Survey of Different AI Models for Music and Lyrics Generation

Dhruthi N Bharadwaj[1], Vaishnavi S[2], Vidyashree C[3], Dr. Laxmi V[4]

[1,2,3] B.E Information Science Engineering, BNM Institute of Technology , Bengaluru, India

[4]Associate Professor, Information Science Engineering, BNM Institute of Technology, Bengaluru, India

Abstract: **The fast development of artificial intelligence (AI) has greatly impacted creative industries, such as music and lyrical composition. This paper provides an extensive comparison of different state-of-the-art AI models developed for music and lyrics generation. The study explores the architecture, capabilities, input-output mechanisms, and creative potential of models such as OpenAI's GPT-3.5, Meta's MusicGen, OpenAI's Jukebox, Google's Magenta (MusicVAE and NSynth), MuseNet, and other industry-relevant tools like AIVA and Amper Music. By analyzing these models across several criteria including quality of generated output, genre adaptability, user control, and real-world applicability, the paper aims to highlight the strengths and limitations of each system. In addition, it speaks to pressing issues like poor emotional depth, biases in the data, issues of copyright infringement, and technical intricacy involved in fusing meaningful lyrics and harmonized sound. The conclusion finds potential research and innovation tracks in future research and development for AI-generated music, highlighting that such technologies are likely to improve human creativity but revolutionize the music industry as well.**

Keywords: **AI-generated Music, Lyrics Generation, Creative Automation, Music composition models**

I. INTRODUCTION

Artificial Intelligence (AI) has travelled a great distance from its early days of developing logic, pattern recognition, and automation to discovering significant applications in creative arts previously the domain of human imagination and intuition. Among these, the domain of music and lyrics generation has proven to be an interesting area of research and development, bridging computational algorithms with creative imagination. It is now possible for AI systems to create melodies, lyrics, and even entire pieces of music that capture emotional resonance, stylistic nuance, and genre conventions. This is a far cry from the statistical and rule-based music generation of yesteryear to the deep learning-based generative models of today. These models are not simply generating content that appears and feels like human-created work, but in most cases, are opening up new possibilities for creativity by generating music and lyrics that are innovative, genre-bending, or stylistically hybrid. The increasing demand for AI-made music and lyrics is visible across entertainment, advertisement, education, and content generation segments. DIY artists leverage such technology as art allies to decorate, transform, and even innovate new pieces. Cinematographers and game designers also utilize AI as tools for fashioning adjustable music for differing sets of scenarios and atmospheres. Social influencers could create distinct tracks and lyrics immediately according to content and individualistic choice. This application, made with advanced AI models has empowered the playing field in music production, making it easy and accessible for non-musicians by challenging the old theories of authorship and originality in the digital era. Due to the increasing trend of artificial intelligence, different AI models and tools have been developed. For lyric composition, the most sophisticated among transformer-based language models are OpenAI's GPT-3.5 and GPT-4, which are being trained on large sets of text and have already demonstrated the ability to generate coherent, creative, and style-suited lyrics from the provided user prompt. These models are aware of rhyme schemes, emotional undertones, and word choices depending on genres so that users can produce tailored lyrical content with excellent fluency. In music composition, Meta's MusicGen has advanced significantly by converting natural language inputs into good quality instrumental pieces, providing control over genre, instrumentation, and tempo. MusicGen utilizes encoded text and audio materials representations, which are combined into a neural process, which produces the desired musical output based on the prompts. OpenAI's Jukebox is another breakthrough, which is a neural network that not only creates instrumental works but also creates full songs with lyrical and vocal elements.

The Magenta project, led by Google, is an important advancement in this area, offering a suite of models including MusicVAE, intended to generate and interpolate musical phrases—and NSynth, to create new sounds using a neural network that has learned from instrument samples. Both models combine elements of music theory with generative modeling, and it's straightforward to musically manipulate and blend melodies and timbres. MuseNet, also an OpenAI project, uses multi-instrumental music generation methodology and has proven to generate four-minute pieces of music with a rich variety of instruments in over a dozen styles, thanks to its advanced deep neural structure. Beyond such research tools, professional AI music tools like AIVA (Artificial Intelligence Virtual Artist) and Amper Music have appeared more significant, enabling media production industry professionals to easily create royalty-free soundtracks and customize them within a matter of seconds. The tools are purposely developed to facilitate use, prioritizing practical functionality and easy embedding in video, game, and advertising campaigns.

In the current developments in the tech world, a serious and comparative analysis of these AI models is a first order priority. Each model has its own unique set of capabilities, includes limitations and areas of application concentration. Differences in their underlying structure designs—be they transformer networks, variational autoencoders, recurrent neural networks, or hybrids—coupled with the heterogeneity and size of their training data, have a significant impact on the quality, internal coherence, and affective value of the generated output. Additionally, factors like input mode (text, MIDI, waveform), level of control over output (genre, emotional content, and tempo), and level of user involvement (prompt-based through complete automation) are significant factors in evaluating creative potential and everyday usability inherent in each instrument.

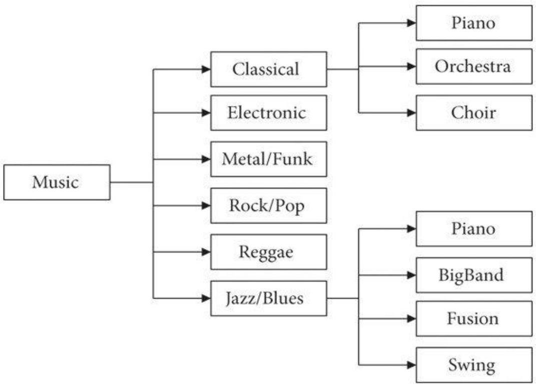

Figure 1. Music Generation Classification

History of artificial intelligence has led to amazing developments in the use of music generation, allowing one to write music in a very wide variety of genres. From what is observed from the diagram, music can be divided into different genres such as Classical, Electronic, Metal/Funk, Rock/Pop, Reggae, and Jazz/Blues. Each of these types branches off into subtypes—i.e., Classical music branches off into Piano and Orchestra pieces, and Jazz/Blues branches off into forms such as BigBand, Fusion, and Swing. Such categorization is used to signify the range of music expression as well as to serve as a firm basis for AI software to mimic or continue from.

More recent advancements in music generation AI models such as OpenAI's MuseNet and Google's MusicLM have been reported to compose music across diverse genres through training from enormous corpora of audio and MIDI. Deep learning architecture, primarily based on transformers, is utilized in these models to understand musical structure, harmony, rhythm, and style-specific data. For example, MuseNet can generate a Mozart-classical sonata for piano or jazz with exact-resolution swing rhythms from a prompt input or training bias. Genre-conditioned models those allow music to be generated that follows particular genres but also adds new variations.

There are trained AI models specifically that specifically strive to give extra attention to specific genres. Software like RiffEdit and Riffusion is especially well suited for generating music in real time with spectrogram-based diffusion, because this is more applicable to genres like Electronic and Funk. AIVA (Artificial Intelligence Virtual Artist) music is voiced specifically to obtain orchestra and film quality sound, which is more suitable to the Classical genre. With the ability to generate music for all genres, the creative professionals are assisted and user experiences in gaming, movie, and therapy businesses are also stimulated. This literature survey is interested in the diversity of music generation across genres and how different AI systems approach these different musical genres.

