Quality Comparison of Classical and Jazz Music Generation using LSTM and Transformer Models

Aika Oki

Received July 23, 2025
Accepted September 27, 2025
Electronic access November 30, 2025

Recently, AI music generation has evolved significantly, alongside other generative AI technologies such as text, image, and time-series prediction. Although numerous studies have analyzed AI music generation in general, a relatively limited amount of research has been conducted comparing the distinct methods for genre-specific music generation. There are two main music generation approaches: long short-term memory (LSTM) and Transformer-based networks. In this paper, we compared the performance of these two artificial neural networks in generating classical and jazz music using 60 MIDI data files (30 files each). We then evaluated the genre accuracy of AI-generated music with LSTM and Transformer classifiers and a human subjective review of generated music for five features (genre accuracy, harmonic accuracy, melodic accuracy, cohesion in form and structure, variation) to determine music generation quality. This study showed that both the LSTM and Transformer had a comparable performance in classical music generation; however, the Transformer was more accurate in jazz music generation as assessed by the Transformer classifier. This tendency, also confirmed by the subjective review, may reflect the higher number of features (note vocabulary) for jazz music, which favored the Transformer's self-attention mechanism and simultaneous processing of features. Even with a small pitch-only dataset, the study clearly showed that LSTM and Transformer generation models can create genre-specific music, and that LSTMs and Transformers may have distinct strengths in different musical genres. Since less is known about their capabilities for specific types of music, this study could provide fundamental insights for realistic AI music generation.

Introduction

Rapid advances in machine and deep learning have facilitated music generation[1,2], along with other generative AI technologies, such as text and image prediction or time-series prediction[3]. Within the past five years, music AI generation has evolved significantly[4–10] as more complex models have been advanced, allowing for greater pattern recognition[2] and effective training[1]. Several different approaches have been attempted for music generation; one of the main approaches is using Recurrent Neural Networks (RNN), especially with long short-term memory (LSTM) networks, which incorporate attention mechanisms into RNNs to process sequential data[1,2].

The data analysis for musical piece has also been facilitated by the Musical Instrumental Digital Interface (MIDI) data format which is a standardized protocol that allows to store and exchange musical performance data[4]. MIDI files usually contain attributes such as note pitches, note velocities, instrumental selections and instrumental changes in a small-sized file enabling the replication of music in a device-independent manner. MIDI files are widely used for music composition, performance, electronic instrument control, and data analysis.

During the development of machine learning for music generation tasks, which are essentially time series prediction tasks, RNNs and their variants were found to be especially effective[1]. Deep learning through LSTMs utilizes memory cells with input and output gates, which allow feedback even with long delays[3]. This, along with the use of linear units or activation functions, mitigates error backflow issues such as vanishing gradients[2]. Previous studies have utilized architectures such as a single LSTM layer, a tied parallel network model of RNN and LSTM, or an LSTM-based RNN[4].

Another primary approach is using Transformer models. Transformer models were initially introduced in 2017 as sequence-to-sequence translation models[11]. Its usage of self-attention mechanisms allows it to better maintain long-term dependencies in data and deal with variable-length inputs[12]. Specifically, it can weigh the importance of different portions of an input sequence and has parallel processing, thereby preventing gradient vanishing or explosion[11]. As shown by the recent successful application of Transformer models in many fields such as Large Language Models (LLMs), drug discovery, or even image generation[11], Transformer models have been extremely versatile and shown to outperform LSTM in areas using sequence-to-sequence (seq2seq) models such as music generation[1] or speech processing[13]. Transformer mod-

els, as well as hybrid LSTM and Transformer models[3] continue to have relevance in time-series prediction, leading to this research on validating these models in specific cases using music classification models.[14]

Some challenges with this method include the lack of inherent structural bias, which can lead to overfitting for small sequences, requiring a large dataset for training. Typically, the encoder-decoder architecture is used for seq2seq mapping in translation. However, the encoder can also be utilized for classification tasks such as music or text classification, and the decoder can be utilized for music generation or language modeling tasks.[12]

Overall, since a self-attention mechanism is used in a Transformer model, it can recognize each piece of information simultaneously and therefore deal with complex data, while an LSTM recognizes the data step-by-step and is better suited for finding patterns in the data. Therefore, both models have their pros and cons; however, there has been limited research on which model should be used in music generation.

In this study, the goal was to determine the efficacy of LSTM and Transformer music generators for classical and jazz music, validating it using corresponding LSTM and Transformer classification models to directly compare the performance of LSTMs and Transformers. This is especially important as much of the current literature has been focused on music generation models trained in just a single technology rather than a comparison of different technologies or methods[1]. For training, we took a sample of files from classical and jazz datasets, using 150 files for classification, 60 files total for generation, and 40 files total for evaluation[13,15].

Thus, we focused on comparing both classical and jazz music generation with LSTMs and Transformers through 3 different evaluation methods (i.e. LSTM and Transformer classifier and subjective review with 5 music features[16]), to produce a more standardized accuracy evaluation. This study elucidates the differences between LSTM and Transformer in music generation with respect to generation of genre-specific features and helps the future development of music generation systems.

Method

We conducted an experimental study of LSTM and Transformer music generation for classical and jazz music; the detailed code is included in the references[17]. The Classical MIDI dataset[15] by Soumik Rakshit on Kaggle, containing 293 MIDI files of 19 famous classical composers as well as the Jazz ML MIDI Dataset[13] by Sai Kayala with 942 MIDI files for jazz were used for training. From this data set of 293 classical and 942 jazz MIDI files, we randomly sampled 150 files for the classifier training set, 30 independent files of each genre for the generator training set, and 20 independent files of each genre for the classification validation set. In order to reduce the computational burden, a simplified approach with seq2seq was taken and only the 'pitch' feature from each MIDI file was used for classification and generation training, given that pitch is the primary feature used in previous research on seq2seq music generation and classification[18]. For the seq2seq approach, during the processing of notes, simultaneous notes in a single instrument part were treated as a chord object. Simultaneous notes across multiple instrument parts were considered separately, with each instrument part connected consecutively to each other. If no instrument parts were present, it was simply flattened chronologically. In addition, the note vocabulary sizes were determined by the number of unique notes and chords in the training sets. Increasing or decreasing the layer size did not lead to any improvements, and in fact may decrease final training accuracy. Google Colab's T4 GPU with 16 GB VRAM was used for both training and generation.

The Classical Piano Composer LSTM generation model[19] by Skúli was used for the LSTM generation, and Transformer layers from a Keras tutorial[20] were used for the Transformer classification and generation architecture. The Transformer architecture utilized a single Transformer layer which seemed ineffective to reduce the loss, so an additional Transformer layer was added.

Classification Architecture

The LSTM Classification model follows the architecture in Figure 16. The first LSTM layer with 32 hidden units initially captures note features. A dropout layer with $dropout = 0.3$ prevents overfitting. The second LSTM layer reduces the sequence to a fixed-size representation of shape (None, 52). Additionally, there is a dropout layer to prevent overfitting. After that, there are two Dense layers: one is a fully connected layer with a ReLU activation function, and the other is an output layer used for genre classification.

The Transformer Classification model, an encoder-only model, follows the architecture in Figure 17. First, there is a Dense embedding layer with 64 units that transforms input sequences into a higher-dimensional space. Then, the Multi-Head Self-Attention layer, with 64 hidden units and 4 heads, helps capture long-term dependencies in music data. The Layer Normalization layer helps stabilize training by adding the original input to the attention output as a residual connection. In the Feedforward Neural Network, there is a Dense layer with 128 units and a ReLU activation function, a Dropout layer with $dropout = 0.3$ to prevent overfitting, and another Dense layer with 64 units. The Global Average Pooling reduces the output sequence to a fixed-size vector, and the

output layer is a Dense layer with 2 units for prediction, using the softmax activation function.

