Exploring the Potential of GANs, LSTM, and VAEs in Advancing Music Generation

G Sai Ram Pavan[1], Akash Varma Kucharlapati[1], N Moneesh[1],
S Abhishek[2], Anjali T [3]
amenu4aie20125@am.students.amrita.edu[1], amenu4aie20141@am.students.amrita.edu[1]
amenu4aie20150@am.students.amrita.edu[1], sabhishek@am.students.amrita.edu[2] anjalit@am.amrita.edu[3]
[1] [2] [3]Department of Computer Science and Engineering, Amrita School of Computing,
Amrita Vishwa Vidyapeetham, Amritapuri, India

***Abstract*—In this paper, we explore three popular approaches: GANs, LSTMs, and VAEs, to explore the fascinating world of computer-generated music. To envision GANs, imagine two creative friends working together to create a musical composition; one works diligently on the composition while the other serves as the discerning judge, ensuring its aesthetic appeal. As a memory reservoir, LSTMs, on the other hand, preserve the essence of previous melodies to serve as an inspiration for and guide in the creation of new harmonious compositions. Like a musical blender, VAEs expertly blend song fragments to produce completely new and melodic musical compositions. Our main goal is to determine which of these techniques produces the most enjoyable musical result while also pointing out any potential drawbacks. We can learn a lot about the future of computer-generated music, and its transformative potential in reshaping our listening experiences and perspectives on music composition, by carefully examining these methodologies.**

***Index Terms*—Computers and music, GANs, LSTMs, VAEs, tunes, melodies.**

I. Introduction

A language that is global and timeless and classic is music. Historically, its evolution has been shaped by innovations ranging from the invention of musical instruments to advancements in recording technology. Now, in the digital era, artificial intelligence (AI) is reshaping the boundaries of music creation and appreciation. The current state of research in AI-driven music generation can be compared to the early days of synthesizers, teetering on the brink of a revolution. While early attempts were algorithmic and often lacked the spontaneity of human creation, the infusion of AI techniques promises a richer tapestry of soundscapes.

The Generative Adversarial Network (GAN)[1] stands as a pivotal innovation in artificial intelligence, carving a noteworthy niche in music generation. In essence, a GAN operates with two intertwined neural networks: the generator, tasked with crafting data, and the discriminator, responsible for its evaluation. This duo engages in a dynamic tug-of-war, with the generator aspiring to concoct music indistinguishable from human-created compositions, prompting its continuous refinement. GAN's design inherently embodies a spirit of competition and perpetual improvement. When applied to music, GANs unfurl a unique palette, facilitating the emergence of novel melodies and rhythms that, while reflecting human ingenuity, bear an algorithmic imprint. This transformative architecture pushes the frontiers of music automation, melding authenticity with fresh innovation. The essence of GANs in music rests in its transformative potential, ushering in a phase where technology plays a co-creator's role in musical artistry.

Long Short-Term Memory networks (LSTMs)[2], a cornerstone in deep learning, have profoundly influenced music generation. Fundamentally, One kind of rnn is the LSTM, it is designed to recognize patterns over prolonged sequences, making it adept at understanding the intricate progressions in music. Its architecture comprises memory cells that can maintain information for extended periods, enabling the capture of long-term dependencies in musical notes and rhythms. When applied to music, LSTMs harness their ability to recall prior sequences, ensuring that generated pieces exhibit continuity and coherence, akin to human compositions. This deep learning framework has redefined the computational approach to music, producing compositions that can span genres while retaining the essence of each. The significance of LSTMs in the musical domain is monumental; they bridge the gap between algorithmic precision and the fluidity of artistic expression, marking a significant stride towards machines comprehending the soul of music.

The Variational Autoencoder (VAE)[3], a transformative concept in artificial intelligence, is making waves in the music generation field. At a foundational level, a VAE is a type of neural network trained to encode and decode data, capturing its underlying structure. Its architecture, balancing between data compression and recreation, allows the extraction of essential musical features, subsequently facilitating the generation of new compositions. In the musical context, VAEs operate by analyzing diverse tracks, internalizing their core elements, and blending them to produce harmonious outputs. This mechanism provides the capacity to generate music that borrows from multiple genres or styles, yet maintains a coherent auditory experience. The value of VAEs in music creation is unparalleled. They foster a confluence of diverse musical inspirations, reshaping the contours of automated composition. In essence, VAEs symbolize the promise of technology: to not just replicate, but to reimagine and enrich the vast tapestry of musical expression.

multiple genres or styles, yet maintains a coherent auditory experience. The value of VAEs in music creation is unparalleled. They foster a confluence of diverse musical inspirations, reshaping the contours of automated composition. In essence, VAEs symbolize the promise of technology: to not just replicate, but to reimagine and enrich the vast tapestry of musical expression.

