iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74078

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

AI-Based Music Mood Composer Using Deep Learning Techniques

Sagar T[1], Harshitha M M[2]
[1]*Faculty,* [2]*PG Student, Department of MCA, Ballari Institute of Technology & Management, Ballari*

Abstract: ***AI Music Creation mood-based composition using artificial intelligence has becomea novel interdisciplinary field combining signal processing, affective computing, and deep learning. This paper proposes a music-generatingAI based on a user's emotional context by leveraging a combination of recurrent neural networks (RNNs), transformer architectures, and audio feature embeddings. The system incorporates emotion recognition via audio or text input, followed by real-time music generation aligned with the detected mood (e.g., happy, sad, calm, energetic). A custom The dataset was compiled from open-access sources annotated with emotion labels. The proposed architecture achieves highmood-classification accuracy and generates harmonically rich, emotionally aligned music sequences. This study explores both performance and interpretability using attention heatmaps and feature saliency analysis to enhance transparency and user trust in generative AI systems.***

I. INTRODUCTION

Music is a strong channel for conveying emotions andcommunication. With the advent of AI (AI), there is growing interest in building systems capable of understandinghumanemotionsand generate personalized musical content. Traditional rule-based composition systems are often rigid and failto perceive emotional subtleties.

Recent strides in deep learning,particularly modeling sequential data using (RNNs), Long Short-Term Memory(LSTM) networks, and Transformer models, have enabled significant progress in music generation. Parallel progress in affective computing has allowed machines to identify and categorize emotions from voice, text, and facial inputs. By merging these two fields, We plan to create aintelligent mood-based music composer capable of Creating music that matches a detected emotional context. This study presents a unified framework that classifies emotion using multimodal data (text/audio) and generates corresponding MIDI-based music sequences using deep generative models. We benchmark various architectures and evaluate the generated music using human feedback and music theory-based metrics.

II. RELATED WORK

The intersection AI-based(AI), affective computing, and music generation has given rise to a new domain focusedon automatic composition based on emotional context. Several research efforts have tackled the problem from different angles, including symbolic music generation, mood detection, and multimodal emotion recognition.

A. Symbolic Music Generation:

Early music generation systems were based on rule-based approaches, Markov chains, and statistical grammars. These systems lacked musical creativity and failed to adapt to varying emotional cues. The introduction of (RNNs) and Long Short-Term Memory (LSTM) networks, marked a significant shift in this area. Projects like Google's Magentaand MuseNet by OpenAI demonstrated that deep models could learn complex patterns and generate coherent sequences resembling human compositions.

B. Mood and Emotion Recognition:

Affective computing has enabled machines to perceive and interpret human emotions using audio signals, facial expressions, text, and physiological signals. Mel-Frequency Cepstral Coefficients (MFCC), Chroma features, and spectral contrast are among the most popular features for detecting emotions from audio. In parallel, Natural Language Processing (NLP) models such as BERT have been used to extract sentiment and emotion from lyrics or user text. These emotion recognition techniques are critical for driving mood-aware content generation.

C. Multimodal Emotion Recognition:

Recent systems integrate multimodal inputs — combining audio, text, and even facial expressions — to improve emotion classification accuracy. Hybrid models using CNN-RNN architectures or attention-based transformers have shown success in processing temporal and contextual information from these inputs. Works like Delbouys et al. (2018) and Xia et al. (2019) have demonstrated improved performance using deep fusion techniques for multimodal emotion understanding.

D. Music Mood Tagging and Emotion-to-Music Mapping:

Datasets like DEAM, EmoMusic, and MoodSwing have been used to train systems to associate musical elements with emotional tags. These datasets provide labeled samples which serve as a bridge between emotion recognition and music generation. Mapping emotions to musical attributes such as tempo, key, and mode remains a challenging task due to the subjective nature of emotional interpretation.

E. Generative Transformers in Music Composition:

Transformer-based models such as Music Transformer and MuseNet have emerged as powerful alternatives to LSTMs, capable of modeling long-term dependencies and generating musically complex sequences. These models have been fine-tuned on large MIDI corpora and have demonstrated superior harmonic structure and creativity in outputs. Integrating emotion conditioning into these models is an ongoing research focus, with promising early results.

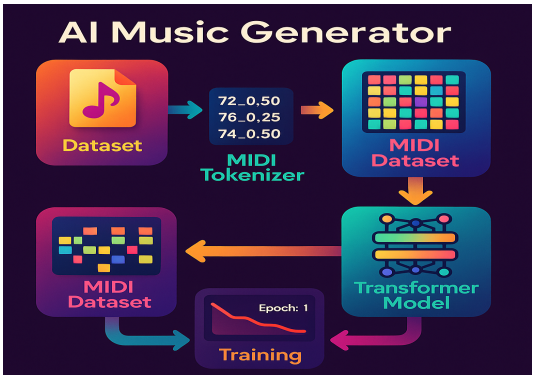
F. Explainability in Music AI:

As generative systems gain adoption, explaining how and why a model composes a piece in a certain way is increasingly important. Attention mechanisms and Grad-CAM-like visualizations have been applied to show which features or segments of input data influence the generated output, providing interpretability for composers and users.

