






Exploring AI Music Generation: A Review of Deep Learning Algorithms and Datasets for Undergraduate Researchers

Isshin Yunoki^(✉), Guy Berreby, Nicholas D'Andrea, Yuhua Lu,
and Xiaodong Qu

Swarthmore College, Swarthmore, PA 19081, USA
{iyunoki1,gberreb1,ndandre1,ylu3,xqu1}@swarthmore.edu

Abstract. This review paper presents an exploration of the deep learning-based music generation literature, designed to offer undergraduate researchers an initiation into the field. This study illustrates prevailing generative models and datasets currently influential in music generation. Four publications have been selected for detailed discussion, representing a spectrum of salient concepts in music generation and potential areas of further inquiry. By focusing on key studies and significant datasets, this review aspires to serve as a guide for undergraduate scholars keen on investigating the intersections of deep learning and music generation.

Keywords: Deep Learning · Generative Model · Transformer · Time Series · Music · MIDI · Audio

1 Introduction

Deep learning, in recent years, has seen transformative progress across various disciplines, notably in the sphere of AI music generation. Current music generation research focuses on generating original and high-quality compositions, which relies on two key properties: Structural awareness and interpretive ability. Structural awareness enables models to generate naturally coherent music with long-term dependencies, including repetition and variation. Interpretive ability involves translating complex computational models into interactive interfaces for controllable and expressive performances [106]. Furthermore, AI music generation exhibits generality, allowing the same algorithm to be applied across different genres and datasets, enabling exploration of various musical styles. [6]

The landscape of music generation presents a unique opportunity for novice researchers to explore and contribute to the field of generative AI. However, navigating the vast amount of research and staying up to date can be challenging. Our survey aims to assess the accessibility and feasibility of music generation

algorithms, providing guidance for undergraduate researchers to establish a solid foundation in this exciting area of study.

We have collected notable research on automatic music generation to provide a starting point for researchers to further investigate. By systematizing each study’s algorithms and datasets, researchers can identify effective architecture-data pairings. We also select and explain four papers that best represent fundamental ideas and popular techniques in music generation. By engaging the factual information in our resources, researchers can learn and further explore the structural awareness properties of music generation.

Moreover, we include content about the interpretive ability of music generation, allowing connections to be made with the area of human interaction. There is vast potential that lies within the user experience and interfaces side of AI music generation: Researchers can explore how algorithms can be designed to be user-friendly and accessible to a human composer, as well as how these algorithms can be integrated into a creative workflow as a powerful tool to expand and enhance musical ideas.

By studying our resources and content, researchers can find new ideas and draw original connections, enabling them to make significant contributions to the music generation field.

1.1 Related Works

Several review papers provide valuable insights into the trends and methodologies for music generation, such as [6, 42, 99, 106]. We further expand our analysis by exploring a range of scholarly articles that employ analogous time-series data [4, 11, 18, 19, 29, 46, 50, 57–62, 73, 75–77, 84, 87, 97, 101–104, 107–109]. The literature highlights that algorithmic performance depends on various factors, including the training data’s quality and diversity, the model’s complexity, and the strategies employed by the model. When it comes to training data, music can be represented in two main formats: Symbolic and raw audio files.

Symbolic audio refers to encoding music information through symbols that represent different aspects of music. The most common form of symbolic music data is MIDI (Music Instrument Digital Interface) which uses discrete values to represent the note pitches and their duration [106].

Raw audio refers to any music file format which encodes an actual audio signal. Such file formats include MP3 files, .WAV files, .FLAC files, and others, which can be used for training algorithms. Raw audio has the advantage of representing expressive characteristics inherent to the original music data, at the cost of being computationally demanding. [6]

1.2 Research Questions

Our investigation seeks to answer the following questions:

- What are recent trends in music generation research?
- Which papers should undergraduate researchers read to gain a thorough understanding of AI music generation?
- Which algorithms and datasets are suited for undergraduate-level research?

2 Methods

2.1 Search Methods for Identification of Studies

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) method was employed to select relevant studies for this review. The search was conducted between January and March 2023 on Google Scholar, Papers with Code, and arXiv. A combination of keywords was used: ('Deep Learning' OR 'Generative Model' OR 'LSTM' OR 'GAN' OR 'VAE' OR 'Transformer' OR 'Attention') AND ('Time Series' OR 'Music' OR 'Music Theory' OR 'jazz' OR 'MIDI' OR 'Audio'). The process of identifying and refining the study collection is illustrated in Fig. 1.

The criteria for selecting appropriate papers are as follows:

- **Task:** Focus on studies with the objective of AI music generation, specifically those that create new music with high fidelity and long-term consistency using existing data. Research on outside areas and tasks such as genre classification, computer vision, and medical applications are excluded.
- **Deep Learning:** Limit the scope to studies that employ deep learning, defined in this review as having multiple layers with more than one hidden layer. Single-layer generative models are excluded.
- **Transformers:** To assist researchers in their exploration of music generation, we recognize the potential for innovation in transformer-based algorithms. As such, we select prominent papers for researchers to learn more about the advancements and capabilities of transformers.

