

## RESEARCH ARTICLE

# The Analysis of Multi-Track Music Generation With Deep Learning Models in Music Production Process

RONG JIANG<sup>1,2</sup> AND XIAOFEI MOU<sup>3</sup><sup>1</sup>School of Humanities and Journalism, Xiamen University Tan Kah Kee College, Zhangzhou, Fujian 364000, China<sup>2</sup>College of Creative Arts, Universiti Teknologi MARA (UiTM), Shah Alam 40450, Malaysia<sup>3</sup>Department of Arts Management, China Conservatory of Music, Beijing 100875, China

Corresponding author: Xiaofei Mou (charles1210@126.com)

This work was supported by the Major Program in Arts Studies of the National Social Science Foundation of China under Grant 23ZD15.

**ABSTRACT** This study aims to explore the application of deep learning models in multi-track music generation to enhance the efficiency and quality of music production. Considering the limited capability of traditional methods in extracting and representing audio features, a multi-track music generation model based on the Bidirectional Encoder Representations from Transformers (BERT) Transformer network is proposed. This model first utilizes the BERT model to encode and represent music data, capturing semantic and emotional information within the music data. Subsequently, the encoded music features are inputted into the Transformer network to learn the temporal relationships and structural patterns among music sequences, thereby generating new multi-track music compositions. The performance of this model is evaluated, revealing that compared to other algorithms, the proposed model achieves an accuracy of 95.98% in music generation prediction, with an improvement in precision by 4.77%. Particularly, the model demonstrates significant advantages in predicting pitch of music tracks. Hence, the multi-track music generation model proposed in this study exhibits excellent performance in accuracy and pitch prediction, offering valuable experimental reference for research and practice in the field of multi-track music generation.

**INDEX TERMS** Deep learning, transformer, music generation, multi-track music, BERT.


## I. INTRODUCTION

### A. RESEARCH BACKGROUND AND MOTIVATIONS

In today's digital era, the music industry is undergoing unprecedented transformation and development. Music production, as a crucial component of music creation, embodies the creativity and emotions of the creators, directly influencing the quality and style of the works [1], [2]. Moreover, with technological advancements, music generation has evolved from simple random note generation to a complex process capable of simulating specific styles or works of artists [3]. However, traditional music production processes face numerous challenges such as high labor costs, long production cycles, and limited creativity, constraining the development and innovation of music creation. Furthermore, traditional

methods have limited capabilities in extracting and representing audio features, failing to effectively capture the advanced features and emotional expressions of music. Additionally, existing automatic music generation systems are mostly limited to single-track generation, and the automatic generation of multi-track music remains an unresolved issue [4].

The generation of multi-track music not only involves the generation of individual notes but also considers the harmony between different tracks, rhythm alignment, and overall musical structure [5], [6], [7]. Deep learning models possess powerful learning and expressive capabilities, enabling them to learn advanced features and patterns of music from vast amounts of music data and generate creative and expressive musical works [8], [9], [10], [11]. Compared to traditional methods, deep learning models can better capture the temporal relationships and emotional expressions of music, achieving more precise and efficient music generation.

The associate editor coordinating the review of this manuscript and approving it for publication was Giacomo Fiumara .

## B. RESEARCH OBJECTIVES

The aim of this study is to explore the integration of deep learning models into the music production process for multi-track music generation, aiming to improve the efficiency and quality of music creation and generate innovative musical works. Therefore, this study innovatively introduces deep learning algorithms and applies them to multi-track music generation models, providing new ideas and methods for the field of music production and promoting the development and application of music generation technology.

## II. LITERATURE REVIEW

The history of music generation technology can be traced back to the early 20th century, initially based on rule-based methods for generating melodies and harmonies, but they often lacked flexibility and creativity. For example, Ramirez et al. [12] proposed a rule-based evolutionary approach to simulate the music performance process. They optimized music performance rules through evolutionary computing techniques to automatically model music performance, providing an effective method for music generation. Hastuti et al. [13] employed rules and genetic algorithms to achieve automatic composition of gamelan music. They designed a set of rules to guide music composition and combined genetic algorithms to optimize and combine music elements, offering new insights for automatic music generation. Hastuti et al. [14] developed a rule-based interactive melody generator for gamelan music composition. They designed a series of rules to support users in quickly generating gamelan music works that met requirements. Wang et al. [15] reviewed the current status of intelligent music generation systems, including rule-based methods, emphasizing their importance and development prospects in the field of intelligent music generation.

With the development of computer technology, music generation has shifted towards two main approaches: template-based and statistical-based methods. These methods can analyze a large number of music works and learn patterns for music generation. For instance, Goienetxea et al. [16] employed statistical-based music generation methods, considering the coherence of rhythm and melody. Zhang [17] proposed a template-based learning adversarial transformer for symbolic music generation, which could enhance the quality and diversity of music generation. Liu [18] reviewed the application of statistical-based methods in intelligent music composition and introduced the role of Generative Adversarial Networks (GANs) music generation.

Entering the 21st century, machine learning technology has brought revolutionary changes to the field of music. Joy et al. [19] developed a music emotion recognition system using machine learning and deep learning, improving recognition accuracy. Sun [20] analyzed the principles and applications of machine learning in music composition, highlighting its potential and development trends in the composition process. Gonzalez and Prati [21] analyzed music timbre similarity using machine learning algorithms, providing technical

support for music retrieval and classification. Kuremoto [22] used machine learning methods to identify guqin music, improving recognition accuracy and efficiency, promoting the protection and inheritance of guqin music.

