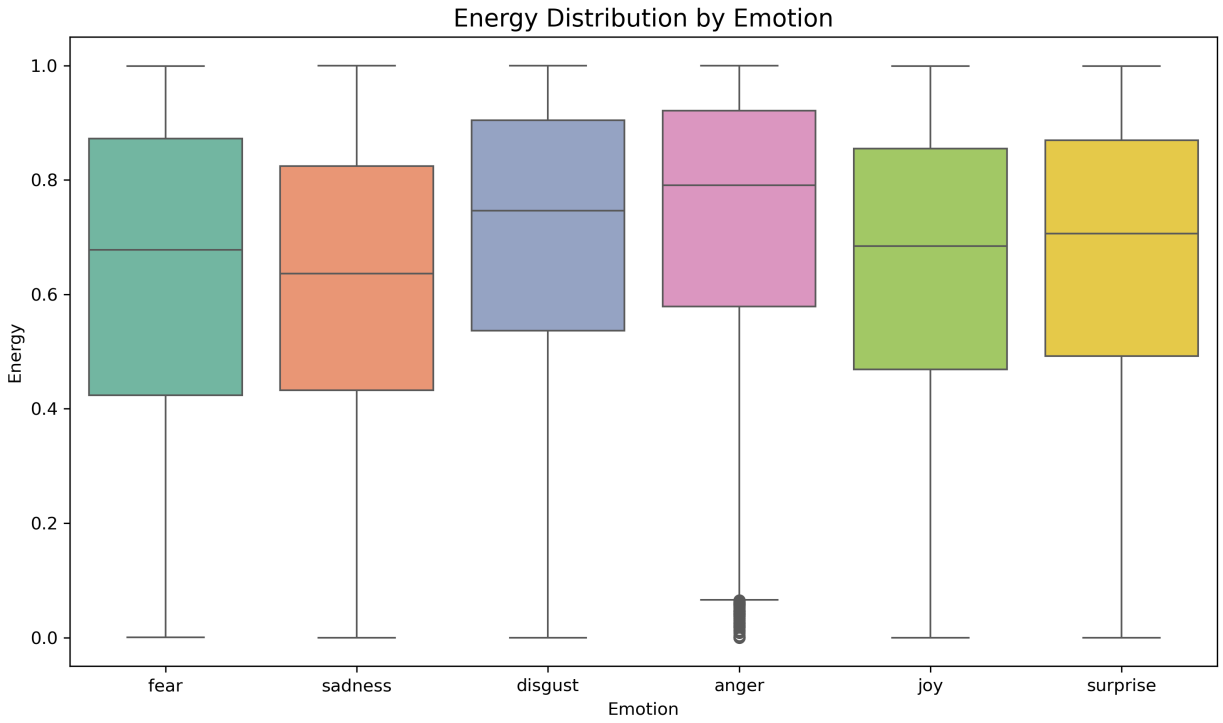

Figure 4: Distribution of `energy` across emotion categories.

`Energy` supports an arousal-based interpretation: high-energy tracks are linked to emotions like *anger* and *surprise*, while lower energy levels are found in *fear* and *sadness*.

In contrast, features such as `key`, `mode`, and `time signature` exhibit minimal emotional separation, suggesting limited utility for predictive modeling. Appendix C contains distributions for other 9 features to support further inspection and model design.

3.5 Visual Overview of the Data Pipeline

To ensure transparency and reproducibility, Figure 5 illustrates the unified data pipeline. It shows the current (DistilRoBERTa-based) label generation paths, which converge into a shared pipeline for metadata matching, Spotify ID retrieval, and audio feature extraction. This visual summary clarifies how weakly supervised labels propagate through the system and contextualizes the dataset's construction process.

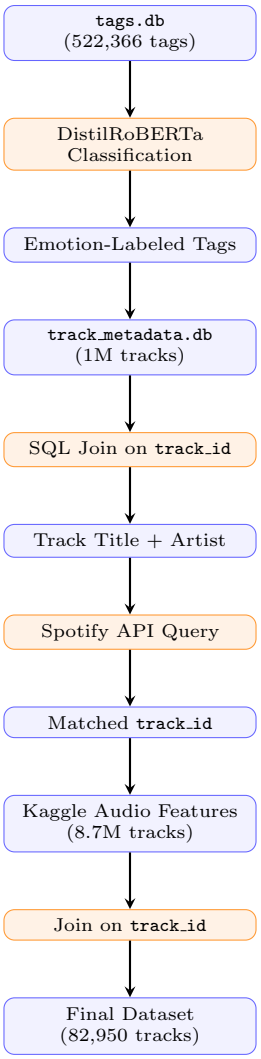

Figure 5: Emotion labeling and dataset construction pipeline.

3.6 Data and Code Availability

All data and code used in this study are available at:

`https://github.com/macs30200-s23/course-project-VeraMao`.

4 Results

4.1 Baseline Model Comparison

To establish benchmark performance for multi-label music emotion classification, this study implemented and evaluated three commonly used classifiers: Multi-Layer Perceptron (MLP), Random Forest, and K-Nearest Neighbors (KNN). All models were trained on the same 12-dimensional audio feature set and evaluated using a top-k dynamic matching strategy, where k corresponds to the number of true labels per instance. This ensures a fairer assessment in the multi-label context and avoids penalizing models for predicting multiple

relevant emotions.

Overall Performance. Among the three models, **Random Forest** achieved the highest micro-average F1 score (**0.6766**) and exact match accuracy (**36.67%**). *Exact match accuracy* refers to the proportion of instances for which the predicted emotion set exactly matches the true multilabel set. This is a strict metric in multi-label classification, and a score of 36.67%, while seemingly modest, is considered relatively strong given the task's complexity and the presence of overlapping emotional labels.

MLP followed closely with an F1 score of 0.6757, benefiting from its ability to capture nonlinear interactions among features. In contrast, **KNN** underperformed across all metrics, particularly in exact match accuracy (30.30%) and Hamming Loss (0.2565), likely due to its sensitivity to feature scaling and difficulty modeling sparse multilabel outputs in high-dimensional acoustic space.

Metric Rationale. Micro-averaged F1 score was selected as the primary evaluation metric because it accounts for all true positives, false positives, and false negatives across the dataset. Unlike macro F1, which treats all classes equally regardless of frequency, micro F1 provides a more balanced estimate in the presence of class imbalance, which is particularly relevant for music emotion classification where categories like *joy* are overrepresented while *fear* and *disgust* are less common.

Hamming Loss was included to capture the average number of incorrect label assignments per sample. It penalizes both false positives and false negatives uniformly, offering a complementary view of model performance at the label level. A lower Hamming Loss indicates more precise multi-label predictions. In our results, Random Forest produced the lowest Hamming Loss (0.2261), reinforcing its advantage in minimizing unnecessary or missed predictions.

Table 8: Overall performance metrics for three baseline classifiers under top-k dynamic matching.

Model	Micro F1	Exact Match (%)	Hamming Loss	Macro F1	Weighted F1
Random Forest	**0.6766**	**36.67**	**0.2261**	0.5379	0.6547
MLP	0.6757	36.27	0.2268	0.5187	0.6436
KNN	0.6332	30.30	0.2565	0.5098	0.6188

Per-Class Analysis. As shown in Figure 6, all models exhibit strong performance on high-frequency emotions like *joy* and *sadness*, with F1 scores exceeding 0.75 across classifiers. In contrast, minority or more ambiguous categories such as *fear*, *disgust*, and *surprise* consistently receive much lower F1 scores—often below 0.40—regardless of model type.

This discrepancy is further clarified in Figure 7, which breaks down the Random Forest classifier's precision, recall, and F1 by emotion. While precision remains moderately stable (typically 0.5–0.6) even for rare emotions, recall is significantly lower, indicating systematic under-prediction. For example, *fear* achieves only 0.22 recall despite a precision of 0.55, and *surprise* reaches just 0.28 recall.

Such imbalances between precision and recall highlight a key challenge in multilabel music emotion recognition: models tend to be conservative in assigning underrepresented labels, often missing relevant instances. These results align with prior findings on label imbalance and weak supervision in affective classification tasks (Ahsan, Kumar, and Jawahar 2015; Uplabdhee et al. 2025).

For completeness, per-class performance matrices for KNN and MLP are provided in Appendix D. These additional visualizations reinforce the observed trend: minority classes suffer from limited recall and are often overshadowed by more frequently occurring emotions.

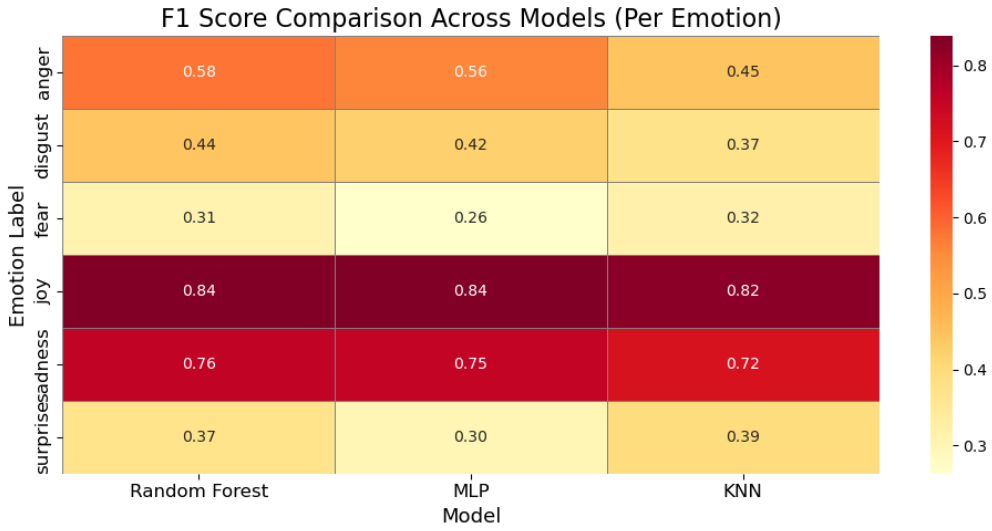

Figure 6: F1 score comparison across emotion categories for three models (Random Forest, MLP, KNN). Joy and Sadness are consistently well-predicted, while minority emotions like Fear and Disgust show much lower F1 performance.

Figure 7: Per-class precision, recall, and F1-score for the Random Forest classifier. Joy and Sadness dominate predictive accuracy, while Fear and Surprise reveal low recall despite moderate precision.

4.2 Emotion Label Co-occurrence Patterns

To elucidate structural dependencies within the multilabel annotation space, a pairwise co-occurrence matrix was computed across all emotion categories. As shown in Figure 8, certain emotion pairs co-occur with high frequency, reflecting common affective blends in listener perceptions. Notably, *joy* and *sadness* co-occur

in over 15,000 tracks, suggesting that many songs convey emotionally ambivalent or bittersweet tones. In contrast, combinations such as *surprise* and *fear* are comparatively rare, with fewer than 1,200 instances.

These patterns reveal several semantically proximal pairings—e.g., *joy* and *surprise*, or *fear* and *sadness*—that are prone to overlap, contributing to inter-label ambiguity. This ambiguity, in turn, may partially explain classifier confusion and low recall on minority categories. The findings underscore the importance of interpretability-driven evaluation in multi-emotion classification, especially for understanding errors arising from overlapping affective semantics.

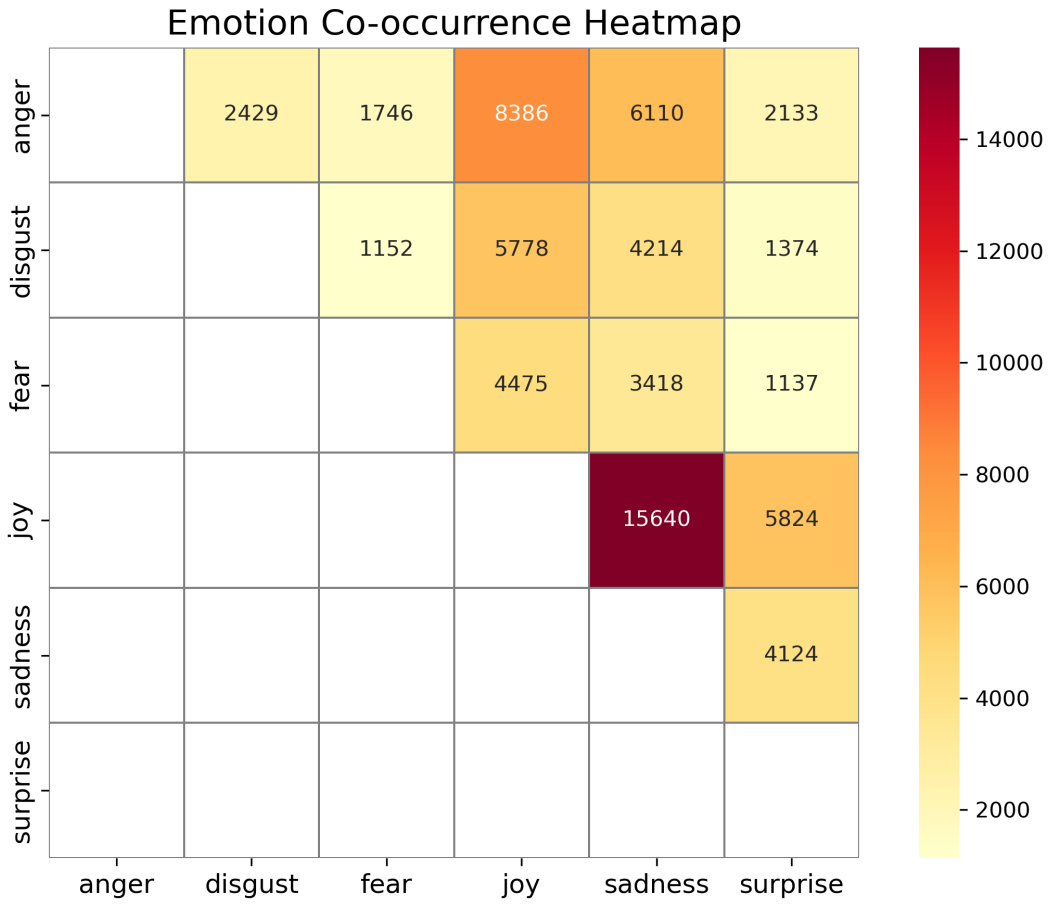

Figure 8: Co-occurrence matrix of emotion labels across multilabel annotations. Values reflect the number of tracks assigned each emotion pair.

4.3 Global and Per-Label Feature Importance Analysis

To assess how different architectures prioritize audio cues in multilabel classification, I computed global and class-specific feature importance across representative model types, including tree-based, distance-based, and neural network classifiers.

Methodology. For linear and distance-based models, permutation importance was estimated using a `OneVsRestClassifier` wrapper and averaged over emotion classes. Tree-based models provided intrinsic importance values (`feature_importances_`), while neural architectures were interpreted via SHAP's Kernel-Explainer applied to a stratified subsample of 100 training examples. Mean absolute SHAP values were used

to quantify both global and emotion-specific contributions.

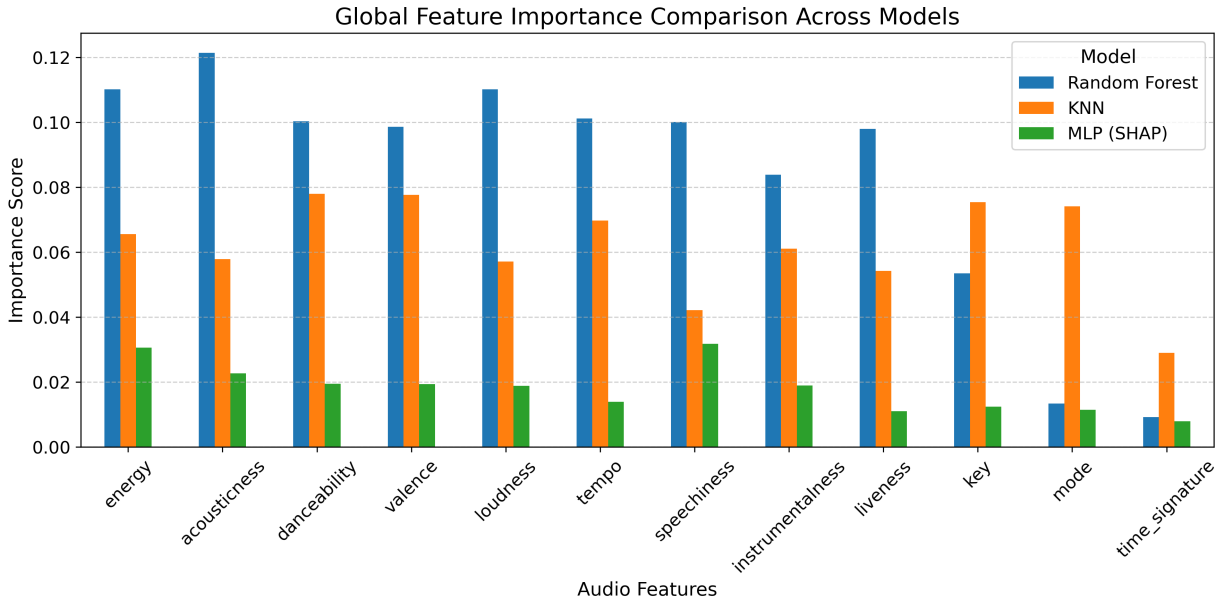

Figure 9: Global feature importance scores across models. MLP importance computed via SHAP; Random Forest and KNN via permutation; Logistic Regression via normalized coefficients.