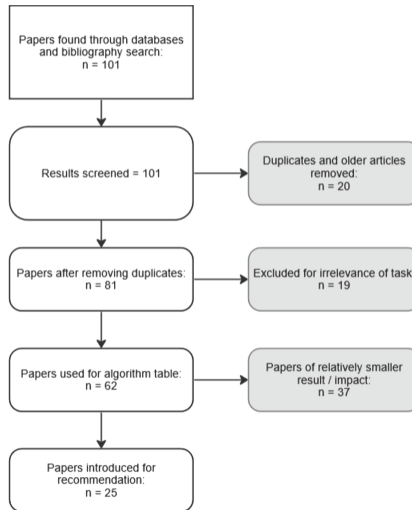


Fig. 1. Selection process for the papers

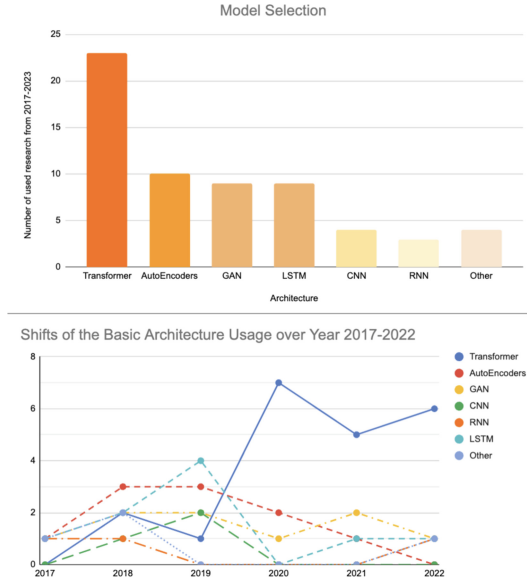


Fig. 2. Visualizations

- **Model Availability:** Prioritize research with publicly available models and datasets. This excludes public music generation websites, as their data and architectures are often undisclosed and subject to change.
- **Target Audience:** The review is tailored for those interested in undergraduate-level research of the intersection of computer science and music. We provide an overview of music generation for students to conduct their own research, developing their skills and knowledge in the field while considering the limited time and experience available to them.
- **Time:** The review only includes studies published in 2017 and onward to account for the rapid progress of the field.

3 Results

We categorize 62 papers' essential algorithms and draw an overall conclusion on their strengths and weaknesses towards the task of music generation. The trends in the model selection are visualized in Fig. 2; transformers have gained significant attention since 2018, while LSTM has shown a decline over the years. AEs and GANs are widely and steadily utilized. This information is summarized in Fig. 3; transformer-based architectures were most commonly used, followed by AEs, GANs, and LSTM neural networks. We also created a list of prominent datasets, in Fig. 4.

4 Discussion

4.1 Papers

We have identified four essential papers that we recommend undergraduate students read to gain an understanding of popular generative architectures.

- Huang et al. [44] introduce a foundational architecture using transformers on MIDI data. Specifically, their architecture is similar to that of the original transformer paper, except they add an innovation they call relative attention. Relative attention modifies the attention mechanism by taking into account how close or far apart two elements of the midi sequence are when determining attention coefficients. This allows for the transformer to generate music which makes more coherent sense on small timescales.
- Dhariwal et al. [21] present an effective combination of the transformer and VAE which operates on raw audio. The overall architecture of their Jukebox model is that of a VQ-VAE. The model has three levels of VQ-VAEs, each of which independently encodes the input data. Thinking of these levels as being vertically stacked, the topmost level is the coarsest, encoding only high

