Master's Thesis Presentation for the 141st Dissertation Evaluation

A Study on Effective Music Skip Prediction based on Information Fusion and Separate Convolution between Track and Session Features

트랙과 세션 특징 간 분리된 컨볼루션을 통한 정보 융합 기반 효과적인 음악 건너뛰기 예측 연구

Sanghyeong Jin

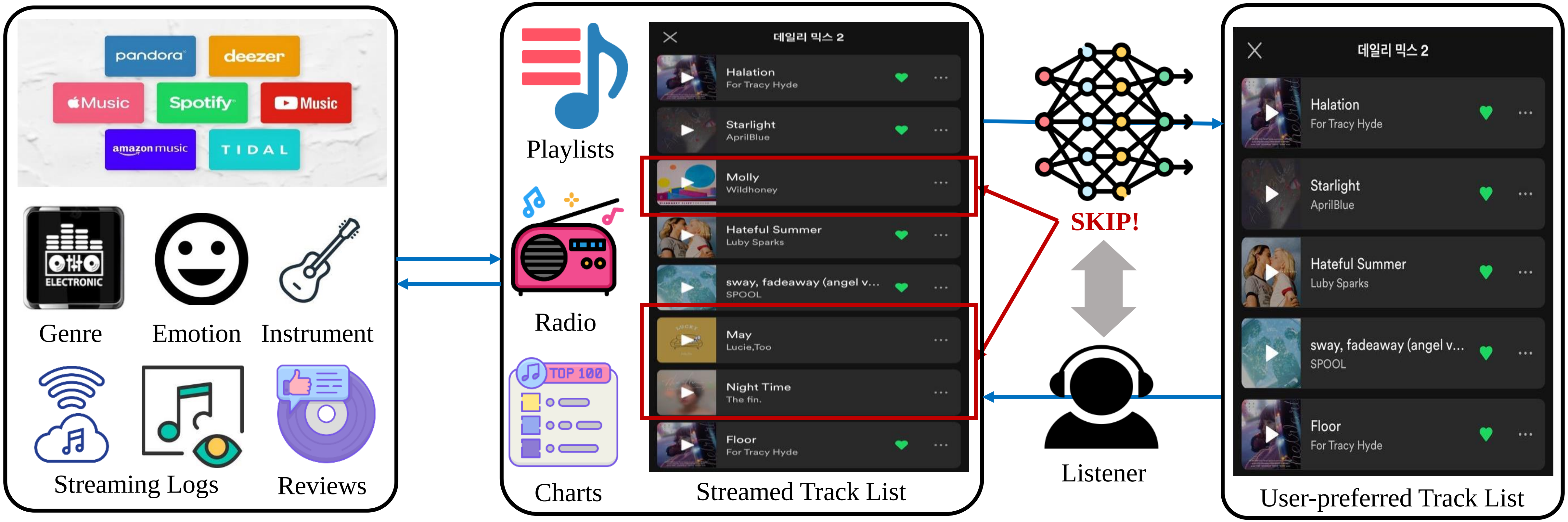
jinsanghyeong@gmail.com

2024. 05. 21.

Introduction

Music skip prediction

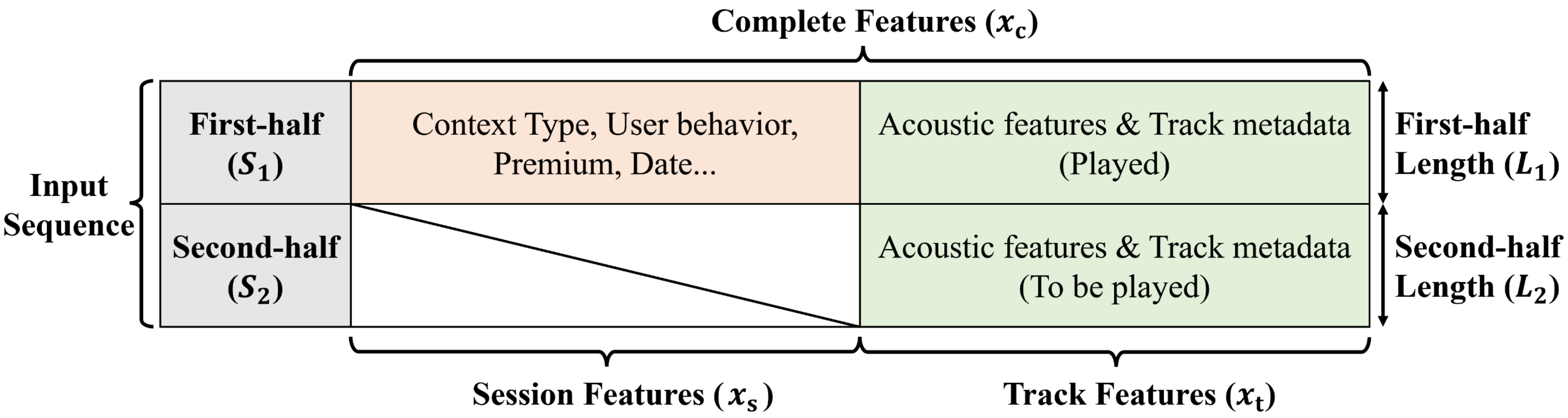
- Binary classification form of session-based recommendation
- Predict **whether a listener will skip upcoming songs or not** during the playlist streaming session
- Enhance listener satisfaction by **creating smooth track transitions** and **tuning of the playlist**



Related Work

Music skip prediction based on Track and Session features

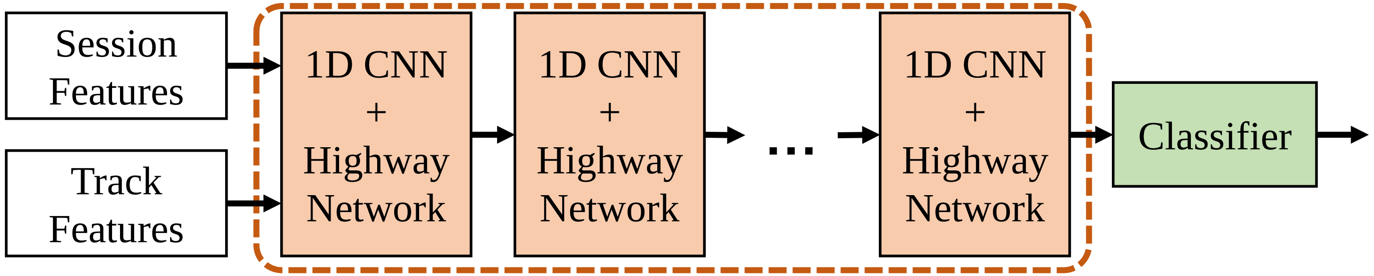
- **Track features** contain acoustic features of music obtained from given playlists
- **Session features** are logged when listener streams the playlist
- **Input sequence** consists of first-half and second-half sequence
- **First-half sequence** contains both session and track features
- **Second-half sequence** consists of only track features to be played without session information



Related Work

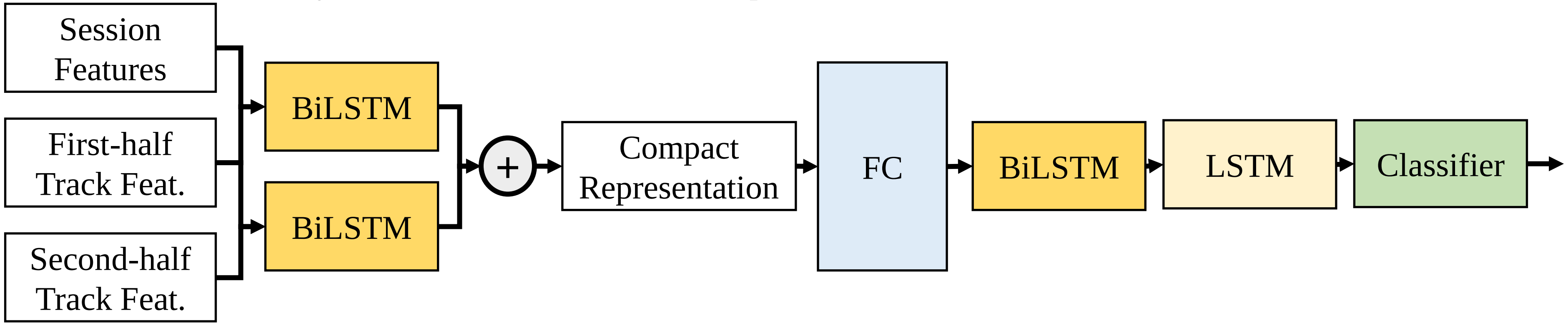
1D CNN-Highway [1]

- Combining 1D CNN and Highway Network
- 1D CNN captures sequential transition and highway network add a shortcut based on the gating function



LSTM Encoder-Decoder [2]

- Encoder-Decoder style architecture based on LSTM and BiLSTM
- Compact representation is obtained by concatenation of the session representation and track features

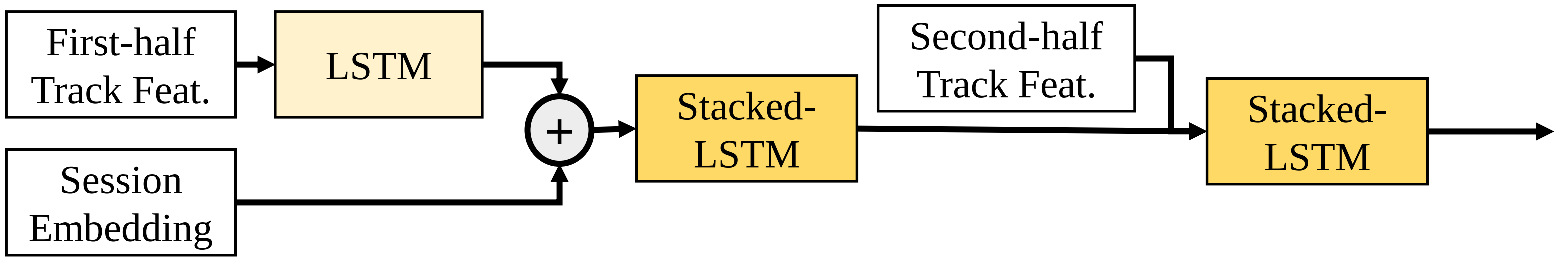


[1] Chang, S., Lee, S., Lee, K., 2019. Sequential skip prediction with few-shot in streamed music contents. In: Proceedings of the 12th ACM International Conference on Web Search and Data Mining Cup Workshop
[2] Adapa, S., 2019. Sequential modeling of sessions using recurrent neural networks for skip prediction. In: Proceedings of the 12th ACM International Conference on Web Search and Data Mining Cup Workshop

Related Work

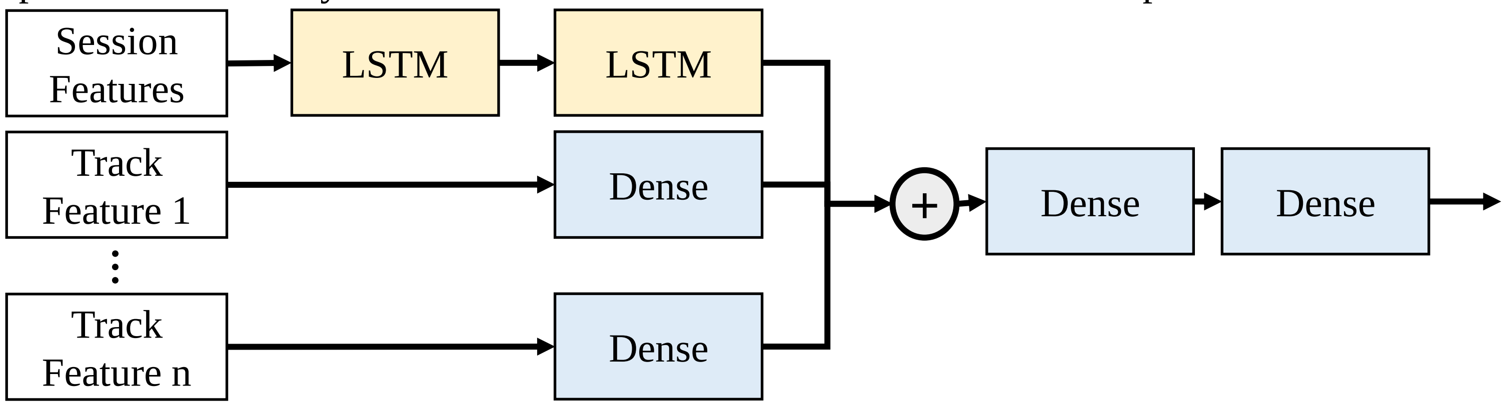
Multi-RNN [3]

- Multi-RNN stacks 2 LSTM layers and First stacked-LSTM encodes the session and track feature combination
- Second stacked-LSTM predicts track skipping based on first stacked-LSTM



RNN-Dense [4]

- Combining LSTM and track representing Dense layers
- Concatenates the output of the Dense layers with LSTM-encoded session features for final prediction

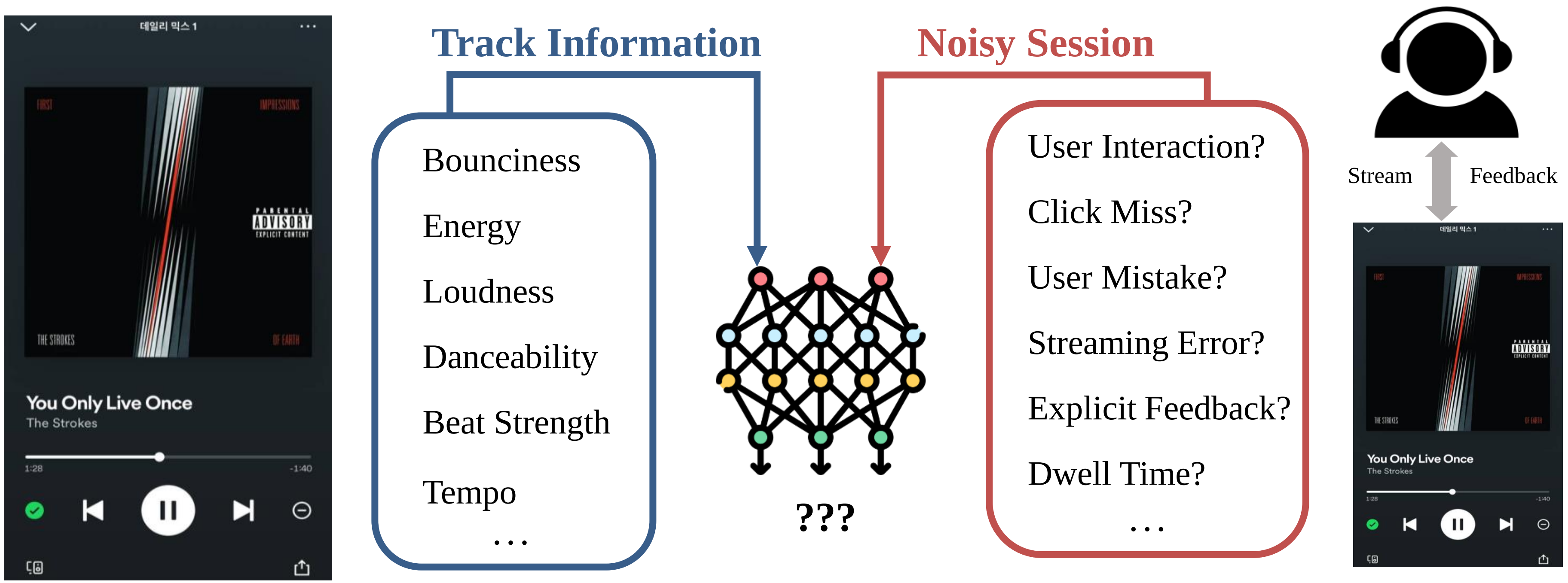


[3] Hansen, C., et al., 2019. Modelling sequential music track skips using a multi-rnn approach. In: Proceedings of the 12th ACM International Conference on Web Search and Data Mining Cup Workshop
[4] Jeunen, O., Goethals, B., 2019. Predicting sequential user behaviour with session-based recurrent neural networks. In: Proceedings of the 12th ACM International Conference on Web Search and Data Mining Cup Workshop

Proposed Method

Common Drawbacks

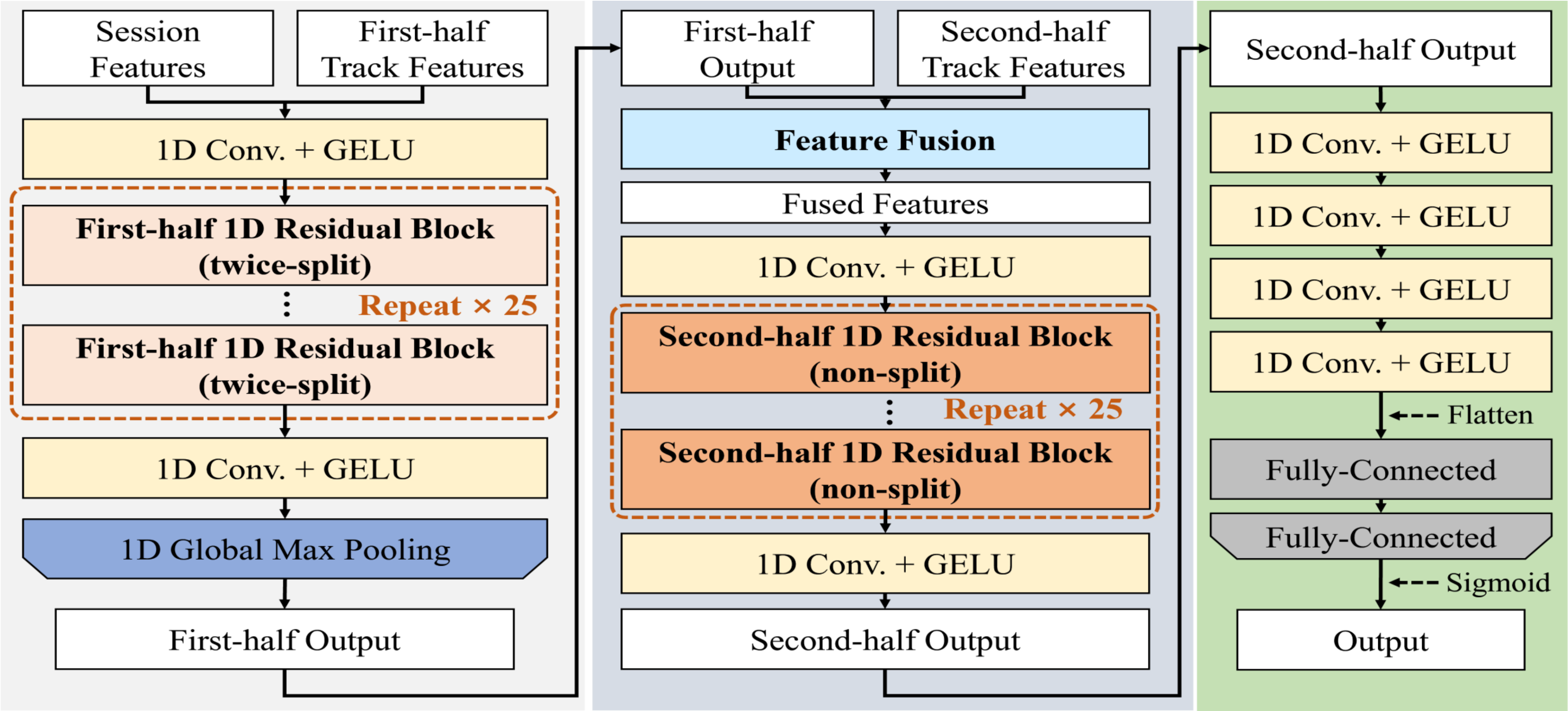
- Using listener interaction features in session logs is essential in session-based recommendation task
- **Session features** may contain considerable noises compared to the track features, and it degrade prediction performance
- **Track features** can be more reliable than session features because the acoustic information hard to be changed
- Music skip prediction model **should utilize track and session features while reducing session noise** from being fused to track features