As shown in Figure 9, Random Forest emphasized `acousticness`, `loudness`, and `energy`—features with sharp thresholds well-suited to tree-based decision boundaries. The high importance of `acousticness` (0.1214) aligns with its role in differentiating unplugged and high-arousal music, consistent with prior affective MIR findings (Huron 2015; McCraty et al. 1998).

KNN placed more weight on `danceability`, `valence`, and `mode`, suggesting that similarity-based classification is more influenced by rhythmic and tonal characteristics—features that likely produce tighter clusters in Euclidean space.

MLP showed more balanced attribution across features, reflecting its ability to capture nonlinear combinations. SHAP results indicated `speechiness`, `energy`, and `acousticness` as particularly influential. These dimensions may encode expressive subtleties (e.g., rap-like vocals or whispered tones) that are difficult for tree-based or distance-based models to capture.

Across all models, `key`, `mode`, and `time_signature` consistently received low importance scores—consistent with the weak correlation patterns observed earlier. These features likely contribute little predictive value in large, genre-diverse datasets and may only matter in specialized musical contexts.

Per-Label Feature Importance. To examine model behavior at the class-specific level, I compared feature importance rankings for the emotion *anger* across three classifiers—Random Forest, KNN, and MLP (SHAP-based).

Anger. All three models agree that `acousticness`, `loudness`, and `energy` are key predictors of *anger*. Random Forest assigns the highest importance to `acousticness`, followed closely by `loudness` and `energy`, suggesting that anger is associated with low-acoustic, high-intensity audio signatures. KNN also emphasizes these features but disproportionately relies on `key` and `mode`, indicating sensitivity to tonal structure that may be confounded by genre or harmonic context. MLP, as revealed by SHAP values, distributes importance more

diffusely, with `valence`, `acousticness`, and `loudness` emerging as modest contributors. The consistently high importance of `tempo` and `danceability` across models reinforces the link between anger and rhythmic drive.

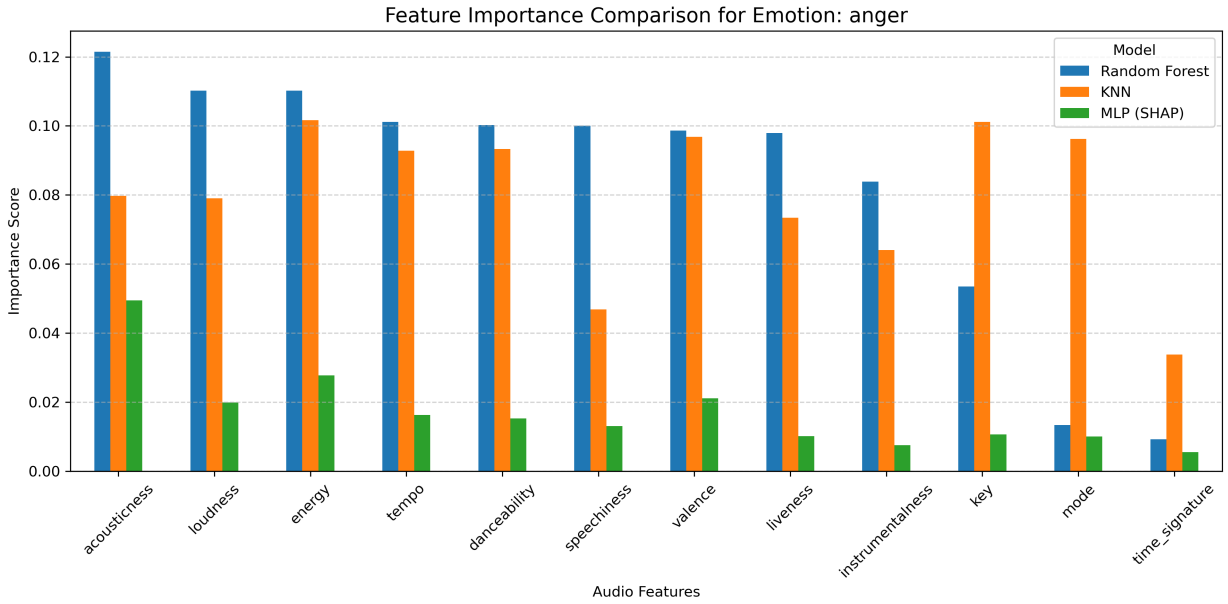

Figure 10: Feature importance comparison for emotion: *anger*.

Feature importance profiles for the remaining emotions (*disgust*, *fear*, *joy*, *sadness*, and *surprise*) are provided in Appendix E. These extended plots and comparisons reveal both shared and emotion-specific acoustic signatures. For instance, high-arousal features such as `energy` and `loudness` consistently appear among the top predictors for *anger*, *joy*, and *surprise*, while low-valence emotions like *fear* and *sadness* are more strongly associated with `acousticness` and `valence`. Additionally, `speechiness` emerges as a reliable cue for *disgust* and *joy*, suggesting that lyrical or spoken-word content may play an outsized role in those categories. Conversely, features like `key`, `mode`, and `time_signature` contribute minimally across all emotions, reinforcing earlier findings from the correlation and distributional analyses. Overall, these per-label comparisons help clarify which acoustic dimensions are broadly salient versus those that operate more selectively, thereby motivating the need for feature-wise ablation and error decomposition in later sections.

Together, these findings reinforce the value of per-label interpretation in multi-label settings. They demonstrate that while some features are broadly predictive, others contribute selectively, depending on the emotional context. This justifies the subsequent ablation and correlation-based redundancy analyses in the following.

4.4 Feature Correlation Analysis

To evaluate potential collinearity among the input variables, a Pearson correlation matrix was computed and visualized via heatmap (Figure 11). This diagnostic offers insight into linear dependencies across audio features and informs subsequent steps in feature selection, normalization, and model interpretation.

Several strong correlations emerged. Most notably, `acousticness` exhibits substantial negative associations with both `energy` ($r = -0.72$) and `loudness` ($r = -0.58$), suggesting that acoustically rich tracks tend to be quieter and less energetic. Conversely, `energy` and `loudness` are strongly positively correlated ($r = 0.76$), consistent with psychophysical theories linking amplitude to perceived arousal. A moderate positive correlation between `danceability` and `valence` ($r = 0.55$) suggests that rhythmically engaging music is more likely to convey positive emotional valence.

In contrast, categorical or structural features such as `key`, `mode`, and `time_signature` display weak or near-zero correlations with other variables. Their limited linear association implies that any emotional signal they encode likely arises through nonlinear interactions or genre-conditioned effects.

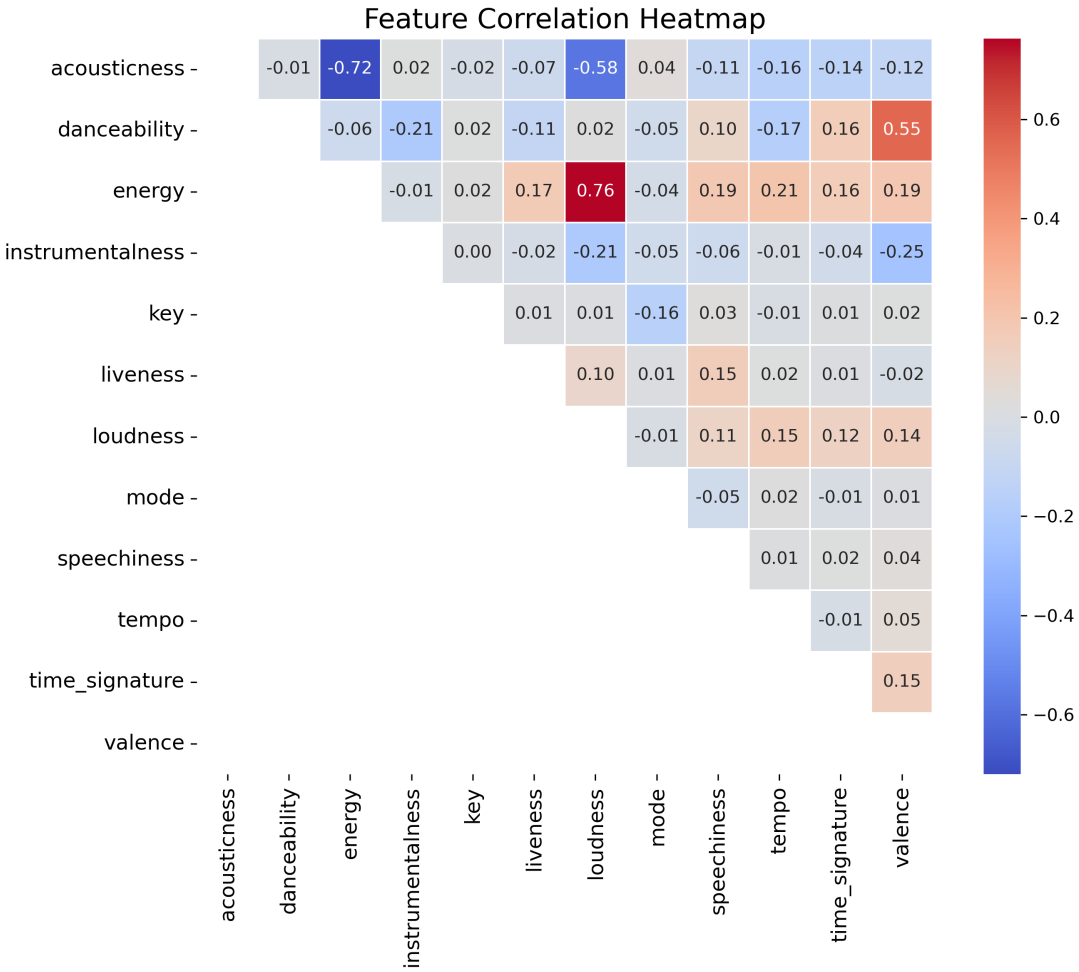

Figure 11: Pearson correlation matrix of the 12 Spotify audio features. High correlations indicate potential redundancy or shared perceptual dimensions.

These findings underscore the limitations of purely linear diagnostics in emotion modeling. To address potential non-additive and interaction-based effects, the modeling pipeline incorporates permutation-based and SHAP-based importance methods, which provide model-aware assessments of feature contributions.

4.5 Ablation Study and Dimensionality Considerations

To evaluate feature redundancy and quantify the marginal contribution of each acoustic variable, I conducted a stepwise ablation study using the Random Forest classifier. Features were removed iteratively in ascending order of global importance (as determined by the model's global feature importance), and the classifier was retrained and evaluated at each step using the multilabel dynamic top-k prediction framework.

Remaining Features	Accuracy	Dropped Features
12	0.2403	[]
11	0.2361	[valence]
10	0.2361	[valence, loudness]
9	0.2182	[valence, loudness, instrumentalness]
8	0.2069	[valence, loudness, instrumentalness, energy]
7	0.1904	[valence, loudness, instrumentalness, energy, acousticness]
6	0.1890	[valence, loudness, instrumentalness, energy, acousticness, danceability]
5	0.1932	[valence, loudness, instrumentalness, energy, acousticness, danceability, speechiness]
4	0.1767	[valence, loudness, instrumentalness, energy, acousticness, danceability, speechiness, liveness]
3	0.2201	[..., tempo]
2	0.2347	[..., key]
1	0.2370	[..., mode]

Table 9: Sequential ablation results. Features are removed based on ascending global importance. Accuracy reflects dynamic top-k prediction performance.

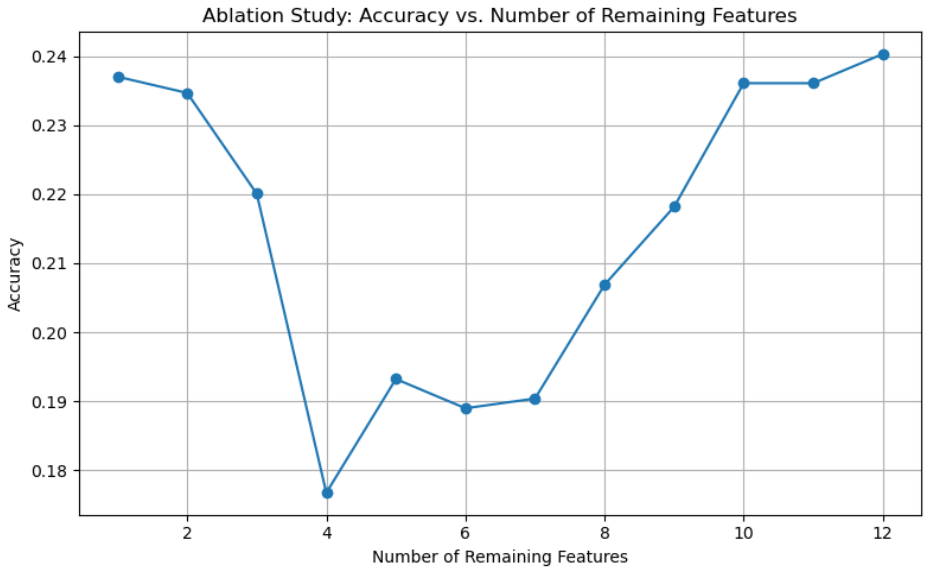

Figure 12: Impact of progressive feature removal on Random Forest classifier accuracy.

Model accuracy decreased sharply upon removing the top five most informative features—`acousticness`, `energy`, `loudness`, `valence`, and `instrumentalness`—dropping from 0.2403 to 0.1767 with only four features retained. This trend highlights that emotional signal is concentrated in a core subset of high-importance acoustic dimensions.

Interestingly, slight performance recovery was observed after excluding low-importance features (e.g., `key`, `mode`, `time_signature`), suggesting that these dimensions may introduce noise or irrelevant variance. These observations align with prior correlation and importance analyses that found minimal contribution from structural features.

PCA Considerations. To further examine the role of dimensionality reduction, I retrained the Random Forest classifier using the original 12-feature input space without PCA transformation. This configuration yielded improved metrics: the micro-averaged F1 score increased to **0.690**, exact match accuracy rose to **38.77%**, and Hamming Loss declined to **0.2166**. Gains were especially notable in underrepresented classes such as *disgust* and *fear*, where precision increased to 0.57.

These results indicate that PCA-based compression may inadvertently discard subtle, emotion-relevant

variance that benefits fine-grained multilabel classification. Accordingly, all final models reported in this study employ the full, unreduced feature set.

4.6 Per-Class Metrics and Misclassification Patterns

Per-class performance varies substantially across emotion categories. To further examine systematic misclassifications, Figure 13 presents a heatmap for the Random Forest model. Notable error patterns include frequent confusion between *Sadness* and *Joy*, as well as underprediction of *Fear* and *Surprise*, both of which are often omitted entirely. These patterns suggest that certain emotional categories possess overlapping acoustic profiles or are more context-dependent, increasing classification difficulty.