Algorithms	Paper (year published)	Strengths	Weaknesses
Transformer	Hawthorne et al. (2018)	<ul style="list-style-type: none"> - Highly parallelizable, enabling fast training - Strong long-range dependency modeling due to attention mechanism - State-of-the-art performance in various tasks 	<ul style="list-style-type: none"> - Can require significant amounts of memory and computational resources - Often require large-scale pre-training for best performance
	Huang et al. (2018)		
	Donahue et al. (2019)		
	Fris and Paquinier (2020)		
	Palea et al. (2020)		
	Jiang et al. (2020)		
	Dhariwal et al. (2020)		
	Zhang (2020)		
	Huang and Yang (2020)		
	Wu and Yang (2020)		
	Yu and Yang (2021)		
	Vierma and Chale (2021)		
	Muhammad et al. (2021)		
	Choi et al. (2021)		
Di et al. (2021)			
Santos et al. (2022)			
Min et al. (2022)			
Santos et al. (2022)			
Yu et al. (2022)			
Shih et al. (2022)			
Dong et al. (2022)			
Sarmiento et al. (2023)			
Agostinelli et al. (2023)			
AutoEncoders	Engel et al. (2017)	<ul style="list-style-type: none"> - Unsupervised learning, requiring no labeled data - VAE can generate diverse, novel outputs by sampling from latent space - Enable efficient representation learning 	<ul style="list-style-type: none"> - Can struggle to generate high-quality outputs, especially with complex data like music - Latent space may not capture all essential features
	Dieleman et al. (2018)		
	Alkhat and Liang (2018)		
	Roberts et al. (2018)		
	Liang et al. (2019)		
	Liang et al. (2019)		
Lattner and Grachten (2019)			
Huang and Huang (2020)			
Choi et al. (2020)			
Grekow and Dimitrova-Grekow (2021)			
GAN	Yang et al. (2017)	<ul style="list-style-type: none"> - Adversarial training enables implicit modeling of data distribution - Encourages creativity in generated outputs 	<ul style="list-style-type: none"> - Training can be unstable and challenging to converge - Mode collapse: GAN may only learn to generate a limited set of outputs
	Dong et al. (2018)		
	Dong and Yang (2018)		
	Jamrani and Berg-Kirkpatrick (2019)		
	Guan et al. (2019)		
LSTM	Nisal et al. (2020)	<ul style="list-style-type: none"> - Capable of learning long-range dependencies better than plain RNNs - Widely used and proven effective in various music generation tasks - Mitigate vanishing and exploding gradient problems with specialized gating mechanisms 	<ul style="list-style-type: none"> - Training can be slow and resource-intensive - Can be more complex to implement and optimize compared to simpler models like RNNs or CNNs
	Brook and Komeet (2021)		
	Caldiz et al. (2021)		
	Tomaz et al. (2022)		
	Brunner et al. (2017)		
	Mao et al. (2018)		
	Défossez et al. (2018)		
Zhao et al. (2019)			
Wang et al. (2019)			
Mangal et al. (2019)			
Gillick et al. (2019)			
Ferrera and Whitehead (2021)			
Keerti et al. (2022)			
CNN	Ehbayaci et al. (2018)	<ul style="list-style-type: none"> - Effective in capturing local and hierarchical features - Invariant to translations, which can be helpful to certain musical tasks 	<ul style="list-style-type: none"> - Limited in modeling long-range dependencies due to fixed-size convolutional filters - Not explicitly designed for sequential data like music
	Schlimbich et al. (2019)		
	Wang and Yang (2019)		
	Nishihara et al. (2023)		
RNN	Hadjeres and Nielson (2017)	<ul style="list-style-type: none"> - Designed for sequential data, making them suitable for music generation - Can model arbitrary-length input and output sequences 	<ul style="list-style-type: none"> - Struggle to learn long-range dependencies due to vanishing or exploding gradient problem - Slower training compared to other models, as processing is inherently sequential
	Car and Zakowski (2018)		
	Jaganathan et al. (2022)		
Other	Yanachenko and Makherjee (2017)	<ul style="list-style-type: none"> - Includes: State Space Model, WaveNet, Moment Matching-Scattering Inverse Network (MM-SIN) 	
	Andrux and Müller (2018)		
	Manzelli et al. (2018)		
	Goel et al. (2022)		

*Categorization is not exhaustive, as many researchers combine multiple models or utilize parts of a model.

Fig. 3. Algorithm usage by papers

Datasets	Description	Original Paper	Citation Number (as of Apr. 2023)
JSB Chorales	MIDI versions of Bach Chorales, short pieces of 4-part vocal music	Boulangier-Lewandowski Bengio, and Vincent 2012	848
MAESTRO	200 hours of both MIDI and raw audio versions of piano pieces	Hawthorne et al. 2019	333
Lakh MIDI dataset	176,581 unique MIDI files	Raffel 2016	279
MagnaTagATune	200 hours of music audio, consisting of 29 second clips (raw audio)	Edith Law and Downie 2009	169
REMI	MIDI-based event sequences; converted files of many popular songs	Huang and Yang 2020	122
Beethoven	Raw Audio recordings of 32 Beethoven piano sonatas	Neuwirth et al. 2018	69
POP909	Multiple versions of piano arrangements of popular songs	Wang et al. 2020	60
The Weimar Jazz Database	MIDI transcriptions of Jazz solos	Pfleiderer et al. 2017	49
GiantMIDI-Piano	10,855 MIDI files of classical piano music	Kong et al. 2020	46
MusicCaps	5,521 music samples, each labeled in English	Agostinelli et al. 2023	31
EWLD	A MIDI dataset of over 5000 music lead sheets	Simonetta et al. 2018	28
NES-MDB	5287 songs from 397 NES games soundtrack in a few symbolic formats	Donahue et al. 2018	23
DadaGP	26,181 Guitar Pro files in 739 genres, a symbolic dataset	Sarmiento et al. 2021	12
840 Piano covers of popular songs	840 Piano covers of popular songs, ranging over 56 hours (raw audio)	Verma and Chafe 2021	11
Projective Orchestral Database	196 pairs of midi files having a piano and orchestral version of music	Crestel et al. 2018	11

Fig. 4. Datasets used by papers

level essential information, while the lowest level encodes the fine details of the music. With these latent spaces, they then train sparse transformers that upsample from a higher level latent space to a lower one. So, to generate music, a sample datapoint from the latent space of the uppermost VQ-VAE, uses transformers to up-sample the datapoint to the latent space of the lower level VQ-VAEs, and then once at the lowest level, use the VQ-VAE decoder to turn the upsampled datapoint into raw audio.

- Dong et al. [27] introduce MuseGan, a GAN architecture for symbolic multi-track piano roll data. MuseGan employs a WGAN-GP framework, which includes modified objective functions and a gradient penalty for the discriminator, leading to faster convergence and reduced parameter tuning. The model consists of two components: the Multitrack model and the temporal model. The Multitrack model incorporates GAN submodels based on three compositional approaches: jamming, composing, and a hybrid of both. Discriminators within these submodels evaluate the specific characteristics of each track. The temporal model comprises two submodels: one for generation from scratch, capturing temporal and bar-by-bar information, and another

for track-conditional generation, using conditional track inputs to generate sequential bars. By combining these models, MuseGan produces latent vectors that incorporate inter-track and temporal information. These vectors are then used to generate piano rolls sequentially.

