CS 4641 Final Presentation

Group 8:
Wendy Pan, Qianyu Zheng, Yurui Wu, Junjie Tang, Yuanming Luo

GTZAN is the dataset we are using

Introduction & Problem

Disorganized library

Without effectively grouping the music, the order of music in library is basically the time of getting in which is very disorganized. The recommendation algorithm will not effectively select the music that the customer most likely will pick.

Economic concerns

When consumers are constantly presented with irrelevant suggestions, poor recommendation algorithms not only make the user experience worse but may also cause streaming platforms to incur financial losses.

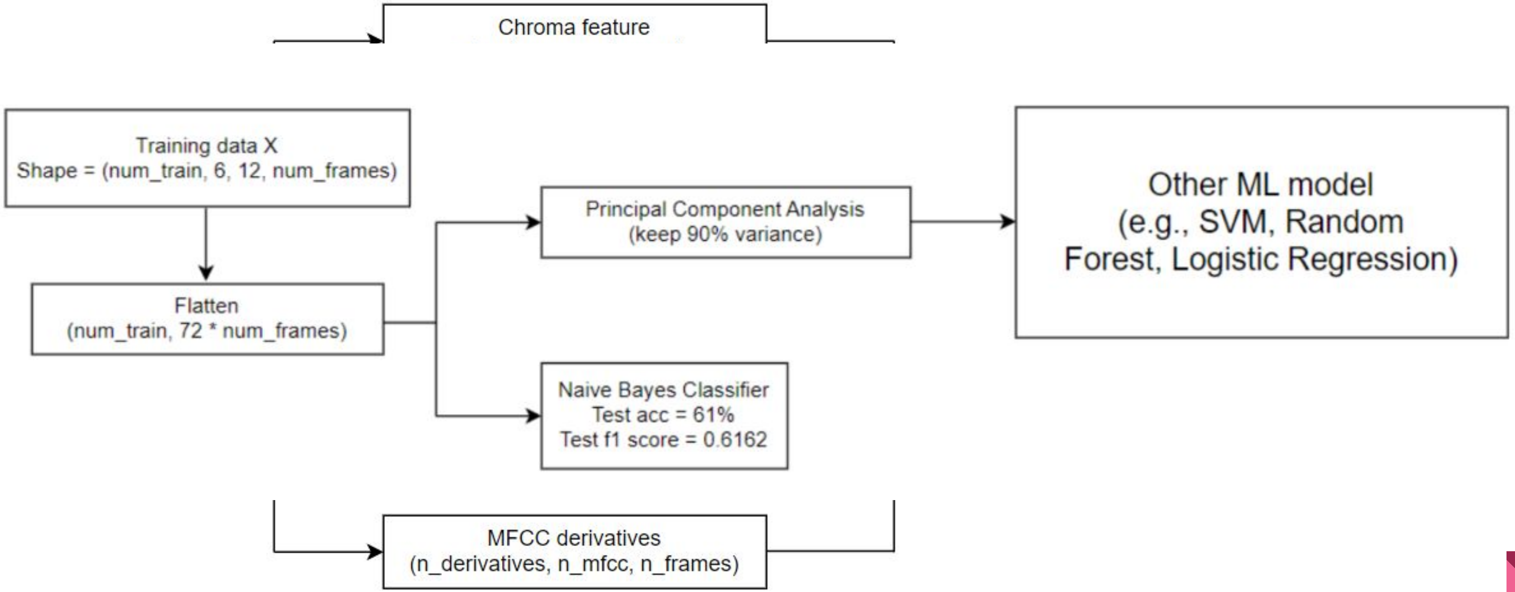
Machine learning

ML algorithm can be used to group music with their genre. Predict classification of new music.

