Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable AutoML-Driven AI-Assisted Traditional Music Composition and Genre Classification via MFCC-Based Acoustic Feature Learning

Ting Zhang

moonshinnlz@outlook.com

Nanchang University

Yang Li

Nanchang University

Article

Keywords: Music genre classification, mel-frequency cepstral coefficients (MFCC), explainable AutoML, ensemble learning, AI-Assisted music composition

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10905943/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explainable AutoML-Driven AI-Assisted Traditional Music Composition and Genre Classification via MFCC-Based Acoustic Feature Learning

Ting Zhang[1,*], and Yang Li[2]

Academy of Arts, Shangluo University, The Gatehouse of a family member of a financial school in Shangzhou [1]District Square, Shangluo, Shaanxi province, 726000, China; moonshinnlz@outlook.com

[2]Department of Student Affairs, CPC Committee, Nanchang University, 999 Xuefu Avenue, Honggutan District, Nanchang City, Jiangxi Province, 330000, P.R.China; moontds1989@163.com

*Correspondence author: moonshinnlz@outlook.com

Abstract: Accurately classifying music genres is important for music information retrieval, recommendation systems, and traditional music heritage conservation. A high-accuracy interpretable music genre classification system was designed. This study proposes an artificial intelligence-aided approach that combines MFCC-based acoustic feature learning, spectral and harmonic descriptors, explainable automated machine learning (AutoML), optimised ensemble learning, and SHAP-based explainability for traditional music genre classification and composition. It was implemented based on 5-fold cross validation with bagging, stacking, and weighted ensemble learning on the considered music genre dataset known as the GTZAN. Furthermore, genre-specific prototypes of acoustics are learned to support music generation based on AI assistance, which can provide feature targets that are representative of the music style for generating music that is consistent with the music style. The experimental results revealed that the WeightedEnsemble_L3 ensemble model achieved a maximum accuracy of 94.59%, whereas the individual classifiers were classified at a lower rate. SHAP makes it easier to identify the most important acoustic characteristics for the model, increasing its transparency and interpretability. Moreover, knowledge-based acoustic prototypes allow the acousability of the generated music to be quantified in comparison to the given music-making genres and transfer this context into AI-assisted music generation. The suggested framework offers a genuine, comprehensible, and viable solution for traditional music recognition and supports culturally motivated AI aids to music production.

Keywords: Music genre classification; mel-frequency cepstral coefficients (MFCC); explainable AutoML; ensemble learning; AI-Assisted music composition

1. Introduction

As detailed by Ding [1], over a century of musical catalogues have been digitised on contemporary cloud-streaming networks, and automated classification has become a key component of modern pipelines for the management and organisation of music. Mycka and Mańdziuk [2] cited the need to create reliable algorithms that can be used to automatically extract stylistic metadata to optimise user recommendation systems, automatic playlist indexing, and digital copyright management networks. Budyputra et al. [3] report that modern deep architectures achieve remarkable taxonomy constellations for western mainstream music, but under such conditions where the music structures are not “western”, but rather belong to ethnic and classical world music traditions or heritage music from Indonesia, the structural complexity of the non-western scale causes significant taxonomic bottlenecks. In Moysis et al.'s ethnomusicological assessment [4], it is common for Music Information Retrieval (MIR) systems to fail on localised acoustic traditions owing to the lack of sensitivity in conventional Music Information Retrieval (MIR) architectural approaches to the specialised instrumentation or the tonalities used are not tempered.

However, as discussed in [5], it is crucial for deep neural networks for sound analysis to maintain an optimal trade-off between the mathematical accuracy of the prediction and interpretability for diagnosis. However, as Liang points out in [6], most of the recent more successful ones are essentially “black boxes” as algorithm models with very high scores during testing and validation but containing secret algorithms. As Naidu et al. [7] stated, this systematic opacity is problematic for clinical, legal, and musical applications, given the lack of transparency

regarding whether any architecture is settling towards the authentic acoustic "structural markers" or rather towards "recording artifacts" in the training sets. Therefore, in many cases, XAI frameworks must be rolled out in their automated workflows to trace back the raw internal decision structure to physical and human-verifiable features.

To obtain strong baseline feature representations to reduce environmental noise, McLoughlin et al. [8] showed that it is important to represent both multiscale temporal dynamics and localised spectral energy structures. Raw waveform processors have gained popularity; however, handcrafted spectral transformations are mathematically more powerful in simulating the logarithmic operation of the human acoustic perception mechanism, as pointed out by Rawat et al. [9]. As demonstrated by Zaman et al. [10], adding rhythmic and harmonic centroids with Mel-Frequency Cepstral Coefficients (MFCCs) to downstream classifiers results in a highly dense feature space that remains stable. Moysis et al. [4] cautioned that the manual treatment of all these complex feature vectors involves excessive manual labour, and suboptimal hyperparameter problems arise when chunking them into manually treated ones in data-poor acoustic traditions.

To overcome these engineering constraints, Abebe et al. [11] showed that it is possible to deploy machine learning frameworks called Automated Machine Learning (AutoML) to systematically characterise the pipeline of data augmentation, balance multiclass representations, and generate elite model topologies autonomously from humans. The ensemble and advanced architectural hybridisation methodologies described by Prajuli et al. [12] can further enhance the statistical stability of these models. In the stacking architectures outlined by Mogonediwa et al. [13], the fusion of diverse classifiers from gradient-boosted trees to deep recurrent networks results in routinely higher generalisation thresholds and a deeply improved immunity from out-of-sample error rates as compared to any one individual network.

In the move from predictive genre classification to generative music composition, Akman et al. [14] conducted a study that addressed the translation of analytical models to creative utilities. Patil et al. [15] note that, if properly retrieved from the system, a particular musical tradition's core feature boundaries, extracted from the musical signals may then be inverted to become the parameter of an active generation engine. Smith and Golcarenarenji [16] claim that explainable feature modeling allows the human creator to remain maximumly in command of the algorithmic synthesis engine by setting up specific “genre prototypes”. This iterative process, as described by Jenifer et al. [17], transforms AI from being a stand-alone after-the-fact labeller into a live part of the creative process to help guide new compositions within the culturally appropriate parameters.