- Huang and Yang [47] provide a promising direction of data conversion with their introduction of REMI (revamped MIDI-derived events). Instead of the traditional MIDI-based music representations, REMI describes musical events with further details to represent the original music with more information. Specifically, REMI adds tempo and chords as part of the data, reinterprets the time grid of the music data from second-based to position- and bar-based, and describes the note duration instead of the ending position of the note for note lengths. REMI helped a transformer-based model output samples with stronger sense of downbeat and natural and expressive uses of chords. The paper also introduces Pop Music Transformer, a transformer-based architecture for music generation. This model differs from traditional transformer models in that it learns to compose music over a metrical structure defined in terms of bars, beats, and sub-beats, through the application of the aforementioned REMI. This approach allows the model to generate music with a more salient and consistent rhythmic structure, and produce musically pleasing pieces without human intervention.

4.2 Algorithms

Our survey suggests that the main algorithms used for music generation in the last five years are transformers and autoencoders such as VAEs, GANS, and LSTMs, with transformers being by far the most popular. Due to the popularity of transformers and their success with music generation, we have decided to focus on them for much of our analysis.

Transformers are applicable in symbolic and raw audio domains with convincing results, offering flexibility for researchers to pursue their research interests. Also, the literature shows transformers have broad functionalities and involve specific components and mechanisms makes them worth exploring individually.

These components can be fine tuned and altered to match the needs of a given task. For example, [41, 43] use a modification known as relative attention, which modifies the attention coefficients based on how far apart two tokens are. There is also the transformer-XL modification used by [24]. This modification adds a recurrence mechanism to hidden states within the Transformer, and has been shown to increase performance [17]. Others, such as [21] used sparse transformers. Sparse transformers introduce sparsity to the attention heads of the transformer, reducing $O(n^2)$ time and memory costs to $O(n\sqrt{n})$ [13].

4.3 Datasets

A variety of both symbolic and raw audio datasets have been used by transformer papers. We observe that certain datasets work best with transformers in capturing complex and long-range dependencies within music sequences:

The LakhMIDI dataset, the largest dataset of symbolic data available, has great potential due to its large training size and MIDI-audio pairings, and was the most popular among transformer papers we surveyed, with five different papers using it. In [26, 34, 78], the authors use LakhMIDI to derive token sequences from MIDI files to create multitrack music transformer models. In [94], a multi-track pianoroll dataset derived from LakhMIDI is used for a transformer as well. [24] maps LakhMIDI tracks onto instrumentation playable by the NES, and then uses a transformer to generate NES versions of songs.

Similar in size and function is the MAESTRO dataset, which is used by the transformer models of [41, 67], and is the second most popular dataset among transformer papers. The MAESTRO dataset contains over 200 h of piano performances, stored in both raw audio and MIDI formats. The MIDI-audio pairings enables music information to be retrieved from the MIDI files and be used as annotations for the matched audio files. Every file is labeled, allowing for easy supervised training.

Besides LakhMIDI and MAESTRO, every other transformer paper we found used a different dataset. In fact, a popular choice among them was to create or scrape their own dataset, which was done by [21, 47, 89, 100].

Overall, for undergraduates working with Transformers, if working with symbolic data we would recommend using the Lakh MIDI dataset due to its large size and relative popularity among other users of transformers. For those who want to use raw audio, we would recommend the MAESTRO dataset, similarly for its large size and popularity.

4.4 Future Work

The future development of music generation technology is increasingly focused on enhancing the ability to control models structurally. In later work, more research and analysis could investigate the data and model decisions behind each study, as well as increase the target research, to better comprehend factors contributing to successful music generation outcomes.

Moving forward, we encourage undergraduate researchers to engage in more experimental and collaborative work, exploring the combination of different algorithms and datasets to develop new approaches to music generation.

5 Conclusion

In conclusion, our review provides a survey of deep learning algorithms and datasets for music generation, with the aim to assist undergraduate researchers interested in the field. Our findings suggest that in the last five years, transformers, GANS, autoencoders, and LSTMs have been the primary algorithms used for AI music generation, with transformers gaining significant popularity in more recent times. We find that the papers use a wide variety of datasets, meaning there is no one single, predominant dataset being used. We suggest four papers that we believe are important for undergraduates to read to get a

solid grasp of the field, and also recommend an algorithm along with datasets for undergraduates to use for their initial adventures into AI music generation.

Author contributions. Isshin, Guy, Nicholas, and Yuhua are the first four authors of this paper, and they contributed equally. Professor Xiaodong Qu is the mentor for this research project.