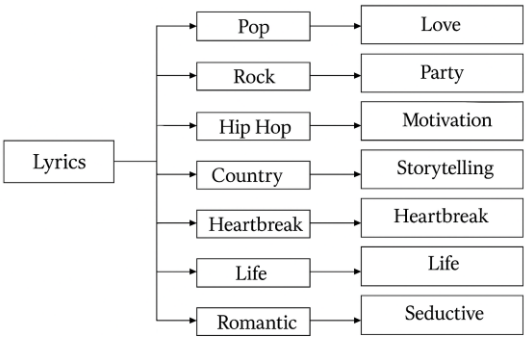

Figure 2. Lyrics Generation Classification

With the rapid development of artificial intelligence in creative domains, lyrics generation has emerged as a promising area where models are trained to create meaningful, emotion-rich textual content. As depicted in the diagram, lyrics can be classified into broad musical genres such as Pop, Rock, Hip Hop, Country, and Romantic. Each genre further branches into specific themes—for example, Pop often revolves around Love, while Hip Hop frequently expresses Motivation. These distinctions are critical in guiding AI models to produce lyrics that not only match a genre's musical structure but also capture the associated emotional tone and thematic essence.

Different AI models specialize in generating lyrics tailored to particular genres and moods. GPT-based models like OpenAI's GPT-3.5 and GPT-4, as well as Google's LyricAI and Jukebox, use massive language datasets to learn lyrical patterns, rhymes, and contextual flow. For instance, models trained on Rock lyrics may produce energetic or rebellious content suited for Party themes, while those trained on Country datasets are adept at crafting vivid storytelling narratives. Emotion-sensitive models can even generate lyrics for more nuanced categories such as Heartbreak or Seductive Romantic content, allowing a diverse range of lyrical compositions that resonate with varied audience preferences. The diversity in lyric generation is essential not only for songwriters and composers but also for entertainment apps, virtual artists, and therapeutic music tools. AI's ability to understand genre-specific tones—from the uplifting vibe of motivational Hip Hop to the raw emotion of Heartbreak ballads—enables it to assist human creators or even act as autonomous lyricists. This literature survey explores how genre-specific training, fine-tuning, and prompt engineering shape the quality and creativity of AI-generated lyrics across different musical styles and emotional themes.

This research essay attempts to conduct a comprehensive review of the most common artificial intelligence models being used in the composition of musical pieces and also lyrical material. It will evaluate these tools on a list of established parameters like input and output modality, the extent of creative freedom utilized, the internal quality of the produced piece, the extent of personalization provided, the technology involved behind, and their functional applications under actual-world conditions. By scrutinizing the architectural building blocks in detail, the training paradigms, and the scope of functional capabilities exhibited by each model, this work seeks to clarify the inherent strengths and the common fallacies within AI-created music. Therefore, in this paper, an attempt is made to add to modern-day academic literature pertaining to computational creativity and present intellectual foundations for future developments that can continue to minimize the current gap between the creative goals of human artists and supporting capabilities of machine intellect.

II. LITERATURE SURVEY

A. LYRICS GENERATION METHODS

The model discussed in [1] is a conditional GPT-2 model that is suited for generating lyrics for a specific musical style with its suitability for singing. The model has the text generation abilities of GPT-2 combined with a syntactic parser and an adapting rhyme module to generate contextually suitable and harmonically consistent lyrics to be sung with the provided music. This gives the model, the ability to create suitable and consistent lyrics. Yet, in contrast to ChatGPT-3.5, the GPT-2-based model is found to lack creativity, awareness of context, and flexibility. With a higher number of model parameters and effective training procedures, ChatGPT-3.5 generates lyrics. Though the GPT-2-based approach is found to work well, ChatGPT-3.5 is superior with better performance and enhanced flexibility in generating vivid lyrics.

This paper [2] tells that GPoeT-2 is a special variant of the GPT-2 architecture, comprehensively optimized for generating specific forms of poetry, specifically limericks. Designed to automate the creative writing process, GPoeT-2 operates on a two-stage generation protocol using forward and backward language modelling methods for strict compliance with the characteristic AABBA rhyme scheme of limericks. Its training protocol uses specially curated poetic datasets, and it generates output without seed phrase inputs or constraints imposed during generation, only depending on the fine-tuned model to generate syntactic coherence, correct rhyming, and thematic coherence. The GPoeT-2 technology behind it is a derivative from transformer-based GPT-2 model technology, which, as adept as it is by nature, has certain limitations in holding context, understanding minute nuances, and maintaining coherence on very lengthy text passages. In comparison to ChatGPT-3.5, with advanced architecture, GPoeT-2 has some limitations. With greater count in parameters, high training corpus, advanced optimization methods, ChatGPT-3.5 provides better linguistic fluidity, adaptability with a variety of writing styles, and handles long-range relations in text. The GPT-3.5 model has better interaction and sense to prompt such that, it can generate poetry based on trained creative parameters around specific themes or emotions provided by users. Whereas, the capability is lacking with GPoeT-2 because of its limited training and more specialized goal. Overall GPoeT-2 is creative in generating structured poetic form automatically but ChatGPT-3.5 is a smarter and more powerful system that is not only superior in the automatic creation of structured poetry but also has great utility and consistency in a wide range of tasks in the natural language generation.

This study in paper [3] explores the ability of GPT-3 to understand and justify musical decisions based on the descriptions of music with the goal of enabling meaningful conversation in human-AI collaborative music. Although GPT-3 is shown to be skilled in many natural language processing tasks, it shows shortcomings in completeness or justifying artistic decisions related to music, thus being a limitation in the simulation of genuine artistic reasoning. The main identified limitation is the lack of advanced datasets containing annotated descriptions of artists' creative processes, hindering GPT-3 interpretability in music tasks. More recent models like ChatGPT-3.5 have improved contextualization and interactive abilities but do not possess a basis of knowledge of musical theory unless they are expressly trained within this subject area. Accordingly, although GPT-3 has opened doors to creative AI engagement, its application in musical reasoning is limited, requiring the use of advanced datasets and training processes to achieve the full capabilities of AI's musical intelligence and collaborative potential.

The paper [4] introduces a deep learning-based approach using Long Short-Term Memory (LSTM) networks to generate lyrics tailored to specific artists and musical genres. While earlier models used RNNs or GRUs, this study enhances the lyric generation process by implementing a multilayer LSTM architecture integrated with bidirectional neurons and BERT for better context understanding. The model operates on seed lyrics and generates outputs that mirror the lyrical style, rhyme patterns, and word variation of the training data. By breaking input into word and rhyme indices and adjusting model parameters, it effectively produces unique and genre-specific lyrical content. Compared to ChatGPT-3.5, which leverages large-scale transformer architecture for general-purpose text generation, the LSTM-based model is more targeted and controlled but less versatile. ChatGPT-3.5 outperforms in maintaining contextual flow and diversity over longer texts, while LSTM remains effective in mimicking stylistic elements. In conclusion, though LSTM models can closely replicate artist styles and generate quality lyrics, transformer-based models like ChatGPT-3.5 provide greater creativity, coherence, and adaptability, making them more suitable for dynamic songwriting applications.

