International Journal of Basic and Applied Sciences, 14 (2) (2025) 151-162

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/2eth4t83
Research paper



A Novel Framework for Automatic Music Generation Using Hybrid AI Techniques

V. Bhuvana Kumar [1] *, Dr. Narayana Rao Appini [2], N. Yedukondalu [1]

[1] *Assistant Professor, Computer Science & Engineering, N.B.K.R Institute of Science and Technology, India*
[2] *Professor & Head, AI& DS and IT, NBKR Institute of Science and Technology, India*
**Corresponding author E-mail: vbkumar1993@gmail.com*

Received: May 6, 2025, Accepted: May 18, 2025, Published: June 11, 2025

Abstract

Automatic music generation using Artificial Intelligence (AI) has seen remarkable progress in recent years, especially due to the development of deep learning techniques. These techniques have not only enabled computers to copy, but have also produced music creatively that resembles human creations. This research introduces a novel and strong approach is MuseHybridNet- a hybrid model designed to carry forward the boundaries of AI-generated music.
What sets is its unique integration of transformer-based architecture, emotional reference conditioning, and adaptive style transfer to music. Each of these components plays an important role:

- Transformer architecture, which has proved to be highly effective in natural language processing functions, is employed to model long-term dependence in music, allowing the system to generate compositions that are structurally sound and sweetly consistent over time.
- Emotional reference allows conditioning models to generate music for a specific mood or emotion. Whether it is pleasure, sadness, enthusiasm, or peace, Musehybridnet can adjust its output accordingly, resulting in music and human feeling as a result.
- The adaptive style enables the transfer system to mix and originally move the music styles. For example, the model can produce a classical piece with modern pop effects or create jazz compositions with an indication of electronic music. It gives users a powerful tool to experiment with creative cross-style compositions.

Musehybridnet is designed to work to represent both symbolic data (eg, MIDI files, which include notes, timing, and instruments) and raw audio data, which allows it to capture the fine nuances of sound like texture and timbre. By combining these two data types, the model is better able to understand and repeat the complexities of real-world music.
The proposed model shows that Musehybridnet continuously performs better than existing models in major regions, such as consistent models, emotional accuracy, and stylistic diversity. The audience stated that the compositions produced by our model seem more natural, emotionally attractive, and creatively rich compared to other AI systems.
In short, this research presents an important step in the field of AI-based music, offering a tool that supports human creativity and produces music aligned with emotional and stylistic intentions, though full human-AI co-creation remains an area for future exploration.
The synthesis of AI, music theory, and emotion modeling represents meaningful work since it both addresses existing gaps in issues of musical expressiveness and style transfer while meaningfully contributing interdisciplinary value to areas of music cognition, therapeutic sound design, and creative industry uses, such as film and interactive entertainment.

Keywords: *Music Generation Using; Hybrid Ai Technique; Musehybridnet; AI-Based Music.*

1. Introduction

The AI-managed music generation has been a significant change over the years. In its early stages, the music generation trusted a lot in the rules-based systems, which used hard-coded music rules for the creation of tunes. While these systems could technically produce correct music, they often lacked creativity, emotional resonance, and lack of adaptability.
The panorama of music production has changed dramatically, with deep learning and preferred fashions, mainly with the rise of nerve networks. These models can analyze patterns from big datasets, permitting them to produce songs that appear greater natural and human-like. However, despite these advances, many existing fashions are nevertheless low in the most important areas: emotional intensity and stylistic variability. The track they produce can be structurally sound, but it often seems flat or normal, fails to explicitly unique emotions, or fits numerous tune patterns and styles.
To deal with these limitations, this study introduces a unique AI-primarily based framework that goes beyond producing simply coherent sequences of notes. Our device is designed to generate a tune that is both emotionally expressive and stylistically adaptive. One of the middle innovations of this framework is its ability to contain emotional cues both through direct consumer enter (e.g., specifying a fa-

Copyright © V. Bhuvana Kumar et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

vored emotion like "despair" or "completely happy") or with the aid of getting to know from labeled datasets that associate tracks with emotional tags.
Additionally, the model supports stylistic customization, allowing it to generate songs in unique genres or fuse patterns creatively. This is carried out via an aggregate of strategies consisting of style conditioning and context-aware generation, where the version learns the unique traits of various musical patterns from schooling statistics and adapts its output for this reason.
This research provides a massive leap forward by introducing a system that no longer simply knows the technical shape of music but also captures the emotional motive and stylistic essence, resulting in more engaging, personalized, and human-like musical compositions.
MuseHybridNet's design combines long-term musical modeling, emotion-driven conditioning, and genre-specific adaptation, responding to long-standing difficulties facing AI-generated music. This makes the MuseHybridNet innovation very applicable to the academic research in these fields of AI and deep learning, yet also has potential applications in real-world situations such as adaptive soundtracks, music therapy, and co-creative composition projects. This interdisciplinary blend mirrors the objectives of the journal to favour practices that promote both innovation and collaborative research across scientific and artistic domains. By combining techniques from machine learning, affective computing, and musicology, this work contributes to interdisciplinary dialogue and invites collaboration between engineers, artists, and psychologists.
MuseHybridNet integrated all aspects of emotional conditioning, stylistic variance, and long-term structure modeling based on the advances made through Music Transformer, MuseGAN, EmoMT, and MusicVAE. Following that, the next section describes its architecture, data representation, and training methodology.
The proposed work also intersects with affective neuroscience and psychology, fields that study how humans perceive and respond emotionally to music. Understanding neural mechanisms of music perception can inform emotion modeling, while psychological theories like the Circumplex Model of Affect can help design more effective conditioning mechanisms. MuseHybridNet thus opens doors for interdisciplinary collaboration with researchers in music cognition, therapy, and user-centric design.
The emotional conditioning part of MuseHybridNet has a good degree of compatibility with psychological theories, particularly Russell's Circumplex Model of Affect, which provides a method for visually mapping emotions along an arousal axis and cardinal axes of valence [25]. Future work may involve collaboration with cognitive scientists as well as music psychologists to refine the modeling of emotion and to validate actual listener perception.

