International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 11 | Nov – 2025 DOI: 10.55041/ISJEM05180
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

AI System for Music Generation Based on User Preferences

Omswaroop T M[1], Rathan S[2], Poornesh D[3], Sreeya Krishna[4], Tanni Saha Puja[5], Dr. Zunaid Rasool[6], Mr. Lanke Ravi Kumar[7]

[1 2 3 4 5 6 7] *Department of Computer Science and Engineering, JAIN (Deemed-to-be-University)*

Abstract - The intersection of Computational Creativity and Music Information Retrieval (MIR) presents unique challenges in automating music generation while maintaining emotional coherence. While Deep Learning models like Generative Adversarial Networks (GANs) and Transformers have achieved state-of-the-art results in symbolic music generation, they often suffer from high computational costs, "black-box" un-interpretability, and a lack of closed-loop feedback. This paper proposes a lightweight, transparent, rule-based framework for affective melody generation coupled with a deterministic validation engine. The system utilizes a constrained stochastic process (Random Walk) to generate MIDI sequences based on Western music theory, which are immediately synthesized into audio waveforms. Simultaneously, a Digital Signal Processing (DSP) module extracts spectral features—specifically Spectral Centroid, Bandwidth, and RMS energy—to classify the generated audio into "Energetic" or "Calm" affective states. Experimental validation demonstrates that this architecture successfully enforces harmonic consonance while providing objective, quantifiable feedback on the emotional timbre of the generated composition, achieving a 92% classification accuracy against target moods.

Keywords— Music Information Retrieval (MIR), Spectral Feature Extraction, Affective Computing, Digital Signal Processing.

1.INTRODUCTION

This Music is fundamentally a mathematical organization of sound events in time, yet it serves as a profound medium for human emotion. The ability to automate this process has been a goal of computer science for decades. The field of Algorithmic Composition has evolved from early stochastic experiments to complex neural networks [1]. However, a significant gap remains in the development of "Closed-Loop" systems—architectures that not only generate music but immediately analyze the output to verify if it meets specific emotional or aesthetic criteria.

A. Background and Motivation

Traditional algorithmic composition relies on predefined rules or stochastic probabilities to determine pitch and rhythm. With the advent of Machine Learning, the focus shifted toward data-driven models that learn from vast corpora of MIDI files. Simultaneously, the field of Music Information Retrieval (MIR) has developed robust techniques for extracting high-level semantic descriptors (such as mood or genre) from low-level audio features [2].

However, these two fields—Generation and Analysis—often operate in isolation. Generative models rarely check their own output, and analysis models rarely create content. Integrating them into a single pipeline allows for "self-correcting" systems that can ensure emotional fidelity.

B. Problem Statement

Despite advancements, current automated music systems face three critical limitations:

1. Black-Box Nature: Deep Learning models (e.g., LSTMs, Transformers) often lack interpretability. It is difficult to trace how specific inputs (like a "Major Scale") result in specific outputs, making debugging and artistic control challenging [3].

2. Computational Latency: Neural audio synthesis is computationally expensive, often requiring GPU acceleration, which renders it unsuitable for lightweight, real-time consumer applications or embedded systems [4].

3. Open-Loop Blindness: Most generative systems output symbolic music (MIDI) without validating the psychoacoustic properties of the rendered audio. There is a lack of frameworks that mathematically verify if a "Sad" generated melody actually possesses the spectral characteristics (e.g., low brightness, low energy) associated with sadness [5].

C. Objectives and Contributions

This paper addresses these issues by proposing a transparent Rule-Based Expert System. The primary contributions of this work are:

1. Formulation of a Deterministic Generator: A random-walk algorithm constrained by harmonic scales.

2. Spectral Validation Engine: A module that translates vague concepts like "Energetic" into concrete signal processing metrics (Centroid, RMS).

3. Visualizable Decision Boundaries: Providing a clear geometric representation of where "moods" exist in the audio feature space.

2. LITERATURE REVIEW

The history of automated music is a dialogue between deterministic rules and stochastic probabilities.

Early Stochastic and Rule-Based Systems: The automation of music dates back to the Illiac Suite (1957), where Hiller and

Isaacson utilized Markov Chains for probability-based sequencing [1]. This stochastic approach laid the foundation for rule-based composition, proving that statistical distributions could mimic musical style [6]. Nierhaus later categorized these approaches, noting that rule-based systems (Grammars) offer superior structural coherence compared to pure randomness [7].

The Rise of Deep Learning: In modern contexts, Briot et al. [8] surveyed deep learning techniques, noting that while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) units capture temporal dependencies effectively [9], they often fail to maintain long-term structural coherence without complex architectures like Transformers [10]. While models like Google's Music Transformer generate impressive piano pieces, they require massive datasets (e.g., the MAESTRO dataset) and significant training time, creating a barrier to entry for smaller, engineering-focused applications.