In recent years, deep learning-based music generation methods have gradually become a research hotspot. These methods utilize deep neural network models to learn the advanced features and patterns of music data and generate creative and expressive musical works. For instance, Ji et al. [23] reviewed the application of deep learning in symbolic music generation, including different representation methods, algorithms, and evaluation methods. Yin et al. [24] compared the performance of different deep learning algorithms in automatic music generation and proposed improvement strategies, providing important references for automatic music generation. Guo et al. [25] utilized deep learning algorithms to achieve music generation and evaluation, offering a new method for music composition. Sharma and Bvuma [26] explored the potential of GANs in creative applications, emphasizing the innovation and diversity of GANs in music generation. Ferreira et al. [27] generated symbolic music works using deep learning models, providing a new method for music composition. Moysis et al. [28] reviewed the latest advances of deep learning methods in music signal processing, offering important references for music technology research. Li et al. [29] utilized deep learning algorithms to achieve drum music generation, demonstrating the potential of deep reinforcement learning in music generation.

Through the analysis of current research, the evolution of music generation technology has been summarized, from rule-based, statistical-based to deep learning-based methods. Deep learning-based music generation methods have various advantages, including the ability to learn complex features and temporal relationships, fully utilize large-scale data, and possess strong generalization capabilities. However, this method still faces challenges, including high data and computational resource requirements, lack of structural and coherent generation results, and poor model interpretability. In response to these challenges, this study proposes an improved deep learning-based music generation method, providing new ideas and methods for intelligent music composition.

## III. RESEARCH METHODOLOGY

### A. DEMAND ANALYSIS FOR MULTI-TRACK MUSIC GENERATION

In modern music compositions, the overall musical landscape is shaped by the careful arrangement of multiple instrument tracks and their interactions and dependencies [30], [31], [32]. For instance, in a rock band, the drums typically establish the rhythm, while keyboards, bass, and guitar work together to form the accompaniment, and vocals undertake the task of singing. Each track needs to present a pleasing melody independently, allowing for harmonious coordination between tracks. However, current music generation models often adopt simplified strategies when dealing with symbolic

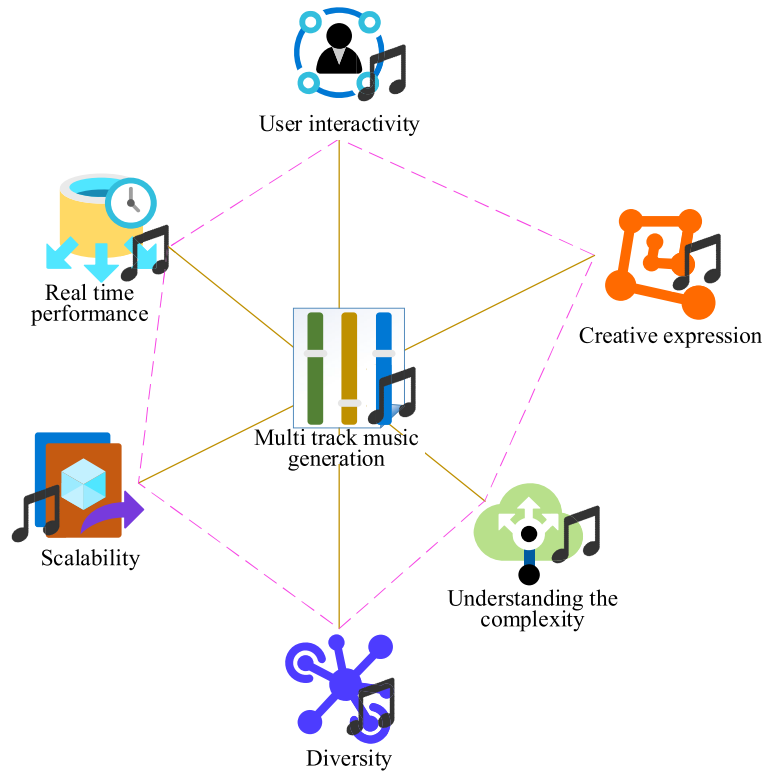


FIGURE 1. Illustration of the demand for multi-track music generation.

music generation [33], such as being limited to single-track music generation, simplifying chords into single notes, ignoring the temporal sequence of notes, etc. This results in generated music lacking fluidity, missing key elements of rhythm, tension, and emotional expression, making melodies less appealing and lacking harmony between instruments. Establishing multi-track music generation is crucial for the intelligent development of the music domain, as depicted in Figure 1.

In multi-track music generation, selecting the appropriate electronic music score format is also crucial. Currently, the three most widely used and popular electronic music score formats include Musical Instrument Digital Interface (MIDI) format [34], MusicXML format [35], and ABC format [36], as compared in Table 1.

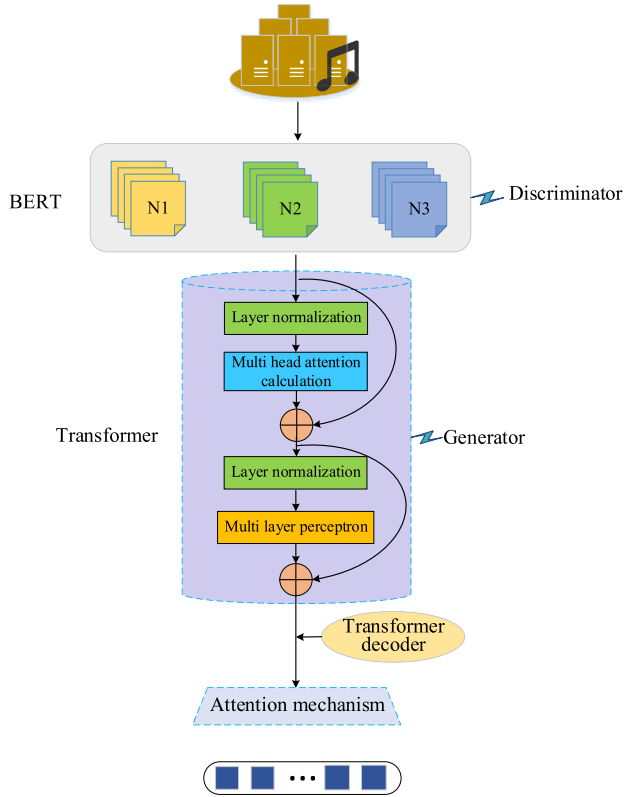
Table 1 compares the three electronic music scores. MIDI format, as the primary format for multi-track music generation, boasts small file size, high editability, strong compatibility, and native support for multi-track. Its advantage in musical expressiveness allows it to record rich musical information. Therefore, this study selects MIDI format as the input for music data, providing more creativity and expressiveness for the generated music. Additionally, by introducing deep learning algorithms, a deep learning model capable of automatically generating high-quality multi-track music is constructed, offering new tools and possibilities for the music industry and music enthusiasts.

