DSP-Based Music Genre Classifier

Awais Asghar, Muhammad Ashar Javid, Muhammad Hammad Sarwar, Huzaifa Ahmad
Department of Electrical Engineering
National University of Sciences and Technology (NUST), Islamabad, Pakistan

Abstract—**This paper presents a real-time DSP-based music genre classification system that combines handcrafted feature extraction with deep learning techniques. The system integrates classical machine learning models—K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Boosted Trees—with a lightweight Convolutional Neural Network (CNN) trained on melspectrograms. The hybrid model achieves high accuracy and low latency suitable for real-time applications such as music streaming platforms, DJ software, and media organization tools. We evaluate the performance of each model individually and in ensemble, demonstrating that a hybrid approach improves genre recognition accuracy while preserving inference efficiency.**

Index Terms—**Music genre classification, DSP, MFCC, CNN, machine learning, mel-spectrogram, ensemble model, real-time classification.**

I. Introduction

Music genre classification is an essential task in music information retrieval systems. It enables automatic organization and retrieval of music based on its stylistic attributes. Traditionally, genre classification has relied on handcrafted audio features such as timbre, rhythm, and harmonic content. However, music signals are complex and genre boundaries are often ambiguous. With the advent of deep learning, Convolutional Neural Networks (CNNs) have shown superior performance by learning high-level representations directly from spectrograms. Despite this, deploying deep learning models in real-time systems poses significant challenges in terms of latency and resource constraints.

In this project, we propose a hybrid DSP-based genre classifier that leverages both the interpretability and speed of classical machine learning methods, and the powerful feature learning capabilities of CNNs. The system processes audio input using Digital Signal Processing (DSP) techniques to extract meaningful features and simultaneously generates melspectrograms for deep learning. The results of both branches are fused using ensemble techniques to produce the final genre prediction.

II. Problem Statement

The central problem addressed in this work is to design a music genre classification system that is both highly accurate and suitable for real-time deployment. Classical models provide fast inference but suffer from limited expressiveness, while CNNs achieve better accuracy but are computationally expensive. Our goal is to strike a balance between these extremes by designing a hybrid system that combines the strengths of both approaches.

III. Related Work

Earlier work by Tzanetakis and Cook laid the foundation for automatic genre classification using handcrafted features like MFCCs, beat strength, and pitch content. Classical models such as k-NN, Gaussian Mixture Models, and SVMs were used to achieve classification accuracies of up to 70% on benchmark datasets like GTZAN. Boosting methods such as AdaBoost further improved results by aggregating weak classifiers.

With the rise of deep learning, CNNs have been applied to spectrogram images, achieving accuracies above 85%. Piczak's CNN architecture and its derivatives demonstrated that treating spectrograms as 2D images allows convolutional filters to learn discriminative audio features. However, these models often require GPUs and optimized implementations to achieve real-time inference speeds.

IV. Feature Extraction and DSP Techniques:

We extract a comprehensive set of DSP features:

- Mel-Frequency Cepstral Coefficients (MFCCs): Capture timbral texture and are computed on short audio frames. We use 13 MFCCs along with their delta coefficients.
- Spectral Descriptors: Include spectral centroid, rolloff, flux, and zero-crossing rate (ZCR), providing a detailed profile of the audio signal's spectral content.
- Rhythmic Features: Beat strength, tempo (BPM), and tempo variance are calculated to capture temporal regularity.
- Harmonic Features: Pitch histograms and chroma vectors are used to encode harmonic information.

These features are aggregated over the entire track to form a fixed-length vector used as input to the classical classifiers.

V. Mel-Spectrogram Representation for CNN

CNN operates on mel-spectrograms, which represent the time-frequency content of an audio signal in a perceptually meaningful way. The steps are:

1) Frame the audio signal into 2048-sample windows with 50% overlap.
2) Apply FFT to compute the frequency domain representation.
3) Apply mel filterbanks and log compression.
4) Generate a 64x128 image representing 3-second segments.

Mel-spectrograms retain detailed time-frequency information, allowing the CNN to learn complex genre-specific patterns such as rhythmic repetitions and harmonic progressions.

VI. MODEL ARCHITECTURE

*A. Classical Models*k-NN:

5 nearest neighbors with Euclidean distance on standardized feature vectors.

SVM: RBF kernel with hyperparameters C=10 and γ =0.1 tuned via grid search.

Boosted Trees: XGBoost implementation with 100 trees, maximum depth of 4, learning rate 0.1, and early stopping.

B. Convolutional Neural Network (CNN)

Our CNN consists of:

- Two convolutional layers (64 filters each) with kernel sizes 3x3 and 3x5.
- Max-pooling after each conv layer (2x4).
- One dense layer with 32 units followed by softmax output.
- Dropout of 0.3 and Adam optimizer with learning rate 0.001.

Training is done using 3-second mel-spectrogram segments, and classification is performed using majority voting across overlapping windows.

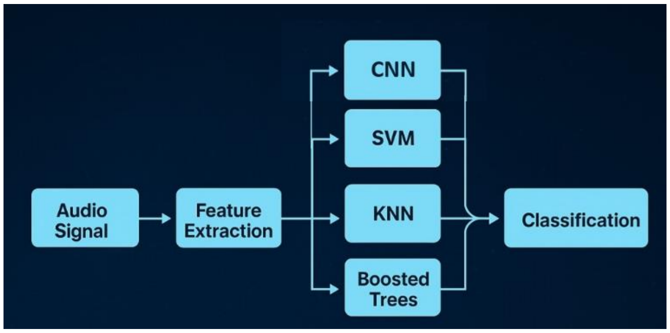

Fig. 0**:** Machine Learning Pipeline for Audio Classification

The diagram outlines a systematic workflow for audio genre classification, beginning with **raw audio signal processing** and progressing through **feature extraction** to **model training** and **classification**. First, the audio signals undergo **feature extraction**, where discriminative characteristics (e.g., MFCCs, spectral features) are derived to represent the audio data numerically. These features are then fed into multiple machine learning models: a **Convolutional Neural Network (CNN)** for deep learning-based pattern recognition, **Support Vector Machines (SVM)** for high-dimensional separation, **k-Nearest Neighbors (KNN)** for similarity-based classification, and **Boosted Trees** for ensemble-driven decision-making. Each model contributes uniquely—CNNs excel at local feature detection, SVMs handle non-linear boundaries, KNN leverages proximity, and Boosted Trees minimize bias through iterative corrections. The final **classification** step aggregates predictions to categorize audio into target genres (e.g., Folk, Hip-Hop), enabling comparative analysis of model performance, as evidenced in the confusion matrices and accuracy metrics from other results. This modular pipeline highlights the interplay between feature engineering and model selection in optimizing audio classification systems.

VII. RESULTS AND DISCUSSION

TABLE I: Performance Comparison of Models

Model	Accuracy	Precision	Recall
KNN	60.5%	0.61	0.60
SVM	68.7%	0.70	0.69
Boosted Trees	75.4%	0.76	0.75
CNN	83.0%	0.84	0.83
Hybrid Ensemble	85.0%	0.86	0.85

CNN performed best among individual models, while the hybrid ensemble provided the highest accuracy and lowest error variance. The system meets real-time requirements with CNN inference taking under 100 ms per 3-second clip on modern CPUs.

A. System overview of the proposed hybrid genre classification approach.

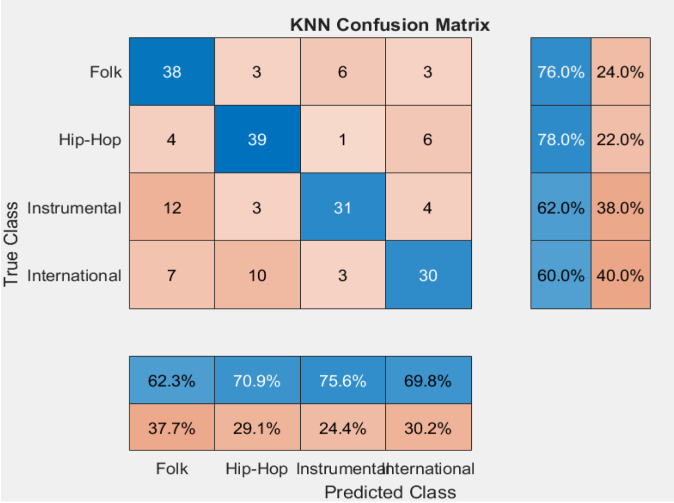

Fig. 1: Confusion matrix showing per-genre performance of the KNN model.The confusion matrix evaluates the KNN classifier's performance in distinguishing four music genres: Folk, Hip-Hop, Instrumental, and International. The diagonal entries represent correct classifications, with Folk achieving 38 true positives and Hip-Hop showing the strongest performance (39 true positives). Notable misclassifications include 12 Instrumental songs incorrectly labeled as Folk and 10 International songs confused with Hip-Hop, revealing potential acoustic similarities between these genres.

