Build Your Own Band (BYOB): Generating instrumental accompaniments for a melody

Bharathi BHAGAVATHSINGH*, Aarthi Vilapakkam SATHISH, Akilan KALAISELVAN, Christina Eunice JOHN

Sri Sivasubramaniya Nadar College of Engineering
Chennai, Tamil Nadu, India

***Corresponding author:** Bharathi BHAGAVATHSINGH
bharathib@ssn.edu.in

Abstract: Artificial Intelligence (AI) has led to advancements in multiple fields of research, and music has always been a field of high interest. Music is an important part of life and various studies have shown the link between better living and listening to music. From completing melodies to composing music from scratch, there are many applications of AI in this domain. This paper aims to analyse one such application of using AI for music generation, specifically for instrumental accompaniment. Instrumental accompaniment is essentially the instrumental music that is composed to support or complement a melody. Creating instrumental accompaniment for music generally requires extensive musical knowledge or forming a band together with skilled instrumentalists. Build Your Own Band (BYOB) attempts to simplify this process with the help of AI. In this research work, three transformer models are employed for training various accompanying instruments. Here, the proposed transformer model accepts a melody line as input and produces an accompanying track with instruments like bass instruments, the guitar and string instruments. One of the main challenges was to make sure that these instruments produce a cohesive sound.

Keywords: Artificial Intelligence, Seq2Seq, Transformer, Music Generation Typesetting and Structure.

1. Introduction

Artificial Intelligence is being used for various applications which involve problem solving, reasoning and general intelligence with greater accuracy and precision than humans. Nowadays, research on the use of AI in applications for simulating human creativity is gaining momentum. One such area is the use of AI in the field of music.

Music composition takes a lot of creativity, energy, skill, knowledge and most importantly streamlined ideas. For a novice music composer this can seem like a daunting task. Relying on another person to help you express your artistic vision can sometimes lead to mismatched tastes. Build Your Own Band (BYOB) aims to use AI to help in the process of musical accompaniment composition and make music more accessible to the layman. The global music publishing business is worth approximately $11.7 billion today. Since BYOB offers multi-instrumental accompaniment, it can help by cutting down on music production costs for composers which can lead to higher profits.

So, this paper proposes to build a system that will take a melody line, which is a piano track as input and generate accompanying instrumental tracks like bass, guitar and strings for this input melody. This paper also aims to overcome one of the main challenges, making the instruments sound cohesive when played together, through the proposed system.

The remainder of this paper is as follows. Section 2 presents an overview of recent studies on generating music with AI. Section 3 describes the proposed methodology in detail. Experimental setup of the proposed work is elaborated in Section 4. Results and performance of the proposed system is explained in Section 5. Section 6 concludes the paper.

2. Related work

There are various ways to generate music using AI. The model proposed by (Hewahi et al., 2019) takes Musical Instrument Digital Interface (MIDI) files and converts them into song files. These song files are encoded and fed as training data to a Long Short-Term Memory (LSTM) network. The model takes an arbitrary note as input and starts amending it until it produces a good

https://doi.org/10.33436/v33i4y202308

piece of music. The data used is from Bach's 'Well-Tempered Clavier Book II' due to its systematic style. The model showed a basic understanding of rhythm and harmony; however, the final music pieces did not have a large structure and some of them were incomplete.

The generation of music is not limited to just composing, there are applications of AI for accompanying melodies provided as user input. The model proposed by (Gillick et al., 2019) transforms abstract musical ideas like scores and rhythms that can be extracted from instruments into expressive drum performances. It uses Sequence-to-sequence (Seq2Seq) and Recurrent Variational Information Bottleneck models on Groove MIDI dataset (Gillick et al., 2019) that contains 13.6 hours, 1,150 MIDI files, and over 22,000 measures of tempo-aligned expressive drumming. The drum patterns are processed over T = 32 timesteps, encoding a drum score to a 256 dimensional vector z with a bidirectional LSTM of 512 layers and decoding it into a performance with a 2-layer LSTM 256 dimensions. The model is able to turn deterministically compressed representations back into complete drum performances. Another single instrument generation model was proposed by (Haki & Jorda, 2019). The model generates a bass line for a given input drum loop. This model used a LSTM-based Word Recurrent Neural Network (RNN) Seq2Seq architecture to learn the encoding between the bass line and the drums vector. The encoder and decoder layers consisted of 128 LSTM units and the decoder was connected to a dense softmax layer to generate the output. The model was trained using 100 loops each with a representation of the drum pattern and baseline transcription for 2 styles. The generated bass lines were rhythmically complex and interlocked with the input drum loop. Melody and chords are important parts of music composition. A model for generating chord sequences from a symbolic melody was proposed by (Lim et al., 2017). The model used a Bidirectional Long Short-Term Memory (BLSTM) architecture with a feature vector of 12 semitones extracted from the notes in each bar of the monophonic melody and was trained on a set of 1802 songs, which consists of 72,418 bars extracted from a lead sheet database with a single chord per bar. All of these models are only capable of generating one instrumental track conditioned on another instrument or melody line. This is a drawback when it comes to the task of generating cohesive multi-track instrumental accompaniments.

