The Role of Dimensionality Reduction and Classification Techniques in Music Genre Recognition Using the GTZAN

Arash Khosropour
dept. Informatica
Università degli Studi di Milano
Milan, Italy
arash.khosropour@studenti.unimi.it

Abstract

Music genre classification is essential for managing digital music libraries and enhancing user experiences through content discovery and personalization. This study evaluates various classification methods, including simple classifiers like Random Forest and SVM and advanced models such as CNN, LSTM, and FCNN. Principal Component Analysis (PCA) was applied as a preprocessing step to reduce features to 39 components, retaining over 95% of the variance. While PCA improved computational efficiency, it disrupted temporal dependencies, reducing the performance of CNN (87% to 82%) and LSTM (88% to 80%). FCNN, with batch normalization, achieved 91% accuracy, matching fine-tuned SVM, which leveraged PCA effectively. Simpler classifiers provided up to 85% accuracy, offering resource-efficient alternatives. This study highlights the trade-offs between dimensionality reduction and model performance, providing insights for optimizing classification pipelines in music genre recognition tasks.

I. Introduction

The exponential growth of digital music libraries has created a pressing need for effective music genre classification systems to enhance content discovery, personalization, and user experience. Music genres serve as a means to categorize audio content based on their distinct musical characteristics, making classification an integral aspect of digital music management. However, achieving high accuracy in music genre classification remains challenging due to the complexity of audio signals and the overlapping of features across genres.

Dimensionality reduction techniques like Principal Component Analysis (PCA) address these challenges by reducing the feature space while retaining most of the variance in the data. By lowering computational complexity, PCA enables the use of advanced models on high-dimensional datasets. However, PCA's transformation can impact models that rely on temporal dependencies, such as CNNs and LSTMs, potentially reducing their effectiveness.

Previous studies have demonstrated the potential of machine learning and deep learning for music genre classification. For instance, Choi (2016) utilized Convolutional Neural Networks (CNNs) in [1], achieving 65% accuracy on a small dataset. In 2018, Elbir introduced a hybrid model combining Long Short-Term Memory (LSTM) networks with Support Vector Machines (SVM) for music genre classification [2]. When tested on the GTZAN dataset, this hybrid model achieved an accuracy of 89%, surpassing the individual performances of standalone LSTM and SVM classifiers. Xu (2020) integrated attention mechanisms with CNNs [3], pushing accuracy to 90.3%, while Saha (2019) applied transfer learning with CNNs and SVMs [4], achieving 80%. Xiaoyu Xie (2024) further advanced the field by combining CNNs with Long Short-Term Memory (LSTM) networks [5], achieving 89.6% accuracy on a large-scale dataset. These advancements underscore the effectiveness of deep learning, albeit with notable computational demands and data requirements.

This study evaluates the impact of PCA preprocessing on music genre classification using the GTZAN dataset. A comprehensive comparison of classifiers—ranging from simple machine learning models to deep learning architecture is presented. The results demonstrate the trade-offs between computational efficiency, feature dependency preservation, and classification accuracy. Insights from this research offer practical guidelines for designing efficient and robust music genre classification systems.

The rest of this paper is structured as follows: Section II describes the dataset and methodology, Section III outlines the experimental setup, and Section IV presents the results and analysis, highlighting key observations and future directions.

II. Description of the Methods

A. Data Preprocessing

Raw audio data from the GTZAN dataset were preprocessed, including normalization and handling missing data. The dataset is split into training(80%) and testing(20%) sets.

