

# MIDI-based generative neural networks with variational autoencoders for innovative music creation

Rosalina<sup>1</sup>, Genta Sahuri<sup>2</sup>

<sup>1</sup>Department of Informatics Engineering, Faculty of Computer Science, President University, Bekasi, Indonesia

<sup>2</sup>Department of Information Systems, Faculty of Computer Science, President University, Bekasi, Indonesia

## Article Info

### Article history:

Received Feb 9, 2024

Revised Feb 23, 2024

Accepted Mar 6, 2024

### Keywords:

Generative neural network

Innovative music creation

Latent space

Musical instrument digital interface

Variational autoencoder

## ABSTRACT

By utilizing variational autoencoder (VAE) architectures in musical instrument digital interface (MIDI)-based generative neural networks (GNNs), this study explores the field of creative music composition. The study evaluates the success of VAEs in generating musical compositions that exhibit both structural integrity and a resemblance to authentic music. Despite achieving convergence in the latent space, the degree of convergence falls slightly short of initial expectations. This prompts an exploration of contributing factors, with a particular focus on the influence of training data variation. The study acknowledges the optimal performance of VAEs when exposed to diverse training data, emphasizing the importance of sufficient intermediate data between extreme ends. The intricacies of latent space dimensions also come under scrutiny, with challenges arising in creating a smaller latent space due to the complexities of representing data in N dimensions. The neural network tends to position data further apart, and incorporating additional information necessitates exponentially more data. Despite the suboptimal parameters employed in the creation and training process, the study concludes that they are sufficient to yield commendable results, showcasing the promising potential of MIDI-based GNNs with VAEs in pushing the boundaries of innovative music composition.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Rosalina

Department of Informatics Engineering, Faculty of Computer Science, President University

Ki Hajar Dewantara Street, Kota Jababeka, Cikarang Baru, Bekasi 17550-Indonesia

Email: [rosalina@president.ac.id](mailto:rosalina@president.ac.id)

## 1. INTRODUCTION

The field of automatic music composition has captivated researchers and practitioners across diverse disciplines for decades [1]–[4]. This interdisciplinary interest stems from the desire to leverage artificial intelligence and computational methods in musical creativity. Automated composition in music serves as a platform for discovering novel musical expressions, pushing the boundaries of conventional composition [5]–[9]. Musicology benefits from studying patterns generated by automated systems, shedding light on the evolution of musical forms and styles. Music philosophy engages in profound inquiries regarding the nature of creativity and the collaborative potential between human composers and algorithmic systems [10]–[12]. Computer science plays a crucial role, providing tools and techniques for implementing generative algorithms and neural networks, and facilitating the automated creation of music [13]–[16]. This collective exploration promises to unveil new dimensions of creativity, bridging the gap between human artistic intuition and the computational capabilities of evolving technologies.

The evolution of music production has undergone a profound transformation owing to rapid strides in artificial intelligence, notably in the field of generative neural networks (GNNs) employing variational

autoencoder (VAE) architectures [17]–[20]. This blend of technology and musical creativity marks a paradigm shift in music composition, ushering in an era of unprecedented possibilities. One illustrative example is the capability of GNNs empowered by VAEs to analyze and generate novel musical compositions based on intricate patterns learned from existing datasets. For instance, a GNN with VAE architecture can seamlessly process and reinterpret musical instrument digital interface (MIDI) based musical data [21], [22], capturing nuances in rhythm, melody, and harmony to produce innovative compositions. Figure 1 illustrates a visual representation of a MIDI file, showcasing the seamless exchange of data facilitated by MIDI. This digital communication standard enables electronic musical instruments, computers, and various devices to communicate and synchronize, fostering harmony among these systems. The graphical depiction assigns the Y-axis to different musical notes, while the X-axis signifies the progression of time, offering a clear insight into the temporal organization of notes and providing a comprehensive view of the musical composition's structure and rhythm. MIDI's role is evident not only in facilitating communication but also in enhancing our understanding of musical elements within the digital domain.

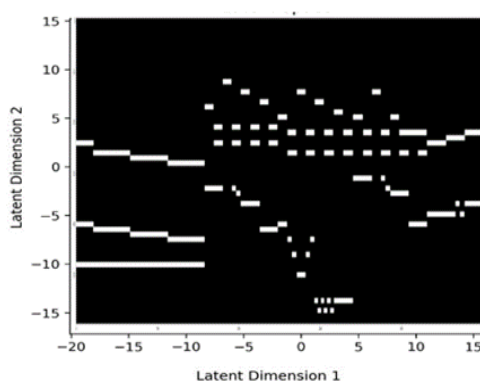


Figure 1. An example of the visualization of a MIDI file

The integration of GNNs with MIDI technologies, particularly harnessing the intelligence of VAEs, responds dynamically to the escalating demand for original and groundbreaking music. The traditional landscape of music composition, deeply rooted in creative expression, has undergone a significant metamorphosis with the advent of MIDI, enabling the digital representation of musical notes and laying the foundation for computational approaches to music creation. The subsequent integration of GNNs into the MIDI-based framework amplifies this transformative journey. GNNs showcase an impressive capacity to generate diverse and innovative content across various domains and hold the potential to revolutionize the essence of music production [23]–[27]. By synthesizing the expressive roots of traditional music composition with cutting-edge technology, the amalgamation of GNNs with MIDI technologies stands at the forefront of reshaping the landscape of music creation. This convergence meets the contemporary demand for originality and serves as a promising focal point for research, exploring the intricate correlation between GNNs and MIDI frameworks, paving the way for novel avenues in the creative realm of music. This research aims to push the boundaries of innovative music creation by harnessing the intelligence of MIDI-based GNNs with VAEs. By delving into the synergy between GNNs and MIDI technologies, this study aims to unlock new possibilities for composers and musicians, providing them with unprecedented tools to explore, experiment, and redefine the frontiers of musical creativity. This fusion of traditional musical roots with cutting-edge technology promises not only to preserve creative expression but also to propel the art of music composition into uncharted and exciting territories.