Generation Architecture

The LSTM Generation model follows this architecture as noted in Figures 18 and 20. The first three LSTM layers have 512 hidden units initially capturing note features, with recurrent $dropout = 0.3$ for the first two. There is a Batch Normalization layer, then a Dropout layer with $dropout = 0.3$ to minimize overfitting. An additional Dense layer with 128 units is followed by a ReLU activation function. Then another Batch Normalization and Dropout layer is used. A Dense layer with 372 or 604 units, corresponding to note vocabulary size, outputs a probability distribution over all the possible notes. The output layer finally uses a softmax activation function.

The Transformer Generation model follows this architecture as noted in Figures 19 and 21. An Input layer with a sequence length of 100 tokens transforms input data. The Token and Position Embedding layer outputs a sequence of token embeddings of size 256. The Transformer encoder layers are multi-head self-attention and feedforward networks that capture long-term dependencies. A Dropout layer with $dropout = 0.3$ is included to reduce overfitting. In the Flatten layer, the output is flattened into a $1D$ vector. Finally, the output layer is an output layer with a softmax activation function over 372 or 604 possible notes.

Classifier/Generator Training

The LSTM classifier was trained on 150 randomly selected classical and 150 randomly selected jazz datasets, for 400 epochs to ensure the training accuracy curve plateaued. The classical and jazz LSTM generation models were trained on 30 randomly selected classical and 30 randomly selected jazz music files for 200 epochs. We utilized 30 files only to ensure a reasonable training time for the LSTM jazz generation, which contains a large note vocabulary.

The Transformer classifier was trained on 150 random classical and 150 random jazz datasets, for 100 epochs to ensure the training accuracy curve plateaued. The classical and jazz Transformer generation models were each trained on 30 random classical and jazz music files for 200 epochs.

Validation

We validated the above classification models using 20 test files each of human-made classical and jazz music (for a total of 40). We also generated 200 files of classical and jazz with LSTM and Transformers generation models and predicted the genre probabilities with the classifier. For each file, we took the average probabilities of all windows. We then conducted a t-test for genre prediction accuracy across 200 generated classical and jazz files for each model, to account for potential unequal variances in the probability distribution of the test files. Afterwards, we included a subjective evaluation of 10 representative classical and jazz-generated files across 5 features (Genre Accuracy, Harmonic Accuracy, Melodic Accuracy, Cohesion in Form and Structure, and Variation)[16]. The subjective analysis was performed by the author, who has more than 10 years of musical performance experience.

Results

Table 1 Data for 6 Neural Networks with Training Times to Accuracy Saturation

Model	Time taken for training (hrs:mins:secs)
LSTM Classification	01:46:46
LSTM Generation: Classical	16:18:26
LSTM Generation: Jazz	15:52:17
Transformer Classification	00:14:36
Transformer Generation: Classical	02:10:08
Transformer Generation: Jazz	03:54:33

Music Classifier Training

For training of the music classifier, Figure 1 shows that the LSTM model improved at around 25 epochs, reaching an accuracy of approximately 90% and at 200 epochs, 96% accuracy (Figure 1). On the other hand, the Transformer model required a few epochs to reach an accuracy of around 80% and remained similar at 100 epochs (Figure 2). This also indicates that the Transformer required around $1/7$ of the time of LSTM generation models for training (Table 1), suggesting that it was far more time-efficient than the LSTM model for training. During training, both models utilized around half of the allotted VRAM or less, with almost no CPU usage. However, the LSTM model achieved a better training accuracy than the Transformer model (Figure 3), and thus, it produced a better classifier overall for Classical and Jazz classification.

When we examined the validation accuracy for the LSTM model, it was a few percentage points lower than the training accuracy, with sharp drops in accuracy toward the end, suggesting slight overfitting. For the Transformer model, the validation accuracy appeared to fluctuate roughly around the training accuracy, albeit irregularly. It is possible that due to the small training size, the transformer classifier had too many parameters and was overfitting and memorizing the training data.

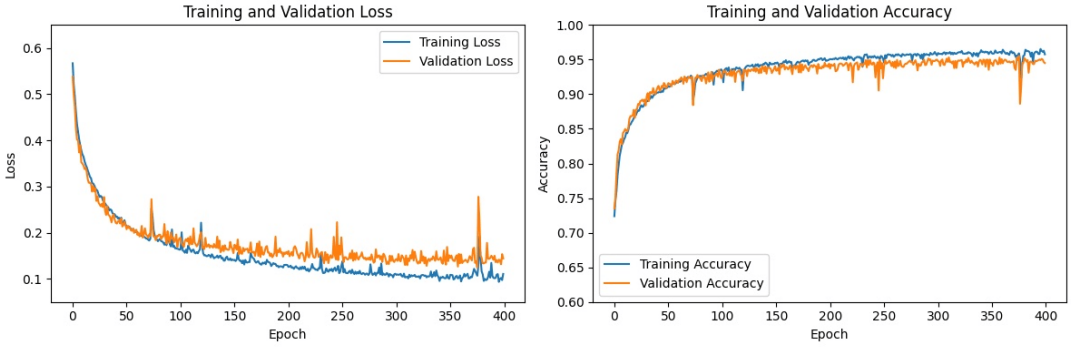

Fig. 1 LSTM Classification Training Graph Comparing Training and Validation Loss and Accuracy vs. Epoch.

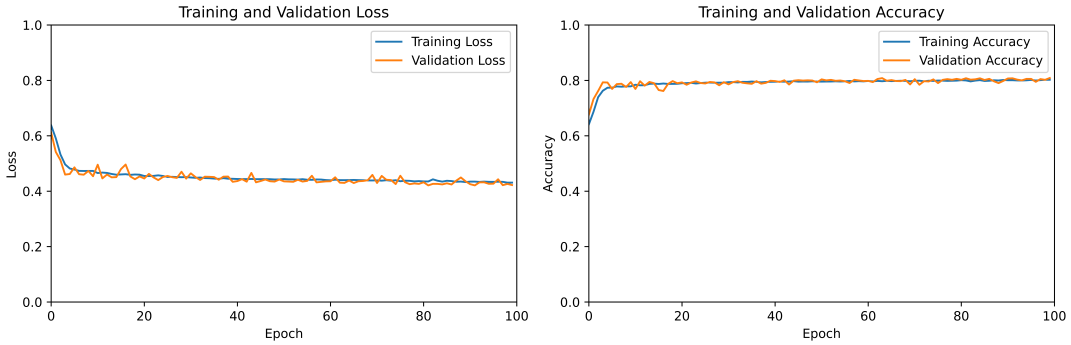

Fig. 2 Transformer Classification Training Graph Comparing Training and Validation Loss and Accuracy vs. Epoch.

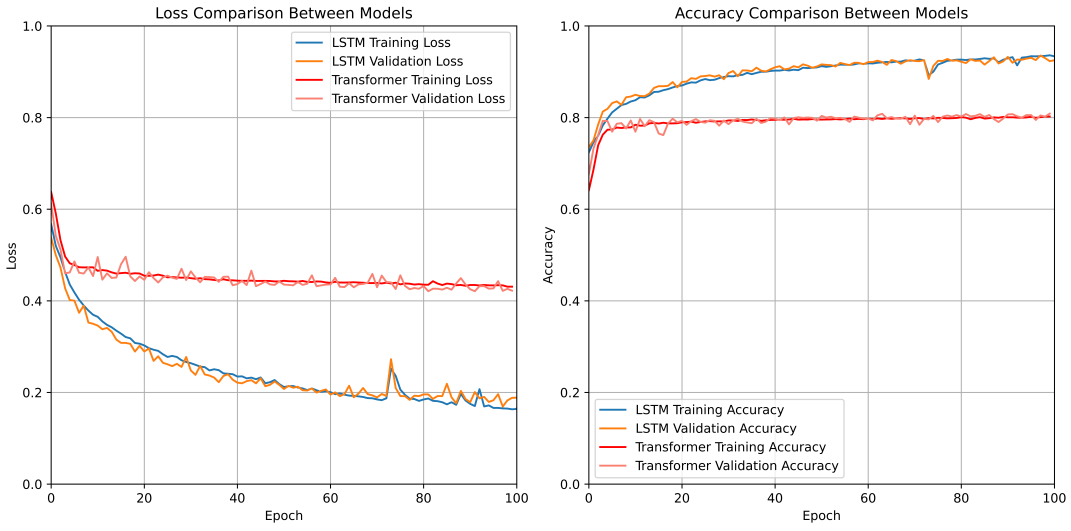

Fig. 3 LSTM and Transformer Classification Training Graph Comparing Training and Validation Loss and Accuracy vs. Epoch.