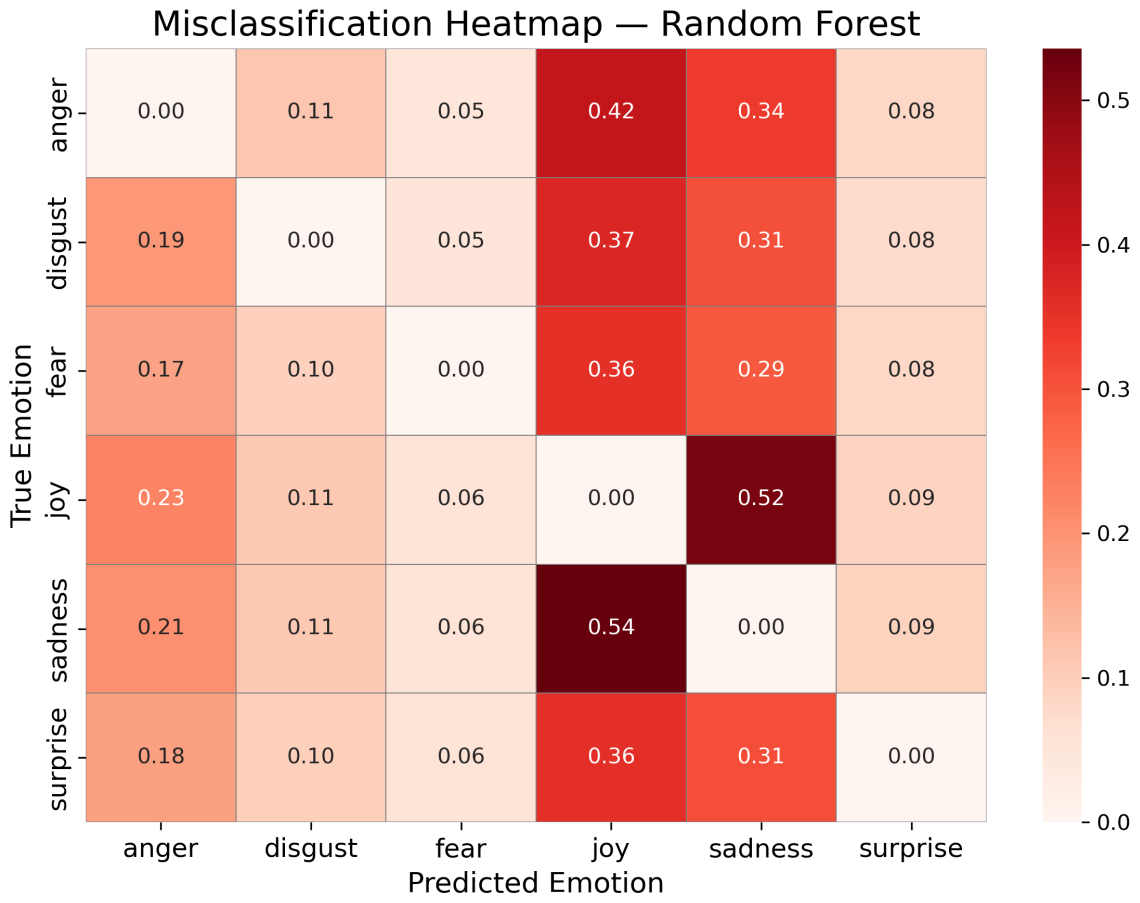

Figure 13: Misclassification heatmap for the Random Forest model. Rows represent true emotion labels; columns represent predicted labels.

Additional misclassification heatmaps for KNN and MLP models are provided in Appendix F.

4.7 Impact of Emotion Co-occurrence

Emotion misclassification may also stem from multilabel co-occurrence effects. To investigate this, I computed the average number of co-labels per instance for each emotion and plotted it against that emotion's false negative rate. As shown in Figure 14, a clear inverse relationship emerges: emotions with higher co-occurrence,

such as *Surprise* and *Fear*, tend to be missed more frequently. In contrast, *Joy* is both distinctive and frequently isolated, yielding lower omission rates. This trend highlights a key challenge in multilabel classification: emotions embedded within complex affective contexts are more difficult to isolate during prediction.

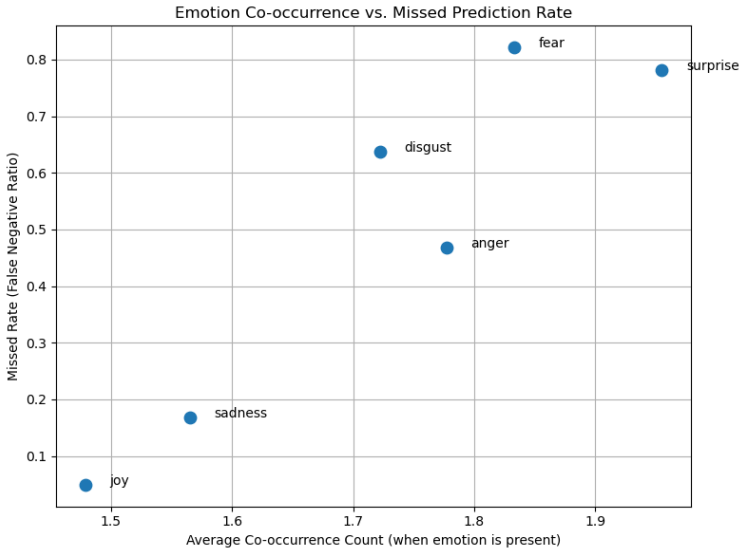

Figure 14: Relationship between average emotion co-occurrence and false negative rate. Emotions that frequently appear alongside others (e.g., *Surprise*, *Fear*) are more likely to be entirely missed.

4.8 SHAP-Based Interpretation of Misclassified Samples

To investigate why certain emotions remain difficult to classify, I conducted a SHAP-based analysis on one misclassified instance per emotion. These samples were selected from predictions that fell just below their respective dynamic top-k confidence thresholds. Each waterfall plot visualizes how audio features contributed to the model's decision, providing insight into failure mechanisms.

Figure 15 illustrates a misclassified *fear* example with a predicted confidence of 0.124. SHAP values show that positive contributions from `speechiness` were neutralized by strong negative effects from `danceability`, `valence`, and `instrumentalness`. The low magnitude and contradictory directions of these contributions indicate the absence of a dominant predictive signal, resulting in indecision.

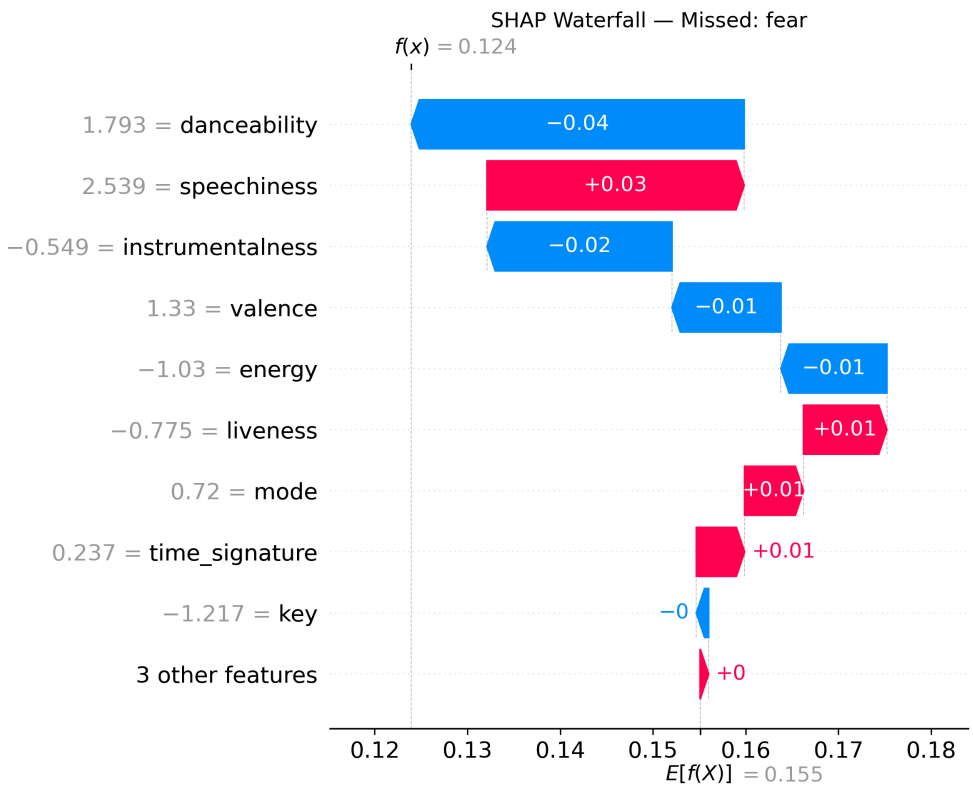

Figure 15: SHAP explanation for misclassified *Fear*. Contributions from key features failed to reach decision threshold.

In contrast, the misclassified *joy* sample (Figure 16) shows strong suppression by low `speechiness` and `tempo`, despite high values in `energy` and `danceability`. This highlights how absence of expected lyrical and rhythmic signals can override otherwise high-arousal acoustic profiles.

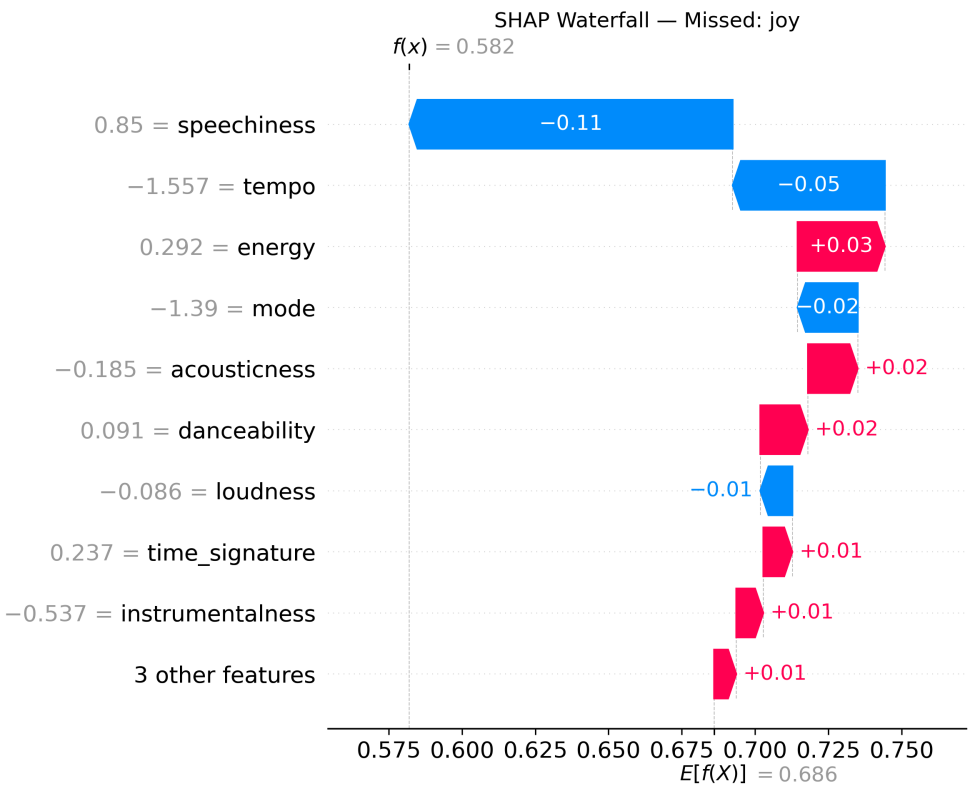

Figure 16: SHAP explanation for misclassified *Joy*. Negative contributions from low `speechiness` and `tempo` suppressed correct prediction.

SHAP analyses for the remaining emotions (*anger*, *disgust*, *sadness*, *surprise*) are provided in Appendix G, with a typology of common failure modes summarized in Table 10.

Failure Mode	**Emotion**	**Explanation**
Overlapping semantics	Sadness → Joy	Bittersweet or nostalgic tone interpreted as positive due to instrumentation
Lack of dominant cues	Fear	Weak signal across all features, leading to model indecision
Feature conflict	Disgust	Noisy combination of low valence but ambiguous rhythmic/timbral signals
Rhythmic mismatch	Anger, Joy	Tempo and time signature inconsistent with model prototype of class
Cultural encoding	Surprise	Subtle emotional valence varies by genre; high acousticness misinterpreted

Table 10: Typology of common failure modes in multilabel music emotion classification.

5 Discussion

This study investigated how supervised machine learning models interpret Spotify-derived acoustic features when classifying songs into emotional categories, within a weakly supervised, multilabel framework. Beyond measuring performance, the analysis prioritized interpretability—focusing on both global feature importance

and local explanation of prediction failures.

Emotion-specific classification disparities. Model performance was notably uneven across emotion categories. *Joy* and *Sadness* consistently achieved high F1 scores, while *Fear*, *Disgust*, and *Surprise* remained error-prone. Heatmaps and label co-occurrence diagnostics suggest that such failures often stem from intrinsic ambiguity in the emotional signal, rather than model incapacity. For example, *Fear* was frequently confused with both *Sadness* and *Joy*, likely due to overlapping acoustic profiles. This aligns with the observed inverse relationship between co-occurrence frequency and false negative rate: emotions that often appear in combination—such as *Surprise* and *Fear*—are more likely to be missed, suggesting that models struggle to isolate context-dependent affective cues.

SHAP-based explanation of failure modes. Local interpretability analyses revealed that many misclassifications arose not from strongly misleading features, but from diffuse, contradictory, or weak signals. In the case of *Fear*, for instance, SHAP values across all features were near zero, leading to model indecision. These cases exemplify what might be called a "semantic vacuum," in which the audio signal lacks a dominant pattern associated with a specific emotion. This finding, echoed in prior work on weakly supervised tags (Artemova et al. 2025; Hannah Kim et al. 2024), reinforces the importance of considering emotional ambiguity as a first-class modeling challenge.

Feature contribution and model heterogeneity. Global importance scores identified perceptually salient features—such as `energy`, `valence`, and `acousticness`—as central to emotion inference. These align with established psychological models of arousal and positivity (Huron 2015; McCraty et al. 1998). However, models differed in how they utilized these cues: MLP distributed weights more evenly and captured nonlinear interactions, while KNN relied on cluster-inducing variables like `mode` and `danceability`, reflecting its sensitivity to local structure.

Feature redundancy and robustness. The ablation study showed that emotional prediction is driven by a compact but semantically rich feature subset. Removing the top five most informative features resulted in a 26% drop in accuracy. Notably, performance partially rebounded when low-importance features—such as `key`, `mode`, and `time_signature`—were removed, indicating that noise-prone or low-variance features may introduce instability. Correlation analysis confirmed that while `energy` and `loudness` are strongly correlated ($r = 0.76$), they contribute unique, nonlinear effects—supporting their joint inclusion despite collinearity.

Failure typologies and cultural encoding. SHAP analyses of misclassified samples revealed recurring patterns. For instance, *Disgust* lacked consistent rhythmic or timbral anchors, while *Surprise* was often misinterpreted as *Sadness* or *Fear* in genres with high `acousticness`, such as jazz. These cases suggest that cultural and genre-specific norms around emotional expression can confound audio-based inference, especially when models lack access to lyrical, historical, or listener context.

Epistemological ambiguity in "error." Manual inspection of borderline cases complicates the notion of model failure. For instance, Billie Holiday's *All of Me* was labeled with *Anger*, *Sadness*, and *Joy*, yet only *Joy* was predicted. While technically incorrect, this prediction is arguably valid—underscoring a key limitation in multilabel music emotion modeling: the blurred boundary between algorithmic misclassification and human perceptual disagreement. Crowd-sourced or weakly labeled ground truth may reflect collective sentiment more than categorical truth.

Methodological contributions. Finally, the study illustrates the value of combining global and local interpretability tools. Permutation importance, SHAP waterfall plots, and co-occurrence diagnostics each revealed distinct failure mechanisms. This multi-angle approach offered insights that single-metric evaluation would have missed—particularly in diagnosing borderline, ambiguous, or context-dependent predictions. As such, interpretability emerges not just as a supplement to performance, but as a diagnostic lens critical for refining emotion-aware systems.

5.1 Limitations

Despite its contributions, this study has several limitations concerning label validity, feature representation, interpretability methods, and evaluation strategy.

The most fundamental limitation lies in the use of automatically generated emotion labels rather than human-annotated ground truth. Labels were inferred from a DistilRoBERTa classifier applied to Last.fm user tags, many of which are genre-descriptive, polysemous, or aesthetically focused (e.g., *lo-fi*, *dream pop*) rather than affective. While transformer-based tag-to-emotion mapping improves semantic coverage, its reliability for music emotion classification remains under-validated. This introduces the risk of semantic drift—where labels reflect inferred sentiment rather than listener-perceived emotion. Moreover, since the same weak labels were used for both model training and evaluation, performance metrics may reflect alignment with the labeling model rather than true generalization.

The use of a dynamic top-k prediction threshold introduces further abstraction. Although it respects label cardinality and improves match quality, it imposes a hard cutoff that excludes borderline yet semantically plausible predictions. This especially penalizes ambiguous or co-occurring emotions such as *fear* and *surprise*, which often fall below the threshold despite moderate model confidence—compromising recall for minority classes.

The emotional taxonomy itself presents constraints. The six-class schema aligns with basic emotion theory but underrepresents the diversity and fluidity of musical affect. Music frequently conveys blended, evolving, or culturally-specific emotions that defy discrete categorization. Rigid label targets may therefore understate affective nuance. Future work should consider fuzzy labeling schemes or continuous representations (e.g., valence-arousal space) to better capture this complexity.

The fixed threshold of 0.8 for emotion label assignment also introduces bias. While this value prioritizes high-confidence tags and reduces semantic noise, it remains an arbitrary hyperparameter not systematically tuned. Appendix A reports a follow-up analysis suggesting that a 0.7 threshold offers improved label coverage—especially for underrepresented emotions like *anger*, *fear*, and *surprise*. However, the downstream effects of threshold choice warrant further empirical validation via full relabeling and ablation studies.

