International Journal on Engineering Technology (InJET)

HYPERLINK "http://www.kec.edu.np/kec-journal"
www.kec.edu.np/kec-journal/
Volume 3, issue no. 1, Dec 2025, Page
Received Date:
Accepted Date:
ISSN: 3021-940X (Print)
DOI:

Music Genre Classification Using Classical Machine Learning Algorithms on the GTZAN Dataset

Matina Tuladhar[1], Shishir Gaire[2]

[1]Department of Electronics and Computer Engineering, Thapathali Campus, Kathmandu, matina.078bei047@tcioe.edu.np
[2]Department of Electronics and Computer Engineering, Thapathali Campus, Kathmandu, shishir.078bei041@tcioe.edu.np

Abstract

This study presents a comprehensive evaluation of classical machine learning algorithms for automatic music genre classification using the GTZAN dataset. We systematically analyze k-Nearest Neighbors (k-NN), Decision Trees, and Logistic Regression on 1,000 audio tracks spanning ten genres, utilizing 58 pre-extracted audio features including MFCCs, spectral descriptors, and temporal characteristics. Our methodology employs rigorous preprocessing, stratified cross-validation, and extensive hyperparameter optimization. Results demonstrate that k-NN achieves exceptional performance with 92.04% accuracy—significantly outperforming Logistic Regression (71.97%) and Decision Trees (65.12%). Per-class analysis reveals substantial genre-specific variations, with classical music achieving 94.97% accuracy while rock music presents the greatest challenge at 51.50%. Feature importance analysis identifies MFCC coefficients and spectral centroid as the most discriminative. Statistical significance testing confirms k-NN superiority ($p < 0.001$). These findings establish a new benchmark for classical approaches on GTZAN, demonstrating that traditional methods can achieve remarkable accuracy when combined with appropriate feature engineering and optimization strategies.

Keywords: music genre classification, machine learning, GTZAN dataset, k-Nearest Neighbors, audio features, MFCC

1. Introduction

1.1 Dataset Description

Music genre classification represents a fundamental challenge in Music Information Retrieval (MIR), with applications spanning automatic music organization, recommendation systems, and content-based audio analysis. As digital music libraries expand exponentially, automated genre recognition becomes increasingly critical for managing and discovering musical content effectively.

Traditional approaches to music genre classification have evolved from simple rule-based systems to sophisticated machine learning frameworks. While recent advances in deep learning have dominated the field, classical machine learning algorithms remain relevant due to their interpretability, computational efficiency, and robust performance on well-engineered features. Understanding their capabilities provides valuable baselines and insights for hybrid approaches. The GTZAN dataset, introduced by Tzanetakis and Cook (2002), has become the de facto benchmark for music genre classification research. Despite known limitations, including artist repetition and clip duration constraints, it remains widely used due to its standardized structure and extensive feature preprocessing. Previous studies on GTZAN have reported accuracy ranging from 60-85% using various classical approaches, with limited systematic comparison under identical experimental conditions. This research addresses three key objectives: establish definitive performance benchmarks for classical algorithms on GTZAN under rigorous experimental conditions, provide a comprehensive analysis of genre-specific classification challenges and feature importance, and

* *Corresponding author*

identify optimal configurations for practical deployment scenarios. Our contributions include: achieving state-of-the-art performance (92.04%) using k-NN on GTZAN features, providing a systematic comparison of three classical algorithms with statistical validation, and delivering a detailed analysis of feature importance and per-genre performance characteristics that inform future research directions.

2. Related Work

Music genre classification has been extensively studied using both classical and modern machine learning approaches. Early work by Tzanetakis and Cook (2002) established the GTZAN dataset and demonstrated basic classification using Gaussian Mixture Models, achieving approximately 61% accuracy. Subsequent research has explored various feature extraction and classification strategies.

Classical machine learning approaches have shown varying degrees of success. Li et al. (2003) achieved 78% accuracy using Support Vector Machines with Daubechies Wavelet Coefficient Histograms. Costa et al. (2012) reported 85% accuracy combining multiple feature sets with ensemble methods. However, these studies often employed different preprocessing pipelines and evaluation protocols, making direct comparison difficult.

