# Modeling Musical Harmony: Sequence Architectures for Chord Recognition

**Yutong He**
Heidelberg University
yutong.he@stud.uni-heidelberg.de

## Abstract

This project aims to compare the performance of three sequence modeling architectures, including BiLSTM, Mamba, and TCN, on the task of automatic chord recognition from audio. Using chroma features extracted from the McGill Billboard dataset as input and simplified chord annotations as labels, we train and evaluate each model under the same experimental setup. Our results show that BiLSTM achieves the highest prediction accuracy and macro-averaged F1 score, while Mamba and TCN offer superior evaluation speed and throughput. These findings highlight the trade-offs between accuracy and efficiency in designing neural models for music analysis tasks. Project code available at GitHub.[1]

## 1 Introduction

The development of neural networks, particularly in sequence-to-sequence learning, has significantly advanced the field of audio and music understanding. Tasks such as automatic speech recognition, machine translation, and music generation have benefited from the ability of neural models to process sequential input and output in a context-aware manner. In this project, we apply such techniques to the task of chord recognition from audio, a foundational problem in music information retrieval (MIR).

Chord recognition involves identifying the underlying harmonic structure of a piece of music by labeling segments of audio with their corresponding chord symbols (e.g., C:maj, A:min). Accurate chord transcription is essential for a range of applications, including automatic music transcription, music recommendation, and educational tools for music learners. Given the sequential nature of harmonic progressions in music, chord recognition is well-suited to sequence modeling approaches, where each input frame is mapped to a chord label in temporal order.

In this project, we focus on chord recognition using chroma features extracted from the McGill Billboard dataset(Burgoyne et al., 2011) [2], a widely used benchmark corpus that provides time-aligned harmonic annotations for hundreds of popular songs. We use the `bothchroma.csv` files as model input and the simplified `majmin.lab` annotations as ground-truth labels.

To evaluate the effectiveness of different sequence modeling approaches, we implement and compare three architectures: a Bidirectional Long Short-Term Memory (BiLSTM) network(Schuster and Paliwal, 1997), the recently proposed state-space model Mamba(Gu and Dao, 2024), and a Temporal Convolutional Network (TCN)(Lea et al., 2016). All models are trained and tested on the same dataset split and evaluated using standard classification metrics. Our goal is to analyze the trade-offs between model performance and computational efficiency in the context of music chord recognition.

## 2 Dataset

When it comes to audio chord recognition, the McGill Billboard dataset (Burgoyne et al., 2011) has become one of the most reliable and high-quality training datasets. This dataset collects time-aligned transcriptions of the harmony in more than a thousand songs, randomly selected from the Billboard "Hot 100" chart in the United States between 1958 and 1991. Among these, 890 entries include chord annotations that are aligned with time, and corresponding audio features have been extracted, which are used by our project. In this project, we use these 890 annotated tracks for model training and evaluation.

For every single track, we use the bothchroma

[1] https://github.com/WitchPuff/chord_reg

[2] The official website of McGill Billboard dataset.

features as input and the majmin-style chord annotations as labels for supervised learning. Specifically, we extract chroma vectors from the bothchroma.csv files, which were computed using the Chordino VAMP plugin (Mauch and Dixon, 2010) [3]. These features provide a time-aligned, pitch-class representation of the harmonic content in each track. The corresponding chord labels are taken from the `majmin.lab` files, which contain time-aligned chord annotations limited to major and minor triads, excluding chord inversions, sevenths, and other complex chord types, providing a simplified and widely used label set for chord recognition tasks. This setting allows for a clean and consistent mapping between acoustic features and chord labels, facilitating effective model training and evaluation on the chord recognition task.

## 3 Methods

To perform chord recognition based on extracted chroma features, we implement and compare three different sequence modeling architectures: a bidirectional Long Short-Term Memory (BiLSTM) network (Schuster and Paliwal, 1997), a state-space model based on Mamba (Gu and Dao, 2024), and a Temporal Convolutional Network (TCN) (Lea et al., 2016). Each model takes as input the time-aligned chroma vectors extracted from the `bothchroma.csv` files and outputs a time-aligned sequence of predicted chord labels aligned with the ground truth annotations.

### 3.1 BiLSTM

The BiLSTM model is a recurrent neural network capable of capturing long-range temporal dependencies in sequential data. It consists of two LSTM layers processing the input sequence in forward and backward directions, allowing the model to integrate both past and future context. This bidirectional structure is particularly suitable for music, where chord progression often depends on surrounding harmonic context. A fully connected output layer with softmax activation maps the hidden states to chord label predictions at each time step.