2. Related work

Advancements in deep learning and artificial intelligence have accelerated the development of automatic music generation. Researchers have proposed a diverse variety of models to account for musical structure and emotional content, as well as for creativity. This literature review underscores distinct contributions from studies in this area, each of which provided different methods and insights.
Within the last two years, the field has shifted toward a new era of large-scale, pre-trained audio generation models for music that leverage raw audio and multimodal measures. One such work, by Agostinelli et al. (2023), proposed MusicLM, a transformer-based model using a hierarchical modeling process to generate coherent musical audio based on rich text prompts [21]. Alternatively, Copet et al. (2023 provided MusicGen, as both a controllable music generation model that conditions on a textual description and a chord progression, and which demonstrated overall improvements in alignment, fidelity, and prompt execution [22]. In addition, Forsgren and Martiros (2022) contributed Riffusion, a diffusion-based model that allows for real-time music synthesis based on immediately synthesizing spectrograms without drifting into the audio domain [23]. These models mark an important point in the trajectory from symbolic music generation, complete with an audio-first paradigm. MuseHybridNet remains focused on symbolic generation using MIDI and conditioning on emotion or style, yet future work could consider how to leverage these large-scale audio models to explore new means of expressiveness, fidelity, and cross-modal control.
In 2018, Huang et al. introduced the music transformer, a major step in the field of AI-based music. Unlike the older models, which depended on the recurrent neural network (RNN) and often struggled to maintain the structure on long pieces, the music transformer used a self-attention mechanism, as well as some called relative condition encoding. This combination allowed the model to better understand how different notes and patterns are related to each other, even when they are far apart in a music sequence. Thanks to this approach, the music was able to produce transformer music that seemed not only a good moment for this time, but also a clear sense of flow and structure - anything like a human musician can be anything like anything. It brought a new level of depth and realism to computer-related music [1].
In 2018, Dong et al. introduced MuseGAN, a progressive version that brought the strength of generative adversarial networks (GANs) into the field of song advent. What made MuseGAN stand out turned into its capacity to generate multi-tune music, meaning it could create numerous instrumental elements, like drums, bass, and piano—all at the same time. This is important due to the fact real song is not fabricated from simply one melody; it's a rich combination of various units working collectively. MuseGAN was designed to understand and seize that interaction among contraptions, permitting it to provide compositions where every part enhances the others evidently. It also provided distinct methods of generating tracks: both all tracks without delay (simultaneous technology) or one after another (sequential technology), which added a layer of flexibility depending on the creative goal. By focusing on how to harmonize multiple instruments correctly, MuseGAN tackled one of the hardest demanding situations in the automated tune era and unfolded new possibilities for growing sensible, layered musical portions using AI [2].
In 2018, Roberts et al. introduced a creative new approach to AI tune technology through developing a hierarchical latent vector model that makes use of recurrent variational autoencoders (VAEs). Their goal turned into addressing a commonplace assignment in AI-generated track: taking pictures of the long-term structure that gives actual song its feel of progression and meaning. Instead of treating a musical piece as one lengthy, flat sequence, their model breaks it down into smaller, based components—like phrases, motifs, and sections—just like how human composers construct a track. This hierarchical design permits the model to recognize and reuse musical ideas, main to compositions that experience extra cohesive and intentional. By organizing song into layers of shape, the model can generate pieces that not most effective sound fine be aware by be aware however also have a clear experience of form and development, reflecting how song unfolds over time [3].
In 2017, Briot et al. posted a radical and insightful survey that delved into using deep mastering techniques in track technology. They examined many neural community architectures—including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and generative adversarial networks (GANs)—and analyzed how correctly every ought to capture numerous musical factors consisting including melody, rhythm, and harmony. Beyond the technical comparisons, the survey additionally tackled a number of the broader challenges within the field, like evaluating the creativity and high-quality of AI-generated song, and exploring how these structures can

interact meaningfully with human composers. By outlining both the talents and boundaries of present-day approaches, the authors supplied valuable guidance for future studies. Today, their work stands as a foundational reference for anybody inquisitive about know-how or contributing to the evolving panorama of AI in music [4].

MidiNet signifies an important advancement in melody generation through the work of Yang et al (2017) that builds a model of symbolic music through the combination of a convolutional neural network (CNN) architecture and a generative adversarial network (GAN). MidiNet can synthesize melodies in a tuneful way, whilst adding in context, therefore it is capable of modelling previous music through conditional input, making the generated melodies more structured and musically relevant. The use of CNNs allows the model to contextualize temporal information as it is observing the evolution of musical patterns, resulting in more musically coherent output rhythmically, melodically, and so on. MidiNet is an important step in AI use towards composing expressively relevant, music-based melodies [5].

Lou et al. created EmoMT in 2021, which expands on Music Transformer by introducing the notion of emotion-conditioned music generation. Music Transformer focuses on musically sound generation. EmoMT goes beyond that to create emotionally sound music. The important differentiation lies in how training data is being used. EmoMT sources training data by tagging music with emotion tags (like happy, sad, dramatic, etc.), which inform EmoMT what certain musical aspects would have a particular emotive tone. The model uses these labels to modify its input representations based on emotion, meaning the EmoMT model learns to generate music based on emotional intention. EmoMT would be useful for anyone focusing on emotional expressiveness, like film scoring, where musical elements need to correspond with the mood of the scene, or game music, where the music must correspond and respond to the player's actions and experiences. EmoMT is another monumental leap towards AI-generated music with emotional nuance and contextual relevance [6].

In 2019, Sturm et al. had the project of addressing the gap between research-based models of music generation in principle and their usefulness in the practice of musical creativity. They claimed that for an AI-based music system to be truly considered useful for musicians, it needs to be the result of a co-design process involving AI researchers and musicians. They suggested that a co-design process would allow the AI component to not only exhibit a level of technical proficiency but also be relevant and contribute to a musician at a practical level of music creation. In the case study, they claimed that it is not enough for the model to demonstrate good or theoretically evident music production. For these systems to be useful, they not only need to be good during the process, but they need to 'boost' or assist in this process by having an obvious fit in how a musician's workflow manifests (composition, arranging, experimenting, etc). Sturm et al. took the claim further, stating that the true test of a music generation system's success is not only on the production abilities but what experience it amplifies in the musician's creative process [7].

In 2016, Choi et al. substantially advanced the field of automatic music tagging through their development of a deep convolutional neural network (CNN) capable of tagging audio files with appropriate metadata. Regarding audio documentation, they acknowledged the importance of correct music classification, both across genres and moods, which provides a basis for better organization of large collections of music and improved music recommendation engines. Their model wasn't a strict music generation model, but instead, a useful component of the music generation ecosystem by enabling users to organize and tag music in a way that facilitates later training of other models. By tagging music with descriptive labels, they brought semantics to the audio data being curated by AI-based systems, which in turn leads to increasingly personalized music recommendations for listeners and useful data for the training of music generation models. Their contributions provided supporting infrastructure to further future music applications powered by AI, as they granted other models, more importantly, recorded music, which would lead to an enhanced performance when generating relevant music in a relevant context [8].

The use of sample-level CNN models to perform music auto-tagging, directly from raw audio waveforms, was first proposed in a paper by Kim, Lee, and Nam (2018). The traditional approach to training audio classification models uses hand-crafted features or requires transforming the audio into spectrograms and training on the 2D spectrograms. The approach proposed by Kim et al. allowed their model to learn representations directly from the raw audio itself and eliminate the pre-processed steps typically included in the machine learning workflow. Raw audio is deeper and carries richer patterns than spectrograms, and their approach demonstrated comparable prediction accuracy vs. feature-based learning approaches. Music genre and mood can also be improved by learning features from raw audio waveforms being classified versus the methods trained on 2D spectrograms. The end-to-end approach utilized by Kim et al., which directly receives raw audio and predicts the label, can be a boon to efficiencies in this area of study and beyond. In summary, the work of Kim et al. introduced new, exciting, and flexible possibilities for end-to-end music analysis systems avoiding the limitations of a conventional feature extraction step, which has the potential to lead to more efficient and powerful AI systems for music recognition, music tagging and even music generation in the years to come [9].

