Inferring Musical Mood from Symbolic MIDI Basslines

Ahmet Batuhan Uluçay

Bocconi University

ahmet.ulucay@studbocconi.it

September 5, 2025

Abstract

Research on musical mood prediction has predominantly relied on audio and lyrical features, leveraging timbral and expressive cues unavailable in symbolic formats. Whether high-level affective attributes such as mood can be inferred from symbolic structure alone remains underexplored in Music Information Retrieval (MIR) [1]. In this work, we investigate mood classification using exclusively MIDI-based representations, without access to audio or lyrics, focusing on basslines as a structurally grounded yet constrained test case. Using 9,205 deduplicated songs from the Lakh MIDI Dataset [3], we construct an end-to-end pipeline comprising bassline extraction, key estimation, heuristic mood labeling based on tempo, mode, and note density, confidence-based label refinement, and token-based sequence modeling. A lightweight self-attention classifier is trained to perform four-way mood classification under artist-level data splits, achieving 0.36 test accuracy and 0.27 macro-F1. Although performance remains modest and is influenced by class imbalance and heuristic label noise, results consistently exceed random baselines and reveal systematic structural cues, particularly for high-energy moods. These findings suggest that symbolic bassline structure encodes partial affective information, while highlighting the limitations of purely structural representations for capturing nuanced emotional states without timbral or melodic context.

1 Introduction

Music Information Retrieval (MIR) has developed robust methods for extracting structural and perceptual attributes from music, including tempo, key, and genre [1, 2]. In contrast, modeling musical mood remains substantially more challenging. Affective perception in music emerges from a complex interaction of harmony, rhythm, instrumentation, timbre, and performance, and is inherently subjective and context-dependent. Consequently, most existing approaches to mood classification rely on audio-domain features or multimodal systems combining audio with lyrical or metadata information [5, 6].

Symbolic representations such as MIDI encode structural musical information—note pitches, durations, timing, and instrumentation—while omitting timbral and performance nuances. Symbolic formats have been widely studied for generative modeling, sequence learning, and structural analysis [7, 8]. However, their capacity to support affective modeling remains underexplored. In particular, it is unclear to what extent high-level emotional attributes can be inferred from structural information alone, independent of timbre and expressive performance cues.

This work investigates whether mood can be inferred from symbolic MIDI representations without access to audio or lyrics. Rather than modeling full arrangements, we focus specifically on basslines as a controlled and structurally grounded test case. Basslines encode rhythmic drive and harmonic foundation, making them musically central yet representationally constrained. By isolating this component, we aim to assess the extent to which structural musical cues alone contribute to affective classification.

The contributions of this work are threefold. First, we provide a systematic evaluation of mood classification using exclusively symbolic bassline representations under artist-level splits. Second, we introduce a weakly supervised labeling and refinement pipeline that enables large-scale symbolic mood modeling without external annotations. Third, we analyze the limits of bassline-based affect modeling, demonstrating measurable but constrained performance and highlighting the role of structural versus timbral information in mood perception.

2 Dataset

We base our study on the Lakh MIDI Dataset (LMD), one of the largest publicly available collections of symbolic music in MIDI format [3]. LMD has been widely used in symbolic music research for sequence modeling, structural analysis, and generative tasks. We use the Clean MIDI Subset, which provides higher-quality MIDI files with associated artist and title metadata, enabling reliable artist-level partitioning.

The Clean subset initially contains 17,232 MIDI files organized by artist. Because LMD includes multiple versions and near-duplicate arrangements of the same songs, we apply a deduplication procedure based on filename and metadata alignment, resulting in 9,205 unique songs. This step reduces redundancy and prevents inflated performance due to multiple arrangements of the same composition appearing across splits.

Following preprocessing and bassline extraction (Section 3), a small number of files are discarded due to missing tempo information, empty bass tracks, or structurally trivial basslines. The remaining dataset retains broad stylistic diversity across genres and decades.