Metadata limitations further constrain dataset completeness. Due to API mismatches, spelling inconsistencies, and missing Spotify metadata, many Last.fm tracks could not be matched. The final dataset of 82,950 songs overrepresents mainstream and Western music, limiting generalizability to niche genres or global musical traditions.

Feature representation also limits affective fidelity. Spotify's 12 audio features capture coarse acoustic properties but omit structural, lyrical, and cultural cues—such as melodic contour, harmonic tension, and lyrical sentiment—that shape listener emotion. As a result, models may over-rely on surface-level features (e.g., `loudness`, `tempo`) whose emotional interpretation is genre-dependent. For instance, high `acousticness` may connote intimacy in folk but surprise or playfulness in jazz.

The interpretability tools used in this study, though valuable, have technical constraints. SHAP's KernelExplainer is computationally intensive and requires small reference sets, limiting scalability. SHAP values are also sensitive to feature collinearity—making importance scores less stable when features like `energy` and `loudness` co-vary. Moreover, axis-aligned models like Random Forests may overemphasize certain features due to their split mechanics, complicating cross-model comparisons.

Dimensionality reduction (e.g., PCA) was intentionally excluded to preserve interpretability. While this helped retain signals for rare emotions such as *disgust*, it may have also preserved noisy or redundant features. Future pipelines could benefit from SHAP-guided pruning or domain-aware selection to balance interpretability and robustness.

The evaluation strategy itself is impacted by labeling incompleteness. Since many songs carry only partial emotion tags, true positive predictions may be penalized as false positives. This weakens the interpretability of metrics like F1 and Exact Match, especially for songs with multilayered or genre-dependent emotional

content.

Finally, cultural and stylistic biases embedded in both the training tags and model predictions remain unresolved. Tags such as "dark ambient" may be mapped to *fear* or *sadness*, although the user's intent may be aesthetic, not emotional. Likewise, genre conventions shape how emotions are acoustically encoded—yet current features are not designed to capture cross-cultural affective variation. Without genre-aware modeling or cultural grounding, classifiers risk overfitting to dominant musical norms and failing to generalize meaningfully.

Addressing these limitations will require multimodal inputs (e.g., lyrics, harmonic embeddings, listener surveys), genre- and culture-aware model designs, and benchmark datasets with verified annotations. Integrating human coding into the labeling pipeline and evaluating sensitivity to threshold and taxonomic design will be essential for building more valid and inclusive emotion recognition systems.

6 Conclusion

This study examined the extent to which supervised machine learning models can classify musical emotion using only Spotify-derived acoustic features. Leveraging a weakly supervised labeling pipeline—where emotion labels were inferred from Last.fm tags via a DistilRoBERTa transformer—the resulting dataset enabled multilabel classification across six basic emotions: anger, disgust, fear, joy, sadness, and surprise.

Among the models evaluated (Random Forest, Logistic Regression, K-Nearest Neighbors, and Multi-Layer Perceptron), Random Forest achieved the highest micro-average F1 and exact match scores. This performance likely reflects both its suitability for thresholdable, high-variance features (e.g., `acousticness`, `energy`, `loudness`) and its advantage from full hyperparameter tuning. However, model performance varied substantially by emotion: *Joy* and *Sadness* were consistently predicted with high confidence, whereas *Fear*, *Disgust*, and *Surprise* posed persistent challenges—often due to low acoustic salience, semantic overlap, and frequent co-occurrence with other emotions.

Interpretability emerged as a key analytical lens. Global feature importance scores revealed divergent inductive biases: Random Forest favored separable, high-variance features; MLPs leveraged nonlinear feature combinations; KNN relied on local clustering. SHAP-based local interpretations revealed that many misclassifications stemmed not from erroneous inputs but from ambiguous or weakly expressive acoustic profiles—highlighting the challenge of modeling fuzzy affective categories from audio alone.

The ablation study confirmed the predictive centrality of a small set of perceptually salient features. Removal of top contributors—`acousticness`, `energy`, `valence`, among others—led to a substantial accuracy drop, underscoring their role in affective inference. At the same time, the elimination of weak or noisy features (e.g., `key`, `mode`) modestly improved robustness, suggesting that excessive feature inclusion may hinder generalization.

Several limitations warrant caution. The emotion labels were inferred rather than empirically annotated, introducing potential biases from genre conventions or tag semantics. Top-k prediction thresholds improved interpretability but suppressed emotional plurality, especially for minority or co-labeled emotions like *Fear* and *Surprise*. Metadata mismatches and API failures further constrained dataset completeness, potentially skewing the corpus toward English-language and Western musical traditions. Finally, SHAP explanations—while insightful—remain sensitive to model architecture and training dynamics, limiting direct cross-model comparisons.

Despite these constraints, the study affirms that Spotify audio features encode meaningful emotional structure. Yet, acoustic data alone is insufficient for high-fidelity emotion classification. Future research should adopt a multimodal framework, integrating lyrics, melodic contour, harmonic progression, and temporal structure. Emotion annotations should reflect probabilistic, fuzzy, or listener-rated constructs rather than discrete labelings. Cross-cultural and genre-aware model design will also be crucial to ensure generalizability and inclusivity.

In sum, while supervised learning over acoustic features offers a promising foundation for music emotion recognition, achieving deeper affective fidelity demands models that move beyond sound—toward context, semantics, and human subjectivity. Doing so not only improves classification fidelity, but also enhances the utility of emotion-aware systems in downstream applications such as personalized music recommendation, affective computing, music therapy, and cultural analytics.

References

Acheampong, Francisca Adoma, Henry Nunoo-Mensah, and Wenyu Chen (Dec. 2021). "Transformer models for text-based emotion detection: a review of BERT-based approaches". en. In: *Artificial Intelligence Review* 54.8, pp. 5789–5829. ISSN: 0269-2821, 1573-7462. DOI: 10.1007/s10462-021-09958-2. URL: https://link.springer.com/10.1007/s10462-021-09958-2 (visited on 04/26/2025).

Ahsan, Hiba, Vijay Kumar, and C.V. Jawahar (Jan. 2015). "Multi-label annotation of music". In: *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*. Kolkata, India: IEEE, pp. 1–5. ISBN: 978-1-4799-7458-0. DOI: 10.1109/ICAPR.2015.7050685. URL: http://ieeexplore.ieee.org/document/7050685/ (visited on 05/14/2025).

Artemova, Ekaterina et al. (Jan. 2025). *Hands-On Tutorial: Labeling with LLM and Human-in-the-Loop*. arXiv:2411.04637 [cs]. DOI: 10.48550/arXiv.2411.04637. URL: http://arxiv.org/abs/2411.04637 (visited on 04/21/2025).

Bandhakavi, Anil et al. (Jan. 2017). "Lexicon Generation for Emotion Detection from Text". In: *IEEE Intelligent Systems* 32.1, pp. 102–108. ISSN: 1941-1294. DOI: 10.1109/MIS.2017.22. URL: https://ieeexplore.ieee.org/abstract/document/7851145?casa_token=BP2dXjPKQEMAAAAA:WQKU9husdL6_1NViXxmikxT2WNKZjLOUuIwpCERQyoEBOYrjinE9JdqY_uMh3vNO7WNOHdsIVf0 (visited on 04/06/2025).

Cano, Erion and Maurizio Morisio (May 2017). "Music Mood Dataset Creation Based on Last FM Tags". In: *Computer Science & Information Technology (CS & IT)*. Academy & Industry Research Collaboration Center (AIRCC), pp. 15–26. ISBN: 978-1-921987-66-3. DOI: 10.5121/csit.2017.70603. URL: http://airccj.org/CSCP/vol7/csit76803.pdf (visited on 04/26/2025).

Casey, M.A. et al. (Apr. 2008). "Content-Based Music Information Retrieval: Current Directions and Future Challenges". In: *Proceedings of the IEEE* 96.4, pp. 668–696. ISSN: 0018-9219, 1558-2256. DOI: 10.1109/JPROC.2008.916370. URL: http://ieeexplore.ieee.org/document/4472077/ (visited on 04/26/2025).

Garg, Anupam et al. (Feb. 2022). "Machine learning model for mapping of music mood and human emotion based on physiological signals". en. In: *Multimedia Tools and Applications* 81.4, pp. 5137–5177. ISSN: 1380-7501, 1573-7721. DOI: 10.1007/s11042-021-11650-0. URL: https://link.springer.com/10.1007/s11042-021-11650-0 (visited on 03/30/2025).

Grosse, Malte (2022). *8+ M. Spotify Tracks, Genre, Audio Features*. URL: https://www.kaggle.com/datasets/maltegrosse/8-m-spotify-tracks-genre-audio-features/data.

Hartmann, Jochen (2022). *Emotion English DistilRoBERTa-base*. URL: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Helmholz, Patrick, Michael Meyer, and Susanne Robra-Bissantz (June 2019). "Feel the moosic: Emotion-based music selection and recommendation". In: DOI: 10.18690/978-961-286-280-0.11.

Huron, David (2015). "Affect induction through musical sounds: an ethological perspective". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1664, p. 20140098. DOI: 10.1098/rstb.2014.0098. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2014.0098.

Juthi, Jannatul Humayra et al. (2020). "Music emotion recognition with the extraction of audio features using machine learning approaches". In: *Proceedings of ICETIT 2019*. Ed. by Pradeep Kumar Singh et al. Cham: Springer International Publishing, pp. 318–329. ISBN: 978-3-030-30577-2.

Kamenetsky, Stuart B., David S. Hill, and Sandra E. Trehub (Oct. 1997). "Effect of Tempo and Dynamics on the Perception of Emotion in Music". en. In: *Psychology of Music* 25.2, pp. 149–160. ISSN: 0305-7356, 1741-3087. DOI: 10.1177/0305735697252005. URL: https://journals.sagepub.com/doi/10.1177/0305735697252005 (visited on 04/26/2025).

Kim, Hannah et al. (Feb. 2024). *MEGAnno+: A Human-LLM Collaborative Annotation System*. DOI: 10.48550/arXiv.2402.18050. URL: http://arxiv.org/abs/2402.18050 (visited on 04/21/2025).

Leubner, Daniel and Thilo Hinterberger (2017). "Reviewing the effectiveness of music interventions in treating depression". In: *Frontiers in Psychology* 8. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01109. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2017.01109.

McCraty, Rollin et al. (Feb. 1998). "The effects of different types of music on mood, tension, and mental clarity". In: *Alternative therapies in health and medicine* 4, pp. 75–84.

Olha Zalutska et al. (2023). "Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network". In: *International Conference on Computational Linguistics and Intelligent Systems*. URL: https://api.semanticscholar.org/CorpusID:258688336.

Parthasarathy, Srinivas, Reza Lotfian, and Carlos Busso (Mar. 2017). "Ranking emotional attributes with deep neural networks". en. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, pp. 4995–4999. ISBN: 978-1-5090-4117-6. DOI: 10.1109/ICASSP.2017.7953107. URL: http://ieeexplore.ieee.org/document/7953107/ (visited on 04/26/2025).

Perlovsky, Leonid (2010). "Musical emotions: Functions, origins, evolution". In: *Physics of Life Reviews* 7.1, pp. 2–27. ISSN: 1571-0645. DOI: https://doi.org/10.1016/j.plrev.2009.11.001. URL: https://www.sciencedirect.com/science/article/pii/S1571064509000438.

Raufi, Bujar, Ciaran Finnegan, and Luca Longo (2024). "A Comparative Analysis of SHAP, LIME, ANCHORS, and DICE for Interpreting a Dense Neural Network in Credit Card Fraud Detection". en. In: *Explainable Artificial Intelligence*. Ed. by Luca Longo, Sebastian Lapuschkin, and Christin Seifert. Vol. 2156. Series Title: Communications in Computer and Information Science. Cham: Springer Nature Switzerland, pp. 365–383. ISBN: 978-3-031-63802-2 978-3-031-63803-9. DOI: 10.1007/978-3-031-63803-9_20. URL: https://link.springer.com/10.1007/978-3-031-63803-9_20 (visited on 04/26/2025).

Rosner, Aldona and Bozena Kostek (Apr. 2018). "Automatic music genre classification based on musical instrument track separation". en. In: *Journal of Intelligent Information Systems* 50.2, pp. 363–384. ISSN: 0925-9902, 1573-7675. DOI: 10.1007/s10844-017-0464-5. URL: http://link.springer.com/10.1007/s10844-017-0464-5 (visited on 03/30/2025).

Russell, James A and James H Steiger (1982). "The structure in persons' implicit taxonomy of emotions". In: *Journal of Research in Personality* 16.4, pp. 447–469. ISSN: 0092-6566. DOI: https://doi.org/10.1016/0092-6566(82)90005-8. URL: https://www.sciencedirect.com/science/article/pii/0092656682900058.

Spotify Web API (n.d.). URL: https://developer.spotify.com/documentation/web-api.

Thierry, Bertin-Mahieux et al. (2011). "The Million Song Dataset". In: *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. URL: https://labrosa.ee.columbia.edu/millionsong.

Uplabdhee, Avni et al. (2025). "A Comparative Analysis of Multi-label Emotion Recognition in Music". en. In: *ICT Systems and Sustainability*. Ed. by Milan Tuba, Shyam Akashe, and Amit Joshi. Vol. 1159. Singapore: Springer Nature Singapore, pp. 201–213. ISBN: 978-981-97-8525-4 978-981-97-8526-1. URL: https://link.springer.com/10.1007/978-981-97-8526-1_16 (visited on 05/14/2025).

Xia, Yu and Fumei Xu (2022). "Study on music emotion recognition based on the machine learning model clustering algorithm". In: *Mathematical Problems in Engineering* 2022.1, p. 9256586. DOI: https://doi.org/10.1155/2022/9256586. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/9256586.

Xu, Liang et al. (Nov. 2021). "Using machine learning analysis to interpret the relationship between music emotion and lyric features". In: *PeerJ Computer Science* 7, e785. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.785. URL: https://doi.org/10.7717/peerj-cs.785.

Yang, Mengxi (Apr. 2024). "Comparison and analysis of prediction accuracy between traditional machine learning algorithms and XGBoost algorithm in music emotion classification". In: *Applied and Computational Engineering* 57, pp. 98–103. DOI: 10.54254/2755-2721/57/20241316.

Yoo, Gilsang, Sungdae Hong, and Hyeocheol Kim (June 2024). "Emotion Recognition and Multi-class Classification in Music with MFCC and Machine Learning". In: *International Journal on Advanced Science, Engineering and Information Technology* 14.3, pp. 818–825. ISSN: 2460-6952, 2088-5334. DOI: 10.18517/ijaseit.14.3.18671. URL: https://ijaseit.insightsociety.org/index.php/ijaseit/article/view/18671 (visited on 03/30/2025).

A Transformer Score Threshold Justification

To determine an appropriate threshold for emotion label assignment from transformer-based tag scores, I conducted a descriptive analysis using the output of the DistilRoBERTa classifier applied to Last.fm tags. As shown in Figure 17, the distribution of maximum emotion scores per tag is highly right-skewed, with the majority of tags receiving relatively low confidence scores. This indicates that a high threshold (e.g., 0.8 or above) may excessively filter out potentially meaningful emotional information.

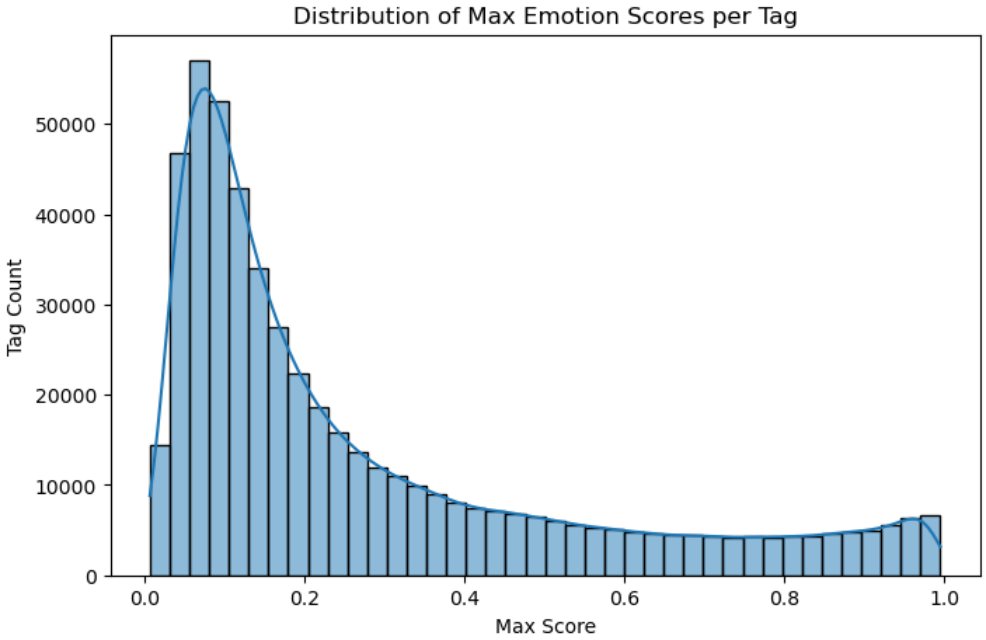

Figure 17: Distribution of maximum emotion scores per tag above threshold 0.5.