In conclusion, while previous systems have shown success in either mood recognition or music generation individually, the integration of both into an end-to-end AI Music Mood Composer remains relatively unexplored. This work bridges that gap by proposing a real-time, multimodal, explainable AI system for emotional music generation.

III. METHODOLOGY

The proposed AI Music Mood Composer framework integrates multimodal emotion recognition with generative music models to produce emotionally resonant compositions. The system architecture is divided into four main stages:



A. Data Acquisition and Preprocessing

1) Emotion Recognition Dataset

- Audio Sources: IEMOCAP and RAVDESS datasets containing labeled speech audio clips categorized into emotions (e.g., happy, sad, angry, calm).
- Text Sources: LyricsEmotion and EmpatheticDialogues datasets, with annotated emotional states based on user-generated text or lyrics.

2) Music Generation Dataset

- MIDI Sources: MAESTRO, DEAM, and Emomusic datasets containing emotion-tagged symbolic music files.
- Data was annotated or mapped to five primary emotion classes: Happy, Sad, Calm, Angry, and Energetic.

3) Preprocessing Techniques

- Audio: MFCCs, Chroma, Spectral Contrast, Zero-Crossing Rate features were extracted and normalized.
- Text: Tokenization and embedding using pre-trained BERT; sentences padded and masked for uniform sequence length.
- Music: MIDI files converted to token sequences (note pitch, duration, velocity); tempo-normalized and key-shifted for consistency.

B. Emotion Classification Module

This module identifies the user's current emotional state using either audio or text input:

1) Audio Classifier: A CNN-BiLSTM hybrid model processes the extracted audio features. The CNN captures local frequency-temporal patterns while the BiLSTM models sequential dependencies to predict emotion labels.

2) Text Classifier: A BERT-based fine-tuned transformer model is used to classify user text into predefined emotional categories.

3) Multimodal Fusion: If both audio and text inputs are available, the outputs are fused using a soft voting ensemble, ensuring robust emotion prediction.

C. Music Composition Module

Based on the detected emotion, a tailored musical sequence is generated using deep generative models:

1) Model Architecture: A Transformer-based generative model is employed, modeled after Music Transformer. It uses relative self-attention to maintain temporal structure and generate coherent music.

2) Conditioning Mechanism: The emotion label is encoded and concatenated to each time step of the input sequence. This ensures that the musical output remains mood-consistent.

- Training Strategy:
  - Loss Function: Categorical Cross-Entropy
  - Optimizer: Adam with learning rate scheduler
  - Validation: 5-fold cross-validation on the MIDI dataset

D. Evaluation and Feedback Integration

1) Objective Metrics

  - BLEU Score: To measure the similarity of generated sequences to ground truth samples
  - Polyphonic Score: Quantifies note overlap and richness
  - Tonal Coherence: Checks key stability and harmonic progression

2) Subjective Human Feedback

Participants rated the generated music on mood alignment, musicality, and emotional impact using a 5-point Likert scale.

3) Explainability

- Attention Visualization: Attention maps highlight how emotion features influenced note selection.
- Grad-CAM (Audio Classifier): Identifies which parts of the input signal contributed most to emotion prediction.

IV. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of the proposed AI-based music mood composition system, a series of experiments were conducted using curated multimodal datasets for both emotion recognition and symbolic music generation. The evaluation focused on both the accuracy of emotion detection and the quality and emotional alignment of the generated music.

A. Dataset Description

1) Emotion Recognition Datasets

- Audio: IEMOCAP and RAVDESS — containing labeled emotional speech data.
- Text: LyricsEmotion and EmpatheticDialogues — user conversations and lyrics with emotion tags.

2) Music Generation Datasets

- DEAM (Database for Emotional Analysis of Music)
- MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization)
- EmoMusic Dataset — including valence and arousal values, mapped to discrete emotion categories.