References

1. Agostinelli, A., et al.: MusicLM: generating music from text (2023)
2. Akbari, M., Liang, J.: Semi-recurrent CNN-based VAE-GAN for sequential data generation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2321–2325. IEEE (2018)
3. Andreux, M., Mallat, S.: Music generation and transformation with moment matching-scattering inverse networks. In: ISMIR, pp. 327–333 (2018)
4. Basaklar, T., Tuncel, Y., An, S., Ogras, U.: Wearable devices and low-power design for smart health applications: challenges and opportunities. In: 2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), p. 1. IEEE (2021)
5. Boulanger-Lewandowski, N., Bengio, Y., Vincent, P.: Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription (2012)
6. Briot, J.P.: From artificial neural networks to deep learning for music generation: history, concepts and trends. *Neural Comput. Appl.* **33**(1), 39–65 (2021)
7. van den Broek, K.: Mp3net: coherent, minute-long music generation from raw audio with a simple convolutional GAN. arXiv e-prints, pp. arXiv-2101 (2021)
8. Brunner, G., Wang, Y., Wattenhofer, R., Wiesendanger, J.: JamBot: music theory aware chord based generation of polyphonic music with LSTMs. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 519–526. IEEE (2017)
9. Cádiz, R.F., Macaya, A., Cartagena, M., Parra, D.: Creativity in generative musical networks: evidence from two case studies. *Front. Robot. AI* **8**, 680586 (2021)
10. Carr, C., Zukowski, Z.: Generating albums with SampleRNN to imitate metal, rock, and punk bands. arXiv preprint [arXiv:1811.06633](https://arxiv.org/abs/1811.06633) (2018)
11. Chen, L., et al.: Data-driven detection of subtype-specific differentially expressed genes. *Sci. Rep.* **11**(1), 332 (2021)
12. Chen, Y.H., Wang, B., Yang, Y.H.: Demonstration of PerformanceNet: a convolutional neural network model for score-to-audio music generation, pp. 6506–6508 (2019). <https://doi.org/10.24963/ijcai.2019/938>
13. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers (2019)
14. Choi, K., Hawthorne, C., Simon, I., Dinulescu, M., Engel, J.: Encoding musical style with transformer autoencoders. In: International Conference on Machine Learning, pp. 1899–1908. PMLR (2020)
15. Choi, K., Park, J., Heo, W., Jeon, S., Park, J.: Chord conditioned melody generation with transformer based decoders. *IEEE ACCESS* **9**, 42071–42080 (2021). <https://doi.org/10.1109/ACCESS.2021.3065831>
16. Crestel, L., Esling, P., Heng, L., McAdams, S.: A database linking piano and orchestral midi scores with application to automatic projective orchestration (2018)

17. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-XL: attentive language models beyond a fixed-length context, pp. 2978–2988 (2019). <https://doi.org/10.18653/v1/P19-1285>
18. Deb, R., An, S., Bhat, G., Shill, H., Ogras, U.Y.: A systematic survey of research trends in technology usage for Parkinson’s disease. *Sensors* **22**(15), 5491 (2022)
19. Deb, R., Bhat, G., An, S., Shill, H., Ogras, U.Y.: Trends in technology usage for Parkinson’s disease assessment: a systematic review. *MedRxiv*, pp. 2021–02 (2021)
20. Défossez, A., Zeghidour, N., Usunier, N., Bottou, L., Bach, F.: Sing: Symbol-to-instrument neural generator. In: *Advances in Neural Information Processing Systems*, vol. 31 (2018)
21. Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A., Sutskever, I.: Jukebox: a generative model for music. arXiv preprint [arXiv:2005.00341](https://arxiv.org/abs/2005.00341) (2020)
22. Di, S., et al.: Video background music generation with controllable music transformer. In: *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2037–2045 (2021)
23. Dieleman, S., van den Oord, A., Simonyan, K.: The challenge of realistic music generation: modelling raw audio at scale. In: *Advances in Neural Information Processing Systems*, vol. 31 (2018)
24. Donahue, C., Mao, H.H., Li, Y.E., Cottrell, G.W., McAuley, J.: Lakhnes: improving multi-instrumental music generation with cross-domain pre-training. arXiv preprint [arXiv:1907.04868](https://arxiv.org/abs/1907.04868) (2019)
25. Donahue, C., Mao, H.H., McAuley, J.: The NES music database: a multi-instrumental dataset with expressive performance attributes. In: *ISMIR* (2018)
26. Dong, H.W., Chen, K., Dubnov, S., McAuley, J., Berg-Kirkpatrick, T.: Multitrack music transformer (2022)
27. Dong, H.W., Hsiao, W.Y., Yang, L.C., Yang, Y.H.: MuseGAN: multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018)
28. Dong, H.W., Yang, Y.H.: Convolutional generative adversarial networks with binary neurons for polyphonic music generation. arXiv preprint [arXiv:1804.09399](https://arxiv.org/abs/1804.09399) (2018)
29. Dou, G., Zhou, Z., Qu, X.: Time majority voting, a PC-based EEG classifier for non-expert users. In: Kurosu, M., et al. *HCI International 2022-Late Breaking Papers. Multimodality in Advanced Interaction Environments. HCII 2022. LNCS*, vol. 13519, pp. 415–428. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17618-0_29
30. Edith Law, Kris West, M.M.: Evaluation of algorithms using games: the case of music annotation. In: *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)* (2009)
31. Elbayad, M., Besacier, L., Verbeek, J.: Pervasive attention: 2D convolutional neural networks for sequence-to-sequence prediction (2018)
32. Engel, J., Agrawal, K.K., Chen, S., Gulrajani, I., Donahue, C., Roberts, A.: Gansynth: Adversarial neural audio synthesis. [arXiv:1902.08710](https://arxiv.org/abs/1902.08710) (2019)
33. Engel, J., et al.: Neural audio synthesis of musical notes with WaveNet autoencoders. In: *International Conference on Machine Learning*, pp. 1068–1077. PMLR (2017)
34. Ens, J., Pasquier, P.: MMM: exploring conditional multi-track music generation with the transformer. arXiv preprint [arXiv:2008.06048](https://arxiv.org/abs/2008.06048) (2020)