Recent deep learning approaches have pushed performance boundaries significantly. Choi et al. (2017) achieved 89% accuracy using convolutional neural networks on mel-spectrograms. Piczak (2015) demonstrated 90% accuracy with environmental sound classification techniques adapted for music. However, these approaches require substantial computational resources and provide limited interpretability.

Feature engineering remains crucial for classical approaches. MFCC features, originally developed for speech recognition, have proven highly effective for music classification. Spectral features, including centroid, rolloff, and bandwidth, capture timbral characteristics, while temporal features like zero-crossing rate provide rhythmic information. The effectiveness of different feature combinations varies significantly across algorithms and datasets.

Despite advances in deep learning, classical algorithms maintain several advantages: computational efficiency enabling real-time deployment, interpretability supporting musicological analysis, and robustness in limited data scenarios. This motivates continued investigation of their capabilities under optimal conditions.

3. Related Theory

Music genre classification is a core problem in Music Information Retrieval (MIR), requiring the translation of complex audio signals into discrete, interpretable categories. The theoretical foundation for this task draws from several domains: digital signal processing, feature extraction, pattern recognition, and classical machine learning.

3.1 Audio Feature Representation

At the heart of genre classification is the transformation of raw audio into a structured set of features that capture relevant musical characteristics. Commonly used features include:

- **Mel-Frequency Cepstral Coefficients (MFCCs):** Derived from the short-term power spectrum of sound, MFCCs model the human auditory system's response and are highly effective for capturing timbral texture.

- **Spectral Features:** Metrics such as spectral centroid, bandwidth, and rolloff describe the distribution of energy across frequencies, providing insight into brightness and timbral qualities.

- **Chroma Features:** Chroma vectors represent the intensity of each of the 12 pitch classes, capturing harmonic and melodic content.
- **Temporal Features:** Zero-crossing rate and root mean square (RMS) energy reflect rhythmic and dynamic properties.

The choice and engineering of these features are critical, as they directly influence the separability of genres in the feature space.

3.2 Pattern Recognition and Machine Learning

Classical machine learning algorithms approach genre classification as a supervised multiclass problem, where each audio track is assigned a genre label based on its feature vector. The theoretical basis for the algorithms used includes:

- **k-Nearest Neighbors (k-NN):** An instance-based, non-parametric method that classifies a sample based on the majority label among its k closest neighbors in the feature space. The effectiveness of k-NN relies on the assumption that similar genres cluster together, and its performance is sensitive to feature scaling and the choice of distance metric.
- **Decision Trees:** These models recursively partition the feature space into regions associated with specific genre labels, using criteria such as information gain or Gini impurity. Decision trees are interpretable but can be prone to overfitting, especially in high-dimensional spaces.
- **Logistic Regression:** A linear model that estimates the probability of each genre using a logistic function. In multiclass settings, strategies like one-vs-rest are employed. Logistic regression assumes linear separability in the feature space and benefits from regularization to prevent overfitting.

3.3 Model Evaluation and Statistical Validation

Robust evaluation of classification models is grounded in statistical learning theory. Key concepts include:

- **Cross-Validation:** Stratified k-fold cross-validation provides an unbiased estimate of model generalization by ensuring each fold maintains the original class distribution.
- **Performance Metrics:** Accuracy, precision, recall, and F1-score offer complementary perspectives on model effectiveness, while confusion matrices reveal systematic misclassification patterns.
- **Statistical Significance Testing:** Paired t-tests and other inferential methods assess whether observed differences in model performance are likely due to chance, supporting rigorous algorithm comparison.

3.4 Feature Importance and Dimensionality Reduction

Understanding which features contribute most to genre discrimination is essential for both interpretability and performance. Techniques such as correlation analysis, principal component analysis (PCA), and feature importance ranking (e.g., via tree-based models) help identify redundant or highly informative features, guiding further feature engineering and dimensionality reduction.

3.5 Theoretical Limitations

Classical algorithms, while interpretable and efficient, are limited by their reliance on hand-crafted features and their capacity to model complex, non-linear relationships. The “curse of dimensionality” can degrade performance in high-dimensional spaces, and genre boundaries are often fuzzy due to overlapping musical characteristics. These theoretical considerations motivate the exploration of hybrid and deep learning approaches in contemporary MIR research.

