International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 06 | June – 2026 DOI:10.55041/ISJEM07601
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Symbolic Music Generation using a Variational Autoencoder and LSTM-Based Sequence Modeling

Ketan Kanjiya[1], Piyush Sonani[2], Upendrasinh Zala[3]

[1]*Chief Research Officer, Kshatrainfotech Pvt Ltd, Ahmedabad, Gujarat, India*
[2]*Chief Technology Officer, Kshatrainfotech Pvt Ltd, Ahmedabad, Gujarat, India*
[3]*Chief Executive Officer, Kshatrainfotech Pvt Ltd, Ahmedabad, Gujarat, India*

Abstract - Symbolic music generation has become an important application of deep learning, enabling computational models to learn musical patterns and generate new compositions directly from data. This paper presents a Variational Autoencoder and Long Short-Term Memory based framework for learning latent representations of symbolic music and generating coherent musical sequences. The proposed model is trained on piano-roll representations derived from the Nottingham MIDI dataset, where latent embeddings capture underlying melodic and temporal structures. The effectiveness of the model is evaluated through reconstruction metrics and latent-space analysis. Experimental results demonstrate stable training behaviour, strong reconstruction performance with an F1-score of 84.0%, and a well-organized latent space that supports meaningful music generation. The findings indicate that the proposed VAE-LSTM architecture effectively learns musical representations and can generate diverse symbolic music while preserving important structural characteristics of the training data.

Key Words: Music Generation, Variational Autoencoder (VAE), Long Short-Term Memory (LSTM), Latent Representation Learning, Deep Learning, MIDI

1. INTRODUCTION

Music is widely regarded as a universal form of artistic expression that communicates ideas, emotions, and cultural identity through melody, rhythm, and harmony. Throughout history, composers have relied on creativity, theory, and experience to craft musical compositions. The concept of algorithmic composition, which involves generating music through formal rules or computational processes, has existed for centuries. Early examples include rule-based musical systems and probability-driven techniques designed to automate parts of the compositional process. With the emergence of modern computing, these approaches evolved into computational models capable of generating musical sequences using mathematical and statistical techniques.

Recent advances in artificial intelligence and deep learning have significantly expanded the capabilities of algorithmic music generation. Instead of relying on manually designed rules, modern systems can learn musical structures directly from large datasets. Deep generative models have demonstrated strong potential for capturing patterns in musical compositions and generating novel melodies that resemble human-composed music. These developments have led to growing interest in applying machine learning techniques to music generation tasks.

Symbolic music representation provides an efficient and structured way to model musical information for computational analysis. In symbolic formats such as MIDI, music is represented as sequences of musical events including pitch, timing, and duration rather than raw audio signals. This representation reduces the complexity of the learning problem while preserving essential musical structure. A commonly used representation derived from MIDI data is the piano-roll format, which represents music as a time pitch matrix suitable for neural network processing.

Among deep generative approaches, Variational Autoencoders (VAEs) have gained attention for their ability to learn meaningful latent representations of complex data distributions. A VAE consists of an encoder that maps input data into a continuous latent space and a decoder that reconstructs data from this representation. By learning structured latent embeddings, VAEs enable the generation of new samples through latent vector sampling and interpolation. Since musical compositions are inherently sequential, models capable of capturing temporal dependencies are essential. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, are widely used for modeling sequential data due to their ability to capture long-range dependencies.

In this work, we propose a deep generative framework for symbolic music generation using a Variational Autoencoder with a bidirectional LSTM encoder and an autoregressive LSTM decoder. The model is trained on the Nottingham MIDI dataset using piano-roll representations of musical sequences. Experimental evaluation includes reconstruction quality metrics, latent space visualization, and statistical comparisons between original and generated music. The results demonstrate that the proposed VAE-LSTM architecture can effectively learn latent musical representations and generate coherent symbolic music sequences.