35. Ferreira, L.N., Whitehead, J.: Learning to generate music with sentiment. arXiv preprint [arXiv:2103.06125](https://arxiv.org/abs/2103.06125) (2021)
36. Gillick, J., Roberts, A., Engel, J., Eck, D., Bamman, D.: Learning to groove with inverse sequence transformations. In: International Conference on Machine Learning, pp. 2269–2279. PMLR (2019)
37. Goel, K., Gu, A., Donahue, C., Ré, C.: It's raw! Audio generation with state-space models. In: International Conference on Machine Learning, pp. 7616–7633. PMLR (2022)
38. Grekow, J., Dimitrova-Grekow, T.: Monophonic music generation with a given emotion using conditional variational autoencoder. *IEEE Access* **9**, 129088–129101 (2021)
39. Guan, F., Yu, C., Yang, S.: A GAN model with self-attention mechanism to generate multi-instruments symbolic music. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–6 (2019). <https://doi.org/10.1109/IJCNN.2019.8852291>
40. Hadjeres, G., Nielsen, F.: Interactive music generation with positional constraints using anticipation-RNNs. arXiv preprint [arXiv:1709.06404](https://arxiv.org/abs/1709.06404) (2017)
41. Hawthorne, C., et al.: Enabling factorized piano music modeling and generation with the maestro dataset (2019)
42. Hernandez-Olivan, C., Beltran, J.R.: Music composition with deep learning: a review. In: Advances in Speech and Music Technology: Computational Aspects and Applications, pp. 25–50 (2022)
43. Huang, C.A., et al.: An improved relative self-attention mechanism for transformer with application to music generation. CoRR abs/1809.04281 (2018). <http://arxiv.org/abs/1809.04281>
44. Huang, C.Z.A., et al.: Music transformer. arXiv preprint [arXiv:1809.04281](https://arxiv.org/abs/1809.04281) (2018)
45. Huang, C.F., Huang, C.Y.: Emotion-based AI music generation system with CVAE-GAN. In: 2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE), pp. 220–222. IEEE (2020)
46. Huang, D., Tang, Y., Qin, R.: An evaluation of PlanetScope images for 3d reconstruction and change detection-experimental validations with case studies. *GISci. Remote Sens.* **59**(1), 744–761 (2022)
47. Huang, Y.S., Yang, Y.H.: Pop music transformer: beat-based modeling and generation of expressive pop piano compositions, pp. 1180–1188 (2020). <https://doi.org/10.1145/3394171.3413671>
48. Jagannathan, A., Chandrasekaran, B., Dutta, S., Patil, U.R., Eirinaki, M.: Original music generation using recurrent neural networks with self-attention. In: 2022 IEEE International Conference On Artificial Intelligence Testing (AITest), pp. 56–63. IEEE (2022)
49. Jhamtani, H., Berg-Kirkpatrick, T.: Modeling self-repetition in music generation using generative adversarial networks. In: Machine Learning for Music Discovery Workshop, ICML (2019)
50. Jiang, C., et al.: Deep denoising of raw biomedical knowledge graph from COVID-19 literature, LitCovid, and PubTator: framework development and validation. *J. Med. Internet Res.* **24**(7), e38584 (2022)
51. Jiang, J., Xia, G.G., Carlton, D.B., Anderson, C.N., Miyakawa, R.H.: Transformer vae: a hierarchical model for structure-aware and interpretable music representation learning. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 516–520. IEEE (2020)