II. Literature Survey

In this comprehensive study, we delve into the latest developments in music generation research, conducting a thorough exploration of the methodologies and outcomes associated with LSTM, GAN, and VAE-based approaches.

Shah, Falak[4] explains As availability of high computing power and the evolution of deep learning architecture the model could easily learn from music sequential data. Used Google Magenta's inbuilt model and Long Short Term Memory model. Many variants have proposed such as extra information about target, differences in the pitches, also changing hyperparameters. Adding additional information will be useful LSTM for prediction.

A. Maduskar[5] explains Autoregression models will generate iterative subsampling will results in generating localized music level. GAN models will generate global music level. By combining auto encoders with GAN we could minimize the volume of training data needed. and eventually speeding up computation and model will generate effective musical structure both in global and local music levels.

J. Tang[6] explains the revolutionary of transformer models, in the creation of music. It is emphasized that the innovative of the random mask module in RM-Transformer is a crucial innovation for improving music quality. The feature extraction that will make good music output. In order to develop the connection between artificial intelligence and human creativity in music composition to more study in areas like multi-track creation and the integration of emotion with music.

J. Wang[7] study indicates the problem of composing computer generated jazz music while attempting to conform created music to align generated music with established music rules. To do this it combines LSTM neural networks with music theory. Information about the length and note category is included in the input data. By using music language parsing, the LSTM model creates note sequences based on transition probabilities. While comparing with other models, it reveals that LSTM performs better on producing jazz music.

S. Sajad[8] explains recurrent neural network to generate melodious pieces of music. This model is build on training the existing melodies or instruments to generate new music. After training the model we will be finding best weight to generate the music. When the model is trained for 4 epoch the accuracy is 90%, but when the model trained for 80 epoch the accuracy is increased to 99%.

III. Dataset

The "GTZAN Dataset for Music Genre Classification" is a notable dataset curated by Andrada Olteanu. Designed primarily for the music genre classification task, this dataset stands out as a comprehensive collection to fuel research in music analysis. Although I can't provide exact figures without directly inspecting the dataset, its size suggests a substantial amount of data, encompassing multiple music genres. This dataset would typically consist of rows representing individual music tracks or samples and columns capturing various features of these tracks, potentially including aspects like tempo, rhythm, and melodic patterns, among others. Andrada's contribution to the community with this dataset provides researchers and enthusiasts with a rich resource to delve deep into the nuances of music genres and their classification.

IV. Methodology

There are some common steps in workflow for all the three algorithms.

Fig. 1. Fundamental workflow of the project.

A. GANs

The research endeavours to understand the capability of GANs in creating unique and harmonious musical outputs. By training on a dataset of musical compositions, we aim to evaluate the GAN's proficiency in capturing the nuances of musical patterns and rhythms and its ability to innovate within the framework of the given data.

1) ***Data Collection and Preprocessing*** *:* Acquired a comprehensive dataset of MIDI files to be the foundational musical data for the GAN.Leveraged tools such as pretty-midi to transform MIDI files into arrays, representing notes, chords, and rhythms. Undertook normalization processes, rendering the data suitable for efficient processing by the GAN.

2) ***GAN Architecture Setup****:* Contains to parts, Generator and Discriminator. **Generator** is composed of Composed layers designed to initiate and craft data, aiming to emulate the musical compositions in the dataset. **Discriminator** is Comprising layers that evaluate and ascertain the authenticity of the music pieces generated by the Generator. Iteratively, the Generator improves its outputs based on feedback from the Discriminator until the generated music is indistinguishable from real compositions.

3) ***Training****:* GANs were trained iteratively, where the Generator aspired to produce music pieces that the Discriminator could not differentiate from genuine compositions in the dataset. Loss functions were employed to gauge the distinction between real and generated compositions, with optimization algorithms like Adam optimizing the neural network weights.

*4) **Relevant Terminology**:*

- **Generative Adversarial Network :** An AI model comprising two neural networks - the Generator and the Discriminator. The Generator crafts data, while the Discriminator evaluates its authenticity.
- **MIDI Files :** Digital files that store musical information, facilitating electronic representation of music notes, rhythms, and instruments.
- **Normalization :** A process to standardize data, ensuring that all inputs have a uniform scale, aiding efficient processing by neural networks.