2. LITERATURE REVIEW

The evolution of algorithmic composition has transitioned from early rule-based systems and formal methods, such as those proposed by Guido of Arezzo in 1025 and Ada Lovelace in 1842, to contemporary deep learning paradigms [1-2]. In recent years, symbolic music generation has emerged as a prominent subfield of artificial intelligence, focusing on representing music as sequences of symbols, such as MIDI, in order to model musical structures, patterns, and genre characteristics [3-4]. This paradigm shift has been driven by the ability of deep neural networks to learn complex data distributions directly from large datasets without requiring extensive manual feature engineering, while automated symbolic music labeling pipelines have further supported the creation of large annotated MIDI datasets for training deep learning models [5].

Recurrent Neural Networks (RNNs) and their variants have historically been among the most widely used models for generating musical sequences due to their inherently sequential nature [6]. Early RNN-based approaches, such as the monophonic music generator proposed by Todd, laid the foundation for sequential prediction in music generation tasks [7]. However, conventional RNN architectures often suffer from the vanishing gradient problem, which limits their ability to capture long-term dependencies in extended musical compositions [8]. To address these limitations, Long Short-Term Memory (LSTM) networks were introduced. LSTM architectures employ specialized gating mechanisms including input, forget, and output gates to regulate the flow of information and retain relevant contextual knowledge over longer time intervals [9-10]. Notable implementations include Google's MelodyRNN, which improved the modeling of long-term musical structures through variants such as lookback and attention-based RNN models [7]. Cross-modal approaches such as Dance2MIDI have also explored generating multi-instrument MIDI music directly from dance video inputs [11].

While LSTMs are effective at maintaining temporal coherence in sequences, Variational Autoencoders (VAEs) provide a principled probabilistic framework for learning compact latent representations of musical data [12]. A VAE typically consists of an encoder that maps input sequences to a latent distribution commonly assumed to follow a standard Gaussian and a decoder that reconstructs the original sequence from sampled latent vectors [12]. This architecture is particularly useful for music generation because it organizes similar musical patterns within a continuous latent space, enabling the generation of diverse musical outputs through latent space sampling and interpolation [13]. A notable example is MusicVAE, which employs a hierarchical decoder structure to address the issue of posterior collapse and to capture multi-scale musical dependencies [14].

The integration of recurrent networks within a VAE framework, forming a VAE-RNN architecture, combines the temporal modeling capability of RNNs with the expressive latent representations learned by VAEs [15]. For example, MIDI-VAE utilizes parallel recurrent encoder-decoder networks that share a common latent space, enabling tasks such as musical style transfer between genres including classical and jazz [7]. Similarly, the MGU-V framework demonstrates that hybrid VAE architectures can achieve strong performance in generating realistic musical sequences by capturing both short-term temporal dynamics and global musical structures [15]. Furthermore, hierarchical RNN-VAE and diffusion-based models have been proposed to balance diversity, coherence, and multimodal consistency across longer musical segments [16].

Recent research has also explored controllable music generation by conditioning latent representations or large language models on specific attributes such as musical style, composer identity, emotional characteristics, and textual prompts [17-18]. To address the subjective nature of musical affect, many approaches employ the Valence-Arousal (VA) dimensional model to provide fine-grained emotional control over generated compositions, including emotion-guided image-to-music generation frameworks based on CNN-Transformer architectures [19]. For instance, models such as the Multimodal Emotion-guided Music Generation Model (MEMGM) project affective signals into a shared latent space to guide the generation of music aligned with particular emotional trajectories [20]. Despite these advances, generating music that simultaneously maintains local rhythmic accuracy and

global structural coherence remains a challenging problem [21]. This study contributes to this area by refining the learning of latent representations through a VAE-LSTM architecture designed to improve both structural consistency and generative diversity in symbolic music generation.