LSTMs are good at handling long range data, and to further enhance it, a concept called attention is used. It was first proposed as a concept by (Bahdanau et al., 2014). Attention tries to mimic human brains and direct focus on certain data while ignoring parts of it. As it can be seen in the paper by (Shih et al., 2019), MIDI files are temporal data and the attention module helps in predicting the next note in a series. Transformer models work best in retaining long-term dependencies (Vaswani et al., 2017). They use encoders and decoders with six layers and a scaled dot product and the multi-head attention approach. This approach opens up a better avenue for exploring this project.

Going a step further, a few models can generate multiple accompanying tracks. This is the concept this research work aims to emulate. (Zhu et al., 2018) proposed a melody and arrangement generation system for pop music. They proposed a novel Chord based Rhythm and Melody Cross-Generation Model (CRMCG) to generate a melody using chord progression and rhythms in pop music and a Multi-Instrument Co-Arrangement Model (MICA) for multi-track music. The model is trained over 5,000 pruned MIDI files which are transposed either to C Major or A Minor using an encoder-decoder RNN. The results were documented for rubrics like rhythm, melody, integrity and singability. The drawback of this model is that it generated accompaniment based on a chord progression. On the other hand, the concept proposed by (Dong et al., 2017) uses three models for symbolic multi-track music generation within the framework of generative adversarial networks (GAN), namely the Jamming Model, Composer Model and Hybrid Model. The Lakh Pianoroll dataset is used for training purposes. The bass and drum tracks of the final piece of music tend to be monotonic while the other three tracks tend to harmonize well. There were a few issues with this particular model for a smoother accompaniment a model was proposed by (Ren et al., 2020) called Pop Music Accompaniment Generation (PopMAG). It uses its own novel Multi MIDI approach where multiple MIDI tracks are converted into a sequence of tokens. These tokens are fed to a Seq2Seq model using the LSTM model. They only train on 4/4 time signature songs and use the Viterbi algorithm and Magenta to recognize chords. On a subjective evaluation the obtained

outputs came close to the ground truth making it a good model. From this survey, it can be seen that transformer models yielded the best results so, for the proposed system, the same model was used.

3. The proposed model

The system follows a multi-step process to generate the accompanying instrumental tracks. The system initially accepts the piano track containing the melody as the input. This is used to generate the guitar instrumental track. Subsequently, the generated guitar track along with the input piano track are taken as the inputs to generate the bass track. The generated bass track is used along with the generated guitar track and the input piano track to generate the strings track. All the instrumental tracks are then combined to produce the final output as a MIDI file.

The model used in this system is a Transformer-based Encoder-Decoder model that is used to perform Sequence-to-Sequence (Seq2Seq) learning. Seq2Seq is defined as a process that takes a sentence as an input and produces another sentence as an output. Transformers enable an optimal performance as they process sentences as a whole and learn relationships between words using multi-head attention mechanisms and positional embeddings.

Each model consists of a number of layers, each of which is divided in three sub-layers: self-attention, encoder-decoder based attention (or cross-attention), and a feed-forward layer.

1) **Model 1:** Trained to produce the Guitar track conditioned on the Piano track.
2) **Model 2**: Trained to produce the Bass track conditioned on the Piano and Guitar tracks taken together.
3) **Model 3:** Trained to produce the Strings track conditioned on the Piano, Guitar and Bass tracks together.