I further evaluated the number of tags retained under various thresholds ranging from 0.5 to 0.9 (Figure 18). At a threshold of 0.8, only around 41,000 tags remain, while at 0.7, over 58,000 tags are preserved—representing a substantial increase in usable emotional annotations without introducing a large volume of low-confidence data.

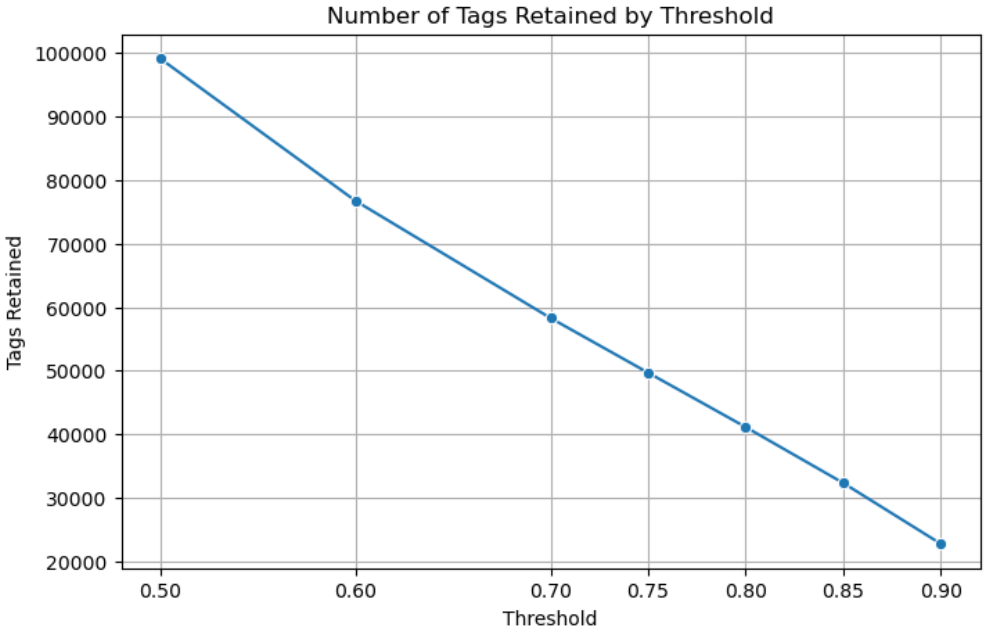

Figure 18: Line Plot Number of Tags Retained by Threshold

Importantly, while joy and sadness dominate across all thresholds, lowering the threshold from 0.8 to 0.7 leads to substantially improved representation of minority emotions such as anger, surprise, and fear. As shown in Appendix Figures 19, 20, and 21, the number of tags assigned to these minority categories increases significantly at lower thresholds without completely distorting the underlying class distribution. Specifically, threshold 0.7 provides a more balanced compromise, retaining the major emotion categories while enabling inclusion of less frequently predicted emotions.

By contrast, a threshold of 0.6 (Appendix Figure 19) introduces a noticeable shift in distribution that may reflect higher label noise, potentially due to the inclusion of tags with lower classifier confidence. This is consistent with the long-tail pattern observed in the score histogram (Figure 17).

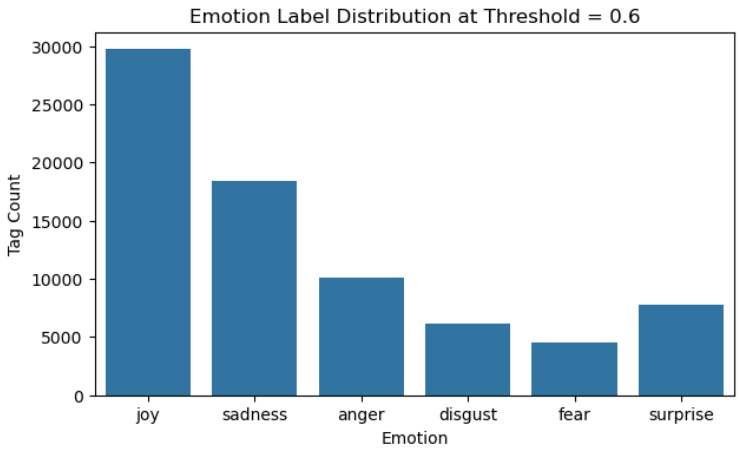

Figure 19: Emotion label distribution at threshold = 0.6.

To evaluate the impact of emotion classification threshold on downstream model performance, we compared the predictive results of three multi-label classifiers (MLP, Random Forest, and KNN) trained on datasets

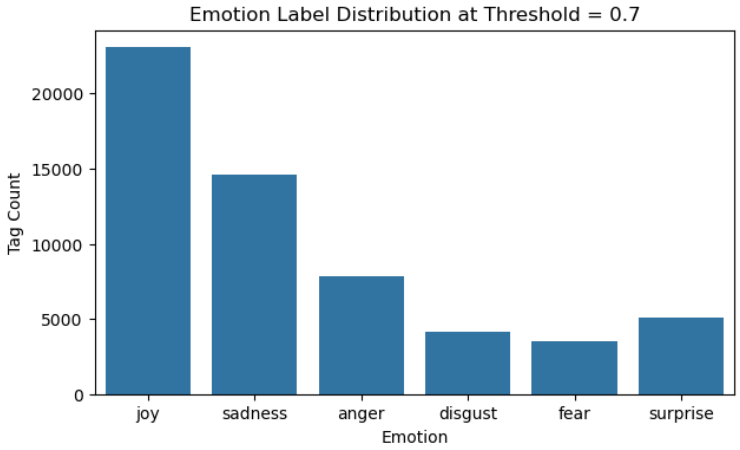

Figure 20: Emotion label distribution at threshold = 0.7.

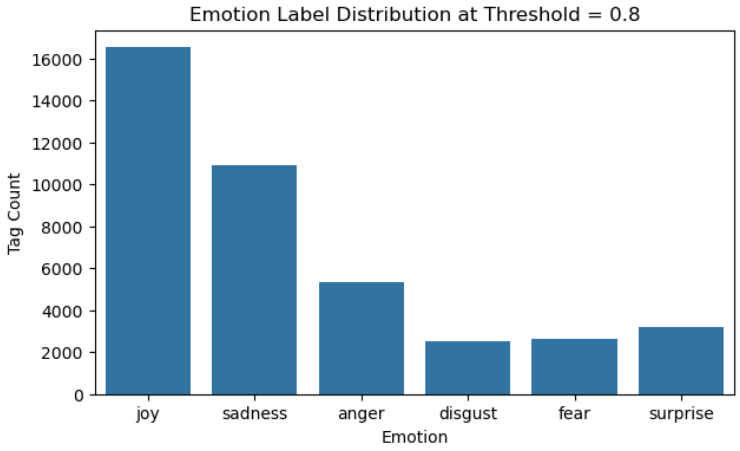

Figure 21: Emotion label distribution at threshold = 0.8.

labeled with thresholds of 0.7 and 0.8. As shown in Table 11, both thresholds yielded similar micro-average F1 scores across models, indicating consistent overall accuracy. However, threshold 0.7 demonstrated several advantages.

First, the 0.7-labeled dataset achieved slightly higher micro F1 scores on MLP and KNN (0.6725 and 0.6350) compared to their 0.8-labeled counterparts (0.6757 and 0.6332), and comparable performance on Random Forest (0.6757 vs. 0.6766). Second, the exact match accuracy—a strict metric measuring full-label agreement—was marginally higher for all three models under threshold 0.7. Notably, the Random Forest model achieved an exact match accuracy of 0.3690 with 0.7 threshold, compared to 0.3667 under 0.8.

More importantly, per-class F1 scores showed improved recall for minority emotions such as *disgust*, *fear*, and *surprise* under the 0.7 setting, suggesting that this threshold supports better emotion diversity in label distribution. For example, MLP's F1-score for *fear* increased from 0.25 to 0.28, and *disgust* from 0.40 to 0.38 despite the trade-off in precision.

Table 11: Comparison of Micro F1 and Exact Match Accuracy under Different Labeling Thresholds

Model	**Threshold**	**Micro F1**	**Exact Match**	**Hamming Loss**
MLP	0.7	0.6725	0.3580	0.2282
MLP	0.8	0.6757	0.3627	0.2268
RF	0.7	0.6757	0.3690	0.2259
RF	0.8	0.6766	0.3667	0.2261
KNN	0.7	0.6350	0.3075	0.2543
KNN	0.8	0.6332	0.3030	0.2565

In summary, while threshold 0.8 yields marginally cleaner labels with high confidence, threshold 0.7 offers a better trade-off between label richness, minority emotion coverage, and predictive performance. Therefore, for future iterations of this classification pipeline, I recommend adopting a threshold of 0.7.

B Raw Dataset Summary Statistics

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Acousticness	82,950	0.240	0.304	0.000	0.004	0.079	0.409	0.996
Danceability	82,950	0.521	0.175	0.000	0.397	0.525	0.648	0.980
Energy	82,950	0.653	0.249	0.000	0.476	0.696	0.868	1.000
Instrumentalness	82,950	0.155	0.290	0.000	0.000	0.001	0.114	0.993
Key	82,950	5.328	3.562	0.000	2.000	5.000	9.000	11.000
Liveness	82,950	0.196	0.160	0.009	0.095	0.130	0.258	1.000
Loudness (dB)	82,950	-8.664	4.338	-50.014	-10.943	-7.779	-5.537	3.744
Mode	82,950	0.653	0.476	0.000	0.000	1.000	1.000	1.000
Speechiness	82,950	0.074	0.079	0.000	0.034	0.045	0.076	0.960
Tempo (BPM)	82,950	122.38	29.48	0.000	99.85	120.11	140.19	219.93
Time Signature	82,950	3.905	0.393	0.000	4.000	4.000	4.000	5.000
Valence	82,950	0.479	0.258	0.000	0.265	0.470	0.689	0.989

Table 12: Descriptive statistics for 12 Spotify audio features across 82,950 records in the raw dataset.

C Boxplots of Acoustic Features by Emotion Category

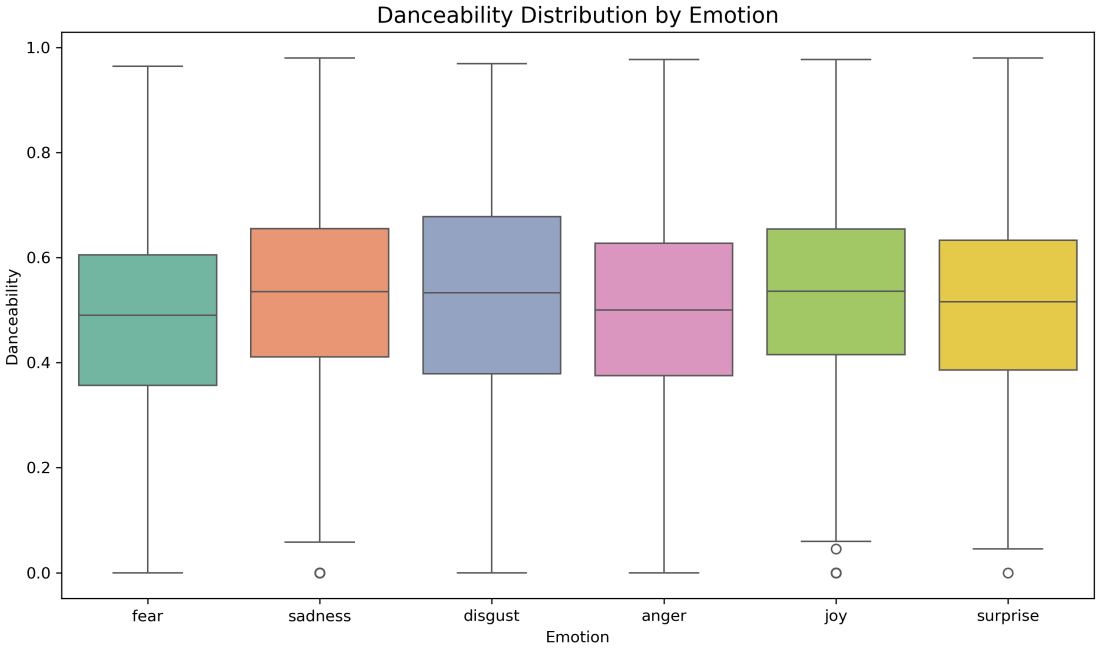

Figure 22: Distribution of `danceability` by emotion category.

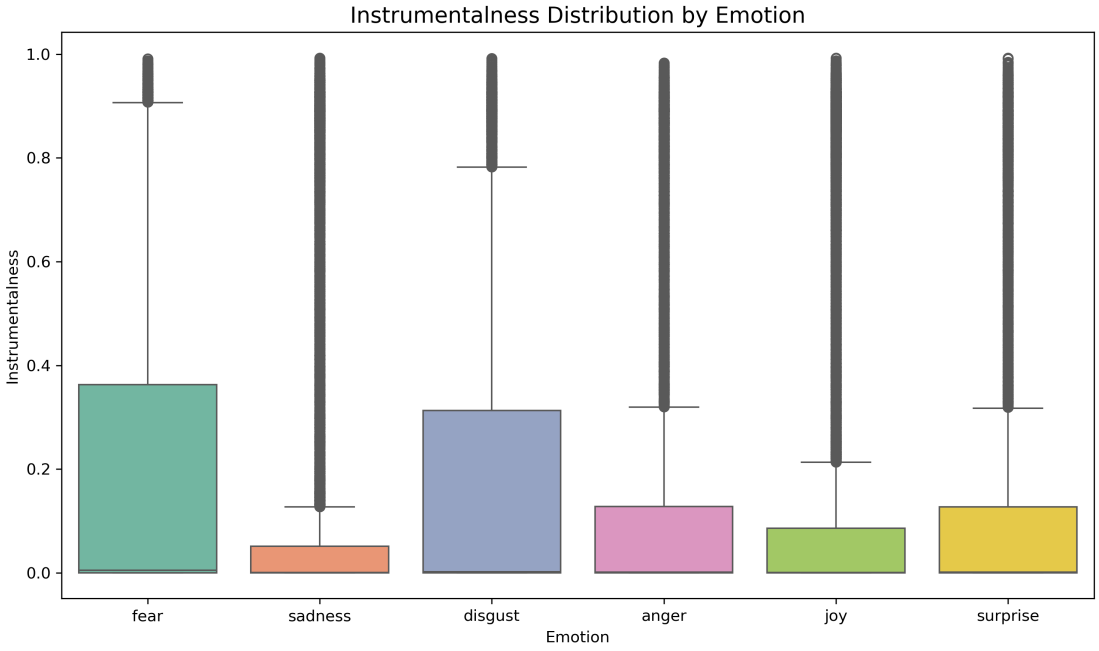

Figure 23: Distribution of `instrumentalness` by emotion category.