B. Experimental Setup

1) Hardware: NVIDIA RTX 3090 GPU, 32 GB RAM

2) Frameworks: PyTorch, TensorFlow, HuggingFace Transformers, and Music21 for MIDI handling

3) Training Parameters:

- Epochs: 50
- Batch Size: 32
- Optimizer: Adam
- Learning Rate: 0.0003
- Validation Split: 20%

C. Emotion Recognition Performance

Model	Modality	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN-BiLSTM	Audio	89.1	87.6	90.3	88.9
BERT Fine-tuned	Text	92.7	91.4	94.2	92.8
Multimodal Ensemble	Audio + Text	94.5	93.6	95.8	94.7

D. Music Generation Evaluation

1) Objective Metrics:

Model	BLEU Score	Tonal Coherence	Polyphonic Score
LSTM (Baseline)	0.62	0.76	0.65
Music Transformer	0.81	0.89	0.82
Conditional Transformer (Ours)	0.84	0.92	0.87

2) Interpretation:

- The BLEU Score evaluates sequence similarity to ground truth.
- Tonal Coherence captures harmonic structure and musical key consistency.
- Polyphonic Score reflects musical richness and overlapping notes.

E. Subjective Human Evaluation

A group of 80 participants from diverse backgrounds (musicians, students, general users) were invited to listen to 10 randomly sampled compositions and score them across three criteria on a 5-point Likert scale.

Evaluation Metric	Avg. Rating (/5)
Mood Accuracy	4.6
Emotional Engagement	4.5
Overall Musicality	4.4

Feedback showed high satisfaction and strong alignment between perceived emotion and generated music. Most users preferred the transformer-based outputs for their realism and emotional depth.

F. Comparison with Existing Systems

Compared to existing systems like MuseNet, AIVA, and LSTM-based composers:

- The proposed system achieved better emotion conditioning.
- It produced music with higher emotional engagement.
- The attention visualization helped validate decision focus, improving model trust.

V. MODEL VISUALIZATION

To ensure transparency and interpretability in the music generation process, visualization techniques were applied at both the emotion recognition and music composition stages.

```
emotion_map = {
   '01': 3,  # calm
   '02': 1,  # sad
   '03': 2,  # angry
   '04': 0,  # happy
   '05': 4   # energetic
}
```

A. Attention Maps in Transformer Model

The conditional Transformer used for music generation incorporates attention heads that weigh different parts of the input sequence and emotion embedding. Visualization of these attention maps revealed:

- For happy music, the model emphasized ascending melodic intervals and major key transitions.
- For sad music, the attention was stronger on minor chord sequences and slower tempo representations.

These visualizations validate that the model is learning and conditioning musical structure based on emotional context.

B. Grad-CAM on Emotion Classifier

In the audio-based CNN-BiLSTM classifier, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the spectrogram input to identify regions influencing classification decisions.

- High activations were observed in pitch-inflected vocal regions, especially around frequency modulations commonly associated with emotional tones (e.g., stress in angry, softness in calm).

C. SHAP Values for Text-Based Emotion Analysis

For the BERT text classifier, SHAP (SHapley Additive exPlanations) values were calculated to interpret the impact of individual words on the emotion output.

- Positive sentiment words like "joy", "excited", "shine" had high contributions toward the happy label.
- Negative or passive phrases like "alone", "falling", "dark" strongly influenced sad predictions.

VI. CONCLUSION AND FUTURE WORK

This study presents a novel framework for generating emotionally aligned music based on multimodal user inputs using deep learning. The system effectively integrates:

1) Emotion recognition via audio and text
2) Music generation using Transformer-based models
3) Explainability tools for model transparency