*5) **Workflow Diagram**:*

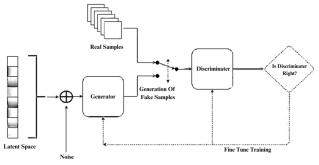

Fig. 2. Schematic representation of the GAN architecture

B. LSTMs

This research focuses on investigating the potential of LSTMs, a special category of recurrent neural networks, in understanding, assimilating, and then generating musical compositions that bear resemblance to a provided dataset. The emphasis lies in assessing the LSTM's capability to recognize, remember, and utilize long-term dependencies in music sequences.

*1) **Data Collection and Preprocessing**:* Extracted a dataset containing MIDI files as the primary source of musical data. Employed utilities, notably pretty-midi, to convert these MIDI files into structured arrays, capturing notes, sequences, and associated rhythms. Conducted data normalization, ensuring that LSTM processes the musical data seamlessly and efficiently.

*2) **LSTM Architecture Construction**:* Structured layers in the LSTM network to handle sequential data, making it equipped to handle the intricacies of music, which is inherently sequential. Incorporated dropout layers to prevent overfitting and ensure the model generalizes well to unseen musical data.

*3) **Training**:* The LSTM was rigorously trained on the processed MIDI files. The objective was to minimize the error between the predicted sequence of notes and the actual sequence. Utilized optimization techniques, particularly the Adam optimizer, to fine-tune and adjust neural network weights.

*4) **Music Generation**:* Leveraging the trained LSTM, sequences of musical notes were generated. The LSTM utilized previously recognized patterns and sequences from the dataset to create these new compositions, imbuing them with elements reminiscent of the training data.

*5) **Workflow Diagram**:*

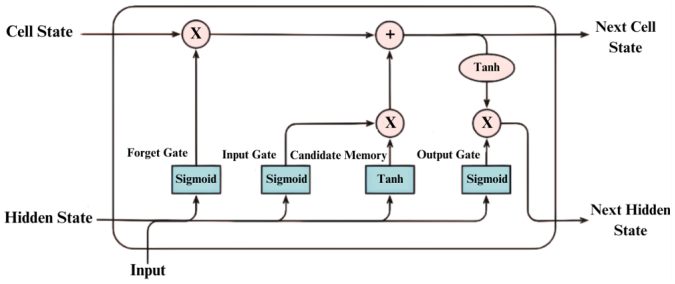

Fig. 3. Schematic representation of the LSTM architecture

C. Variable Auto Encoders VAEs

The goal of our research is to determine the efficacy of a Variational Autoencoder (VAE) in generating coherent and melodically sound music compositions. We aim to understand the architecture's ability to encode the core elements of existing music tracks, blend them, and subsequently decode this information to produce harmonious outputs.

*1) **Data Collection and Preprocessing**:* Sourced a dataset of MIDI files to serve as the musical input. Used tools like pretty-midi to convert MIDI files into arrays of notes and rhythms. Normalized the data to ensure the VAE can process it efficiently.

*2) **VAE Architecture Setup**:* Implemented the Encoder with layers that capture the essential features of the music. The encoder compresses the musical data into a latent space representation. Designed the Decoder to reconstruct music from this compressed form. It takes the latent representation and generates a new composition. Incorporated the reparameterization trick to sample from the latent space, ensuring gradient descent optimization remains feasible.

*3) **Training**:* The VAE was trained on the processed MIDI files, minimizing the difference between the original and the VAE's reconstructed output. This was done using reconstruction loss and the KL divergence, a measure of the difference between the trained model and a standard Gaussian distribution. Used optimization algorithms, such as the Adam optimizer, to reduce the loss.

*4) **Music Genration**:* Sampled from the latent space of the trained VAE to produce new compositions. This involves generating new latent vectors and passing them through the decoder.

By following the methodology detailed above, the VAE is trained and utilized for the generation of new musical compositions, leveraging the inherent patterns from the source MIDI files. We will discuss the indepth comparison in results, analysis