Music Generator Training

For classical generation, the LSTM model required around 100 epochs to reach above 70% model accuracy and 180 epochs for 90% model accuracy, as shown in Figure 4. In contrast, for jazz generation, it improved quickly and reached above 80% model accuracy earlier on such as 50 epochs and around 90% at 100 epochs, as shown in Figure 5. The final accuracy for classic was 91% and the one for jazz was 94% through Figure 4 and Figure 5. Thus, the final accuracy seems comparable between classical and jazz. Interestingly, training for classical music required slightly more epochs than jazz music, but the total training time for classical or jazz music was similar to each other for the LSTM model (Table 1).

For classical generation, the Transformer model required 25 epochs to reach above 80% model accuracy and 90% model accuracy at 75 epochs (Figure 6). In contrast, for jazz generation, it improved quickly and reached above 70% model accuracy earlier on, such as 5 epochs and around 90% at 20 epochs (Figure 7). Thus, the training efficiency was better for jazz music.

Comparing the classical and jazz generation in Figures 3, 4, 5, and 6 respectively, the transformer showed a dramatic rise in accuracy at 10 epochs (instead of around 75 epochs), whereas the LSTM showed a gradual improvement. However, the final accuracies for both generators were comparable. Interestingly, during the training for music generation using the Transformer, several dips in accuracy were observed for both classical and jazz which were not seen in the training of the LSTM, suggesting slight overfitting, although the overall accuracy remained high.

Generator Validation Using LSTM Classifier

Before validating the two types of music generators (LSTM and Transformer) with our LSTM classifier, we validated the LSTM classifier with 20 independent data sets for classical and jazz music to check how many correct predictions could be made. The cutoff for a correct prediction was defined as more than 50% predicted probability. For classical, the average predicted probability across all windows for the test set was 0.850, with $20/20$ correct predictions. For jazz, the average predicted probability across all windows for the test set was 0.817, with $18/20$ correct predictions.

Next, the genre prediction accuracy scores of the generated music by LSTM and Transformer Generator were compared, for classical and jazz music ($n = 200$ for each genre), as noted in Tables 2. A t-test was conducted, and statistical significance was based on 2-sided $P \leq 0.05$.

For classical, the LSTM mean genre prediction accuracy was $M = 0.62$, $SD = 0.20$, while the Transformer mean accuracy was $M = 0.62$, $SD = 0.24$. The results of the t-test did not indicate any difference, with $p = 0.89$. For jazz, the LSTM mean genre prediction accuracy was $M = 0.55$, $SD = 0.25$, while the Transformer mean accuracy was $M = 0.54$, $SD = 0.23$. The results of the t-test did not indicate any difference, with $p = 0.70$. There was no statistical significance between Transformer and LSTM generation.

Generator Validation Using Transformer Classifier

Similarly to the LSTM classifier, we validated the Transformer classifier. For classical, the average predicted classical probability across all windows for the test set was 0.695, with $18/20$ correct predictions, while for jazz, the average predicted probability across all windows for the test set was 0.686, with $16/20$ correct predictions.

We also compared the genre prediction accuracy scores of the LSTM and Transformer Generator, for classical and jazz music ($n = 200$), as noted in Table 3. For classical, the LSTM mean genre prediction accuracy was $M = 0.55$, $SD = 0.18$, while the Transformer mean accuracy was $M = 0.54$, $SD = 0.23$. The results of the t-test indicated no significant difference between the groups, with $p = 0.45$. For jazz, the LSTM mean genre prediction accuracy was $M = 0.56$, $SD = 0.21$, while the Transformer mean accuracy was $M = 0.63$, $SD = 0.17$. The results of the t-test indicated a significant difference between the groups, with $p = 0.0007$.

The Transformer-based assessment for classical music revealed no statistical difference between LSTM and Transformer Generation, confirming the results obtained by the LSTM-based assessment.

On the other hand, for the jazz generation, there was a statistical difference between Transformer Generation and LSTM Generation, with higher prediction accuracy for Transformer generation. This suggests that the Transformer may be superior to the LSTM for jazz generation.

Subjective Analysis

Finally, we also conducted a subjective analysis of 10 representative generated classical/jazz music files for the LSTM and Transformer models based on 5 features: Genre Accuracy, Harmonic Accuracy, Melodic Accuracy, Cohesion in Form and Structure, and Variation. Although mentioned as a possible evaluation criterion in a previous paper[16] , rhythmic nuance was not analyzed. This was because only pitch was considered as a feature during generation, not the duration of notes or start time, meaning exact rhythms may have been inaccurate. Additionally, the playback tempo was adjusted to best fit the character of the music. The analysis was conducted, and each individual feature was rated manually.

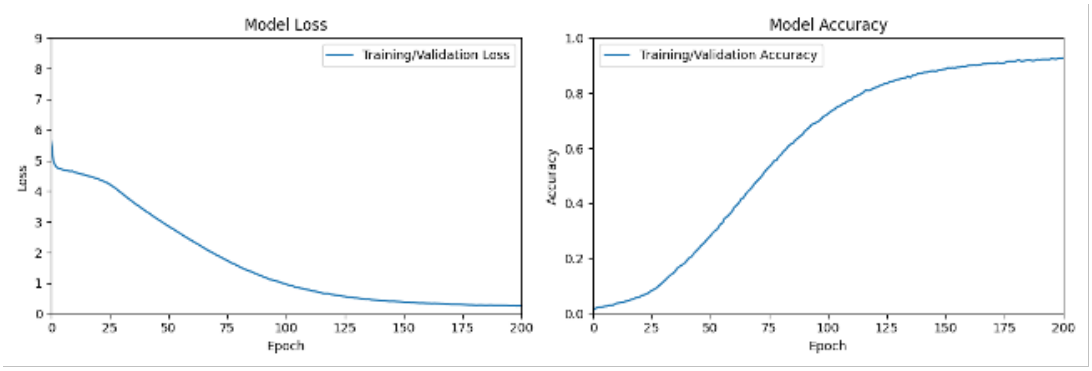

Fig. 4 LSTM Classical Generation Training Graph Comparing Training Loss and Accuracy vs. Epoch

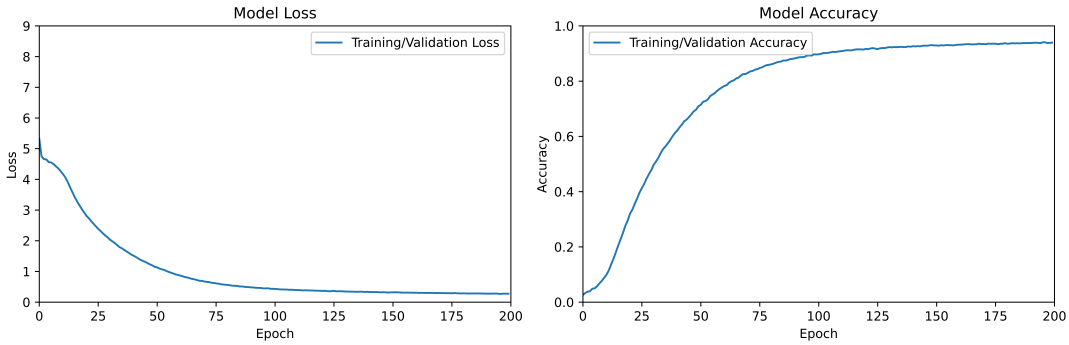

Fig. 5 LSTM Jazz Generation Training Graph Comparing Training Loss and Accuracy vs. Epoch