TABLE 1. Comparison of various electronic music score formats.

Features/Format	MIDI	MusicXML	ABC
File size	Small	Middle	Small
Editorial	High	High	Middle
Compatibility	High	High	Middle
Multi-track support	Yes	Yes	No
Expressiveness	High	Middle	Low
Score display	None	Have	Have
Music production	Suitable	Suitable	Suitable
Musical education	Suitable	Suitable	Suitable
Music analysis	Suitable	Suitable	Suitable
Versatility	Widely	Standard	Specific
Client	Music producer, music software	Music educator, music score publisher	Hobbyists, specific software

**B. ANALYSIS OF THE CONSTRUCTION OF A MULTI-TRACK MUSIC GENERATION MODEL BASED ON BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS INTEGRATED WITH TRANSFORMER NETWORK**

In deep learning algorithms, Bidirectional Encoder Representations from Transformers (BERT) algorithm possesses the



**FIGURE 2.** Illustration of the model framework for multi-track music generation based on bert-fused transformer network.

capability to deeply understand the semantics and emotions of music data, accurately assessing the semantic coherence and emotional expression of generated music works [37], [38], [39]. This provides a solid foundation for the quality of generated music works. Meanwhile, the Transformer algorithm can learn the temporal relationships and structural patterns between music sequences, possessing powerful generation capabilities, making the generated music works more diverse and creative [40], [41], [42].

This study adopts the idea of GANs [43], [44], with BERT serving as the discriminator and the Transformer algorithm acting as the generator. It constructs a multi-track music generation model based on BERT integrated with Transformer network, as depicted in Figure 2.

In Figure 2, the BERT model is utilized to encode and represent music data, capturing semantic and emotional information within the music data. The BERT model algorithm employs Masked Language Model (MLM) and Span Boundary Objective (SBO) to predict missing notes in the input sequence. Initially, given a music note sequence  $S = (s_1, s_2, \dots, s_n)$  as input to the model, the model generates contextually relevant vector representations for each note as shown in Equation (1):

$$\text{enc}(s_1, s_2, \dots, s_n) = S_1, S_2, \dots, S_n \quad (1)$$

In Equation (1),  $\text{enc}(\cdot)$  denotes the quantization function,  $S$  denotes the note sequence,  $s_1, s_2, \dots, s_n$  represent  $n$  notes, and  $S_1, S_2, \dots, S_n$  represent vector representations of  $n$  notes respectively.

Subsequently, given the masking range  $(s_r, \dots, s_e) \in Y$  for the notes,  $(r, e)$  denotes the starting and ending positions, the computation of masked notes  $s_i$  is formulated as in Equation (2):

$$y_i = f(S_{r-1}, S_{e+1}, P_{i-r+1}) \quad (2)$$

In Equation (2),  $S_{r-1}$  and  $S_{e+1}$  denote boundary note vectors,  $P_{i-r+1}$  denotes the position embedding of the target note. The function  $f$  is implemented by a two-layer feedforward network, with ReLU used as the activation function. The process of implementing the masking range  $Y$  is described in Equations (3) to (5):

$$h_0 = [S_{r-1}; S_{e+1}; P_{i-r+1}] \quad (3)$$

$$h_1 = \text{LayerNorm}(\text{GeLU}(W_1 h_0)) \quad (4)$$

$$y_i = \text{LayerNorm}(\text{GeLU}(W_2 h_1)) \quad (5)$$

The vector  $y_i$  predicts the character  $s_i$ , and the loss function calculation is formulated as in Equation (6):

$$\begin{aligned} L(s_i) &= L_{MLM}(s_i) + L_{SBO}(s_i) \\ &= -\log P(s_i|S_i) - \log(s_i|y_i) \end{aligned} \quad (6)$$

Through the discriminative ability of the BERT model, experiments can more accurately assess whether the generated music conforms to the semantic and emotional characteristics of the music data.

Next, the encoded music features are input into the Transformer network to learn the temporal relationships and structural patterns between music sequences and generate new multi-track music compositions. Assuming two consecutive music segments  $c_T$  and  $c_{T+1}$ , where the  $n$ th layer hidden vector sequence of the  $T$ th segment is denoted as  $h_T^n$ . Let  $L$  denote the length of the music segment,  $d$  denote the dimension of the hidden vector, then the  $n$ th layer hidden vector sequence  $h_{T+1}^n$  of the  $T+1$  segment is obtained from Equations (7) to (8):

$$\tilde{h}_{T+1}^{n-1} = \left[ SG \left( h_T^{n-1} \cdot h_{T+1}^{n-1} \right) \right] \quad (7)$$

$$h_{T+1}^n = \text{Transformer} - \text{Layer} \left( q_{T+1}^n, k_{T+1}^n, v_{T+1}^n \right) \quad (8)$$

In Equations (7) to (8),  $n$  denotes the network layer, function  $SG(\cdot)$  denotes the stop-gradient function,  $q_{T+1}^n, k_{T+1}^n, v_{T+1}^n$  correspondingly represent the transformation matrices for query, key, and value, and  $\tilde{h}_{T+1}^{n-1}$  denotes concatenation of the two hidden vector sequences along the length direction. The calculation of attention  $\text{Attention}(Q, K, V)$  is shown in Equation (9):

$$\text{Attention}(Q, K, V) = \text{soft max} \left( \frac{Q \cdot K^T}{\sqrt{d_k}} \right) \cdot V \quad (9)$$