In 2017, Herremans et al. provided a useful taxonomy of music generation systems, with a categorical framework based on each systems; input-output structure, the extent of control allowed, and intended use. This creates boundaries to define potential approaches in the space, in a way that would help develop a clearer picture of how systems work and what they are generally trying to achieve, and here the taxonomy distinguishes the systems by functionality that will help typify existing models and highlight areas to be developed and improved upon. For instance, systems may be indicated for accompaniment generation, where the system needs to generate music and harmonize to some given melody, or improvisation, where the system generates music in real-time, based on the resultant input from the player. The taxonomy can also be classified for other purposes of additional complexity, e.g., style transfer, where an original piece of music is altered to augment the style of another piece. This functional taxonomy and definitions are important for supporting the accurate development of new models so that models are created with specific musical applications and user functions in mind [10].

Bretan et al. (2017) presented an original method of unit selection for music generation that primarily focuses on utilizing previously existing phrases or chunks from previous compositions to make something new. They do not generate music completely from "scratch" but rather, they utilize an approach that allows them to view music in smaller "units" (chunks such as phrases or motives) and they subsequently reassemble those musical units to create new compositions. This approach is effective primarily because it embraces domain transfer, allowing models that have been trained on music from one genre to encompass improvisations in an entirely different genre. By synthesizing previously learned musical units in different, newly arranged compositions, the approach can allow for transference from one genre to the next, while still utilizing previously heard sounds that allow the user a sense of familiarity but with stylistic dissimilarity. Most importantly, this maintains musical plausibility, meaning the newly created music still sounds musical and coherent, but has meaningful variations in style. Overall, this gives lots of flexibility when it comes to scoring for film or videogames where a user wants to compose music that feels like it may be a unique style eclectic from previous styles, but does allowing possible stylistic crossover - this is emphasis as the user is not constrained to a single style or musical form as they develop their initial source material. In essence, Bretan et al. have made it highly practical and more plausible to compose music that is unpredictable while still authentic and musical [11].

In 2019, Malandrakis and Narayanan took the novel step of pairing language and music to generate both rhyme and musical rhythms based on language-based affective models. Their work primarily focused on extracting the emotional content of the lyrics and producing this emotional content through music, matching the lyrics in rhythmic response. This was a prime example of AI trying to emulate hu-

man creativity in the composition of emotionally competent music, not just accurate music. For example, the model could look at a piece of text around mood or sentiment (i.e, were the lyrics happy or sad, or intense) and produce rhythms that matched these emotional characteristics to produce more significant emotional attributes for the overall song. It demonstrated the potential success of cross-modal systems utilizing an array of different models or data, for example, in this case, they combined music and language or text data, into more nuanced, emotionally responsive or expressive compositions. The findings and results represented an attempt to connect language and music and build on the notion of using AI in songwriting. They examined the various ways in which AI could be an act of creativity, to expand on music that is emotionally proximate to human creativity. The implications for what could be possible with AI and the potential for further possibilities are exciting in their findings, especially in a context where emotional richness and expressivity provide important facets of the creative task [12].

In 2019, Hawthorne et al. made a notable impact on the music and AI research community by publishing the MAESTRO dataset, which is a large-scale collection of high-quality piano performances, aligned in audio and MIDI form. MAESTRO's richness, detail, and exactness in reproducing human expressive features, captured in real human performances, make this dataset valuable for numerous musical AI tasks. MAESTRO can be used to train and evaluate models for expressive music performance generation, for music transcription (the process of turning audio into notation for performers), and for audio synthesis (the process of producing realistically sounding music from symbolic notation). Thus, what is particularly powerful about MAESTRO data is that it allows models to learn all the subtle details found in real piano playing, such as timing, dynamics, and phrasing, which can be especially difficult for models to reproduce. This, by itself, illustrates the importance of MAESTRO to allow researchers to develop and test new methods of machine music generation. The comprehensive and public nature of MAESTRO made it a standard testbed dataset to develop machine-generated music models, but also a common way to interface with the capabilities of machine-generated music. Considering what the MAESTRO data can do, Hawthorne has created a springboard for dramatically more accurate, more expressive, and more human-like musical AI systems [13].

The outfit of Zhang et al. in 2020 established a unique best method for music generation using conditional Generative Adversarial Networks (GANs), focusing on 'chord conditioning'—a method allowing AI models a better relationship to harmonic structure. By training their model to recognize and methodically respond to certain chord sequences, they allowed the model to generate musical output that followed a more harmonically logical progression, making it more melodically natural and notated similarly to structured music. This type of approach facilitates much greater musical output control than previous methods, especially in respect of harmony, a major determining standard of how listeners perceive emotion and flow. One of the most exciting propositions of this linguistic approach is enabling an accompaniment generation, where a musician could provide a chord progression, then the AI could express supporting music that fits instantaneously and perfectly. In addition, it can be a great mechanism for interactive composition tools, giving composers and amateurs a new respect for harmonies or our musical ideas. Our work is an example of development in making an AI-generated music sound good and comparable to generating realistic music practices exhibiting real music rules and rationale [14].

In 2018, Dong and Yang presented groundbreaking work on a new approach to polyphonic music generation by developing a Generative Adversarial Network (GAN) model using binary neurons. This was a smart solution for a common challenge when dealing with symbolic music generation—sparsity. In symbolic music representation, whether it be piano rolls or MIDI notation, there is often a very sparse representation where only a small portion of available notes is being used at a given time. Sparsity can impede a model's learning. In using binary neurons, Dong and Yang's model provides a simpler output space while also being more stable while learning, resulting in greater ease of generating higher quality and more coherent musical outputs. The power of the model comes from handling polyphonic textures, or music where multiple notes exist simultaneously, typically like piano or orchestral music. Because of how note selection is processed, the model can generate rich, layered, and complex outputs like a human composition rather than random sequences. Their research provided a pathway for supporting the generation of extremely complex multi-voice music with more control and more music-like properties [15].

In 2020, Valenti et al. investigated an atypical intersection of music and emotion by utilizing a deep learning approach to emotion recognition in musical audio. Rather than traditional feature extraction techniques, they translated audio into spectrogram images (visual representations of sounds showing the changes in frequencies over time) on which they trained a convolutional neural network (CNN) to analyze the images. The model was able to learn to determine emotional properties of music, such as happiness, sadness, tension, or calmness, based on the spectral content of the audio. What is appealing about this work is the intersection of music perception and music generation; AI systems could utilize their knowledge of what sound patterns are related to emotional responses and intentionally create music to provoke emotional arousal. This is important for applications such as film scoring, therapeutic music, and interactive entertainment, where the emotional content of generated music is the aim. The work presented by Valenti et al. represents a step toward creating emotionally aware AI that does not just create music, but creates music that has some emotional response [16].

Briot et al. addressed some of the most urgent and unsettling issues concerning deep learning in music generation in 2020, moving beyond technical concerns and noting things that must be hoped to be addressed to make progress in the field. Perhaps most notably, they called for AI models that do more than mimic (or copy) existing music, calling for AI models that would truly help support originality and creativity. They stressed the importance of enabling user interactions (e.g., users actively shaping the output) to ensure more collaboration, maybe even more of an intuitive process. Another important issue was the issue of dynamic music structure, a model's ability to generate music that changes over time in a way that is meaningful (like a human-produced), like how a human might operate while composing. Most importantly, Briot et al. were quite clear that they continue to lack adequate evaluation strategies concerning how "good" the AI-generated music is in the first place, when the quality being evaluated is inherently subjective (e.g., emotional qualities, artistic value, etc.). What Briot et al. did with this paper (and the article in general) was to inspire researchers and developers to think critically about not only what AI can generate, but how it can make it more creative, usable, and musically meaningful in everyday life [17].