*5) **Workflow Diagram**:*

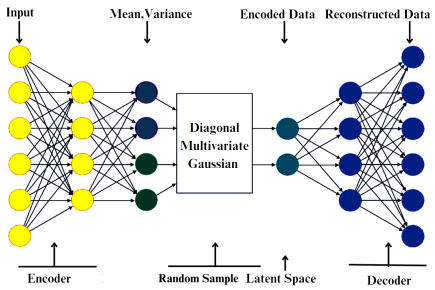

Fig. 4. Schematic representation of the VAE architecture

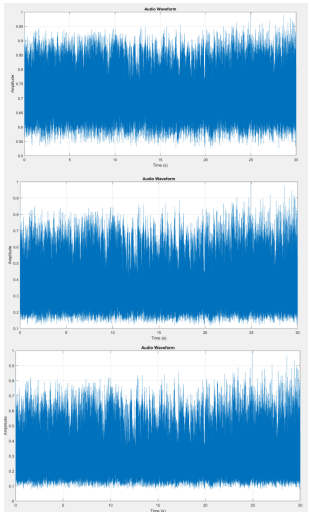

Fig. 5. Visual representation of music generated by a Variational Autoencoder

V. Results and Analysis

After working with the algorithms on our piano dataset, we embarked on a detailed evaluation of the produced music files. The outcome is a clear and distinct quality ranking emerged. The music created by the LSTM algorithm was, in our eyes, top-notch, easily surpassing the tunes from the GAN method. Interestingly, GAN compositions also had an edge over those from the VAE method. To make this data more digestible, we decided to use MATLAB and create insightful plots. These visual aids, which we'll showcase below, simplify our findings for everyone.

Diving deeper, in our pursuit to perfect algorithmic piano music, we gave priority to two factors: energy and mean values. Here's the thing: For serene, soft piano tunes, you'd expect minimal energy. As for the mean value? It's crucial for it to be as close to a central baseline as possible. This ensures the absence of any undue bias towards extreme sound levels, whether too high or too low. By holding these factors in high regard, we can accurately gauge the music's alignment with our ideal characteristics. It also equips us to compare and contrast varied algorithmic techniques in the fascinating world of piano music generation.

A. *Variabe Auto Encoders*

We analysed that comparision with other approaches, the Variational Autoencoder (VAE) produced more errors in the music. So, we made the decision to strengthen our method or assessment and we used the VAE method to create new music samples to do this. By expanding our dataset, we want to understand how VAE generates music as well as to have more data to analyze. This action is important for making sure the accuracy and dependability of our study findings.

B. *Generative Adversarial Network*

In our study, we mainly looked at the GAN method. To really understand how GAN works for creating music, we picked three music pieces made by our GAN program. We chose these pieces because they show the different kinds of music that the GAN can make. This helps us judge what GAN can do.

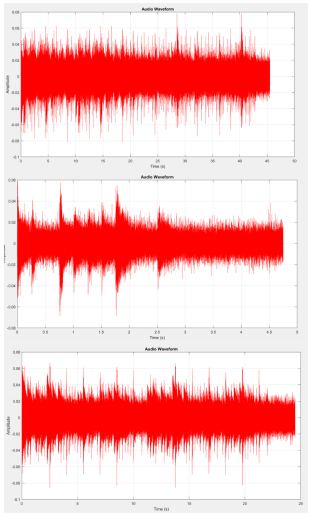

Fig. 6. Visual representation of music generated by a Generative Adversarial Network

C. *Long Short-Term Memory*

In the context of our research, we have dedicated significant attention to our algorithm, which has demonstrated superior performance in music generation. To offer a comprehensive evaluation of its capabilities, we meticulously curated and assessed three distinct music samples produced by our algorithm. These carefully chosen samples serve as exemplars of our algorithm's excellence in music generation, and we

present visual representations of these samples to illustrate our findings.

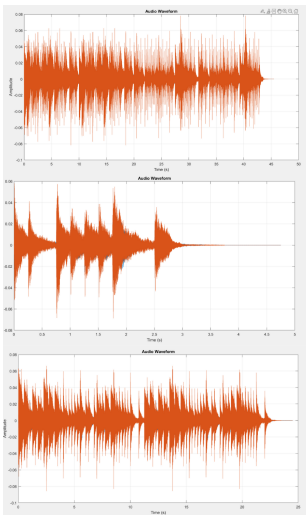

Fig. 7. Visual representation of music generated by a Long Short-Term Memory

Below we can see the detailed mathematical values of the signals that are generated by our algorithms for clear understanding

Algorithm	Energy	Mean
VAE	4.6085e+04 dBFS	0.7116 dB
VAE	1.2842e+04 dBFS	0.4031 dB
VAE	9.2387e+03 dbFS	0.7011 dB
GAN	150.3314 dbFS	-0.4634 dB
GAN	106.7336 dBFS	-0.4270 dB
GAN	113.0019 dBFS	-0.3910 dB
LSTM	92.0556 dBFS	-0.1270 dB
LSTM	4.7353 dBFS	-0.0973 dB
LSTM	49.3611 dBFS	-0.1079 dB