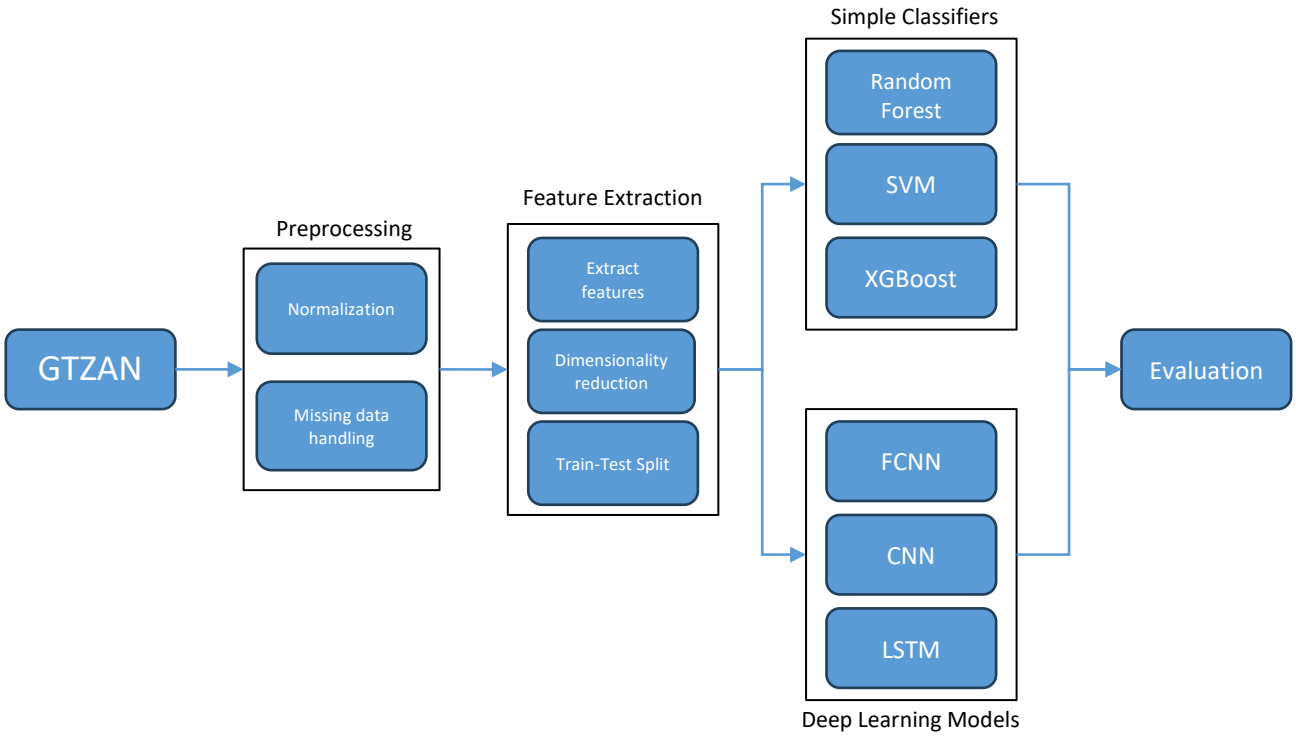

Fig.1. Block Diagram of The Methods

B. *Feature Extraction*

Key audio features are extracted for genre representation, including chroma, spectral characteristics, zero-crossing rate, and Mel-frequency cepstral coefficients (MFCCs). PCA is applied to reduce the feature space to 39, retaining 95% of the variance.

C. *Modeling*

A variety of classification methods are employed:

- Simple Classifiers: Random Forest, SVM and XGBoost.
- Deep Learning Models: Convolutional Neural Network (CNN), Fully Connected Neural Networks (FCNN) and Long Short-Term Memory Networks (LSTM), fine-tuned for optimal performance.

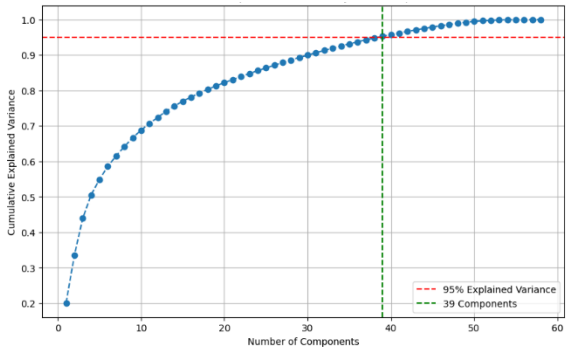

Fig.2. Cumulative Explained Variance by PCA Components

D. *Evaluation*

The models are trained on the reduced feature set, and their performance is evaluated using metrics like accuracy, precision, recall, and F1-score.