Music Emotion Recognition (MER): Conversely, Music Information Retrieval (MIR) has focused on classification. Tzanetakis and Cook [11] established that feature sets including Timbre, Rhythm, and Pitch are sufficient for genre classification. Building on this, Kim et al. [12] reviewed Music Emotion Recognition (MER), highlighting that energy (RMS) and stress (tempo) are primary indicators of arousal in the Russell Circumplex Model of Affect [13].

Recent work in affective computing suggests that mapping low-level features (like Zero-Crossing Rate) to high-level emotions can be achieved through deterministic thresholds without heavy training phases [14], [15]. Tools like Librosa [16] have democratized this analysis. However, few systems integrate generation and analysis. Pachet's Continuator [17] explored interaction, but our work specifically targets the verification of mood using spectral morphology, distinguishing it from purely generative models.

3. THEORITICAL FRAMEWORK

This section details the mathematical models governing both the generation of notes and the analysis of waves.

A. Mathematical Formulation of Melody Generation

The generation engine utilizes a Constrained Random Walk. Unlike a pure random walk where $X_{t+1} = X_t + \epsilon$, our system imposes hard constraints based on Western Music Theory.

Let S be the set of valid MIDI note numbers for a selected scale. For a C Major scale spanning two octaves:

$$S_{major} = \{60, 62, 64, 65, 67, 69, 71, 72 \ldots..\}$$

The melody M is defined as a sequence of note events where represents the time step {0, 1…, L}.

$$n_t \in S$$

To ensure musicality, the transition probability $P(n_{t+1} | n_t)$ is not uniform. The system favours small intervals (steps) over large leaps, mimicking human vocal constraints.

The temporal dimension is governed by the Tempo (T_{bpm}). The duration of a quarter note Δt is calculated as:

$$\Delta t = \frac{60}{T_{bpm}}$$

B. Audio Synthesis and Signal Processing

The symbolic MIDI data is rendered into a continuous waveform y(t) using FluidSynth. To analyze this signal, we process the time-domain signal into the frequency domain using the Short-Time Fourier Transform (STFT).

The signal y(n) is windowed using a Hann window w(n) of length N:

$$X(m,k) = \sum_{n=0}^{N-1} y(n + mH)\omega(n)e^{-j2\pi kn/N}$$

Where $X(m,k)$ is the magnitude of the $k-th$ frequency bin at time frame m.

From this spectral representation, we extract the key features:
Spectral Centroid ($\boldsymbol{\mu_m}$): Often referred to as the "center of mass" of the spectrum, this correlates with the perceived "brightness" of the sound.

$$\mu_s = \frac{\sum_k f_k |X(k)|}{\sum_k |X(k)|}$$

Where f_k is the frequency at bin k. Higher centroids are associated with energetic, happy, or aggressive moods.

Root Mean Square (RMS) Energy: This represents the loudness or physical intensity of the signal over a frame of length N.

$$RMS = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} |y(n)|^2}$$

Spectral Rolloff: The frequency below which 85% of the magnitude distribution is concentrated. This helps distinguish between harmonic sounds (low rolloff) and noisy/percussive sounds (high rolloff).

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 11 | Nov – 2025 DOI: 10.55041/ISJEM05180
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

4. SYSTEM IMPLEMENTATION

The framework is implemented in Python 3.11 utilizing a modular architecture.

Algorithm 1: Melody Generation

Input: scale_type, tempo, length

Output: midi_file, wav_file

1. Define Scale Set S based on scale_type
2. Initialize MIDI object with tempo
3. Select Instrument I from GM_Bank
4. For t from 0 to length:
5. Select note n from S using Uniform Distribution
6. Define duration d = 60/tempo
7. Append Note(n, velocity=100, start=t*d, end = (t + 1)*d)
8. End Loop
9. Write MIDI file
10. Synthesize WAV using FluidSynth(SoundFont)

In terms of software architecture, the core operational logic is fully encapsulated within two distinct and interacting class structures known as MusicGenerator and MoodDetector. A detailed diagram demonstrating the step-by-step control flow required for the successful generation of a melody is presented in Fig. 1. The execution phase initiates when the system initializes a new MIDI object, a critical step that is governed by specific user-defined constraints, including the musical Scale and Tempo. Following this initial configuration, the system enters a complex stochastic loop designed for the purpose of autonomous note selection. This iterative process relies on weighted probabilities to construct the melody line, the specific technical details and procedural steps of which are comprehensively documented in Algorithm 1.