In Equation (9),  $Q$  refers to the matrix of “query” vectors,  $K$  refers to the matrix of “key” vectors,  $V$  refers to the matrix

```

Start
Input: MIDI format music dataset
Output: Multi track music generation
# BERT Discriminator
def encode_with_BERT(music_sequence):
    # Encode and represent music data using BERT model
    encoded_music_features = BERT_encoder(music_sequence)
    return encoded_music_features
def discriminator(encoded_music_features):
    # Evaluate the semantics and emotions of the generated music
    evaluation_result = BERT_discriminator(encoded_music_features)
    return evaluation_result
# Transformer generator
def generate_with_Transformer(encoded_music_features):
    # Learn the temporal relationships and structural patterns of music sequences using Transformer network
    learned_music_sequence = Transformer_generator(encoded_music_features)
    return learned_music_sequence
# Loss function calculation
def calculate_loss(evaluation_result, learned_music_sequence, true_music_sequence):
    # Calculate the losses for the generator and discriminator
# Complete model computation process
def model_process(true_music_sequence)
End

```

FIGURE 3. Pseudocode flow of BERT-fused transformer network applied to multi-track music generation model.

of “value” vectors.  $Attention(Q, K, V)$  is denoted by  $A_{i,j}^{rel}$ , as shown in Equation (10):

$$A_{i,j}^{rel} = E_{x_i}^T W_q^T W_{k,E} E_{x_j} + E_{x_i}^T W_q^T W_{k,R} R_{i-j} + U_i^T W_q^T W_{k,E} E_{x_j} + U_i^T W_q^T W_{k,R} R_{i-j} \quad (10)$$

In Equation (10),  $E_{x_i}, E_{x_j}$  denote word embeddings,  $U_i$  represents the position encoding of the  $i$ th character, and  $R_{i-j}$  represents the relative position of the  $j$ th character.  $E_{x_i}^T W_q^T W_{k,E} E_{x_j}$  represents content-based addressing,  $E_{x_i}^T W_q^T W_{k,R} R_{i-j}$  represents content-related positional deviation,  $U_i^T W_q^T W_{k,E} E_{x_j}$  represents global content deviation, and  $U_i^T W_q^T W_{k,R} R_{i-j}$  represents global positional deviation. Introducing a trainable vector  $u$  to replace  $U_i^T W_q^T W_{k,E} E_{x_j}$  in  $U_i^T W_q^T$  and  $v$  to replace  $U_i^T W_q^T W_{k,R} R_{i-j}$  in  $U_i^T W_q^T$ . The vectors  $u$  and  $v$  are learned. Thus, Equation (10) can be transformed into Equation (11):

$$A_{i,j}^{rel} = E_{x_i}^T W_q^T W_{k,E} E_{x_j} + E_{x_i}^T W_q^T W_{k,R} R_{i-j} + u^T W_{k,E} E_{x_j} + v^T W_{k,R} R_{i-j} \quad (11)$$

Therefore, the complete calculation process of the generator model is as described in Equations (12) to (17).

$$h_T^{n-1} = \left[ SG \left( m_T^{n-1} \cdot h_T^{n-1} \right) \right] \quad (12)$$

$$q_T^n, k_T^n, v_T^n = h_T^{n-1} W_q^{nT}, h_T^{n-1} W_{k,E}^T, h_T^{n-1} W_v^{nT} \quad (13)$$

$$A_{T,i,j}^n = q_{T,i}^n k_{T,j}^n + q_{T,i}^n W_{k,R}^n R_{i-j} + u^T k_{T,j}^n + v^T W_{k,R}^n R_{i-j} \quad (14)$$

$$a_T^n = \text{Masked} - \text{Softmax} \left( A_T^n \right) v_T^n \quad (15)$$

$$o_T^n = \text{LayerNorm} \left( \text{Linear} \left( a_T^n \right) + h_T^{n-1} \right) \quad (16)$$

$$h_T^n = \text{Positionwise} - \text{Feed} - \text{Forward} \left( o_T^n \right) \quad (17)$$

Thus, the generator part of the Transformer network is responsible for generating multi-track music sequences based on the input music features, thereby achieving creative music generation. The loss function of the BERT-fused Transformer network can be represented as Equations (18) to (20):

$$L_{gen} = -E_{S^f \sim p_\theta} \left[ D_\phi \left( S^f \right) \right] \quad (18)$$

$$L_{disc} = -E_{S^r \sim p^r} \left[ D_\phi \left( S^r \right) \right] + E_{S^f \sim p_\theta} \left[ D_\phi \left( S^f \right) \right] \quad (19)$$

$$L_{reg} = E_{\hat{S} \sim p_{\hat{X}}} \left[ \left( \left\| \nabla_{\hat{S}} D_\phi \left( \hat{S} \right) \right\|_2 - 1 \right)^2 \right] \quad (20)$$

In Equations (18) to (20),  $D_\phi(\cdot)$  denotes a one-dimensional Lipschitz function, and  $\hat{S}$  denotes the mean music character sequence between sampling points  $p^r$  and  $p_\theta$  in the embedding space.

The fusion of BERT and Transformer networks fully leverages the feature extraction capability of the BERT model on music data and the modeling capability of the Transformer network on music sequences, thereby achieving more precise and efficient multi-track music generation. This study further applies a global attention mechanism to the generation model. Under the Attention mechanism, the weighted combination of various music character elements based on their importance yields the semantic encoding  $R_i$  as shown in Equation (21):

$$R_i = \sum_{j=0}^{T_s} a_{ij} f(s_j) \quad (21)$$

In Equation (21), the parameter  $i$  denotes the moment,  $j$  denotes the  $j$ -th element in the sequence,  $T_s$  denotes the length of the sequence,  $f(\cdot)$  denotes the encoding of element  $s_j$ , and  $a_{ij}$  denotes the weight, reflecting the importance of element



$s_j$  to the semantic encoding  $R_i$ . The calculation of  $a_{ij}$  is shown in Equation (22):

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_s} \exp(e_{ik})} \quad (22)$$

In Equation (22),  $e_{ij}$  reflects the matching degree between the element to be encoded and other elements; the higher the matching degree, the greater the influence of the element.