This research approach involves a detailed investigation into the convergence of latent spaces within VAEs, specifically focusing on the impact of training data variation. This study aims to scrutinize the challenges associated with latent space dimensions, seeking avenues to improve convergence and optimize the representation of musical data. Through the meticulous refinement of parameters and the incorporation of insights derived from the unique characteristics of MIDI data, the objective is to present an enhanced approach for MIDI-based GNNs utilizing VAEs in the realm of music composition. The innovation of this research lies in advancing the nuanced understanding and capabilities of MIDI-based GNNs with VAEs, thereby pushing the boundaries of achievable outcomes in music composition. By systematically addressing the identified challenges, this research not only contributes to the academic understanding of neural network applications in music but also holds practical implications for the development of tools tailored for

composers and musicians. The anticipated outcomes include the creation of a more refined and adaptable system, empowering musicians to explore previously uncharted territories in music composition. Ultimately, this research endeavors to bridge the gap between artificial intelligence and human creativity, facilitating a harmonious integration of technological innovation into the creative landscape of music.

## 2. RESEARCH METHOD

In our research methodology, we adopt a layered strategy to condense data into a latent space, employing convolutional neural networks (CNNs) for processing spectrogram inputs. The pivotal role of CNN layers lies in capturing nuanced spatial hierarchies inherent in music data, facilitating effective feature extraction. The subsequent decoder network intricately replicates the encoder's architecture, adeptly transforming condensed latent representations into coherent, meaningful musical outputs. By harnessing the spatial analysis capabilities of CNNs, our method efficiently compresses and reconstructs complex musical features, establishing a robust framework for innovative music creation. This not only enhances information representation efficiency but also contributes to the generation of novel and compelling musical compositions. The synergy between CNNs and latent space transformations is visually represented in Figure 2, providing a comprehensive illustration of our methodology.

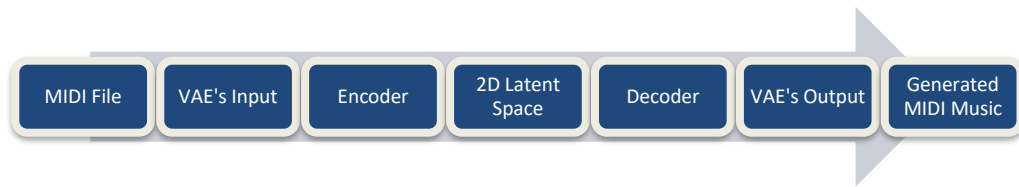


Figure 2. Schematic representation of the layered approach in MIDI-based GNNs with VAEs

### 2.1. Musical instrument digital interface file

We employ music files in the MIDI format, a symbolic representation akin to sheet music. MIDI files consist of multiple tracks, each capable of being active with a specific pitch, velocity, and duration sustained over multiple time steps, or inactive, representing silence. Moreover, each track is assigned a specific instrument. This format provides a flexible and structured way to encode musical information, allowing for the representation of complex compositions with various instruments and dynamic elements. The MIDI's versatility in handling multiple tracks enables the nuanced portrayal of musical intricacies, capturing the expressive nuances of each instrument. The use of MIDI as a symbolic representation facilitates the synthesis of diverse musical elements, making it an ideal choice for our research focused on innovative music creation through GNNs. The MIDI data can be represented as (1).

$$MIDI = \{(P_{t,i}, V_{t,i}, D_{t,i})\} | 1 \leq t \leq T, 1 \leq i \leq N_t \quad (1)$$

Where  $T$  is the total number of tracks in the MIDI file,  $N_t$  as the number of notes in track  $t$ ,  $P_{t,i}$  as the pitch of the  $i$ -th note in the track  $t$ ,  $V_{t,i}$  as the velocity of the  $i$ -th note in track  $t$ ,  $D_{t,i}$  as the duration of the  $i$ -th note of the track  $t$ .

### 2.2. Variational autoencoder's input

In the next step, the music in the MIDI file undergoes a transformation, transitioning from the audio waveform to a mel-spectrogram, an image-based representation of the audio signal. This mel-spectrogram serves as input for the VAE, a type of neural network. The VAE processes this spectrogram, generating an output that reflects the encoded features of the musical information. This conversion from audio to mel-spectrogram provides a condensed and structured representation, allowing the VAE to capture essential musical characteristics and patterns. The utilization of mel-spectrograms enhances the model's ability to extract meaningful features, facilitating the creation of innovative music through the generative capabilities of the VAE. The construction of the VAE model will adhere to the architectural framework depicted in Figure 3. This figure illustrates the specific layers, connections, and configurations integral to the VAE's structure. The model design encompasses the encoder section, responsible for mapping input data into a latent space, and the decoder section, adept at reconstructing the input from the encoded latent representations.

