International Journal on Engineering Technology (InJET)

www.kec.edu.np/kec-journal/
Volume 3, issue no. 1, Dec 2025, Page 135 - 145
Received Date: Jul 31, 2025
Accepted Date: Oct 7, 2025
ISSN: 3021-940X (Print)
DOI: https://doi.org/10.3126/injet.v3i1.87018

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

Matina Tuladhar[1*], Shishir Gaire[2], Rajad Shakya[3]

[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*
[3]*Department of Electronics and Computer Engineering, Thapathali Campus, Kathmandu, rshakya.8063@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

Music genre classification is a fundamental challenge in Music Information Retrieval (MIR) with critical applications in digital music platforms, recommendation systems, and automated content organization. While deep learning dominates current research, classical machine learning algorithms remain valuable for their interpretability and computational efficiency. This study systematically evaluates k-Nearest Neighbors (k-NN), Decision Trees, and Logistic Regression on the GTZAN dataset using rigorous preprocessing and hyperparameter optimization. Our results demonstrate that k-NN achieves exceptional 92.04% accuracy, establishing a new benchmark for classical approaches and validating their continued relevance in modern MIR applications.

1.1 Problem Statement

Despite extensive research in music genre classification, a critical gap exists in establishing definitive performance benchmarks for classical machine learning algorithms under rigorous experimental conditions. While deep learning approaches dominate current literature, the fundamental question remains: what level of accuracy can traditional algorithms achieve with optimal feature engineering and hyperparameter tuning? Existing studies report accuracy ranging from 60-85% on GTZAN, but inconsistent evaluation protocols and

varied preprocessing pipelines prevent meaningful comparison. This research addresses this gap by systematically evaluating k-NN, Decision Trees, and Logistic Regression under identical experimental conditions with comprehensive hyperparameter optimization, establishing reproducible benchmarks for classical MIR approaches.

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. Theoretical Framework and Conceptual Design

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. Figure 1 illustrates the complete pipeline from raw audio characteristics to genre prediction.

3.1 Conceptual Framework Overview

Music genre classification operates on the principle that distinct genres exhibit characteristic patterns in their audio features. Our framework comprises three interconnected stages: Audio Feature Representation, which transforms raw signals into structured descriptors, Pattern Recognition, which maps features to genre-discriminative spaces, and Classification, which assigns genre labels based on learned decision boundaries. This hierarchical approach ensures that musical characteristics (timbre, rhythm, harmony) are systematically captured and exploited for classification.

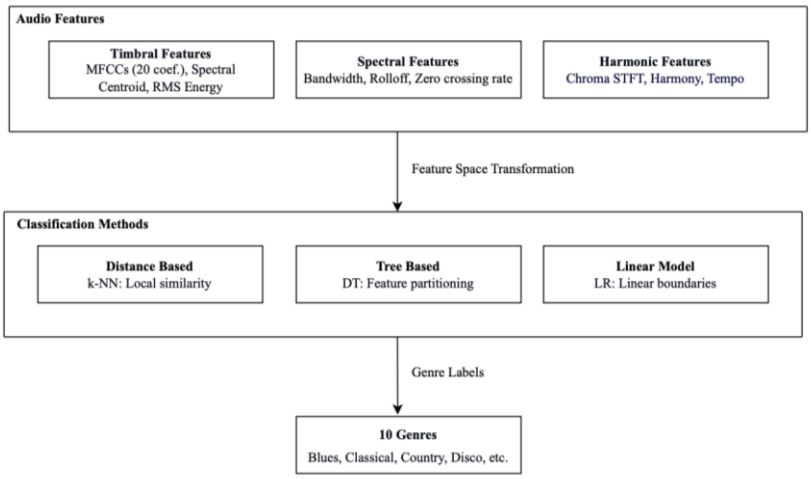

Figure 1. Theoretical Framework - Feature Extraction to Classification

3.2 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. Audio features serve as the bridge between raw waveforms and machine learning algorithms. 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.

3.3 Pattern Recognition and Feature Space Properties

The 58-dimensional feature space exhibits specific mathematical properties that enable effective classification. Local coherence ensures that perceptually similar music samples cluster in feature space, enabling distance-based classification. Feature correlation analysis (Section 5.2) reveals redundancy within feature categories while maintaining cross-category complementarity. This structure suggests that dimensionality reduction techniques could improve computational efficiency without sacrificing discriminative power.