In the model constructed in this study, the specific pseudocode of the BERT-fused Transformer network applied to the multi-track music generation model is illustrated in Figure 3.

## IV. EXPERIMENTAL DESIGN AND PERFORMANCE EVALUATION

### A. DATASETS COLLECTION

The data for this study is sourced from The Lakh Pianoroll Dataset ([https://opendatalab.com/OpenDataLab/Lakh\\_Pianoroll\\_Dataset](https://opendatalab.com/OpenDataLab/Lakh_Pianoroll_Dataset)). This dataset is a derivative version of the Lakh MIDI Dataset, represented in piano roll format, containing 174,154 piano roll files with multiple instrument tracks. In the dataset selected for this study, each multi-track piano roll is compressed into 5 tracks: bass, drums, guitar, piano, and violin.

### B. EXPERIMENTAL ENVIRONMENT

In the experimental environment, a 64-bit Windows 10 operating system is used, and the software is built using the TensorFlow 2.1 framework. Python 3.6 is used for data preprocessing and algorithm implementation details. Additionally, an NVIDIA GeForce RTX 2060 GPU with 16GB of memory is utilized.

### C. PARAMETERS SETTING

For the neural network models constructed in this study, the following hyperparameters need to be set: The discriminator BERT consists of 5 one-dimensional convolutional layers and one fully connected layer, with ReLU used as the activation function. The number of convolutional kernels in each layer of the encoder is limited to 16 to compress the representations of inter-track dependencies. The number of iterations is set to 100, the optimizer is Adam, and the initial learning rate is set to 0.001.

### D. PERFORMANCE EVALUATION

To evaluate the performance of the model proposed in this study, the algorithm is compared with BERT [45], Transformer [46], GAN [47], and the model algorithm proposed by Li et al. [29] from related fields. Evaluation is conducted based on metrics such as accuracy, precision, recall, F1 score, and Num Chroma Used (UPC). UPC refers to the number of different pitch classes contained in each measure of a music sample, ranging from 0 to 12.

Firstly, the comparison results of accuracy, precision, recall, and F1 score under each algorithm are shown in Figures 4 to 7.

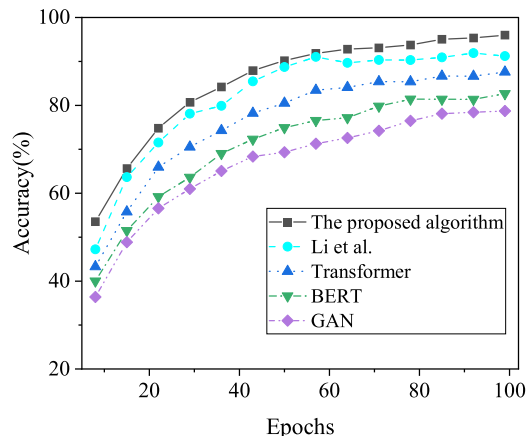


FIGURE 4. Variation of music generation prediction accuracy with iteration cycles under different algorithms.

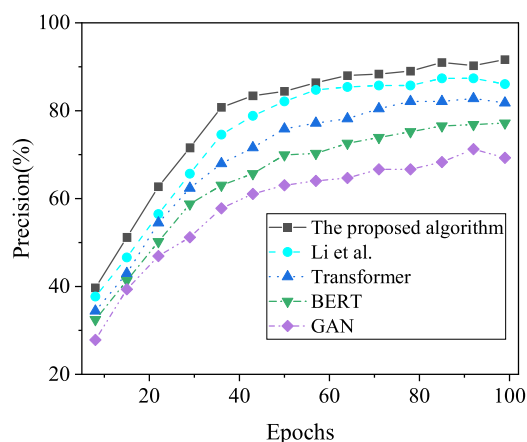


FIGURE 5. Variation of music generation prediction precision with iteration cycles under different algorithms.

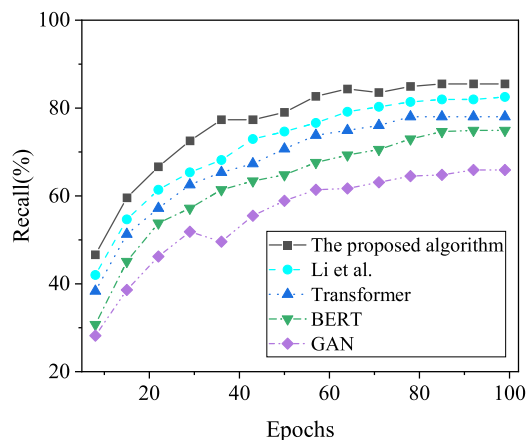


FIGURE 6. Variation of music generation prediction recall with iteration cycles under different algorithms.

In Figures 5 to 7, with the increase in iteration cycles, the accuracy, precision, recall, and F1 score of the model algorithm proposed in this study, as well as those of BERT, Transformer, GAN, and the model algorithm proposed by Li et al. [29] from related fields, show a trend of initially

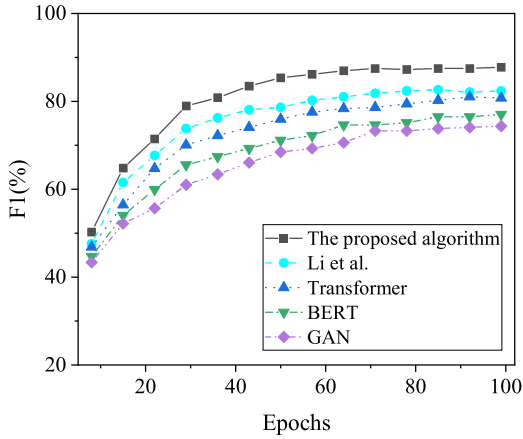


FIGURE 7. Variation of music generation prediction F1 score with iteration cycles under different algorithms.

increasing and then stabilizing. Compared with other algorithms, the music generation prediction accuracy of the model proposed in this study reaches an Accuracy value of 95.98%, representing a minimum improvement of 4.77% in generation accuracy. Moreover, the prediction accuracy of each algorithm for multi-track music generation ranks from highest to lowest as follows: the model algorithm constructed in this study > the algorithm proposed by Li et al. [29] > Transformer > BERT > GAN. Further analysis of the Precision, Recall, and F1 score results of each algorithm shows that the prediction results of the model algorithm proposed in this study all exceed 85.49%, with an improvement in prediction accuracy exceeding 3%. Therefore, the multi-track music generation model based on BERT-fused Transformer network proposed in this study can accurately predict multi-track music, thereby achieving more precise and efficient multi-track music generation.