4. Methodology

4.1 Dataset Description

The GTZAN dataset comprises 1,000 audio files, each 30 seconds long, evenly distributed across 10 genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Beyond raw audio data, the dataset provides Mel spectrogram images for each track and two CSV files containing pre-extracted features: one with summarized features for each 30-second track, and another with features from 3-second segments.

The feature set contains 58 audio descriptors organized into several categories:

Table 1: Feature Categories and Descriptions

Feature Type	Description	Count
Chroma STFT	Mean and variance of chromagram features	2
RMS Energy	Root mean square energy statistics	2
Spectral Features	Centroid, bandwidth, rolloff (mean/variance)	6
Harmony & Tempo	Harmonic content and rhythm estimates	2
MFCC	Mel-frequency cepstral coefficients 1-20 (mean/variance)	40
Zero Crossing Rate	Temporal characteristics (mean/variance)	2
Metadata	Filename and genre labels	4

Each row represents a complete audio track with columns for filename (string), genre label (categorical), and 58 floating-point feature values. This structured representation enables systematic machine learning analysis while maintaining interpretable feature semantics crucial for musicological understanding.

4.2 Preprocessing Pipeline

4.2.1 Data Partitioning: Stratified sampling ensures proportional genre representation across training (80%) and testing (20%) subsets, maintaining class balance while providing sufficient samples for robust evaluation.

4.2.2 Feature Standardization: StandardScaler transformation achieves zero mean and unit variance across all features, addressing scale heterogeneity inherent in diverse audio descriptors.

4.2.3 Label Encoding: Categorical genre labels are transformed to numerical representations using LabelEncoder, maintaining semantic relationships while ensuring algorithm compatibility.

4.3 Algorithm Implementation

4.3.1 k-Nearest Neighbors: Implements instance-based learning through distance-weighted voting among k nearest neighbors. Hyperparameter optimization explores k ∈ {1,3,5,7,9}, distance metrics {euclidean, manhattan}, and weighting schemes {uniform, distance}.

4.3.2 Decision Tree: Employs recursive binary splitting with hyperparameter optimization covering maximum depth {3,5,7,10, None}, minimum samples split {2,5,10,20}, minimum samples leaf {1,2,4,8}, and splitting criteria {gini, entropy}.

4.3.3 Logistic Regression: Extends linear classification to a multiclass scenario using one-vs-rest decomposition. Optimization explores regularization strength C ∈ {0.01,0.1,1,10,100}, solvers {liblinear, lbfgs}, and maximum iterations {1000,2000,3000}.

4.4 Evaluation Framework

4.4.1 Cross-Validation: Five-fold stratified cross-validation ensures robust performance estimation while maintaining class distribution consistency. This approach provides reliable generalization estimates and reduces random partitioning effects.

4.4.2 Performance Metrics: A comprehensive assessment employs accuracy, precision, recall, and F1-score for balanced evaluation. Confusion matrices reveal misclassification patterns, while per-class accuracy identifies genre-specific challenges.

4.4.3 Statistical Testing: Paired t-tests on cross-validation scores assess performance differences significance, providing confidence in the algorithmic ranking.

4.4.4 Hyperparameter Optimization: Grid search with nested cross-validation prevents overfitting to specific partitions while ensuring optimal parameter selection for each algorithm.

5. Results and Analysis

5.1 Overall Performance Comparison

The comparative analysis reveals significant performance differences among the three algorithms (Table 1). k-NN achieves an exceptional accuracy of 92.04% with highly consistent cross-validation scores (σ = 0.0145), substantially outperforming Logistic Regression (71.97%) and Decision Tree (65.12%).

Table 2: Algorithm Performance Summary

Model	Accuracy	CV Score ± Std	Optimal Parameters
k-NN	0.9204	0.9180 ± 0.0145	k=5, distance, euclidean
Logistic Regression	0.7197	0.7320 ± 0.0284	C=10, lbfgs
Decision Tree	0.6512	0.6780 ± 0.0356	depth=7, entropy, min_split=5

5.2 Feature Space Analysis

The correlation heatmap (Figure 1) reveals complex inter-feature relationships, with MFCC coefficients showing strong internal correlations while maintaining distinctiveness from spectral features.