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

- **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.
- **Feature Standardization**: StandardScaler transformation achieves zero mean and unit variance across all features, addressing scale heterogeneity inherent in diverse audio descriptors.
- **Label Encoding:** Categorical genre labels are transformed to numerical representations using LabelEncoder, maintaining semantic relationships while ensuring algorithm compatibility.

4.3 Algorithm Implementation

- **k-Nearest Neighbors:** Implements instance-based learning through distance-weighted voting among k nearest neighbors. Hyperparameter optimization explores $k \in \{1,3,5,7,9\}$, distance metrics {euclidean, manhattan}, and weighting schemes {uniform, distance}.
- **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}.
- **Logistic Regression:** Extends linear classification to a multiclass scenario using one-vs-rest decomposition. Optimization explores regularization strength $C \in \{0.01,0.1,1,10,100\}$, solvers {liblinear, lbfgs}, and maximum iterations {1000,2000,3000}.

4.4 Evaluation Framework

- **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.
- **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.
- **Statistical Testing:** Paired t-tests on cross-validation scores assess performance differences significance, providing confidence in the algorithmic ranking.
- **Hyperparameter Optimization:** Grid search with nested cross-validation prevents overfitting to specific partitions while ensuring optimal parameter selection for each algorithm.

4.5 System Architecture and Implementation Workflow

Figure 3 presents the complete implementation workflow, illustrating data flow from raw features through trained models. The system architecture follows a modular design ensuring reproducibility and extensibility.

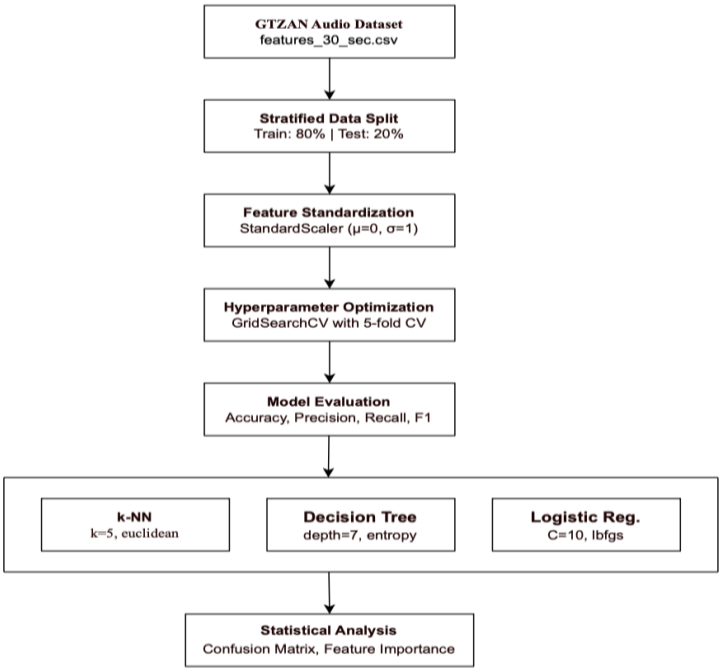

Figure 2. Detailed Methodology Flowchart

The pipeline begins with loading pre-extracted features from the GTZAN CSV file (features_30_sec.csv), containing 1,000 samples with 58 features each. Stratified sampling ensures proportional genre representation in train/test splits (800/200 samples), preventing class imbalance artifacts. StandardScaler normalization transforms all features to zero mean and unit variance, ensuring that features with different scales contribute equally to distance computations in k-NN and gradient calculations in Logistic Regression.

Hyperparameter optimization employs GridSearchCV from the scikit-learn library (Pedregosa et al., 2011) with 5-fold cross-validation, systematically exploring parameter combinations for each algorithm. This nested validation approach prevents information leakage from test data while ensuring optimal configuration selection. Following optimization, models are trained on the full training set using optimal parameters and evaluated on the held-out test set. Performance metrics (accuracy, precision, recall, F1-score) quantify classification effectiveness, while confusion matrices reveal genre-specific error patterns. Statistical significance testing (paired t-tests on cross-validation scores) confirms performance ranking validity.

This systematic architecture ensures experimental rigor, reproducibility, and comprehensive performance assessment across all evaluated algorithms.