This research directly responds to these basic needs and presents an imaginative, transparent, and powerful traditional music management framework. Based on the identified needs and experiences from the field, the following core research questions (RQs) and hypotheses (H) are articulated in this study to investigate the integration of accurate genre classification, explainability, and generative composition systematically:

RQ1): What is a new way to improve the accuracy and stability of the prediction of complex traditional music structures?

- H1: An optimised ensemble learning framework based on X-AutoML techniques and optimised using high-dimensional MFCC extractions along with spectral descriptors will substantially outperform single deep learning models in multiclass classification tasks.

RQ2: How can music classification make black box models transparent and interpretable for human users?

- H2: The SHAP-based post-hoc explainability metric will correctly identify the weights of the features that were used for a particular prediction while being able to trace automated classification decisions back to structurally explainable acoustic features (e.g. timbral and spectral signatures) without compromising the model accuracy.

RQ3: Can passive predictive classification models be used to inform culturally sensitive AI-guided music composition?

- H3: The system can help model the transformations of learned classification boundaries into explicit prototypes of genre-specific acoustic features, which can then produce measurable geometric feature targets that serve as effective guides for the generative process of traditional music.

The main contributions of this study are outlined as follows:

- Multi-Scale Feature Learning Framework (MSLF): high-dimensional MFCC extractions and rhythmic + harmonic spectrocentroids (RC+HC) for an advanced acoustic parsing pipeline enabling the mapping of subtle tonal boundaries of traditional music tracks.
- We describe and use a stacked model with third-layer weighting optimisations to improve multiclass classification accuracy on a collection of baseline deep models to assess AutoGluon's usefulness and provide examples of its usage.
- We construct math-based explainability metrics in a somewhat principled fashion to capture the weights of our features and then directly discuss such explainability metrics as they relate to music-specific physics metrics in a localised, post-hoc fashion.
- Creative Prototype Synthesis: We transformed the learned classification boundaries into geometric feature targets, which is a new way of evaluating the consistency in structure and directing the culturally faithful generation of traditional music.

2. Literature Review

2.1 Foundations of Music Genre Recognition

The history of Music Genre Recognition (MGR) has evolved from the traditional and limited manual classification of genres to advanced computational intelligence systems [1]. Mycka and Mańdziuk [2] noted that early systems were unable to achieve both processing efficiency and a sense of taxonomy. With the growth of digital streaming catalogues, scalable and accurate recognition systems with transparent decision-making algorithms have become crucial in the field of Music Information Retrieval (MIR) research, as emphasised by Senatori et al. [18]. Deep Learning architectures are a major breakthrough, and they have been integrated within the framework presented by Moysis et al. [19] for an analysis of the acoustic structure with an unprecedented level of detail, both in the time and frequency domains. This paradigm has recently been extended to include more than an audio description, for example, symbolic scores and imagery, and in the system devised by Zhang and Li [20] to tackle taxonomic bottlenecks, this leads to a deep multi-modal system.

Despite this progress in predictions, it is still difficult to make a complex computational match between the outputs of an automated classification process of a high-dimensional class and the subjective cognitive perception of humans, as presented by Guo [5]. Liang [6] entrusted with the duties of music production, a modern deep learning model often functions like a "black box", and there is a lack of transparency surrounding the internal mapping between the model and the physics of the music. Naidu et al. [21] argue that transparent and explainable user-profiling mechanisms are critical for ensuring the validity of the diagnosis and for building user trust in AI-based systems for music curation, retrieval, or even automated recommendation architectures.

2.2 Feature Engineering and Spectral Representation

Although there have been substantial advances in the selection and mathematical extraction of audio descriptors, as established by McLoughlin et al. [8], the selection of descriptors and their mathematical extraction remain key for effective algorithmic audio classification. Following previous work by researchers such as Rawat et al. [9], who developed non-continuous features of audio data, the authors attempted to establish timbral, rhythmic, and pitch classes on a standardised basis. Zaman et al. [10] have demonstrated significant improvements in a polyphonic identification problem using a basic fusion strategy resulting from fusing symbolic representations obtained by MIDI with continuous (spectral) descriptors.

Moysis et al. [4] is a good example, as they noted that Mel-Frequency Cepstral Coefficients (MFCCs) are one of the most prevalent methods because they have the great ability to mathematically model human hearing. It is also noted that traditional spectral representations are highly limited in use if there is a lack of data or large stylistic nuances [22]. To circumvent these limitations, Papaioannou et al. [23] successfully (and robustly) applied automatic

data augmentation pipelines to significantly lower out-of-sample error rates. With the emergence of AutoML architectures, as demonstrated by Abebe et al. [11], researchers have now been able to systematically addressed the challenging task of maximising the utilisation of the audio dataset to be processed, optimising the selection of audio features, balancing multi-class representations, and minimising engineering bottlenecks for performing downstream style-conversion tasks without manual effort.

2.3 Optimization and Explainability Frameworks

Recently, MGR systems have increasingly relied on automated workflows for training while using more advanced interpretability toolboxes to ensure deployment reliability, as proposed by Prajuli et al. [12]. Mogonediwa [13] concluded through methodological investigations that optimised ensemble learning frameworks (particularly those based on voting or consensus) provide tremendous statistical stability compared to isolated deep learning models. However, in the architecture proposed by Kamuni et al. [24], frameworks such as SHAP-LM incorporate explainable AI (XAI) measures to direct automated feature selection, thereby reducing the CPU complexity by up to 75% of the total without compromising the classification ability.