Proposed Method

Overall of Proposed Method

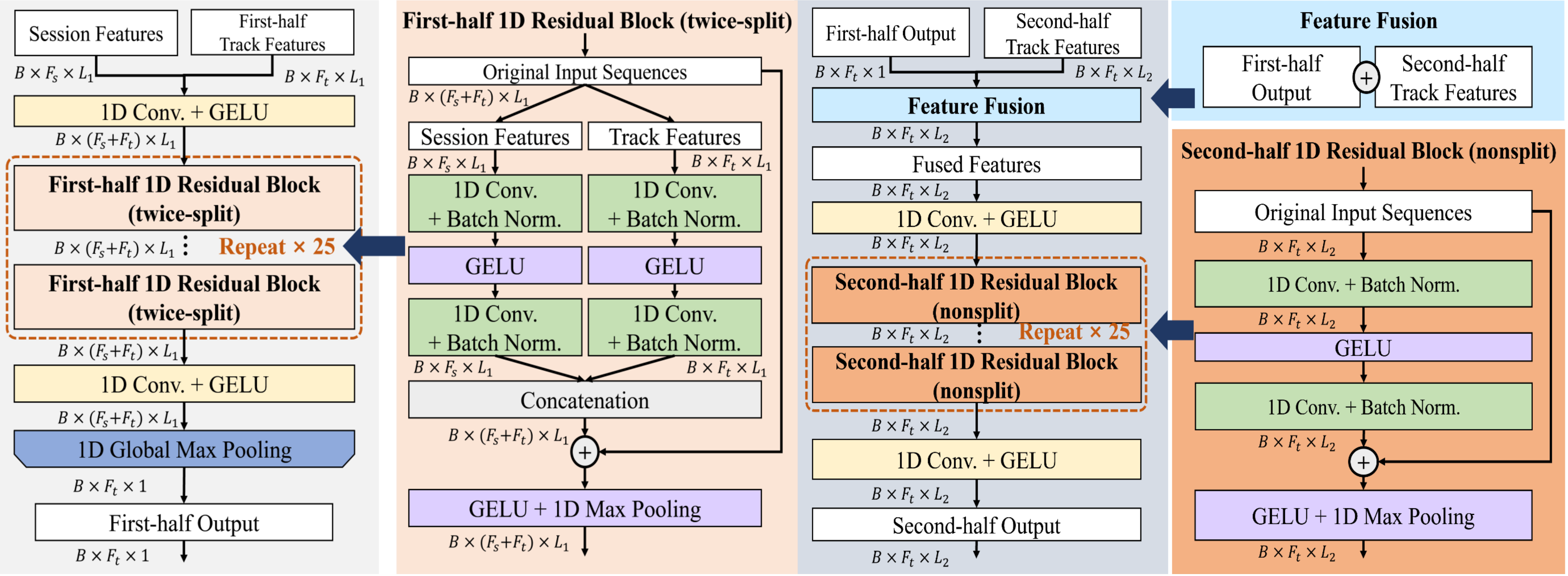
- Consists of First-half network / Second-half (fusion) network / Classifier



Proposed Method

Difference of First-half & Second-half network

- First-half : feature extraction with twice-split 1D residual block
- Second-half : feature fusion for reflecting first-half information and then capture sequential patterns



Experimental Settings

Experimental Settings

- Datasets
  - Music Streaming Sessions (MSS) Dataset
  - Number of Classes: 2 (Skip / Not Skip)
  - Input : (Batch, 74, 20) - (Session & Track Features, Sequence Length)
    → Sequence that less than 20 includes zero padding
  - Output : (Batch, 10) - Binary classification results for the 10 tracks in second half
- Cross-Validation for Test Performance
  - 30 iterations / Data Split Ratio : (Train : Test) = (0.9: 0.1) with random split
- Hyper Parameter Setting:

Model	**Batch Size**	**Loss Function**	**Optimizer**
Proposed	1024	Binary Cross Entropy	Adam
1D CNN-Highway [1]			
LSTM Encoder-Decoder [2]			RMSprop
Multi-RNN [3]			Adam
RNN-Dense [4]			

Experimental Settings

MAA (Mean Average Accuracy)

- D : Test set, n : number of input sequences in D
- $X_i = \{x_{i,1}, x_{i,2}, x_{i,3}, \ldots, x_{i,|x_i|}\}$: Input sequence of length $|X_i| \le 20$
- AA(X_i) : Average Accuracy (AA) of X_i input sequence
- B($x_{i,j}$) : Boolean indicator that whether j-th prediction of X_i sequence was correct (0, 1)

$$MAA(D) = \frac{\sum_{i=1}^{n} AA(X_i)}{n} \qquad AA(X_i) = \frac{\sum_{j=11}^{|X_i|} A(x_{i,j}) \cdot B(x_{i,j})}{|X_i| - 10} \qquad A(x_i, j) = \sum_{k=1}^{j} \frac{B(x_{i,k})}{j}$$

$$\frac{\left(\frac{1}{1} * 1 + \frac{2}{2} * 1 + \frac{3}{3} * 1 + \frac{4}{4} * 1 + \frac{4}{5} * 0\right)}{5} = 0.8000$$

$$\frac{\left(\frac{1}{1} * 1 + \frac{2}{2} * 1 + \frac{2}{3} * 0 + \frac{3}{4} * 1 + \frac{4}{5} * 1\right)}{5} = 0.7100$$

$$\frac{\left(\frac{0}{1} * 0 + \frac{1}{2} * 1 + \frac{2}{3} * 1 + \frac{3}{4} * 1 + \frac{4}{5} * 1\right)}{5} = 0.5433$$

Ground Truth	Second-half Prediction	AA	MAA
1, 1, 1, 1, 1	1, 1, 1, 1, 0	0.8000	
1, 1, 1, 1, 1	1, 1, 1, 0, 1	0.7600	
1, 1, 1, 1, 1	1, 1, 0, 1, 1	0.7100	0.6913
1, 1, 1, 1, 1	1, 0, 1, 1, 1	0.6433	
1, 1, 1, 1, 1	0, 1, 1, 1, 1	0.5433	

Experimental Results

Student *t*-test Result

Model	MAA Mean ± Standard Deviation	Average Rank
Proposed	**0.6713 ± 0.0178**	**1.48**
1D CNN-Highway [1]	0.5976 ± 0.0885 ▼	3.13
LSTM Encoder-Decoder [2]	0.5357 ± 0.0912 ▼	3.87
Multi-RNN [3]	0.6059 ± 0.0583 ▼	3.03
RNN-Dense [4]	0.5820 ± 0.0810 ▼	3.48

▼ indicates corresponding model is significantly worse than proposed method at significance level α=0.05

Wilcoxon signed-rank test Result

	Comparison methods			
	1D CNN-Highway	LSTM Encoder-Decoder	Multi-RNN	RNN-Dense
Proposed ***versus*** (*p*-value)	**Win** (5.974e-06)	**Win** (5.588e-09)	**Win** (2.349e-06)	**Win** (7.252e-05)