3. DATASET

The experiments in this study were conducted using the Nottingham MIDI Dataset, a widely used benchmark dataset for symbolic music generation. The dataset contains approximately 1,036 MIDI files comprising traditional British and American folk melodies represented in symbolic MIDI format. MIDI files encode musical information such as note pitches, durations, velocities, and timing events, providing a structured representation of musical sequences suitable for machine learning applications. Unlike raw audio recordings, symbolic MIDI data explicitly represents musical events and relationships, making it particularly effective for modeling melodic and temporal patterns. The dataset includes compositions of varying lengths and musical structures, providing diverse musical sequences for training and evaluation.

4. METHODOLOGY

4.1 SYSTEM OVERVIEW

The proposed framework utilizes a Variational Autoencoder–Long Short-Term Memory (VAE-LSTM) architecture for symbolic music generation. The system consists of four main stages: data preprocessing, latent representation learning, sequence reconstruction, and music generation. MIDI files are first converted into piano-roll representations and segmented into fixed-length sequences. These sequences are provided to a bidirectional LSTM encoder, which extracts temporal features and maps them to a latent probability distribution defined by mean and variance vectors. Latent vectors are sampled using the reparameterization technique and passed to an LSTM decoder for sequence reconstruction. During training, the model optimizes a combined reconstruction and KL divergence loss to learn meaningful latent representations. After training, new musical sequences are generated by sampling latent vectors and decoding them into piano-rolls, which are then converted into MIDI files.

4.2 DATA PREPROCESSING

Prior to training, all MIDI files were transformed into piano-roll representations to provide a structured input format suitable for sequence modeling. The conversion process discretized musical events into a temporal resolution of four time steps per beat, where each time step represents the activation state of notes within the selected pitch range. To ensure consistency across samples, only pitches between C3 (MIDI 48) and C6 (MIDI 84) were retained, resulting in a 36-dimensional binary representation at each time step. Percussion tracks were excluded from processing to focus exclusively on melodic information. The generated piano rolls were segmented into fixed-length sequences of 32 time steps, corresponding to approximately two measures of music at the chosen resolution. Silent segments containing no active notes were discarded. Finally, the processed sequences were randomly divided into training and validation sets using 85:15 ratio for model development and evaluation.

4.3 PROPOSED VAE-LSTM ARCHITECTURE

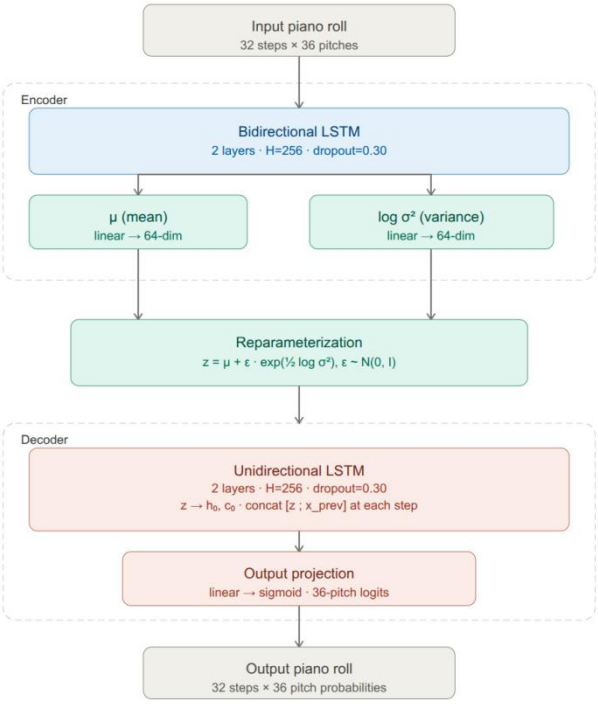

Figure 1: Overall architecture of the proposed VAE-LSTM music generation model.