This article [5] is concerned with the problem of Music Emotion Recognition (MER) from lyrics, a subfield of Music Information Retrieval (MIR). It presents one of the most applicable use cases of lyric analysis—recognizing the emotional content in songs to aid music streaming and recommending services like Spotify and YouTube. The research suggests that a system for detecting the mood of a song such as happy, sad, relaxed, or angry can be done by deriving emotionally significant features from audio signals and lyrics with emphases on values such as valence and arousal. The methodology strongly matches transfer learning with the pre-trained BERT model to improve semantic content understanding in lyrics. Lyrics from the Music4All dataset were used for training and testing purposes. By fine-tuning BERT on the lyrics with emotion-specific labels, the model attained a high accuracy of 92%. The research is novel in the literature because it incorporates deep learning and natural language processing to enhance the emotional understanding of song lyrics to be of use for next-generation music recommendation systems.

This paper [6] investigates the problem of automatic lyrics continuation with the use of the T5 (Text-to-Text Transfer Transformer) model, a pre-trained language model that has been found to be highly adaptable in NLP tasks. The authors implement two variants of the model: the SA (Specific Author) variant that takes into account the stylistic traits of a given artist, and the NSA (Non-Specific Author) variant that is concerned with overall song formats. The model has eight different decoding approaches, such as Beam Search, Contrastive Search, Top-k and Top-p sampling, among others, to facilitate experimentation and flexibility in the generation process. The model intends to maintain such lyrical attributes like rhyme and form, while streamlining the songwriting process. Quantitative analysis was performed using BLEU, Rouge-L, and Rouge-N scores, based on varying decoding approaches. Beam Search and its variants such as Beam Sampling proved to be the top-performing methods in terms of sustaining lyrical coherence and quality. Contrastive Search also performed well, beating the baseline Greedy Search in the majority of the cases. Nevertheless, Top-p and Top-k approaches had varied performance based on the generation task. The NSA model proved more stable in its outputs compared to the SA variant, although the latter performed relatively well when author-specific training was balanced. In comparison to other traditional lyric generation methods or less complex RNN-based models, this T5-based model provides greater control, diversity, and fluency in lyric continuation and is a much more efficient and flexible solution for real-world songwriting applications.

This research [7] investigates the problem of automatic continuation of lyrics through the T5 (Text-to-Text Transfer Transformer) model, a language model pre-trained for its versatility across NLP tasks. Two model variants are proposed by the authors: the SA (Specific Author) variant which takes into account the stylistic characteristics of an individual artist, and the NSA (Non-Specific Author) variant which addresses general song patterns. The system allows for eight different decoding methods, such as Beam Search, Contrastive Search, Top-k and Top-p sampling, among others, so flexibility and experimentation can be allowed during the generation process. The model tries to maintain lyrical features like rhyme and structure, as well as speed up the process of songwriting. Quantitative assessments were conducted on models using BLEU, Rouge-L, and Rouge-N scores as benchmark for various decoding techniques. Beam Search and its extensions such as Beam Sampling have proved to be the top-performing methods in terms of preserving lyrical quality and coherency. Contrastive Search also performed better than the baseline Greedy Search in all cases. Top-p and Top-k methods, however, had varied results based on the generation scenario. The NSA model produced more consistent outputs compared to the SA variant, although the latter performed well when author-specific training data were balanced. Compared to conventional lyric generation methods or easier RNN-based models, this T5-based system has more control, diversity, and fluency in lyric continuation, thus making it a much more efficient and versatile solution for practical songwriting tasks.

This paper [8] describes a comprehensive review of BERT (Bidirectional Encoder Representations from Transformers), a model which has radically changed Natural Language Processing by harnessing deep bidirectional context in order to understand better linguistic patterns. BERT's architecture relies on the transformer encoder and is trained with masked language modeling and next sentence prediction, allowing it to capture complex contextual dependencies within text. This research not just analyzes BERT's structural architecture and training process but also investigates its performance on a broad range of NLP tasks such as Sentiment Analysis, Named Entity Recognition, Question Answering, Machine Translation, and Cross-lingual Transfer Learning.

This chapter [9] discusses the two main types of architectures of attention-based language models, which characterize the distribution of tokens in texts: Autoencoders like BERT take an input text and output a contextual embedding for every token. Autoregressive language models like GPT take a subsequence of tokens as input. They output a contextual embedding for each token and predict the next token. In this manner, all tokens of a text can be generated successively. Transformer Encoder-Decoders are responsible for translating an input sequence into another sequence, e.g. for language translation. Initially they produce a contextual embedding for each input token through an autoencoder. Subsequently, these embeddings are fed into an autoregressive language model, which produces the output sequence tokens sequentially. These models tend to be first pre-trained in a large general training set and subsequently fine-tuned for an application. So, they come under the definition of Pre-trained Language Models (PLM). As the parameters of these models become large in number, then often they tend to be able to be prompted by instructions and are referred to as Foundation Models. In other sections we detailed information on methods of optimization and regularization applied at the time of training. Finally, we discuss the uncertainty of model predictions and how predictions can be explained.

This work [10] is on melody-conditioned lyrics generation, a problem where the system learns the semantic and rhythmic correlation between musical melodies and the respective textual lyrics. The authors solve two prominent issues prevalent in this area: the absence of inter-dependency modeling among the attributes of melodies (such as pitch, duration, etc.) and the degradation of generated text in terms of syllable consistency and logical coherence. In order to counter these challenges, the paper proposes a Semantic Dependency Network, which uses two primary components to improve generation quality.

The initial part, an N-gram CNN block, compresses and combines information from single and multiple melody features to more effectively capture their inter-dependencies. With this, the model allows the produce musically coherent and structurally consistent lyrics. The second part shows unlikelihood training, an approach towards training that scores off negative or nonsense predictions during training, to enforce syllabic consistency and lyrical coherence. This avoids problems such as mismatched syllables and disrupted semantic continuity in produced text. In comparison with the current state-of-the-art methods in melody-conditioned lyric generation, this method shows significant enhancement. Experimental assessment on a massive dataset reveals that the model generates more harmonic and semantically consistent lyrics with both musical and text quality, showing improvement over previous models. Its advantage is it has the potential to bridge the gap between melody and meaning, an aspect that past models commonly suffer from as a result of detached feature modeling or naive generation paradigms.

This work [11] introduces a melody-conditioned lyrics generation model from Sequence Generative Adversarial Networks (SeqGAN), a state-of-the-art deep learning framework with adversarial training that enhances sequence generation quality. The model is trained to produce a line of lyrics at a time, conditioned on the related melody, which is a huge improvement over previous models that usually disregarded melodic features or oversimplified their impact on lyrics. This end-to-end model learns from data directly without relying on hand-designed rules or musical domain knowledge. What distinguishes this model is that it has a dual conditioning mechanism — it takes melody as its main input and optionally provides a thematic or topical context for the lyrics. With theme guidance incorporated, the system improves the coherence and meaning of the produced lines without compromising performance. Empirical tests verify that this extra context has no adverse impact on baseline evaluation metrics (such as BLEU and ROUGE), yet qualitative analysis reveals enhanced thematic coherence and creativity. In contrast to older rule-based or simple neural models, this SeqGAN-based system exhibits improved lyrical quality, improved melody-text coherence, and improved adaptability in generating useful content. It also fares competitively with past melody-conditioned systems through the use of adversarial learning, which compels the generator to produce outputs indistinguishable from actual lyrics. Therefore, this model reconciles musical structure and lyrical semantics better than past methods.