As tracked by Franzoni [25], edge applications and mobile implementations have also recently reaped the rewards of the big związzek with Neural Architecture Search (NAS) which allows for the discovery of extremely compressed, always-on, and energy-efficient audio processors. Comanducci et al. [26] provided evidence that hyperparameter optimisation spaces have been extensively optimised to isolate impulsive sound events from highly unstable and noisy backgrounds. In the context of creative software, Becker et al. [27] demonstrated that the internal configuration of latent spaces, for example, within a variational autoencoder, has come to the fore in recent years, as the internal state that explains what is being generated must be retained in a dynamic and predictable way, enabling humans to exert meaningful control over an AI-powered system of synthesis.

2.4 Neural Architectures and Temporal Patterns

However, the defects of such models, such as the restriction on the model's capacity to capture the complex temporal sequences in music, mean that musical recordings have fallen out of favour since the analysis performed by their models is limited to parallel time. However, as analysed by Akman et al. [14], the capacity of deep learning models to capture complex multi-scale temporal sequences has led to a major paradigm shift in how musical recordings are now computationally processed, with that capacity being limited to parallel time evaluations. Patil et al. [15] through post-hoc XAI techniques on Convolutional Neural Networks (CNNs) validate that these models are learning acoustically salient patterns and not superficial data artefacts. Because traditional interpretability methods were created for and work best with visual media, Smith and Golcarenarenji [16] highlighted the critical need for audio-specific XAI frameworks that are essential for explaining highly dense frequency array data.

When examining raw acoustic wave patterns from a frequency perspective, Jenifer et al. [17] demonstrated that deep neural networks could provide objective physical explanations[17]. Lyberatos et al. [28] described that an attention mechanism enables a transformer model to directly extract long-range semantic patterns from unsegmented streams, whereas Gated Recurrent Units (GRUs) are superior at capturing the high-level rhythmic details of stream segments. Transformer models are typically better at capturing long-range semantic patterns directly from unsegmented streams of data, whereas Gated Recurrent Units (GRUs) are generally more effective at handling highly localised rhythmic nuances. MSGE [29] asserts that it is a strong background for modern streaming platforms, as it removes the old limitations imposed by manual feature engineering.

2.5 Hybrid Models and Ensemble Learning

Sharma [30] showed that the highest classification threshold for various audio corpora when using different neural architectures resulted from hybridisation. Schuller et al. [31] used deep CNN layers with bidirectional gated recurrent units (Bi-GRUs), which process the local spectral frame and capture the macro-temporal sequence of Mel-

spectrograms. To minimise background interference and isolate vocal and instrumental passages, Cataltepe et al. [32] improved this technique by adding subcomponent-level attention mechanisms to it.

As verified by Ramírez and Flores [33], intra-song instance bagging methods can be effectively applied to account for the internal temporal fluidity and improve label stability over varying stanzas of music while keeping parametric dimensions unchanged. Sun et al. [34] performed an experimental evaluation of hundreds of baseline classifier permutations, which revealed that one way of processing highly diverged audio streams is stacking (stacked generalisation). These new ensemble topologies are tailored to the multi-label classification problem, which is gaining increasing attention because musical genres are starting to merge conventional boundaries in modern music [35].

2.6 Specialized Applications and Semantic Mapping

Recent MGR systems are shifting away from the general categorisation of consumers to more specialised systems for tracking copyrights and generative processing pipelines for particular niche traditions, as Wang et al. [36] demonstrate. Speckhard et al. [37] use pre-trained networks (e.g., VGGish) and transfer learning architectures to learn with success the highly complex tonal and melodic shapes in the specialized traditions (e.g., Raga music of northern India). Mendes et al. [38] proposed a combination of structured knowledge graphs and acoustic feature models for detecting complex copyright violations in large multimodal databases.

The deployment of these models comes with technical trade-offs in terms of how these biases are propagated [39], especially when pre-trained audio embeddings do not have access to musical intervals outside the Western calendar. To ensure the accurate use of audio conversion, Chakrabarty et al. [40] proposed a solution that combines Short-Time Fourier Transforms (STFT) with the Inception V3 deep architecture.

2.7 State-of-the-Art Core Trends

In contrast, Wang and Jiu [41] showed that raw acoustic representation is gradually combined with lyrics in the text to form a deep multimodal kernel network for capturing nonlinear structural dependencies. To achieve optimal synergy between channel attention and multi-head cross-attention frameworks, Li and Du [42] used ECAPA-TDNN parallel aggregation architectures and reduced dynamic feature extraction to 32.22 s. To better represent human affect and match it to raw features, Sun [43] showed that current self-supervised transformer models, such as HuBERT, outperform representations based on raw waveforms, with up to 90% classification accuracy in overlapping acoustic environments.

Prompt-based deep networks are reshaping pedagogical and composition processes, as described by Cheng [44] regarding trends in algorithmic generation. Wang and Zhang [45] opened their research into the psychological mechanism of collaborative AI creative tools with a concrete example of visual-app-based systems of real-time feedback of features as an important catalyst for novice composers to trap into a higher level of self-efficacy. To handle such interactions safely, Zhang [46] proposed an active approach to self-regulated LSTM training learning, taking a creative rate and control of the stylistic boundary generation constraint beyond passive labelling and into active and real-time creative interaction with AI.

2.8 Identification of Literature Gaps

Although music genre recognition systems have made tremendous progress in the past few years, several important issues have not been addressed in the existing literature. There is also a distinct lack of active frameworks that connect the dots between predictive inference and creative generation; in particular, classification as an endpoint does not equip the acoustic space to project into workable composite targets valued by the engines. Most active models rely on the disposal of static parameters with only one label and do not consider the natural fluidity of the present ethnologistic reality, which is, by nature, elusive of categorisation and more of a continuum than a taxonomy.

There is a gap in understanding between engineering measures and human audibility, in which computational models tend to focus on the details of spectral changes but neglect the macro-level rhythmic/harmonic aspects of sounds that are important to human listeners. The creation of lightweight, easily understood models that can provide live, instantaneous, and real-time feedback to musicians during improvisation or compositional processes is still in its early stages.