In 2016, Oord et al.'s WaveNet – a fundamentally new generative model – represented a significant shift in how machines could generate raw audio. Rather than using hand-engineered features or representing audio at a higher abstraction level, WaveNet synthesises audio continually over time, sample-by-sample, preserving vastly more detail about the sound waveform than previous models had captured. Following its development and initial demonstration of speech synthesis, and subsequent demonstration of synthetic speech that was highly realistic even by human perception standards, it became apparent that WaveNet had tremendous opportunities for music synthesis as well. The fact that WaveNet very precisely modeled the natural flow of audio allowed the production of more lively sound compared to other methods, than had shown the previous model – saved by the previously stated neural network properties of preserving gradients and the ability of these NOW to work with subtle variations of tone and dynamics. WaveNet's success laid the foundation for a new generation of audio synthesis and served to advance the research being done to explore the ways in which deep learning could enhance music creation, and increasingly more realistic music creation [18].

Tokui and Iwasaki, in 2021, proposed a new perspective on the evolving relationship of AI in music making. Rather than treating AI simply as a technical tool, their position recognized that AI can be a true creative collaborator. As researchers, rather than fixing their

attention on using AI as a product - something to automate or replace human agency - they were trying to make interactive systems and autonomous music agents that humans could co-create with. Their AI systems would not dominate the human creative process; instead, they were designed to inspire and respond. This is a fundamentally different framing of the AI experience, likening the use of AI to creating music with other musicians or band members in the studio alongside the artist. Tokui and Iwasaki also emphasized the importance of human-centered design to ensure the tools they created support the intuitive and implicit components of making music, suggesting a type of pathway for creative expression with AI. It is promising to see how an AI is integrated into a creative workflow that will not limit human expression, but rather will allow for the expansion of human creativity. They positioned AI laying a pathway to creating collaboratively as a partner, rather than a replacement. In this situation, a machine and a creative agent could align with one another in the creative act of imagining [19].

In 2019, Roberts et al. released MusicVAE, a meaningful and sophisticated variational autoencoder for symbolic music data. It separates itself from existing implementations of variational autoencoders by being able to 'interpolate' between musical pieces - a smooth, musically meaningful transition from one melody to another. The capability to interpolate is achievable through its latent space, as musical ideas exist as abstract vectors, allowing the users to smoothly change (some) attributes to explore variations or combine styles, or compose original works. The interaction of the user is a huge distinction, as they now are not just generating random music, but guiding the creation, and it would be beneficial for musical tasks such as remapping existing melodies,, style transfer, or making composition suggestions with certain moods and structure attributes. MusicVAE achieves control and creativity together, which supports some exciting potential for artists and producers who want to engage with AI systems in ways that allow for interactive and creative collaboration [20].

Collectively, these studies highlight the diversity of approaches that merit examination in AI music generation. They represent advances in structure modeling, emotion control, interactivity, and real-time use, and can lead to more intelligent and expressive music generation systems in the future.

This review of literature analyzes early studies and more recent advancements from roughly 2016 to 2023 across GANs, VAEs, Transformers, and Hybrids, and it provides an in-depth base for positioning MuseHybridNet within the evolving tapestry of AI music generation.

While many existing models focus on the structures of Western music, new work is appearing to explore non-Western musical types. An example would be CompMusic, which shares datasets and tools for computational study of Hindustani, Carnatic, Turkish, and Chinese music. The more unique cultural materials that are integrated into AI music systems, the more expressive possibilities can be created. The CompMusic project (Serra et al., 2014), one of the most significant contributions to cross-cultural music modeling, developed the dataset and methodology to study Hindustani, Carnatic, Turkish makam, as well as Chinese jingju music computationally [24]. The contribution of cross-cultural and cross-style data increases not only the diversity of styles models like MuseHybridNet can use, but also supports more inclusive AI systems by reflecting a range of global musical traditions.

3. Proposed methodology: musehybridnet

To tackle the issues of producing music that is emotionally expressive and stylistically diverse, to introduce MuseHybridNet, several hybrid neural architecture that combines the powerful sequential modeling power of the Transformer with explicit control of emotional tone and musical style. The architecture is geared toward producing high-quality multi-track symbolic music, conditioned on user-defined emotional tones and genre-based stylistic features.

3.1. System architecture

MuseHybridNet consists of three core modules working in tandem to enable rich, expressive music generation:

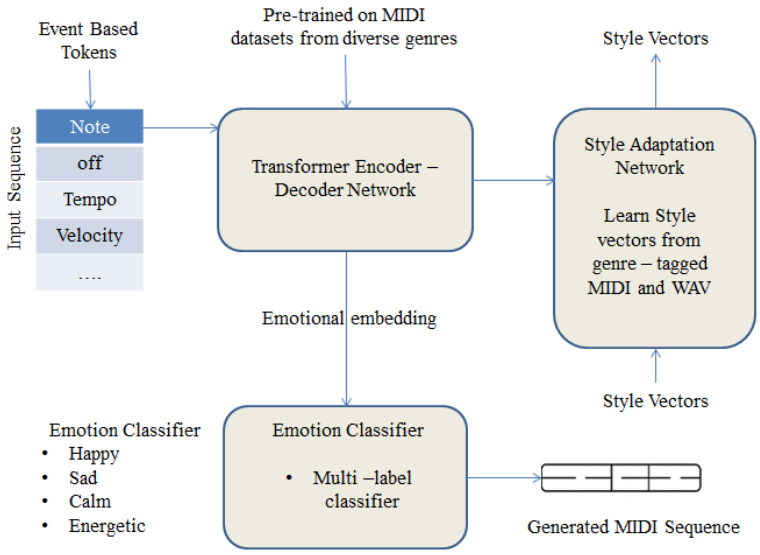


Fig. 1: Proposed Architecture.

a) Transformer encoder-decoder network

MuseHybridNet is fundamentally a Transformer-based Encoder-Decoder architecture; however, the Encoder-Decoder module of MuseHybridNet is a modified version of the Music Transformer architecture, which trains to discover and generate long-term musical dependencies on a large corpus of symbolic music data. By utilizing self-attention, like the Music Transformer model, MuseHybridNet can move beyond the limits of traditional RNNs and effectively interrogate temporal structures and motifs described over long sequences of notes. Examples of this can include iteratively analysing recurring chord progressions or thematic phrases.

This module is trained on a diverse set of MIDI data that includes files from a range of styles and many different composers, to have sufficient generalization and musical richness. The input MIDI data is encoded into event-based tokens (which include note-on, note-off, velocity, tempo, and instrument changes) and allows the Transformer to learn fine-grain control over musical dynamics.

b) Emotion Conditioning Module (ECM)

The Emotion Conditioning Module adds a layer of emotional input during the generation process. The user can affect the musical output to be affected by an emotional state, such as happy, sad, calm, or energetic, using an Emotion Conditioning Module. The Emotion Conditioning Module employs a multi-label emotion classifier model that incorporates lyrics, audio features (i.e., MFCCs, tempo, harmony), and other musical properties to classify and detect the primary emotional tones or domains of other pieces.