significance level α=0.05

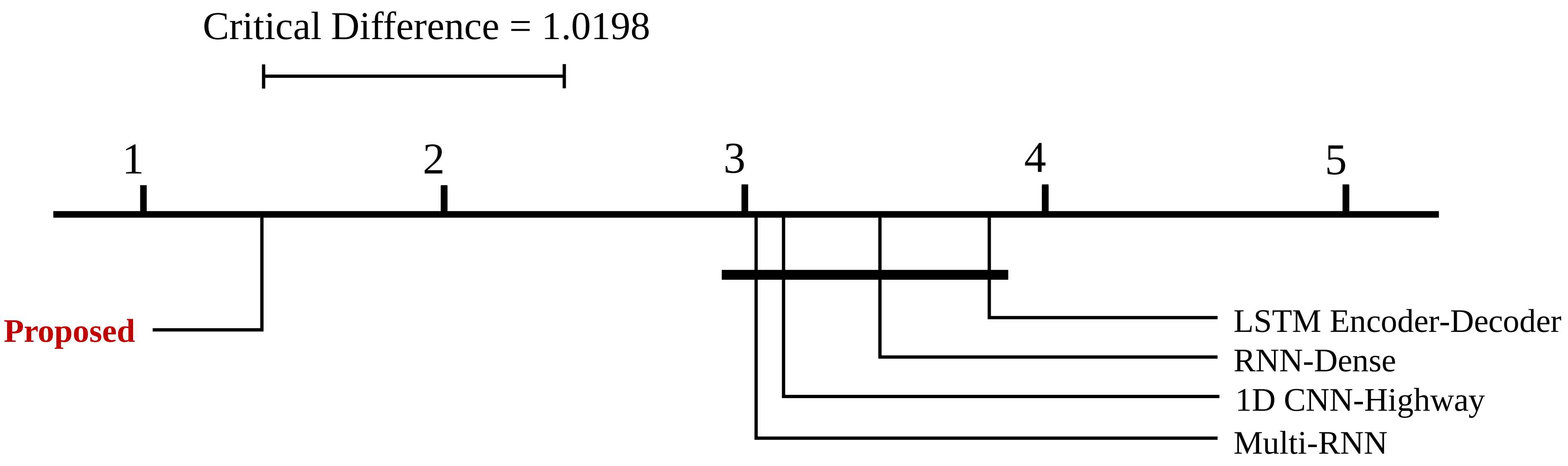
Experimental Results

Friedman test Result

Evaluation Measure	Friedman Statistics	Critical Value
MAA	39.713	2.534

significance level α=0.05

Bonferroni-Dunn test Result



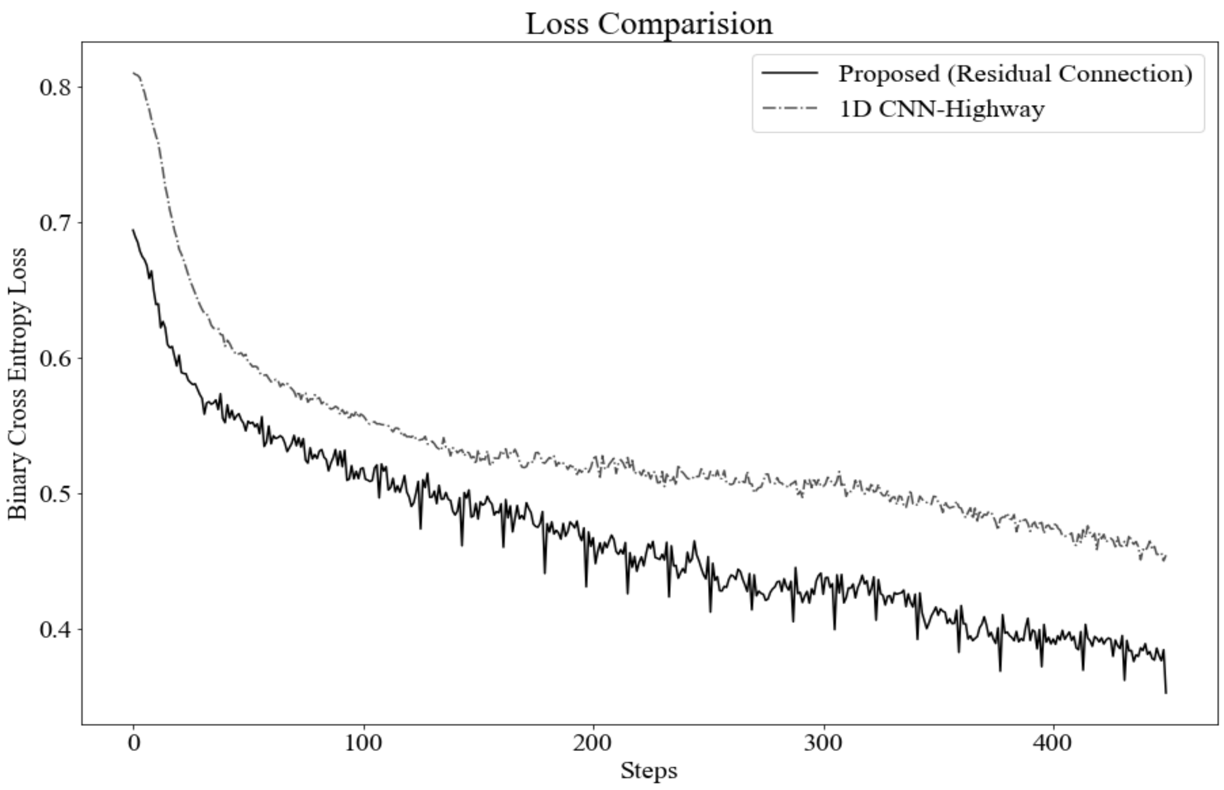
In-depth Analysis

Comparison with Traditional approaches and 1D CNN-Highway

- Comparing the proposed method to the *k*NN-based recommendation methods for validating the superiority of recent Neural Network models
- Comparing the training loss of the proposed method and 1D-CNN Highway to determine the effectiveness of the residual connection
- Rapid convergence of the proposed method through residual connections has a more promising performance

Model	MAA Mean ± Standard Deviation
Proposed	**0.6713 ± 0.0178**
User-*k*NN	0.5102 ± 0.0059 ▼
Item-*k*NN	0.5205 ± 0.0015 ▼

▼ indicates corresponding model is significantly worse than proposed method at significance level α=0.05