As it is illustrated in Figure 1, the system consists of 3 such Transformer Seq2Seq models and for each model the number of tracks given as input is increased by one and the next instrumental track is produced as an output:

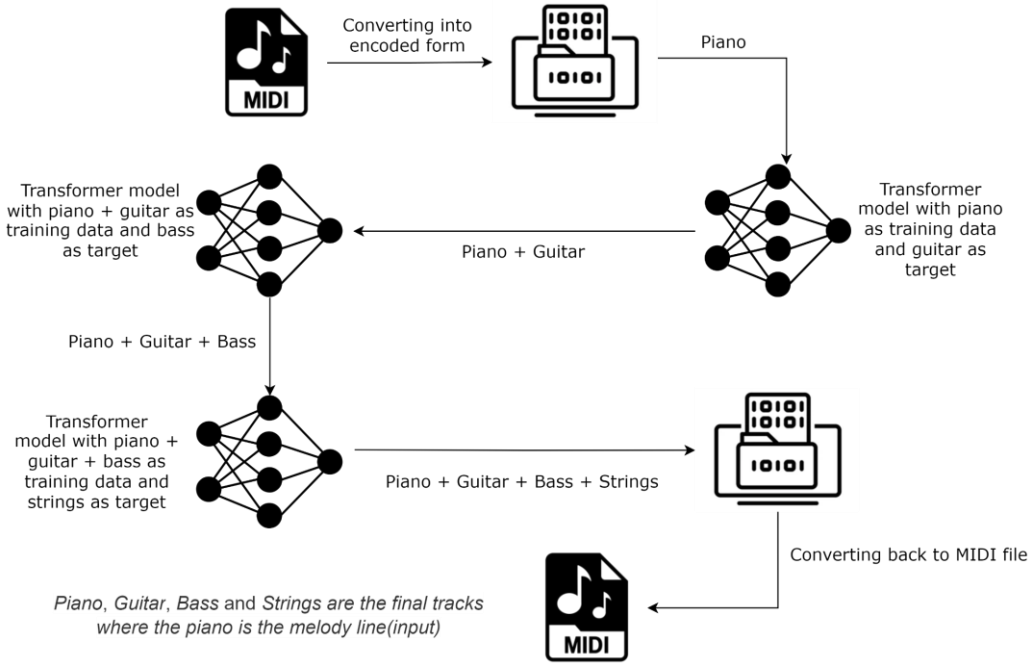

Figure 1. The proposed system architecture

The cross-attention layer of the transformer can be used to incorporate multiple encoders in a single model. This is useful in Seq2Seq tasks where the target sequence is conditioned on more

than one source sequence. This mechanism is used in the proposed system for models 2 and 3 to produce an instrumental track conditioned on more than one instrumental track.

The combination strategies, as they are proposed in (Libovický et al., 2018) that can be used to combine encoders include:

1) Serial combination;
2) Parallel combination;
3) Flat combination;
4) Hierarchical.

This system uses the Serial strategy, as it is shown in Figure 2, for the Bass and Strings Transformer as it was shown to produce the best results as they are presented in (Libovický et al., 2018).

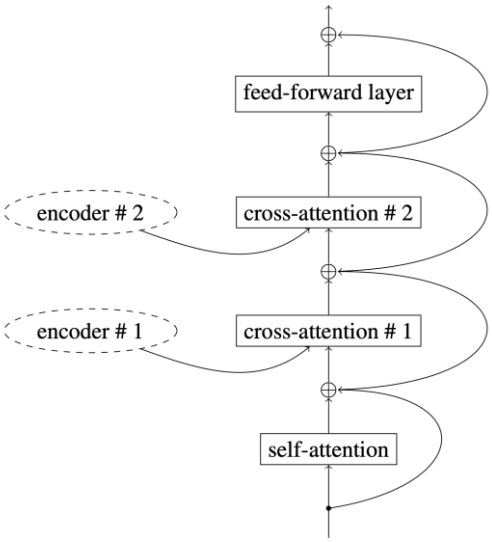

Figure 2. Serial strategy

3.1. Data encoding

A pianoroll was used for data representation in the proposed system. The pianoroll is a multidimensional datatype with dimensions [5, t, 128]. To transform the data into a single-dimensional tensor, embedding is done, as it was proposed in (Shaw, 2019). The pianoroll of each instrumental track is separated and converted back into a MIDI file. These files go through a chord encoding which encodes the data into a sparse numpy matrix of dimensions [Timesteps x Tracks x 128]. This matrix represents which notes are played at each timestep. This chord encoding is then converted into a dense matrix representation to only keep track of when and how long a note is played. The dense matrix is ultimately converted into the one-dimensional tensor using a vocabulary indexed tensor encoding, which is shown in Figure 3. 'xxbos' and 'xxpad' are inserted at the beginning of each sequence as special tokens.