III. Experiments

A. *Dataset Description*

The experiments were conducted using the GTZAN dataset, a benchmark dataset for music genre classification. It consists of:

- Genres: 10 distinct genres (e.g., Classical, Pop, Jazz, Rock).
- Samples: 1,000 audio clips, each 30 seconds long (100 clips per genre).
- Features: Pre-extracted features from the features_3_sec.csv file, including spectral, rhythmic, tonal, and MFCC descriptors. Each clip was segmented into 3-second windows, resulting in 30 segments per clip.

B. *Experimental Setup*

B.1. Data Preprocessing

- Feature Selection: All 59 features in the features_3_sec.csv file were used.
- Dimensionality Reduction: PCA reduced the features to 39 principal components, retaining over 95% of the variance.
- Data Splitting: The dataset was stratified into training (70%), validation (15%), and testing (15%) sets for

deep learning models and into training (80%) and testing (20%) for other classifiers. Validation was used to prevent overfitting in deep learning models.

B.2. Hardware and Tools

- Hardware: Experiments were conducted on a system with an NVIDIA RTX 3060 GPU, 32GB RAM, and Intel Core i7 processor.
- Software:
  - Python (3.10)
  - Scikit-learn (for classical classifiers)
  - PyTorch (for deep learning models)

C. Performance Metrics

The following metrics were used to evaluate the models:

- Accuracy: Percentage of correctly predicted genres.
- Precision: Proportion of true positive predictions for each genre.
- Recall (Sensitivity): Proportion of actual instances correctly identified for each genre.
- F1-Score: Harmonic mean of precision and recall, providing a balance between them.
- Confusion Matrix: Visualized the performance of each model across genres.

D. Comparison of Models

D.A. Random forest

Fined tuned Random Forest Classifier, got the best results with the following parameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}. The total accuracy of the model at the end was 82%, with the lowest in the country genre (73%) and the highest with metal with 88%.

D.B. SVM

We used SVC here, the main difference with SVM is that SVM focuses on similarity relationships between data points, while SVC focuses on creating a decision boundary for classification.

model was optimized via grid search with the following parameters: {'C': 10, 'gamma': 'auto', 'kernel': 'RBF'} and achieved an impressive 91% accuracy, demonstrating its robustness with PCA-reduced features. The lowest accuracy was with rock songs with 85% and the highest on metal with 96%.

D.C. XGBoost

XGBoost is an optimized distributed gradient boosting. The tuned parameters are: {'colsample_bytree': 0.8, 'learning_rate': 0.2, 'max_depth': 6, 'n_estimators': 200, 'subsample': 0.8}. This model achieved an accuracy of 85% with the lowest between the classes being rock (71%) and the highest, classical and metal both with the accuracy of 91%.

D.D. Fully Connected Neural Network (FCNN)

This architecture is a Fully connected deep feedforward neural network designed to capture complex patterns in the input data through progressively increasing and decreasing hidden layer sizes. Each layer consists of a linear transformation followed by batch normalization, ReLU activation, and dropout for regularization. The results were gathered from testing different drop out and different numbers of hidden layers. The model achieved the same accuracy as the SVM (91%). Batch normalization contributed significantly to stabilizing training and improving performance. The architecture of the model is as follows:

Input Layer:

- Input size:
  Matches the number of features in the PCA
- Hidden Layers:
  5 layers that have the Batch Normalization the ReLU activation and the dropout of 0.2.
- Output Layer:
  The 10 output neurons correspond to the 10 genres to be classified.

D.E. Convolutional Neural Network (CNN)

This architecture is a 1D Convolutional Neural Network (CNN) designed to process sequential data, combined with a series of fully connected layers to capture both local and global patterns in the input data. The model uses 1D convolutional layers to extract spatial features, followed by max-pooling to reduce dimensionality, and a stack of fully connected layers with batch normalization, ReLU activation, and dropout for regularization. The model achieved 82% accuracy, which is lower than the FCNN's performance, primarily due to the use of PCA for dimensionality reduction. PCA, while reducing computational complexity, may have discarded some important features necessary for the CNN to achieve higher accuracy. Batch normalization and dropout still played a significant role in stabilizing training and improving generalization. the architecture of the model is as followed:

- Input size:
  Matches the number of features in the PCA
- Convolutional Layers:
  2 layers that have the Batch Normalization and the ReLU with MaxPooling.
- Output Layer:
  A FCNN with BatchNorm and Dropout and ReLU that for last layer it has 10 output

neurons correspond to the 10 genres to be classified.