This work [12] introduces a new Conditional Hybrid Generative Adversarial Network (C-Hybrid-GAN) to tackle the task of generating high-quality melodies from textual lyrics. Unlike most state-of-the-art music generation systems, which find the discrete nature of musical features like pitch, duration, and rest challenging, this model innovatively addresses the discrete sequence generation task through the Gumbel-Softmax approximation. This approach enables the model to produce discrete values in direct output, enhancing the fidelity of the output melodies. C-Hybrid-GAN architecture is hybrid in the sense that three generator branches that independently produce each of the major melody features (pitch, duration, rest) condition on one and the same common set of lyrics. For a contrast, discriminator acts over concatenated outputs in order to examine consistency and the truthfulness of output sequences. In addition, a Relational Memory Core (RMC) is used for improving temporal and cross-attribute consistency—providing intra-sequence coherence and inter-sequence dependency. In contrast to state-of-the-art approaches, C-Hybrid-GAN drastically improves in producing musically coherent and structurally consistent melodies, which is confirmed by experimental metrics like Maximum Mean Discrepancy (MMD), average rest value, and MIDI transition quality. In contrast to other models such as GAN-based or rule-based models, this system more accurately captures the subtlety of discrete musical components while along with maintaining semantic consistency with the input lyrics. The hybrid approach and focused conditioning provide it with a significant advantage both in terms of flexibility and generation precision, establishing it as a state-of-the-art method for lyric-to-melody generation tasks.

This paper [13] deals with tackling the creative and structural challenges of songwriting by designing a semi-automatic lyric generation system for English songs. The authors acknowledge the need for both inspiration and linguistic creativity in writing lyrics and try to bridge the gap by employing a model based on methods from Artificial Intelligence (AI) and Natural Language Processing (NLP). Prior to constructing the model, the researchers performed in-depth analysis in order to grasp what makes "good" lyrics — analyzing patterns, themes, structure, and emotional appeal across current songs. Although the model does not have the same level of knowledge as the transformer-based models or the state-of-the-art GAN models, it is higher in terms of usability and practical songwriting aid. The focus on human judgment also shows the practical use in the real world. In contrast to completely automatic lyric generation systems, the semi-automatic model also provides a more cooperative and adaptable tool for artists particularly, those looking for creative inspiration as opposed to complete songs.

This research [14] investigates the application of a fine-tuned pre-trained GPT-2 language model to the task of lyric writing in multiple languages. The study addresses generating English and Portuguese lyrical lines and shows how transformer language models may be transferred to the musical and poetry text domain. Through fine-tuning GPT-2 on two different corpora of lyrics, the model learns to acquire the stylistic and syntactic patterns typical of musical texts, including poetic phrasing, rhyme, idiomaticity, and metaphorical speech. The authors conduct both quantitative and qualitative analyses of the output lyrics. They test the outputs for spelling accuracy, syntactic naturalness, and semantic coherence, and also try to identify patterns across different samples output by the model. The analysis thus underscores the GPT-2 model's fluency and language comprehension strengths, as well as its weaknesses in processing extended figurative language such as metaphors and metonymy—comprising parts in songwriting. The system was not only evaluated based on traditional NLP evaluation metrics but was also examined from literary and musical angles. Compared to traditional rule-based systems and earlier architectures like LSTM and RNNs, GPT-2 showcases a profound step ahead in quality of lyric generation, especially when handling contextual coherence and lexicality. However, it still lags behind in understanding deeper poetic structures and inter-line coherence, which is most important to music-based content. In spite of those shortcomings, the study is a testament to the potential of GPT-2 as an extremely powerful tool for multilingual, poetry-form text generation and that domain adaptation and fine-tuning can turn it even more creatively powerful. The study also stresses the role of diverse datasets and cultural context in improving lyric authenticity.

This work [15] proposes a music and affect-aware Chinese song lyrics generation model that takes into account musical and emotional aspects in text generation. The work intends to enhance one of the weaknesses of existing models, the seed-based or word-prompt generation, but not the melodic and affective suitability of music and lyrics. To shatter that, the authors propose a two-stage framework coupled with Support Vector Regression (SVR) and sequence-to-sequence (Seq2Seq) deep learning methods. At the first stage, SVR is used to predict melody emotions from musical input, i.e., to quantify the emotional tone conveyed by a melody. At the second stage process, the Seq2Seq model is trained to generate lyrics by conditioning the musical notes by recognizing melody emotions and available lyrics. This allows the model to learn not just the syntactic structure, but also emotional coherence, generating more coherent lyrics that are closer to musical context. Experimental results validate that the system produces semantically consistent and coherent sentences, indicating an improvement from the earlier models that were not emotionally integrated. This SVR + Seq2Seq hybrid configuration is superior to standard models like LSTM or even prompt-based transformer models because it produces a richer output that is emotionally and musically richer. It is particularly well-suited for application in scenarios where emotional expression in music is a prerequisite. This approach is much better at generating affectively-coherent lyrics, especially in a tonal language like Chinese where syllable harmony and melody matter. It is thus a valuable contribution to multi-modal lyrics generation, uniting affective computing and NLP to improve the music generation pipeline.

In contrast to unidirectional models, BERT has contextualized embeddings provide better comprehension of polysemy and sentence semantics which has resulted in better task performances. The review also showcases the superiority of BERT models over other previous models such as RNNs and LSTMs across different benchmarks as a result of its deeper representations of language and scalability. Along with it, the study even shows the flexibility and limitations of BERT-based systems, including computational resources needed and the requirement for fine-tuning methods. Through the compilation of results from various sources and demonstrations of BERT's dominance in applications such as fake review detection and grammar correction, this paper situates BERT as a baseline not just for contemporary NLP solutions but also as a driving force behind the creation of faster and more intelligent future language models.

In this literature survey, our primary objective is to systematically examine and categorize existing models used for automated lyrics generation. To ensure a comprehensive analysis, we adopted a structured methodology focused on grouping the literature based on model architecture, conditioning techniques, and output evaluation methods. The surveyed works are first classified by the underlying model types—such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, and transformer-based architectures like GPT-2 and GPT-3.5. w

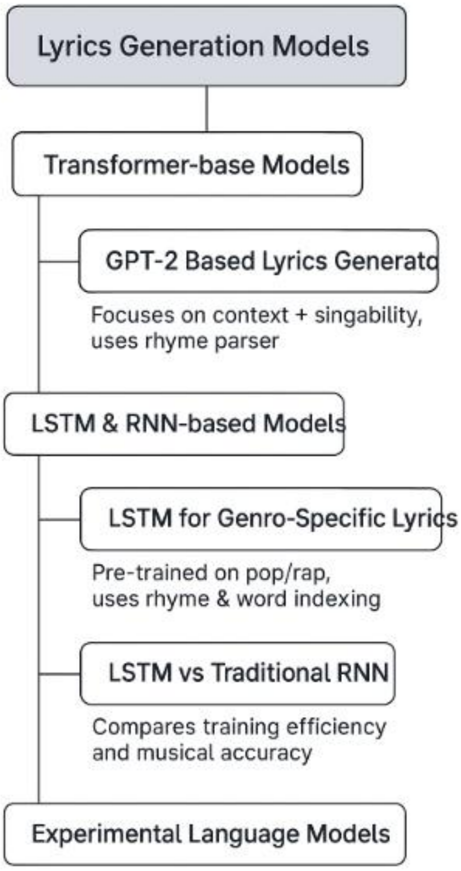
Lyrics Generation Models
Transformer-base Models
GPT-2 Based Lyrics Generato
Focuses on context + singability, uses rhyme parser
LSTM & RNN-based Models
LSTM for Genro-Specific Lyrics
Pre-trained on pop/rap, uses rhyme & word indexing
LSTM vs Traditional RNN
Compares training efficiency and musical accuracy
Experimental Language Models

Figure 3. Lyrics Generation Models

Beyond architectural classification, we also group the models based on their functional enhancements, including syntactic parsers for structural refinement and rhyme enforcement modules that ensure singability. This classification helps highlight how different approaches address key challenges in lyrics generation, such as maintaining rhyme schemes, structural coherence, and contextual relevance. Furthermore, we assess each model's performance using both objective metrics—like BLEU scores, perplexity, and rhyme accuracy—and subjective human evaluations where available, such as creativity, fluency, and listener appeal. This layered approach allows us to draw meaningful comparisons between models and identify common strengths, limitations, and research gaps in the field of AI-powered lyrics generation.