To fill these gaps, the present study presents an optimised AutoGluon-guided ensemble pipeline. Passive classification data can be transformed into explicit multidimensional classification targets of highly representative "genre prototypes" extracted from verified sets of audio data, and can more effectively guide culturally consistent traditional music composition.

3. Methodology

The overall process of the proposed framework is depicted in Figure 1, which includes acoustic feature extraction, XAI-AutoML training, ensemble optimisation, explainability analysis, and AI-assisted music composition.

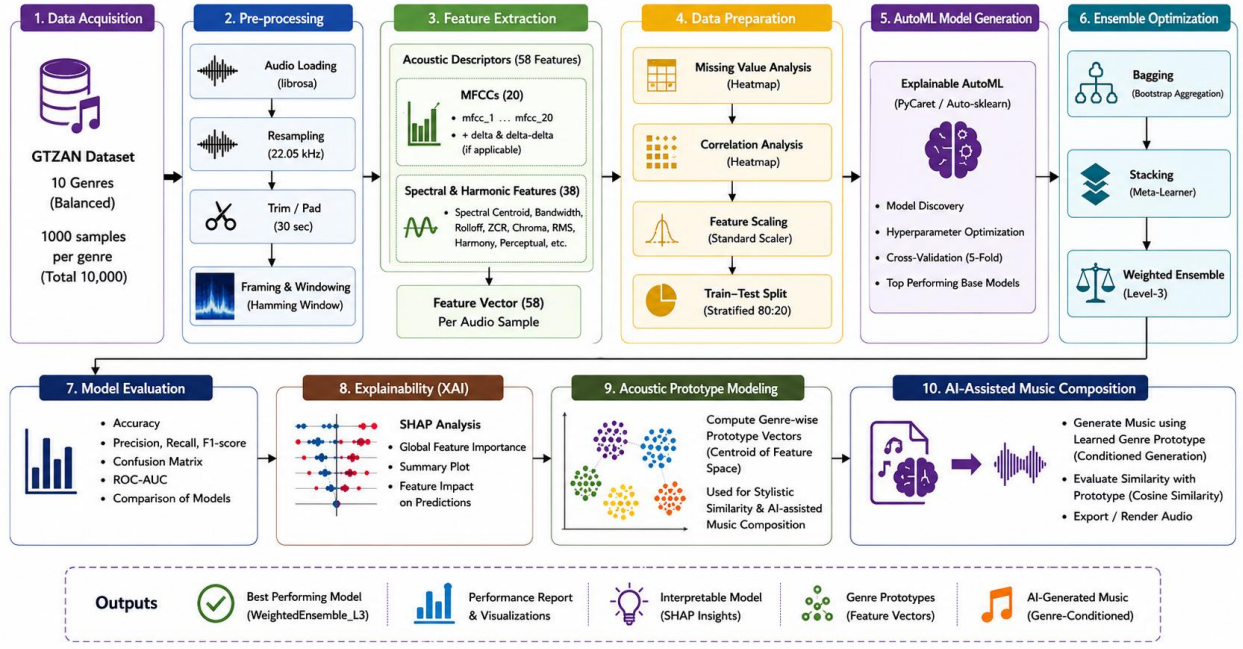


Figure 1. AI-assisted traditional music genre classification and composition using MFCC-based feature learning, Explainable AutoML, ensemble optimisation, SHAP explainability, and genre prototype modelling (as shown in Figure 1) are suitable for traditional music classification. The proposed methodology pipeline for AI-assisted traditional music classification and composition through MFCC-based feature learning, Explainable AutoML, ensemble optimisation, SHAP explainability, and genre prototype modelling are suitable choices for traditional music classification.

3.1 Research Framework

This study proposes an integrated artificial intelligence framework for traditional music genre classification and AI-assisted music composition using MFCC-based acoustic feature learning, Explainable Automated Machine Learning (AutoML), ensemble optimisation, and genre prototype modelling. The framework was designed to simultaneously achieve a high classification accuracy and provide interpretable acoustic representations that can guide music generation systems. Unlike conventional genre classification studies that terminate with predictive inference, the proposed framework extends the learned acoustic feature space to composition-oriented applications by constructing genre-specific acoustic-prototypes. Let the complete dataset be represented as

$$D = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

where $x_i \in R^d$ denotes the acoustic feature vector of the i^{th} music segment, y_i denotes the corresponding genre label, d represents the dimensionality of the feature space, and N denotes the total number of observations. The objective was to learn the following predictive function:

$$f : R_d \rightarrow G \quad (2)$$

that maps acoustic descriptors to a predefined genre set

$$G = \{g_1, g_2, \dots, g_C\} \quad (3)$$

where C denotes the number of genres.

The plan underlying this framework comprises five sequential steps. First, descriptors were extracted and analysed using exploratory data analysis of the acoustic data. Second, procedures are involved in preprocessing and standardising the data to make them suitable for machine learning. Third, multiple candidate models were automatically built and optimised using AutoML. Fourth, ensemble optimisation is achieved by stacking and bagging high-performing learners. Finally, two analysis methods are applied to help interpret and provide guidance for music composition using AI/persona: explainability analysis and genre-prototype learning. The overall structure takes the acoustic content and maps it into discriminative representations of the genre and simultaneously builds a feature space with knowledge that can be applied to music generation.

3.2 Dataset Description

For the experiments, a popular benchmark dataset for music information retrieval studies, the GTZAN Music Genre Classification Dataset (https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification), was used in this study. The dataset contains ten types of music: Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, and Rock. Each audio recording was divided into several three-second clips, to raise the degree of observation and increase the statistical power. The data were organised as follows:

$$D = \{(x_i, y_i)\}_{i}^{19990} \quad (4)$$

where 19990 indicates 19990 segmented audio samples. A set of acoustic parameters describing the acoustic signal corresponding to each sample was used to represent each sample. This is expressed by the feature matrix as follows.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_N d \end{bmatrix} \quad (5)$$

where $N = 9990$ and $d = 58$. The set of competing variables comprised 10 balanced genre types:

$$G = \{Blues, Classical, Country, Disco, HipHop, Jazz, Metal, Pop, Reggae, Rock\} \quad (6)$$

This guarantees that the classification performance is not severely affected by the unbalanced distribution of classes. Therefore, the dataset is a good ground truth for assessing multiclass music genre recognition systems.