To ensure realistic generalization, we perform artist-level splits into training, validation, and test sets, preventing songs by the same artist from appearing across partitions. This avoids stylistic leakage and provides a stricter evaluation of mood generalization in the symbolic domain.

3 Data Extraction and Preprocessing

3.1 Bassline Extraction

Our first preprocessing step is bassline extraction. Earlier iterations of this project explored chord-based representations, but large-scale chord inference across thousands of MIDI files proved computationally expensive and susceptible to harmonic estimation errors [7]. Basslines provide a more tractable alternative: they encode rhythmic drive and harmonic grounding while remaining comparatively simple to isolate.

Bass tracks are identified using instrument metadata and program numbers corresponding to bass instruments in the General MIDI specification. In cases where multiple candidate tracks are present, the lowest-pitched track with consistent note activity is selected. MIDI files without identifiable bass content are discarded.

Using the `PrettyMIDI` and `mido` libraries, each file is converted into a structured JSON representation containing:

- **tempo**: the global estimated tempo in beats per minute,
- **grid_subdiv**: the number of quantization steps per quarter note,
- **bassline**: a sequence of quantized bass note events defined by start_tick, duration_tick, pitch, and velocity.

All note events are quantized to a fixed temporal grid to reduce expressive timing variation and ensure comparability across songs. To normalize sequence alignment, bassline timing is shifted such that the first note begins at tick 0. This yields a consistent symbolic representation independent of original MIDI offsets.

3.2 Key Estimation

In addition to rhythmic and pitch information, we extract tonal context using the `music21` library. For each song, a global key is estimated using a correlation-based method over pitch-class distributions. Key is represented as a tuple consisting of tonic and mode (major or minor). While global key estimation may not capture local modulations, it provides a coarse tonal descriptor suitable for large-scale heuristic labeling. The estimated key features are appended to the JSON representation, completing the symbolic preprocessing pipeline.

4 Tokenization

Following preprocessing, each bassline is converted into a discrete token sequence suitable for sequence-based modeling. Rather than representing music as a fixed time-step piano roll, we adopt an event-based representation in which symbolic events are serialized into a linear sequence, analogous to token sequences in natural language processing [8]. This design preserves temporal ordering while avoiding sparsity introduced by fixed-grid encodings.

Each song is represented as an ordered sequence of tokens drawn from a predefined vocabulary comprising:

- **Global tokens**: key (tonic and mode) and coarse tempo category,
- **Pitch tokens**: MIDI pitch values observed in the bassline,
- **Duration tokens**: quantized note lengths in grid units,
- **Velocity tokens**: discretized dynamic levels (binned into fixed ranges),
- **Time-shift tokens**: relative temporal offsets between consecutive note events.

Global tokens are inserted at the beginning of each sequence to provide contextual conditioning. Subsequent tokens are serialized in event order as alternating time-shift and note-attribute tokens, preserving the chronological structure of the bassline. Velocity values are discretized into coarse bins to reduce vocabulary size and mitigate expressive noise.

Sequences are truncated or padded to a fixed maximum length to enable batch training. Padding tokens are masked during training to prevent spurious learning signals. The resulting vocabulary consists of discrete tokens spanning pitch, duration, velocity, time-shift, and contextual descriptors, yielding a compact yet expressive symbolic representation.

In addition to token sequences, we compute summary features for each song—tempo, mode, tonic, and estimated note density (notes per bar)—which are used exclusively for heuristic mood labeling and are not directly provided to the final classifier.

5 Mood Labeling and Refinement

5.1 Heuristic Label Generation

Because large-scale symbolic datasets lack reliable affect annotations, we adopt a weakly supervised labeling strategy based on interpretable musical heuristics. Rather than relying on external mood tags, we generate labels using combinations of structural features extracted during preprocessing: global tempo, tonal mode (major/minor), and bassline note density (notes per bar). These features are commonly associated with perceived musical energy and affective valence in prior music psychology and MIR literature.