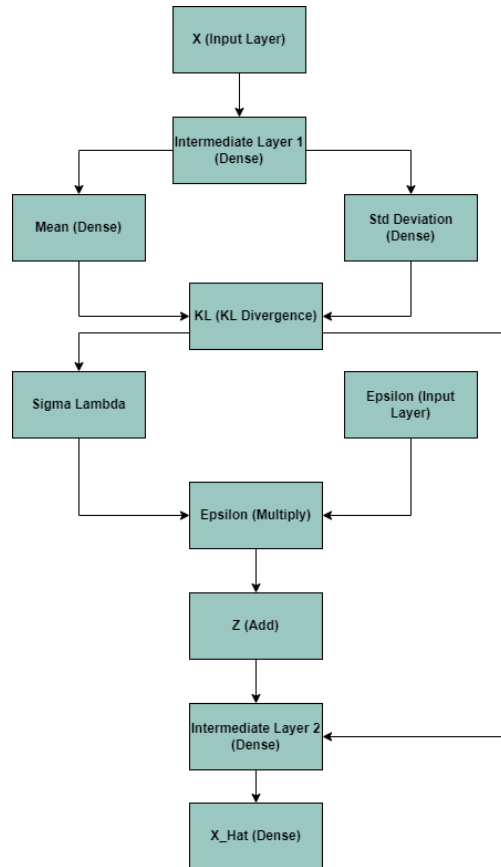


Figure 3. The model of the VAE

### 2.3. Encoder

The audio variational autoencoder (VAE) encoder consists of four layers, commencing with the Input layer that receives mel spectrogram data, representing the temporal frequency distribution of a MIDI piece. The input layer receives the mel spectrogram data, which represents the temporal frequency distribution of a MIDI piece. Let's denote the input mel spectrogram as  $X$ , which has dimensions  $W \times H$ , where  $w$  is the width (time steps) and  $H$  is the height (frequency bins) of the spectrogram. The flattened layer is applied to convert the 2D mel spectrogram into a 1D tensor, which reshapes the spectrogram into a vector of size  $W \times H$ .

$$X_{Flat} = Flatten(X) \quad (2)$$

The flattened mel spectrogram is then passed through two Dense layers. The first Dense layer computes the mean ( $\mu$ ) and the second Dense layer calculates the variance ( $\sigma^2$ ) of the features within the data. Let  $h_1$  and  $h_2$  be the outputs of the first and second Dense layers, respectively. The mean ( $\mu$ ) and the log variance ( $\log(\sigma^2)$ ) are computed as (3) to (4).

$$\mu = Dense_1(X_{Flat}) \quad (3)$$

$$(\log(\sigma^2)) = Dense_2(X_{Flat}) \quad (4)$$

To sample from the learned distribution, the model uses a reparameterization trick. It samples from a standard normal distribution ( $\epsilon$ ) and then scales and shifts the samples by the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) to obtain the latent representation ( $z$ ).

$$z = \mu + \sigma \odot \epsilon \quad (5)$$

Where  $\odot$  represents element-wise multiplication.

The sampled  $z$  represents a point in the 2D latent space. Each point in this space encapsulates the encoded mean and variance of an audio sample from the dataset. The encoder strategically organizes these

points so that similar data points, sharing musical features like pitch or rhythm, are proximately located, while dissimilar ones are more distantly positioned.

#### 2.4. 2D latent space

The 2D latent space in a VAE is a lower-dimensional representation where each point corresponds to the encoded mean and variance of an input sample. This space is constructed by the encoder part of the VAE and serves as a compressed, abstract representation of the input data. Figure 4 illustrates a 2-dimensional latent space graph, offering a visual representation of diverse data points. In this graphical depiction, each point corresponds to a unique position within the transformed latent space, encapsulating the encoded mean and variance of audio samples.

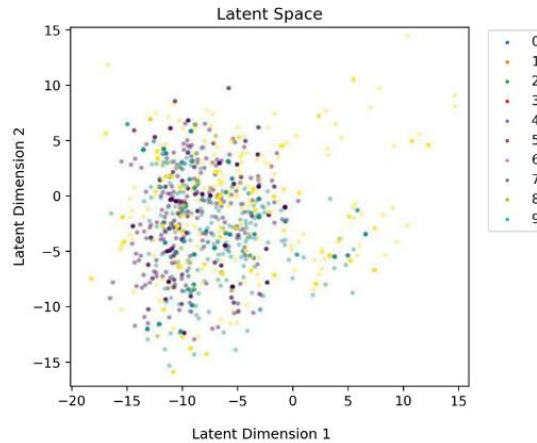


Figure 4. A graph depicting various data in a 2-dimensional latent space

The intentional arrangement of these points ensures that similar data, characterized by shared musical features like pitch or rhythm, are situated in proximity, fostering a coherent representation. In the decoder phase, the process begins with input derived from the latent space, characterized by its latent\_dim dimensionality. This latent representation serves as the foundation for reconstructing the original music data. Subsequently, a Dense layer, activated by a sigmoid function, is utilized to map these latent representations back into the comprehensive feature space of the music, effectively reversing the compression process initially executed by the encoder. This layer meticulously restores the intricate details embedded within the music, ensuring the faithful reproduction of its characteristics. The final layer of the decoder is crucial in reshaping the output from the Dense layer to align with the original spectrogram dimensions. This pivotal step essentially completes the reversal of the encoding process, providing a reconstructed representation of the original music data. Throughout this reconstruction process, the decoder leverages the encoded means extracted from actual music data within the latent space, ensuring that the generated outputs encapsulate the essential characteristics of the input.