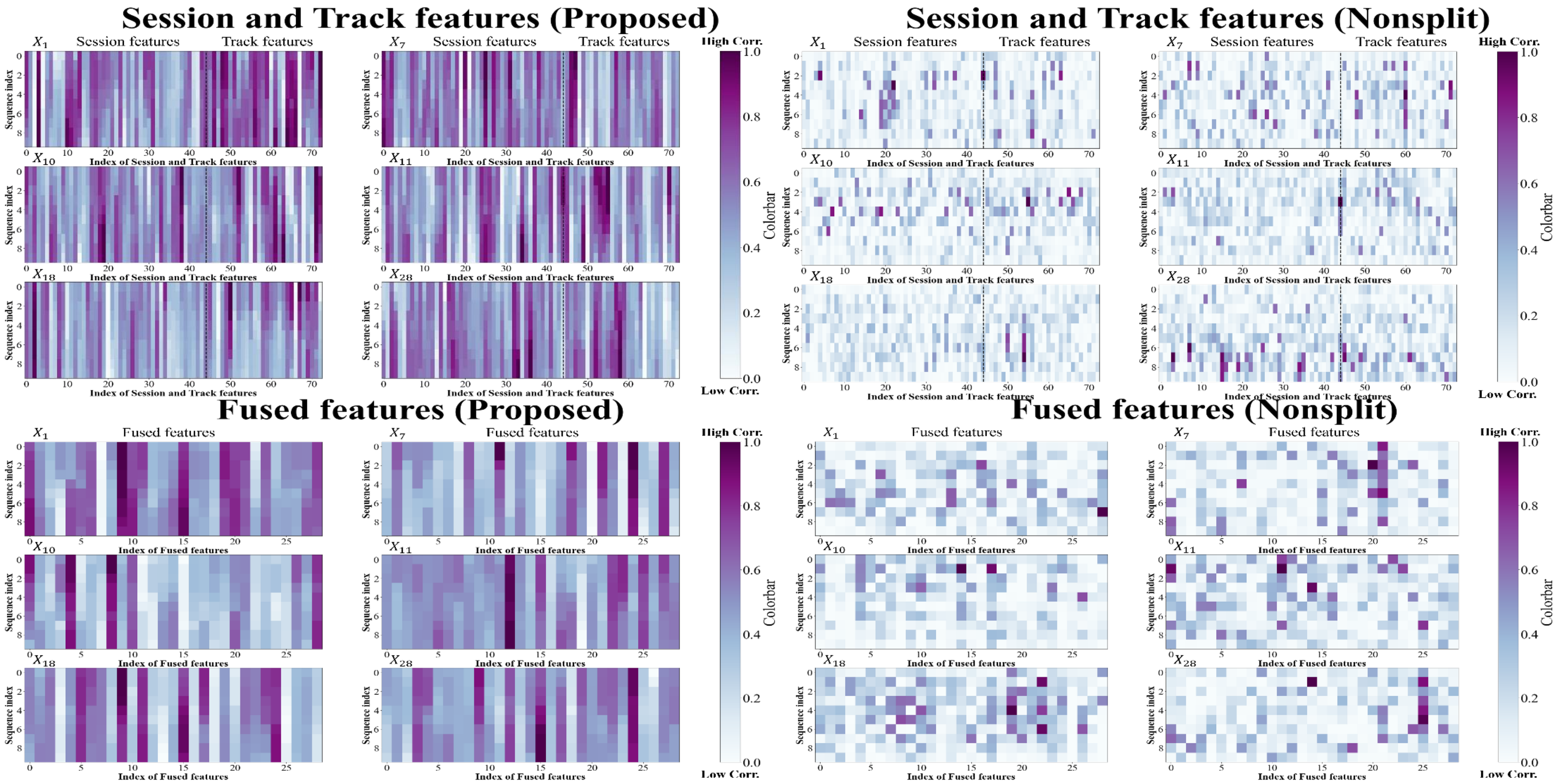


In-depth Analysis

Feature maps of Proposed method and Nonsplit method

- The proposed method shows high emphasis, while the nonsplit model indicates less feature emphasis

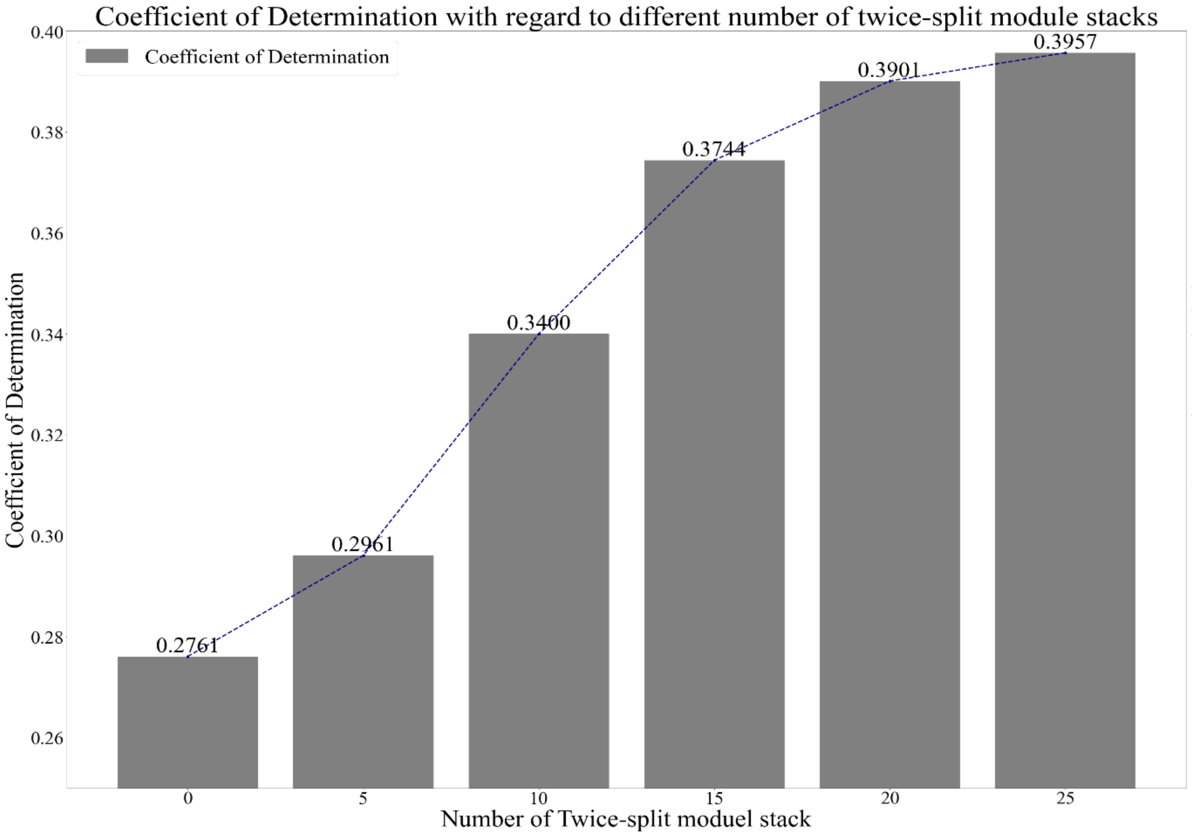
- Nonsplit method weakening the first-half information reflected in feature fusion and resulting in low performance

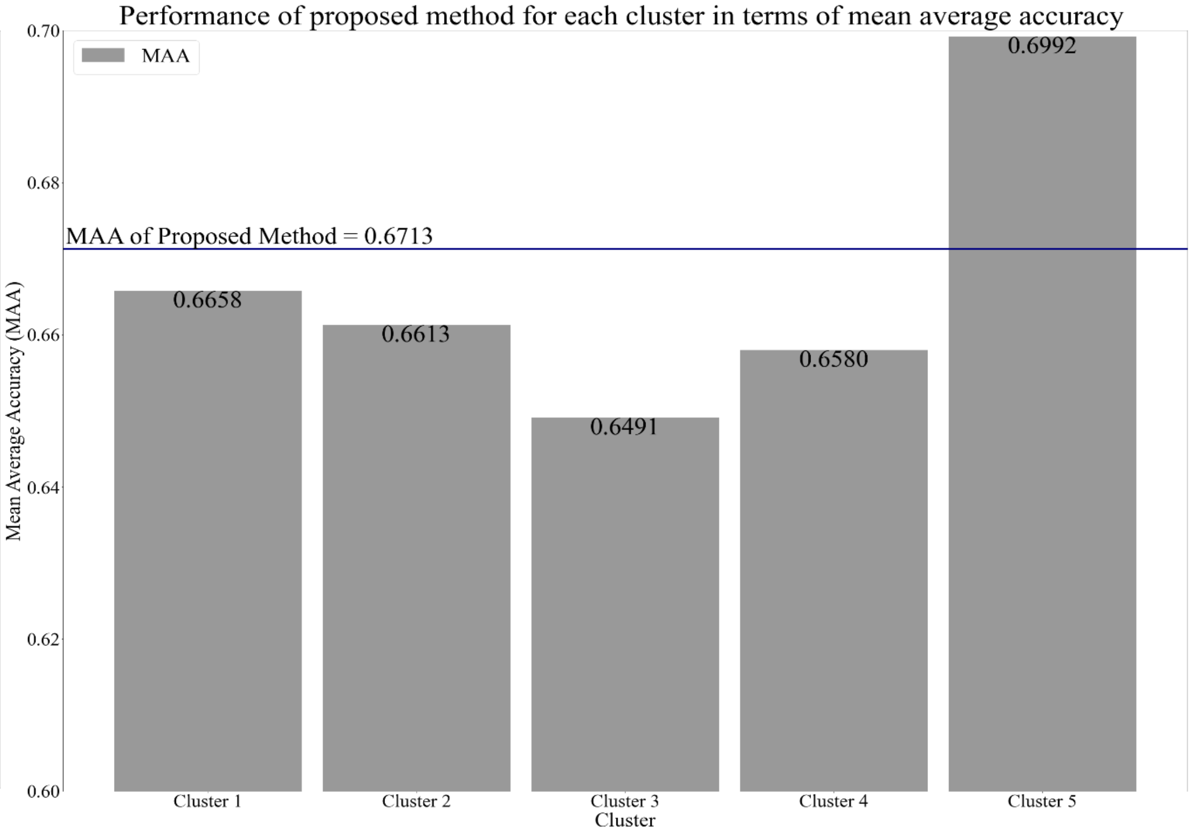


In-depth Analysis

Quantitative result of noise reduction and Robustness in different subgroups of music