Figure 3. The embedded tensor

4. The experimental setup

4.1. Data pre-processing

For each entry in the dataset, the individual instrumental tracks are encoded using the method mentioned in the previous section.

4.2. Model training

Neural Monkey (Helcl & Libovický, 2017) is an open-source toolkit for training neural models for sequence-to-sequence tasks. This toolkit has been successfully used for machine translation, multi-modal machine translation, and automatic post-editing tasks. It can also be used for many other tasks that involve sequence to sequence learning like image captioning, part-of-speech tagging, sequence classification etc. Neural Monkey allows fast prototyping and easy extension, which makes it very simple to implement and modify recently published techniques. It also provides implementations for the attention combination strategies for the Transformer Decoder used by the models in the proposed system, which is why it was chosen to be used as the toolkit for building the models for the proposed system.

Each Transformer Encoder has 4 self-attention heads for 5 sublayers with a number of 25 feed-forward sublayers. Each Transformer Decoder has 4 self-attention heads and 2 attention heads over each encoder for 5 sublayers with a number of 25 feed-forward sublayers.

All the models were trained with a learning rate of 0.1 using a Lazy Adam Optimizer. The evaluation metric used for all the models is the Mean Squared Error (MSE) Evaluator. This evaluator was chosen due to the similarity of the notes that are close to each other, for example, n65 is close to n66 which should result in a lower error rate. This dataset being in an integer format it is also computationally more efficient in comparison with the usage of strings.

There are three transformer Models, the Guitar Model, the Bass Model and the Strings Model. The Guitar Model takes the Piano input and produces a guitar track. The Bass Model trains over the Piano and Guitar to produce the Bass track. In a similar fashion the Strings Model takes the Bass, Guitar and Piano input to generate Strings. This is where the aforementioned attention combination concept comes into effect. The models described above are summarised in Table 1.

Table 1. Model parameters

Description	**Guitar Model**	**Bass Model**	**Strings Model**
Input	Piano Track	Piano and Guitar Tracks	Piano, Guitar and Bass Tracks
Output	Guitar Track	Bass Track	Strings Track
Trainable parameters	156	248	340
Total parameters	297338	479671	662004

4.3. Evaluation metrics

Since music is a subjective art, a quantitative analysis of the generated music is difficult to carry out. Hence, to analyse the composed accompaniment, four parameters were suggested. Experts in the field of music scored the generated accompaniment on a scale of one to five. These parameters are:

1. **Tunefulness:** How in tune with the music is the accompaniment. In other words, how the instruments follow the flow of the given melody.

2. **Harmonic Accuracy:** How harmonic the instruments sound together. Harmonies are generally formed when instruments play different notes but when heard together they sound pleasant and almost like one sound. That is, the notes, though varied, do not sound out of place.

3. **Timing:** How on the beat the instruments are to the main melody. This is a measure of

how the generated music follows the timing of the given melody and how the notes are not played out of sync.

4. **Cohesiveness:** How the melody and the accompaniment sound together as a whole. This measure ultimately relies on subjective listening and how good the instruments sound together.

5. The obtained results

For the subjective evaluation, three songs generated from different melody lines were chosen. The MIDI files of the generated accompaniments are shown below in Figure 4, Figure 5 and Figure 6. In these images, the blue line is the Piano Track, the red line is the Guitar Track, the green line is the Strings Track and the yellow line is the Bass Track.