Strategy 1

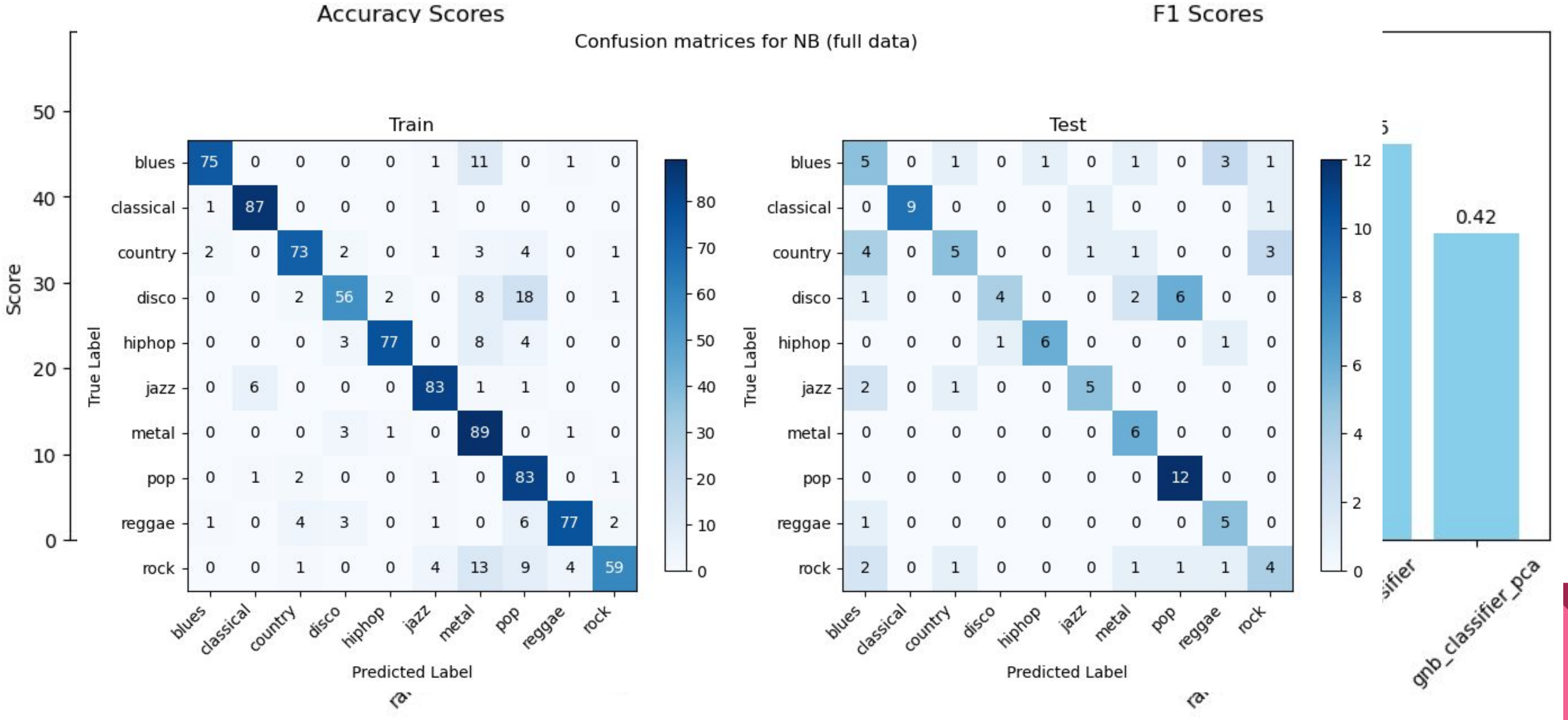
Stacked MFCC & Chroma

Preprocessing Workflow



In Strategy 1, n_chroma = n_mfcc = 12, and n_derivate = 2

Results


Accuracy Scores
F1 Scores
Confusion matrices for NB (full data)
Train
Test
True Label
Predicted Label
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
Score
0.42
gnb_classifier_pca