52. Keerti, G., Vaishnavi, A., Mukherjee, P., Vidya, A.S., Sreenithya, G.S., Nayab, D.: Attentional networks for music generation. *Multimed. Tools Appl.* **81**(4), 5179–5189 (2022)
53. Kong, Q., Li, B., Chen, J., Wang, Y.: GiantMIDI-Piano: a large-scale midi dataset for classical piano music (2022)
54. Lattner, S., Grachten, M.: High-level control of drum track generation using learned patterns of rhythmic interaction. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 35–39. IEEE (2019)
55. Liang, X., Wu, J., Cao, J.: MIDI-sandwich2: RNN-based hierarchical multi-modal fusion generation VAE networks for multi-track symbolic music generation. *CoRR* abs/1909.03522 (2019). <http://dblp.uni-trier.de/db/journals/corr/corr1909.html#abs-1909-03522>
56. Liang, X., Wu, J., Yin, Y.: MIDI-sandwich: multi-model multi-task hierarchical conditional VAE-GAN networks for symbolic single-track music generation. arXiv preprint [arXiv:1907.01607](https://arxiv.org/abs/1907.01607) (2019)
57. Liu, C., Li, H., Xu, J., Gao, W., Shen, X., Miao, S.: Applying convolutional neural network to predict soil erosion: a case study of coastal areas. *Int. J. Environ. Res. Public Health* **20**(3), 2513 (2023)
58. Lu, Y., Wang, H., Wei, W.: Machine learning for synthetic data generation: a review. arXiv preprint [arXiv:2302.04062](https://arxiv.org/abs/2302.04062) (2023)
59. Lu, Y., et al.: Cot: an efficient and accurate method for detecting marker genes among many subtypes. *Bioinform. Adv.* **2**(1), vbac037 (2022)
60. Luo, X., Ma, X., Munden, M., Wu, Y.J., Jiang, Y.: A multisource data approach for estimating vehicle queue length at metered on-ramps. *J. Transp. Eng. Part A Syst.* **148**(2), 04021117 (2022)
61. Ma, X.: Traffic Performance Evaluation Using Statistical and Machine Learning Methods. Ph.D. thesis, The University of Arizona (2022)
62. Ma, X., Karimpour, A., Wu, Y.J.: Statistical evaluation of data requirement for ramp metering performance assessment. *Transp. Res. Part A Policy Pract.* **141**, 248–261 (2020)
63. Mangal, S., Modak, R., Joshi, P.: LSTM based music generation system. arXiv preprint [arXiv:1908.01080](https://arxiv.org/abs/1908.01080) (2019)
64. Manzelli, R., Thakkar, V., Siahkamari, A., Kulis, B.: An end to end model for automatic music generation: combining deep raw and symbolic audio networks. In: *Proceedings of the Musical Metacreation Workshop at 9th International Conference on Computational Creativity, Salamanca, Spain* (2018)
65. Mao, H.H., Shin, T., Cottrell, G.: DeepJ: style-specific music generation. In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp. 377–382. IEEE (2018)
66. Min, J., Liu, Z., Wang, L., Li, D., Zhang, M., Huang, Y.: Music generation system for adversarial training based on deep learning. *Processes* **10**(12), 2515 (2022)
67. Muhamed, A., et al.: Symbolic music generation with transformer-GANs. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 408–417 (2021)
68. Neuwirth, M., Harasim, D., Moss, F., Rohrmeier, M.: The annotated Beethoven corpus (ABC): a dataset of harmonic analyses of all Beethoven string quartets. *Front. Digital Human.* **5**, 16 (2018). <https://doi.org/10.3389/fdigh.2018.00016>
69. Nishihara, M., Hono, Y., Hashimoto, K., Nankaku, Y., Tokuda, K.: Singing voice synthesis based on frame-level sequence-to-sequence models considering vocal timing deviation. arXiv preprint [arXiv:2301.02262](https://arxiv.org/abs/2301.02262) (2023)

70. Nistal, J., Lattner, S., Richard, G.: DrumGAN: synthesis of drum sounds with timbral feature conditioning using generative adversarial networks. arXiv preprint [arXiv:2008.12073](https://arxiv.org/abs/2008.12073) (2020)
71. van den Oord, A., et al.: Wavenet: a generative model for raw audio (2016)
72. Palea, D., Zhou, H.H., Gupta, K.: Transformer bard: music and poem generation using transformer models (2020)
73. Peng, X., Bhattacharya, T., Mao, J., Cao, T., Jiang, C., Qin, X.: Energy-efficient management of data centers using a renewable-aware scheduler. In: 2022 IEEE International Conference on Networking, Architecture and Storage (NAS), pp. 1–8. IEEE (2022)
74. Pfeleiderer, M., Frieler, K., Abeßer, J., Zaddach, W.G., Burkhart, B. (eds.): Inside the Jazzomat - New Perspectives for Jazz Research. Schott Campus (2017)
75. Qu, X., Liu, P., Li, Z., Hickey, T.: Multi-class time continuity voting for EEG classification. In: Frasson, C., Bamidis, P., Vlamos, P. (eds.) BFAL 2020. LNCS (LNAI), vol. 12462, pp. 24–33. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60735-7_3
76. Qu, X., Liukasemsarn, S., Tu, J., Higgins, A., Hickey, T.J., Hall, M.H.: Identifying clinically and functionally distinct groups among healthy controls and first episode psychosis patients by clustering on eeg patterns. *Front. Psych.* **11**, 541659 (2020)
77. Qu, X., Mei, Q., Liu, P., Hickey, T.: Using EEG to distinguish between writing and typing for the same cognitive task. In: Frasson, C., Bamidis, P., Vlamos, P. (eds.) BFAL 2020. LNCS (LNAI), vol. 12462, pp. 66–74. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60735-7_7
78. Raffel, C.: Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching (2016). <https://colinraffel.com/projects/lmd/>
79. Roberts, A., Engel, J., Raffel, C., Hawthorne, C., Eck, D.: A hierarchical latent vector model for learning long-term structure in music. In: International Conference on Machine Learning, pp. 4364–4373. PMLR (2018)
80. Santos, G.A.C., Baffa, A., Briot, J.P., Feijó, B., Furtado, A.L.: An adaptive music generation architecture for games based on the deep learning transformer mode. arXiv preprint [arXiv:2207.01698](https://arxiv.org/abs/2207.01698) (2022)
81. Sarmiento, P., Kumar, A., Carr, C., Zukowski, Z., Barthet, M., Yang, Y.H.: DadaGP: a dataset of tokenized GuitarPro songs for sequence models. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference (2021). <https://archives.ismir.net/ismir2021/paper/000076.pdf>
82. Sarmiento, P., Kumar, A., Chen, Y.H., Carr, C., Zukowski, Z., Barthet, M.: GTR-CTRL: instrument and genre conditioning for guitar-focused music generation with transformers. In: Johnson, C., Rodríguez-Fernández, N., Rebelo, S.M. (eds) Artificial Intelligence in Music, Sound, Art and Design. EvoMUSART 2023. LNCS, vol. 13988, pp. 260–275. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-29956-8_17
83. Schimbinschi, F., Walder, C., Erfani, S.M., Bailey, J.: Synthnet: learning to synthesize music end-to-end. In: IJCAI, pp. 3367–3374 (2019)
84. Shen, X., Sun, Y., Zhang, Y., Najmabadi, M.: Semi-supervised intent discovery with contrastive learning. In: Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pp. 120–129 (2021)
85. Shih, Y.J., Wu, S.L., Zalkow, F., Muller, M., Yang, Y.H.: Theme transformer: symbolic music generation with theme-conditioned transformer. *IEEE Trans. Multimed.* **14**, 1–14 (2022)