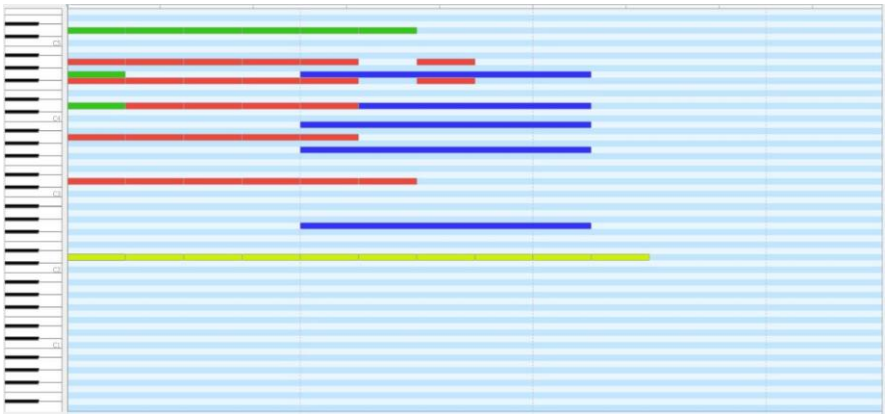

Figure 4. MIDI representation of the first song

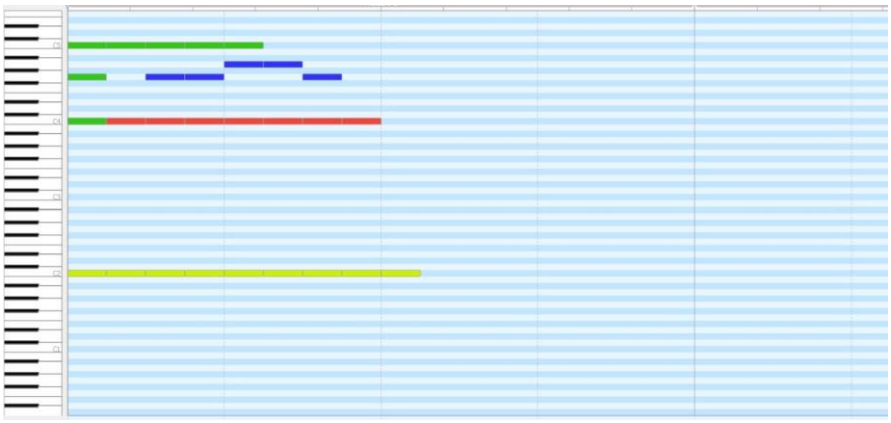

Figure 5. MIDI representation of the second song

Figure 6. MIDI representation of the third song

The generated songs were evaluated by 7 amateur musicians for the aforementioned metrics. The aggregated scores for each song are illustrated in Figure 7, Figure 8 and Figure 9.

5.1. Song 1

In the first song's evaluation the most varied scores can be seen with Harmonic accuracy being the highest-rated feature. The melody for song 1 was taken from the testing subset. On average the obtained scores lie between 3 and 4 out of a maximum score of 5. Figure 4 shows the MIDI representation for the generated music and Figure 7 shows the average score for each of the evaluated metrics namely Tunefulness, Harmonic Accuracy, Timing and Cohesiveness. From Figure 7, it can be observed that Harmonic Accuracy is the highest-rated feature, while Timing is the lowest-rated one.

5.2. Song 2

In the second song the scores, as it can be seen in Figure 8 dipped when there were no strings but increased once the strings track was added to the song. This shows that as a whole all the instruments play a part in making the song more cohesive and well-rounded. This song is a small snippet from the popular nursery rhyme "twinkle, twinkle, little star". The average rating is seen in Figure 8 the strings track. Here it can be observed that Harmonic Accuracy and Cohesiveness are the highest-rated parameters, while once again Timing is the lowest-rated one.

5.3. Song 3

This song's evaluation showed the least variation in score. The highest-rated feature of this song is its cohesiveness. The input for this song is a newly composed snippet that was not taken from any existing song. Hence, the scoring consistency for this song reiterates how well this model generates musical accompaniment. The MIDI representation for Song 3 is depicted in Figure 6 and the average scores are illustrated in Figure 9. From Figure 9, it can be seen that Cohesiveness is the highest-rated feature, while Tunefulness is the lowest-rated one.