- Coefficient of determination of each regression model is compared to analyze the reduction degree of the session noise
- Applying *k*-means clustering to the track features, and MAA of each cluster is compared to analyze the performance for different subgroups





In-depth Analysis

Generalizability of the proposed method

- Conduct additional experiment on the KKBOX dataset
- The proposed method outperforms other methods, similar to MSS dataset results

Model	MAA Mean ± Standard Deviation
Proposed	**0.5560 ± 0.0179**
1D CNN-Highway	0.5106 ± 0.0106 ▼
LSTM Encoder-Decoder	0.5123 ± 0.0127 ▼
Multi-RNN	0.5077 ± 0.0111 ▼
RNN-Dense	0.5068 ± 0.0078 ▼

▼ indicates corresponding model is significantly worse than proposed method at significance level α=0.05

Conclusion

Summary

1. Proposed twice-split 1D residual module to reduce the influences of noisy session information

2. Proposed method significantly outperforms than conventional music skip prediction methods

3. The twice-split module can capture high-correlated information and effectively reflect first-half information to the second-half track features without being influenced by session noise

Achievement

- Effective Music Skip Prediction based on Late Fusion Architecture for User-Interaction Noise, ***Expert Systems with Applications***, 2023
- Employing Multi-branch Convolution Structure in 1D ResNet for Music Skip Prediction, ***ICEIC***, 2023

Thank you

References

[1] Chang, S., Lee, S., Lee, K., 2019. Sequential skip prediction with few-shot in streamed music contents. In: Proceedings of the 12th ACM International Conference on Web Search and Data Mining Cup Workshop

[2] Adapa, S., 2019. Sequential modeling of sessions using recurrent neural networks for skip prediction. In: Proceedings of the 12th ACM International Conference on Web Search and Data Mining Cup Workshop

[3] Hansen, C., et al., 2019. Modelling sequential music track skips using a multi-rnn approach. In: Proceedings of the 12th ACM International Conference on Web Search and Data Mining Cup Workshop

[4] Jeunen, O., Goethals, B., 2019. Predicting sequential user behaviour with session-based recurrent neural networks. In: Proceedings of the 12th ACM International Conference on Web Search and Data Mining Cup Workshop

Appendix

Appendix – Session & Track Features

Summary of Session Features

Feature Name	Description
Session ID	Unique ID for the playlist session.
Session Position	Position of the played track within the session. (1-20)
Session Length	The length of the session. (11-20)
Track ID	Unique ID for the played track.
Context Switch	Boolean indicator of the changed context between previous track and current track.
No Pause Before Play	Boolean indicator of the pausing before playing the current track.
Short Pause Before Play	Boolean indicator of the short pausing before playing the current track.
Long Pause Before Play	Boolean indicator of the long pausing before playing the current track.
Hist User Behavior Seekfwd	The number of seek forward times within the track.
Hist User Behavior Seekback	The number of seek back times within the track.
Hist User Behavior Shuffle	Boolean indicator of shuffle.
Hour of Day	The hour of day. (0-23)
Date	YYYY-MM-DD formatted date.
Premium	Boolean indicator of premium account.
Context Type	Type of context the playback occurred.
Hist User Behavior Reason Start	Reason for track starting.
Hist User Behavior Reason End	Reason for track ending.

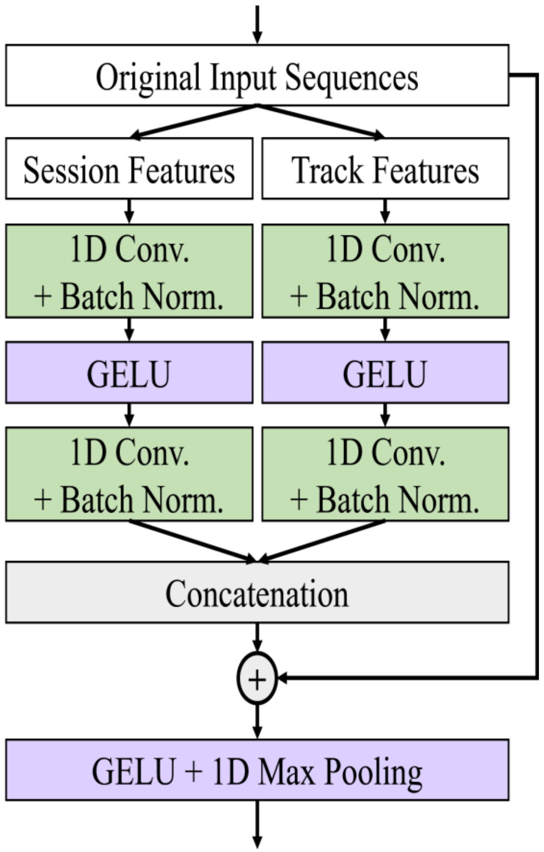
Summary of Track Features

Type	Feature Name
Acoustic Features	Acousticness, Beat Strength, Bounciness, Danceability, Dynamic Range Mean, Energy, Flatness, Instrumentalness, Key, Liveness, Loudness, Mechanism, Mode, Organism, Speechiness, Tempo, Time Signature, Valence, Acoustic Vectors
Track Metadata	Track ID, Duration, Release Year, US Popularity

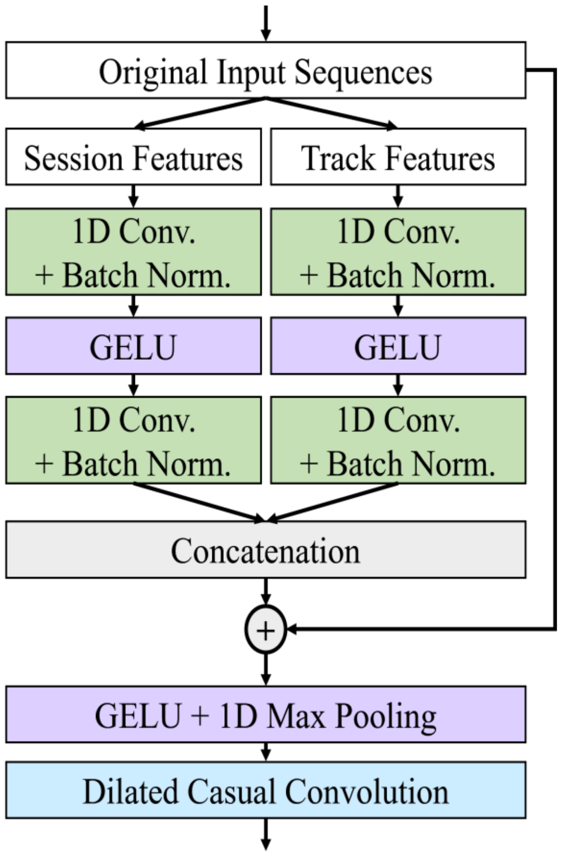
Appendix – Variations of 1D Residual Block