#### 2.5. Decoder

During the reconstruction phase, the decoder utilizes the encoded means from the latent space to sample points and reconstructs the input data. The reconstruction loss, often calculated using the mean squared error (MSE) for image data such as spectrograms, measures the disparity between the input data and the data reconstructed by the decoder. Simultaneously, the KL-divergence loss quantifies the difference between the learned latent distribution and a prior distribution, typically a standard normal distribution. This loss term encourages the latent space to align with the prior distribution, promoting the learning of a disentangled and continuous latent space. Mathematically, the KL-divergence loss is computed as (6).

$$L_{KL} = -\frac{1}{2} \sum_{i=1}^N (1 + \log(\sigma^2) - \mu_i^2 - \sigma_i^2) \quad (6)$$

Where  $\mu_i$  and  $\sigma_i$  are the mean and standard deviation of the learned latent distribution for the  $i$ -th data point, respectively.

### 3. RESULTS AND DISCUSSION

In previous studies, the impact of various factors on musical composition and generation has been explored. However, there is a notable gap in explicitly addressing the influence of the latent space quality in VAE models on the generation of coherent and meaningful musical compositions. While prior research has focused on the technical aspects of VAEs and their application in music generation, few studies have delved into the specific characteristics of the latent space that contribute to the quality and creativity of the generated music. To fill this gap, this study evaluates the effectiveness of the model through various criteria, emphasizing the construction of a robust latent space representation and accurate reconstruction of input data with minimal loss of information. A methodology is proposed to gauge the model's improvement over epochs, measuring progress iteratively during training and using checkpoints to evaluate performance. This approach, leveraging Keras' callback function, enables the generation of the latent space and corresponding outputs using matplotlib graphs, facilitating a detailed analysis. The evaluation of latent space quality focuses on two key aspects: the latent space's discriminative capability and its convergence to the true posterior distribution. By assessing these aspects, the study aims to gain insights into the latent space's capacity to capture meaningful representations of input data and its ability to learn the underlying structure of the data distribution.

Throughout the model testing process, the effectiveness of the model is evaluated through various criteria. It is imperative that the model constructs a robust latent space representation and accurately reconstructs the input data with minimal loss of information. Since assessing these aspects cannot be simply represented by a binary outcome, a methodology is proposed to gauge the model's improvement over epochs. Instead of evaluating the model on a per-scenario basis, its progress is measured iteratively during training. Leveraging Keras' callback function, checkpoints are set to evaluate the model's performance at designated intervals during training. At these checkpoints, the model generates its latent space and corresponding outputs using matplotlib graphs. For the output, an 8x8 2-dimensional matrix is generated, with each element containing coordinates to sample from in the latent space, mapped from 0 to 1. Figure 5 illustrates these coordinates.

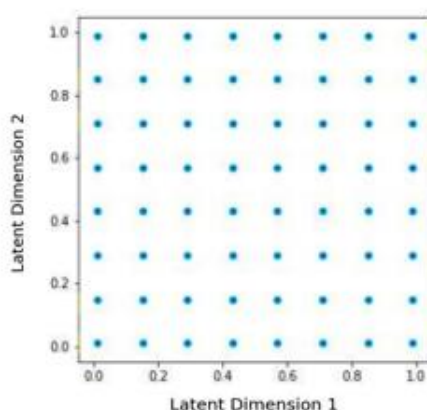


Figure 5. Sample coordinates to draw from the latent space

The latent space quality will be assessed based on its ability to classify snippets from different training data accurately and its convergence to closely approximate the true posterior distribution. This evaluation entails two key aspects: first, the latent space's discriminative capability, measured by its effectiveness in distinguishing between snippets originating from distinct training data sources. Secondly, the latent space's convergence to the true posterior distribution is gauged by its proximity to the distribution of latent variables conditioned on the observed data. These assessments provide insights into the latent space's capacity to capture meaningful representations of the input data and its ability to learn the underlying structure of the data distribution. By scrutinizing these aspects, we gain a comprehensive understanding of the latent space's quality and its suitability for generating coherent and representative outputs during the model's training process.

Figure 6 shows illustrates the latent space in its initial state. Both the latent space (depicted in Figure 6(a)) and the generated result (depicted in Figure 6(b)). It's evident that the latent space appears to lack meaningful structure, resembling randomness, while the generated result exhibits characteristics akin to noise. These observations indicate that the model has not yet effectively learned the underlying patterns and

features of the input data. The latent space, intended to encapsulate meaningful representations of the data, appears disorganized and devoid of discernible patterns. Similarly, the generated output lacks coherence and fails to capture the salient features of the original music data. These results suggest that further refinement and optimization of the model architecture and training process are necessary to achieve the desired outcomes of meaningful latent space representations and coherent generative outputs.