The proposed model combines the sequence modeling capability of Long Short-Term Memory (LSTM) networks with the latent representation learning framework of Variational Autoencoders (VAEs) to generate symbolic music. The architecture consists of three main components: a bidirectional LSTM encoder, a latent representation module, and an LSTM-based decoder. As illustrated in Figure 1, the encoder transforms an input piano-roll sequence into a compact latent representation, which is subsequently sampled and decoded to reconstruct the original sequence or generate new musical patterns.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 06 | June – 2026 DOI:10.55041/ISJEM07601
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

4.3.1 BIDIRECTIONAL LSTM ENCODER

The encoder is designed to capture temporal dependencies present in musical sequences. A bidirectional LSTM processes the input piano-roll segment in both forward and backward directions, enabling the model to utilize contextual information from past and future time steps simultaneously. Each input segment consists of 32 time steps, where every step is represented by a 36-dimensional binary pitch vector.

The final hidden states from the forward and backward LSTM layers are concatenated to form a comprehensive sequence representation. This representation is then projected through two fully connected layers to estimate the mean vector (μ) and logarithmic variance vector ($\log\sigma^2$) of the latent probability distribution. These parameters characterize the latent space from which musical representations are sampled.

4.3.2 LATENT SPACE LEARNING AND REPARAMETERIZATION

To learn a continuous and structured latent representation, the encoder outputs are modeled as a Gaussian distribution parameterized by μ and $log\,\sigma^2$. During training, latent vectors are generated using the reparameterization technique:

$$z = \mu + \sigma \bullet \epsilon, \quad \epsilon \sim N(0, I) \quad (1)$$

where z denotes the latent vector and ϵ represents random noise sampled from a standard normal distribution. This formulation enables gradient propagation through the sampling process and allows the model to learn meaningful latent representations. The latent space captures high-level musical characteristics and facilitates the generation of diverse musical sequences through latent vector sampling.

4.3.3 LSTM DECODER

The decoder reconstructs musical sequences from the sampled latent vector z. The latent representation is first transformed into the initial hidden and cell states of a multi-layer LSTM decoder. During training, teacher forcing is employed, where the previous ground-truth note vector is provided as input at each decoding step. The decoder additionally receives the latent vector at every time step, allowing global musical information to influence sequence generation throughout the reconstruction process.

The decoder produces pitch activation probabilities for each time step, which are converted into binary piano-roll representations using a predefined threshold. During inference, latent vectors sampled from the learned distribution are decoded autoregressively to generate new symbolic music sequences, which are subsequently converted into MIDI format for playback and evaluation.

4.4 LOSS FUNCTION AND TRAINING STRATEGY

The proposed VAE-LSTM model is trained by minimizing a composite objective function consisting of a reconstruction loss and a Kullback–Leibler (KL) divergence regularization term. The total loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \beta\mathcal{L}_{KL} \quad (2)$$

where $\mathcal{L}_{recon}$ measures the quality of sequence reconstruction, $\mathcal{L}_{KL}$ regularizes the latent space distribution, and β controls the contribution of the KL term during training.

4.4.1 RECONSTRUCTION LOSS

Since piano-roll representations consist of binary note activation values, reconstruction quality is evaluated using Binary Cross-Entropy (BCE) loss. Let x denote the original piano-roll sequence and $\hat{x}$ the reconstructed output. The reconstruction loss is computed as:

$$\mathcal{L}_{recon} = -\frac{1}{N}\sum_{i=1}^{N}[x_i \log(\hat{x}_i) + (1 - x_i)\log(1 - \hat{x}_i)] \quad (3)$$

where N represents the total number of note activations in a sequence. To compensate for the sparsity of musical data, positive note activations are assigned higher importance through weighted BCE loss, encouraging accurate reconstruction of active notes.

4.4.2 KL DIVERGENCE LOSS

To ensure that the learned latent representations follow a standard Gaussian distribution, KL divergence regularization is applied between the encoder distribution q(z|x) and the prior distribution p(z). The KL loss is defined as:

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum_{j=1}^{d}(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2) \quad (4)$$

where d denotes the latent dimension, while μ and σ represent the mean and standard deviation vectors generated by the encoder. This regularization promotes a smooth and continuous latent space, enabling meaningful

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 06 | June – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI:10.55041/ISJEM07601

interpolation and diverse music generation through latent sampling.