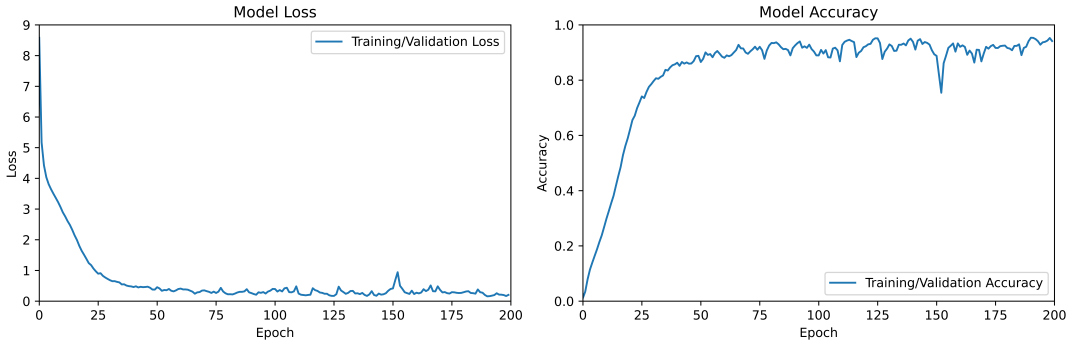

Fig. 6 Transformer Classical Generation Training Graph Comparing Training Loss and Accuracy vs. Epoch

Table 2 Validation with LSTM and Transformer generators using an LSTM classifier

	LSTM Classical	**Transformer Classical**	**LSTM Jazz**	**Transformer Jazz**
Sample	200	200	200	200
Mean	0.62	0.62	0.55	0.54
Standard Deviation	0.20	0.24	0.25	0.23
95% Confidence interval	(0.59, 0.65)	(0.59, 0.65)	(0.52, 0.58)	(0.51, 0.57)
p-value (LSTM VS Transformer)	0.89	0.89	0.70	0.70

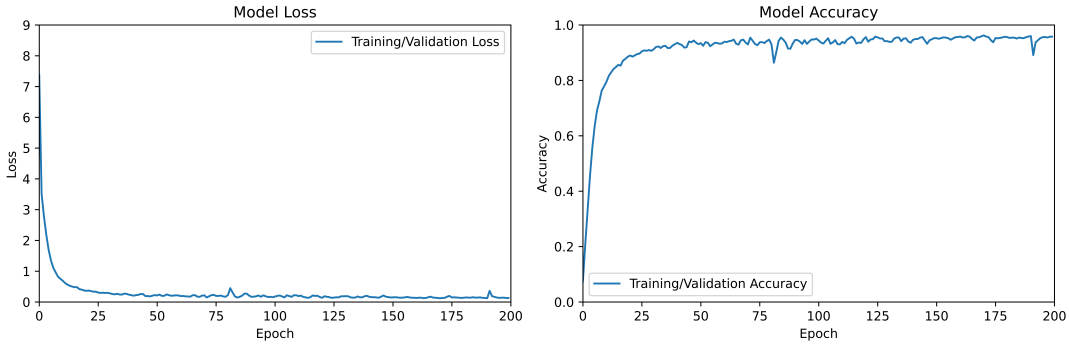

Fig. 7 Transformer Jazz Generation Training Graph Comparing Training Loss and Accuracy vs. Epoch

Table 3 Validation with LSTM and Transformer generators using a Transformer classifier

	LSTM Classical	**Transformer Classical**	**LSTM Jazz**	**Transformer Jazz**
Sample	200	200	200	200
Mean	0.55	0.54	0.56	0.63
Standard Deviation	0.18	0.23	0.21	0.17
95% Confidence Interval	(0.53, 0.57)	(0.51, 0.57)	(0.53, 0.59)	(0.61, 0.65)
p-value (LSTM VS Transformer)	0.45	0.45	0.0007	0.0007

The rating ranges from 1 to 10, with 10 being the best, and a description and rubric of the features is presented below:

Feature Descriptions and Rubrics:

Genre Accuracy: How close the composition is overall to other compositions in the same genre of music (overall rating).

Rubric:

- **10**: Clearly pertaining to selected genre; similar to human-made music of genre;
- **5**: Plausibly fits into genre, but with noticeable inconsistencies or oversimplifications;
- **1**: Features of music do not align at all with genre or music seems randomly composed

Harmonic Accuracy: Accuracy and complexity of harmonies and harmonic progressions.

Rubric:

- **10**: Skillfully utilizes major harmonic progressions or structures, is deliberate and varied in usage with complex harmonies over time
- **5**: Harmonic progressions are recognizable, but may be less varied or used haphazardly over time
- **1**: Harmonic progressions and structures are incoherent or misleading, with no pattern or organization

Melodic Accuracy: Accuracy and complexity of melodic lines and phrasing.

Rubric:

- **10**: Phrases and melodic structures are deliberate and memorable, with development and exploration of the melody over time
- **5**: Melodies are frequently used but may be repetitive or hard to recognize
- **1**: Melodic and phrasing structure is absent beyond a few coincidental notes

Cohesion in Form and Structure: Coherence of composition over a long duration and clear structure.

Rubric:

- **10**: Consistently retains a clearly recognizable structure and a balance of sections throughout the composition and shorter segments, as well as skillfully repeating melodies or harmonies
- **5**: Retains a clear structure and repeated sections in the short term, but becomes imbalanced or unclear over time
- **1**: Lacks categorization or sections, with little consistency or integrity as a singular composition

Variation: Variability of melodies, harmonies, and phrases over time within a composition.

Rubric:

- **10**: Exhibits engaging variation across melodic and harmonic structures as well as unique developments in composition as a whole
- **5**: Is varied in some melodies or harmonies but not consistently; variation may pertain to only some phrases and not others
- **1**: Is constantly repetitive on the same few notes or phrase, with little musical value beyond a few lines

Classical Music

Genre Accuracy:

LSTM (6/10): The generation varied depending on the composition. Sometimes, the combination of different chords seemed to mix contemporary, classical, and romantic periods, making it not seem like one particular composition, especially with its frequent use of minor second chords. Sometimes the harmonic rhythm fluctuated too much. However, when successful, there were clear scale-based melodies and harmonies of the classical style.

Transformer (6/10): The Transformer was similar to the LSTM. It occasionally hit upon some interesting, complex harmonies, but was less adept at consistently using natural-sounding motives. Although it was more cohesive than the LSTM over time, there were sometimes too many unnatural melodic or harmonic fluctuations in the beginning few lines.

Harmonic Accuracy:

LSTM (7/10): For LSTM, chords and harmonies were used extensively, which usually progressed naturally in the short term (e.g., modulations, cadential embellishments such as V7-iii6-I). Basic cadences were often used for classical-era music. However, under the surface, there were some inconsistencies. Sometimes, in a more contemporary fashion, chords changed very quickly, such as every bar or so, including non-diatonic tones and secondary dominants with little context, which could feel slightly confusing. The LSTM usually began with a more classical-period-like style, and sometimes transitioned to a contemporary music style, or vice versa. Sometimes, this transition to unexpected minor seconds or chromaticism felt novel; however, other times it was a little abrupt. This may have been because many chords were arpeggiated (likely due to the sequential instead of simultaneous flattening of notes in different parts), which made the relative harmonic rhythm and sense of direction a bit off.

Transformer (7/10): The Transformer was similar; however, it used more adventurous short-term harmonic progressions than the LSTM. Particularly, it utilized many surprising chords such as IV7, iv° , or II, with only brief resolutions. Sometimes this made the harmonies more confusing, especially when it switched between multiple musical periods. This may have contributed to some of the inaccuracies with the classifier, which struggled to distinguish between contemporary classical style and jazz progressions. Thus, a possible further research development could be to train the classical generator based on only contemporary-style music and see if it distinguishes it from jazz music.

Melodic Accuracy:

LSTM (8/10): Long melodic lines were used, with frequent melodic independence and counterpoint between various voices. There were more arpeggios compared to jazz, sometimes with clearly distinct melodic and harmonic components, with clear tonal center (when relevant) and variations of motives. Even though the melodic aspects were obviously less pronounced in the more contemporary-style music, it still managed to create a sense of direction in the short term.