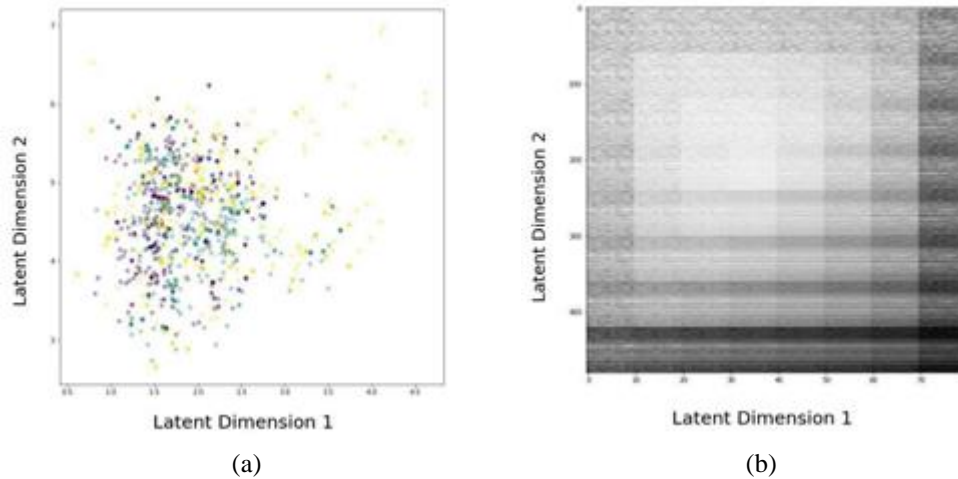


Figure 6. Illustrates the latent space in its initial state, showcasing the initial result in (a) the initial latent space and (b) the initial result

Figure 7 shows illustrates both the latent space (Figure 7(a)) and the generated result (Figure 7(b)) obtained at the 10th epoch. Notably, the encoder component of the VAE is discernibly attempting to classify the training data across the latent space. This suggests that the latent space representation is evolving towards capturing more meaningful and structured representations of the input data. Additionally, compared to earlier epochs, the generated results exhibit reduced noise, indicating improved coherence and fidelity to the original music data. These developments signify the progress of the VAE model in learning and representing the underlying patterns and features of the input music data. However, further optimization and refinement may still be necessary to achieve even more robust latent space representations and higher-quality generative outputs.

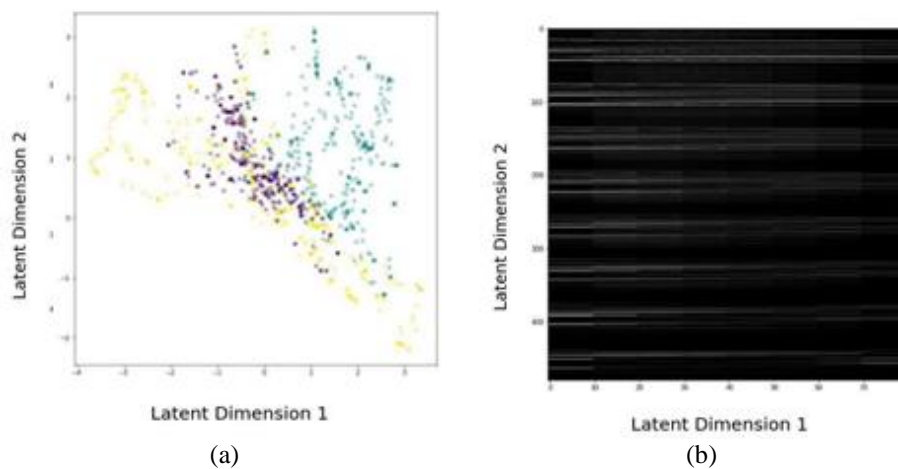


Figure 7. Illustrates (a) the latent space at the 10th epoch and (b) the result at the 10th epoch

Figure 8 displays both the latent space (Figure 8(a)) and the generated result (Figure 8(b)) obtained at the 1000th epoch. It appears that the encoder is encountering difficulties in creating a more compact latent space. This challenge may arise from the random selection of training data with varying tempos, note ranges, and other factors, leading to a complex and diverse input distribution. Despite these challenges, the results are beginning to exhibit structural characteristics. This suggests that the VAE model is gradually learning to extract meaningful features from the input data and generate outputs with discernible patterns. While the latent space may not yet be fully optimized, the emergence of structure in the generated results indicates progress in the model's training. Continued iterations and adjustments may further refine the latent space representation and enhance the quality of the generated outputs.

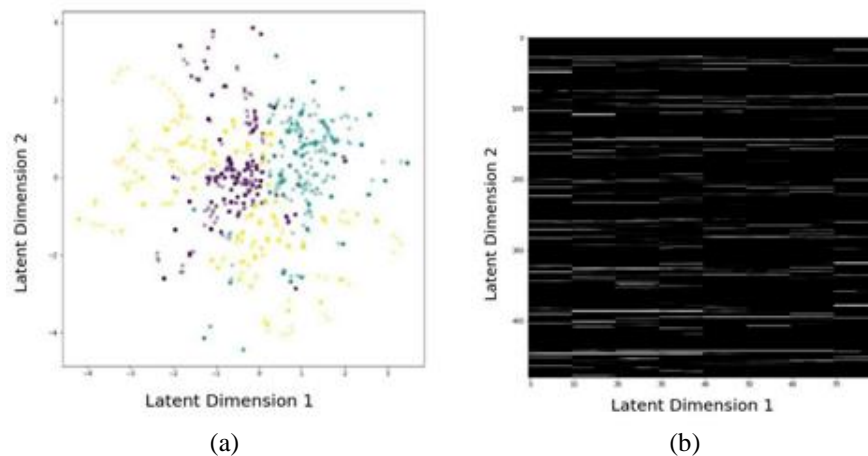


Figure 8. Depicts (a) the latent space at the 1000 th epoch and (b) the result at the 1000 th epoch