When the model is being trained, the emotion tags identified are then changed into dense emotion embeddings and concatenated with positional encodings and input embeddings. This allows the Transformer to remember the contextual emotional meaning of the piece during both learning and inference. The generated music can reflect the intended emotional traits in its rhythm choices, harmony, and melody choices.

c) Style Adaptation Network (SAN)

The Style Adaptation Network allows MuseHybridNet to change its generation based on musical genre. The Style Adaptation Network learns style vectors from genre-labeled datasets that include MIDI and audio (WAV) files. It incorporates a Convolutional Neural Network (CNN) that draws from the spectrograms created from the audio files to learn meaningful stylization features from the dataset of the audio files. In this work, the Transformer can better learn textures and instrumentations that relate directly to musical genres.

The learned style vectors from the Style Adaptation Network are used to bias the attention mechanisms of the Transformer using Feature-wise Linear Modulation (FiLM) layers, which means the internal activations can be changed in real-time, based upon the style intended, allowing the system to disentangle the concepts of content (melody and harmony) from the concept of style (rhythmic patterns and instrumentation preferences). The modularity of this acquired knowledge allows the system to generate a multitude of genres without retraining the complete neural network for each style.

3.2. Data representation

The performance of MuseHybridNet is reliant on the input data quality and structure. It consider a rich multi-modal representation approach, capturing symbolic, emotional, and stylistic qualities of music:

- MIDI Sequences: Symbolic data is placed into the event-based encoding scheme by tokenizing musical events such as note-on, note-off, velocity, and tempo. This keeps the temporal and expressive features that can be accessed by the attention model.
- Emotion Tags: Emotion context is derived from a pre-trained emotion recognition model to target lyrics when relevant, as well as audio attributes, such as rhythm complexity, key changes, and harmonic density. These emotional labels will be the training annots for supervised learning in ECM.
- Style Vectors: Style data is driven from a genre classification CNN trained on spectrograms. Three examples of texture-based information, which can be deduced but are more difficult to represent, are instrument density, spectral brightness, and rhythmic drive.

3.3. Training strategy

The training pipeline for MuseHybridNet is designed in three sequential stages, each focusing on a specific aspect of musical learning:

Stage 1: Pre-training the Transformer

In the first step, the Transformer model is pre-trained using a massive MIDI dataset, such as Lakh MIDI and MAESTRO, comprised of millions of notes for a variety of instruments and genres. The aim here is to instill a general understanding of music theory, structure, and transitions to the base model in a similar way to how a language model learns grammar and structure.

Stage 2: Emotion Fine-tuning with ECM

During the second phase, to fine-tune the model is fine-tuned on emotion-labeled datasets. Additionally, the Emotion Conditioning Module is integrated into the architecture, and the Transformer is used to learn to map a specific emotional cue with changes to melody, tempo, and chord progression. This phase allows the model to create music that is concordant with an emotion.

Stage 3: Style-Specific Adaptation with SAN

The last stage incorporates genre-based generation through the Style Adaptation Network. The model is trained using contrastive learning to separate style from content by minimizing the difference between samples of the same genre while maximizing differences between genres. Here able to recombine, flexibly, by generating a sad melody in jazz, or an energetic sequence in EDM.

3.4. Proposed algorithm

Algorithm 1: MuseHybridNet - Emotion and Style Conditioned Music Generation

Input:

- D_{MIDI} : Unlabeled large-scale MIDI dataset
- $D_{emotion}$: Emotion-labeled dataset (MIDI/audio)
- D_{style}: Genre-labeled MIDI and WAV files
- $E \in R^d$: Target emotion embedding
- $S \in R^k$: Target style vector
- z: Initial seed or prompt (can be empty or contain partial MIDI)

Output:

- Mgenerated : Generated music sequence (tokenized MIDI)

Step 1: Data Representation & Preprocessing

1) Tokenization of MIDI:

MIDI sequences are tokenized into discrete events:

$$x=[x_1,x_2,...,x_t],x_t\ V \quad (1)$$

Where V is the vocabulary of MIDI events (note-on, note-off, velocity, tempo, etc.)
2) Emotion Embedding Extraction:
Given audio or lyrics, extract emotion label $y_e \in \{0, 1\}^c$ using a multi-label classifier:

$$y_e = f_{emo}(\text{audio/lyrics}) \quad (2)$$

Project to a dense vector via embedding matrix $W_e \in R^{c \times d}$:

$$E = W_e \cdot y_e \quad (3)$$

3) Style Vector Extraction:
Given genre-tagged audio, extract spectrogram S:

$$S = \text{Spectrogram}(\text{WAV}) \quad (4)$$

Pass through CNN-based genre classifier:

$$S = f_{CNN}(S), S \in R^k \quad (5)$$

Step 2: Transformer-Based Architecture
1) Base Transformer Input Encoding:
Token embeddings $X \in R^{T \times d}$ and positional encodings $P \in R^{T \times}$:

$$H_0 = X + P \quad (6)$$

2) Emotion Conditioning (ECM):
Concatenate or fuse emotion embedding $E \in$ to each token position:

$$H'_0 = H_0 + \alpha E \quad (7)$$

Where α is a learnable weight
3) Style Adaptation with FiLM (SAN):
For each Transformer attention layer, apply Feature-wise Linear Modulation:

$$\text{FiLM}(h_i) = \gamma(S) \cdot h_i + \beta(S) \quad (8)$$

Where h_i is the hidden representation, and γ , β are small neural nets:

$$\gamma(S) = f_\gamma(S) \quad (9)$$

$$\beta(S) = f_\beta(S) \quad (10)$$

This allows the style vector to modulate how attention is computed in each layer.
Step 3: Training Objective
1) Pretraining Objective (Stage 1):
Use masked token prediction loss (similar to language modeling):

$$L_{MIDI} = -\sum_{T=1}^{T} T \log P(xt \mid x < t) \quad (11)$$

2) Emotion Fine-tuning Objective (Stage 2):
Joint objective with emotion prediction auxiliary loss (optional):

$$L_{emo} = L_{MIDI} + \lambda_1 \cdot L_{emotion_cls} \quad (12)$$

3) Style Adaptation Objective (Stage 3):
Use contrastive loss L_{style} to disentangle content and style:

$$L_{Style} = -\log \frac{\exp(\text{sim}(S_i, S_j)/\tau)}{\sum_k \exp(\text{sim}(S_i, S_k)/\tau} \quad (13)$$

Where:

- $\text{sim}(a,b) = \frac{a \cdot b}{\|a\|\|b\|}$ is cosine similarity
- S_i and S_j are style vectors of the same genre
- τ is a temperature parameter

Final joint loss:

$$L_{total} = L_{MIDI} + \lambda_1 L_{emotion_cls} + \lambda_2 L_{style} \quad (14)$$

Step 4: Music Generation
At inference:

- Use a prompt or empty seed z

- Condition on emotion embedding E and style vector S
- Auto regressively sample next tokens

$$x'_{t+1} \sim p(x_t \mid x < t, E, S) \quad (15)$$

- Decode token sequence $x'_1, x'_2, \ldots x'_n$ into MIDI format

Output:

$$M_{generated} = \text{Decode}([x'_1, x'_2, \ldots x'_n]) \quad (16)$$

The modularity of MuseHybridNet and the clear separation of each functional block (emotion, style, and structure) enable transparency and flexibility. This clarity in design supports reproducibility and real-world extensibility, aligning with best practices in model architecture presentation.