Discussion and Reflection

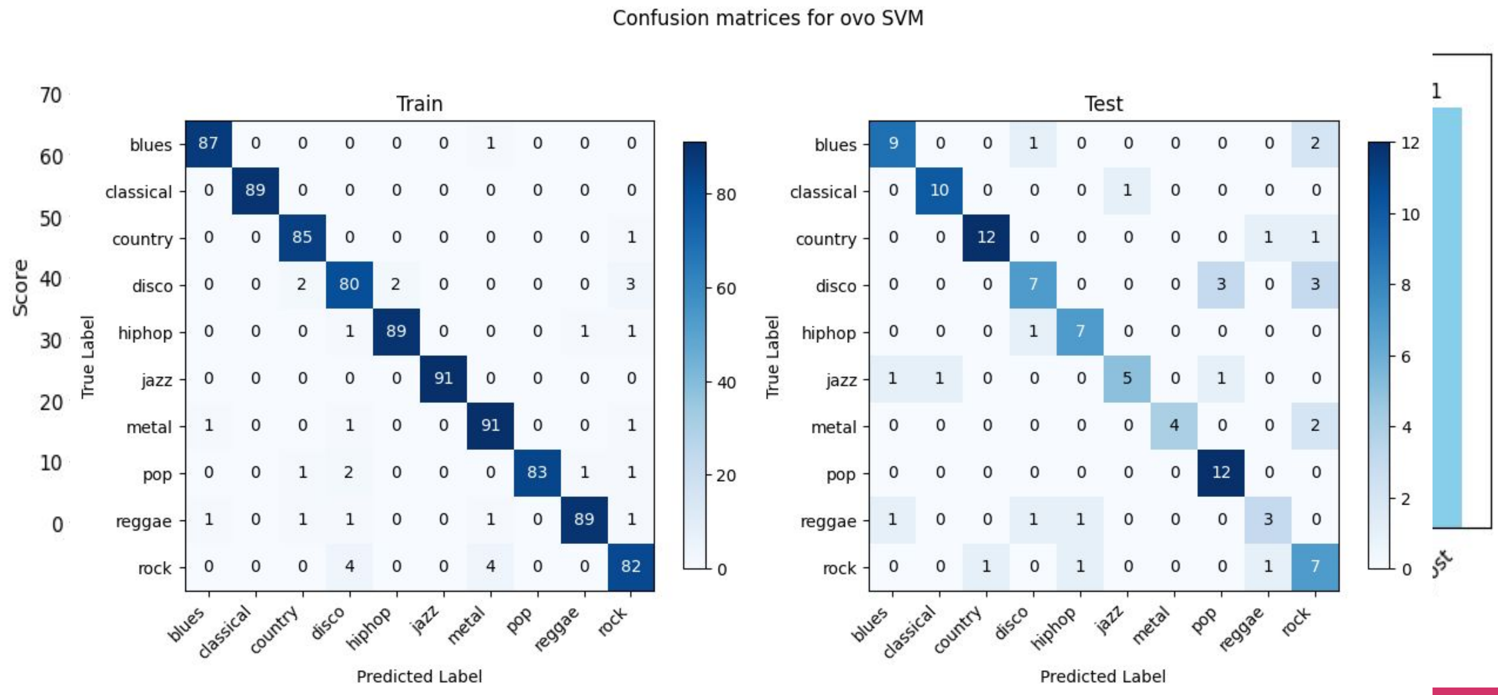
- Initial model using full MFCC and Chroma graphs resulted in 60% accuracy possibly due to high dimensionality that hindered learning and generalization capabilities of the models
- Proposed solution:
  - Reduce feature set by extracting key characteristics (e.g., statistical measures, domain-specific metrics) from MFCC and Chroma graphs.
  - Introduce additional features such as spectrograms, tempo graphs, and RMS values to enrich the feature set.
- Objective: Improve model performance by balancing complexity and representational power of the features, enhancing model's ability to generalize and learn effectively.