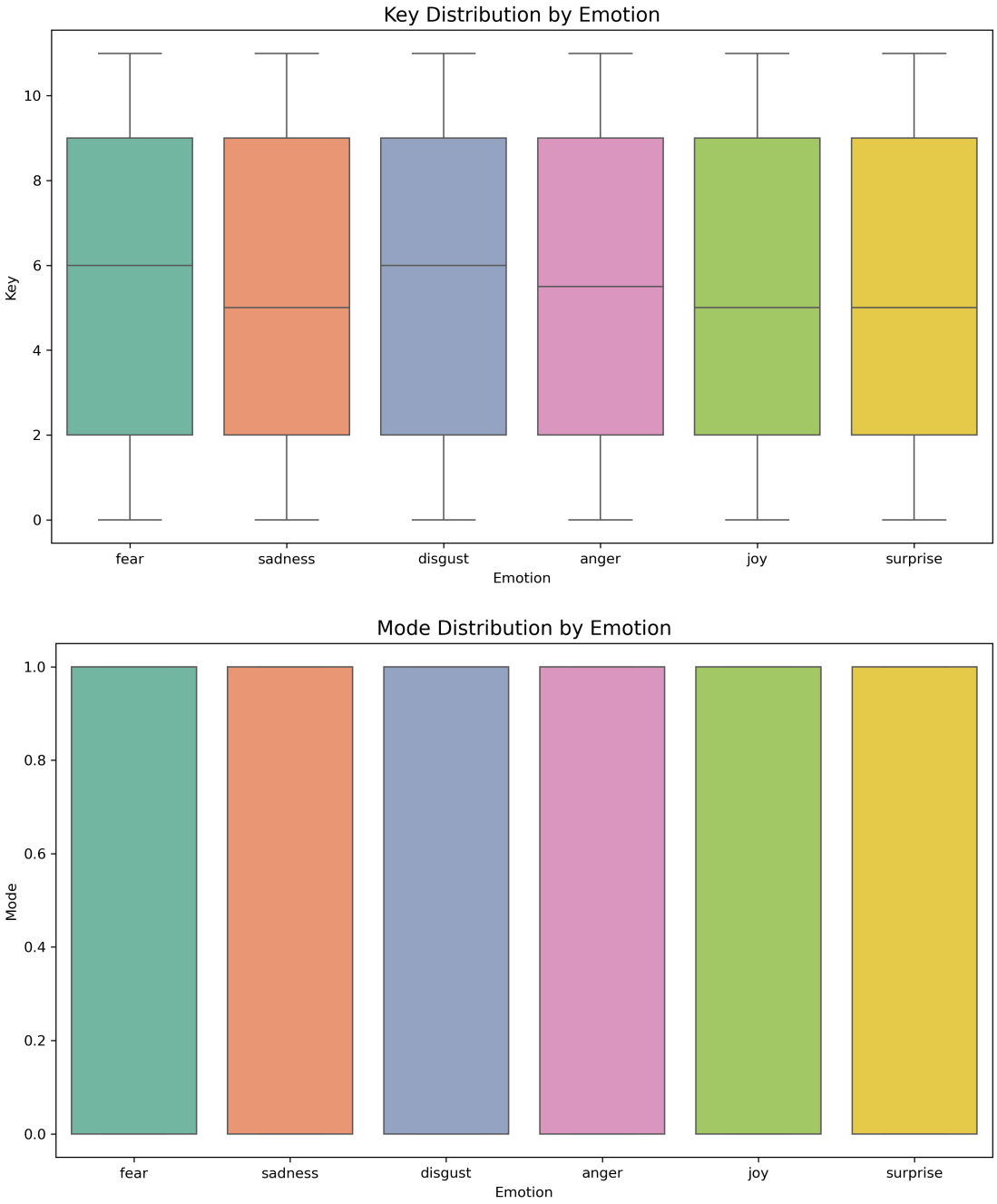

Figure 24: Distribution of `key` and `mode` by emotion category.

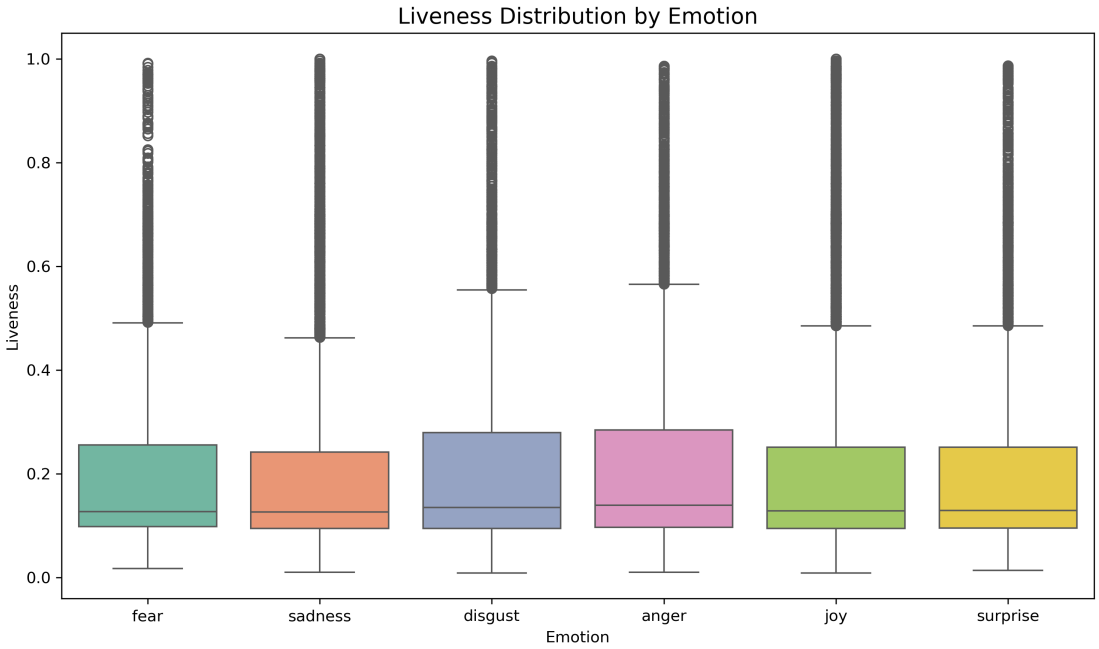

Figure 25: Distribution of `liveness` by emotion category.

Figure 26: Distribution of `loudness` by emotion category.

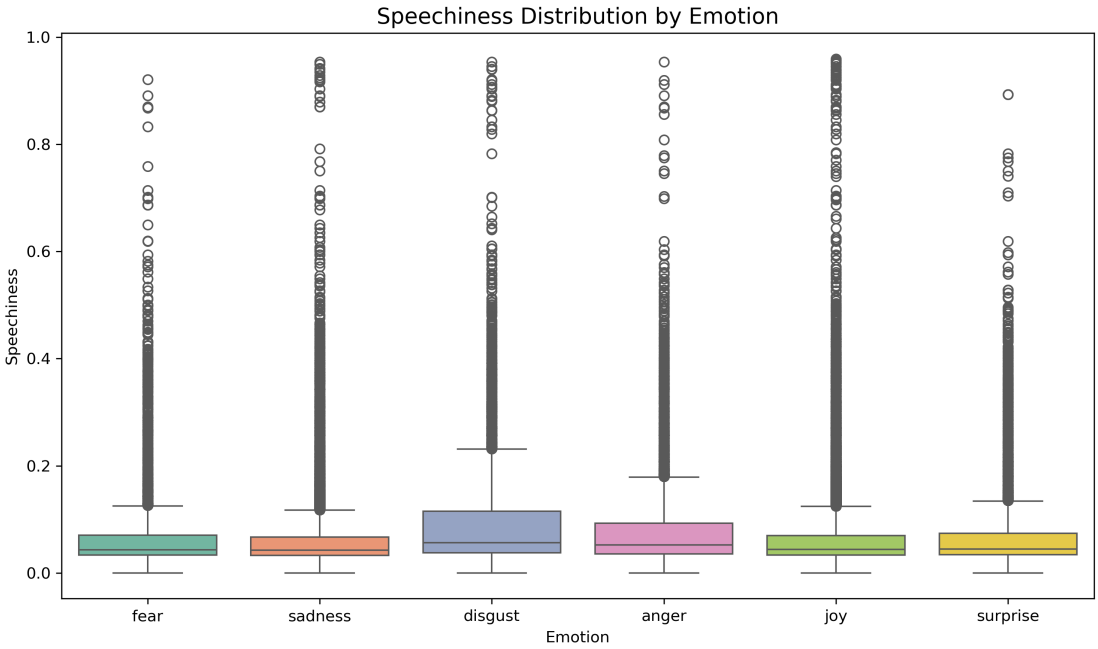

Figure 27: Distribution of `speechiness` by emotion category.

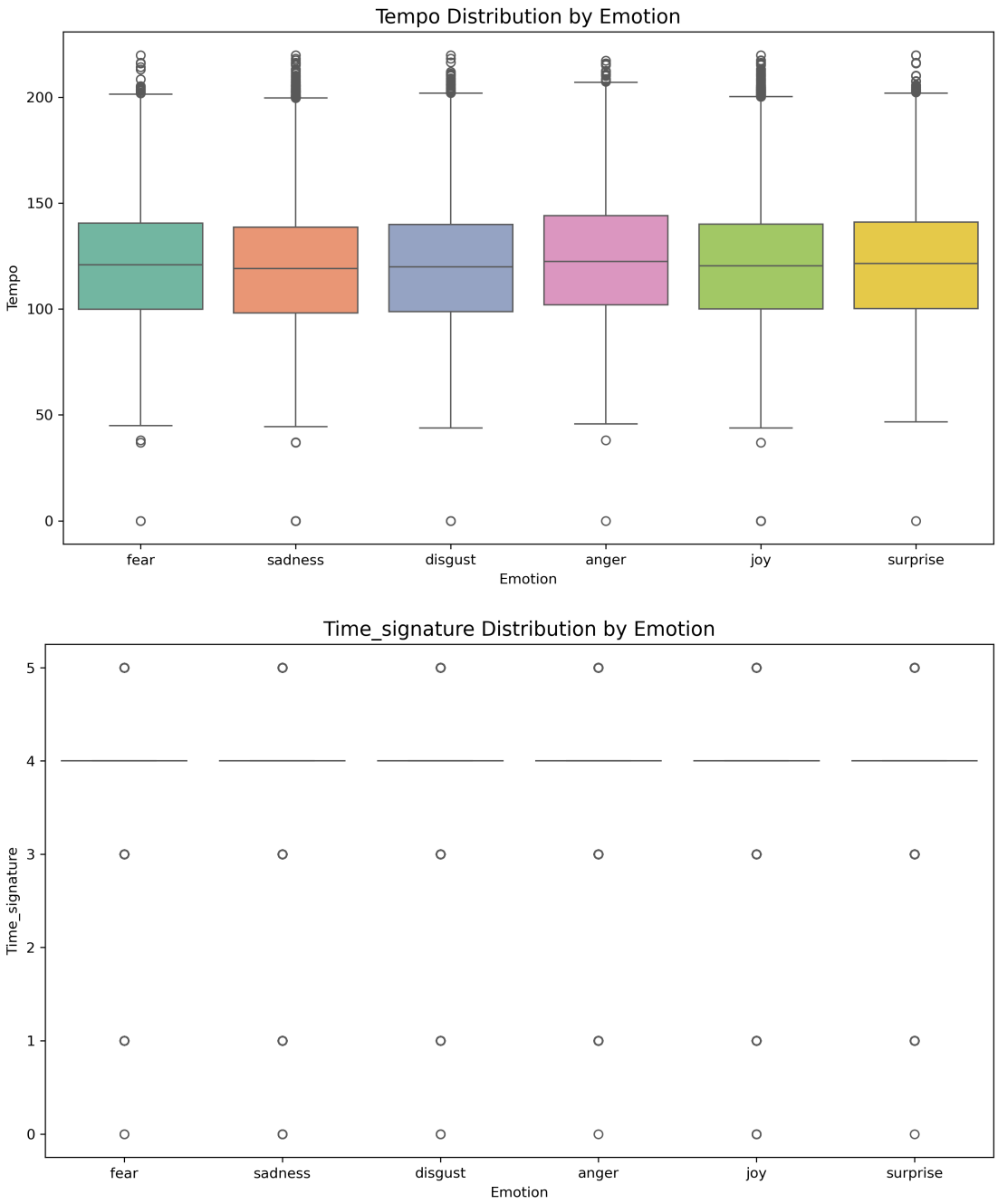

Figure 28: Distribution of `tempo` and `time signature` by emotion category.

D Per-Class Performance Metrics (KNN and MLP)

K-Nearest Neighbors (KNN). As shown in Figure 29, KNN achieved high recall and F1-score for *Joy* and *Sadness*, but struggled significantly on low-frequency emotions such as *Fear* and *Disgust*.

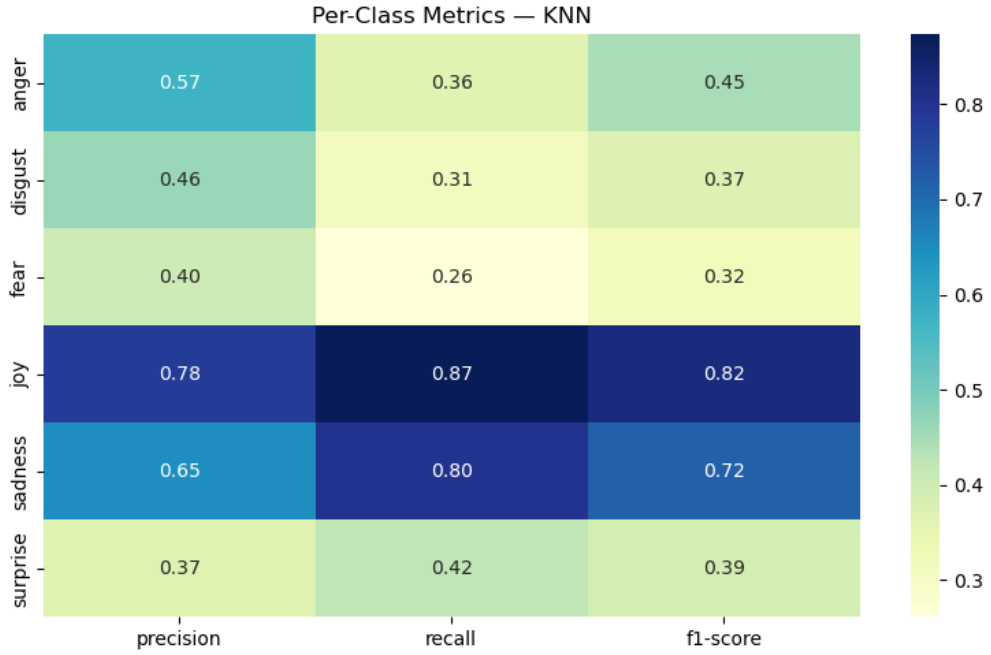

Figure 29: Per-class metrics for K-Nearest Neighbors classifier.

Multi-Layer Perceptron (MLP). Figure 30 shows that the MLP slightly improved over KNN for minority emotions, though recall remained low for *Fear* and *Surprise*.

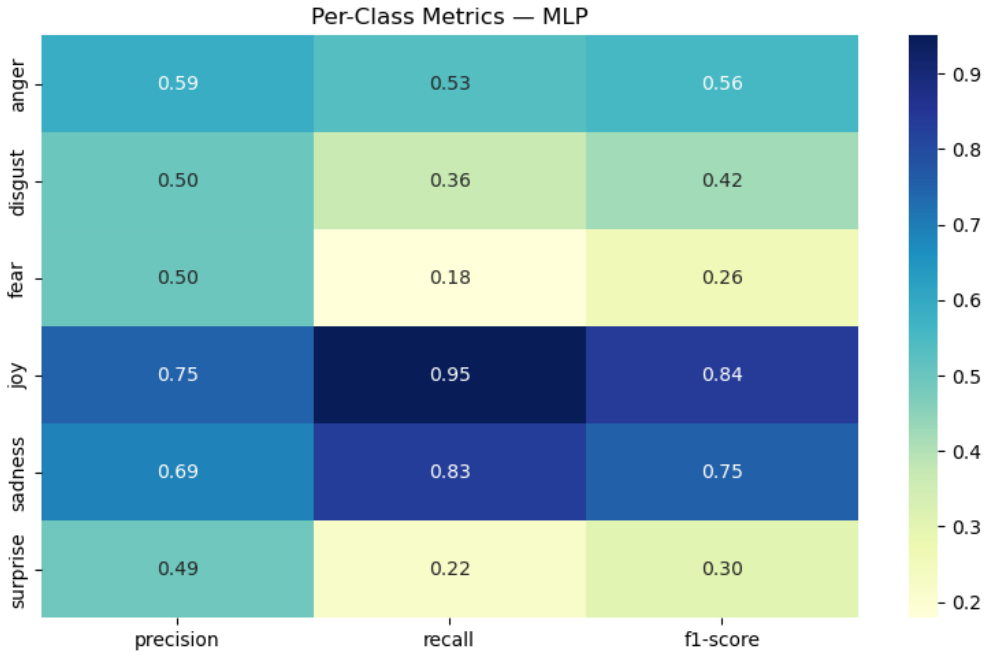

Figure 30: Per-class metrics for Multi-Layer Perceptron classifier.

E Per-Label Feature Importance by Emotion Category

To complement the global feature analysis, I examined per-label feature importance across all six emotion categories using Random Forest, KNN, and SHAP-based MLP classifiers. Visualizations and discussions for each emotion are presented below.

Disgust. For *disgust*, all models highlight `speechiness`, `acousticness`, and `valence` as important predictors. Random Forest places the greatest weight on `speechiness` (0.117), suggesting that vocal texture—such as harsh or spoken delivery—may convey disgust. KNN agrees on this ranking but assigns relatively high importance to `mode` and `key`, indicating a possible reliance on harmonic context that may reflect genre patterns more than affective signals. MLP SHAP values, although smaller in magnitude, still elevate `speechiness` and `danceability`, reinforcing the notion that disgust is acoustically expressed through rhythmic and vocal cues rather than tonal structure. Compared to emotions like *joy* or *sadness*, feature contributions for *disgust* appear more evenly distributed, which may reflect the emotion's more context-dependent or ambiguous nature in music.