Transformer (7/10) The Transformer's melodic accuracy depended on the generation. Sometimes it created coherent melodies; however, it could also be a quite simplistic collection of very similar notes. When successful, it was similar to the LSTM, producing melodic sections effectively, and sometimes utilizing arpeggiated melodic sections.

Cohesion in Form and Structure:

LSTM (6/10): It was somewhat coherent on a small timescale, using similar rhythmic structures and motifs, but did not have clear separate sections or beginnings and endings. For example, when it moved to a distinct section, it often progressed without returning to the original pattern later on. This highlighted some of the difficulties of LSTM in capturing long-range dependencies.

Transformer (7/10): The Transformer had better long-term cohesion and structure. The Transformer often returned to a similar tonal center and a similar rhythmic structure throughout. Of course, much of it may have been due to its tendency to repeat similar sections throughout the piece haphazardly without much change.

Variation:

LSTM (7/10): There was more harmonic variation than in jazz, and it shifted keys over time. It could also have many contrasting sections.

Transformer (6/10): In the first few lines, the variation was very similar to the LSTM. However, later on it sometimes became stuck in a loop of the same note over an extended period. It is true that the chord progressions could at times be longer and more complex than in LSTM, often spanning multiple lines, and ultimately resolving. However, many of these singular sections, which were fine on their own, could be repeated excessively throughout the piece.

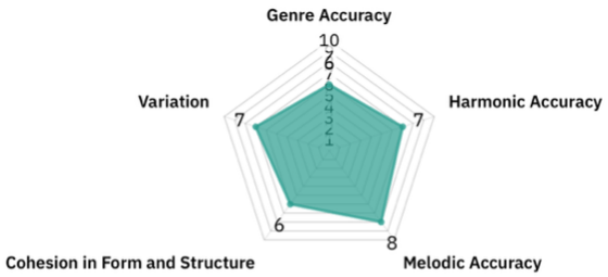

Fig. 8 LSTM Classical: Radar chart showing subjective ratings of 10 MIDI songs for LSTM classical music generation across 5 features.

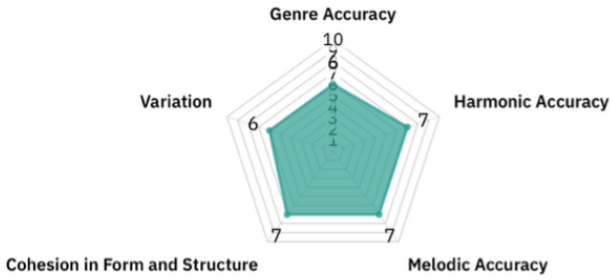

Fig. 9 Transformer Classical: Radar chart showing subjective ratings of 10 MIDI songs for Transformer classical music generation across 5 features.

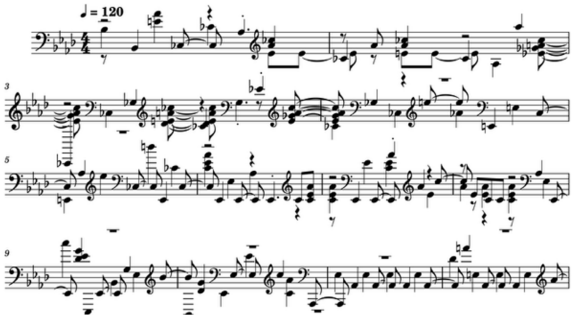

Fig. 10 LSTM Generation: Example of LSTM-generated classical music.

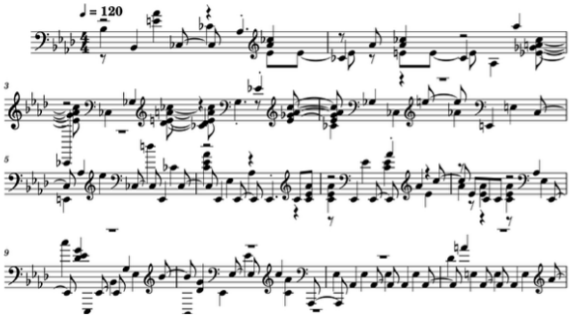

Fig. 11 Transformer Generation: Example of Transformer-generated classical music.

Jazz Music

Genre Accuracy:

LSTM (5/10): Overall, several dissonant chords and harmonies created a nuance. However, it sometimes lacked a clear sense of direction and could tend to be more like contemporary classical music rather than jazz music.

Transformer (6/10): The Transformer was similar to the LSTM. It featured more jazz-like harmonies, and occasionally a syncopated feel through chord placement, but passages often tended to be too repetitive and simplistic, outlining one harmony rather than creating melodic complexity.

Harmonic Accuracy:

LSTM (6/10): The LSTM was slightly better at capturing complicated harmonies than the Transformer through increased chord usage. However, it seemed to include fewer unique jazz-like chords.

Transformer (7/10): The composition featured many jazz-like chord progressions such as II-V-I, and incorporated other elements, including tritone substitutions or chromaticism, which contributed to the harmonic variety.

Melodic Accuracy:

LSTM (6/10): The melodic contours were sometimes challenging to distinguish amid the dense chords.

Transformer (6/10): There were discernible melodies; however, they could be quite simplistic with repetitions of scale-like passages.

Cohesion in Form and Structure:

LSTM (5/10): Short phrases often have coherent chord progressions. Sometimes, there was a defining motif that was repeated, and similar variations over half a page or so. However,

more so than in classical music, phrases often lacked a definite beginning or ending, possibly due to the difficulty of capturing small phrasing nuances within rapidly shifting chords.

Transformer (5/10): Individual phrases were usually complete, but they could diverge significantly from the original. Sometimes, it reverted to constant repetitions of similar notes that have little connection to the original music.

Variation:

LSTM (6/10): The first few lines could often be varied, and sometimes novel chords or improvisation-like structures were introduced. However, the constant usage of chords could become repetitive at times.

Transformer (5/10): When successful, it generates creative alternative sections. Similar to classical, however, the Transformer seemed to excessively repeat similar combinations over an extended period of time. In addition, in a single phrase, there was less variation in the types of notes compared to the LSTM.

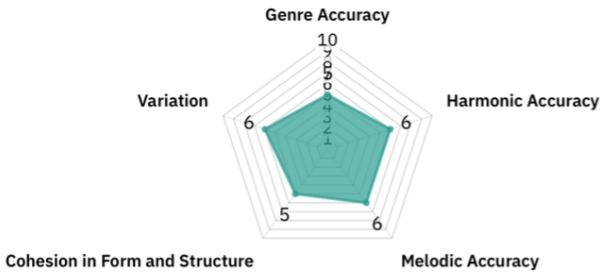

Fig. 12 LSTM Jazz: Subjective Ratings Radar Chart

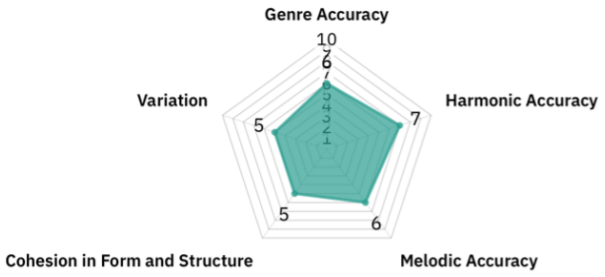

Fig. 13 Transformer Jazz: Subjective Ratings Radar Chart

There was no obvious difference between the LSTM and Transformer approaches in terms of jazz music generation: there were no features with greater than 2 rank differences. This was aligned with the assessment with classifiers.

Discussion

Firstly, we developed the LSTM and Transformer music genre classifiers using 150 classical and 150 jazz music files for training. Comparing the two methods, LSTM and Transformer, the LSTM model produced a classifier with higher accuracy than the Transformer model. However, the Transformer model was able to reach training saturation much faster than the LSTM model as shown in Figures 1 and 2. Music classifier performance was also confirmed by 20 independent classical and jazz data sets, and while both models achieved more than 90% accuracy and 80% specificity, the LSTM classifier had a better accuracy and specificity: LSTM accuracy 100% LSTM sensitivity 90%, Transformer accuracy 90% specificity 80% regarding identification of classical music.