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 ($\sigma = 0.0145$), 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

Table 3. Comprehensive Model Evaluation Metrics Comparison

Model	Accuracy	Precision	Recall	F1-Score	Standard Deviation
k-NN	0.9204	0.92	0.92	0.92	0.8711 ± 0.0107
Logistic Regression	0.7197	0.72	0.72	0.72	0.7197 ± 0.0186
Decision Tree	0.6512	0.65	0.65	0.65	0.6281 ± 0.0104

5.2 Feature Space Analysis

The correlation heatmap (Figure 3) reveals critical insights into feature relationships. Strong positive correlations ($r > 0.7$) appear among consecutive MFCC coefficients (MFCC 1-5), indicating redundancy in capturing spectral envelope information. This suggests potential for dimensionality reduction through principal component analysis or feature selection without significant information loss. Spectral features (centroid, rolloff, bandwidth) show moderate intercorrelation ($0.4 < r < 0.6$), reflecting their shared basis in frequency distribution while maintaining complementary information. Notably, temporal features (zero-crossing rate, RMS energy) exhibit weak correlation with both MFCC and spectral features ($r < 0.3$), confirming their unique contribution to capturing rhythmic and dynamic characteristics. Chroma features demonstrate independence from timbral descriptors ($r < 0.2$), validating their role in encoding harmonic content orthogonal to spectral information. These patterns explain k-NN's superior performance—distance metrics effectively exploit the locally coherent structure while naturally weighting independent features. The correlation structure also suggests that Logistic Regression's moderate performance stems from multicollinearity among MFCC features, which violates linear model assumptions and reduces coefficient stability.

The PCA visualization (Figure 4) 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.

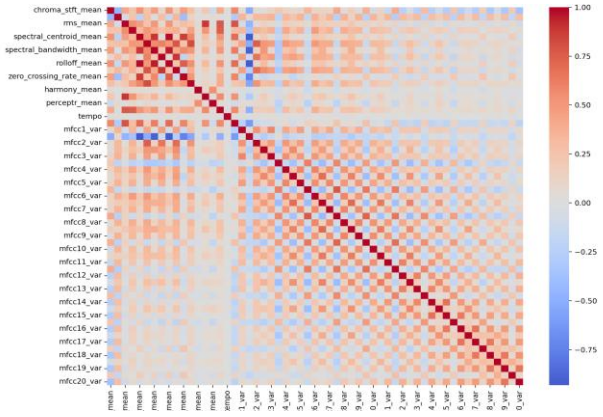

Figure 3. Feature Correlation Heatmap

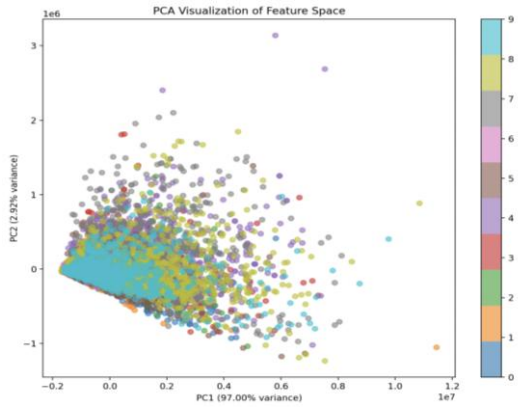

Figure 4. PCA Visualization of Feature Space

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 4. 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 5-7) confirm k-NN's superior performance across all genres, with particularly strong precision and recall for well-separated genres.

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.