B. MUSIC GENERATION METHODS

This study [16] presents a music genre classification model using Artificial Neural Networks (ANN), diverging from the conventional use of deep Convolutional Neural Networks (CNN) that typically rely on spectrogram images and require high computational resources. Instead, the model utilizes key extracted audio features—such as Mel-frequency Cepstral Coefficients (MFCC), Chroma features, and Spectral Contrast—as inputs to a lightweight ANN, significantly reducing complexity while preserving accuracy. Trained on the GTZAN dataset, the model achieved 98.41% training accuracy and 97.05% validation accuracy. The paper argues that this selective feature-based approach enables high-performing classification suitable for real-time and low-resource environments. While not intended for direct music generation like systems such as MusicGen, the work demonstrates the importance of efficient genre identification in applications like music recommendation and analysis, emphasizing that precision can be achieved without the overhead of deep neural architectures.

This study [17] presents an automatic music generation system that utilizes Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), to generate and evaluate music from MIDI files. The MIDI inputs are transformed into MIDI matrices that encode note sequences for model training. The generation process explores both single-layer and double-stacked variants of LSTM and GRU architectures. To assess the musicality of the output, the study employs dual evaluation methods: an objective classification task that assigns compositions to historical musical eras, and a subjective listener survey involving participants from various musical disciplines. The double-stacked GRU model performs best, with a 70% recall score in classification and a subjective listener rating of 6.85/10, indicating its capacity to produce musically coherent outputs. Although GRU-based models show strong pattern learning and temporal awareness in symbolic music, the study acknowledges their limitations in controllability and audio fidelity when compared to systems like MusicGen. Still, the research reinforces the relevance of RNN-based methods in generating stylistically grounded symbolic music.

This study [18] investigates the use of Long Short-Term Memory (LSTM) networks for AI-based music generation, focusing on sequence continuation from partial MIDI inputs. The LSTM model, well-suited for capturing long-term dependencies in sequential data, is trained to extend existing musical compositions in a stylistically coherent manner. The authors improve model performance by analyzing the relationship between training loss and epochs to fine-tune learning effectively. A comparative analysis with standard RNNs further highlights LSTM’s superiority in maintaining musical structure over longer time spans. While the generated melodies show promise in terms of sequence prediction and structural coherence, the method falls short in producing high-fidelity audio compared to advanced models like MusicGen, which incorporate multi-stream outputs and text-based conditioning. Nonetheless, the study underscores LSTM's utility in foundational tasks within symbolic music generation and its role in advancing creative AI systems.

This study [19] introduces MuseNet, a music generation model that employs a two-part architecture featuring a discriminator and a generator for structured, abstractive composition. The discriminator predicts the first chord of a new bar based on the previous one, ensuring harmonic continuity, while the generator builds on this chord to complete the current bar with musically dense and coherent notes. This method enables MuseNet to maintain long-form musical flow without sacrificing creativity. A key advantage of the model is its lightweight training process, as it avoids complex components like CTC loss layers or GANs, allowing for faster and more efficient model development. Though MuseNet lacks the granularity and multi-stream capabilities of advanced systems like MusicGen, which support stereo outputs and fine-grained token conditioning, it remains a compelling solution for coherent and efficient music generation in simpler use cases.

This study [20] examines the role of artificial intelligence in music education through the integration of the AIVA (Artificial Intelligence Virtual Artist) platform into an interactive learning environment. Conducted as a quasi-experimental study, it compares AI-assisted instruction with traditional lecture-based methods, using ANCOVA to assess student outcomes. The findings indicate that students using AIVA demonstrated faster mastery of musical concepts and greater enthusiasm for theoretical content, showcasing the platform’s intuitive and engaging design. However, the study also notes potential drawbacks, such as decreased attention and increased fatigue in the absence of human interaction. While AIVA differs from generative models like MusicGen—which focus on producing controllable, high-fidelity music from textual or melodic input—it plays a unique role in educational settings by enhancing learning engagement and conceptual understanding. This positions AIVA as both a creative and pedagogical tool, capable of modernizing music education practices.

This study [21] presents MusicGen, a state-of-the-art model for conditional music generation built on a single-stage transformer-based language model. It operates on compressed discrete audio representations encoded as tokens, eliminating the need for multi-stage or hierarchical generation methods. MusicGen supports input conditioning via both textual descriptions and melodic prompts, offering users a high degree of creative control. Through efficient token interleaving strategies, the model generates coherent, high-quality mono and stereo audio samples in a single pass. Empirical evaluations—both automatic and human—demonstrate that MusicGen significantly outperforms existing models on standard text-to-music benchmarks. Its unified architecture and flexible conditioning mechanisms set a new standard for AI-generated music, making it especially effective in tasks requiring expressive and stylistically accurate outputs.

This study [22] is a qualitative study of two artificial intelligence-based music composition platforms—Jukedeck and WaveAI (the developer of the ALYSIA app). These companies are among the first to create music based on artificial intelligence, aiming to produce melodies and songs through machine learning algorithms. Although these platforms apply basic neural forms and probabilistic models for the purpose of generating musical pieces or the lyrics, the author says that such computationally advanced models fail to capture human imagination, creativity, and expression of feelings. Instead of examining the inner mechanisms of the individual models technically, the Socio-cultural and philosophical critique is shown in the paper. It shows that the symbolic and emotional nature of music resists complete transposition into the "impassive language of mathematics." Referring to theorists like Jean-François Lyotard and Alan Turing, the research places AI music as a cyborgian by-product of human imagination—a machine with limited but growing expressive ability. Relative to other quantitative-performance-based models such as melody-harmony congruence or BLEU scores (e.g., MusicGen, Jukebox, or C-Hybrid-GAN), this view recognizes that Jukedeck and ALYSIA are not necessarily more technically wasteful, but they are reflective of the inherent limitations of today's AI in replicating human creativity and cultural sensitivity. The paper calls not for the exclusion of such tools, but for the embracing of AI music as a complementary social artifact, not a substitute for human artistry.