86. Simonetta, F., Carnovalini, F., Orio, N., Roda, A.: Symbolic music similarity through a graph-based representation (2018). <https://doi.org/10.1145/3243274.3243301>
87. Tang, Y., Song, S., Gui, S., Chao, W., Cheng, C., Qin, R.: Active and low-cost hyperspectral imaging for the spectral analysis of a low-light environment. *Sensors* **23**(3), 1437 (2023)
88. Tomaz Neves, P.L., Fornari, J., Batista Florindo, J.: Self-attention generative adversarial networks applied to conditional music generation. *Multimed. Tools Appl.* **81**(17), 24419–24430 (2022)
89. Verma, P., Chafe, C.: A generative model for raw audio using transformer architectures. In: 2021 24th International Conference on Digital Audio Effects (DAFx), pp. 230–237. IEEE (2021)
90. Wang, B., Yang, Y.H.: Performancenet: score-to-audio music generation with multi-band convolutional residual network. In: Proceedings of the AAAI Conference on Artificial Intelligence 33(01), 1174–1181 (2019). <https://doi.org/10.1609/aaai.v33i01.33011174>, <https://ojs.aaai.org/index.php/AAAI/article/view/3911>
91. Wang, J., Wang, X., Cai, J.: Jazz music generation based on grammar and LSTM. In: 2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 1, pp. 115–120. IEEE (2019)
92. Wang*, Z., Chen*, K., Jiang, J., Zhang, Y., Xu, M., Dai, S., Bin, G., Xia, G.: Pop909: a pop-song dataset for music arrangement generation. In: Proceedings of 21st International Conference on Music Information Retrieval, ISMIR (2020)
93. Wu, S.L., Yang, Y.H.: The jazz transformer on the front line: exploring the shortcomings of AI-composed music through quantitative measures. arXiv preprint [arXiv:2008.01307](https://arxiv.org/abs/2008.01307) (2020)
94. Wu, S.L., Yang, Y.H.: Musemorphose: full-song and fine-grained music style transfer with one transformer VAE. arXiv preprint [arXiv:2105.04090](https://arxiv.org/abs/2105.04090) (2021)
95. Yanchenko, A.K., Mukherjee, S.: Classical music composition using state space models. arXiv preprint [arXiv:1708.03822](https://arxiv.org/abs/1708.03822) (2017)
96. 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](https://arxiv.org/abs/1703.10847) (2017)
97. Yi, L., Qu, X.: Attention-based CNN capturing EEG recording’s average voltage and local change. In: Degen, H., Ntoa, S. (eds.) Artificial Intelligence in HCI. HCI 2022. LNCS, vol. 13336, pp. 448–459. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05643-7_29
98. Yu, B., et al.: MuseFormer: transformer with fine-and coarse-grained attention for music generation. arXiv preprint [arXiv:2210.10349](https://arxiv.org/abs/2210.10349) (2022)
99. Zhang, H., Xie, L., Qi, K.: Implement music generation with GAN: a systematic review. In: 2021 International Conference on Computer Engineering and Application (ICCEA), pp. 352–355. IEEE (2021)
100. Zhang, N.: Learning adversarial transformer for symbolic music generation. *IEEE Transactions on Neural Networks and Learning Systems* (2020)
101. Zhang, S., Zhao, Z., Guan, C.: Multimodal continuous emotion recognition: a technical report for ABAW5. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5763–5768 (2023)
102. Zhang, Y., et al.: Biotic homogenization increases with human intervention: implications for mangrove wetland restoration. *Ecography* **2022**(4), 1–12 (2022)
103. Zhang, Z., et al.: Implementation and performance evaluation of in-vehicle high-way back-of-queue alerting system using the driving simulator. In: 2021 IEEE

- International Intelligent Transportation Systems Conference (ITSC), pp. 1753–1759. IEEE (2021)
104. Zhang, Z., Tian, R., Sherony, R., Domeyer, J., Ding, Z.: Attention-based interrelation modeling for explainable automated driving. *IEEE Trans. Intell. Veh.* **18**, 1564–1573 (2022)
 105. Zhao, K., Li, S., Cai, J., Wang, H., Wang, J.: An emotional symbolic music generation system based on LSTM networks. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 2039–2043. IEEE (2019)
 106. Zhao, Z., et al.: A review of intelligent music generation systems (2022)
 107. Zhao, Z., Chopra, K., Zeng, Z., Li, X.: Sea-net: squeeze-and-excitation attention net for diabetic retinopathy grading. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 2496–2500. IEEE (2020)
 108. Zhao, Z., et al.: BIRA-net: bilinear attention net for diabetic retinopathy grading. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 1385–1389. IEEE (2019)
 109. Zong, N., et al.: Beta: a comprehensive benchmark for computational drug-target prediction. *Brief. Bioinform.* **23**(4), bbac199 (2022)