4.4.3 KL ANNEALING AND OPTIMIZATION

Direct optimization of the full VAE objective can lead to posterior collapse, where the decoder ignores latent representations and relies primarily on sequence modeling. To mitigate this issue, a KL annealing strategy is employed, gradually increasing the weighting factor β from 0 to 1 during the initial training phase. This allows the model to first learn accurate sequence reconstruction before enforcing latent space regularization.

The model is optimized using the Adam optimizer with a learning rate of 3×10^{-4} and weight decay of 10^{-4}. Gradient clipping is applied to stabilize training, while dropout regularization is incorporated within the LSTM layers to reduce overfitting. Additionally, an 85:15 training-validation split, adaptive learning rate scheduling, model checkpointing, and early stopping are utilized to improve generalization performance and training stability.

4.5 EVALUATION METRICS

The performance of the proposed VAE-LSTM model was evaluated using both reconstruction-based and statistical measures. Reconstruction quality was assessed through note-level Precision, Recall, and F1-score by comparing reconstructed piano-roll sequences with their corresponding ground truth inputs. Precision measures the proportion of correctly predicted note activations, Recall quantifies the ability of the model to recover actual note events, and the F1-score provides a balanced assessment of both metrics.

In addition to reconstruction accuracy, the learned latent space was analyzed using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). These visualization techniques were employed to examine the organization and continuity of latent representations learned by the model.

4.5.1 TRAINING DYNAMICS

Figure 2 illustrates the evolution of the total loss, reconstruction loss, and KL divergence during training and validation. The loss curves exhibit a rapid decrease during the initial training epochs, followed by a gradual convergence phase, indicating stable optimization of the proposed VAE-LSTM architecture. The close alignment between training and validation losses throughout training suggests effective generalization and limited overfitting.