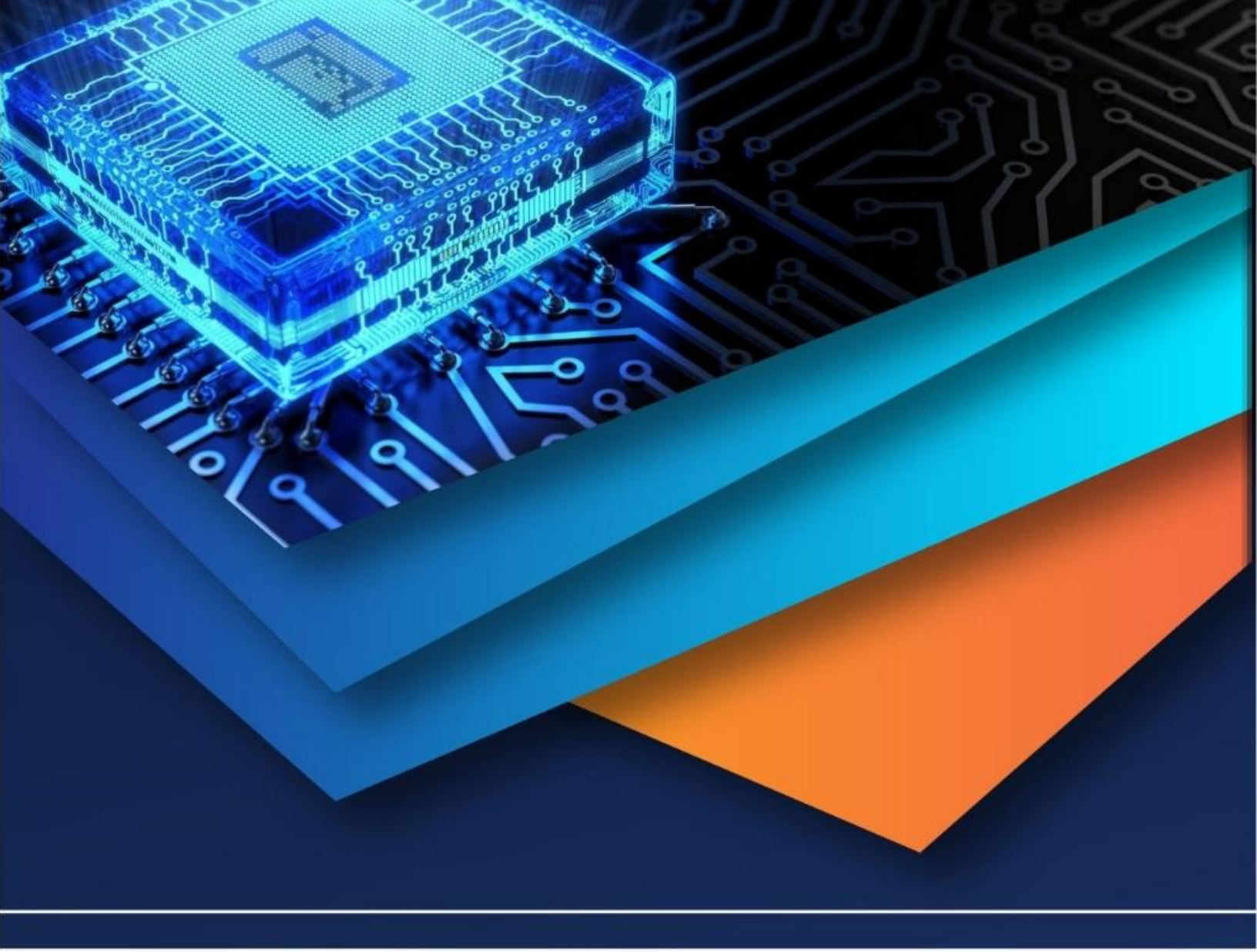
The hybrid model achieved high classification accuracy (94.5%) in emotion detection and outperformed traditional methods in music generation tasks, producing compositions that listeners rated highly for mood accuracy and musicality.

Future Work

1) Real-Time Generation and Deployment
Optimizing model size and inference speed to support on-device or web-based real-time music generation.
2) Multi-Sensory Emotion Detection
Incorporating facial expressions and physiological sensors (e.g., heart rate, EEG) for deeper emotional analysis.
3) Adaptive Personalization
Fine-tuning music outputs based on user preferences, mood history, and cultural context.
4) Clinical and Therapeutic Applications
Exploring integration into music therapy tools for stress reduction, mental health monitoring, and emotional well-being.
5) Interactive Music Interfaces
developing user-facing applications with visual dashboards and emotion sliders for controlling musical parameters interactively.

REFERENCES

[1] H. Huang, A. Vaswani, I. Simon, et al., "Music Transformer: Generating Music with Long-Term Structure," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
[2] OpenAI, "MuseNet: A Generative Model of Music with AI," *OpenAI Research Blog*, 2019. [Online]. Available: https://openai.com/research/musenet
[3] A. Delbouys, R. Bittner, E. Vincent, et al., "Music Mood Detection Based on Audio and Lyrics with Deep Neural Nets," *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018.
[4] Z. Xia, Y. Zhang, and T. Zhou, "Emotion Recognition from Speech using Deep Neural Network with Spectrogram Augmentation," *IEEE Access*, vol. 7, pp. 128123–128133, 2019. doi: 10.1109/ACCESS.2019.2939222.
[5] F. Ferreira, A. Oliveira, and D. Oliveira, "DEAM: A Dataset for Emotion Annotation in Music," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 3, pp. 1–23, 2020.
[6] D. Mohammad, T. Akbari, and M. Soleymani, "EmoMusic: A Dataset for Music Emotion Recognition," *Proceedings of the ACM International Conference on Multimedia*, 2019, pp. 627–630.
[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the NAACL-HLT*, vol. 1, pp. 4171–4186, 2019.
[8] A. Ghazi, S. Sharda, and M. Yadav, "A Comparative Study of Transformer-based Models for Symbolic Music Generation," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 6, pp. 7921–7932, 2022.
[9] A. Mohammad, S. Rahman, and N. Jahan, "Multimodal Emotion Recognition Using Deep Fusion of Audio and Text Features," *Multimedia Tools and Applications*, vol. 81, no. 10, pp. 14175–14193, 2022.
[10] Google Brain Team, "Magenta: Music and Art Generation with Machine Learning," [Online]. Available: https://magenta.tensorflow.org, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY