AI Charting for Music Game Cytoid

•••

Final presentation

Yuchen Song

Content

Cytoid Game Logo[1]

[1] Cytoid. https://cytoid.io. [Accessed: Nov. 6, 2024]

Introduction

Cytoid

Open-source music rhythm game[1]

Active community for player

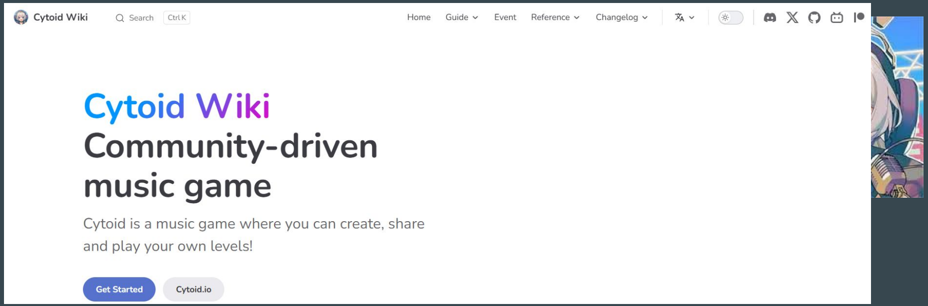
Nice charts for all difficulties

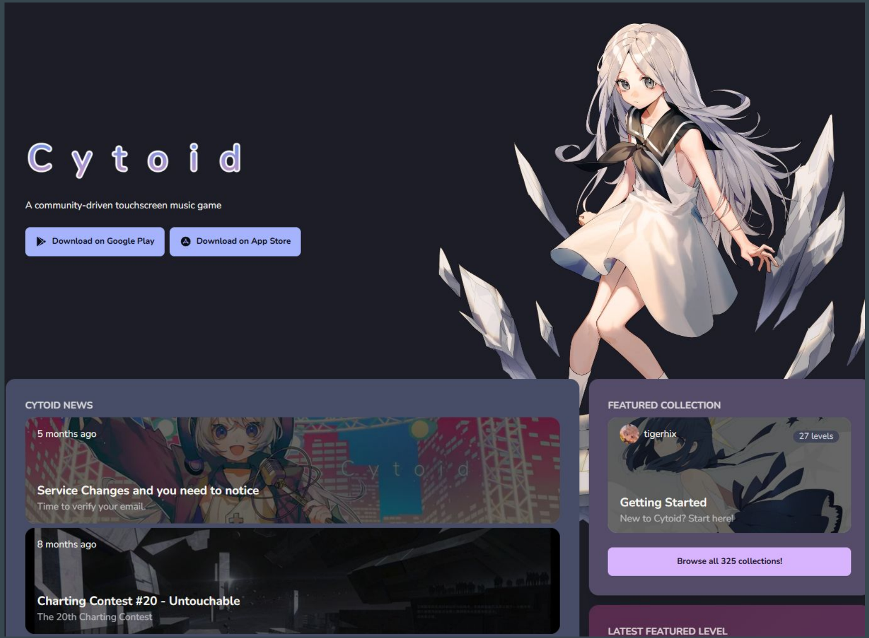
Challenges:

Complex and Time consuming

for players to create new charts.

[1] Cytoid. https://cytoid.io. [Accessed: Nov. 6, 2024]



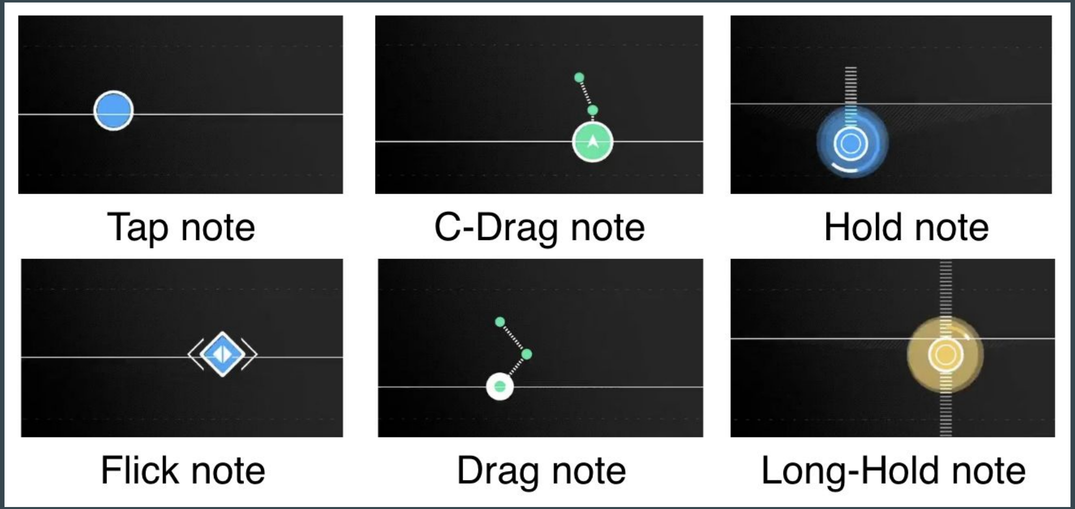


Charting for Music Games

Appear time

Key types

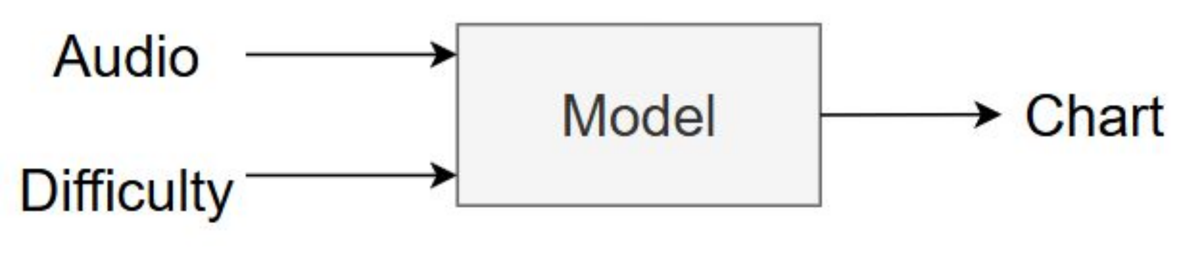
Place position on scanning line



[2] Cytoid Community, "Cytoid Gameplay," Google Sites. [Online]. Available: https://sites.google.com/site/cytoidcommunity/gameplay/cytoid-gameplay?authuser=0. [Accessed: Dec. 9, 2024].

Problem Statement

With the input of any song audio and desired difficulty level, the model will output a cytoid-compatible chart that can be played.



Related Work

Related Work

Dance Dance Convolution (2017)[3]:

- Step placement & step selection
- CNN & LSTM

TaikoNation (2021)[4]:

- human-like chart pattern
- BiLSTM

Genelive (2023)[5]:

- Beat information
- Multi-scale convolutional network LSTM

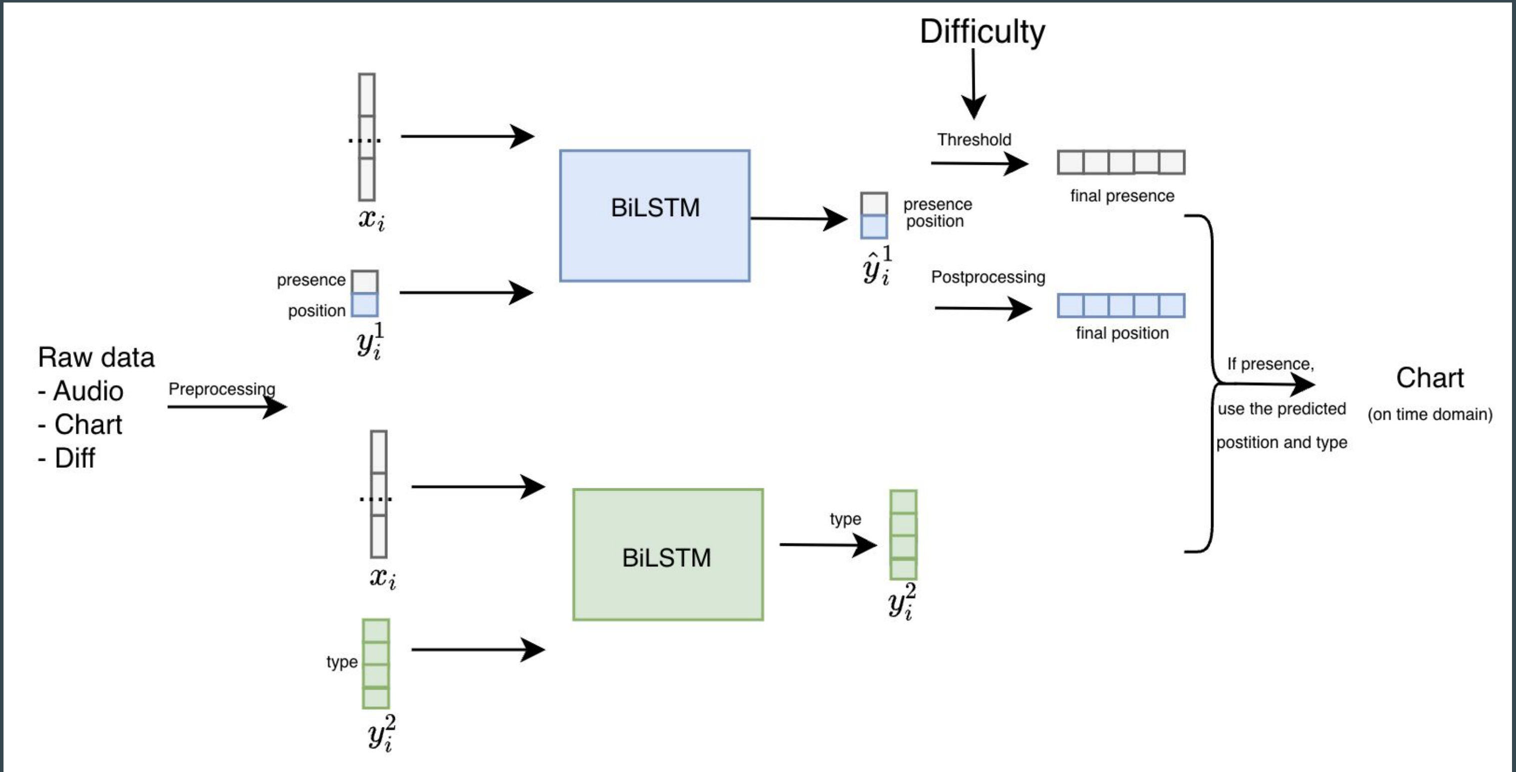
Mug Diffusion (2023)[6]

- Diffusion model

[3] D. Donahue, K. Simonyan, A. Zisserman, and G. Vondrick, “Dance Dance Convolution: Learning Generative Models for Dance,” Proc. IEEE Int. Conf. on Computer Vision, 2017, pp. 1-9.
[4] S. Siami-Namini, N. Tavakoli, and A. S. Namin, ”The performance of LSTM and BiLSTM in forecasting time series,” in Proc. 2019 IEEE Int. Conf. on Big Data (Big Data), pp. 3285–3292, 2019
[5] A. Takada, D. Yamazaki, Y. Yoshida, N. Ganbat, T. Shimotomai, N.Hamada, L. Liu, T. Yamamoto, and D. Sakurai, ”Gen´elive! Generating Rhythm actions in Love Live!,” in Proc. AAAI Conf. on Artificial Intelligence, vol. 37, no. 4, pp. 5266–5275, 2023.
[6] C. MX, "Mug-Diffusion: High-quality and controllable charting AI for rhythm games," GitHub repository, 2024. [Online]. Available: https://github.com/Keytoyze/Mug-Diffusion. [Accessed: Dec. 8, 2024].

Proposed Methodology

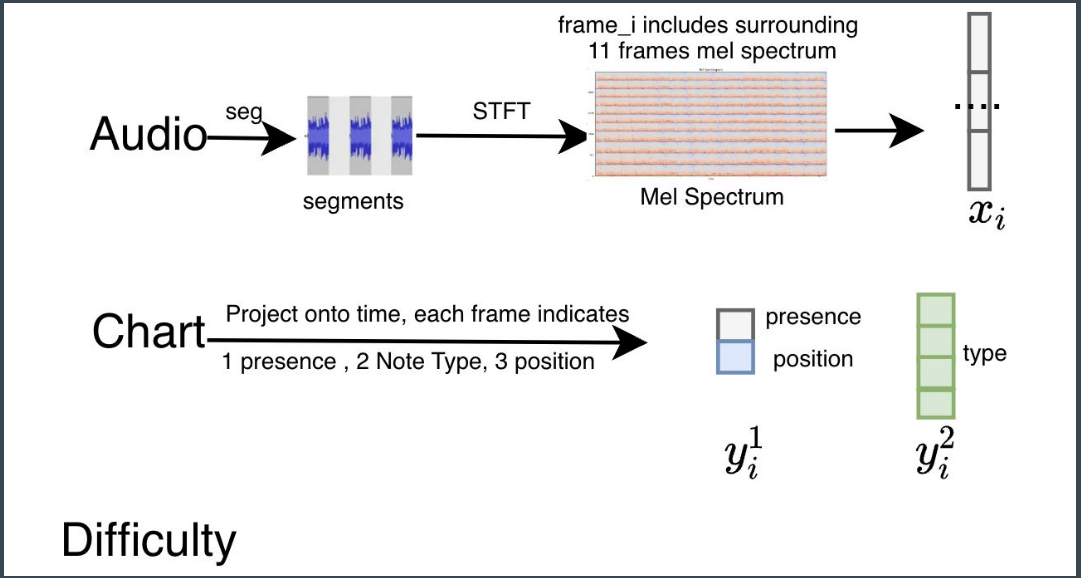
Original Model with BiLSTM



Preprocessing for Original Model

Preprocessing:

- Segmentation
- Short Time Fourier Transform
- One-hot encoding for key types
- (no key is also a type)



Follow structure of Genelive[5] (2023)

[5] A. Takada, D. Yamazaki, Y. Yoshida, N. Ganbat, T. Shimotomai, N.Hamada, L. Liu, T. Yamamoto, and D. Sakurai, "Gen´elive! Generating Rhythm actions in Love Live!," in Proc. AAAI Conf. on Artificial Intelligence, vol. 37, no. 4, pp. 5266–5275, 2023.

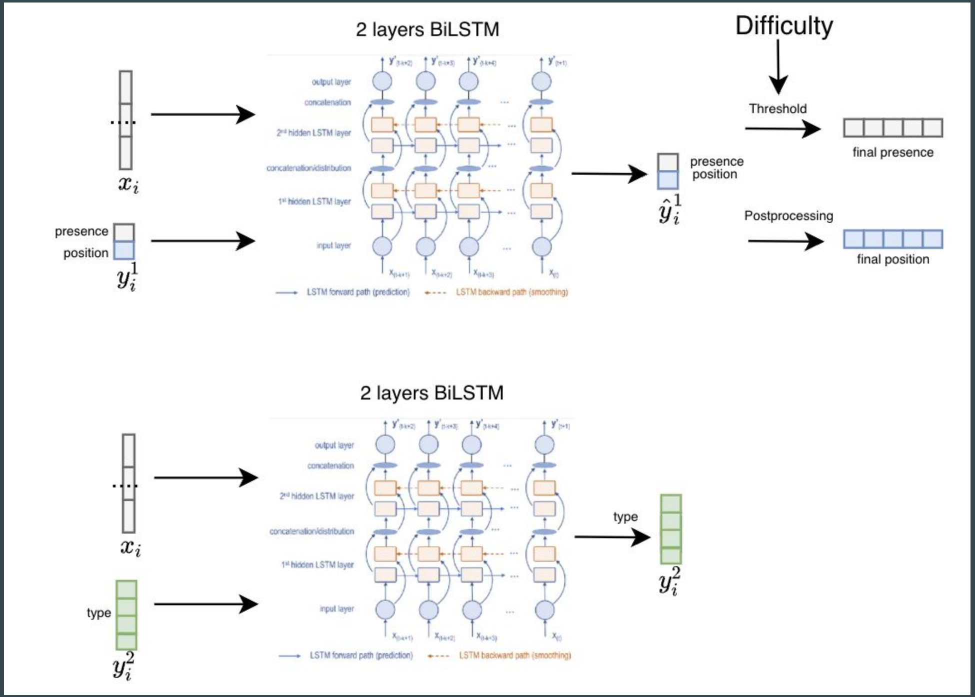
Original BiLSTM Model

Note Presence Prediction:

- Bidirectional LSTM model
- Apply threshold based on difficulty

Note Type Prediction:

- Separate LSTM model
- Type 0 indicates no note



Threshold Selection for Difficulty levels

(1) Average threshold based:

Calculate average threshold for each difficulty level. (follow DDC[3])

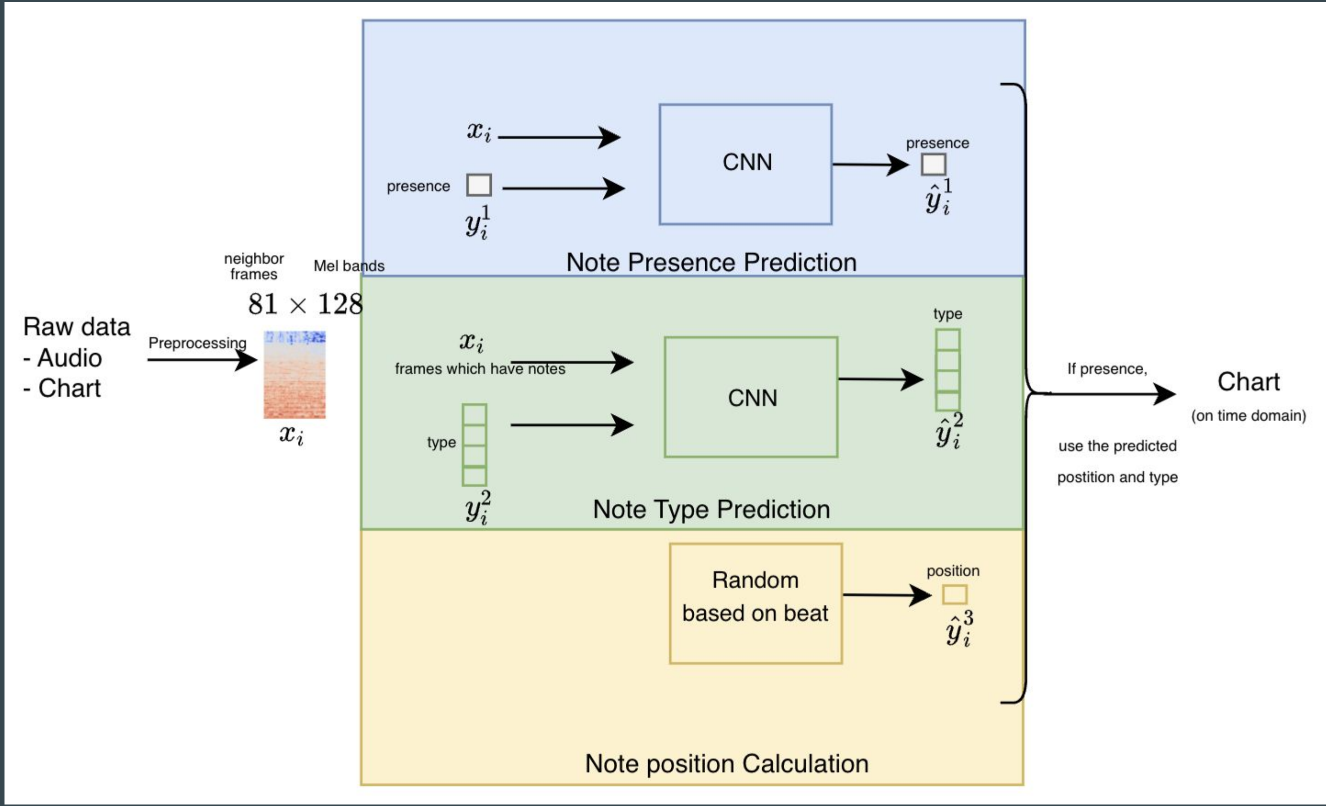
(2) Note Number based:

Selecting the top 'n' notes based on predicted probabilities.

'n' being the average notes number for the difficulty level.

[3] D. Donahue, K. Simonyan, A. Zisserman, and G. Vondrick, “Dance Dance Convolution: Learning Generative Models for Dance,” Proc. IEEE Int. Conf. on Computer Vision, 2017, pp. 1-9.

Final Model with CNN



Model Details

CNN Presence Prediction:

- Feature Extraction: 81 frames of mel band per frame
- Model: Three-layer CNN with sigmoid activation function

CNN Type Prediction:

- Note Type Prediction: Only frames with notes as input (Excludes empty frames)
- Softmax activation function
- Use Focal Loss[7] (Weighted version of BCE) to solve Type Imbalance

Random Note position:

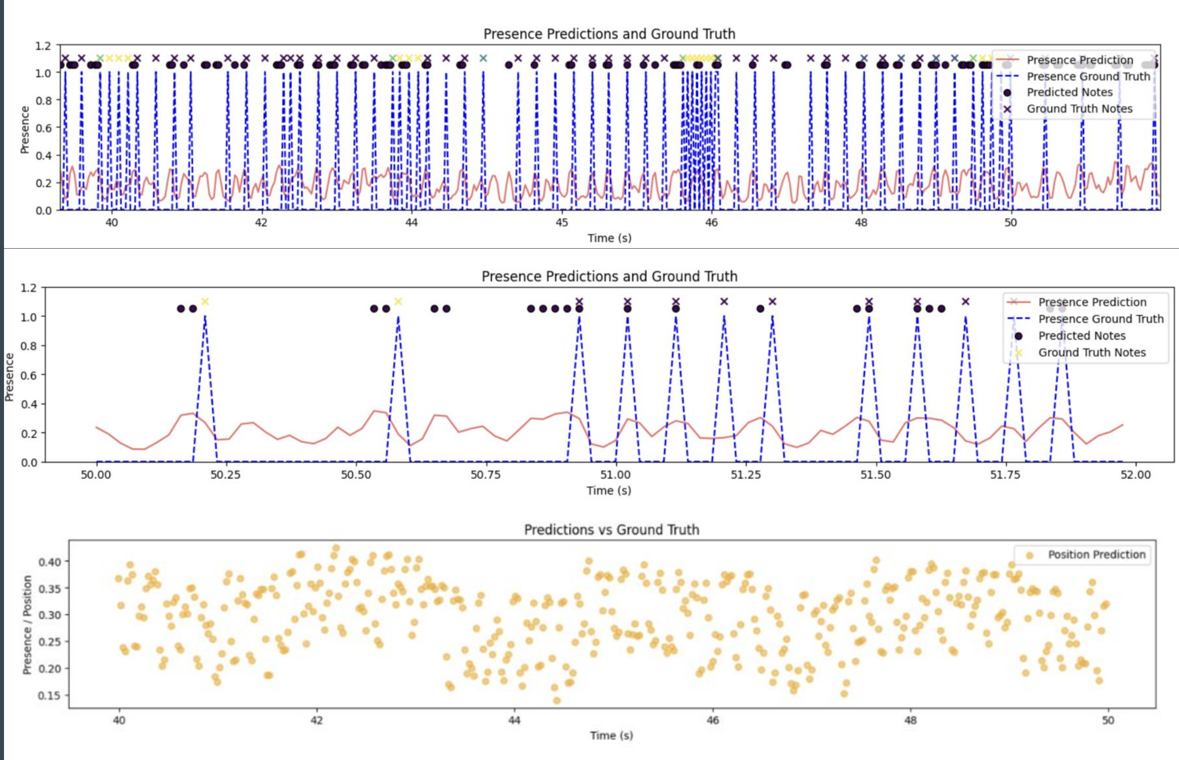
- Apply uniform distribution within measures with random disturbance

[7] Lin, T. "Focal Loss for Dense Object Detection." arXiv preprint arXiv:1708.02002 (2017).

Performance Evaluation

BiLSTM Model: Note Presence Prediction

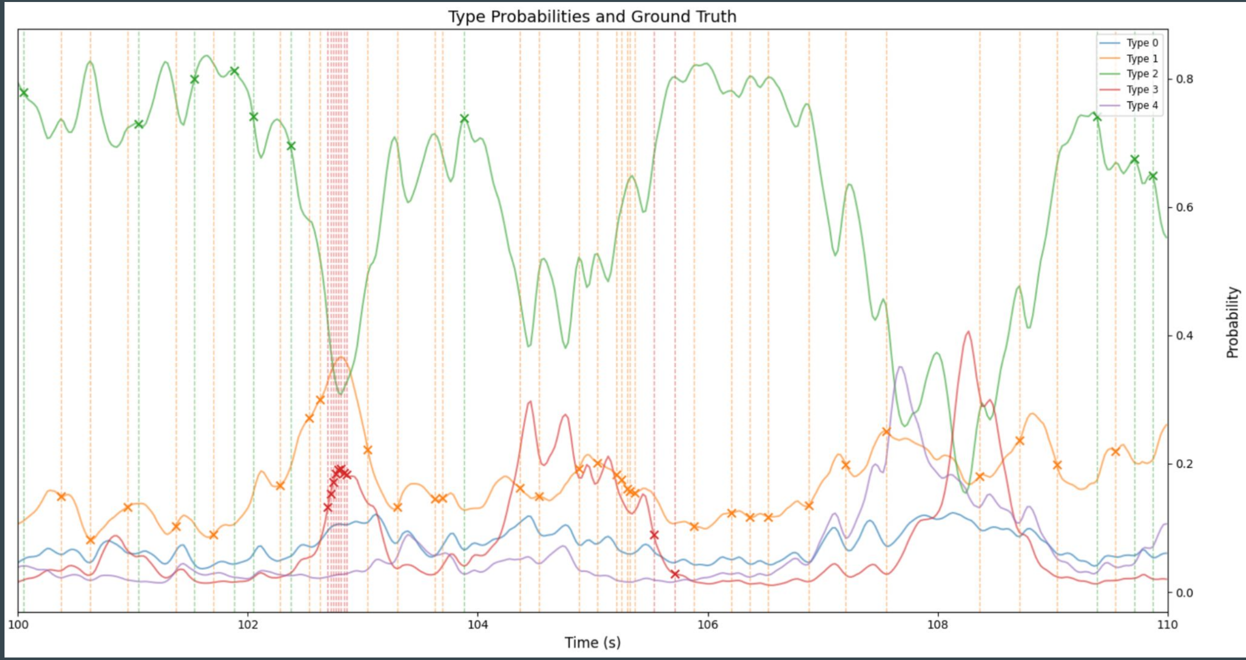
- Threshold selection impacts chart quality.
- Initial improvements showed potential
- Poor Prediction result



BiLSTM Model: Note Type Prediction

Type 0 (no note) dominates (~90%)

Model struggled to learn effectively



Threshold Selection Result

(1) Average threshold:

Inconsistent across songs

Difficulty 16: Best Threshold = 0.12, F1 Score = 0.3470
Difficulty 15: Best Threshold = 0.16, F1 Score = 0.3738
Difficulty 13: Best Threshold = 0.18, F1 Score = 0.3231
Difficulty 11: Best Threshold = 0.10, F1 Score = 0.0899
Difficulty 10: Best Threshold = 0.10, F1 Score = 0.1464
Difficulty 9: Best Threshold = 0.14, F1 Score = 0.2978

(2) Top 'n' Selection:

Fluctuating predictions

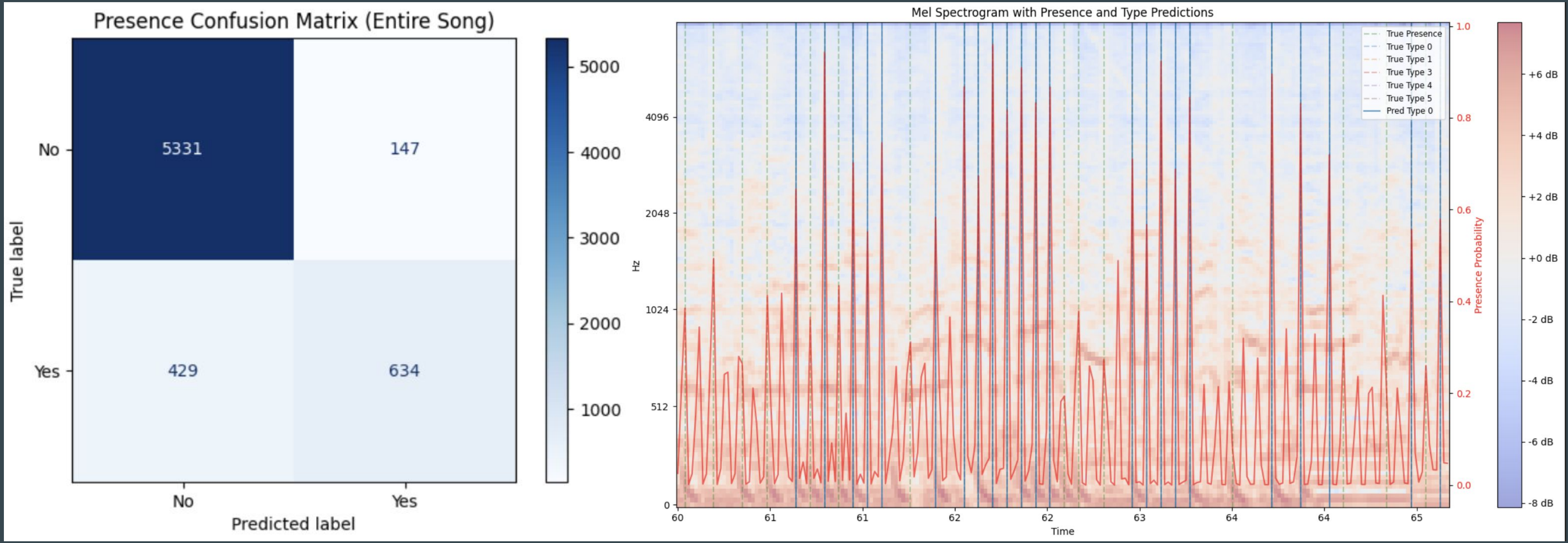
Difficulty	Threshold	Pred. Count	Target Count
16	0.27	1536	1533.60
15	0.26	1150	1145.58

Difficulty	Accuracy	Precision	Recall	F1
16	0.8036	0.1739	0.0694	0.0992
15	0.8271	0.1616	0.0376	0.0610

CNN Model: Presence Prediction

Higher F1 score

Wider prediction range (0-1)