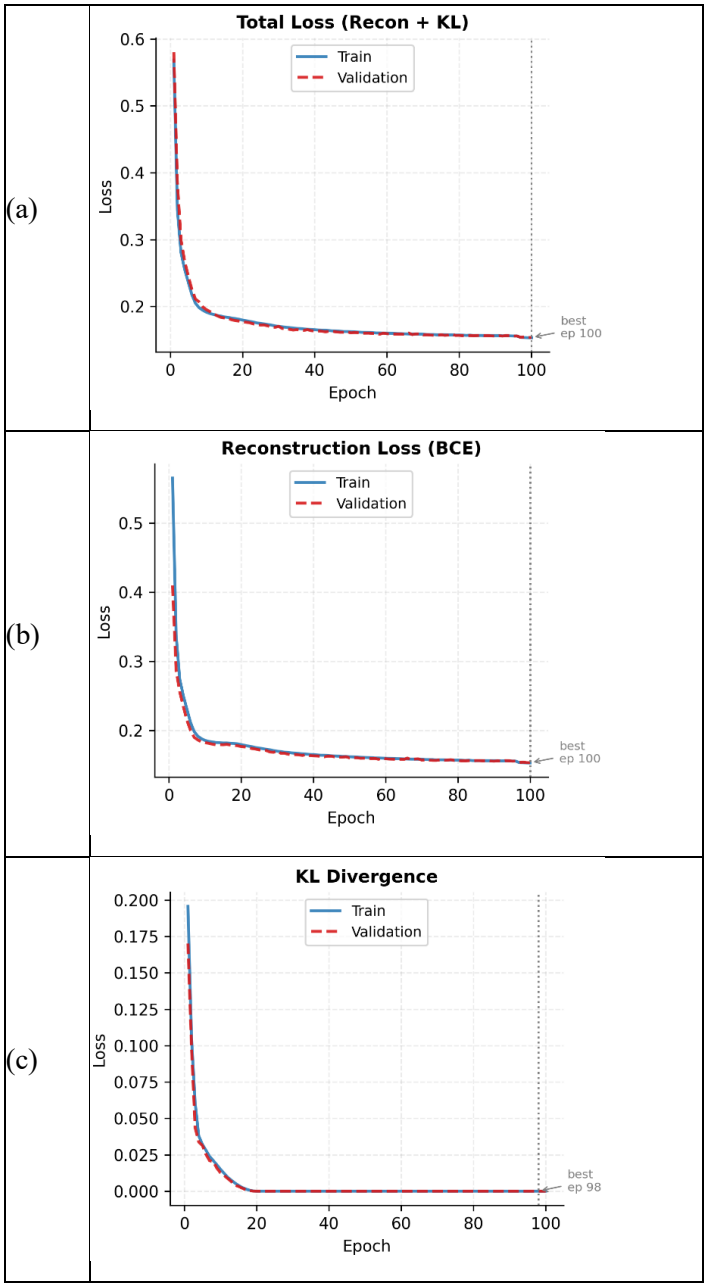

Figure 2: Training dynamics of the proposed VAE-LSTM model showing (a) total loss, (b) reconstruction loss, and (c) KL divergence for training and validation sets.

Furthermore, the KL divergence term progressively approaches a stable value while remaining well-behaved, demonstrating successful latent-space regularization without evidence of posterior collapse. Overall, the observed training behaviour confirms that the model effectively balances reconstruction quality and latent representation learning.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 06 | June – 2026 DOI:10.55041/ISJEM07601
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

4.5.2 RECONSTRUCTION PERFORMANCE

Table 1. Performance Metrics of the Proposed VAE-LSTM Model

Metric	Value
Precision	87.9 %
Recall	80.4 %
F1 Score	84.0 %

The reconstruction capability of the proposed model was evaluated using note-level precision, recall, and F1-score metrics on the validation set, as summarized in Table 1. The obtained results indicate that the model successfully captures the temporal and harmonic structures present in the symbolic music sequences while maintaining a balanced trade-off between correctly reconstructed notes and reconstruction completeness.

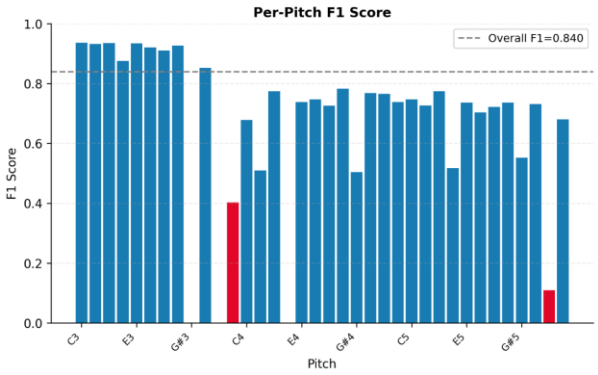

Figure 3: Per-pitch F1-score distribution across the pitch range.

To further analyse reconstruction quality across the pitch range, Figure 3 presents the per-pitch F1-score distribution. The results show that reconstruction performance remains relatively consistent across most pitches, indicating that the model does not excessively favor a narrow subset of notes. Minor variations among pitches can be attributed to differences in note occurrence frequencies within the training corpus. The overall distribution demonstrates that the learned latent representation effectively preserves melodic and harmonic information across the considered pitch range.

4.5.3 LATENT SPACE ANALYSIS

To investigate the structure of the learned latent representation, the latent vectors of validation samples were projected into two dimensions using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), as shown in Figure 4. The visualizations reveal a continuous and smoothly distributed latent space without isolated regions or severe fragmentation, suggesting that the encoder successfully maps musically related sequences into a coherent representation space.

The absence of abrupt discontinuities in both PCA and t-SNE projections indicates that neighbouring latent vectors correspond to structurally similar musical patterns. Such organization is desirable for generative modeling, as it facilitates smooth interpolation and sampling within the latent space. Additionally, the gradual variation of note-density characteristics across the latent manifold suggests that meaningful musical attributes are encoded within the learned representation, supporting the model's capability to generate diverse yet musically coherent sequences.