### 3.2 Mamba

Mamba is a recently proposed state-space sequence model that combines the efficiency of convolutional approaches with the long-range memory capacity of recurrent structures. Unlike RNNs, which rely on recursive updates, Mamba employs structured state-space representations that allow for highly parallelizable training and efficient long-sequence modeling. In our implementation, we use Mamba as a backbone to process chroma sequences and output chord predictions, benefiting from its ability to model global dependencies while maintaining low computational cost.

### 3.3 TCN

The Temporal Convolutional Network (TCN) is a causal convolutional model designed for sequence modeling tasks. It replaces recurrence with dilated 1D convolutions, allowing for a large receptive field while maintaining stable gradients during training. TCNs offer efficient parallelization and are particularly well-suited for temporal tasks like audio processing. In our setup, the TCN processes the input chroma sequence through multiple residual blocks with increasing dilation rates, followed by a linear output layer for chord classification.

## 4 Experiments

### 4.1 Setup

We evaluate three different model architectures for chord recognition: BiLSTM, TCN, and Mamba. For each model, we experiment with three different hidden dimensions: 64, 128, and 256. All models are trained for 30 epochs with a batch size of 32 and a learning rate of 5e-3. By default, the HuggingFace `Trainer` applies a linear learning rate scheduler with no warm-up steps, gradually decaying the learning rate to zero over the course of training. We split the dataset into training, validation, and test sets with a ratio of 8:1:1. The total number of data points is 890.

We use the cross-entropy loss function for training, and optimization is performed using the AdamW optimizer, following the default configuration of the HuggingFace `Trainer` API. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

Model performance is evaluated using multiple metrics, including macro-averaged F1 score, accuracy. We log all training and evaluation metrics using Weights & Biases (`wandb`) for reproducibility and visualization.

### 4.2 Results

We evaluate all models on the test set using two main classification metrics: macro-averaged F1

[3] http://www.isophonics.net/nnls-chroma

score and accuracy. Additionally, we measure evaluation efficiency using average runtime and samples processed per second during validation.

Table 1 shows that BiLSTM consistently achieves the best macro F1 scores across all hidden dimensions, with the highest score of **0.7417** at hidden dimension 256. Mamba and TCN also improve with increasing capacity, reaching their peak F1 scores at 256 and 128 hidden dimensions respectively. However, their performance still falls short of BiLSTM, suggesting that bidirectional recurrent structures are more effective at capturing harmonic dependencies in chord sequences.

The accuracy results in Table 2 show a similar trend: BiLSTM again outperforms the other models across all configurations, with a maximum accuracy of **0.7618**. Mamba follows with 0.7376, and TCN reaches 0.7333 at its best. These results are further supported by the training curves in Figures 1 and 2, which show that BiLSTM not only achieves higher final performance but also converges more smoothly.

In terms of computational efficiency, Table 3 shows that Mamba is the fastest model, with an average evaluation runtime of only **1.5834s** per epoch, slightly outperforming TCN (1.5891s) and significantly faster than BiLSTM (1.7715s). Similarly, Table 4 indicates that Mamba processes the most validation samples per second on average (**56.59**), followed closely by TCN (56.34), and then BiLSTM (50.31). This trend is visualized in Figures 3 and 4, where Mamba consistently leads in throughput and runtime efficiency.

Overall, the results reveal a trade-off between performance and efficiency. While BiLSTM achieves the best prediction accuracy and F1 scores, Mamba and TCN offer faster inference speeds. Depending on application needs, such as the trade-off between real-time processing and accuracy-sensitive scenarios, different architectures may be preferred.

## 5 Limitations

However, the overall performance across all models remains limited, likely due to the small size of the training dataset. We believe this issue could be mitigated by incorporating a larger and more diverse dataset to improve generalization. Additionally, increasing model capacity, particularly for BiLSTM, by using higher hidden dimensions may further enhance performance. Beyond model architecture,

| Model | Hidden Dim | F1 (macro) |
|---|---|---|
| BiLSTM | 64 | 0.7162 |
| BiLSTM | 128 | 0.7328 |
| BiLSTM | 256 | **0.7417** |
| Mamba | 64 | 0.6777 |
| Mamba | 128 | 0.7030 |
| Mamba | 256 | **0.7116** |
| TCN | 64 | 0.6814 |
| TCN | 128 | **0.7040** |
| TCN | 256 | 0.6939 |

Table 1: Test set F1-macro scores for different models and hidden dimensions.

| Model | Hidden Dim | Accuracy |
|---|---|---|
| BiLSTM | 64 | 0.7424 |
| BiLSTM | 128 | 0.7543 |
| BiLSTM | 256 | **0.7618** |
| Mamba | 64 | 0.7087 |
| Mamba | 128 | 0.7292 |
| Mamba | 256 | **0.7376** |
| TCN | 64 | 0.7244 |
| TCN | 128 | **0.7333** |
| TCN | 256 | 0.7122 |

Table 2: Test set accuracy for different models and hidden dimensions.

improvements in feature representation could also contribute significantly. For example, replacing handcrafted chroma features with learned representations from models like MusicBERT, which encodes symbolic music as sequences of octuples, may offer a richer understanding of musical structure. However, such approaches require substantially more computational resources and access to symbolic (e.g., MIDI) data, which may not always be feasible in practical settings.