VI. Conclusion

A noteworthy finding has been made in the area of creating music using three different algorithms: Long Short-Term Memory networks (LSTMs), Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs), and visualizing the results through plotted waveforms in MATLAB. The generated music's noise characteristics were represented intelligibly and clearly by the plotted waveforms. Notably, the orange-colored music produced by LSTMs consistently displayed relatively lower noise levels and also energy. The red-hued GAN-generated music occupied a middle ground, with noise levels between LSTMs and VAEs. It's interesting to note that the music produced by VAEs, shown in blue, displayed a noticeably higher level of noise throughout the plotted waveforms. This visual and the mathematical distinction, which can be seen in the plotted music waveforms, highlights the crucial part that algorithm selection plays in the creation of music. It emphasizes the importance of making a thoughtful algorithm selection that is in line with particular quality and noise criteria, tailored to the desired musical outcome. In the end, this observation highlights how crucial it is to choose the right algorithm in order to generate music, with the visual representations serving as a useful tool for evaluating noise levels and overall musical fidelity.

References

[1] J. Vasudha, Iniya, S., Iyshwarya, G., and Dr. Jeyakumar G., "Computer Aided Music Generation Using Genetic Algorithm and Its Potential Applications ". 2011.

[2] R. Vinayakumar, Dr. Soman K. P., and Poornachandran, P., "Long Short-term Memory based Operation log Anomaly Detection", in 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017.

[3] Mukesh K, IppatapuVenkataSrisurya, ChereddySpandana, Anbazhagan E, I R Oviya(2022). Paper titled A VariationalAutoencoder – General Adversarial Networks (VAE-GAN) based model for ligand designing presented in the 5th International Conference on Innovative Computing and Communication (ICICC-2022), organized by Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi, India in association with National Institute of Technology Patna, India and Korea Institute of Digital Convergence, South Korea and University of Valladolid, Spain on 19 February, 2022 [SCOPUS indexed].

[4] Shah, Falak et al. "LSTM Based Music Generation." 2019 International Conference on Machine Learning and Data Engineering (iCMLDE) (2019): 48-53.

[5] Maduskar, Advait et al. "Music Generation using Deep Generative Modelling." 2020 International Conference on Convergence to Digital World - Quo Vadis (ICCDW) (2020): 1-4.

[6] Tang, Jiandong & Yin, Lanqing & Yu, Jinming. (2022). Generation of Western Piano Music Based on Deep Learning. 524-527. 10.1109/ISAIEE57420.2022.00113.

[7] J. Wang, X. Wang and J. Cai, "Jazz Music Generation Based on Grammar and LSTM," 2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 2019, pp. 115-120, doi: 10.1109/IHMSC.2019.00035.

[8] S. Sajad, S. Dharshika and M. Meleet, "Music Generation for Novices Using Recurrent Neural Network (RNN)," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2021, pp. 1-6, doi: 10.1109/ICSES52305.2021.9633906.

[9] Minu, R I & Nagarajan, Govidan & Bhatia, Rishabh & Kunar, Aditya. (2022). Music Generation Using Deep Learning. 10.1007/978-981-16-8739-6_54.

[10] S Sreenivasa Chakravarthi and R. Jagadeesh Kannan, "Non-linear Dimensionality Reduction-based Intrusion Detection using Deep Autoencoder" International Journal of Advanced Computer Science and Applications(IJACSA), 10(8), 2019

[11] A. D. Reddy, M. Kumar, A., Dr. Soman K. P., G.R., M. Reddy, V.S., R., and V.K., P., "LSTM based paraphrase identification using combined word embedding features", in Advances in Intelligent Systems and Computing, 2019, vol. 898, pp. 385-394.

[12] S. Viswanathan, M. Kumar, A., and Dr. Soman K. P., "A Sequence-Based Machine Comprehension Modeling Using LSTM and GRU", in Lecture Notes in Electrical Engineering, 2019, vol. 545, pp. 47-55.

[13] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, Antonio Torralba; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4502-4511

[14] Choi, Keunwoo, György Fazekas, Mark Sandler, and Kyunghyun Cho. "A comparison of audio signal preprocessing methods for deep neural networks on music tagging." In 2018 26th European Signal Processing Conference (EUSIPCO), pp. 1870-1874. IEEE, 2018.

[15] Yu, Yong, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. "A review of recurrent neural networks: LSTM cells and network architectures." Neural computation 31, no. 7 (2019): 1235-1270.