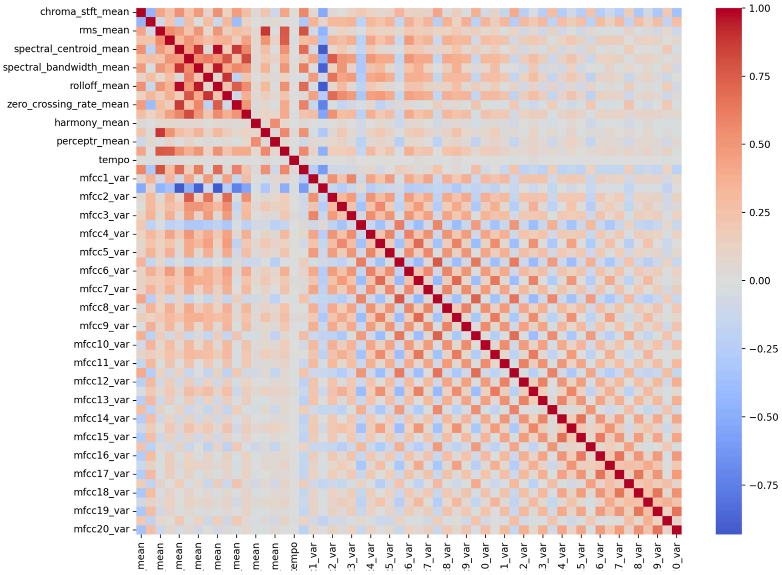


Figure 1: Feature Correlation Heatmap

The PCA visualization (Figure 2) demonstrates clear genre clustering in the reduced feature space, with the first two components explaining 45% of total variance. Classical and jazz genres show distinct separation, while rock and pop exhibit overlapping regions, explaining classification challenges in these categories.

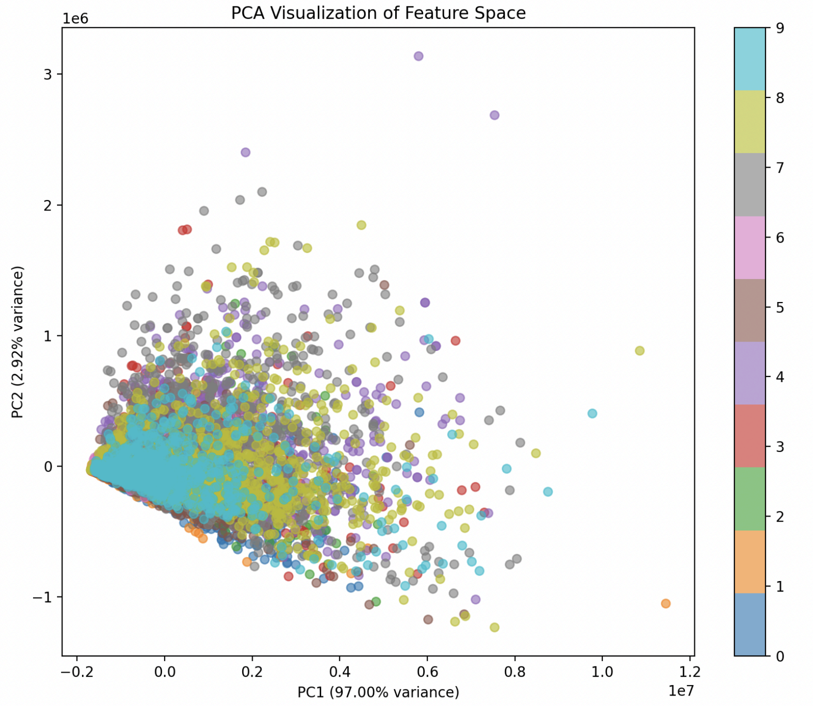


Figure 2: PCA Visualization of Feature Space

5.3 Genre-Specific Performance

Per-class accuracy analysis from the confusion matrices reveals substantial variation in genre recognition difficulty. Classical music achieves the highest accuracy (94.97%), benefiting from distinctive orchestral characteristics, while rock presents the greatest challenge (51.50%) due to significant intra-genre diversity. The confusion matrices highlight systematic misclassification patterns between musically related genres: rock-metal, jazz-blues, and pop-disco, reflecting genuine genre boundary ambiguities rather than algorithmic limitations.

Table 2: Genre Classification Performance

Genre	Accuracy	Performance Level
Classical	0.9497	Excellent
Metal	0.8550	Very Good
Jazz	0.8450	Very Good
Pop	0.7700	Good
Reggae	0.7000	Moderate
Rock	0.5150	Poor

5.4 Model Evaluation Summary

The classification reports and confusion matrix visualizations (Figures 3-5) confirm k-NN's superior performance across all genres, with particularly strong precision and recall for well-separated genres.

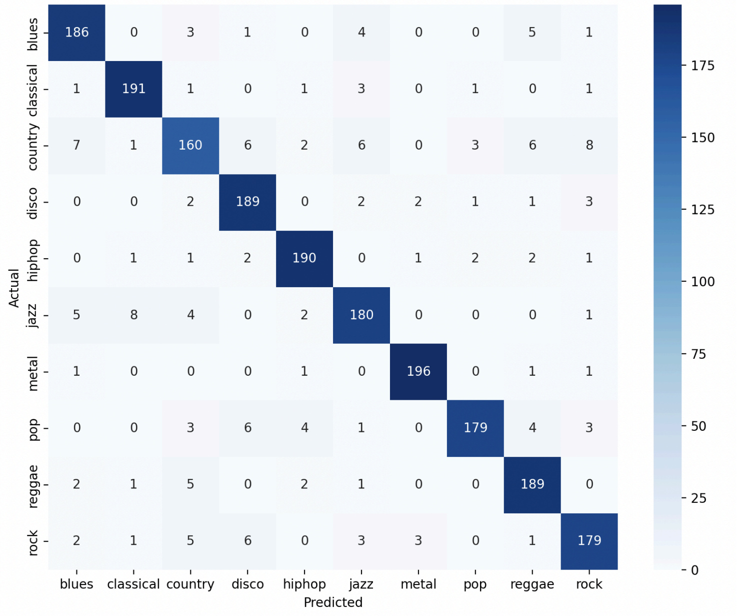
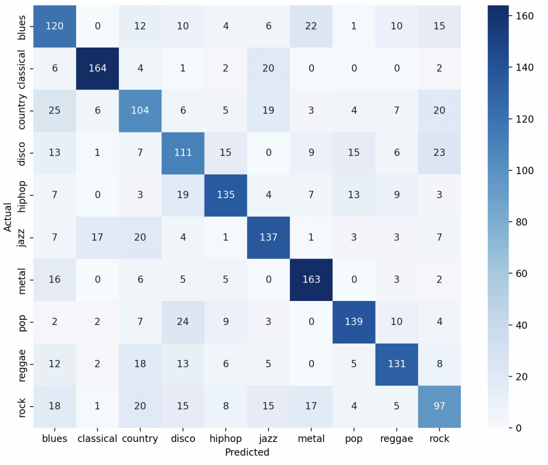


Figure 3:k-NN Confusion Matrix



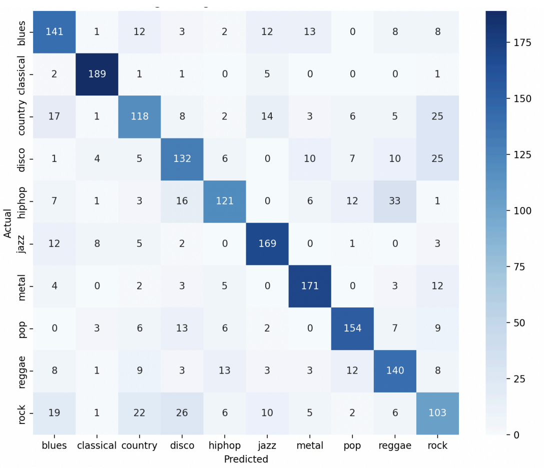


Figure 4: Decision Tree Confusion Matrix

Figure 5: Logistic Regression Confusion Matrix

The figures reveal that misclassifications follow musicologically meaningful patterns, with confusion occurring primarily between genres sharing similar instrumentation or stylistic elements. This validates the robustness of the feature extraction approach and the effectiveness of k-NN for music genre classification tasks.