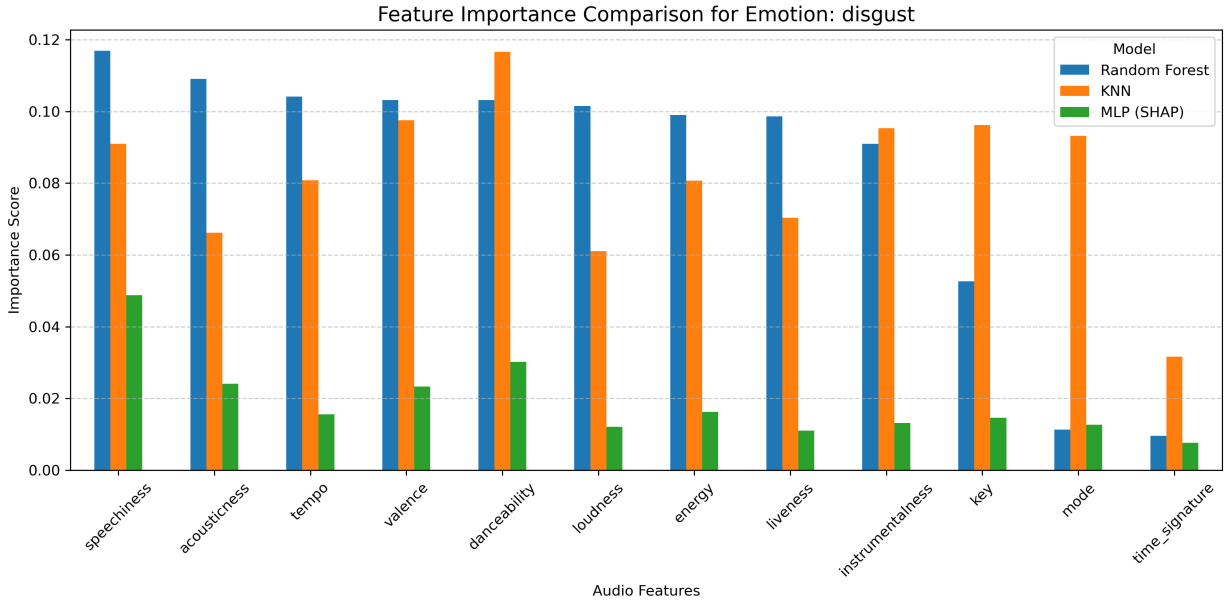

Figure 31: Feature importance (MLP SHAP) for emotion: *disgust*

Fear. For *fear*, models show consistent reliance on `valence`, `loudness`, and `energy`, aligning with psychological literature that defines fear as a low-valence, high-arousal emotion. Random Forest assigns highest importance to `valence` (0.111) and `loudness` (0.108), suggesting that low-intensity, dark, or ambiguous tracks are indicative of fear responses. KNN emphasizes similar features but also places considerable weight on `mode` and `instrumentalness`, potentially capturing modal dissonance or texture typical in suspenseful music. In contrast, MLP assigns relatively flat SHAP scores across features, with slightly higher values for `valence`, `energy`, and `liveness`. This indicates that MLP struggles to isolate dominant acoustic signals for *fear*, which may help explain its poor classification performance on this category. The absence of strong, consistent cues likely reflects the contextual and genre-dependent nature of fear in music.

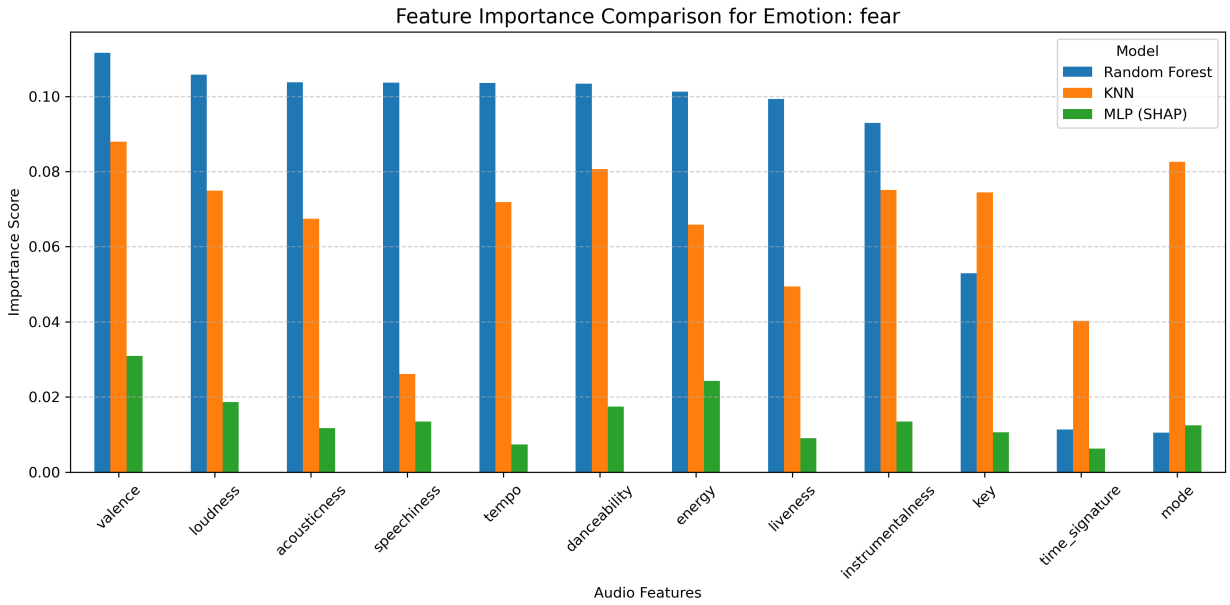

Figure 32: Feature importance (MLP SHAP) for emotion: *fear*

Joy. Across all models, `speechiness` emerges as the most important feature for predicting *joy*. This suggests that joyful songs often involve pronounced lyrical or spoken-word elements—consistent with upbeat, expressive vocal delivery. Random Forest assigns the highest weight to `speechiness` , followed by `acousticness`, `tempo`, and `danceability`, indicating that timbral brightness and rhythmic energy are jointly predictive of positive affect. KNN highlights similar features but with reduced magnitude, and additionally emphasizes `mode` and `key`, aligning with traditional tonal-emotion mappings that associate major keys with happiness. In contrast, MLP SHAP values show `speechiness` as most predictive, with moderate contributions from `energy` and `valence`. These differences reflect MLP's tendency to spread importance across features and learn subtle nonlinear combinations. The convergence across models underscores the reliability of rhythmic and vocal cues in signaling musical joy.

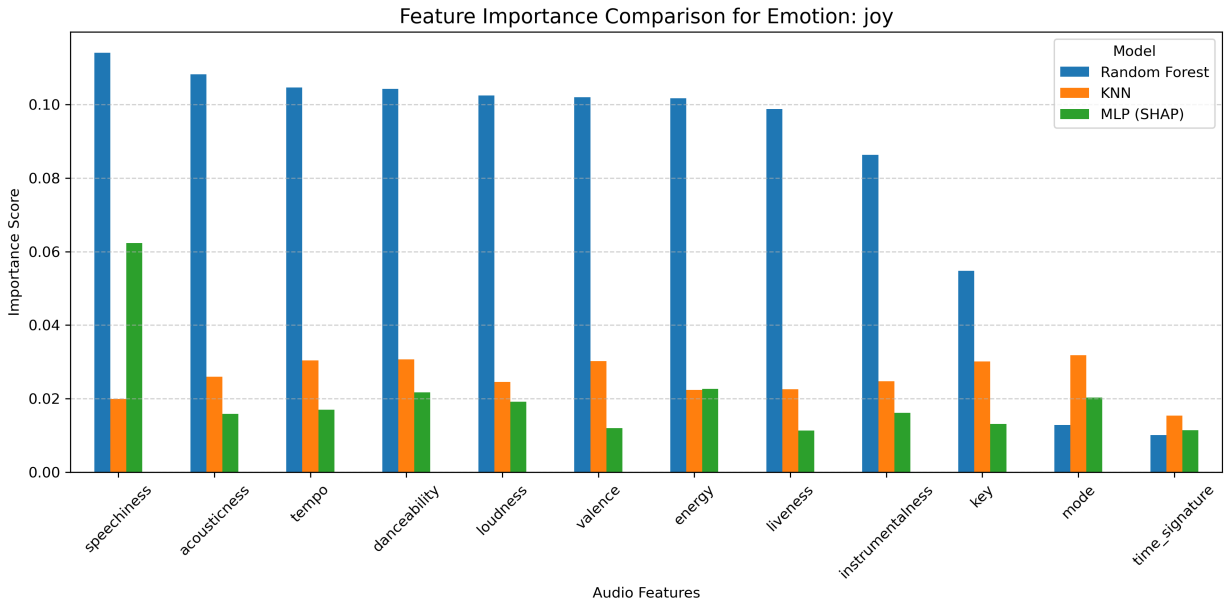

Figure 33: Feature importance (MLP SHAP) for emotion: *joy*

Sadness. All three models consistently rank `energy` as the most influential feature for classifying *sadness*, supporting the view that lower energy levels are a hallmark of melancholic or reflective tracks. Random Forest assigns high importance to `speechiness`, `acousticness`, and `loudness`, indicating that soft, lyric-rich, and minimally amplified compositions are common among sad songs. SHAP values from the MLP model highlight a similar trend, especially the heightened role of `energy` (0.078), reinforcing that reduced dynamic intensity is predictive of sadness. KNN also places moderate emphasis on `tempo`, `valence`, and `danceability`, although its importance scores are more diffuse. Interestingly, all models treat `mode` and `time_signature` as minimally informative, suggesting that formal tonal structure plays a limited role in sadness detection within this dataset. These patterns collectively affirm the acoustic softness and emotional introspection typically associated with sad musical affect.

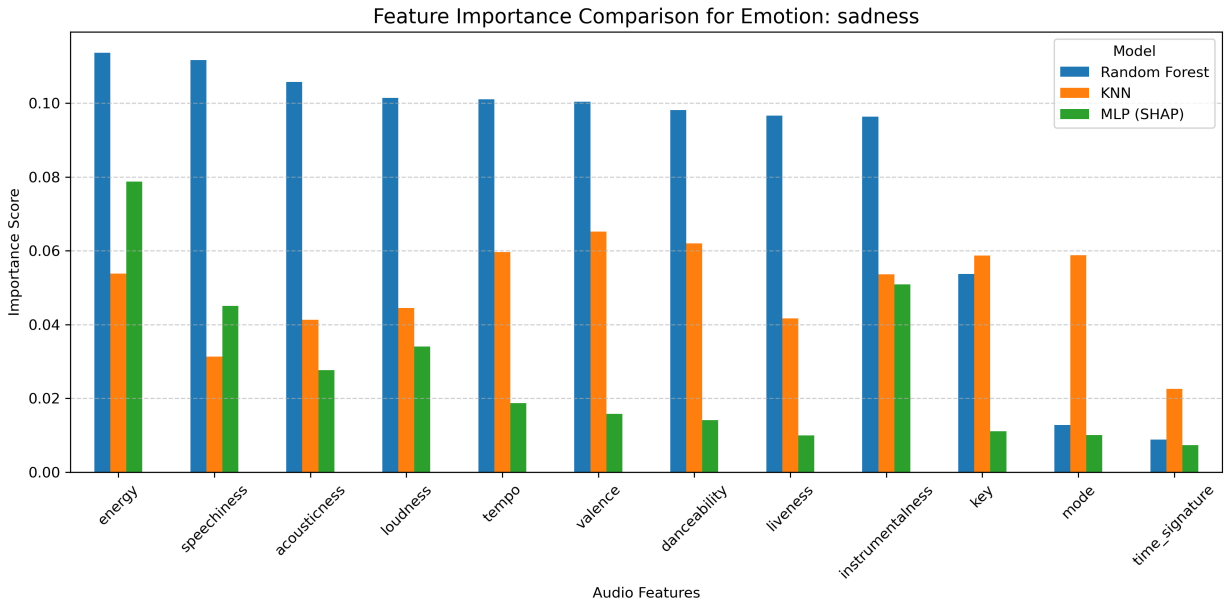

Figure 34: Feature importance (MLP SHAP) for emotion: *sadness*

Surprise. For *surprise*, all three models converge on a core set of predictive features—namely, `loudness`, `tempo`, `danceability`, and `energy`. Random Forest assigns nearly equal importance to this cluster of high-arousal descriptors, suggesting that *surprise* is encoded through dynamic intensity and rhythmic drive. KNN reinforces this trend with elevated weights on `tempo` and `valence`, while also placing substantial emphasis on `mode` and `key`—potentially overfitting to harmonic novelty. The MLP model, via SHAP analysis, identifies moderate importance for `loudness` and `energy` but shows flatter distributions overall, implying more diffuse activation in non-linear interactions. Notably, all models deem `speechiness` and `instrumentalness` less critical, perhaps reflecting that lyrics and texture are secondary to rhythmic and intensity cues when detecting musical surprise. These patterns align with prior work linking *surprise* to heightened arousal and irregular structural transitions in music.

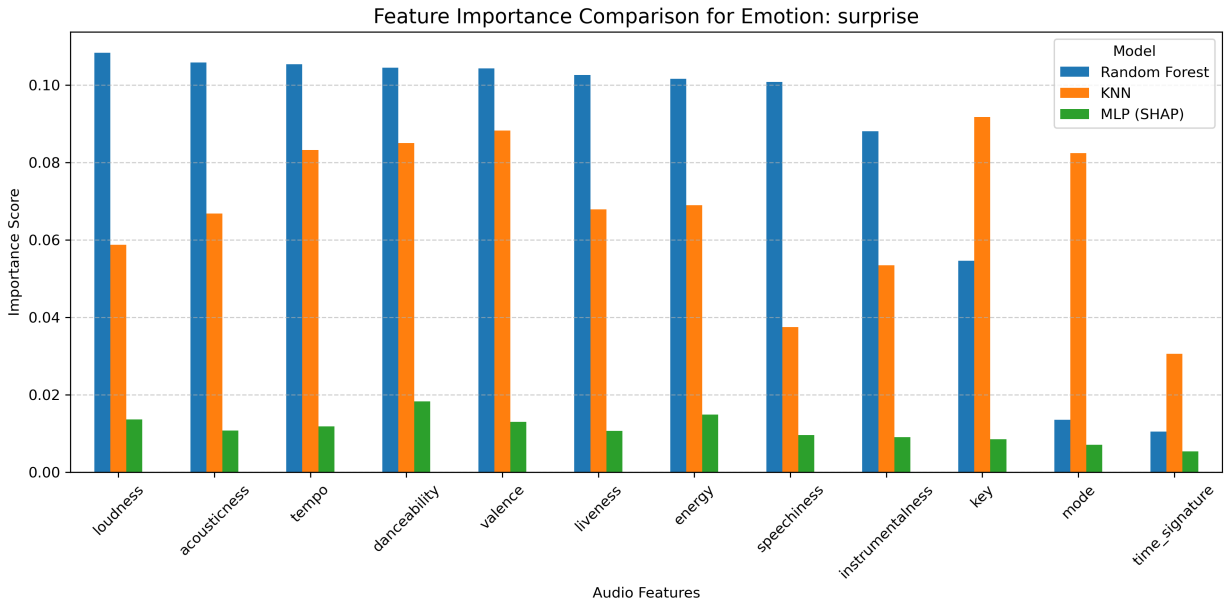

Figure 35: Feature importance (MLP SHAP) for emotion: *surprise*

F Supplementary Misclassification Heatmaps

Figures 36 and 37 visualize misclassification patterns for the KNN and MLP models. Error distributions resemble those seen in the Random Forest classifier, with consistent confusion between semantically close emotions.