3.3 Acoustic Feature Representation

Music genre classification requires a set of acoustic features that can accurately capture the spectral, temporal, harmonic, and timbral features of music signals. Therefore, the proposed framework employs a hybrid feature of MFCCs and complementary spectral features. The feature vector for each music segment is as follows:

$$x_i = [c_1, c_2, \dots, c_d]^T \quad (7)$$

Parameter c_j represents the value of the j^{th} acoustic feature.

3.3.1 Chroma Features

Compensation of the "harmonic content" is shown by chroma descriptors, which are the projections of spectral energy on the 12 semitone classes of the musical octave. The visualization is in the form of a chroma vector that is represented as

$$C = [c_1, c_2, \dots, c_{12}] \tag{8}$$

This is for the arbitrary placement of each part in one pitch class. Chroma features are particularly effective in identifying differences between genres with different structures in their harmonies and progressions.

3.3.2 Root Mean Square Energy

The perceived loudness and intensity of the signal were measured using the Root Mean Square (RMS) energy. It is computed as:

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} s_n^2} \tag{9}$$

where s_n denotes the signal amplitude at sample n. In general, the RMS values of various genres tend to be greater if they have a more aggressive instrument profile and a higher rhythmic structure.

3.3.3 Spectral Centroid

The spectral centroid is the centre of gravity of the frequency spectrum and is a measure of what might be called audio brightness.

$$SC = \frac{\sum_{k=1}^{K} f_k X(k)}{\sum_{k=1}^{K} X(k)} \tag{10}$$

where f_k is the frequency and $X(k)$ is the spectral magnitude. The centroids tended to be higher for more luminous sounds.

3.3.4 Spectral Bandwidth

The measure of the spread of energy around the centroid in the spectrum is called spectral bandwidth.

$$SB = \sqrt{\frac{\sum_{k=1}^{K} (f_k - SC)^2 X(k)}{\sum_{k=1}^{K} X(k)}} \tag{11}$$

This descriptor is applicable to timbral complexity and frequency dispersion.

3.3.5 Spectral Rolloff

The point at or below which the percentage of the total spectral energy is concentrated is designated as the spectral rolloff frequency.

$$\sum_{k=1}^{R} X(k) = 0.85 \sum_{k=1}^{K} X(k) \tag{12}$$

where R denotes the roll off frequency index.

3.3.6 Zero-Crossing Rate

Zero Crossing Rate (ZCR) is a measure of waveform complexity and is defined as

$$ZCR = \frac{1}{2N} \sum_{n=1}^{N} |sign(s_n) - sign(s_{n-1})| \quad (13)$$

where $sign(.)$ represents the sign function.

3.3.7 Mel-Frequency Cepstral Coefficients

MFCCs are the most common acoustic representations used in the present study because they are closely related to what humans hear. The Mel-frequency transformation is given by

$$M(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad (14)$$

where f denotes frequency in Hertz. After mel-filter bank processing and logarithmic compression, the Discrete Cosine Transform generates cepstral coefficients:

$$MFCC_n = \sum_{m=1}^{M} \log(S_m) \cos[\frac{\pi n}{M}] (m - 0.5)] \quad (15)$$

where S_m represents the energy picked up by the m^{th}, Mel filter. MFCCs are an effective feature representation for timbral features and are thus regarded as the main feature representation in the proposed framework.

3.4 Data Preprocessing

Data preprocessing was performed to ensure consistency between the numbers of each acoustic description while simultaneously stabilising the model. Data preprocessing was performed to avoid information leakage and ensure numerical consistency between acoustic descriptors to stabilise the model. Label encoding across genre labels was used to convert the genre labels into numbers. Let

$$G = \{g_1, g_2, \dots, g_{10}\} \quad (16)$$

denote the genre set. A mapping function

$$E: G \rightarrow \{0,1, \dots ,9\} \quad (17)$$

Categorical labels were used and learned as class identifiers using this method. Label encoding was done and then there was standardization of the features for the elimination of differences in numerical scales. This conversion was performed for each feature based on the following:

$$z = \frac{x-\mu}{\sigma} \quad (18)$$

where μ and σ are the mean and standard deviation calculated from the training data. The data were then split into subsets based on stratified sampling as follows:

$$D = D_{train} \cup D_{test} \quad (19)$$

subject to

$$D_{train} \cap D_{test} = \emptyset \quad (20)$$

The chosen parameter was 80:20, which yielded 7,992 and 1, 998 training and test samples, respectively. The amount of each genre in both subsets was preserved in the stratification to ensure an unbiased performance assessment.

3.5 Cross-Validation Strategy

To obtain a reliable estimate of the model performance and to decrease overfitting of the model, stratified five-fold cross validation was integrated into the learning process. The training data were represented as follows:

$$Dtrain = \{(xi, yi)\}_i^{N_t} \tag{21}$$

where $N_t = 7992$. The training set was split into five non-overlapping parts.

$$D_{train} = \bigcup_{k=1}^{5} D_k \tag{22}$$

such that:

$$D_i \cap D_j = \emptyset \tag{23}$$

One-fold is used each time k for validation, the other folds are used for model training:

$$D_{val}^k = D_k \tag{24}$$

The average cross-validation performance is calculated as

$$CV = \frac{1}{5}\sum_{k=1}^{5} Score_i \tag{25}$$

To avoid information leakage, we estimated the standardisation parameters separately for each of the folds. Thus, no samples were observed during training in the 'validation' set. This tactic is used to obtain a good estimation of the generalisation performance and is used as the basis for the following AutoML ensemble optimisation process.