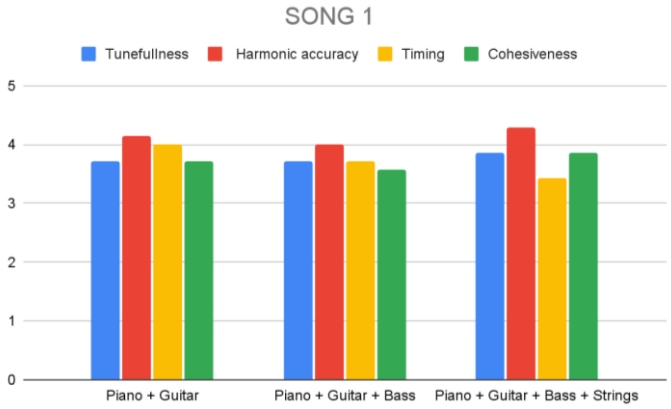

Figure 7. The evaluation of the first song

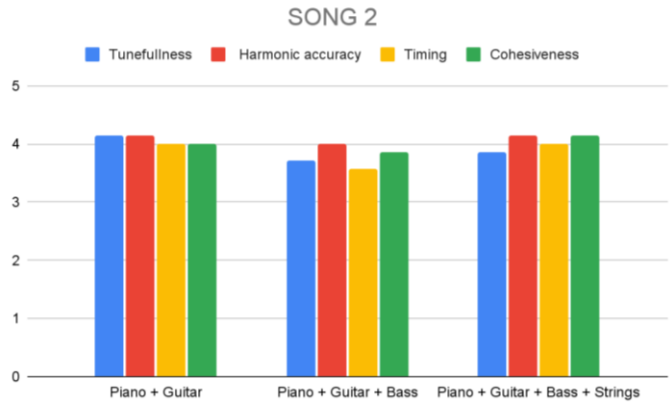

Figure 8. The evaluation of the second song

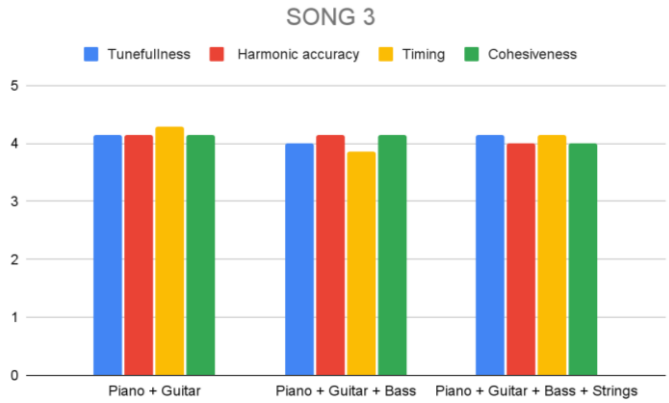

Figure 9. The evaluation of the third song

Overall, when analysing the obtained scores, a trend can be observed. The highest-rated features are **Harmonic Accuracy** and **Cohesiveness** while the lowest-rated feature is **Timing**. Scoring well in those two features further drives home the benefits of using transformers for music accompaniment generation. The timing of the accompanying music is an area that the proposed model needs to work on and it can be done in the context of future iterations.

6. Conclusions

This work presented a model for generating accompaniment for a given input melody line. The idea of using various transformers to train the guitar based on the piano, and training the bass based on the piano and guitar and so on helped in making the harmonies work well, which further enhanced the cohesiveness of the generated music. The generated music scored fairly well in terms of subjective evaluation. The feature of the generated music that was the most appreciated was its cohesiveness, which is one of the most important features as it refers to the overall sound of the evaluated song. Through this analysis, it can be confidently concluded that transformers aid in generating the best music as they help generate sequences. However, the training time and dataset should be increased to obtain longer music snippets and optimal results. Once it would be further enhanced, this model could be practically used to help generate musical accompaniment.

As further enhancements certain areas were identified where this model could be improved. One possible enhancement could be attained by using Beat Wise Positional Encoding. Here the positional encoding should be done based on the note separator token instead of the index of the token in the tensor. This would help the proposed model to understand the musical timing much better and faster. Another enhancement would be to use a custom evaluator function. A custom evaluator score could be created instead of using the BLEU score or MSE as evaluators. These evaluators work well for languages but may not be optimal for music, which gives rise to the need for a custom evaluator. Beyond this, this model could also be turned into an application with a dedicated user interface which could give users a better experience and aid them in customizing their music as well.

REFERENCES

Bahdanau, D., Cho, K. & Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. To be published in *3rd International Conference on Learning Representations, ICLR 2015*. [Preprint] https://arxiv.org/abs/1409.0473 [Accessed 23rd May 2022].

Dong, H.-W., Hsiao, W.-Y., Yang, L.-C. & Yang, Y. (2017) MuseGAN: Demonstration of a Convolutional GAN Based Model for Generating Multi-track Piano-rolls. *Proceedings of Late-Breaking/Demo Session of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), 23-27 October, 2017, Suzhou, China*. pp. 34-41.

Gillick, J., Roberts, A., Engel, J., Eck, D. & Bamman, D. (2019) Learning to Groove with Inverse Sequence Transformations. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June, 2019, Long Beach, CA, USA*. pp. 2269-2279.