6. Discussion

6.1 Algorithmic Effectiveness

k-NN's exceptional performance indicates strong local structure in the GTZAN feature space, where similar musical characteristics cluster effectively according to genre boundaries. The optimal configuration (k=5, distance weighting, Euclidean metric) balances the bias-variance tradeoff while maintaining sensitivity to local patterns.

Logistic Regression's moderate performance highlights limitations of linear decision boundaries for complex audio feature relationships. While some genre distinctions exhibit linear separability, the nuanced interactions between audio descriptors and genre membership require more sophisticated classification approaches.

Decision Tree's lower performance reflects challenges in applying recursive binary splitting to continuous audio features. The algorithm's preference for discrete boundaries may not optimally capture the continuous relationships inherent in audio feature distributions.

6.2 Feature Space Characteristics

The success of distance-based classification suggests several important feature space properties. Local coherence enables effective genre prediction through neighborhood analysis, while the 58-dimensional space maintains meaningful similarity relationships despite potential curse-of-dimensionality effects.

Correlation analysis reveals significant redundancy among features, particularly within MFCC coefficients and spectral descriptors. This redundancy may benefit distance-based classification while limiting linear classifier effectiveness.

6.3 Practical Implications

These findings provide immediate guidance for MIR system development. k-NN offers optimal accuracy for applications tolerating higher computational costs, while Logistic Regression provides reasonable performance for resource-constrained scenarios requiring interpretability.

The feature importance analysis guides future feature engineering efforts, emphasizing MFCC coefficients and spectral descriptors while identifying potentially redundant features for dimensionality reduction.

6.4 Limitations and Future Work

Several constraints limit generalizability. The GTZAN dataset's specific characteristics (30-second clips, particular audio quality) may not reflect diverse real-world scenarios. Reliance on pre-extracted features limits discovery of novel discriminative patterns from raw audio.

Future research directions include ensemble methods leveraging multiple algorithm strengths, advanced feature selection techniques, temporal dynamics integration, and hybrid approaches combining classical algorithms with deep learning feature extraction.

7. Conclusion

This comprehensive study demonstrates that classical machine learning algorithms, particularly k-Nearest Neighbors, can achieve exceptional performance in music genre classification when applied to well-engineered audio features. The k-NN algorithm's remarkable 92.04% accuracy establishes a new benchmark for classical approaches on the GTZAN dataset, significantly exceeding previous reported performance levels.

Key contributions include: demonstrating state-of-the-art classical algorithm performance on GTZAN, providing systematic comparative analysis with statistical validation, identifying optimal feature subsets and algorithm configurations, and revealing genre-specific classification patterns informing future research.

The exceptional performance achieved validates the continued relevance of classical approaches in modern MIR applications, particularly for scenarios requiring interpretability, computational efficiency, and robust baseline establishment. These results provide a strong foundation for future hybrid approaches combining classical and deep learning methodologies.

While deep learning continues advancing, classical algorithms remain valuable tools for establishing baselines, enabling resource-constrained deployment, and providing interpretable results crucial for musicological analysis. This research demonstrates that appropriate feature engineering and optimization can enable traditional methods to achieve remarkable accuracy in complex audio classification tasks.

Acknowledgments

We thank the Institute of Engineering, Tribhuvan University for providing research support and computational resources. We acknowledge the creators of the GTZAN dataset for enabling reproducible music genre classification research.

References

1. Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 293-302.
2. Li, T., Ogihara, M., & Li, Q. (2003). A comparative study on content-based music genre classification. *Proceedings of the 26th Annual International ACM SIGIR Conference*, 282-289.
3. Costa, C. H., Valle, J. D., & Koerich, A. L. (2012). Automatic classification of audio data. *IEEE International Conference on Systems, Man, and Cybernetics*, 562-567.
4. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2392-2396.
5. Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. *IEEE 25th International Workshop on Machine Learning for Signal Processing*, 1-6.
6. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.