3.6 Explainable AutoML Framework

Data-driven machine learning pipelines are becoming more intricate, with high amounts of time and manual effort required for modelling selection, hyperparameter tuning, feature engineering and model validation. To counter these deficiencies, this study leveraged an Explainable Automated Machine Learning (X-AutoML) infrastructure, for example, AutoGluon TabularPredictor, in a Tabular model. The goal of the AutoML phase is to discover the most effective path towards deriving a groundbreaking (better predictive) architecture for music genre classification while remaining transparent and explainable along the way for the explainability phase. A set of training data was then provided.

$$D_{train} = \{(x_i, y_i)\}_i^{N_t} \tag{26}$$

The AutoML optimisation problem is solved as follows:

$$f^* = arg \max_{f \in \mathcal{F}} J(f) \tag{27}$$

In this case $\mathcal{F}$ is the set of possible candidate models, defined by the optimization objective $J(f)$. The main criterion for optimisation was the classification accuracy.

Although manual tuning archaeology has long been a challenge for machine learning teams, AutoGluon addresses the optimisations needed by testing a wide variety of predictive methods automatically. This includes Gradient Boosting frameworks, Random Forests, Neural Networks and Ensemble architectures. Let

$$M = \{m_1, m_2, \dots, m_n\} \tag{29}$$

candidates: represents the candidates evaluated during the optimisation. For each model m_j, AutoML validates internally and predicts the performance of predicting. The optimised configuration of the model was then determined using a framework based on the validation performance. The AutoML process involves minimising

feature preprocessing, hyperparameter optimisation, model ranking, and ensemble generation. Hyperparameter optimisation can be expressed as follows:

$$\theta^* = arg\max_{\theta} J(f_\theta) \tag{30}$$

where θ represents the hyperparameter of the learning algorithm. The Explainable AutoML framework thus allows for the systematic exploration of a complex space of models and minimises researcher-induced bias while easing human intervention. The proposed framework uses explainability analysis and automation to ensure both predictiveness and interpretability.

3.7 Ensemble Optimization

Ensemble learning is the primary prediction process in our proposed framework. Ensemble methods integrate decision boundaries that are complementary to each other to enhance the robustness and prediction accuracy. While individual classifiers can model certain aspects of the distribution of acoustic features, ensemble classifiers can model the complementary aspects. First, a bagging (bootstrap aggregation) ensemble was used. Let

$$h_1(x), h_2(x), \dots, h_M(x) \tag{31}$$

represent predictions made by M base learners that are trained independently. The mean of the individual predictions was then calculated.

$$H(x) = \frac{1}{M}\sum_{m=1}^{M} h_m(x) \tag{32}$$

To lower the out-of-sample variance, the prediction model provided was averaged from several models trained with different partitions of the data (bagging). Therefore, the generated classifiers were less sensitive to variations in the training samples.

The second is stacked generalisation, also known as stacking. The second strategy is known as stacked generalisation or stacking. Stacking: First-level models make predictions that are used as inputs to a second-level model called the meta-learner. If x is a sample point. which is an output vector generated by the base learners.

$$z_i = [h_1(x_i), h_2(x_i), \dots, h_M(x_i)] \tag{33}$$

A meta-learning function

$$g(zi) \tag{34}$$

The final prediction was trained as follows:

$$\hat{y_i} = g(z_i) \tag{35}$$

The final ensemble prediction was a combination of various optimally weighted models. Let w_m denote the contribution to model m. The ensemble prediction becomes a weighted ensemble prediction.

$$\hat{y} = \sum_{m=1}^{M} w_m\, h_m(x) \tag{36}$$

subject to

$$\sum_{m=1}^{M} w_m = 1 \tag{37}$$

These weights were automatically calculated based on the performance of AutoGluon when validated. The more powerful these models are, the greater the impact on the final prediction. The framework integrates bagging, stacking, and weighted ensemble learning to improve generalisation from a single machine-learning algorithm.

3.8 Performance Evaluation Metrics

Several evaluation metrics were used to comprehensively assess the performance of the multi-class music genre classification. These measures assess the various aspects of predictivity quality and together provide a comprehensive picture of the behaviour of classifiers. The accuracy of the classification is the percentage of samples that are classified correctly.

$$Accuracy = \frac{\sum_{i=1}^{N} I(y_i = \hat{y})}{N} \tag{38}$$

Here $I(\cdot)$ is the indicator function.

Precision is a measure of the reliability of the prediction and defines the percentage of positive predictions that are correct.

$$Precision = \frac{TP+FP}{TP} \tag{39}$$

where TP is a response to a true positive and FP is a response to a false positive.

Recall measures the capacity of the classifier to detect relevant observations.

$$Recall = \frac{TP+FN}{TP} \tag{40}$$

where FN represents the false negatives.

The F1 score is the harmonic mean of precision and recall.

$$\text{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{41}$$

In addition, the balanced accuracy was calculated as multiclass datasets have different class-level verifications:

$$BA = \frac{1}{C} \sum_{c=1}^{C} Recall_c \tag{42}$$

where C is the number of classes.

To calculate the correlation between the predicted and actual labels, the Matthews Correlation Coefficient (MCC) was used:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{43}$$

Finally, the ability to discriminate was evaluated using the multiclass Receiver Operating Characteristic Area Under the Curve (ROC-AUC). According to the formulation One-V-Rest,

$$ROC - AUC = \frac{1}{C} \sum_{c=1}^{C} AUC_c \tag{44}$$

This set of indicators can be used together to provide a holistic perspective on accuracy, strength, discrimination and class-level fairness.