Haki, B. & Jorda, S. (2019) A Bassline Generation System Based on Sequence-to-Sequence Learning. In: *Proceedings of 19th International Conference on New Interfaces for Musical Expression, NIME 2019, 3-6 June, 2019, Porto Alegre, Brazil*. pp. 204-209.

Helcl, J. & Libovický, J. (2017) Neural Monkey: An Open-source Tool for Sequence Learning. *The Prague Bulletin of Mathematical Linguistics*. 107(1), 5-17. doi: 10.1515/pralin-2017-0001.

Hewahi, N., AlSaigal, S. & AlJanahi, S. (2019) Generation of music pieces using machine learning: long short-term memory neural networks approach. *Arab Journal of Basic and Applied Sciences*. 26(1), 397-413. doi: 10.1080/25765299.2019.1649972.

Libovický, J., Helcl, J. & Mareček, D. (2018) Input Combination Strategies for Multi-Source Transformer Decoder. In: *Proceedings. of the Third Conference on Machine Translation, 31 October - 1 November, 2018, Brussels, Belgium*. Volume 1: Research Papers, pp. 253-260.

Lim, H., Rhyu, S. & Lee, K. (2017) Chord Generation from Symbolic Melody Using BLSTM Networks. In: *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), 23-27 October, 2017, Suzhou, China*. pp. 621-627.

Ren, Y., He, J., Tan, X., Qin, T., Zhao, Z. & Liu, T.-Y. (2020) PopMAG ag: Pop music accompaniment generation. In: *Proceedings of the 28th ACM International Conference on Multimedia, MM '20, 12-16 October, 2020, Seattle, WA, USA*. New York, NY, Association for Computing Machinery. pp. 1198-1206.

Shaw, A. (2019) *Creating a Pop Music Generator with the Transformer.Medium,* https://medium.com/towards-data-science/creating-a-pop-music-generator-with-the-transformer-5867511b382a [Accessed: 20th May 2022].

Shih, S.-Y., Sun, F.-K. & Lee, H. (2019) Temporal pattern attention for multivariate time series forecasting. *Machine Learning*. 108(8-9), 1421-1441. doi: 10.1007/s10994-019-05815-0.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017) Attention is All you Need. In: von Luxburg, U., Guyon, I., Bengio, S., Wallach, H. and Fergus, R. (eds.) *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, 4-9 December, 2017, Long Beach, CA, USA*. Red Hook, NY, Curran Associates Inc. pp. 6000-6010.

Zhu, H., Liu, Q., Yuan, N. J., Qin, C., Li, J., Zhang, K., Zhou, G., Wei, F., Xu, Y. & Chen, E. (2018) Xiaolce band: A melody and arrangement generation framework for pop music. In: *Proceedings of 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, 19-23 August, 2018, London, UK*. New York, NY, Association for Computing Machinery. pp. 2837-2846.

Bharathi BHAGAVATHSINGH working as an Associate Professor at the Department of Computer Science and Engineering of Sri Sivasubramaniya Nadar College of Engineering in Chennai, Tamil Nadu, India. She received her Ph.D. in Information and Communication Engineering from Anna University in Chennai, India. Her areas of research interest include Speech processing, Natural language processing, and Music processing. She published around 95 papers in reputable journals and conference proceedings.

Aarthi Vilapakkam SATHISH obtained her Bachelor's Degree in Computer Science and Engineering in 2022, at Sri Sivasubramaniya Nadar College of Engineering in Chennai, Tamil Nadu, India. She is currently working as a Software Development Engineer at Amazon. The product that she is working on is the AppStore application for Amazon’s devices. This has further increased her interest in the field of productising AI/ML systems.

Akilan KALAISELVAN obtained his Bachelor's Degree in Computer Science and Engineering in 2022, at Sri Sivasubramaniya Nadar College of Engineering in Chennai, Tamil Nadu, India. He is working as a Software Engineer at Optum, which is the tech wing of UnitedHealth Group. Since he has recently started working on real-world problems especially in the healthcare industry, he is intrigued by the application of ML in this specific field.

Christina Eunice JOHN obtained her Bachelor's Degree in Computer Science and Engineering in 2022, at Sri Sivasubramaniya Nadar College of Engineering in Chennai, Tamil Nadu, India. She is working as a Software Engineer at Optum, which is the tech wing of UnitedHealth Group. She is interested in the applications of AI for making healthcare more accessible.