Next, we trained a classical and jazz generator using two methods: LSTM and Transformer, with datasets comprising 30 classical and 30 jazz music files. The LSTM and Transformer generation had similar training accuracies of around 95% and 96% accuracy for classical, respectively, and 94% and 96% for jazz, respectively. Table 1 clearly indicates that in music generation, the Transformer was generally superior in terms of training time, with $1/7$ of the training time compared to the LSTM. As we noted in musical generation, the Transformer also required $1/5$ of the time for LSTM training. Thus, the Transformer was far superior to the LSTM in training time for both generation and classification. This may have been due to the self-attention mechanism that was used in the Transformer model, which allowed the system to recognize each piece of information simultaneously, compared to the LSTM, which recognizes the data step-by-step.

Finally, we compared two types of music generators (LSTM vs Transformer) to determine whether they can generate the correct genre of music (classical vs jazz) by assessing the generated music using LSTM and Transformer classifiers. For the classical music generation, the assessment with both LSTM and Transformer classifiers showed that there was no statistical difference between LSTM generation vs Transformer generation, indicating they had a similar performance. Also, the subjective analysis revealed that a similar musical structure was created by both generators (although individual files varied significantly) (Figures 8, 9, 10, and 11), possibly explaining the lack of statistical difference. On the other hand, for the jazz music generation, there was a statistical difference in the Transformer classifier, potentially implying the Transformer generator's superior music generation. Subjective analysis

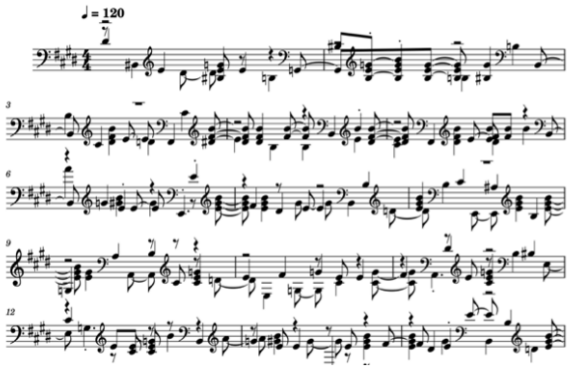

Fig. 14 LSTM-Generated Jazz Music Example

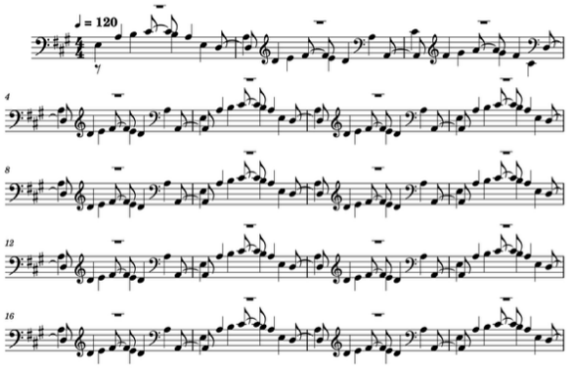

Fig. 15 Transformer-Generated Jazz Music Example

also showed that the Transformer generator seemed to be good at generating jazz-specific harmonic progressions, which were exemplified in the better performance in the Harmonic Accuracy and Genre Accuracy features, as shown in Figures 12, 13, 14 and 15. Interestingly, this result was not consistent with the LSTM assessment, possibly due to architectural or evaluation differences. Overall, the LSTM classifier seemed to be more robust across both classical and jazz in predicting the test sets. However, there didn't seem to be a significant difference between the LSTM and Transformer Generator. There could definitely be improvements.

In this study, with small-scale datasets of 60 MIDI files in total, both the LSTM and Transformer generation were similar for classical music, while the Transformer generation seemed to be superior for jazz music for the Transformer classifier and subjective analysis. This may have been due to the note vocabulary size of 604 in jazz, compared to 372 for classical, and due to sequential processing, the LSTM had a relatively hard time learning with many parameters compared to the Transformer, which made use of a self-attention mechanism.[2]

Interestingly, the differences between LSTM and Transformer were not obvious, especially in classical music generation. This might have been because we used a small-sized data set and focused on pitch as the only training feature, and thus, the relatively lower parameter study may have not been the best setting to evaluate the performance of the Transformer generator. In the subjective analysis, there is some room for improvement for jazz music generation: the small-scale variations were relatively well captured while broader structural elements and organization could be improved. This was true for both LSTM and Transformer generation. This was aligned with the previous evaluation of LSTM models, where Skúli mentioned the lack of beginnings or endings.[19]

The structure of the compositions seemed to be in 2 or 3 large sections of around 30 bars long (4-8 phrases), with each section having some related musical ideas, but rather disconnected from the other sections. For classical, the sections seemed to consist of a continued sequence of phrases of around 4-5 bars long (approx. 32-40 notes), often combined into several longer 8-bar phrases. This is surprisingly consistent with standard short-term phrase lengths in human-made classical music. However, the short phrases sometimes lacked a clear direction (such as toward or away from a climax), meaning the phrase lengths could be unpredictable and only discernable when looked at retrospectively. For jazz, the phrases seemed to be a similar length of around 4 bars while rather than melodic phrases (such as those in the classical music files), most of the phrasing was done through the harmonic rhythm, sometimes making the phrases less immediately identifiable or improvisatory.

Notably, this study clearly showed that even with a small-sized data set, both the LSTM and Transformer could create music that fits the genre of music that was used for training, and that LSTMs and Transformers may have different strengths in different musical genres, meeting the objectives. Although LSTMs and Transformers have been studied extensively for general music generation in the past, their capabilities for specific types of music have remained less understood. This could inform the realistic generation of music pertaining to a particular style. In the future, this may be helpful for composers to edit their compositions in a certain genre or for music educators to highlight analytical aspects specific to a genre[21,22].

There are some limitations. The genre accuracies of classic and jazz music generated by LSTM and Transformers were both 0.54 - 0.63, indicating that further improvements are possible. As noted in the subjective analysis, variation in the music could be improved, especially jazz, and classical music could have less mixing between styles from separate classical periods. This is because the model only used jazz and classical genres of music for training, instead of including other

genres such as blues, pop, or rock. Parameters such as dropout rates may be reconsidered to prevent overfitting. Additionally, due to computational and time constraints, a small dataset was used, and only pitch was considered as a feature in order to reduce the note vocabulary size for training. This also limited the quality of music generated.

In order to further improve the generator, future work would benefit from considering features such as note velocity or duration to increase the variation of generated music, as well as using the chordify() function to parse notes across different[18,22]. This would aid in incorporating rhythmic structures and nuance and enhancing realism of generated music. During this process, a larger dataset, such as the MAESTRO dataset, which contains 200 hours of virtuosic piano performances, may be beneficial to account for the increased note vocabulary dimension size. For the Transformer model, techniques such as L2 regularization could help reduce overfitting by penalizing the weights[12,21]. Creating multiple instrument parts may additionally be interesting. Furthermore, due to the difficulty of producing coherent and interesting music as a whole, other applications such as short-term editing or part-writing may be particularly helpful to explore further as practical applications of AI-based music generation in the near term.[5,23,24]

In addition, due to the disorienting combination of training files from composers of different periods, it might also be useful to stratify by time period, such as Baroque/Classical and Romantic/Contemporary, to create a truly classical-sounding piece. Learning forms and phrasing explicitly may be useful, as well as hybrid LSTM/Transformer models to capture their respective strengths.[4,6]

Finally, we parsed the music files together to train the classifier and generator. This method allowed the creation of a format easily used for training, which made the beginnings and endings of each music obscure. It is also good to consider exploring ways to enhance the cohesion of each piece by training based on separate files.

For future studies, considering these aspects could broaden the generation and classification to more genres while increasing the accuracy and specificity of generated music. In addition, this research could be extended by fine-tuning[13,18,25,26].