Strategy 2.1

Manual Feature Engineering

For each audio generate 77 Features:
Mean and variance of root-mean-square of frames, spectral centroid, spectral bandwidth, spectral rolloff, zero crossing rate, harmony, 20 MFCC and 12 Chromas, and tempo.

Results



Confusion matrices for ovo SVM
Train
blues 87 0 0 0 0 0 1 0 0 0
classical 0 89 0 0 0 0 0 0 0 0
country 0 0 85 0 0 0 0 0 0 1
disco 0 0 2 80 2 0 0 0 0 3
hiphop 0 0 0 1 89 0 0 0 1 1
jazz 0 0 0 0 0 91 0 0 0 0
metal 1 0 0 1 0 0 91 0 0 1
pop 0 0 1 2 0 0 0 83 1 1
reggae 1 0 1 1 0 0 1 0 89 1
rock 0 0 0 4 0 0 4 0 0 82
True Label
Predicted Label
blues classical country disco hiphop jazz metal pop reggae rock
0 20 40 60 80
Test
blues 9 0 0 1 0 0 0 0 0 2
classical 0 10 0 0 0 1 0 0 0 0
country 0 0 12 0 0 0 0 0 1 1
disco 0 0 0 7 0 0 0 3 0 3
hiphop 0 0 0 1 7 0 0 0 0 0
jazz 1 1 0 0 0 5 0 1 0 0
metal 0 0 0 0 0 0 4 0 0 2
pop 0 0 0 0 0 0 0 12 0 0
reggae 1 0 0 1 1 0 0 0 3 0
rock 0 0 1 0 1 0 0 0 1 7
True Label
Predicted Label
blues classical country disco hiphop jazz metal pop reggae rock
0 2 4 6 8 10 12
Score
0 10 20 30 40 50 60 70