Figure 9 illustrates the implementation result, focusing on the generated result display. Within this display, sliders are utilized to represent the values used to draw samples from the latent space, thus generating musical compositions. Each slider corresponds to a dimension within the latent space. Additionally, the latent space graph is depicted, showcasing the distribution of training data points within the latent space. This graph also indicates the position from which the user will draw samples, aligned with the slider values. The generated result display exhibits the outcomes of the generation process, with the X-axis representing time and the Y-axis denoting the note values. This visualization provides users with a comprehensive understanding of the generation process, enabling them to interactively manipulate latent space parameters to produce desired musical compositions.

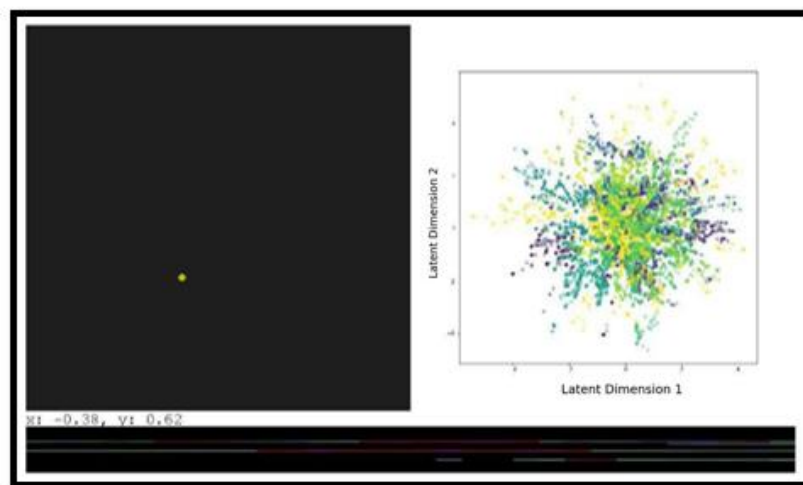


Figure 9. MIDI-based GNNs with VAE implementation result



Figure 10 depicts the effect of a very high tolerance level on the generated musical composition. In this scenario, notes are only positioned where the neural network prediction for note placement approaches 1, indicating a high level of confidence in those placements. When the user presses the up arrow on the keyboard, the application adjusts by increasing the tolerance level of the neural network. This alteration prompts the neural network to generate output with a broader tolerance for variations or deviations from the input data. Consequently, the generated output may display heightened diversity or creativity, incorporating more unconventional or unique musical elements. This adjustment allows for a broader exploration of musical possibilities, potentially leading to the creation of compositions with novel and innovative characteristics.

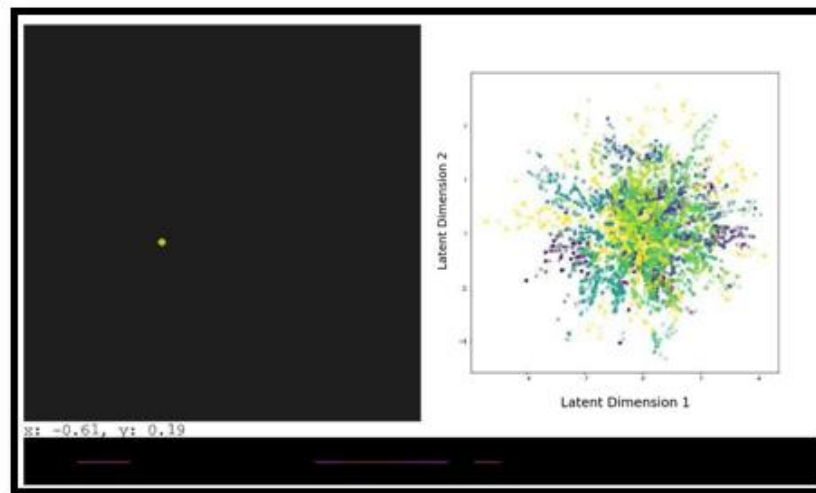


Figure 10. The effect of a very high tolerance level on the generated musical composition

Conversely, when the user presses the down arrow on the keyboard, the application adjusts by decreasing the tolerance level of the neural network. This modification prompts the neural network to generate output with a reduced tolerance for variations or deviations from the input data. Consequently, the generated output may demonstrate greater adherence to the patterns and characteristics of the training data. In Figure 11, representing a very low tolerance level, the resulting musical composition features notes placed only where the neural network exhibits some confidence that a note should be positioned there. This conservative approach ensures that notes are added to the composition only when the neural network's predictions are relatively certain. As a result, the generated output tends to closely resemble conventional or familiar musical styles, aligning more closely with the patterns learned from the training data.

