Genre Classification using Lyrics and Audio Features

GROUP 19:

CHEN WEI, LIAO HUNG CHIEH, LOW XU YAN JESSICA, WANG CHUXIN, ZHU XUEYING

Project Overview

Objective: Genre classification using both lyrics and Spotify audio features

Goal: Compare how different data modalities impact model performance

Methods: Implemented and evaluated multiple models

Dataset Introduction

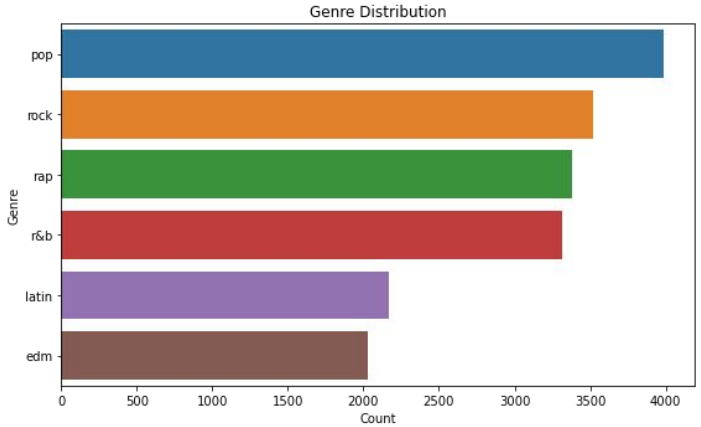
Initial Dataset :

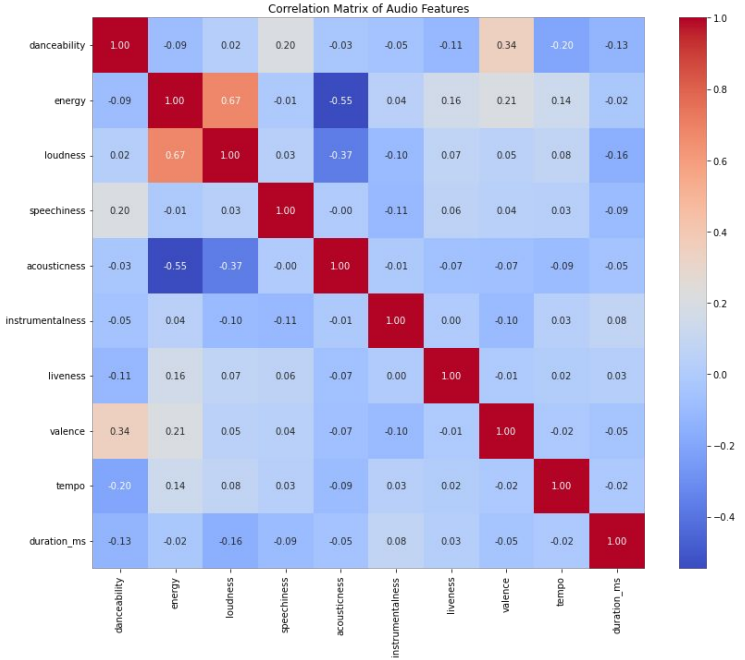
- Kaggle dataset with numerical labels
- Unclear genre mapping
- Imbalanced after filtering lyrics
- Finally discarded

Current Dataset:

- Kaggle (Nakhaee, 2020), 18,454 entries, 25 features
- Target: playlist_genre (6 genres)
- Data types: Lyrics + Audio features (e.g. danceability, energy, valence…)