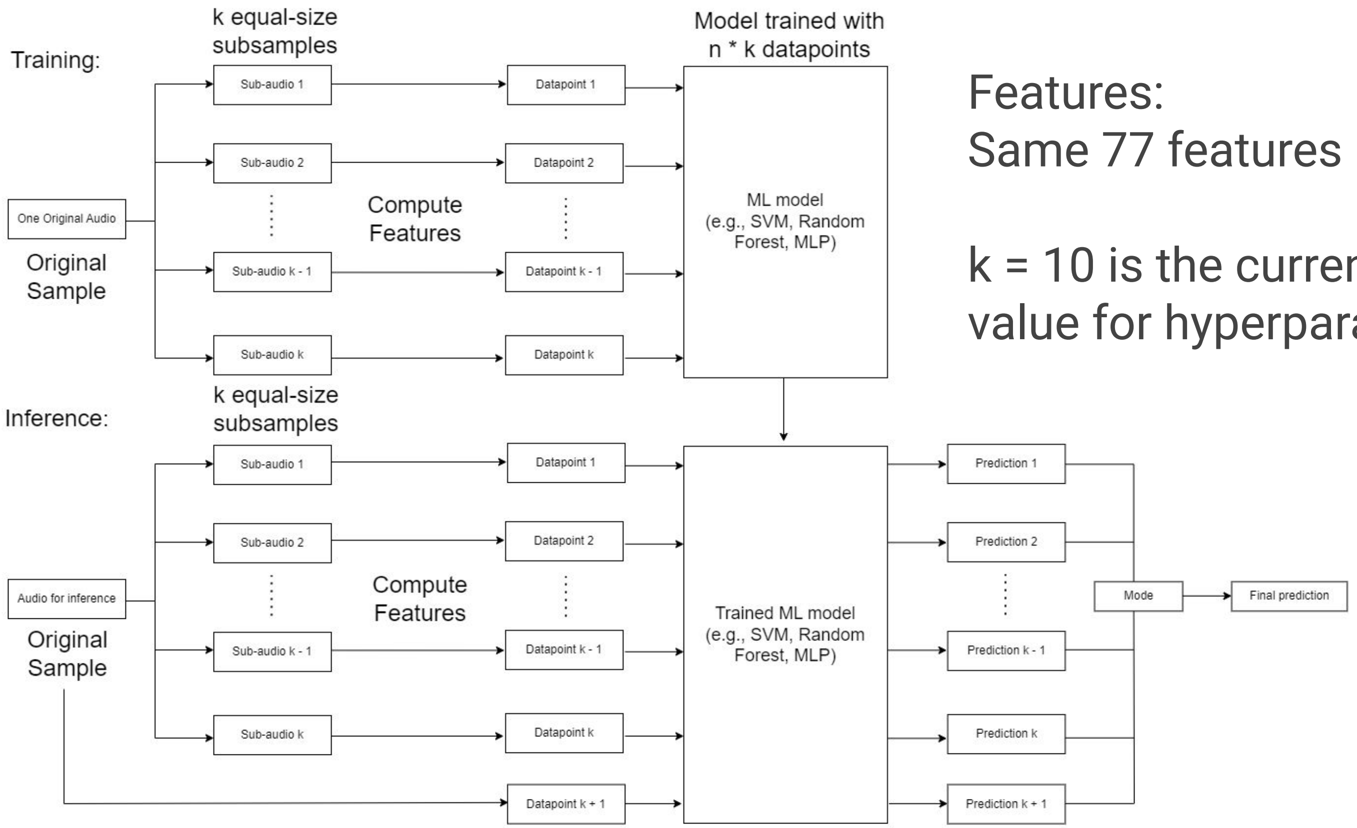
Discussion and Reflection

- The SVM's RBF kernel is effective in handling the non-linearity in audio data, utilizing the higher dimensions to construct a genre-separating hyperplane with maximum margin
- Regularization parameter C in SVM helps balance the trade-off between maximizing the margin and minimizing classification errors, enhancing robustness and generalization.
- Despite its strengths, manual feature selection remains challenging and time-consuming, necessitating domain expertise and significant experimentation.