D.F . Long Short-Term Memory Network (LSTM)

The LSTM model consists of an LSTM layer that processes sequences of features. It includes fully connected layers for reducing dimensions and making genre predictions. ReLU activation and dropout layers are applied to enhance performance and prevent overfitting. The final layer outputs probabilities for classifying genres, with an optional bidirectional approach to capture temporal dependencies. The model achieved 80% accuracy. Its underperformance was attributed to the limited temporal resolution of PCA components. the architecture of the model is as followed:

- 1 LSTM layer (with optional bidirectional processing).
- 2 Fully Connected layers (fc1 and fc2), each followed by ReLU and dropout. First layer reduces the LSTM output size. Final output is the number of genres

Models	Accuracy
Random forest	0.82
SVM	0.91
XGBoost	0.85
FCNN	0.91
LSTM	0.80
CNN	0.82

Table.1. Accuracy Report of The Models

IV. Analysis

The application of Principal Component Analysis (PCA) as a preprocessing step reduced the feature space to 39 components, retaining over 95% of the variance while improving computational efficiency. However, PCA disrupted temporal dependencies, which negatively impacted models reliant on sequential data, such as CNN and LSTM. The accuracy of CNN decreased from 87% to 82%, while LSTM's accuracy dropped from 88% to 80%, demonstrating PCA's limitation in preserving temporal structure. In contrast, FCNN and SVM effectively leveraged PCA, both achieving 91% accuracy. FCNN benefited from batch normalization, which stabilized training and improved generalization, while SVM's grid search optimization enabled robust classification using PCA-reduced features. Simpler models, such as Random Forest and XGBoost, achieved 82% and 85% accuracy, respectively, providing resource-efficient alternatives for music genre classification.

Genre-specific analysis revealed that metal and classical achieved consistently higher accuracies across all models, likely due to their distinct acoustic features. In contrast, overlapping genres such as rock and country posed classification challenges, with frequent misclassifications observed in the confusion matrix. For example, rock songs were often misclassified as country, and vice versa, due to similarities in instrumentation and rhythm.

V. Conclusion

This study evaluated the impact of dimensionality reduction and various classification methods on music genre recognition using the GTZAN dataset. Principal Component Analysis (PCA) effectively reduced the feature space and improved computational efficiency. However, PCA disrupted temporal dependencies, negatively affecting the performance of models like CNN and LSTM, with accuracy reductions of 5% and 8%, respectively. Simpler classifiers such as Random Forest and XGBoost demonstrated resource-efficient alternatives, achieving accuracies of 82% and 85%. Meanwhile, SVM and Fully Connected Neural Networks (FCNN), which utilized PCA effectively, achieved the highest accuracy of 91%. Genre-specific analysis revealed high classification accuracies for metal and classical genres, while overlapping genres like rock and country posed challenges due to feature similarities. These results highlight the trade-offs between computational efficiency and feature dependency preservation, providing practical insights for optimizing music genre classification pipelines. Future work could explore methods to preserve temporal dependencies during dimensionality reduction or develop hybrid architectures to balance efficiency and accuracy.

References

[1] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 2392–2396.

[2] A. M. Elbir, G. Aydin, and O. Yilmaz, "A hybrid model combining LSTM networks and SVM for music genre classification," in Proc. IEEE Signal Processing and Communications Applications Conf. (SIU), Izmir, Turkey, 2018, pp. 1–4 .

[3] Y. Xu and W. Zhou, "A deep music genres classification model based on CNN with Squeeze & Excitation Block," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 2020, pp. 332–338.

[4] S. S. Saha, R. Hossain, and M. Rahman, "Music Genre Classification Using Machine Learning," International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 4, pp. 3703–3708, Apr. 2020.

[5] X. Xie, "A Hybrid CNN-LSTM Architecture for Enhanced Music Genre Classification," IEEE Xplore.

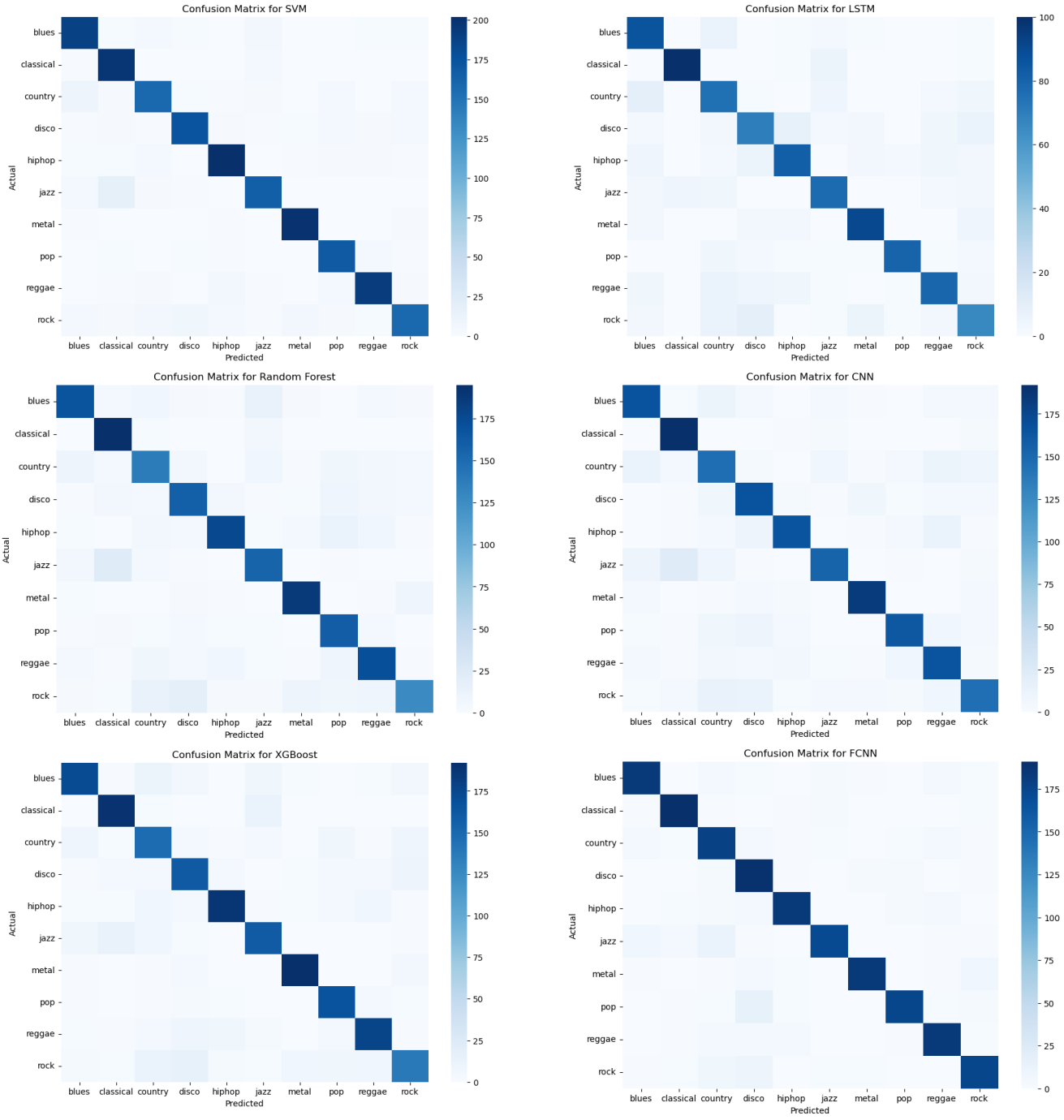

Fig.3. Confusion Matrix of Models