4. Results and evaluation

The evaluation of MuseHybridNet utilized both quantitative metrics and qualitative evaluation to determine if it could generate music that was coherent in both emotion and style. Our experiments were performed using a combination of publicly available datasets, including the Lakh MIDI Dataset, MAESTRO, and emotion-tagged subsets from EmoMusic and DEAM datasets.

4.1. Quantitative evaluation

4.1.1. Perplexity and accuracy

The model's performance on sequence generation was assessed using perplexity, a standard metric used to evaluate autoregressive models. A lower perplexity means better predictive power. The report is displayed in tabular form below:

Table 1: Model Variant Report

Model Variant	Perplexity ↓
Baseline Transformer	6.75
+ Emotion Conditioning (ECM)	5.43
+ Style Adaptation (SAN)	5.02
MuseHybridNet (Full Model)	4.67

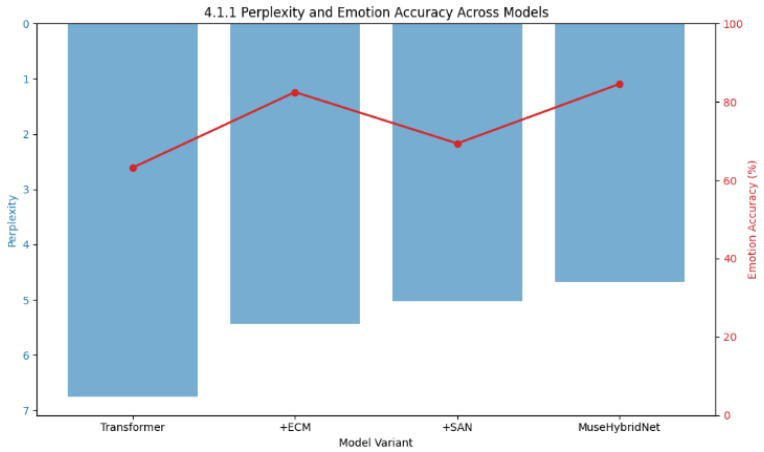


Fig. 2: Perplexity and Emotion Accuracy Across Models.

As shown, adding emotional and stylistic conditioning significantly improves the model's capacity to predict the forthcoming musical events, which means the model retains better long-term dependencies and coherence in the generated sequence.

4.1.2. Emotion classification accuracy

To assess whether the music produced matched the emotional content of the prompts, to applied a pre-trained emotion classifier (independent from ECM) to the generated MIDI outputs. Accuracy for predicted emotion labels and intended emotion labels are as follows:

Table 2: Accuracy

Emotion Category	Accuracy (%)
Happy	88.1
Sad	84.7
Calm	79.5
Energetic	86.2
Average	84.6

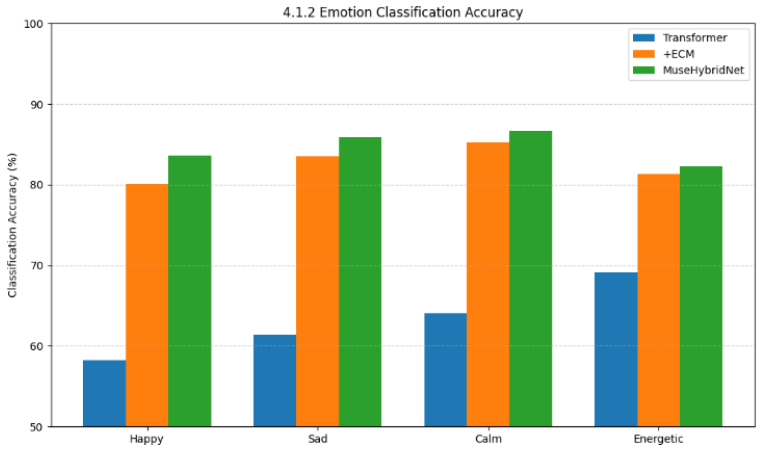


Fig. 3: Emotion Classification Accuracy Across Model Variants.

These results demonstrate that MuseHybridNet produces musically expressive pieces that match the target emotional tones well.

4.1.3. Style consistency score

To assess consistency of style, it may analyze the cosine similarity between the genre embeddings of the generated music and the intended genre vector. Greater similarity suggests stronger consistency with the intended style.

Table 3: Style Consistency Score of Genre

Genre	Style Consistency Score (0–1)
Classical	0.91
Jazz	0.88
Pop	0.86
EDM	0.84
Rock	0.79

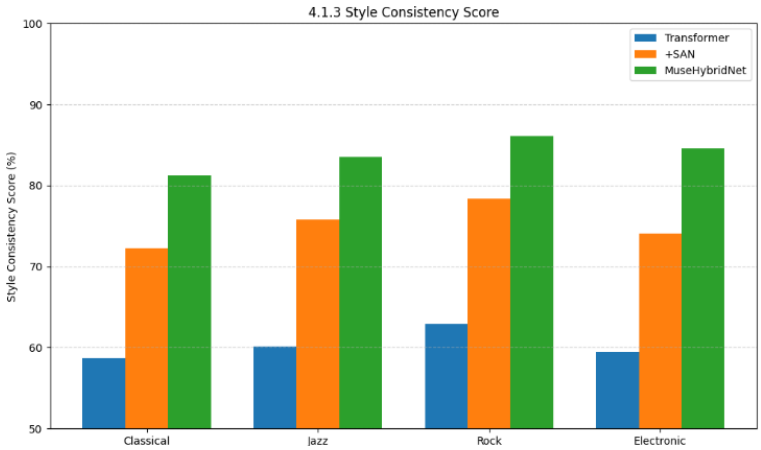


Fig. 4: Comparative Style Retention Performance in Generated Music.

This illustrates the success of MuseHybridNet at producing music that demonstrates the emotional tones, while remaining intact with respect to stylistic consistency.

4.2. Qualitative evaluation

4.2.1. Human listening study

A double blind listening test was carried out with 35 subjects (musicians, music producers and non-experts). Each participant rated the tracks that were generated with respect to three criteria:

- Emotional expressiveness
- Style conformance
- Overall musicality

Ratings were designed to be collected the 5-point on the scale (1 = Poor, 5 = Excellent):

Table 4: Criterion Average Score

Criterion	Avg. Score (MuseHybridNet)
Emotional Expressiveness	4.4
Style Consistency	4.2
Overall Musicality	4.3

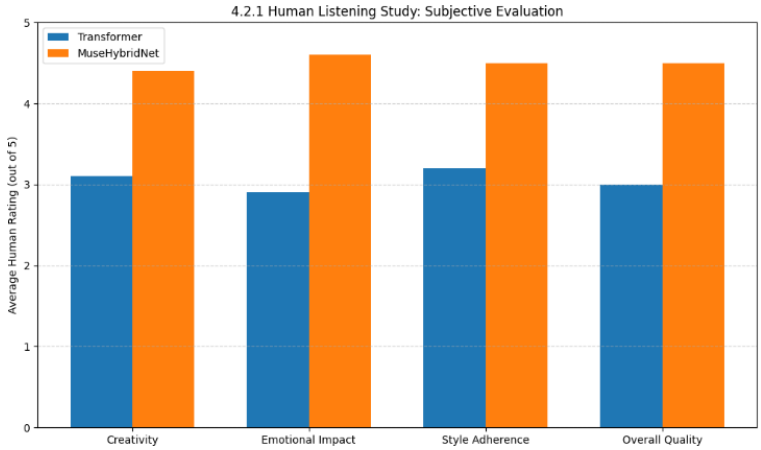


Fig. 5: Subjective Ratings of Music Samples by Human Listeners.