## 6 Conclusion

In this work, we conducted a comparative study of three sequence modeling architectures: BiLSTM, Mamba, and TCN, for the task of chord recognition using chroma features extracted from the McGill Billboard dataset. Our experiments show that BiLSTM consistently achieves the highest performance in terms of both macro-averaged F1 score and accuracy across all model sizes. Mamba and TCN, on the other hand, demonstrate significantly better efficiency, with faster evaluation runtimes and higher throughput measured in samples per second.

These results highlight a trade-off between pre-

| Model | Average Runtime (s) |
|---|---|
| BiLSTM | 1.7715 |
| Mamba | **1.5834** |
| TCN | 1.5891 |

Table 3: Average evaluation runtime per epoch for each model with hidden dimension 256.

| Model | Average Samples/Sec |
|---|---|
| BiLSTM | 50.31 |
| Mamba | **56.59** |
| TCN | 56.34 |

Table 4: Average number of evaluation samples processed per second for each model with hidden dimension 256.

dictive performance and computational efficiency. While BiLSTM is more accurate, it requires more time to evaluate, whereas Mamba and TCN offer faster inference suitable for real-time or resource-constrained applications. Future work could explore larger datasets, more expressive model architectures, and richer input representations, such as symbolic music embeddings, to further advance performance in chord recognition tasks.

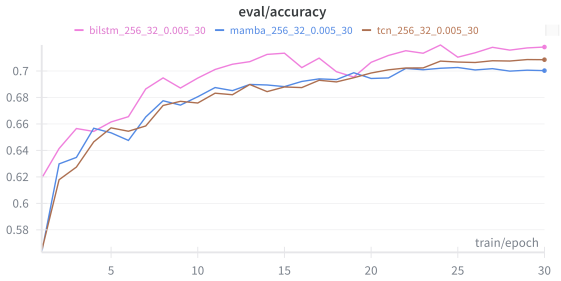


Figure 1: Accuracy on validation set during training for different models.

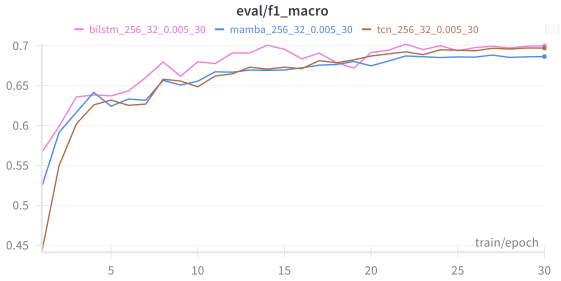


Figure 2: Macro F1 on validation set during training for different models.

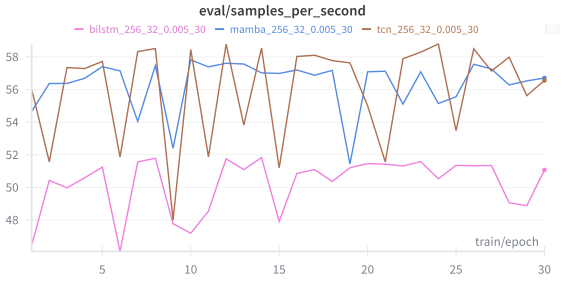


Figure 3: Samples per second on validation set during training for different models.

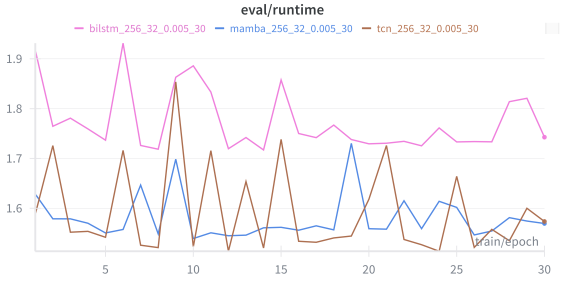


Figure 4: Runtime on validation set during training for different models.

## References

John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. 2011. An expert ground truth set for audio chord recognition and music analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 633–638, Miami, FL, USA.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. *Preprint*, arXiv:2312.00752.

Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. 2016. Temporal convolutional networks for action segmentation and detection. *Preprint*, arXiv:1611.05267.

Matthias Mauch and Simon Dixon. 2010. Approximate note transcription for the improved identification of difficult chords. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 135–140, Utrecht, Netherlands.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.