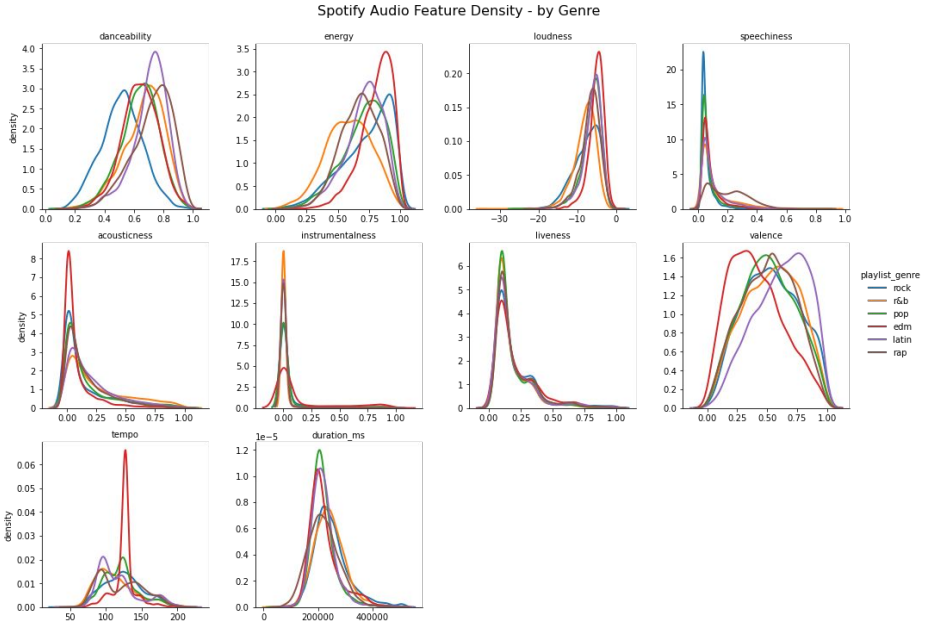
Data Preprocessing & Analysis



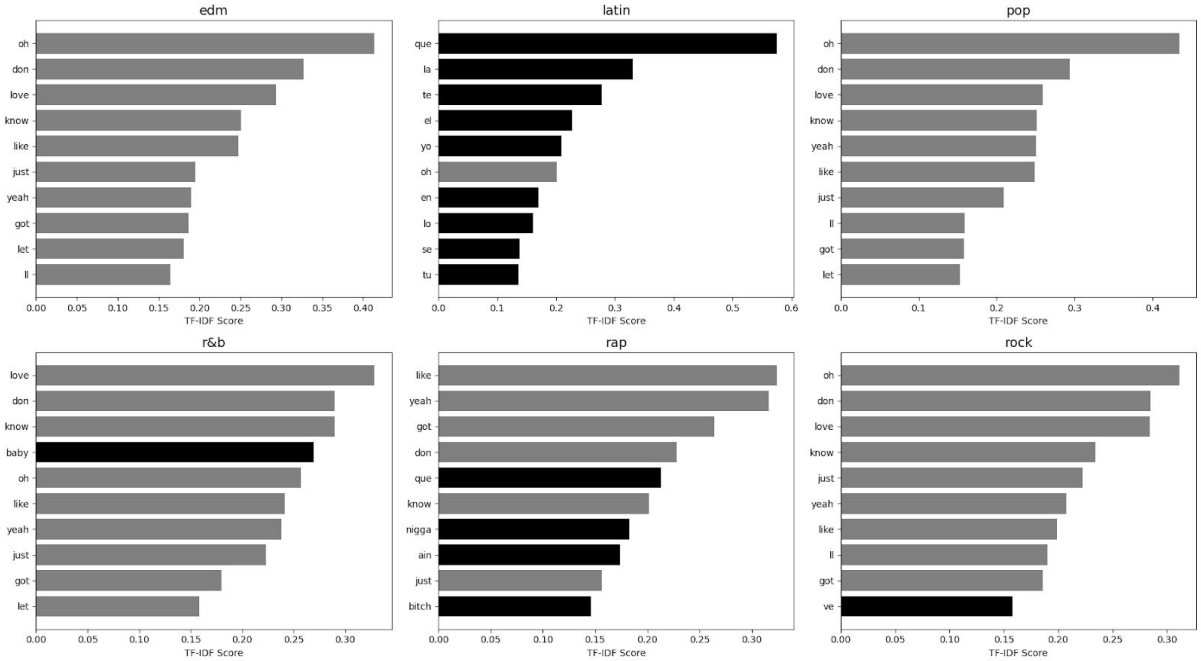


Data Preprocessing & Analysis

Genre	Speechiness	Energy	Danceability	Valence	Tempo / Loudness
Rap	High	Moderate	Low	–	–
EDM	Low	High	High	–	High Tempo
Pop	–	–	High	High	–
Rock	–	High	Low	–	High Loudness
Latin	Low	–	High	High	–
R&B	–	Low	–	Average	Low Loudness



Data Preprocessing & Analysis

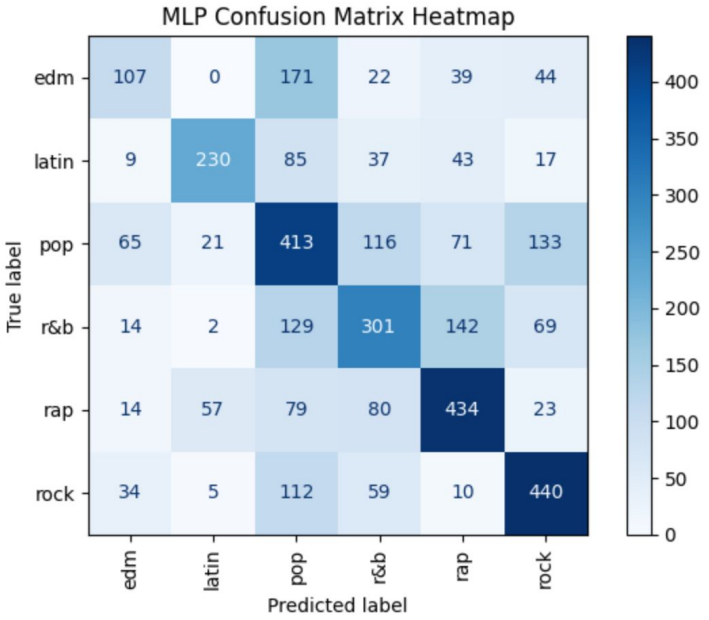


Baseline Model

Baseline 1 - MLP (Audio)