The UPC results of each algorithm are compared as shown in Figure 8.

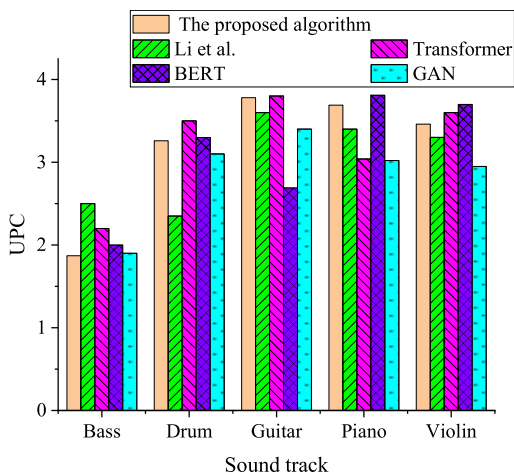


FIGURE 8. UPC results under different algorithms.

In Figure 8, the algorithm proposed in this study demonstrates significant advantages in the UPC values of various

instruments. Particularly, in the drums and guitar instruments, the UPC values of the algorithm in this study are notably higher than those of other algorithms, reaching 3.26 and 3.78 respectively. These data highlight the excellent performance of the model proposed in this study in generating multi-track music. Furthermore, the UPC values of piano and violin instruments are also slightly higher than those of other algorithms, indicating the diversity and complexity of the music generated by the algorithm in this study for these instruments. Although the UPC value of the bass instrument for the algorithm in this study is slightly lower than that of other algorithms, it still remains at a relatively low level, consistent with the characteristics of bass instrument music. Overall, the algorithm proposed in this study exhibits outstanding performance in terms of pitch accuracy, making valuable contributions to research and practice in the field of multi-track music generation.

### E. DISCUSSION

Through evaluating the performance of our model, a comparison with the model algorithms proposed by BERT, Transformer, GAN, and Li et al. [29] reveals that with the increase in iteration cycles, the music generation prediction accuracy of our model algorithm reached an Accuracy value of 95.98%, representing a minimum improvement of 4.77% compared to other algorithms (BERT, Transformer, GAN, and the model algorithm proposed by Li et al. [29]). Thus, the performance of the model proposed in this study in the field of multi-track music generation has been validated, consistent with the findings of Kang et al. [48] and Wu et al. [49]. Furthermore, through further comparison of the UPC results of each algorithm, since the piano and guitar instruments usually play chords and often express emotions with melodies with larger pitch ranges in music, the UPC values of the algorithm proposed in this study are more concentrated across various instruments. Moreover, since the bass tends to play lower tones, the UPC values of the model in this study demonstrate significant advantages, consistent with the views of Tang et al. [50].

Thus, the model proposed in this study provides valuable references and insights for research and practice in the field of multi-track music generation. By combining BERT and Transformer networks and adopting innovative algorithm design, the model in this study not only improves the accuracy and efficiency of music generation but also expands the understanding and application of music generation technology. This is of great significance and value for advancing the development of music generation technology, enhancing the quality and diversity of music creation, and enriching the content and forms of the music industry.

## V. CONCLUSION

### A. RESEARCH CONTRIBUTION

This study addresses the prevailing tendency in music generation to adopt simplified strategies and proposes a multi-track

music generation model based on BERT-fused Transformer networks. This model combines two advanced deep learning technologies to efficiently encode and generate music data. Through comparative experiments with other algorithms, the study validates that the proposed model achieves an accuracy of over 95% in music generation prediction and demonstrates significant advantages in predicting pitch accuracy. The data indicates the significant contribution and importance of the proposed model in the field of multi-track music generation.

## B. FUTURE WORKS AND RESEARCH LIMITATIONS

However, this study also has certain limitations. Firstly, although the model exhibits good performance in experiments, there are still prediction errors under specific circumstances, possibly due to the characteristics of the music data and limitations in model design. Secondly, the experimental dataset of this study may lack generalizability in specific domains due to the diversity and complexity of music data. Therefore, future studied will explore more advanced deep learning technologies and attempt to integrate research findings from other domains to further enhance the quality and diversity of music generation. Additionally, efforts will be made to construct richer and more diverse music datasets to validate and assess the generalizability and robustness of the model, bringing more innovation and opportunities to music creation and industry.