This study [22] is a qualitative study of two artificial intelligence-based music composition platforms—Jukedeck and WaveAI (the developer of the ALYSIA app). These companies are among the first to create music based on artificial intelligence, aiming to produce melodies and songs through machine learning algorithms. Although these platforms apply basic neural forms and probabilistic models for the purpose of generating musical pieces or the lyrics, the author says that such computationally advanced models fail to capture human imagination, creativity, and expression of feelings. Instead of examining the inner mechanisms of the individual models technically, the Socio-cultural and philosophical critique is shown in the paper. It shows that the symbolic and emotional nature of music resists complete transposition into the "impassive language of mathematics." Referring to theorists like Jean-François Lyotard and Alan Turing, the research places AI music as a cyborgian by-product of human imagination—a machine with limited but growing expressive ability. Relative to other quantitative-performance-based models such as melody-harmony congruence or BLEU scores (e.g., MusicGen, Jukebox, or C-Hybrid-GAN), this view recognizes that Jukedeck and ALYSIA are not necessarily more technically wasteful, but they are reflective of the inherent limitations of today's AI in replicating human creativity and cultural sensitivity. The paper calls not for the exclusion of such tools, but for the embracing of AI music as a complementary social artifact, not a substitute for human artistry.

This special issue [23] of journal Neural Computing and Applications provides an overview of recent progress of deep learning for music and audio tasks, in response to the first International Workshop on Music and Audio at IJCNN 2017. State-of-the-art methods applied to music generation, music transcription, voice separation, emotion recognition, and other tasks are discussed, focusing on the evolving relationship between neural networks and musical creativity. Several model architectures are described: e.g., RNN-based models (Oore et al.) demonstrated strong potential in generating expressive music, and Hadjeres and Nielsen employed user constraints in an RNN to harmonize soprano lines in Bach-style chorales. Dean and Forth explored post-tonal improvisation using recurrent architectures, demonstrating AI's potential to venture into niche musical forms. Other new models include deep convolutional networks for singing voice separation (Lin et al.), LSTM-based time-frequency models for audio restoration (Deng et al.), and word2vec-based models that learned harmonic structures like key and chords (Chuan et al.). What makes this problem unique is its multi-aspect assessment—not only comparing performance across models for tasks such as chord labeling or voice separation, but also defining more abstract notions that networks can learn (e.g., emotion in sound, harmony structure, low-SNR acoustic event detection). Models such as CNNs and hybrid deep structures are demonstrated to surpass conventional methods in accuracy and flexibility, particularly for tasks involving subjective musical taste or low-quality inputs. In contrast to standalone music generation tools like Jukedeck or MusicGen, this work is more integrated in its approach because it unites generation, classification, and restoration under one framework. In doing so, it sets the stage for a unified integrated understanding of how deep learning can revolutionize audio technologies.

This paper [24] offer an editorial overview of the Special Issue on Deep Learning for Music and Audio in the journal Neural Computing and Applications. Triggered by the increased interest in deep learning for audio applications, the issue collects a broad set of research in the areas of music generation, transcription, voice separation, emotion recognition, and style transfer. The editorial identifies critical research areas such as modeling long-term musical structures, handling ambiguity in chord labeling, and designing novel architectures that effectively represent music’s complex semantics. The models that are addressed consist of a broad range of deep learning methods: Hadjeres and Nielsen propose a recurrent neural network (RNN) design that reconciles soprano lines based on constraints provided by users, while Oore et al. show that RNNs are capable of producing expressive music that was received well by human musicians. In the meantime, Lin et al. employ CNNs with perfect binary masks to separate singing voices, and Dean & Forth address post-tonal improvisation, demonstrating good performance in a specialized musical style. The issue also explores audio feature learning. For instance, word2vec models are utilized by Chuan et al. to represent musical concepts such as chords and key signatures, and CNNs used for emotion recognition by Wieser et al., identifying their capacity for detecting emotional content in audio. Other studies propose architectures such as Time-Frequency LSTM (Deng et al.) for audio restoration and CNNs for low-SNR acoustic event detection (Kiskin et al.), showcasing the versatility of deep learning under different noise conditions and audio complexities. In contrast to models such as MusicGen that are mainly concerned with end-to-end melody synthesis, the contributions in this issue venture into the realms of granular and interpretable learning—emphasizing deep learning's applicability across both generative and discriminative tasks. This makes the work not only creative but also technically diagnostic, allowing for precise control and comprehension of musical material.

This paper [25] looks behind ten years of AI-assisted music composition research, showing the achievements and legacy of the Flow Machines project. The chapter goes back to the initial motivation for the project which is employing artificial intelligence to support and stimulate musical creativity. The paper also evaluates its technological and artistic impact years later upon its completion. Flow Machines used machine learning algorithms to produce stylistically consistent music, with the goal of combining human creativity with AI-generated novelty. The output of the project is not just generated pieces and music generation software, but also new forms of collaboration between humans and machines. It consists of systems that learn compositional rules from a database of pre-existing songs and produce new material that conforms to those stylistic patterns. The chapter even suggests new taxonomies of music, generated from the capacity of AI to create non-traditional content that also dismantles traditional genre boundaries. In contrast to music generation systems such as MusicGen, where melodic and instrumental generation from text or audio input is given top priority, Flow Machines emphasized style transfer, rule-based composition, and collaborative generation with users exerting explicit control over the creative direction. This half-automatic process enabled a richer human–AI interaction, not just affecting the technical but also musicological and philosophical aspects regarding authorship and artistic agency in the era of smart machines.

[26] is an interactive tool that is designed to assist non-technical users and designers alike in incorporating machine learning functionality—namely, gesture recognition—into web applications. It assists users to create body gestures, train classifiers and test, and implement gesture recognition into an end-to-end functional template application without prior knowledge of ML. Its intention is to make ML experiential, intuitive, and accessible to web developers as well as to user experience designers. An empirical evaluation with UX practitioners and MSc students indicated that the tool was effective in allowing individuals to try out ML concepts using visualizations and trial-and-error exploration. Individuals could access classifiers independently, and the direct manipulation supported their understanding of ML capabilities and limitations. The technique enhances experiential learning by abstracting ML processes into concrete steps in a design process. As compared to music generation systems like MusicGen, which hide the model training complexity and are more end-user-focused solutions, this system is user-in-the-loop model training and thus an efficient prototyping and learning tool. The paper also provides design principles for designing next-generation ML-based creative tools, specifically early-stage ideation and prototyping tools for non-experts.

[27] evaluates the performance of deep learning models on automatic music generation (AMG) with an extensive listening study between deep learning models and non-deep learning models. Evaluation was carried out on six music dimensions: stylistic success, aesthetic enjoyment, repetition/self-reference, melody, harmony, and rhythm. Experiment included the generation of 30-second excerpts of music from Classical string quartet and classical piano improvisation styles, by using both deep learning models (such as a reimplement of Music Transformer) and classical algorithms (such as MAIA Markov). Fifty music-expert listeners with high music expertise ranked the human-made and generated excerpts without being aware of their source. The experiment, using non-parametric Bayesian hypothesis testing, demonstrated that deep learning models were not better than non-deep learning models. Surprisingly, the performance of the Music Transformer was statistically comparable to the MAIA Markov model. This contradicts the belief that deep models are better than non-deep models for music generation. Furthermore, the paper points out a long-standing gap between human and machine music, stressing that neither shallow nor deep models yet are as creative as humans. In contrast to models such as MusicGen, which are based on large-scale pretraining and neural audio synthesis, this paper presents a symbolic-level comparison and stresses the subtle assessment of musical quality, which cannot be deduced directly from model complexity.