Test Loss: 1.2099
Overall Accuracy: **53.07%**

Genre	Precision	Recall	F1-Score
edm	0.44	**0.28**	**0.34**
latin	0.73	0.55	**0.62**
pop	0.42	0.50	0.46
r&b	0.49	0.46	0.47
rap	0.59	0.63	**0.61**
rock	0.61	0.67	**0.63**



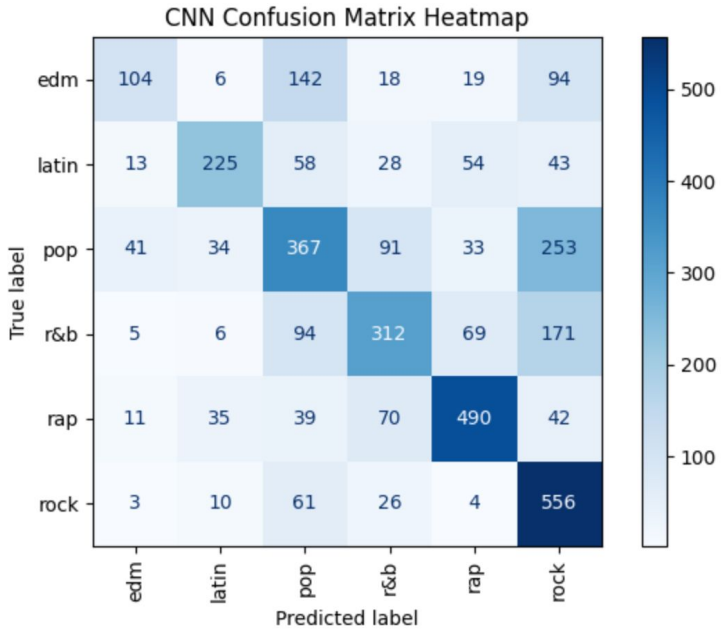
Baseline 2 - CNN (Lyrics)

Layer Block	Layer Type	Input Dim	Output Dim	Activation
Embedding Layer	Embedding	(seq_len,) → (128)	(seq_len, 128)	—
Conv Layer (3-gram)	Conv1d	128	256	—
Conv Layer (4-gram)	Conv1d	128	256	—
Conv Layer (5-gram)	Conv1d	128	256	—
Concat & Flatten	—	3 × 256	768	—
Fully Connected	Linear	768	256	ReLU
Output Layer	Linear	256	6	—

Baseline 2 - CNN (Lyrics)

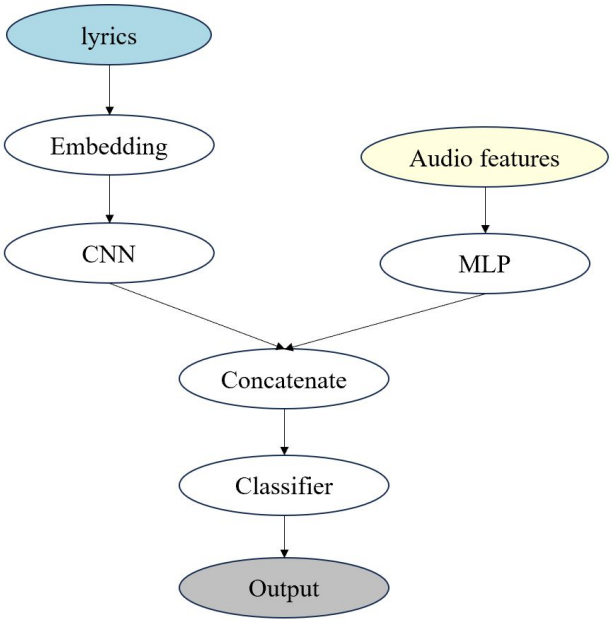
Test Loss: 1.2321
Overall Accuracy: **56.63%**

Genre	Precision	Recall	F1-Score
edm	0.59	0.27	**0.37**
latin	0.71	0.53	0.61
pop	0.48	0.45	0.46
r&b	0.57	0.47	0.52
rap	0.73	0.71	**0.72**
rock	**0.48**	**0.84**	0.61