Conclusion

In the context of AI Music Generation, LSTM and Transformer models work well for generating classical and jazz music. This study demonstrated the differences between LSTMs and Transformers in their application to music by comparing their performance in classical and jazz music generation.

Evaluation through genre classification models demonstrated that the Transformer was superior for jazz music generation compared to the LSTM, while both were comparable for classical music generation. These results imply that tailoring the models to the genres (e.g. a hybrid system with LSTM and transformer) would lead to the best performance in music generation[4,9]. While the results and their implications need to be validated in future studies with robust data sets and larger genres, this study provided useful information that could be a foundation for the generation of music pertaining to a particular style.[10]

Acknowledgements

I would like to thank my mentors for guiding me through the process of writing this paper. Their expertise and support have contributed greatly to my project's completion.

References

1 I. Agchar, I. Baumann, F. Braun, P. A. Perez-Toro, K. Riedhammer, S. Trump and M. Ullrich, *arXiv preprint arXiv:2402.15294*, 2024.

2 M. Arektout, *MUSIC GENERATION USING RNN-LSTM*, 2024.

3 M. Waqas and U. W. Humphries, *MethodsX*, 2024, **13**, 102946.

4 R. Mayya, V. Venkataraman, N. Darapaneni *et al.*, *arXiv preprint arXiv:2404.05765*, 2024.

5 A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, *arXiv preprint arXiv:2301.11325*, 2023.

6 S. Ji, X. Yang and J. Luo, *ACM Computing Surveys*, 2023, **56**, 1–39.

7 Y.-J. Shih, S.-L. Wu, F. Zalkow, M. Müller and Y.-H. Yang, *IEEE Transactions on Multimedia*, 2022, **25**, 3495–3508.

8 G. Mittal, J. Engel, C. Hawthorne and I. Simon, *arXiv preprint arXiv:2103.16091*, 2021.

9 D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata and X. Serra, Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 855–858.

10 N. Ndou, R. Ajoodha and A. Jadhav, 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021, pp. 1–6.

11 E. Y. Zhang, A. D. Cheok, Z. Pan, J. Cai and Y. Yan, *Sci*, 2023, **5**, 46.

12 T. Lin, Y. Wang, X. Liu and X. Qiu, *AI open*, 2022, **3**, 111–132.

13 S. Kayala, *Jazz ML ready MIDI*, Kaggle, 2018, `https://www.kaggle.com/datasets/saikayala/jazz-ml-ready-midi`.

14 H. Bahuleyan, *arXiv preprint arXiv:1804.01149*, 2018.

15 S. Rakshit, *Classical music MIDI*, Kaggle, 2019, `https://www.kaggle.com/datasets/soumikrakshit/classical-music-midi/data`.

16 Z. Xiong, W. Wang, J. Yu, Y. Lin and Z. Wang, *arXiv preprint arXiv:2308.13736*, 2023.

17 A. Oki, *Music-LSTM-Transformer*, GitHub, `https://github.com/Vacuum1234/Music-LSTM-Transformer`.

18 Z. Jiang, S. Li and Y. Sung, *Mathematics*, 2022, **10**, 2747.

19 M. Bajor and M. Niemiec, *Information & Security*, 2020, **47**, 261–275.

20 A. Nandan, *Text classification with Transformer*, Keras Examples, 2020, `https://keras.io/examples/nlp/text_classification_with_transformer/`.

21 P. Ferreira, R. Limongi and L. P. Fávero, *Applied Sciences*, 2023, **13**, 4543.

22 W.-Y. Hsiao, J.-Y. Liu, Y.-C. Yeh and Y.-H. Yang, Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 178–186.
23 H. Sun, X. Wang, Y. Wang and P. Lu, *Scientific Reports*, 2025, **15**, 19943.
24 P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford and I. Sutskever, *arXiv preprint arXiv:2005.00341*, 2020.
25 L.-C. Yang and A. Lerch, *Neural Computing and Applications*, 2020, **32**, 4773–4784.
26 M. Zhang, *Scientific Reports*, 2025, **15**, 28007.

Supplementary Material

Supplementary Figures

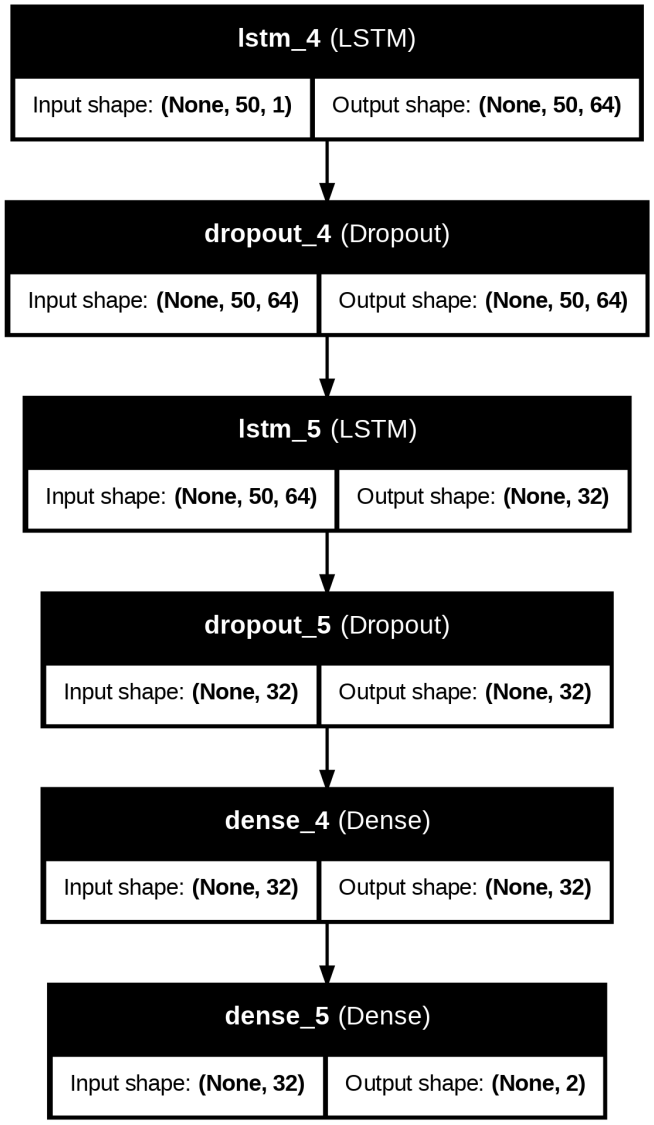

Fig. 16 LSTM Classification Architecture. The model takes input sequences of length 50 and processes them through LSTM layers to produce genre classification outputs.