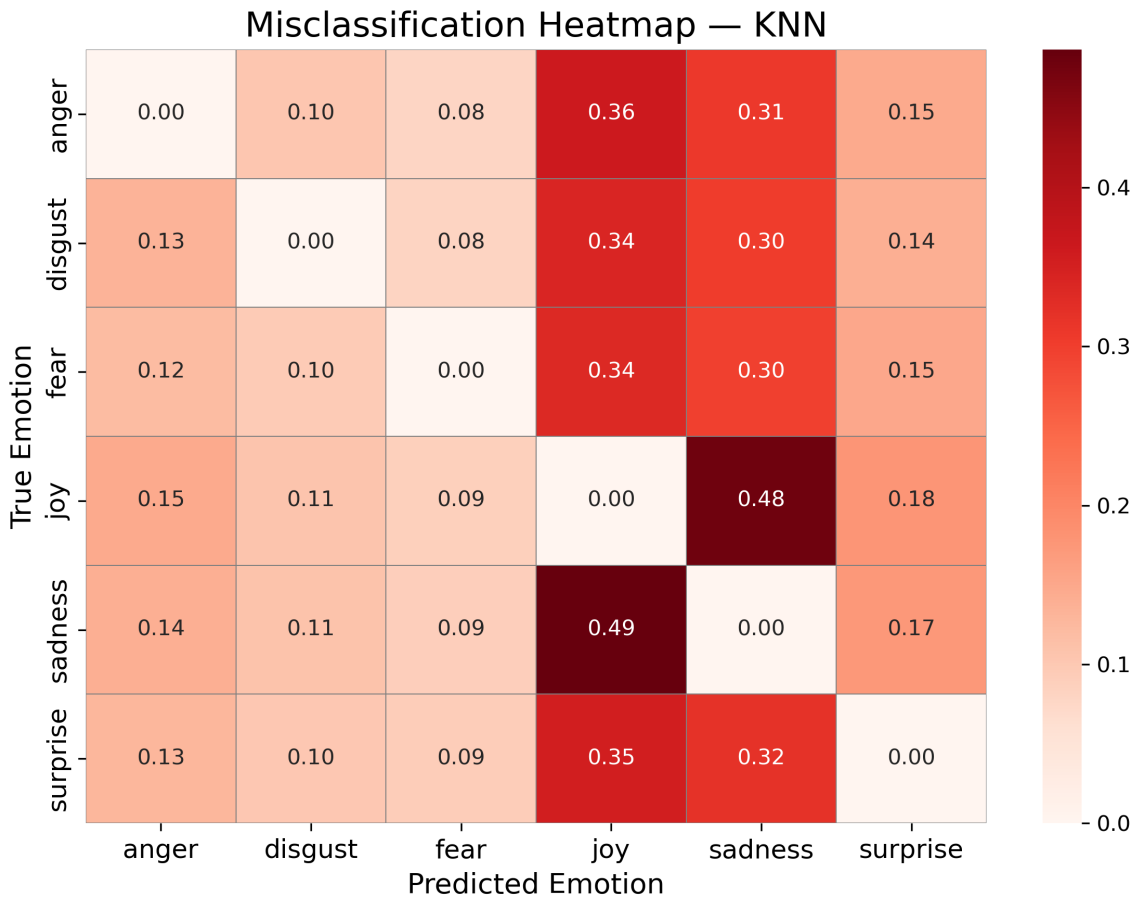

Figure 36: Misclassification heatmap for the KNN model. *Sadness* is frequently confused with *Joy* and *Fear*.

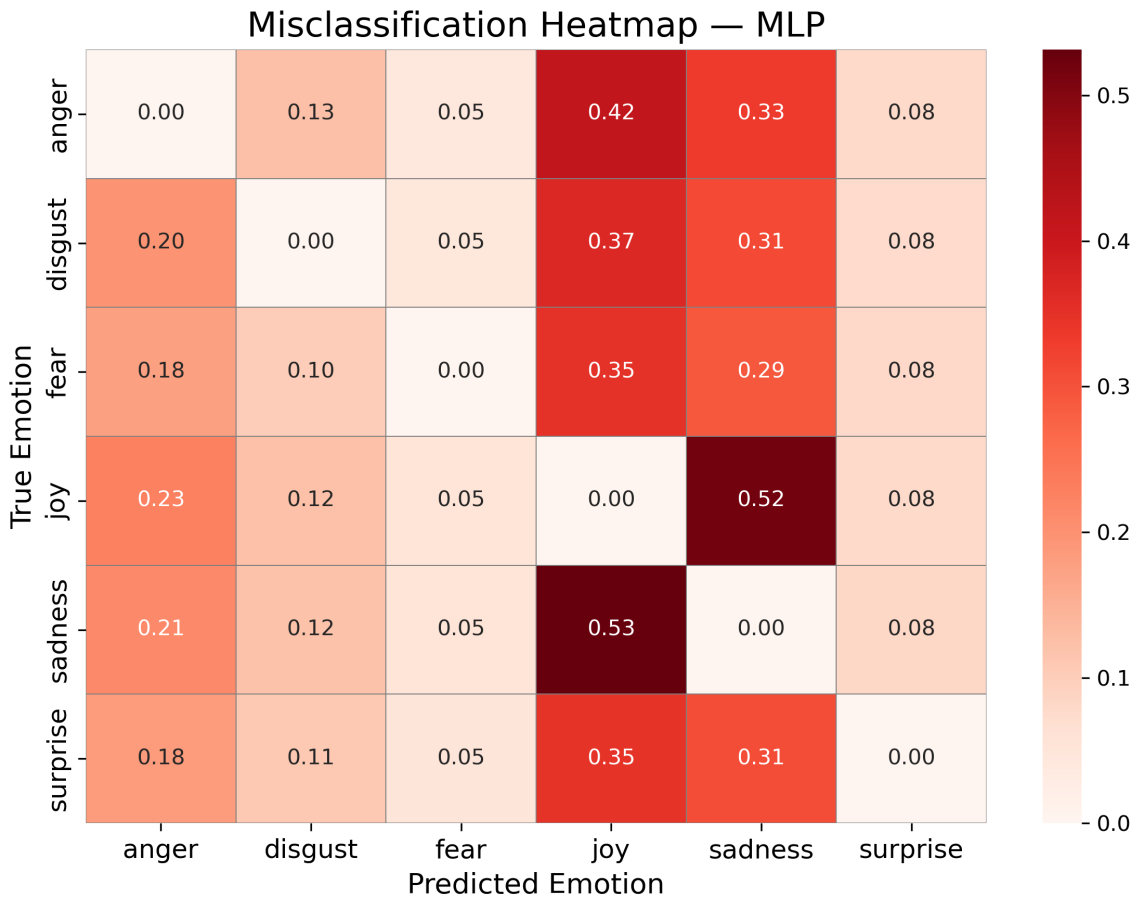

Figure 37: Misclassification heatmap for the MLP model. Similar to KNN, common misclassifications occur between *Sadness*, *Joy*, and *Fear*.

G SHAP Waterfall Visualizations for Misclassified Emotions

Anger: The model predicted a confidence score of 0.583 but still fell short of the top-k cutoff. High `energy` and low `valence` were positively associated with anger—consistent with aggressive delivery and negative emotional tone. However, low `tempo` and abnormal `time_signature` contributed negatively, suggesting rhythmic ambiguity may mask anger cues.

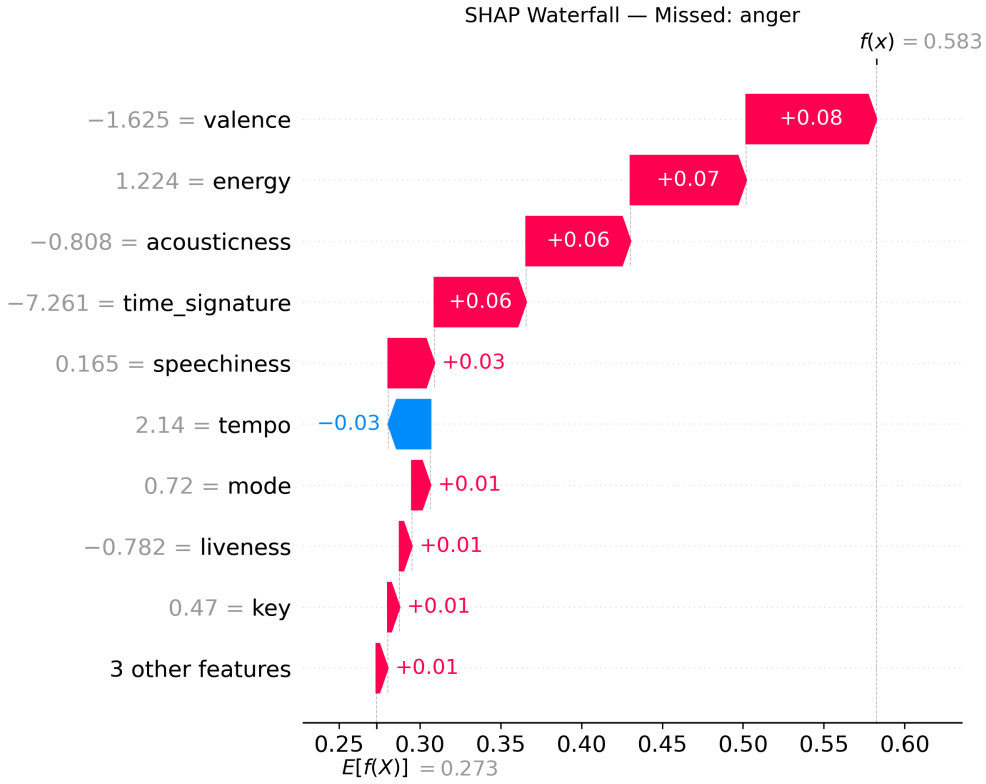

Figure 38: SHAP explanation for misclassified *Anger*.

Disgust: Moderate `speechiness`, `energy`, and low `valence` were aligned with disgust, but weak contributions from `loudness` and `danceability`, and negative influence from `instrumentalness`, indicate that disgust lacks consistent acoustic signatures.

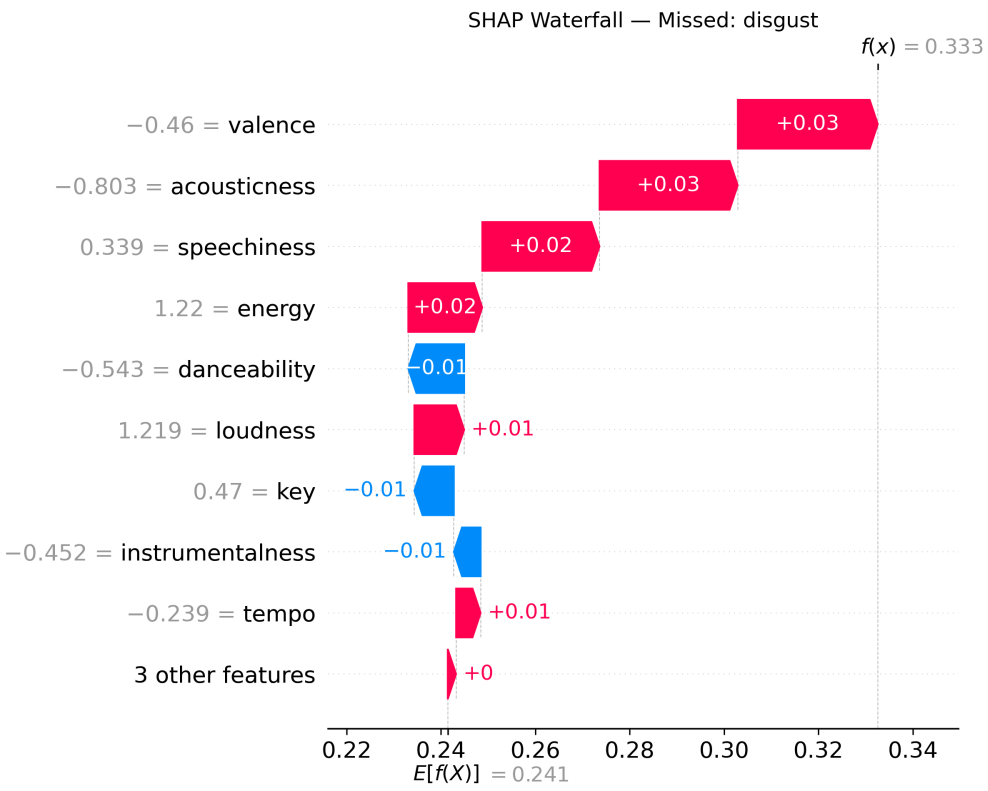

Figure 39: SHAP explanation for misclassified *Disgust*.

Sadness: Surprisingly, low `tempo` and `energy` pushed the score downward. Although expected to correlate with sadness, the model may rely more heavily on `acousticness` or `instrumentalness`, which were absent in this case.

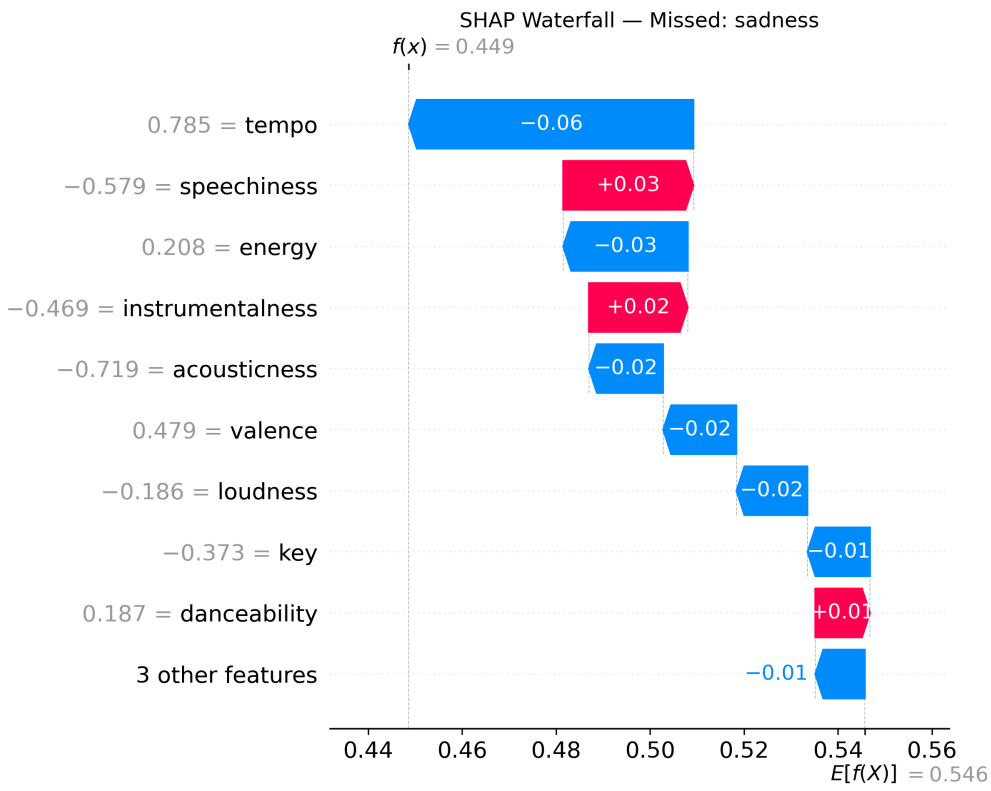

Figure 40: SHAP explanation for misclassified *Sadness*.

Surprise: Despite moderate `loudness`, `mode`, and `danceability`, high `acousticness` suppressed the prediction. This suggests the model interpreted the track as introspective rather than surprising, highlighting ambiguity in acoustic cues.

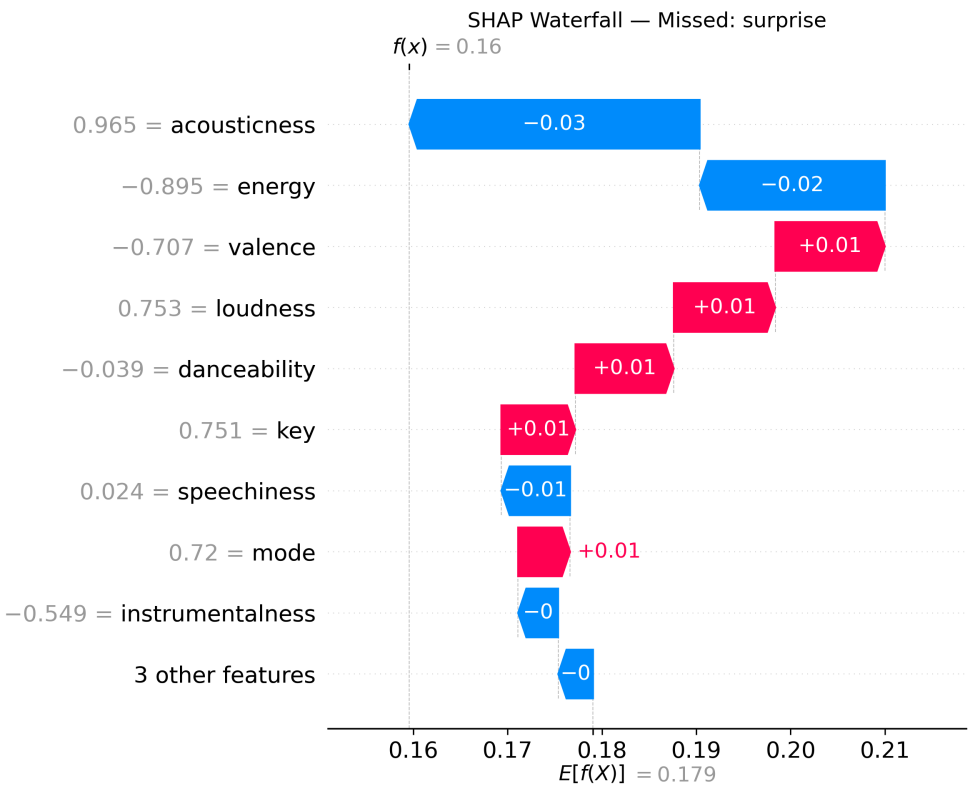

Figure 41: SHAP explanation for misclassified *Surprise*.