CNN_MLP

Multimodal CNN + MLP



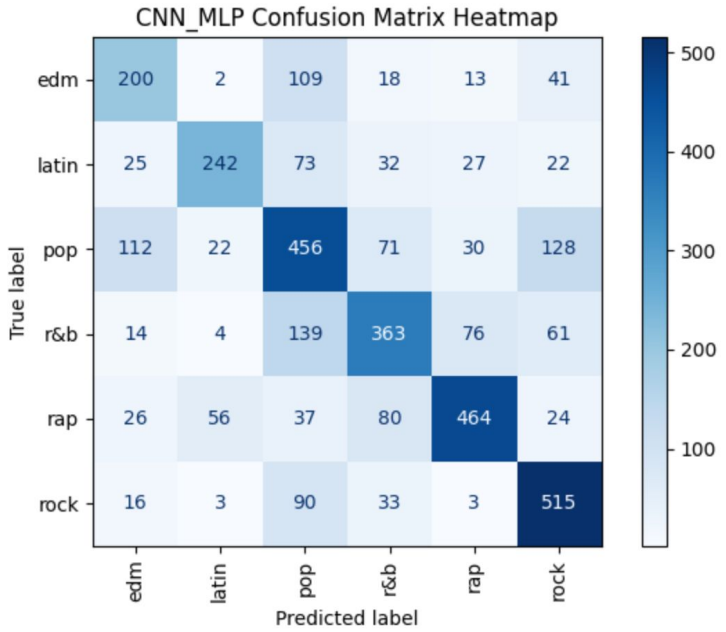
Module	Layer Type	Input Dimension	Output Dimension	Activation
Text Embedding	Embedding	(seq_len,) → 128	(seq_len, 128)	—
CNN (Lyrics)	Conv1d (k=3)	128	128	—
	Conv1d (k=4)	128	128	—
	Conv1d (k=5)	128	128	—
Concat Lyrics Features	—	3 × 128	384	—
MLP (Audio)	Linear (2 layers)	11 → 64→ 32	32	ReLU
Multimodal Fusion	Concatenation	384(lyrics) + 32 (audio)	416	—
Fully Connected	Linear (2 layers)	416→ 128→ 6	6	ReLU

≈ 1.5M Parameters

Multimodal CNN + MLP

Test Loss: 1.0425
Overall Accuracy: **0.6176**

Genre	Precision	Recall	F1-Score
edm	0.51	0.52	**0.52**
latin	0.74	0.57	0.65
pop	0.50	0.56	0.53
r&b	0.61	0.55	0.58
rap	0.76	0.68	**0.71**
rock	0.65	0.78	**0.71**



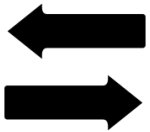
BiLSTM

With Attention Layer

Intuition

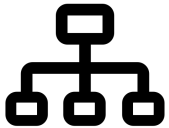
Lyrics are like sentences, genres are classes in classification.

BiDirectional LSTM



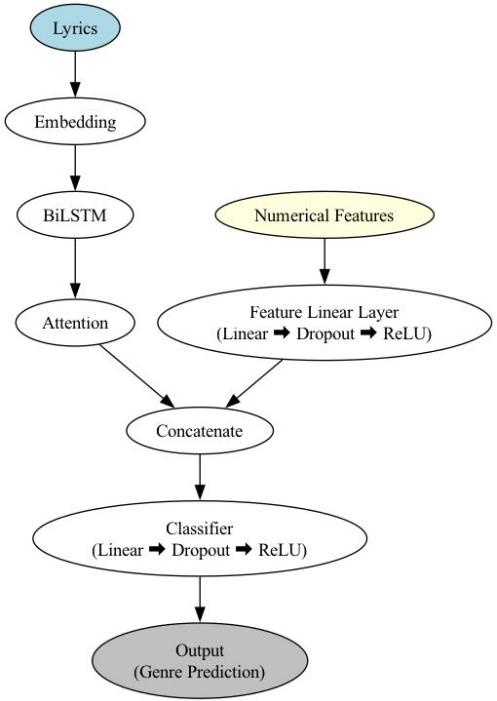
Learn word patterns forward & backward

Attention Layer



Focuses on the most important words

Attention-BiLSTM Structure

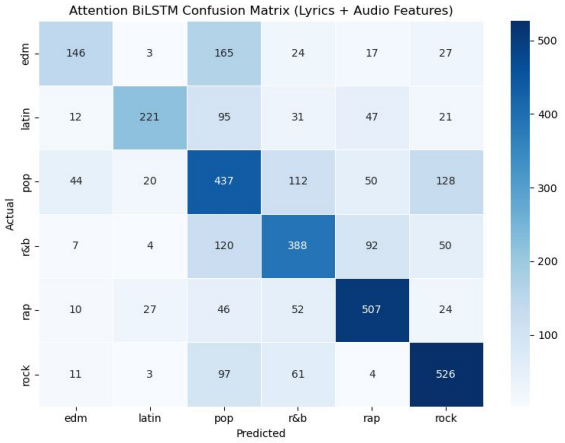


Module	Layer Type	Input Dimension	Output Dimension	Dropout / Activation
Text Embedding	Embedding	(seq_len,) → 128	(seq_len, 128)	—
BiLSTM	LSTM (bidirectional = True)	128	128	Dropout
Attention	Linear	256	256	—
Feature Linear	Linear	11	128	Dropout, ReLU
Concat Lyrics Features	—	(256, 128)	384	—
Classifier	Linear (2 layers)	384 → 128 → 6	6	Dropout, ReLU

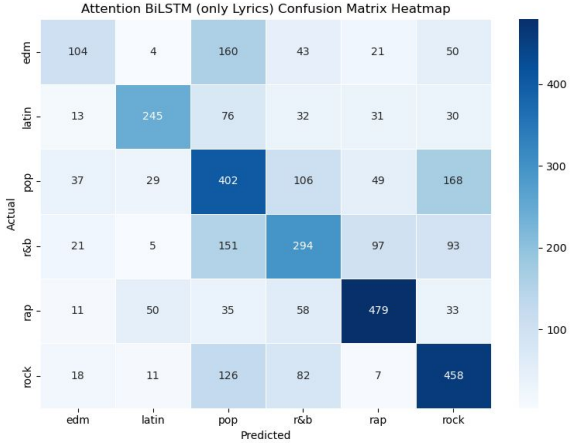
≈ 1.59M Parameters

BiLSTM: Model Performance

Lyrics + Audio Features:
1.59M parameters | **Accuracy: 0.61**



Lyrics Only:
1.57M parameters | **Accuracy: 0.55**



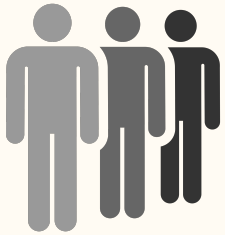
Findings:

- Audio features do **boost** performance.
- **Pop** is hardest to classify.
- **Rock & R&B** benefited most from audio features.

GRU

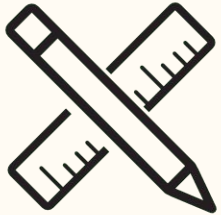
WITH Cross Attention

Why GRU for lyrics?



Lyrics Sequence

Lyrics naturally have word order and structure. Models like RNNs, which capture sequential information, are well-suited for lyrics



Lyrics Length

The median lyrics length is ~300 words — too long for a VRNN to handle. GRU can process longer sequences without suffering from vanishing gradients

Dataset Size

GRU is lighter and more efficient than LSTM and Transformer models. Given our dataset size (~18k songs), GRU provides a good balance between model complexity and performance.

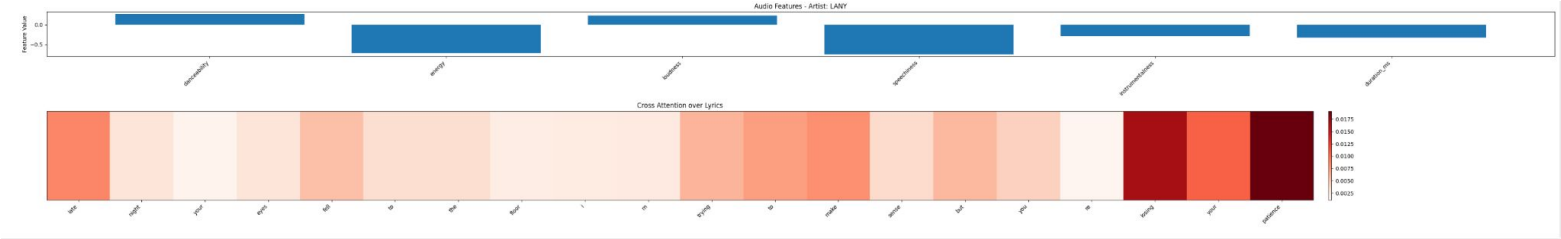
Cross Attention

Cross Attention allows model to **dynamically** identify which parts of the lyrics are most relevant to the sound.

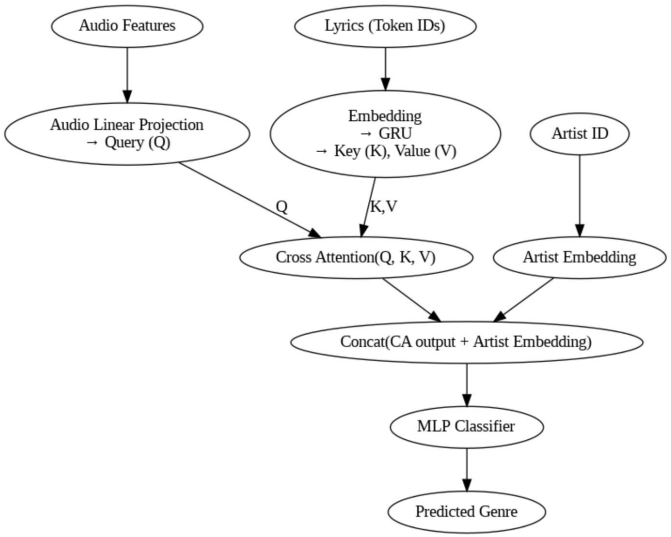
it's like asking "**Given how this song sounds, which words should I pay attention to**?"