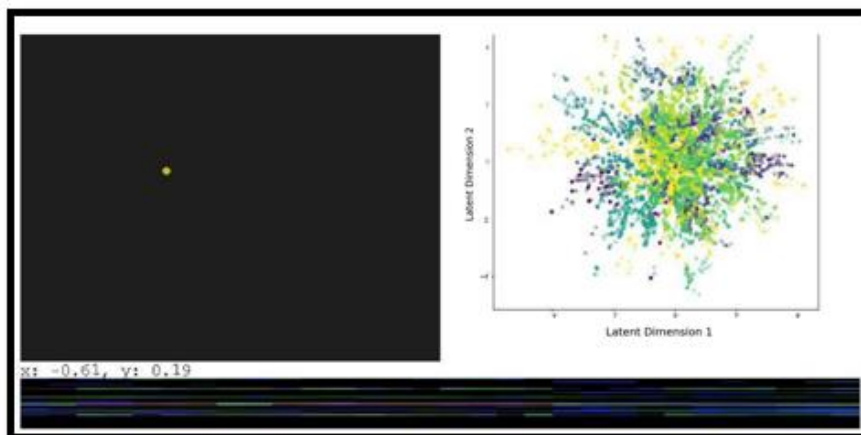


Figure 11. The effect of a very low tolerance level on the generated musical composition

#### 4. CONCLUSION

In conclusion, the integration of MIDI-based GNNs with VAEs represents a transformative approach to innovative music creation. This research journey has unveiled the potential of leveraging advanced artificial intelligence techniques to push the boundaries of traditional music composition. By synthesizing the expressive roots of music with cutting-edge technology, this fusion offers unprecedented opportunities for composers and musicians to explore new realms of creativity. Through meticulous experimentation and iterative refinement, the study has demonstrated the evolving capabilities of GNNs with VAEs in generating novel musical compositions. The evaluation of latent space representations and generated outputs has provided valuable insights into the model's learning process and its ability to capture meaningful patterns from input music data.

Moreover, the exploration of tolerance levels within the neural network's generation process has offered nuanced control over the diversity and coherence of generated compositions. This adaptive approach allows for the exploration of a spectrum of musical possibilities, from highly structured compositions to more experimental and innovative pieces. Looking ahead, the development of enhanced methodologies for MIDI-based GNNs with VAEs holds promise for further advancing the field of automatic music composition. By continuing to bridge the gap between artificial intelligence and human creativity, future research can unlock even greater potential for musical expression and exploration.





#### REFERENCES

- [1] D. Conklin, "Multiple viewpoint systems for music classification," *Journal of New Music Research*, vol. 42, no. 1, pp. 19–26, Mar. 2013, doi: 10.1080/09298215.2013.776611.
- [2] C. Plut and P. Pasquier, "Generative music in video games: State of the art, challenges, and prospects," *Entertainment Computing*, vol. 33, p. 100337, Mar. 2020, doi: 10.1016/j.entcom.2019.100337.
- [3] T. Y. Yi and S. Thiruvavur, "Understanding the potential of music learning application as a tool for learning and practicing musical skills," *International Journal of Creative Multimedia*, vol. 2, no. 1, pp. 42–56, Apr. 2021, doi: 10.33093/ijcm.2021.1.3.
- [4] Rosalina, A. Sengkey, G. Sahuri, and R. Mandala, "Generating intelligent agent behaviors in multi-agent game AI using deep reinforcement learning algorithm," *International Journal of Advances in Applied Sciences*, vol. 12, no. 4, pp. 396–404, Dec. 2023, doi: 10.11591/ijaas.v12.i4.pp396-404.
- [5] D. Rivero, I. Ramírez-Morales, E. Fernández-Blanco, N. Ezquerro, and A. Pazos, "Classical music prediction and composition by means of variational autoencoders," *Applied Sciences (Switzerland)*, vol. 10, no. 9, p. 3053, Apr. 2020, doi: 10.3390/app10093053.
- [6] H. T. Hung, C. Y. Wang, Y. H. Yang, and H. M. Wang, "Improving automatic jazz melody generation by transfer learning techniques," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, IEEE, Nov. 2019, pp. 339–346, doi: 10.1109/APSIPAASC47483.2019.9023224.
- [7] A. Jagannathan, B. Chandrasekaran, S. Dutta, U. R. Patil, and M. Eirinaki, "Original music generation using recurrent neural networks with self-attention," in *Proceedings - 4th IEEE International Conference on Artificial Intelligence Testing, AITest 2022*, IEEE, Aug. 2022, pp. 56–63, doi: 10.1109/AITest55621.2022.00017.
- [8] J. López-Montes, M. Molina-Solana, and W. Fajardo, "GenoMus: representing procedural musical structures with an encoded functional grammar optimized for metaprogramming and machine learning," *Applied Sciences (Switzerland)*, vol. 12, no. 16, p. 8322, Aug. 2022, doi: 10.3390/app12168322.
- [9] N. Hewahi, S. AlSaigal, and S. AlJanahi, "Generation of music pieces using machine learning: long short-term memory neural networks approach," *Arab Journal of Basic and Applied Sciences*, vol. 26, no. 1, pp. 397–413, Jan. 2019, doi: 10.1080/25765299.2019.1649972.
- [10] Y. W. Wen and C. K. Ting, "Recent advances of computational intelligence techniques for composing music," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 2, pp. 578–597, Apr. 2023, doi: 10.1109/TETCI.2022.3221126.
- [11] M. Dua, R. Yadav, D. Mamgai, and S. Brodiya, "An improved RNN-LSTM based novel approach for sheet music generation," *Procedia Computer Science*, vol. 171, pp. 465–474, 2020, doi: 10.1016/j.procs.2020.04.049.
- [12] S. Li and Y. Sung, "Inco-gan: Variable-length music generation method based on inception model-based conditional gan," *Mathematics*, vol. 9, no. 4, pp. 1–16, Feb. 2021, doi: 10.3390/math9040387.
- [13] J. P. Briot and F. Pachet, "Deep learning for music generation: challenges and directions," *Neural Computing and Applications*, vol. 32, no. 4, pp. 981–993, Feb. 2020, doi: 10.1007/s00521-018-3813-6.
- [14] J. P. Briot, "From artificial neural networks to deep learning for music generation: history, concepts and trends," *Neural Computing and Applications*, vol. 33, no. 1, pp. 39–65, Jan. 2021, doi: 10.1007/s00521-020-05399-0.
- [15] S. Angioni, N. Lincoln-DeCusatis, A. Ibba, and D. R. Recupero, "A transformers-based approach for fine and coarse-grained classification and generation of MIDI songs and soundtracks," *PeerJ Computer Science*, vol. 9, p. e1410, Jun. 2023, doi: 10.7717/PEERJ-CS.1410.
- [16] P. S. Yadav, S. Khan, Y. V. Singh, P. Garg, and R. S. Singh, "A lightweight deep learning-based approach for jazz music generation in MIDI format," *Computational Intelligence and Neuroscience*, pp. 1–7, Aug. 2022, doi: 10.1155/2022/2140895.
- [17] D. Devyatkin and I. Trenev, "Data generation with variational autoencoders and generative adversarial networks," in *Engineering Proceedings*, Basel Switzerland: MDPI, Jun. 2023, p. 37, doi: 10.3390/engproc2023033037.
- [18] F. Roche, T. Hueber, M. Garnier, S. Limier, and L. Girin, "Make that sound more metallic: towards a perceptually relevant control of the timbre of synthesizer sounds using a variational autoencoder," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, pp. 52–66, May 2021, doi: 10.5334/tismir.76.
- [19] J. Grekow and T. Dimitrova-Grekow, "Monophonic music generation with a given emotion using conditional variational autoencoder," *IEEE Access*, vol. 9, pp. 129088–129101, 2021, doi: 10.1109/ACCESS.2021.3113829.
- [20] C. Jin et al., "A transformer generative adversarial network for multi-track music generation," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 3, pp. 369–380, Sep. 2022, doi: 10.1049/cit.12065.
- [21] J. Grekow, "Generating polyphonic symbolic emotional music in the style of bach using convolutional conditional variational autoencoder," *IEEE Access*, vol. 11, pp. 93019–93031, 2023, doi: 10.1109/ACCESS.2023.3309639.