Strategy 2.2

Manual Feature Engineering with Split Training

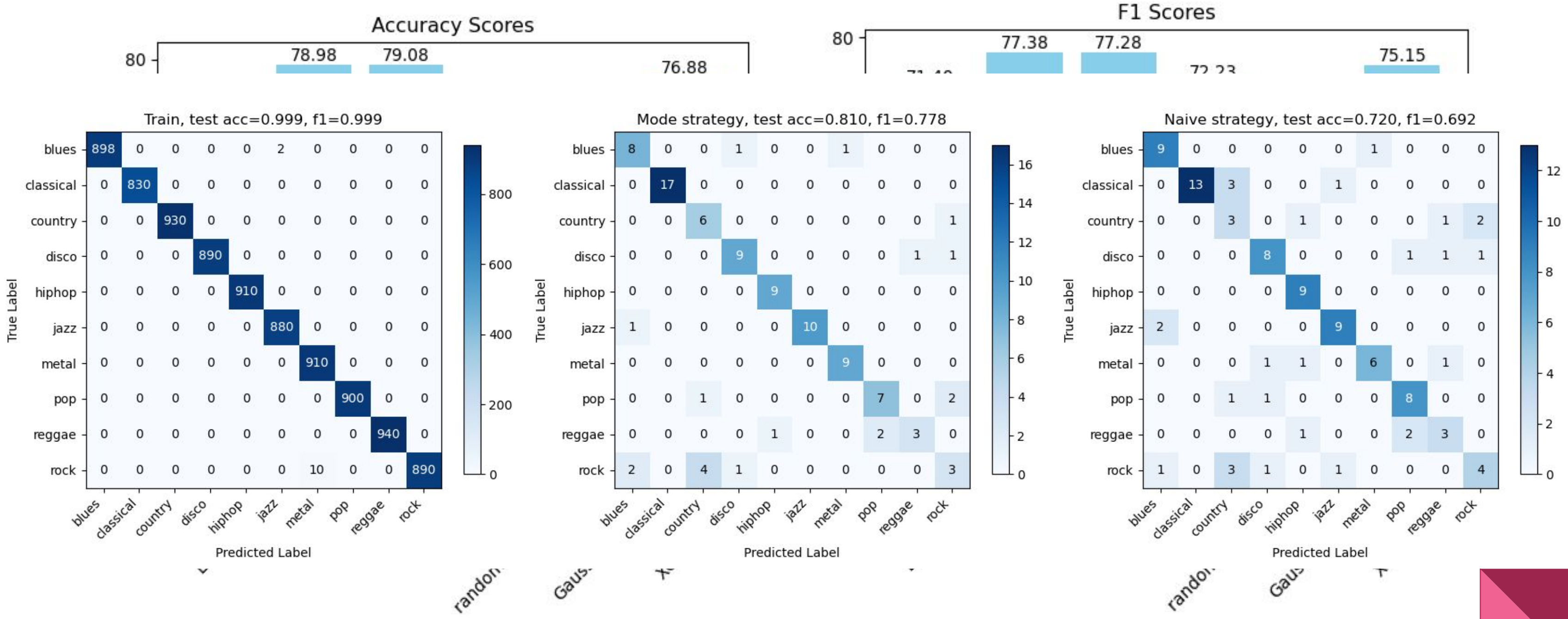
Pipeline



Features:
Same 77 features as in Strategy 2.1

k = 10 is the current best-performing value for hyperparameter k.

Results



5% improvement across models!

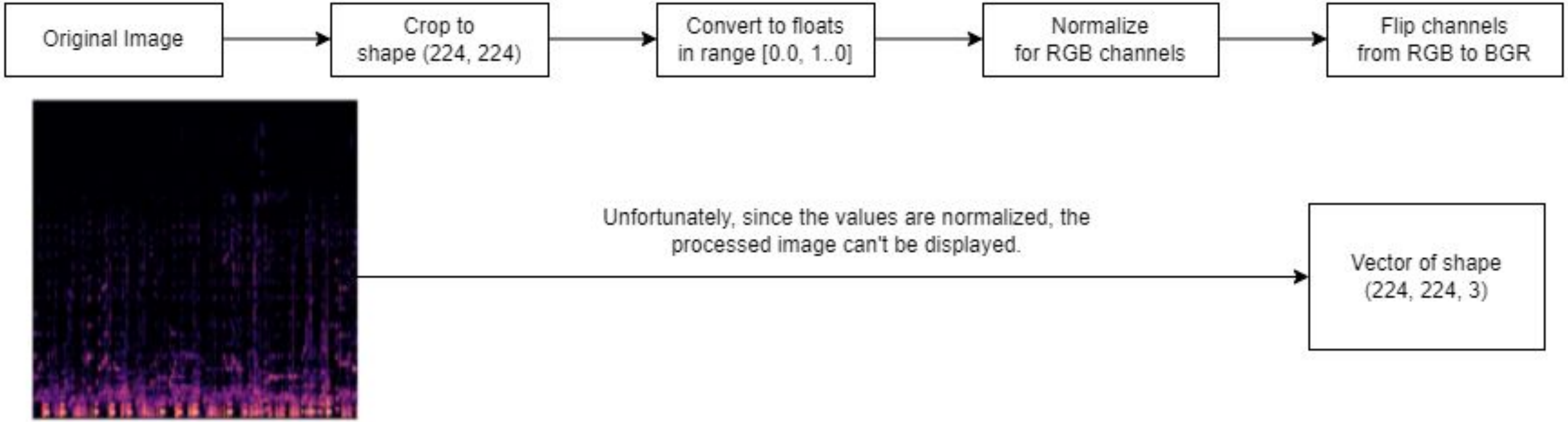
Discussion and Reflection

- Advantages of local feature extraction on using subsamples:
  - Local pattern recognition
  - Increased training data and reduced overfitting
  - Aggregation of predictions
- Proposed challenge:
  - Genre mismatch in subsamples