Q: audio Features

K,V: Lyrics



GRU+Cross Attention Structure



≈ 1.6M Parameters

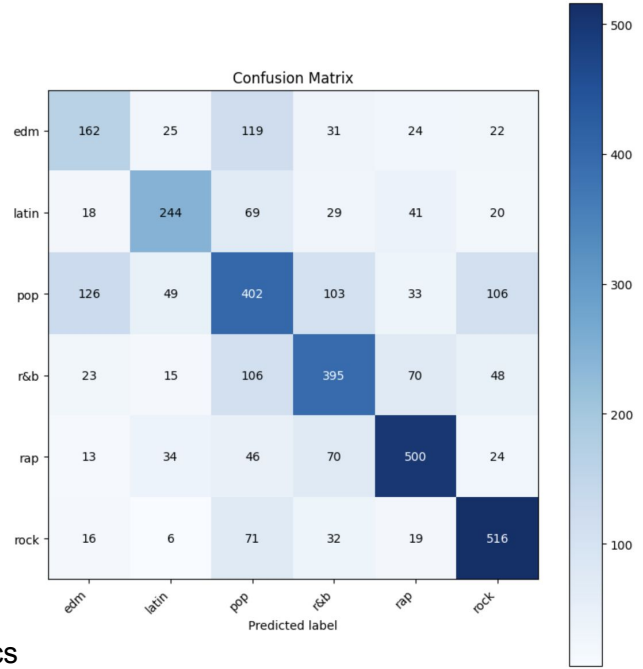
Module	Layer Type	Input Dimension	Output Dimension	Dropout / Activation
Text Embedding	Embedding (Lyrics)	(Batch, Lyrics Length)	(Batch, Lyrics Length, 128)	Nozne
Text Encoder	GRU (Lyrics Encoder)	(Batch, Lyrics Length, 128)	(Batch, Lyrics Length, 128)	None
Audio Projection	Linear (Audio Features)	(Batch, 7) *(6 audio features + missing flag)*	(Batch, 128)	None
Cross Attention	Multihead Attention	Query: (Batch, 1, 128) Key/Value: (Batch, Lyrics Length, 128)	(Batch, 1, 128)	None
Artist Embedding	Embedding (Artist)	(Batch,)	(Batch, 32)	None
Classifier	Linear (Fusion MLP - 1st Layer)	(Batch, 160) *(128 + 32 fusion)*	(Batch, 128)	ReLU + Dropout(0.3)
Classifier	Linear (Fusion MLP - 2nd Layer)	(Batch, 128)	(Batch, Num Genres)	None

GRU+Cross Attention

Overall Accuracy: **0.6118**

Classification Report:

	precision	recall	f1-score	support
edm	0.4525	0.4230	0.4372	383
latin	0.6542	0.5796	0.6146	421
pop	0.4945	0.4908	0.4926	819
r&b	0.5985	0.6012	0.5998	657
rap	0.7278	0.7278	0.7278	687
rock	0.7011	0.7818	0.7393	660
accuracy			0.6118	3627
macro avg	0.6048	0.6007	0.6019	3627
weighted avg	0.6092	0.6118	0.6098	3627



Findings:

- Better performance than baseline models
- **Pop and edm** are hardest to classify.
- **Pop, R&N, Rock & Rap** benefited most from lyrics

BERT

WITH CHUNKED LYRICS

What is BERT?

BERT stands for **Bidirectional Encoder Representations from Transformers**

- Pre-trained on a massive amount of text

Masked Language Modelling

Next Sentence Prediction

3 Models Using BERT

Split the lyrics into overlapping chunks to go through all the lyrics in a song

1

BERT-Only

After each chunk has been classified, we used majority voting to determine the genre prediction for the song

2

BERT + MLP

Concatenate the average embeddings from BERT with audio features to train a small MLP classifier

3

BERT + Random Forest

Instead of passing through a MLP in the previous model, we pass it through Random Forest

BERT Results

1

BERT-Only

Accuracy: 59%
F1-Score: 57%
Recall: 56%

2

BERT + MLP

Accuracy: 59%
F1-Score: 58%
Recall: 58%

3

BERT + Random Forest

Accuracy: 57%
F1-Score: 58%
Recall: 59%

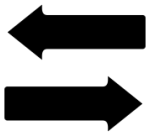
Transformer + MLP

WITH MULTI-HEAD SELF-ATTENTION

Intuition

To make use of strengths from each method,
leveraging both textual and numerical information.