- [22] L. C. Yang, S. Y. Chou, and Y. H. Yang, "Midinet: A convolutional generative adversarial network for symbolic-domain music generation," *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, pp. 324–331, 2017.
- [23] O.-G. Cosma *et al.*, "Automatic music generation using machine learning," in *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, IEEE, Nov. 2024, pp. 1–9. doi: 10.1109/icecet58911.2023.10389273.
- [24] H. Zhang, L. Xie, and K. Qi, "Implement music generation with GAN: a systematic review," in *Proceedings - 2021 International Conference on Computer Engineering and Application, ICCEA 2021*, IEEE, Jun. 2021, pp. 352–355. doi: 10.1109/ICCEA53728.2021.00075.
- [25] H. W. Dong, W. Y. Hsiao, L. C. Yang, and Y. H. Yang, "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," *32nd AAI Conference on Artificial Intelligence, AAI 2018*, vol. 32, no. 1, pp. 34–41, Apr. 2018, doi: 10.1609/aaai.v32i1.11312.
- [26] C. Chauhan, B. Tanawala, and M. Hasan, "Multi-genre symbolic music generation using deep convolutional generative adversarial network," *ITM Web of Conferences*, vol. 53, p. 02002, Jun. 2023, doi: 10.1051/itmconf/20235302002.
- [27] P. Ferreira, R. Limongi, and L. P. Fávero, "Generating music with data: Application of deep learning models for symbolic music composition," *Applied Sciences (Switzerland)*, vol. 13, no. 7, p. 4543, Apr. 2023, doi: 10.3390/app13074543.

## BIOGRAPHIES OF AUTHORS



**Rosalina**     a respected lecturer in the Informatics Study Program at President University exemplifies dedication to her field. Her commitment to academic excellence is underscored by her master's degree in informatics from President University. With a strong educational background and a passion for informatics, she has a significant impact on students' academic journeys. Her expertise in the subject matter, combined with her ability to explain complex concepts, creates a dynamic and enriching learning environment. She can be contacted at email: [rosalina@president.ac.id](mailto:rosalina@president.ac.id).



**Genta Sahuri**     a dedicated lecturer at President University's Information Systems Study Program brings a wealth of knowledge and expertise to his role. Holding a master's degree in Informatics from the same institution underscores his commitment to academic excellence. With a strong educational background and a passion for his field, He plays a crucial role in shaping the academic journey of his students. His adeptness in conveying complex concepts fosters a dynamic and enriching learning environment. His dedication to his field and his ability to inspire students make him a valuable asset to President University. He can be contacted at email: [genta.sahuri@president.ac.id](mailto:genta.sahuri@president.ac.id).