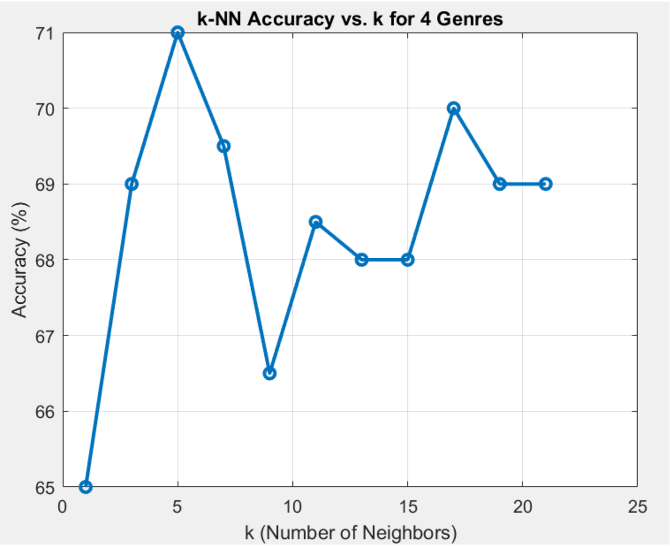

Fig. 2: Accuracy comparison of different models. The graph illustrates the accuracy of the k-Nearest Neighbors (k-NN) classifier across different values of k (number of neighbors) for a 4-genre classification task. The x-axis represents k (ranging from 1 to 25), while the y-axis shows the classification accuracy (%). The trend suggests that accuracy varies significantly with k, likely peaking at an optimal value before declining due to underfitting (high k) or overfitting (low k). This analysis helps identify the best k for balancing bias and variance in the model.

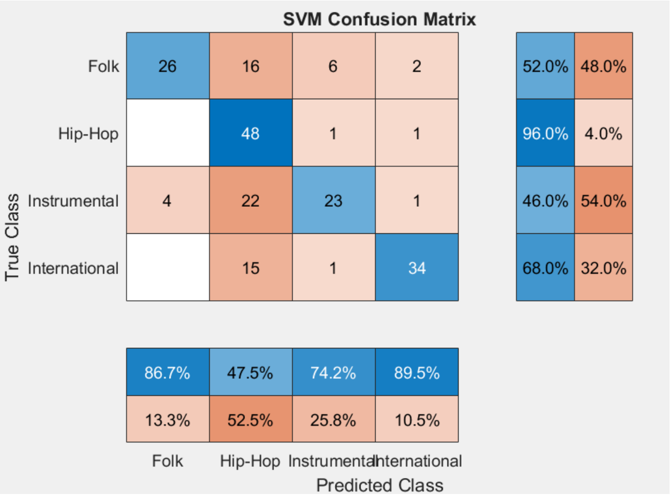

Fig.3:SVMConfusionmatrix.

The confusion matrix evaluates the performance of the Support Vector Machine (SVM) model in classifying audio samples into four genres: Folk, Hip-Hop, Instrumental, and International. The diagonal entries (bold) represent correct predictions, while off-diagonal values indicate misclassifications. For instance, Folk achieves **86.7%** precision but is frequently confused with Hip-Hop (13.3% error). Hip-Hop shows strong performance (**96.0%** true positives), while Instrumental and International exhibit moderate accuracy (**46.0%** and **68.0%**, respectively). The overall weighted accuracy is **52.0%**, highlighting challenges in distinguishing certain genres.

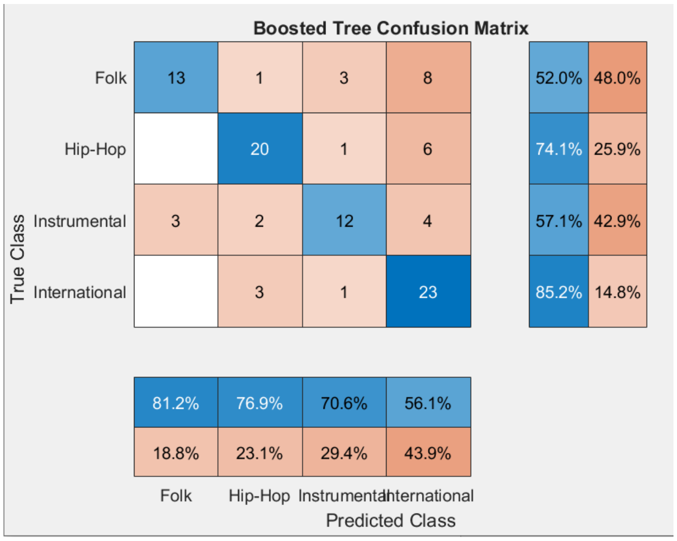

Fig. 4: Top importances identified by Boosted Trees.

This confusion matrix assesses a Boosted Tree classifier's performance. The model struggles with Folk (only **52.0%** accuracy) and Instrumental (**57.1%**), often misclassifying them as International or Hip-Hop. Hip-Hop performs better (**74.1%** accuracy), while International achieves the highest precision (**85.2%**). The uneven distribution of errors (e.g., 8 Folk samples misclassified as International) suggests feature overlap or insufficient training data for certain genres.