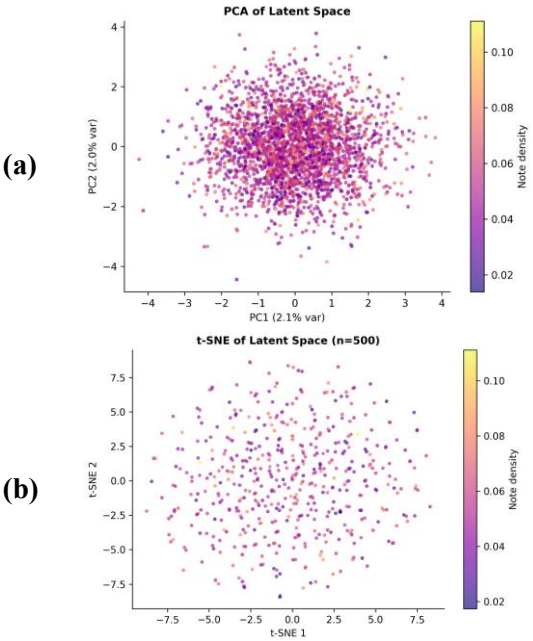

Figure 4: Visualization of the learned latent space of the proposed VAE-LSTM model using (a) PCA projection and (b) t-SNE projection, with points colored according to note density.

5. CONCLUSIONS

This study presented a VAE-LSTM architecture for symbolic music generation that combines the latent representation learning capability of Variational Autoencoders with the sequence modeling strength of Long Short-Term Memory networks. Using piano-roll representations derived from the Nottingham MIDI dataset, the proposed model successfully learned meaningful latent embeddings capable of reconstructing

and generating musically coherent note sequences. Experimental results demonstrated stable training behaviour, effective latent-space regularization, and strong reconstruction performance, indicating that the model can capture important melodic and temporal characteristics of symbolic music. Furthermore, latent-space visualizations revealed a continuous and well-structured representation that supports smooth sampling and generation of diverse musical patterns.

Future work will focus on extending the architecture to generate longer and more structurally complex musical compositions through hierarchical decoding mechanisms and attention-based sequence modeling. The integration of conditional latent variables for style, genre, or emotion control may further enhance the flexibility of the generation process. Additionally, training on larger and more diverse symbolic music datasets could improve the diversity and generalization capability of the generated compositions.