Transformer

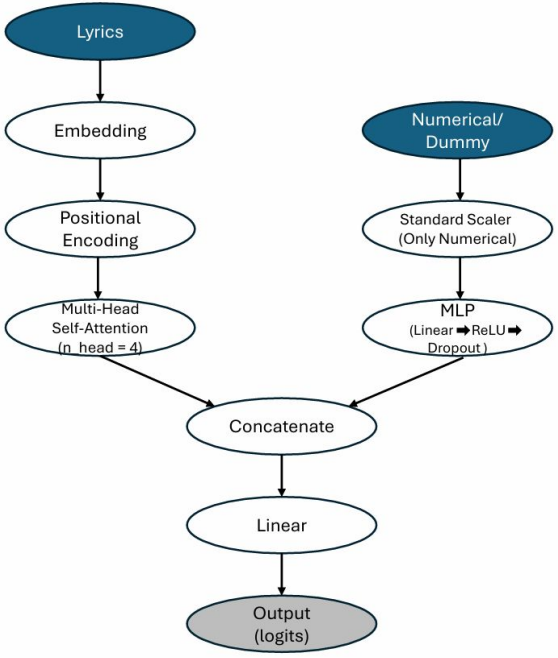


Long-range dependencies and contextual patterns

MLP

Extract patterns from structured data

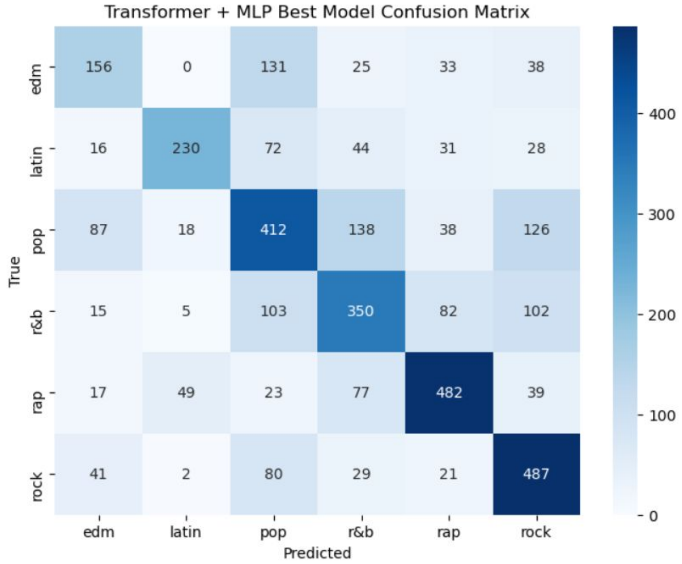
Transformer + Multi-Head Structure



Module	Layer Type	Input Dimension	Output Dimension	Dropout / Activation
Text Embedding	Embedding	(seq_len,) → 64	(seq_len, 64)	—
Transformer Encoder	2× TransformerEncoder Layer	(seq_len, 64)	(seq_len, 64)	Dropout(0.1)
Feature Linear	Linear	11	128 → 64 → 32	Dropout (0.1), ReLU
Concat Lyrics & Features	—	(64 + 32) = 96	96	—
Classifier	Linear	96	6	—

≈ 10.28M Parameters

Transformer: Model Performance



Lyrics + Audio Features:
10.28M parameters | **Accuracy: 0.58**

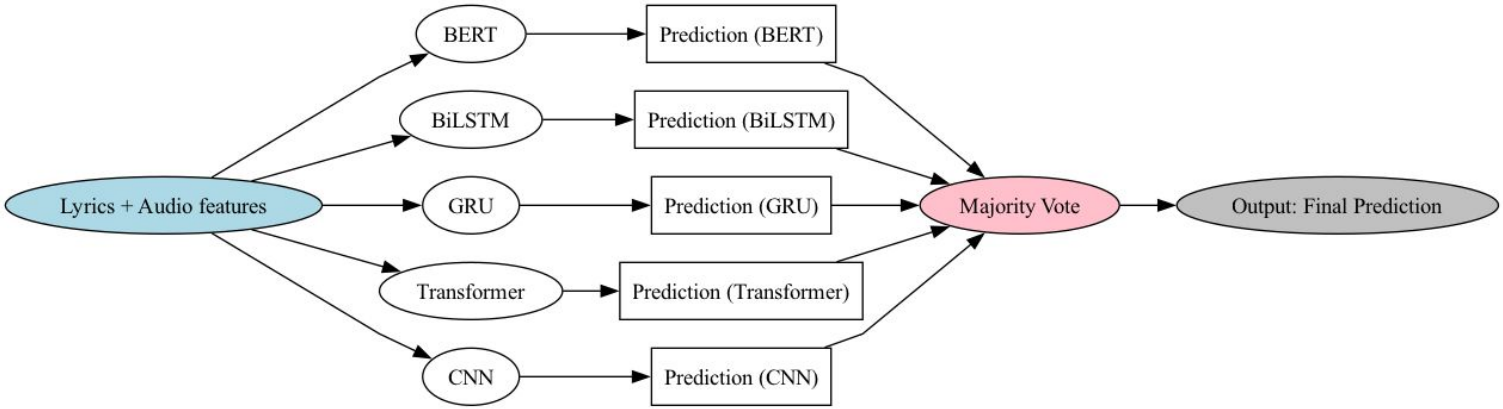
Findings:

- **Pop** is hardest to classify.
- Significant misclassification among **pop, r&b,** and **rock**
- **Rock** and **rap** are well classified

Ensemble

With Majority Voting

Ensemble Model

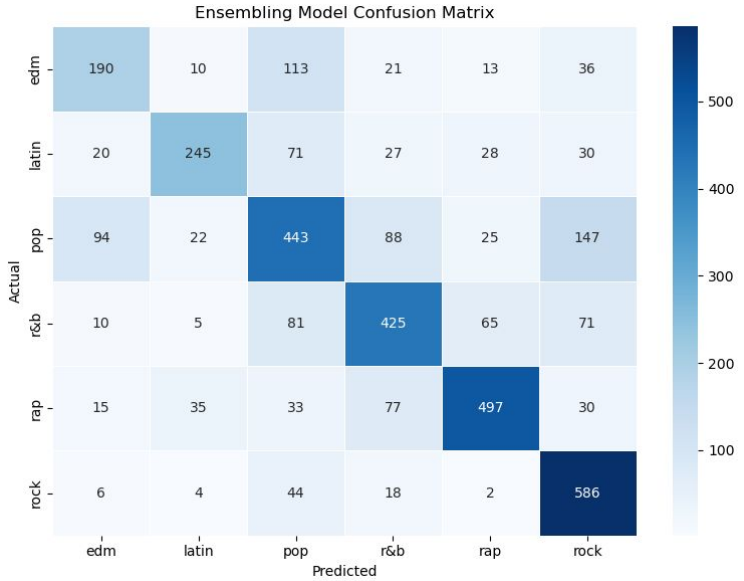


Mitigate individual bias!