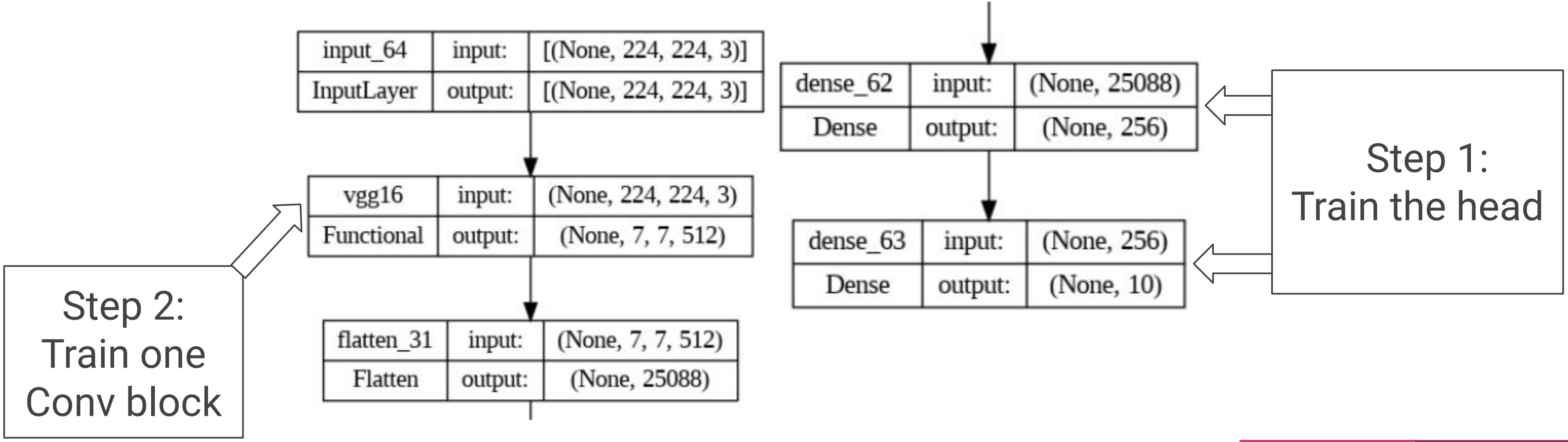
Strategy 3

Two-Step VGG Fine Tuning

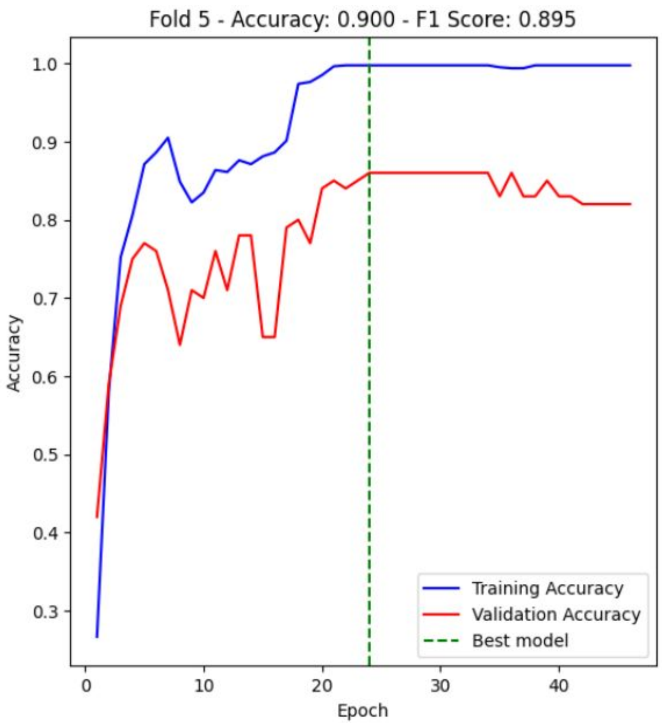
Spectrogram processing

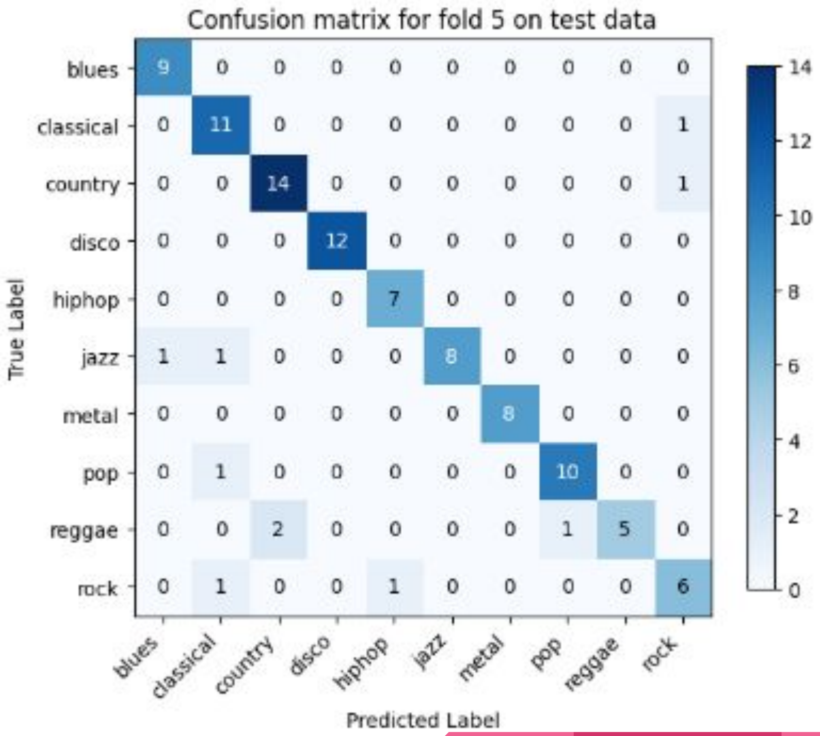


Model structure



Results


Fold 5 - Accuracy: 0.900 - F1 Score: 0.895
Accuracy
Epoch
Training Accuracy
Validation Accuracy
Best model


Confusion matrix for fold 5 on test data
True Label
Predicted Label
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock

Discussion and Reflection

- 2-step fine-tuning process - adapt the model to specific tasks while leveraging pre-trained ImageNet features
- Focuses on fine-tuning the later convolutional layers
- Combines audio-to-image conversion, transfer learning, gradual layer adaptation, and precise preprocessing to enhance model effectiveness, showing promising results for the specified task

Comparison

Performance: (best) 3 > 2 > 1 (worst)

Computational Cost: (cheapest) 1 < 2 < 3 (most expensive)

Requirement for Domain-Specific Knowledge: (need most) 2 > 1 > 3 (need least)

Model Interpretability: 2

References

[1] Tzanetakis, G. and Cook, P. (2002) 'Musical genre classification of Audio Signals', IEEE Transactions on Speech and Audio Processing, 10(5), pp. 293–302. doi:10.1109/tsa.2002.800560.

[2] Li, T., Ogihara, M., & Li, Q. (2003, July). A comparative study on content-based music genre classification. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 282-289).

[3] Burred, J. J., & Lerch, A. (2003, September). A hierarchical approach to automatic musical genre classification. In Proceedings of the 6th international conference on digital audio effects (pp. 8-11).

[4] Ndou, N., Ajoodha, R., & Jadhav, A. (2021, April). Music genre classification: A review of deep-learning and traditional machine-learning approaches. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (pp. 1-6). IEEE

[5] Bahuleyan, H. (2018). Music Genre Classification using Machine Learning Techniques. arXiv:1804.01149v1

Thank You