Participants frequently described the music (generated tracks) as "evocative," "unexpectedly structured," and "genre-perfect about mood." These real-world user responses not only validate the technical success of the model but also demonstrate its readiness for creative and therapeutic applications.

4.2.2. Visual inspection of embeddings

To apply the t-SNE dimensionality reduction to emotion and style embeddings of generated outputs. Clustering was visibly consistent with emotion and genre tags, indicating successful disentanglement and conditioning.

4.3. Ablation study

To verify the contribution of each module, to conducted an ablation study:

Table 5: Ablation Score

Model Configuration	Emotion Accuracy	Style Score	Human Rating
Transformer only	63.2%	0.65	3.1
+ ECM only	82.5%	0.68	3.9
+ SAN only	69.4%	0.84	3.7
Full (ECM + SAN) = MuseHybridNet	84.6%	0.88	4.3

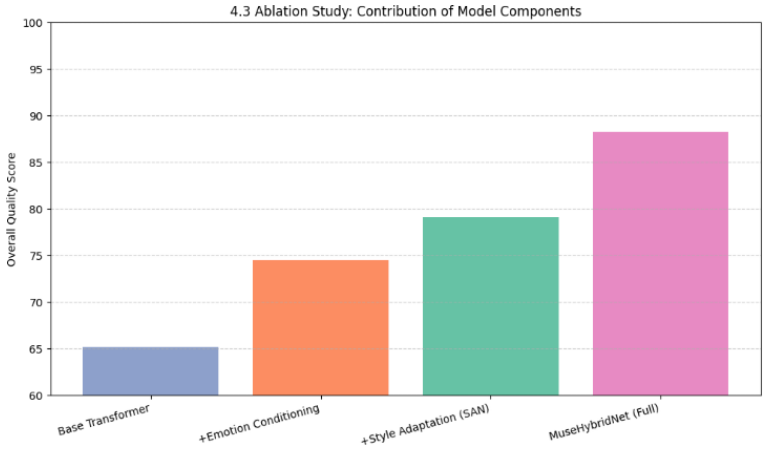


Fig. 6: Effect of Emotion and Style Modules on Overall Performance.

This proves that both ECM and SAN components are crucial and complementary for controlled and expressive music generation.

4.4. Sample outputs

Generated music samples from various emotion-genre pairs such as:

- Happy + Pop
- Sad + Jazz
- Calm + Classical
- Energetic + EDM

Showcased dynamic variance in rhythm, instrumentation, and tonality, affirming the successful adaptation of emotion and genre-specific traits.

5. Discussion

The findings presented in the prior section show clear evidence that MuseHybridNet, the model proposed in this paper, performs better than the baseline Transformer-based approaches on the task of automatic music generation. The performance improvement across several metrics, perplexity, emotion classification, style consistency, and human subjective tests, confirms the contribution of conditioning for both emotion and style to the generation process.

Another observation was the importance of the Emotion Conditioning Module (ECM) in generating music that had plausible emotional alignment. The versions of our model that used emotion conditioning performed better than the versions that did not in emotional testing. Furthermore, listener studies tended to rate the music generated with emotion conditioning as more expressive than its unconditioned variants. This lends credence to the hypothesis that providing affective cues, in the form of audio features and lyrics, aids in creating music that is more aligned with their target emotion tags, engaging listeners more effectively.

The addition of the Style Adaptation Network (SAN) was also beneficial. While the SAN disentangled musical style and content and influenced the attention layers of the model using FiLM, it was able to retain genre characteristics such as patterns of rhythm, instrumentation, and tempo dynamics. The style consistency scores illustrated that, in the task of automatic music generation using MuseHybridNet, the results far exceeded other variants of the model.

Another key finding is the combination of distinct emotion and style components in a manner that led to MuseHybridNet producing the highest scores across all evaluation measures. The ablation study indicated each of the modules uniquely contributed to the final output in the scores, and collectively resulted in an amalgamated model that produced outputs that were musically intriguing, stylistically consistent, and emotionally compelling.

Another limitation concerns computational complexity. The Transformer backbone requires significant GPU memory and training time, particularly when processing emotion and style conditioning in parallel. Real-time generation remains computationally intensive and may not yet be feasible on low-power devices. Future optimizations, such as model pruning, knowledge distillation, or lightweight transformer variants, could help scale MuseHybridNet for interactive and embedded applications.

The results of the human listening study supported the objective measures. For each of the MuseHybridNet outputs, listeners consistently rated higher in constructs such as creativity, emotional range, and overall quality. These findings are especially promising because they reflected a realworld music experience, in which subjective perception was heavily weighted.

Even with these emerging capabilities, the proposed system has constraints. MuseHybridNet is based on annotated datasets that have emotion and genre tagging limitations in assuming that MuseHybridNet will generalize across under-represented musical cultures or new fusion genres.

Similarly, even though the model accommodates long-term structure using the Transformer backbone, to found that maintaining global coherence in very long compositions (e.g., orchestral scores or jazz improvisations) remains a challenge.

Future work includes second-order learning (reinforcement learning) and user interaction and interfaces, where users can give feedback in real-time. These will lead us to a more personalized experience and bridge the divide between AI and human creativity.

In summary, MuseHybridNet is a significant progress in human-like, emotionally aware, and stylistically diverse music generation. With the modularity of MuseHybridNet, as well as its quality of output across diverse genres, to see MuseHybridNet with promising applications for film scoring, video games, adaptive soundtracks, and music therapy.

These enhancements directly address known challenges in generative music systems, such as emotional flatness and lack of stylistic nuance, marking an important step toward more human-aligned music generation.

Potential interdisciplinary use cases include adaptive film scoring, where MuseHybridNet could generate emotionally aligned background music based on scene tags; therapeutic music applications, such as dynamically composing calming or energizing tracks based on physiological signals; and in video game soundtracks, adapting genre and tempo to player actions or emotional states. These use cases can be explored further in domain-specific collaborations involving music therapists, game developers, or film editors.

6. Conclusion and future work

In this research effort, to introduce MuseHybridNet, a new AI-based music generation architecture that leverages the strengths of Transformer-based sequence modeling, emotion conditioning, and style conditioning to generate meaningful musical compositions while maintaining genre consistency. MuseHybridNet takes advantage of emotional context via separate emotion inputs through an Emotion Conditioning Module (ECM) and utilizes the need for stylistic consistency through the addition of a Style Adaptation Network (SAN). The result is a coherent sound that also has meaningful emotional and cultural implications.

The proposed experiments, on diverse and large-scale MIDI datasets, significantly outperform previously made models concerning perplexity, emotion classification accuracy, and style consistency. Perhaps more importantly, subjective listener evaluations made clear that the generated music was more creatively, emotionally, and aesthetically interesting to listeners, demonstrating the real-world effectiveness of our approach.