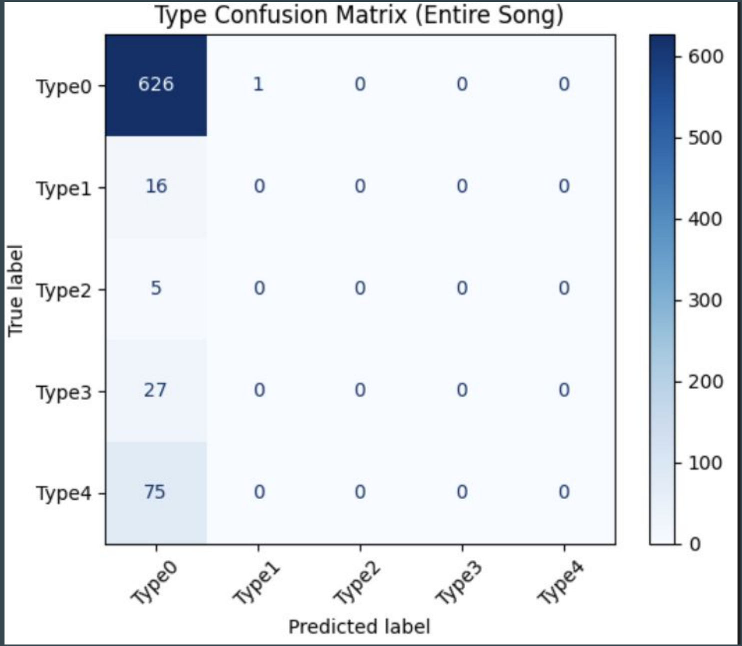


CNN Model: Type Prediction

Class imbalance issues:

- Majority classified as type 0
- Appear good but misleading

	Accuracy	Precision	Recall	F1
Note Presence	0.9039	0.8628	0.5166	0.6463
Note Type	0.9219	0.8613	0.9219	0.8906



Compared with Baseline

Baseline in DDC[3]

Table 2. Results for step placement experiments

Model	Dataset	PPL	AUC	F-score^c	F-score^m
LogReg	Fraxtil	1.205	0.601	0.609	0.667
MLP	Fraxtil	1.097	0.659	0.665	0.726
CNN	Fraxtil	1.082	0.671	0.678	0.750
C-LSTM	Fraxtil	**1.070**	**0.682**	**0.681**	**0.756**
LogReg	ITG	1.123	0.599	0.634	0.652
MLP	ITG	1.090	0.637	0.671	0.704
CNN	ITG	1.083	0.677	0.689	0.719
C-LSTM	ITG	**1.072**	**0.680**	**0.697**	**0.721**

Our Model

	Accuracy	Precision	Recall	F1
Note Presence	0.9039	0.8628	0.5166	0.6463
Note Type	0.9219	0.8613	0.9219	0.8906

[3] D. Donahue, K. Simonyan, A. Zisserman, and G. Vondrick, "Dance Dance Convolution: Learning Generative Models for Dance," Proc. IEEE Int. Conf. on Computer Vision, 2017, pp. 1-9.

Conclusion and Future Work

Conclusion

Current Status: CNN model for presence and type task shows promising results.

A Conjecture why CNN performs better:

- Relatively Small dataset.
- Since it is community based, the charts are diverse and the pattern is less obvious. (A holistic view seems conservative but effective here)

Key Takeaways:

- Extensive trial and error process
- Gained insights into ML and rhythm games

Future Work

1. Collect a dataset with better chart quality.
2. Develop a model to predict the note position that the trajectory are more made human-like.
3. Improve the Note Type prediction quality by applying certain constraints.

Q & A

Reference

[1] Cytoid. https://cytoid.io. [Accessed: Nov. 6, 2024]

[2] Cytoid Community, "Cytoid Gameplay," Google Sites. [Online]. Available: https://sites.google.com/site/cytoidcommunity/gameplay/cytoid-gameplay?authuser=0. [Accessed: Dec. 9, 2024].

[3] D. Donahue, K. Simonyan, A. Zisserman, and G. Vondrick, “Dance Dance Convolution: Learning Generative Models for Dance,” Proc. IEEE Int. Conf. on Computer Vision, 2017, pp. 1-9.

[4] S. Siami-Namini, N. Tavakoli, and A. S. Namin, ”The performance of LSTM and BiLSTM in forecasting time series,” in Proc. 2019 IEEE Int. Conf. on Big Data (Big Data), pp. 3285–3292, 2019

[5] A. Takada, D. Yamazaki, Y. Yoshida, N. Ganbat, T. Shimotomai, N.Hamada, L. Liu, T. Yamamoto, and D. Sakurai, ”Gen´elive! Generating Rhythm actions in Love Live!,” in Proc. AAAI Conf. on Artificial Intelligence, vol. 37, no. 4, pp. 5266–5275, 2023.

[6] C. MX, "Mug-Diffusion: High-quality and controllable charting AI for rhythm games," GitHub repository, 2024. [Online]. Available: https://github.com/Keytoyze/Mug-Diffusion. [Accessed: Dec. 8, 2024].

[7] Lin, T. "Focal Loss for Dense Object Detection." arXiv preprint arXiv:1708.02002 (2017).