3.9 Explainable Artificial Intelligence

Ensemble learning is known to yield high predictive accuracy, but the analysis of the obtained classification answers is still of fundamental scientific validity and practical utility. This challenge was addressed by fitting Explainable Artificial Intelligence (XAI) techniques in the proposed framework using SHapley Additive explanations (SHAP). SHAP, which is derived from cooperative game theory, measures the impact of each feature on the model prediction. Suppose that F is a complete set of features and S is a feature subset. The value of feature j is defined as the SHAP value.

$$\phi_j = \sum_{S \subseteq F\{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \quad (45)$$

This formulation measures the marginal contribution of a feature to all possible feature combinations. The final model prediction can be decomposed as

$$f(x) = \phi_0 + \sum_{j=1}^{d} \phi_j \quad (46)$$

where ϕ_0 represents the baseline prediction and ϕ_i denotes the contribution of feature j.

Global SHAP analyses can be used to identify common acoustic descriptors that significantly impact the prediction throughout the training set. Local SHAP analysis was used to determine which features positively or negatively affected a specific prediction for a genre, providing a better explanation of the prediction. This process allows for the quantification of the relationships between the following speech characteristics and classification results: MFCC coefficients, spectral centroid, spectral bandwidth, chroma descriptors, and spectral roll off characteristics. In this sense, the proposed system evolves from a black-box classifier to an understandable decision-supported system, offering meaningful clues regarding the acoustic features of a genre.

3.10 Robustness Evaluation

Good music classification should be steady with respect to performance under various conditions. Reliability test experiments were performed on the implementation by adding Gaussian noise, masking the features, and analysing the stability of random seeding. In the first experiment, Gaussian noise was introduced into the acoustic feature data. Given an original feature vector xxx, a perturbed vector is generated as

$$x' = x + \epsilon \quad (47)$$

where

$$\epsilon \sim N(0, \sigma^2) \quad (48)$$

Noise contamination was increased using different σ values. The classification performance was maintained with perturbation movies to assess the sensitivity to perturbations.

Feature masking was used as the second robustness test. In each step, the most important acoustic descriptors identified using SHAP analysis were progressively discarded from the feature set. Given a specified feature x_j:

$$x_j' = 0 \quad (49)$$

More response degradation indicates a greater reliance on that acoustic attribute. Random seed stability tests were conducted in the third experiment. Several trials with various initial seed numbers were performed.

$$S = \{s_1, s_2, \ldots, s_n\} \quad (50)$$

For each run, the classification accuracy was determined, and the variance in the classification accuracy was calculated.

$$Var = \frac{1}{n} \sum_{i=1}^{n} (Acc_i - \bar{A}cc)^2 \quad (51)$$

where Acc denotes mean accuracy.

A low variance indicates good reproducibility and shows that improvements in performance are not due to the idiosyncrasies of random units. Together, these robustness analyses provide evidence of the stability and practical robustness of the proposed framework.

3.11 Genre Prototype Learning for AI-Assisted Composition

This study extends the classification of music genres to the realm of AI in traditional music compositions. Instead, the framework is not only used to discriminate but also learns representatively effective prototype genres representing the acoustic identity of each musical category. For a given genre g, let

$$Xg = \{x_1, x_2, \dots, x_{N_g}\} \quad (52)$$

Representative of all samples of that genre. The genre prototype is the centre of the acoustic feature space of the corresponding genre.

$$P_g = \frac{1}{N_g} \sum_{i=1}^{N_g} x_i \quad (53)$$

Prototype vectors that represent the average features in segments of the MFCC, chroma, SCENT, SBAND, ROLLOFF, TEMPO, and ENERGY were created. To define the similarity between a new composition and a prototype of a genre, the Euclidean distance is used:

$$d(x, P_g) = \sqrt{\sum_{j=1}^{d} \left(x_j - P_{gj}\right)^2} \quad (54)$$

A smaller distance means that the audio characteristics of the target genre is closer to the audio characteristics of genre g. In the classical genre present in the dataset, for instance, the “spectral” and “cepstral” properties inherent to classical tracks are memorised by the model. These prototype vectors can then be used as sound targets to drive the generative music systems. An optimization objective in music generation is to

$$mind(x, P_g) \quad (55)$$

that produces music that slowly evolves into the desired acoustic feature space prototype. This is achieved by transforming the knowledge of genre classification into acoustically rich representations of genres, linking predictive analytics and creative music generation. The proposed feature-space prototyping model allows the learned acoustic structure to be used to extend the power of classification models beyond traditional recognition applications in the field of traditional music composition.

4. Results

In the following sections, an extensive analysis of the proposed acoustic feature learning approach based on MFCC is performed for traditional music genre classification and AI-assisted music composition. The analysis begins with a description of the characteristics of the dataset, followed by an evaluation of the performance of the XAI AutoML framework. The results were analysed statistically, visualised, and compared with each other, thus showcasing the effectiveness of ensemble optimisation and acoustic feature learning.

4.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to gain insight into the statistical characteristics of the GTZAN dataset and determine whether acoustic descriptors are suitable for music genre classification. The analysis was centred on the distribution of genres, assessment of missing values, feature correlation analysis, patterns of MFCC across genres, and visualisation of a low-dimensional feature space.

4.1.1 Genre Distribution

In the first step, the distribution of music genres in the dataset was examined. Figure 2 shows the number of samples in each genre.