* DCC : Dilated Casual Convolution

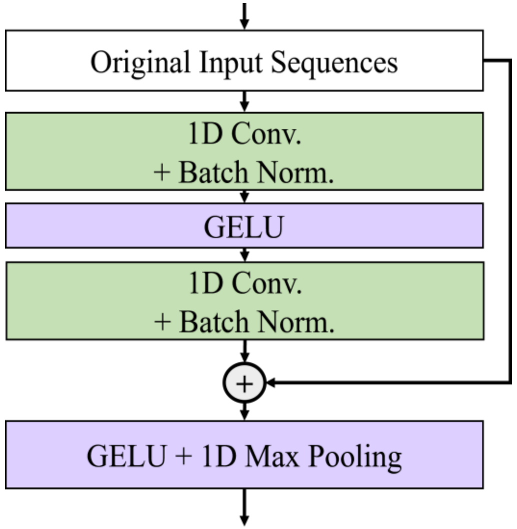
1. Proposed



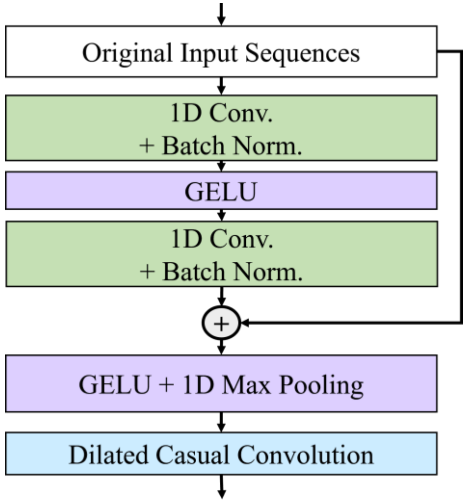
2. Proposed + DCC



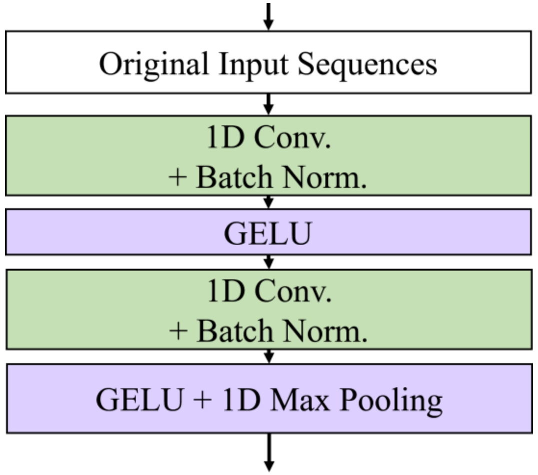
3. Non-split 1D Residual



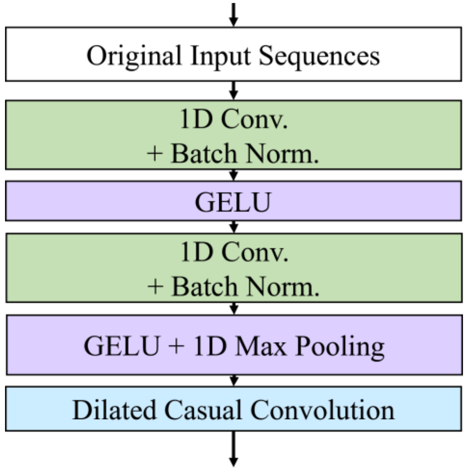
4. Non-split 1D Residual + DCC



5. w/o Residual



5. w/o Residual + DCC



Appendix – Variations of 1D Residual Block

Student *t*-test Result

Model	MAA Mean ± Standard Deviation	Average Rank
Proposed	**0.6713 ± 0.0179**	**1.43**
Proposed + DCC	0.6629 ± 0.0380	1.56
Non-split 1D Residual	0.5445 ± 0.0070 ▼	3.5
Non-split 1D Residual + DCC	0.5399 ± 0.0241 ▼	3.73
w/o Residual	0.4616 ± 0.0431 ▼	5.47
w/o Residual + DCC	0.4615 ± 0.0434 ▼	5.3

▼ indicates corresponding model is significantly worse than proposed method at significance level α=0.05
* DCC : Dilated Casual Convolution

Appendix – Variations of 1D Residual Block

Friedman test Result

Evaluation Measure	Friedman Statistics	Critical Value
MAA	130.038	2.421

significance level α=0.05

Wilcoxon signed-rank test Result

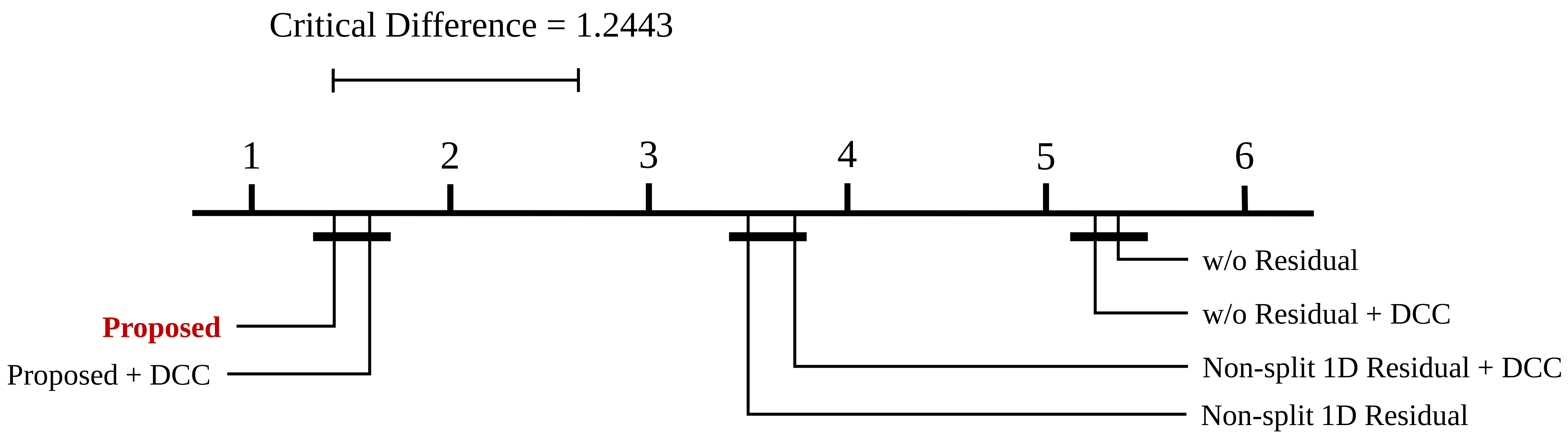
* DCC : Dilated Casual Convolution

	Comparison methods				
	Proposed + DCC	Non-split 1D Residual	Non-split 1D Residual + DCC	w/o Residual	w/o Residual + DCC
Proposed ***versus*** (*p*-value)	Tie (0.3285)	**Win** (1.86e-09)	**Win** (1.86e-09)	**Win** (1.86e-09)	**Win** (1.86e-09)

significance level α=0.05

Appendix – Variations of 1D Residual Block

Bonferroni-Dunn test Result



* DCC : Dilated Casual Convolution