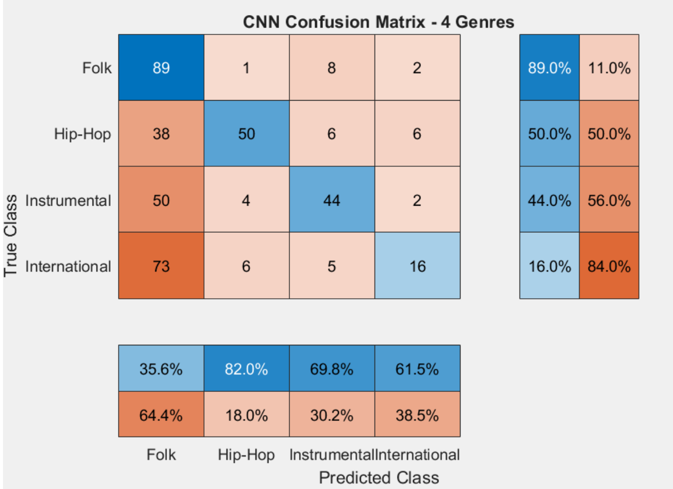

Fig. 5: Confusion matrix of the CNN model on the test dataset.

The Convolutional Neural Network (CNN) confusion matrix reveals significant misclassifications, particularly for International (**84.0%** error) and Instrumental (**56.0%** error). Folk achieves **89.0%** accuracy, but Hip-Hop shows a balanced but suboptimal performance (**50.0%**). The high error rates for International (frequently confused with Folk) suggest that the CNN may struggle with distinguishing subtle acoustic features between these genres, possibly due to dataset limitations or model architecture.

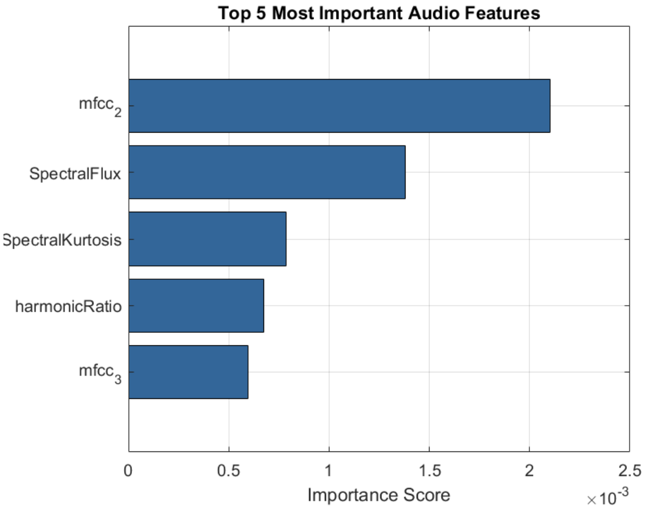

Fig. 6: Top 5 important audio features.

The bar chart ranks the top five most influential audio features for genre classification, measured by importance scores. The top two features—**MFCC$_2$** (Mel-Frequency Cepstral Coefficient) and **SpectralFlux**—dominate, indicating their critical role in capturing timbral and spectral dynamics. **SpectralKurtosis**, **harmonicRatio**,
and **MFCC$_3$** follow, suggesting that harmonicity and spectral shape are also key discriminators. These findings align with prior research, as MFCCs are widely used in audio classification tasks.

VIII. CONCLUSION AND FUTURE WORK

We designed a DSP-based hybrid music genre classifier that effectively balances accuracy and real-time feasibility. By combining handcrafted feature-based models with a CNN trained on mel-spectrograms, we achieve a robust classification system suitable for live applications. Future improvements include:

Extending the model to handle sub-genres and multilingual datasets.

Integrating LSTM layers for temporal dependencies. Optimizing CNN for mobile platforms using pruning or quantization.

REFERENCES

[1] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, B. Kegl, Aggregate features´ and a da b oost for music classification, Machine learning 65 (2006) 473–484.

[2] D. Kostrzewa, M. Ciszynski, R. Brzeski, Evolvable hybrid ensembles for musical genre classification, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, 2022, pp. 252–255.

[3] L. Oudre, Y. Grenier, C. Fevotte, Template-based chord recognition:´ Influence of the chord types., in: ISMIR, 2009, pp. 153–158.

[4] S. Sigtia, S. Dixon, Improved music feature learning with deep neural networks, in: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2014, pp. 6959–6963.

[5] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Transactions on speech and audio processing 10 (5) (2002) 293– 302.

[6] A. L. Uitdenbogerd, International conference on music information retrieval 2003 (2004).