## REFERENCES

- [1] M. Alfaro-Contreras, J. M. Iñesta, and J. Calvo-Zaragoza, "Optical music recognition for homophonic scores with neural networks and synthetic music generation," *Int. J. Multimedia Inf. Retr.*, vol. 12, no. 1, p. 12, Jun. 2023.
- [2] S.-S. Weng and H.-C. Chen, "Exploring the competitive advantages of an innovative online music production framework combined with deep learning," *Int. J. Electron. Commerce Stud.*, vol. 13, no. 1, p. 01, Sep. 2021.
- [3] Y.-J. Shih, S.-L. Wu, F. Zalkow, M. Muller, and Y.-H. Yang, "Theme transformer: Symbolic music generation with theme-conditioned transformer," *IEEE Trans. Multimedia*, vol. 25, pp. 1–16, 2022.
- [4] C. Jin, T. Wang, X. Li, C. J. J. Tie, Y. Tie, S. Liu, M. Yan, Y. Li, J. Wang, and S. Huang, "A transformer generative adversarial network for multi-track music generation," *CAAI Trans. Intell. Technol.*, vol. 7, no. 3, pp. 369–380, Sep. 2022.
- [5] E. Deruty, M. Grachten, S. Lattner, J. Nistal, and C. Aouameur, "On the development and practice of AI technology for contemporary popular music production," *Trans. Int. Soc. Music Inf. Retr.*, vol. 5, no. 1, p. 35, Feb. 2022.
- [6] R. Brøvig-Hanssen and E. Jones, "Remix's retreat? Content moderation, copyright law and mashup music," *New Media Soc.*, vol. 25, no. 6, pp. 1271–1289, Jun. 2023.
- [7] V. Chaturvedi, A. B. Kaur, V. Varshney, A. Garg, G. S. Chhabra, and M. Kumar, "Music mood and human emotion recognition based on physiological signals: A systematic review," *Multimedia Syst.*, vol. 28, no. 1, pp. 21–44, Feb. 2022.
- [8] C. Zhu, Z. Xu, C. Hou, X. Lv, S. Jiang, D. Ye, and Y. Huang, "Flexible, monolithic piezoelectric sensors for large-area structural impact monitoring via MUSIC-assisted machine learning," *Structural Health Monitor.*, vol. 23, no. 1, pp. 121–136, Jan. 2024.
- [9] M. Majidi and R. M. Toroghi, "A combination of multi-objective genetic algorithm and deep learning for music harmony generation," *Multimedia Tools Appl.*, vol. 82, no. 2, pp. 2419–2435, Jan. 2023.
- [10] G. Keerti, A. N. Vaishnavi, P. Mukherjee, A. S. Vidya, G. S. Sreenithya, and D. Nayab, "Attentional networks for music generation," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 5179–5189, Feb. 2022.
- [11] M. Civit, J. Civit-Masot, F. Cuadrado, and M. J. Escalona, "A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends," *Expert Syst. Appl.*, vol. 209, Dec. 2022, Art. no. 118190.
- [12] R. Ramirez, E. Maestre, and X. Serra, "A rule-based evolutionary approach to music performance modeling," *IEEE Trans. Evol. Comput.*, vol. 16, no. 1, pp. 96–107, Feb. 2012.
- [13] K. Hastuti, A. Azhari, A. Musdholifah, and R. Supanggih, "Rule-based and genetic algorithm for automatic gamelan music composition," *Int. Rev. Model. Simulations (IREMOS)*, vol. 10, no. 3, p. 202, Jun. 2017.
- [14] K. Hastuti, P. N. Andono, G. F. Shidik, E. Noersangko, and A. M. Syarif, "Gamelan composer: A rule-based interactive melody generator for gamelan music," *Int. J. Eng. Appl. (IREA)*, vol. 8, no. 4, p. 148, Jul. 2020.
- [15] L. Wang, Z. Zhao, H. Liu, J. Pang, Y. Qin, and Q. Wu, "A review of intelligent music generation systems," *Neural Comput. Appl.*, vol. 36, no. 12, pp. 6381–6401, Apr. 2024.
- [16] I. Goienetxea, I. Mendialdua, I. Rodríguez, and B. Sierra, "Statistics-based music generation approach considering both rhythm and melody coherence," *IEEE Access*, vol. 7, pp. 183365–183382, 2019.
- [17] N. Zhang, "Learning adversarial transformer for symbolic music generation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1754–1763, Apr. 2023.
- [18] W. Liu, "Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition," *J. Supercomput.*, vol. 79, no. 6, pp. 6560–6582, Apr. 2023.
- [19] R. P. Joy, M. R. Thanka, J. P. M. Dhas, and E. B. Edwin, "Music mood based recognition system based on machine learning and deep learning," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 2, pp. 904–911, Sep. 2023.
- [20] Y. Sun, "Analysis of the principle and application of machine learning for music composition," *Highlights Sci., Eng. Technol.*, vol. 85, pp. 556–562, Mar. 2024.
- [21] Y. Gonzalez and R. C. Prati, "Similarity of musical timbres using FFT-acoustic descriptor analysis and machine learning," *Eng.*, vol. 4, no. 1, pp. 555–568, Feb. 2023.
- [22] T. Kuremoto, "Guqing music recognition by machine learning methods," *Impact*, vol. 2024, no. 1, pp. 40–42, Jan. 2024.
- [23] S. Ji, X. Yang, and J. Luo, "A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges," *ACM Comput. Surv.*, vol. 56, no. 1, pp. 1–39, Jan. 2024.
- [24] Z. Yin, F. Reuben, S. Stepney, and T. Collins, "Deep learning's shallow gains: A comparative evaluation of algorithms for automatic music generation," *Mach. Learn.*, vol. 112, no. 5, pp. 1785–1822, May 2023.
- [25] Y. Guo, Y. Liu, T. Zhou, L. Xu, and Q. Zhang, "An automatic music generation and evaluation method based on transfer learning," *PLoS One*, vol. 18, no. 5, May 2023, Art. no. e0283103.
- [26] S. Sharma and S. Bvuma, "Generative adversarial networks (GANs) for creative applications: Exploring art and music generation," *Int. J. Multi-disciplinary Innov. Res. Methodol.*, vol. 2, no. 4, pp. 29–33, 2023.
- [27] P. Ferreira, R. Limongi, and L. P. Fávero, "Generating music with data: Application of deep learning models for symbolic music composition," *Appl. Sci.*, vol. 13, no. 7, p. 4543, Apr. 2023.
- [28] L. Moysis, L. A. Iliadis, S. P. Sotiroudis, A. D. Boursianis, M. S. Papadopoulou, K. D. Kokkinidis, C. Volos, P. Sarigiannidis, S. Nikolaidis, and S. K. Goudos, "Music deep learning: Deep learning methods for music signal processing—A review of the state-of-the-art," *IEEE Access*, vol. 11, pp. 17031–17052, 2023.
- [29] P. Li, T. M. Liang, Y. M. Cao, X. M. Wang, X. J. Wu, and L. Y. Lei, "A novel Xi'an drum music generation method based on Bi-LSTM deep reinforcement learning," *Appl. Intell.*, vol. 54, no. 1, pp. 80–94, Aug. 2024.
- [30] Y. Ghatas, M. Fayek, and M. Hadhoud, "A hybrid deep learning approach for musical difficulty estimation of piano symbolic music," *Alexandria Eng. J.*, vol. 61, no. 12, pp. 10183–10196, Dec. 2022.
- [31] T. S. Ahmedovich, "Fundamentals of the use of inavatory programs of music culture in teaching as a subject," *Asia Pacific J. Marketing Manage. Rev.*, vol. 11, no. 11, pp. 171–174, 2022.
- [32] M. Blaszkę and B. Kostek, "Musical instrument identification using deep learning approach," *Sensors*, vol. 22, no. 8, p. 3033, Apr. 2022.
- [33] G. Song and Z. Wang, "An efficient hidden Markov model with periodic recurrent neural network observer for music beat tracking," *Electronics*, vol. 11, no. 24, p. 4186, Dec. 2022.
- [34] S. Shen and K. Wu, "Solfeggio teaching method based on MIDI technology in the background of digital music teaching," *Int. J. Web-Based Learn. Teach. Technol.*, vol. 18, no. 1, pp. 1–18, Sep. 2023.