REFERENCES

1. Xu, Y.: Enhancing Music Generation With a Semantic-Based Sequence-to-Music Transformer Framework. Int. J. Semant. Web Inf. Syst. 20(1) (2024). https://doi.org/10.4018/IJSWIS.343491
2. Gunawan, A.A.S., Iman, A.P., Suhartono, D.: Automatic Music Generator Using Recurrent Neural Network. Int. J. Comput. Intell. Syst. 13(1) (2020) 645–654. https://doi.org/10.2991/ijcis.d.200519.001
3. Chieppa, S., Brutti, P., Pedro Paiva, R.: Automatic Guitar Transcription With Deep Neural Networks. IEEE Access 13 (2025) 113573–113585. https://doi.org/10.1109/ACCESS.2025.3583646
4. Cataltepe, Z., Yaslan, Y., Sonmez, A.: Music Genre Classification Using MIDI and Audio Features. EURASIP J. Adv. Signal Process. 2007(1) (2007) 036409. https://doi.org/10.1155/2007/36409
5. Ghatas, Y.S., Fayek, M.B., Hadhoud, M.M.: Generic Symbolic Music Labeling Pipeline. IEEE Access 10 (2022) 76233–76242. https://doi.org/10.1109/ACCESS.2022.3192462
6. Yang, L.-C., Chou, S.-Y., Yang, Y.-H.: MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation. arXiv preprint arXiv:1703.10847 (2017). https://arxiv.org/abs/1703.10847
7. Ji, S., Yang, X., Luo, J.: A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges. ACM Comput. Surv. 56(1) (2024) Article 7. https://doi.org/10.1145/3597493
8. Mangal, S., Modak, R., Joshi, P.: LSTM Based Music Generation System. IARJSET 6(5) (2019) 47–54. https://doi.org/10.17148/IARJSET.2019.6508
9. Sulun, S., Davies, M.E.P., Viana, P.: Symbolic Music Generation Conditioned on Continuous-Valued Emotions. IEEE Access 10 (2022) 44617–44626. https://doi.org/10.1109/ACCESS.2022.3169744
10. Yadav, P.S., Khan, S., Singh, Y.V., Garg, P., Singh, R.S.: A Lightweight Deep Learning-Based Approach for Jazz Music Generation in MIDI Format. Comput. Intell. Neurosci. 2022 (2022) 2140895. https://doi.org/10.1155/2022/2140895
11. Han, B., Li, Y., Shen, Y., Ren, Y., Han, F.: Dance2MIDI: Dance-driven multi-instrument music generation. Comput. Vis. Media 10(4) (2024) 791–802. https://doi.org/10.1007/s41095-024-0417-1
12. Mitra, R., Zualkernan, I.: Music Generation Using Deep Learning and Generative AI: A Systematic Review. IEEE Access 13 (2025) 18079–18106. https://doi.org/10.1109/ACCESS.2025.3531798
13. Brunner, G., Konrad, A., Wang, Y., Wattenhofer, R.: MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer. In: Gómez, E., Hu, X., Humphrey, E., Benetos, E. (eds.): Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018), pp. 747–754 (2018). https://doi.org/10.3929/ethz-b-000292318
14. Civit, M., Civit-Masot, J., Cuadrado, F., Escalona, M.J.: A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. Expert Syst. Appl. 209 (2022) 118190. https://doi.org/10.1016/j.eswa.2022.118190
15. Kumar Bairwa, A., Bhat, S., Sawant, T., Manoj, R.: MGU-V: A Deep Learning Approach for Lo-Fi Music Generation Using Variational Autoencoders With State-of-the-Art Performance on Combined MIDI Datasets. IEEE Access 12 (2024) 143237–143251. https://doi.org/10.1109/ACCESS.2024.3471918
16. Liu, L., Gong, R., Yang, Y.: MusDiff: A multimodal-guided framework for music generation. Alexandria Eng. J. 129 (2025) 128–136. https://doi.org/10.1016/j.aej.2025.05.053
17. Wang, W., Li, J., Li, Y., Xing, X.: Style-conditioned music generation with Transformer-GANs. Front. Inf. Technol. Electron. Eng. 25(1) (2024) 106–120. https://doi.org/10.1631/FITEE.2300359
18. Wu, S.-L., Kim, Y., Huang, C.-Z.A.: MIDI-LLM: Adapting Large Language Models for Text-to-MIDI

Music Generation. arXiv preprint arXiv:2511.03942 (2025). https://arxiv.org/abs/2511.03942

19. Kundu, S., Singh, S., Iwahori, Y.: Emotion-Guided Image to Music Generation. In: Proceedings of the 2024 7th Artificial Intelligence and Cloud Computing Conference, AICCC '24, pp. 323–330. Association for Computing Machinery, New York (2025). https://doi.org/10.1145/3719384.3719430

20. Sun, X., Han, X., Yao, F., Xu, J.: Emotion-aware cross-modal music generation based on multimodal emotion recognition. Alexandria Eng. J. 133 (2025) 254–270. https://doi.org/10.1016/j.aej.2025.11.020

21. Zhao, H., Min, S., Fang, J., Bian, S.: AI-driven music composition: Melody generation using Recurrent Neural Networks and Variational Autoencoders. Alexandria Eng. J. 120 (2025) 258–270. https://doi.org/10.1016/j.aej.2025.02.013