Ensemble: Model Performance

Accuracy: 0.66 (Highest among all models)

- Each model captures different patterns and has different bias.
- Ensemble can correct individual bias and give more reliable results



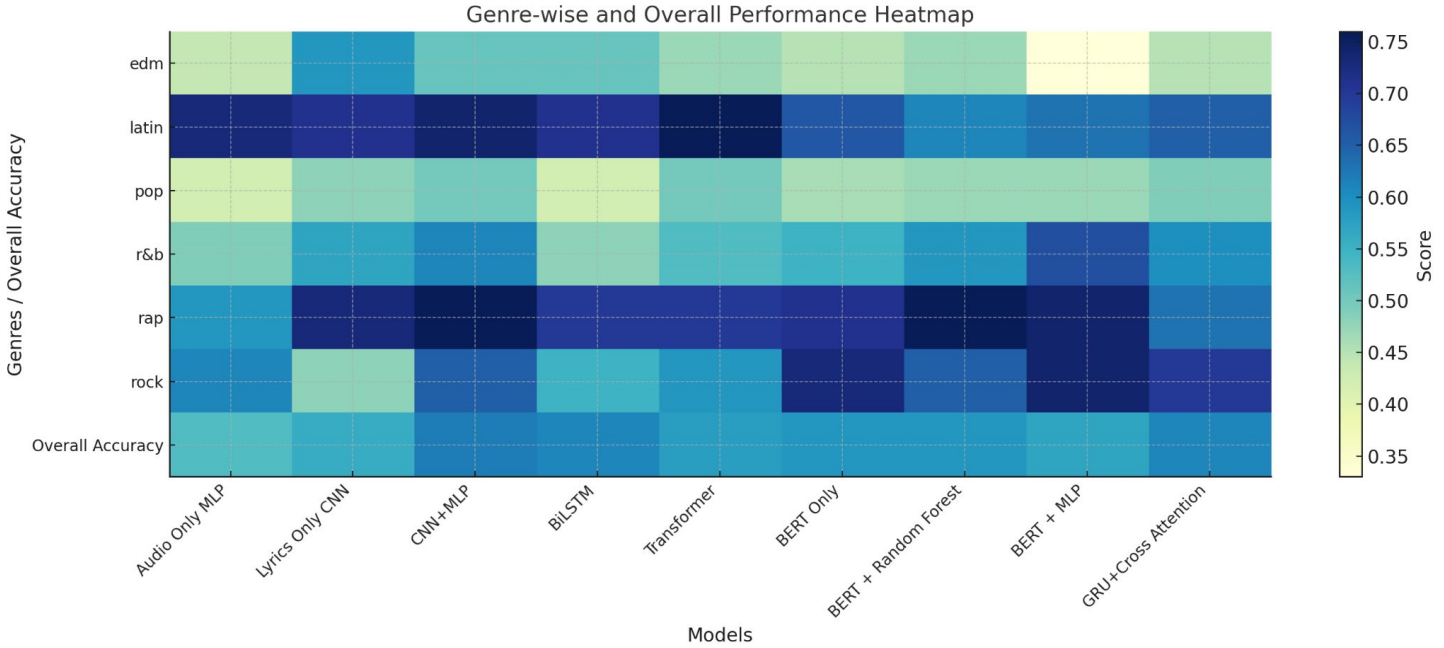
Result Comparison

Model Performances

Model Name	Parameters	Accuracy
MLP	2390	0.53
CNN	1.87M	0.56
CNN + MLP	1.53M	0.62
BiLSTM (lyrics only)	1.57M	0.55
BiLSTM (lyrics + audio)	1.59M	0.61
GRU with Cross Attention	1.60M	0.61
Transformer + MLP	10.27M	0.58
BERT-Only	109.49M	0.59
BERT + Random Forest	109.49M	0.59
BERT + MLP	109.6M	0.57
Ensemble Model (Majority Voting)	—	0.66

Conclusion

Conclusion


Genre-wise and Overall Performance Heatmap
Genres / Overall Accuracy
edm
latin
pop
r&b
rap
rock
Overall Accuracy
Audio Only MLP
Lyrics Only CNN
CNN+MLP
BiLSTM
Transformer
BERT Only
BERT + Random Forest
BERT + MLP
GRU+Cross Attention
Models
Score
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35

Future Improvement

Improvements

- Parameter Efficiency
- Advanced Fusion Strategies (attention-based cross-modal interactions)
- Incorporation of Explainable AI
- Multi-Task Learning