- [35] W. Seeyo, W. Seekhunlio, and S. Chuangprakhon, "Bridging Thai music notation to western music scores through innovative conversion and evaluation," *Multidisciplinary Sci. J.*, vol. 6, no. 5, Nov. 2023, Art. no. 2024073.
- [36] J. Yang, "Musi-ABC for predicting musical emotions," *IEEE Access*, vol. 11, pp. 79455–79465, 2023.
- [37] S. Li and Y. Sung, "MRBERT: Pre-training of melody and rhythm for automatic music generation," *Mathematics*, vol. 11, no. 4, p. 798, Feb. 2023.
- [38] A. H. Oliiae, S. Das, J. Liu, and M. A. Rahman, "Using bidirectional encoder representations from transformers (BERT) to classify traffic crash severity types," *Natural Lang. Process. J.*, vol. 3, Jun. 2023, Art. no. 100007.
- [39] H. Wang, S. Hao, C. Zhang, X. Wang, and Y. Chen, "Motif transformer: Generating music with motifs," *IEEE Access*, vol. 11, pp. 63197–63204, 2023.
- [40] L. B. Cesar, M. Manso-Callejo, and C. I. Cira, "BERT (bidirectional encoder representations from transformers) for missing data imputation in solar irradiance time series," *Eng. Proc.*, vol. 39, no. 1, p. 26, 2023.
- [41] A. Areshey and H. Mathkour, "Transfer learning for sentiment classification using bidirectional encoder representations from transformers (BERT) model," *Sensors*, vol. 23, no. 11, p. 5232, May 2023.
- [42] W. Wang, J. Li, Y. Li, and X. Xing, "Style-conditioned music generation with transformer-GANs," *Frontiers Inf. Technol. Electron. Eng.*, vol. 25, no. 1, pp. 106–120, 2024.
- [43] Y. Yu, Z. Zhang, W. Duan, A. Srivastava, R. Shah, and Y. Ren, "Conditional hybrid GAN for melody generation from lyrics," *Neural Comput. Appl.*, vol. 35, no. 4, pp. 3191–3202, Feb. 2023.
- [44] J. Huang, X. Huang, L. Yang, and Z. Tao, "Dance-conditioned artistic music generation by creative-GAN," *IEICE Trans. Fundamentals Electron., Commun. Comput. Sci.*, vol. 107, no. 5, pp. 836–844, 2024.
- [45] A. S. Sams and A. Zahra, "Multimodal music emotion recognition in Indonesian songs based on CNN-LSTM, XLNet transformers," *Bull. Electr. Eng. Informat.*, vol. 12, no. 1, pp. 355–364, Feb. 2023.
- [46] Y. Qin, H. Xie, S. Ding, B. Tan, Y. Li, B. Zhao, and M. Ye, "Bar transformer: A hierarchical model for learning long-term structure and generating impressive pop music," *Int. J. Speech Technol.*, vol. 53, no. 9, pp. 10130–10148, May 2023.
- [47] W. Huang and F. Zhan, "A novel probabilistic diffusion model based on the weak selection mimicry theory for the generation of hypnotic songs," *Mathematics*, vol. 11, no. 15, p. 3345, Jul. 2023.
- [48] J. Kang, S. Poria, and D. Herremans, "Video2Music: Suitable music generation from videos using an affective multimodal transformer model," *Expert Syst. Appl.*, vol. 249, no. 4, Sep. 2024, Art. no. 123640.
- [49] G. Wu, S. Liu, and X. Fan, "The power of fragmentation: A hierarchical transformer model for structural segmentation in symbolic music generation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, no. 1, pp. 1409–1420, Jun. 2023.
- [50] H. Tang, Y. Zhang, and Q. Zhang, "The use of deep learning-based intelligent music signal identification and generation technology in national music teaching," *Frontiers Psychol.*, vol. 13, Jun. 2022, Art. no. 762402.



**RONG JIANG** was born in Longyan, Fujian, China, in 1990. She received the bachelor's and master's degrees from China Conservatory of Music, and the master's degree from the University of York, U.K. She is currently pursuing the Ph.D. degree with the College of Creative Arts, UiTM, Malaysia. Her research interests include opera production, music production, and big data analysis.



**XIAOFEI MOU** was born in Qingdao, Shandong, China, in 1986. He received the bachelor's and master's degrees in arts management and the Ph.D. degree in musicology from China Conservatory of Music. He is currently working as an Associate Professor of arts management with China Conservatory of Music. His research interests include music history, arts management, and music production.

...