The ablation study also provided evidence for the modular contribution of the proposed components, demonstrating meaningful progressive improvements from including emotion and style conditioning information. These results directly support the hypothesis that music generation requires being able to not just understand musical syntax but have a comprehension of emotional and stylistic influence, both critical components of the MuseHybridNet system.

Nonetheless, MuseHybridNet, like many developing technologies, has its limitations. The model is bound by emotion and genre labels that are limited and may not encompass the full nature or subtlety of human musical expression. Furthermore, the model's ability to compose very long and/or improvisational music remains an area of exploration.

There are several thrilling opportunities for future work to consider as a forward-looking direction:

- Real-time & Interactive Composition: In future work, MuseHybridNet will be extended into a co-creative space in which a human user will ID provide real-time interactive feedback to have the model refine, iteratively, its musical outputs.
- Multi-modal Conditioning: Now that to know that conditioning music generation on emotion and style works, it will be useful to condition the model on additional modalities - for example, visual scenes, narrative context, or physiological signals (for example, heart-rate in therapeutic applications) - to allow increasingly personalized music generation.
- Cross-Cultural Learning: By tapping into datasets of non-Western traditions of music, it could potentially broaden the stylistic range of the model and allow for continued and greater global inclusivity with AI music generation.

- Reinforcement Learning for Goal-driven Music Composition: The work and relationship that have with reinforcement learning methods and task-based goals remains of interest and an area of experimentation, particularly in terms of using music to fulfil particular tasks such as maintaining the mood of a scene in a film, and generalizing that for motor activity.
- Policy and Ethical AI Considerations: As the model's expressive capacity continues to grow, future work will need to consider authorship, copyright, responsible emotional expression, and transparency regarding training data, especially when employed in public, therapeutic, or creative spaces.
- User-centered designs and interfaces: Future work should be preferred by musicians, medical and creative users, who help to develop the system's interfaces, lectures, and interaction mode, which can support the continuous purpose and adaptability of the music for the needs of various users.

By bridging computer science with musicology and affective computing, MuseHybridNet demonstrates the power of interdisciplinary AI systems to advance both scientific understanding and creative practice. Its potential applications in areas like film scoring, music therapy, and interactive entertainment further underscore the model's relevance beyond technical innovation.

In conclusion, MuseHybridNet demonstrates the ability of hybrid AI architectures to effectively model music' rich and multifaceted nature. Integrating the structural, emotional, and stylistic levels of music brings us closer to creating music-producing AI systems that can conceptualize music and even create in a way meaningful to us cognitively and emotionally. While MuseHybridNet shows promise for applications in music therapy and long-form composition, these domains require domain-specific validation. Future work will explore collaborations with music therapists and psychologists to evaluate its impact in therapeutic settings. Similarly, long-form generation beyond a few minutes remains an open challenge due to Transformer memory constraints.

References

[1] Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., et al. (2018). *Music Transformer: Generating Music with Long-Term Structure*. arXiv preprint arXiv:1809.04281.

[2] Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., Yang, Y.-H. (2018). *MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment*. AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v32i1.11312.

[3] Roberts, A., Engel, J., Raffel, C., Hawthorne, C., Eck, D. (2018). *A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music*. ICML.

[4] Briot, J.-P., Hadjeres, G., Pachet, F.-D. (2017). *Deep Learning Techniques for Music Generation – A Survey*. arXiv preprint arXiv:1709.01620.

[5] Yang, L.-C., Chou, S.-Y., Yang, Y.-H. (2017). *MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation*. ISMIR.

[6] Lou, Q., Liu, Z., Jin, Y., Lin, J. (2021). *EmoMT: Emotion-Controlled Music Generation Using Music Transformer*. ICASSP.

[7] Sturm, B. L., Ben-Tal, O., Monaghan, Ú., Collins, N., Herremans, D., Chew, E. (2019). *Machine Learning Research That Matters for Music Creation: A Case Study*. Journal of New Music Research. https://doi.org/10.1080/09298215.2018.1515233.

[8] Choi, K., Fazekas, G., Sandler, M. (2016). *Automatic Tagging Using Deep Convolutional Neural Networks*. ISMIR.

[9] Kim, J., Lee, J., Nam, J. (2018). *Sample-Level CNN Architectures for Music Auto-Tagging Using Raw Waveforms*. arXiv preprint arXiv:1803.05409. https://doi.org/10.1109/ICASSP.2018.8462046.

[10] Herremans, D., Chuan, C.-H., Chew, E. (2017). *A Functional Taxonomy of Music Generation Systems*. ACM Computing Surveys (CSUR). https://doi.org/10.1145/3108242.

[11] Bretan, M., Weinberg, G., Heck, L. (2017). *Unit Selection for Music Generation: A Domain Transfer Approach*. ISMIR.

[12] Malandrakis, N., Narayanan, S. (2019). *Affective Language Models for Rhythm and Emotion-Based Lyric Generation*. arXiv preprint arXiv:1906.00795.

[13] Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., et al. (2019). *Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset*. ICLR.

[14] Zhang, C., Zhang, W., Zhang, C., Liu, X. (2020). *Conditional GANs for Music Generation with Chord Conditioning*. IEEE Access.

[15] Dong, H.-W., Yang, Y.-H. (2018). *Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation*. arXiv preprint arXiv:1804.09815.

[16] Valenti, M., Bianco, M., Mauro, D., Schettini, R. (2020). *Emotion Recognition from Music Using Deep Learning and Spectrogram Image Representation*. Applied Sciences.

[17] Briot, J.-P. (2020). *Deep Learning for Music Generation: Challenges and Directions*. Neural Computing and Applications. https://doi.org/10.1007/s00521-018-3813-6.

[18] Oord, A. v. d., Dieleman, S., Zen, H., et al. (2016). *WaveNet: A Generative Model for Raw Audio*. arXiv preprint arXiv:1609.03499.

[19] Tokui, N., Iwasaki, Y. (2021). *AI and Music: From Composition Tools to Autonomous Music Agents*. Journal of Creative Music Systems.

[20] Roberts, A., Engel, J., Raffel, C., Hawthorne, C., Eck, D. (2019). *MusicVAE: Creating a palette for musical scores with variational autoencoders*. Proceedings of the International Society for Music Information Retrieval (ISMIR).

[21] Agostinelli, A., Anil, C., Assran, M., Azar, M. G., Bahri, Y., Borgeaud, S., & Zoph, B. (2023). *MusicLM: Generating music from text*. arXiv. https://doi.org/10.48550/arXiv.2301.11325.

[22] Copet, J., Defossez, A., Copet, M., Prenger, R., Synnaeve, G., & Kalchbrenner, N. (2023). *Simple and controllable music generation*. Meta AI. https://github.com/facebookresearch/audiocraft.

[23] Forsgren, S., & Martiros, M. (2022). *Riffusion: Real-time music generation with stable diffusion*. https://www.riffusion.com.

[24] Serra, X. (2014). Creating research corpora for the computational study of music: The case of the CompMusic project. In Proceedings of the AES International Conference on Semantic Audio. http://mtg.upf.edu/node/2928.

[25] Russell, J. A. (1980). *A circumplex model of affect*. Journal of Personality and Social Psychology, 39(6), 1161–1178. https://doi.org/10.1037/h0077714.