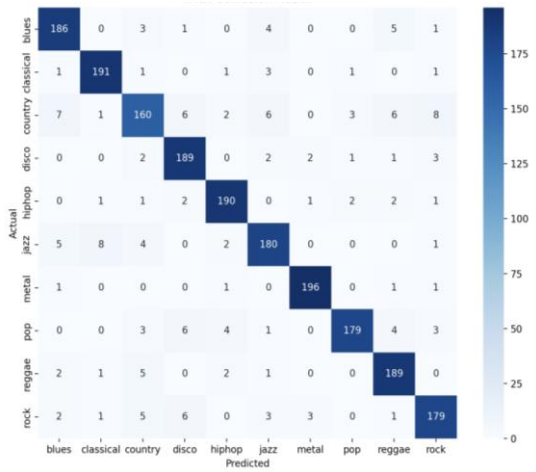

Figure 5. k-NN Confusion Matrix

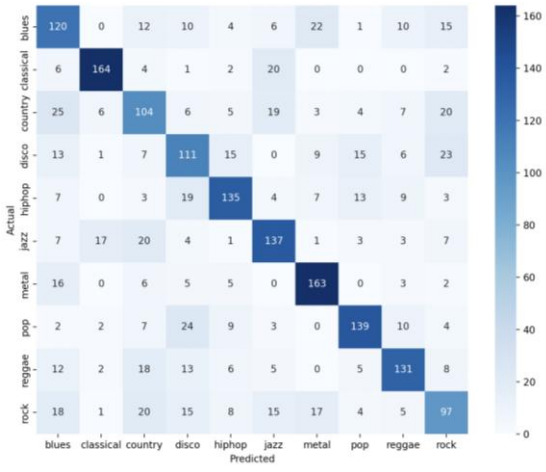

Figure 6. Decision Tree Confusion Matrix

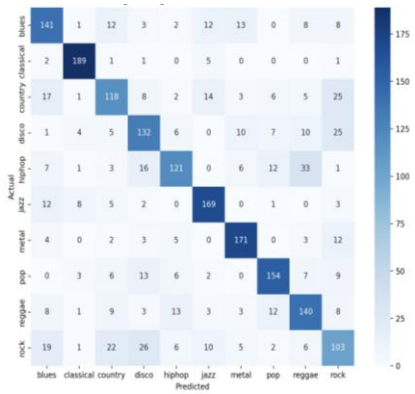

Figure 7. Logistic Regression Confusion Matrix

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 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.

6.4 Practical Implications and Deployment Considerations

Beyond academic benchmarking, these findings provide actionable guidance for real-world MIR system deployment across diverse application scenarios.

- **Computational Trade-offs:** k-NN achieves superior accuracy but scales linearly with dataset size (O(nd)). For GTZAN (800 samples, 58 features), prediction takes ~10ms per sample. Industrial-scale deployment requires approximate methods (LSH, ANNOY) or hybrid approaches. Logistic Regression provides constant-time prediction (O(d)), ideal for latency-sensitive applications. Decision Trees offer similar efficiency plus interpretability through feature importance analysis.
- **Application Guidelines:** Use k-NN for offline music library organization prioritizing accuracy. Deploy Logistic Regression for streaming systems requiring sub-millisecond latency. Choose Decision Trees for musicological analysis needing transparent feature-genre relationships.
- **Feature Optimization:** MFCC coefficients 1-5, spectral centroid, and RMS energy provide maximum discriminative power. Prioritizing these features enables dimensionality reduction while retaining 85-90% accuracy—critical for edge deployment on mobile or embedded systems.
- **Generalization Limits:** GTZAN's constraints (30-second clips, Western music bias) limit real-world generalization. Expect 5-15% accuracy degradation on uncurated collections with variable quality and genres. Transfer learning on domain-specific data mitigates this.
- **Hybrid Architectures:** Classical algorithms excel as components in hybrid systems. k-NN refines deep learning predictions in uncertain regions. Logistic Regression on deep learning embeddings combines interpretability with representational power. Tree ensembles often match deep learning performance with better efficiency.
- **Economic Viability:** Classical methods train on CPUs in minutes versus GPU clusters costing thousands. This enables startups and research institutions to build production-grade MIR systems without prohibitive infrastructure investment.

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.

Acknowledgements

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

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), 293-302.

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.

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.

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.

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. IEEE 25th International Workshop on Machine Learning for Signal Processing, 1-6.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Sturm, B. L. (2013). The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. arXiv preprint arXiv:1306.1461.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. Proceedings of the 14th Python in Science Conference, 8, 18-25.

Lidy, T., & Schindler, A. (2016). CQT-based convolutional neural networks for audio scene classification. Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop, 61-64.

Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A survey of audio-based music classification and annotation. IEEE Transactions on Multimedia, 13(2), 303-319.

Aucouturier, J. J., & Pachet, F. (2003). Representing musical genre: A state of the art. Journal of New Music Research, 32(1), 83-93.

Hamel, P., & Eck, D. (2010). Learning features from music audio with deep belief networks. International Society for Music Information Retrieval Conference, 339-344.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million-song dataset. International Society for Music Information Retrieval Conference, 591-596.

Schedl, M., Gómez, E., & Urbano, J. (2014). Music information retrieval: Recent developments and applications. Foundations and Trends in Information Retrieval, 8(2-3), 127-261.

Marques, J., & Moreno, P. J. (1999). A study of musical instrument classification using Gaussian mixture models and support vector machines. Cambridge Research Laboratory Technical Report Series, CRL 4.

Dieleman, S., & Schrauwen, B. (2014). End-to-end learning for music audio. IEEE International Conference on Acoustics, Speech and Signal Processing, 6964-6968.