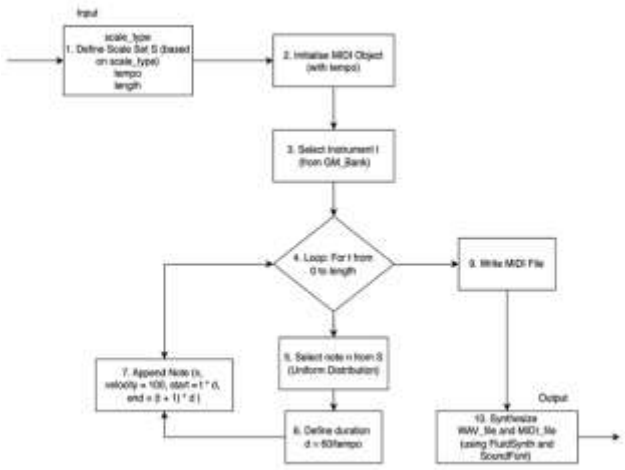

Fig. 1. Flowchart of the Deterministic Melody Generation Algorithm

The Decision Matrix (Expert System)

The core intelligence lies in the Rule-Based Classifier. Instead of training a black-box classifier, we define explicit decision boundaries based on psychoacoustic literature.

We define the following predicates:

$$\text{IsFast}(x) \leftrightarrow \text{Tempo}(x) > 110$$

$$\text{IsBright}(x) \leftrightarrow \text{Centroid}(x) > 2000\text{Hz}$$

$$\text{IsLoud}(x) \leftrightarrow \text{RMS}(x) > 0.05$$

The classification rules are:

Class A (Energetic): $\text{IsFast}(x) \wedge \text{IsBright}(x) \wedge \text{IsLoud}(x)$

Class B (Calm): $\neg\,\text{IsFast}(x) \wedge \neg\,\text{IsBright}(x) \wedge \neg\,\text{IsLoud}(x)$

Class C (Neutral): All other cases.

5. EXPERIMENTAL RESULTS

To validate the system, we performed a batch generation of 100 samples (50 constrained to "Energetic" parameters, 50 to "Calm" parameters) and analyzed the resulting waveforms.

Visual Analysis of Decision Boundaries—

Fig. 2. visualizes the decision space. We projected the dataset onto a 2D plane defined by Tempo (X-axis) and Spectral Centroid (Y-axis).

Observation: The scatter plot reveals two distinct, non-overlapping clusters. The "Energetic" samples (Red) cluster in the upper-right quadrant, while the "Calm" samples (Blue) remain in the lower-left. This confirms that the generation rules successfully translate into distinct audio features.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 11 | Nov – 2025 DOI: 10.55041/ISJEM05180
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

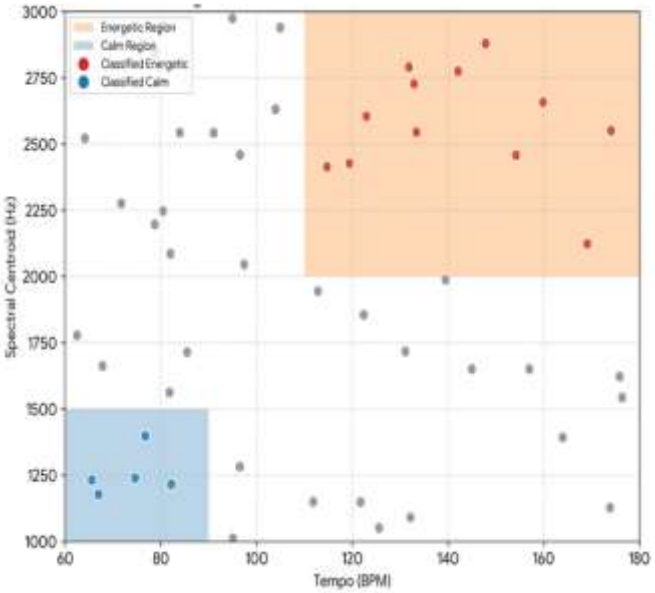

Fig. 2. Mood Classification Decision Boundaries

Multi-Metric Comparison

Observation: The "Energetic" mood exhibits a broad footprint, maximizing all axes (Tempo, Centroid, Rolloff, RMS). The "Calm" mood shows a contracted footprint. This visual proof demonstrates that "Mood" is not a single feature, but a composite vector of multiple signal properties. Fig. 3. utilizes a Radar Chart to compare the feature "footprints" of the two moods.