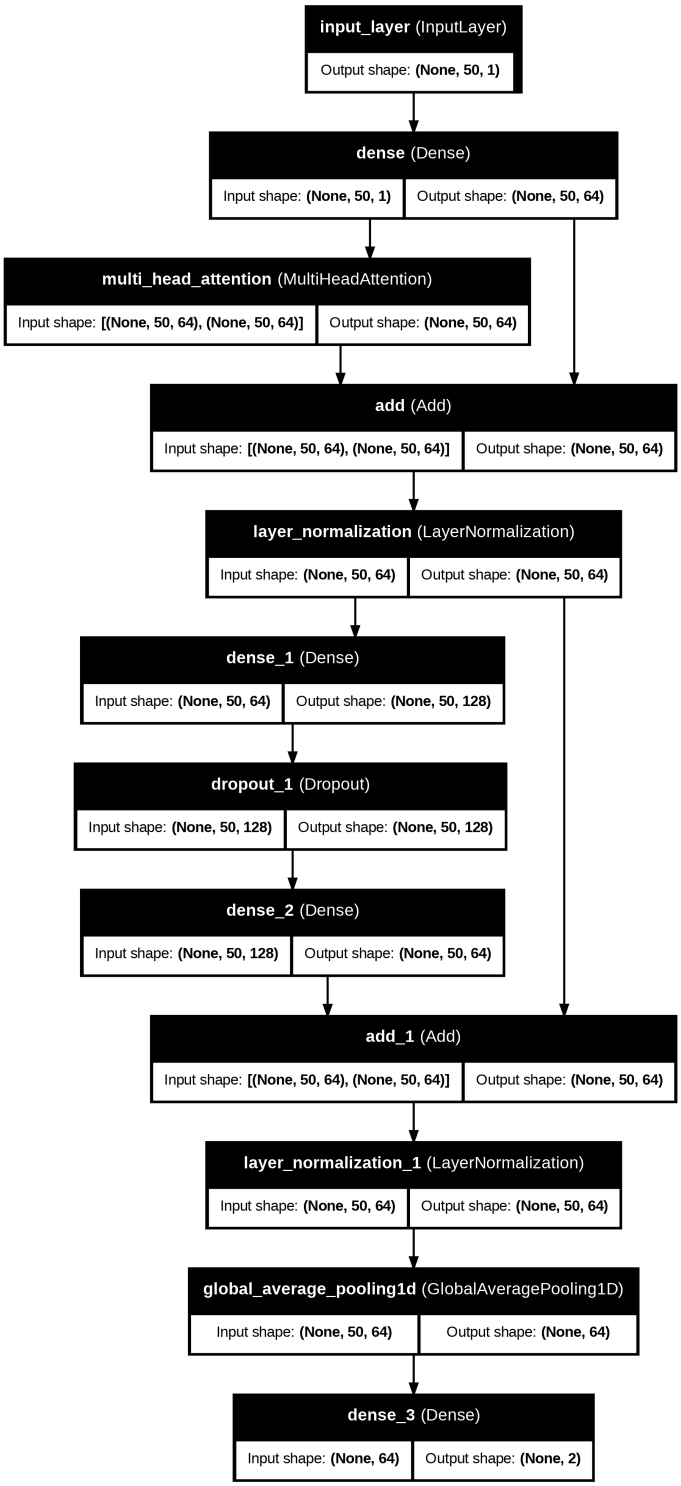

Fig. 17 Transformer Classification Architecture. The model takes input sequences of length 50 and processes them through embedding, multi-head attention, and feed-forward layers to produce genre classification outputs.

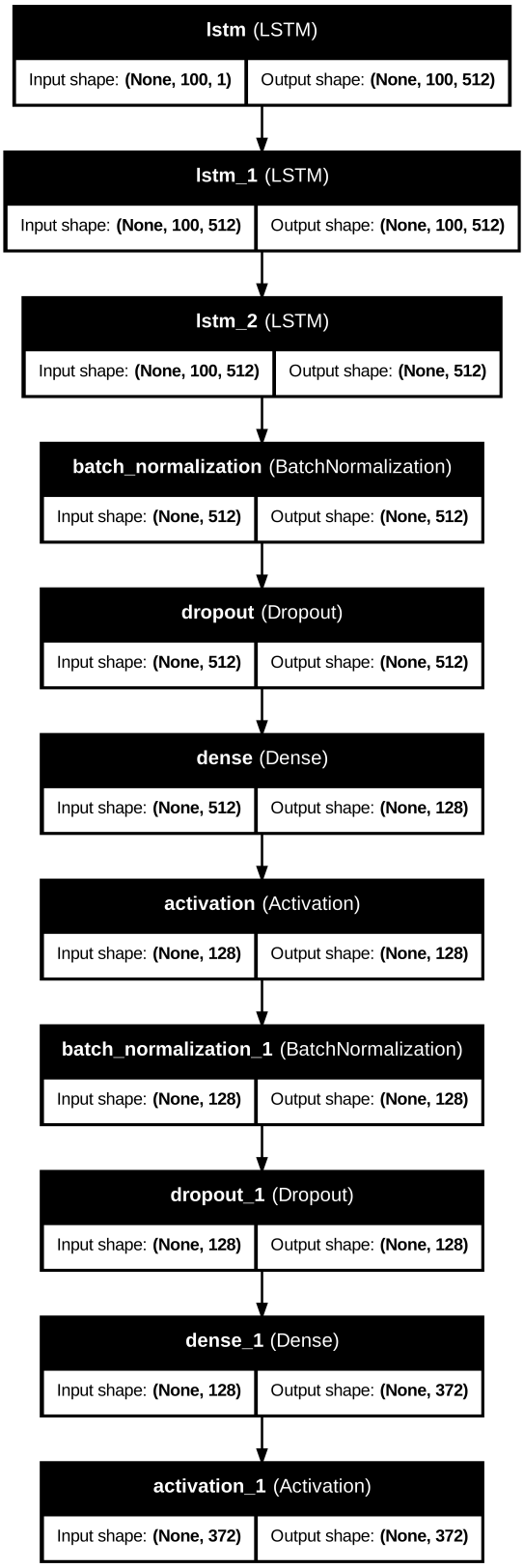

Fig. 18 LSTM Classical Generation Architecture. The model uses three LSTM layers with dropout and batch normalization to generate jazz music sequences, outputting probability distributions over 372 possible notes.

Fig. 19 Transformer Classical Generation Architecture. This model generates classical music sequences using token and position embeddings followed by transformer blocks and dense layers to predict the next note in the sequence.

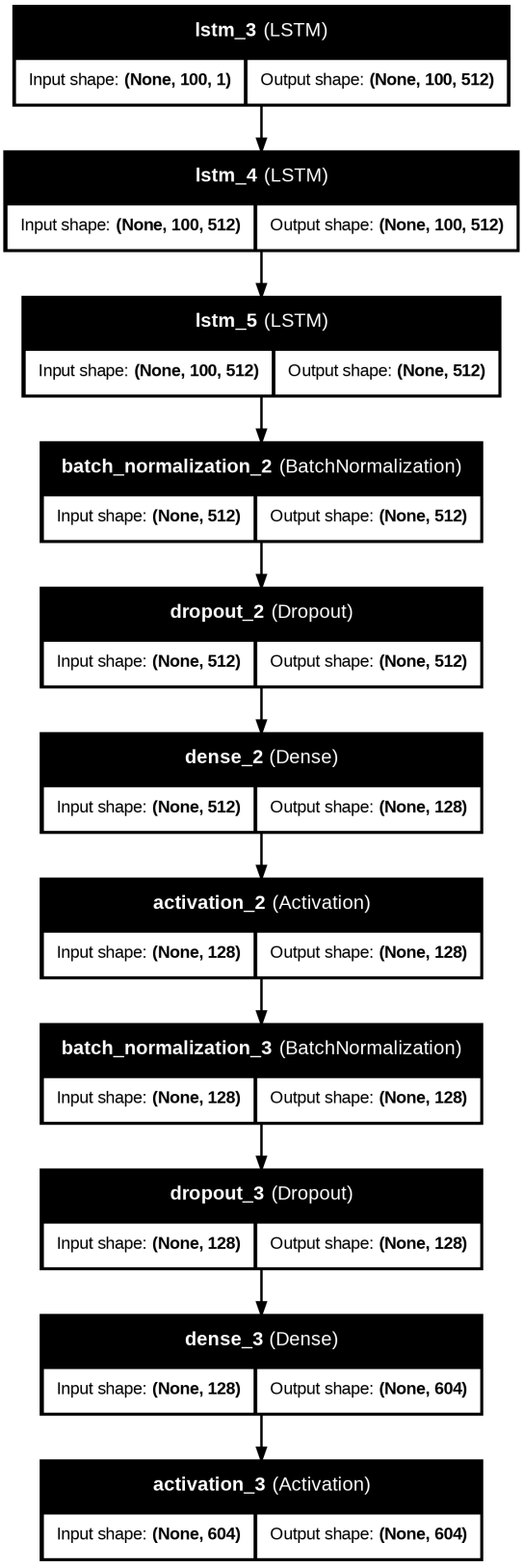

Fig. 20 LSTM Jazz Generation Architecture. The model uses three LSTM layers with dropout and batch normalization to generate jazz music sequences, outputting probability distributions over 604 possible notes.

Fig. 21 Transformer Jazz Generation Architecture. Similar to the classical generator but trained on jazz music data, this model uses transformer blocks to capture long-range dependencies in jazz compositions.