In the initial labeling pass, songs are assigned to six mood categories according to rule-based thresholds over these features: *joyful*, *angry*, *melancholic*, *calm*, *easygoing*, and *wistful*. For example, fast tempo, major mode, and high note density correspond to joyful, while slow tempo, minor mode, and low density correspond to melancholic. Songs that do not satisfy any rule with sufficient confidence are assigned a fallback **neutral** label. This neutral category accounts for approximately 32% of the dataset, reflecting the limited expressivity of rigid structural heuristics and the continuous nature of musical affect.

5.2 Neutral Class Refinement

To reduce the size of the neutral category while preserving label reliability, we perform a secondary refinement step. A lightweight classifier is trained exclusively on the subset of songs assigned one of the six heuristic mood labels. Neutral songs are then treated as unlabeled data and passed through this model to obtain predicted mood labels with associated confidence scores.

Only predictions exceeding a predefined confidence threshold are incorporated into the dataset; lower-confidence predictions remain neutral. Importantly, this refinement model is used solely for label expansion and is not reused in final evaluation. A subset of low- and mid-confidence examples is manually inspected, and files containing trivial or structurally insignificant basslines are removed. This process reduces label noise while avoiding direct contamination of the downstream classifier.

5.3 Class Merging

Following refinement, class imbalance and semantic overlap remain pronounced among certain mood categories. In particular, *calm* and *easygoing*, as well as *melancholic* and *wistful*, exhibit strong similarity in structural descriptors and are frequently confused in preliminary experiments. To improve stability and reduce ambiguity, these pairs are merged into two broader categories: **calm_easygoing** and **melancholic_wistful**. The final task therefore consists of four mood classes, providing a more balanced and semantically coherent target space for classification.

6 Model

To evaluate whether symbolic bassline representations alone support mood prediction, we employ a lightweight token-based classification architecture designed to model global structural patterns rather than local event dependencies.

Each song is represented as a token sequence (Section 3). Tokens are mapped to learnable embeddings of dimension $d = 128$. The embedded sequence is processed by a single self-attention block comprising multi-head attention with 4 heads and a position-wise feedforward layer. Residual connections and layer normalization are applied following standard Transformer design principles.

To obtain a fixed-length representation, we apply mean pooling over the sequence dimension, producing a global embedding intended to capture aggregate rhythmic and harmonic structure. This pooled vector is passed through a two-layer feedforward classifier with ReLU activation and dropout before producing four class logits.

Sequences are truncated or padded to a maximum length of L_{max} tokens. Padding positions are masked during attention computation. The total model contains approximately ∼1–2 million trainable parameters, ensuring a lightweight configuration that minimizes overfitting to dataset-specific artifacts.

Training Procedure

Data is partitioned at the artist level into training, validation, and test sets to prevent stylistic leakage. Songs labeled as neutral are excluded from final training and evaluation. Class imbalance is addressed using weighted sampling proportional to the inverse class frequency.

The model is trained using cross-entropy loss with label smoothing ($\epsilon = 0.1$) and optimized using AdamW with an initial learning rate of 10^{-4}. Training proceeds for up to 30 epochs with early stopping based on validation macro-F1. The model with the highest validation macro-F1 is selected for final evaluation.

Performance is reported using overall accuracy and macro-averaged F1 score, with macro-F1 serving as the primary metric due to residual class imbalance.

7 Results

We evaluate the model on a four-way mood classification task using artist-level splits. Performance is reported in terms of overall accuracy and macro-averaged F1 score, with macro-F1 serving as the primary metric due to residual class imbalance.

The model achieves a test accuracy of 0.36 and a macro-F1 score of 0.27, exceeding the random baseline of 0.25 for a four-class task. Although absolute performance remains modest, the improvement over chance indicates that symbolic bassline structure encodes measurable affective information.