The work [28] introduces a new approach to automated music generation based on a sophisticated Generative Adversarial Network (GAN) model called MTM-GAN (Multi-Track Music GAN) for solving the costly and intricate nature of traditional music generation. Five parallel music tracks—bass, drums, guitar, piano, and strings—are generated by the model towards the simulation of realistic multi-instrument music compositions. MTM-GAN is contrasted with MuseGAN, a state-of-the-art multi-track GAN architecture, to evaluate improvements in convergence, realism, and musical quality. Experimental outcomes show that MTM-GAN generates more agreeable and fluid music clips. The Consistency Term (CT) is a significant enhancement, which enables the model to converge faster and more stably compared to MuseGAN. The parameter distribution learned by MTM-GAN has a numerical space with fewer numbers which reduces the risk of overfitting. In blind listening tests, 62.8% of the listeners were unable to discern AI-generated music from authentic music, which testified to the excellent generative ability of the model. Compared to other models like MusicGen, founded on transformer architecture and big-data pretraining, MTM-GAN aims structured symbolic music creation with improved stability through adversarial training. Thus, it can be a prospective architecture for real-time, multi-instrument music synthesis, particularly for use where interpretability and control of track structure are more important.

This work [29] introduces a hybrid music generation method that integrates an array of deep learning methods, i.e., Variational Autoencoders (VAEs), Long Short-Term Memory (LSTM) networks, and Transformers, to produce rich and purposeful musical experiences. The work begins with mining a rich collection of features from a vast database of music samples across varied genres and emphasizing the spectral features, rhythmic patterns, and tonal structures. These features are used as input to train models to generate music that is genre-specific as well as personalized. VAEs are used to map music into a continuous latent space in a manner that allows reconstruction of new but style-coherent samples. In contrast, LSTMs and Transformers are used to model and reconstruct intricate temporal relationships and sequential patterns in music. This aids the system in maintaining musical coherence over time. The entire architecture is used to prioritize flexibility and personalization in order to accommodate personal listening behavior. As a comparison to models like MusicGen, which are very transformer-based in attempting to output audio end-to-end, the hybrid model used in this project has structured control and explainability, especially within symbolic or MIDI-based representations. Although the system is not necessarily superior to current state-of-the-art systems within quantitative testing, it does present real potential for use within music recommendation systems and composer-assisted creative tools, precisely because it is structured and can handle multiple musical dimensions adequately.

This work [30] presents DiffuseRoll, a novel image-based music generation technique employing diffusion models for generating polyphonic, multi-track, and multi-attribute music, tackling most of the shortcomings of the majority of the previous techniques that worked on primarily monophonic or homophonic generation. The strategy is to represent music as piano-rolls, and then synthesize them in terms of diffusion. These piano-rolls are subsequently transformed back into MIDI for musical output. For encoding musical complexities, a color-coding approach is used where note pitch, velocity, tempo, and instrument are encoded by color and pixel position. The technique efficiently bridges the gap between discrete music events and continuous image data to facilitate the use of high-speed image-based generative methods. A post-processing module, Music Mini Expert System (MusicMES), is also utilized to optimize and enhance the generated output to achieve greater musical quality. In contrast to text-to-audio transformer-based models like MusicGen, DiffuseRoll provides a visually-oriented, interpretable, and modular approach to generating multi-attribute symbolic music. In contrast to raw audio delivery-focused MusicGen, DiffuseRoll is targeting symbolic composition with a terminal focus on both structural coherence and orchestration. Seven music-based subjective scores on Coherence, Diversity, Harmoniousness, Structureness, Orchestration, Overall Preference, and Average demonstrate considerable improvement over previous image-based solutions and suggest that the diffusion models have promise as an instrument within the structured music generation domain.

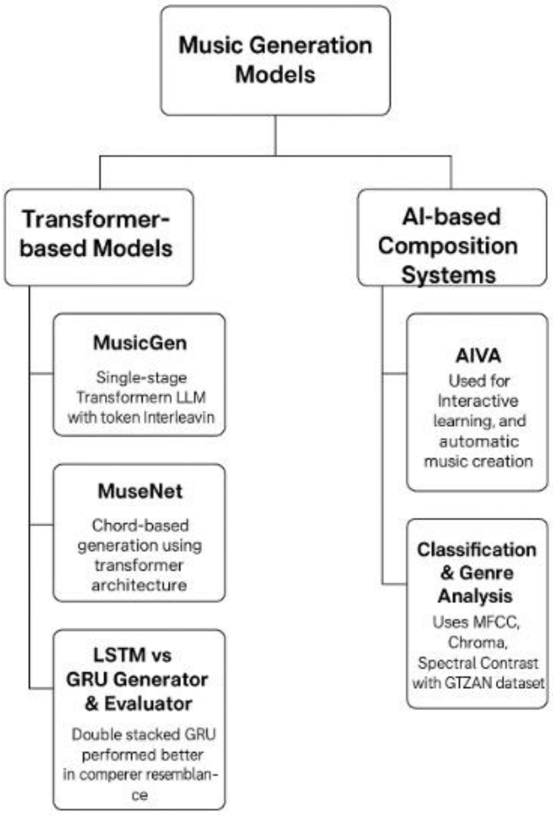
Music Generation Models
Transformer-based Models
MusicGen
Single-stage Transformern LLM with token Interleavin
MuseNet
Chord-based generation using transformer architecture
LSTM vs GRU Generator & Evaluator
Double stacked GRU performed better in comperer resemblan-ce
AI-based Composition Systems
AIVA
Used for Interactive learning, and automatic music creation
Classification & Genre Analysis
Uses MFCC, Chroma, Spectral Contrast with GTZAN dataset

Figure 4. Music Generation Models

In carrying out the survey of music generation models, we followed a systematic process to classify the literature according to the underlying neural architectures and their respective applications within the music generation field. The main classification criterion was the model type employed—Transformer-based models, Recurrent architectures such as LSTM and GRU, and application-oriented platforms or tools like AIVA. Each model was examined for how it could create music conditionally or independently with MIDI inputs, melody cues, or text descriptions. We also examined their training methods, including single-layer or stacked architectures, bidirectional neuron arrangements, and the incorporation of other language models such as BERT.

A further layer of comparison concerned the performance of these models in automatic and human-evaluation environments, specifically musicality, coherence, and creative imitations of composers. This structured categorization allows us to highlight technological evolution and identify emerging patterns, as well as to assess the growing role of AI in music composition and education.

III. CONCLUSION

The development of AI-powered lyrics generation models has significantly advanced the creative scope of songwriting. Leveraging large language models like GPT-2 and GPT-3.5, researchers and developers have created systems that not only generate grammatically correct text but also capture thematic depth, emotional tone, rhyme schemes, and lyrical structure. These models, when fine-tuned for music-related tasks, are capable of producing lyrics that align with specific genres, moods, or styles, reflecting the nuances of human creativity. With added modules for syntactic parsing and rhyme control, these models can ensure that the lyrics are both coherent and singable, matching the musicality required for successful integration into songs. Human and automatic evaluations of such frameworks reveal that AI-generated lyrics can be original, contextually meaningful, and musically compatible. As these models continue to improve, they offer promising support for artists, composers, and content creators seeking inspiration or looking to speed up the songwriting process.

On the music generation front, advanced models such as MusicGen, MuseNet, and LSTM-based architectures have made remarkable strides in producing high-quality, stylistically accurate compositions. Unlike earlier systems that relied on hierarchical or multi-stage architectures, recent models like MusicGen use single-stage transformer architectures with efficient token representations to generate coherent musical outputs. These models can be conditioned on textual prompts, melodies, or even stylistic inputs to generate music that closely mimics human compositions in terms of structure, harmony, and rhythm. Additionally, GRU and LSTM-based models trained on MIDI datasets have demonstrated the ability to extrapolate musical themes and emulate specific composers' styles. Evaluations—both subjective and objective—have indicated that such AI-generated music is not only listenable but also emotionally engaging, resonating well with listeners from varied backgrounds.