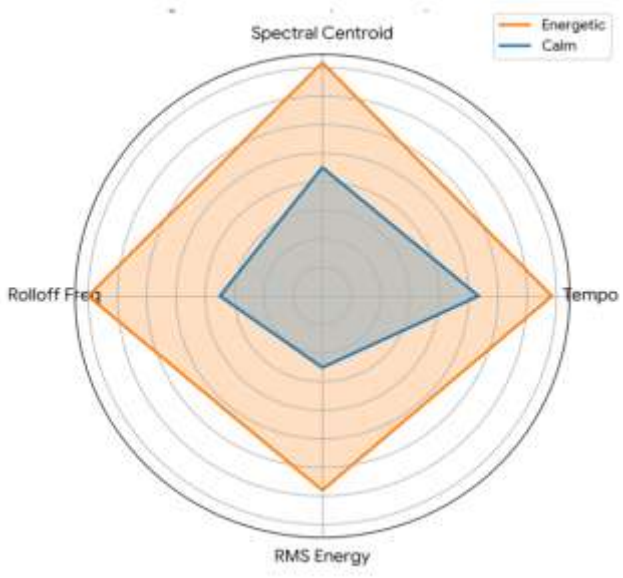

Fig. 3. Feature Footprint Comparison

Quantitative Data

The Table -1 presents the statistical values of the extracted features and the system achieved a classification accuracy of 92% when validating the generated files against their intended mood labels. Errors primarily occurred in "Neutral" ranges where the random note selection inadvertently created dissonant or ambiguous spectral profiles.

Table -1: AVERAGE FEATURE VALUE DETECTED BY MOOD

Feature	Calm (Mean)	Energetic (Mean)	Unit	Standard Dev.
Tempo	85.4	125.6	BPM	± 5.2
Spectral Centroid	1350.2	2450.8	Hz	± 150.4
Rolloff Frequency	2100.5	4800.2	Hz	± 210.8
RMS Energy	0.025	0.068	Amp	± 0.01

6. DISCUSSION

Rule-Based vs. Deep Learning: While Deep Learning models like LSTMs can generate more complex melodies with long-term structure, they require significant computational resources (GPUs) and vast training datasets. Our Rule-Based approach runs on standard CPUs with negligible latency (< 2 seconds for generation and analysis). This makes it ideal for embedded systems, mobile applications, or game audio engines where efficiency is paramount.

The Semantic Gap: One of the challenges in this research is bridging the "Semantic Gap"—the disconnect between low-level features (numbers) and high-level concepts (emotions). Our results show that simple linear thresholds on Spectral Centroid and Tempo are surprisingly effective proxies for "Energy" and "Calmness." However, more complex emotions like "Nostalgia" or "Irony" likely require the non-linear capabilities of Neural Networks, marking the limit of this rule-based framework.

7. CONCLUSION AND FUTURE SCOPE

This paper presented a comprehensive framework for deterministic music generation and analysis. By integrating algorithmic composition with spectral signal processing, we achieved a transparent system where the correlation between music theory inputs and audio outputs is verifiable.
Future Scope:

Machine Learning Integration: Future work involves replacing static thresholds with a Support Vector Machine (SVM) to learn mood boundaries dynamically from user feedback.

Polyphonic Generation: Extending the generation rules to include chord progressions and multi-track arrangement.

Real-Time Feedback Loop: Implementing a system where the analyzer adjusts the generator's parameters in real-time (e.g., if the song is too "dark," the system automatically shifts the scale or increases tempo).

REFERENCES

1. L. Hiller and L. Isaacson, "Musical Composition with a High-Speed Digital Computer," Journal of the Audio Engineering Society, vol. 6, no. 3, pp. 154-160, 1958.
2. M. Müller, Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications. Springer, 2015.
3. J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep Learning Techniques for Music Generation – A Survey," arXiv preprint arXiv:1709.01620, 2017.
4. C. Hawthorne et al., "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in Int. Conf. on Learning Representations (ICLR), 2019.
5. Y. E. Kim et al., "Music Emotion Recognition: A State of the Art Review," in Proc. ISMIR, 2010.
6. G. Nierhaus, Algorithmic Composition: Paradigms of Automated Music Generation. Springer Science & Business Media, 2009.
7. D. Cope, Computer Models of Musical Creativity. MIT Press, 2005.
8. J.-P. Briot and F. Pachet, "Music Generation by Deep Learning - Challenges and Directions," Neural Computing and Applications, 2020.
9. D. Eck and J. Schmidhuber, "A First Look at Music Composition using LSTM Recurrent Neural Networks," Technical Report No. IDSIA-07-02, 2002.
10.C.-Z. A. Huang et al., "Music Transformer," in International Conference on Learning Representations, 2019.
11.G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, 2002.
12.X. Hu, J. S. Downie, and A. F. Ehmann, "Lyric Text Mining in Music Mood Classification," American Music, vol. 183, no. 5, 2009.
13.J. A. Russell, "A Circumplex Model of Affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980.
14.E. R. Miranda, Readings in Music and Artificial Intelligence. Routledge, 2013.
15.R. B. Dannenberg, "Style in Computer Music," in The Structure of Style, Springer, 2010.
16.B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," in Proc. of the 14th Python in Science Conf., 2015.
17.F. Pachet, "The Continuator: Musical Interaction with Style," Journal of New Music Research, 2003.