Per-class analysis reveals substantial variation across mood categories. The **joyful** class achieves the highest F1 score (approximately 0.50), suggesting that faster tempi and higher bassline density provide strong structural cues that are readily captured by the model. In contrast, **melancholic_wistful** exhibits the lowest performance (F1 $\approx$ 0.11) and is frequently misclassified as **calm_easygoing**.

This asymmetry suggests that high-energy moods are more strongly correlated with structural descriptors such as tempo and note density, whereas lower-energy or introspective moods may depend more heavily on timbral, melodic, or harmonic information that is not present in bassline-only symbolic representations.

We evaluated several architectural and optimization variations, including increased embedding dimensionality, modified dropout rates, alternative sequence lengths, and class-weighted loss functions. None of these variations yielded consistent improvements in macro-F1, suggesting that performance limitations are primarily representational rather than architectural.

8 Discussion and Conclusion

This work investigated whether musical mood can be inferred from symbolic MIDI basslines alone, without access to audio or lyrical information. Our results demonstrate that structural representations encode limited but measurable affective cues, as evidenced by performance exceeding chance on a four-way mood classification task under artist-level evaluation. In particular, high-energy moods such as joyful are more reliably predicted, indicating that tempo and note density provide strong structural correlates of perceived affect.

At the same time, performance remains modest overall, and low-energy or introspective moods exhibit substantial confusion. This asymmetry highlights a central limitation of purely symbolic bassline representations: while structural descriptors capture coarse energy characteristics, they lack the timbral, melodic, and expressive detail that often shapes nuanced emotional perception. The inability of architectural modifications to significantly improve performance further suggests that these limitations are representational rather than model-specific.

Beyond classification accuracy, this study contributes a weakly supervised pipeline for large-scale symbolic mood modeling, including bassline extraction, heuristic labeling, confidence-based refinement, and token-based sequence learning. By isolating basslines as a controlled structural test case, we provide empirical insight into the role of harmonic and rhythmic foundations in affective modeling.

Overall, our findings suggest that symbolic structure alone forms a meaningful but incomplete component of musical mood representation. Future work may extend this framework by incorporating additional symbolic layers, such as melody or harmonic context, or by integrating symbolic and audio modalities to jointly model structural and timbral aspects of affect.

References

[1] D. Bainbridge and T. Bell, “Music Information Retrieval,” *IEEE MultiMedia*, vol. 8, no. 3, pp. 86–89, Jul.–Sep. 2001.

[2] C. C. Liem, “An Overview of Music Information Retrieval Methods and Challenges,” *IEEE Trans. Multimedia*, vol. 22, pp. 3–17, 2020.

[3] C. Raffel, “The Lakh MIDI Dataset v0.1,” 2016. [Online]. Available: `https://colinraffel.com/projects/lmd/`

[4] I. Sparsh, “Lakh MIDI Dataset Clean,” *Kaggle Dataset*, 2019. [Online]. Available: `https://www.kaggle.com/datasets/imsparsh/lakh-midi-clean`

[5] M. Soleymani, J.-P. Caro, and K. Pun, “A Survey of Music Emotion Recognition: Features, Classification Schemes, and Challenges,” *IEEE Trans. Affective Computing*, vol. 8, no. 3, pp. 273–289, Jul.–Sep. 2017.

[6] Y. Ouyang and S. Doraisamy, “Comparative Analysis of Music Mood Classification Methods,” in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2021, pp. 205–212.

[7] H. Zhang, Z. Duan, and S. Dixon, “Symbolic Music Representations for Classification Tasks: A Systematic Evaluation,” in *Proc. Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2023, pp. 789–795.

[8] D.-V.-T. Le, A. Rigoulot, and J. Pachet, “Natural Language Processing Methods for Symbolic Music Generation and Information Retrieval: A Survey,” *IEEE Access*, vol. 12, pp. 45 523–45 544, 2024.