Taken together, the combined power of AI in both lyrics and music generation marks a significant step forward in the evolution of creative technologies. These systems bridge the gap between human expression and machine intelligence, enabling semi-autonomous or fully automated music production processes. When integrated thoughtfully, lyric-generation models and music-generation frameworks can collaborate to create full-length songs that align in theme, tone, and rhythm—transforming the way music is composed, produced, and experienced. As these technologies continue to mature, they hold the potential to democratize music creation, provide personalized composition tools for artists, and open up new forms of human-AI collaboration in the creative arts.

IV. REFERENCES

[1] Chang, J., Hung, J. C., & Lin, K. (2021). Singability-enhanced lyric generator with music style transfer. *Computer Communications*, *168*, 33–53. https://doi.org/10.1016/j.comcom.2021.01.002

[2] Lo, K., Ariss, R., & Kurz, P. (2022). GPOET-2: A GPT-2 based poem generator. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2205.08847

[3] Krol, S. J., Llano, M. T., & McCormack, J. (2022). Towards the Generation of Musical Explanations with GPT-3. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2206.08264

[4] Dhandapani, A., Ilakiyaselvan, N., Mandal, S., Bhadra, S., & Viswanathan, V. (2023). Lyrics Generation using LSTM and RNN. In *Lecture notes in electrical engineering* (pp. 371–388). https://doi.org/10.1007/978-981-99-1051-9_24

[5] Rajendran, R. V., Pillai, A. S., & Daneshfar, F. (2022). LyBERT: Multi-class classification of lyrics using Bidirectional Encoder Representations from Transformers (BERT).

Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-1501499/v1

[6] Mediakov, O., & Vysotska, V. (2024). SONGS CONTINUATION GENERATION TECHNOLOGY BASED ON TEST GENERATION STRATEGIES, TEXTMINING AND LANGUAGE MODEL T5. Radio Electronics Computer Science Control, 4, 157. https://doi.org/10.15588/1607-3274-2023-4-15

[7] Shahriar, S., & Roken, N. A. (2022). How can generative adversarial networks impact computer generated art? Insights from poetry to melody conversion. International Journal of Information Management Data Insights, 2(1), 100066. https://doi.org/10.1016/j.jjimei.2022.100066

[8] Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing: a review. Artificial Intelligence Review, 58(6). https://doi.org/10.1007/s10462-025-11162-5

[9] Paaß, G., & Giesselbach, S. (2023). Pre-trained language models. In Artificial intelligence: foundations, theory, and algorithms/Artificial intelligence: Foundations, theory, and algorithms (pp. 19–78). https://doi.org/10.1007/978-3-031-23190-2_2

[10] Duan, W., Yu, Y., & Oyama, K. (2023). Semantic dependency network for lyrics generation from melody. Neural Computing and Applications, 36(8), 4059–4069. https://doi.org/10.1007/s00521-023-09282-6

[11] Y. Chen and A. Lerch, "Melody-Conditioned Lyrics Generation with SeqGANs," 2020 IEEE International Symposium on Multimedia (ISM), Naples, Italy, 2020, pp. 189-196, doi: 10.1109/ISM.2020.00040.

[12] Yu, Y., Zhang, Z., Duan, W. et al. Conditional hybrid GAN for melody generation from lyrics. Neural Comput & Applic 35, 3191–3202 (2023). https://doi.org/10.1007/s00521-022-07863-5

[13] S. Pudaruth, S. Amourdon and J. Anseline, "Automated generation of song lyrics using CFGs," 2014 Seventh International Conference on Contemporary Computing (IC3), Noida, India, 2014, pp. 613-616, doi: 10.1109/IC3.2014.6897243.

[14] Rodrigues, M. A., Oliveira, A., Moreira, A., & Possi, M. (2022). Lyrics Generation supported by Pre-trained Models. Proceedings of the . . . International Florida Artificial Intelligence Research Society Conference, 35. https://doi.org/10.32473/flairs.v35i.130607

[15] Y. -F. Huang and K. -C. You, "Automated Generation of Chinese Lyrics Based on Melody Emotions," in IEEE Access, vol. 9, pp. 98060-98071, 2021, doi: 10.1109/ACCESS.2021.3095964.

[16] Gunawan, A. a. S., Iman, A. P., & Suhartono, D. (2020). Automatic music generator using recurrent neural network. *International Journal of Computational Intelligence Systems*, *13*(1), 645. https://doi.org/10.2991/ijcis.d.200519.001

[17] Sood, A., Rathee, T., Bansal, P., Garg, H., Aggarwal, A., & Tyagi, A. (2024). Music Genre Classification using Artificial Neural Networks. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-4428600/v1

[18] Arya, P. K., Kukreti, P., & Jha, N. (2022). Music Generation Using LSTM and Its Comparison with Traditional Method. In *Advances in transdisciplinary engineering*. https://doi.org/10.3233/atde220793

[19] Pal, A., Saha, S., & Anita, N. (2020). Musenet : Music Generation using Abstractive and Generative Methods. *International Journal of Innovative Technology and Exploring Engineering*, *9*(6), 784–788. https://doi.org/10.35940/ijitee.f3580.049620

[20] Yin, L., & Guo, R. (2024). An Artificial Intelligence-Based Interactive Learning Environment for music education in China: Traditional Chinese music and its contemporary development as a way to increase cultural capital. *European Journal of Education*. https://doi.org/10.1111/ejed.12858

[21] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). Simple and controllable music generation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.05284

[22] Cole, R. (2020). The problem with AI music: song and cyborg creativity in the digital age. Popular Music, 39(2), 332–338. https://doi.org/10.1017/s0261143020000161

[23] Cole, R. (2020). The problem with AI music: song and cyborg creativity in the digital age.

Popular Music, 39(2), 332–338. https://doi.org/10.1017/s0261143020000161

[24] Herremans, D., & Chuan, C. (2019). The emergence of deep learning: new opportunities for music and audio technologies. Neural Computing and Applications, 32(4), 913–914. https://doi.org/10.1007/s00521-019-04166-0

[25] Pachet, F., Roy, P., & Carré, B. (2020). Assisted music creation with Flow Machines: towards new categories of new. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2006.09232

[26] Winter, M., Jackson, P., & Fallahkhair, S. (2023). Gesture Me: a machine learning tool for designers to train gesture classifiers. In Communications in computer and information science (pp. 336–352). https://doi.org/10.1007/978-3-031-49425-3_21

[27] Yin, Z., Reuben, F., Stepney, S. et al. Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation. Mach Learn 112, 1785–1822 (2023). https://doi.org/10.1007/s10994-023-06309-w

[28] Liu, W. Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition. J Supercomput 79, 6560–6582 (2023). https://doi.org/10.1007/s11227-022-04914-5

[29] Pricop, T., & Iftene, A. (2024). Music Generation with Machine Learning and Deep Neural Networks. Procedia Computer Science, 246, 1855–1864. https://doi.org/10.1016/j.procs.2024.09.692

[30] Wang, H., Zou, Y., Cheng, H. et al. DiffuseRoll: multi-track multi-attribute music generation based on diffusion model. Multimedia Systems 30, 19 (2024). https://doi.org/10.1007/s00530-023-01220-9