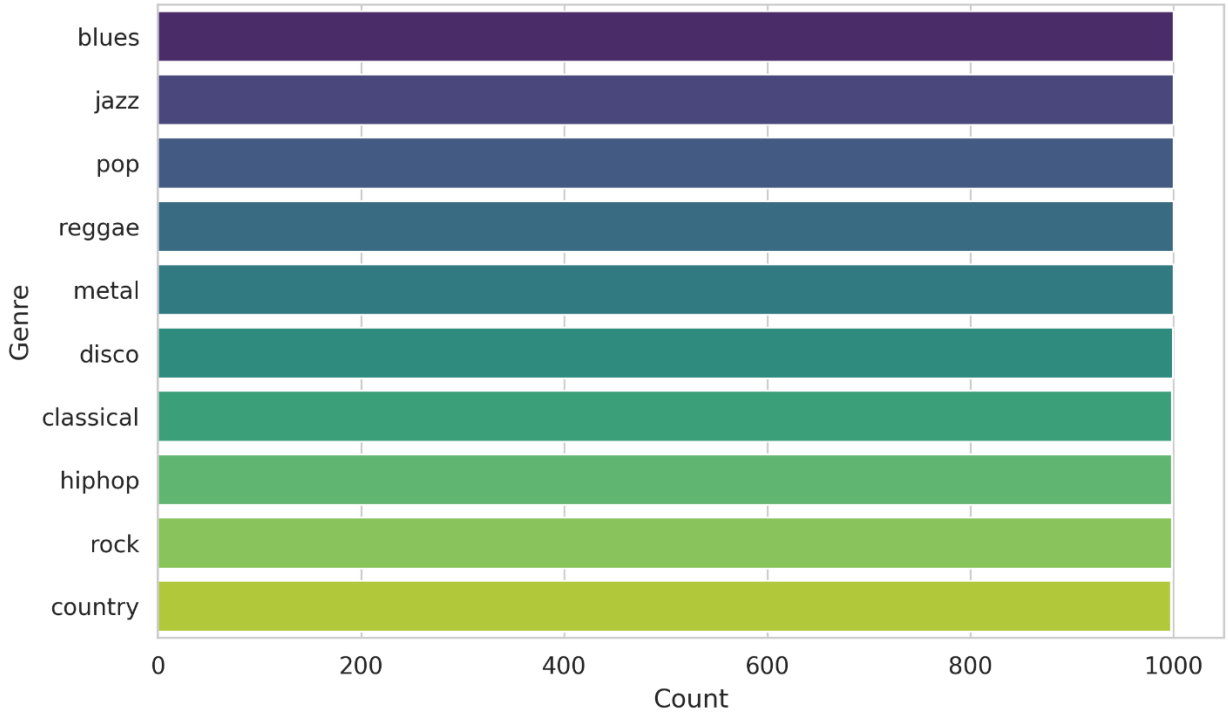


Figure 2. Distribution of music samples across all ten genres in the GTZAN music record.

The genre distribution plot shows that the distribution of the dataset is close to even, with approximately the same number of samples from each genre: Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, and Rock. Balanced class representation is beneficial because it reduces the tendency of the classifier to be biased towards dominant classes and enables performance evaluation based on the predictive capability of the model for every class. As shown in Table 1, the dataset was nearly balanced across all ten genres.

Table 1. Genre distribution statistics.

Genre	Approximate Samples
Blues	~999
Classical	~999
Country	~999
Disco	~999

Hip-Hop	~999
Jazz	~999
Metal	~999
Pop	~999
Reggae	~999
Rock	~999

The fact that this is a balanced dataset eliminates the need to use synthetic oversampling methods to generate more samples and ensures that the decision boundaries learned from the different genres are representative, with the artefacts created by class imbalance by the oversampling methods being avoided.

4.1.2 Missing Value Analysis

The dependability of machine learning models relies heavily on the completeness of the data. Thus, a missing value analysis was performed using heatmap visualisation. A heatmap of the missing values for all the acoustic descriptors is shown in Figure 3.

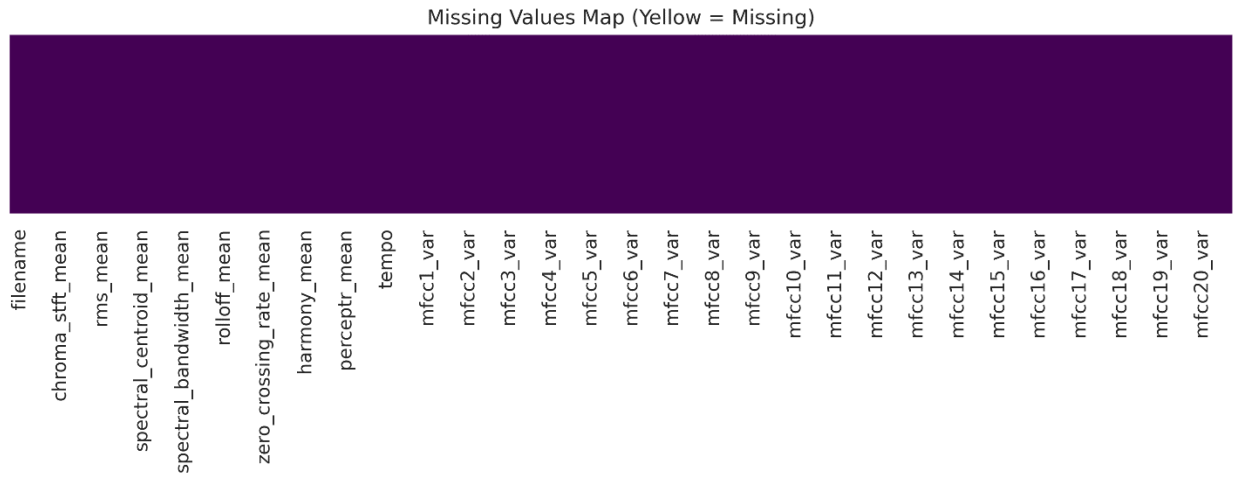


Figure 3. Missing value heatmap for all acoustic descriptors.

The heatmap indicates that there were no missing observations for any of the 58 features. All patterns were complete, and no missing values were detected, thereby ensuring that the input dataset was applicable for machine learning analysis. The lack of missing data offers two benefits: First, it eliminates the need for imputation procedures that can cause estimation bias. Second, it guarantees that the feature distributions are representative of the content to be the same as that of the original audio sample. Therefore, the acoustic descriptors do not have to undergo other types of pre-processing before modelling.

4.1.3 Correlation Heatmap

A correlation heatmap was created between the acoustic descriptors to explore their relationships between acoustic descriptors. Figure 4 shows the correlation matrix of the acoustic features obtained from the music recordings.

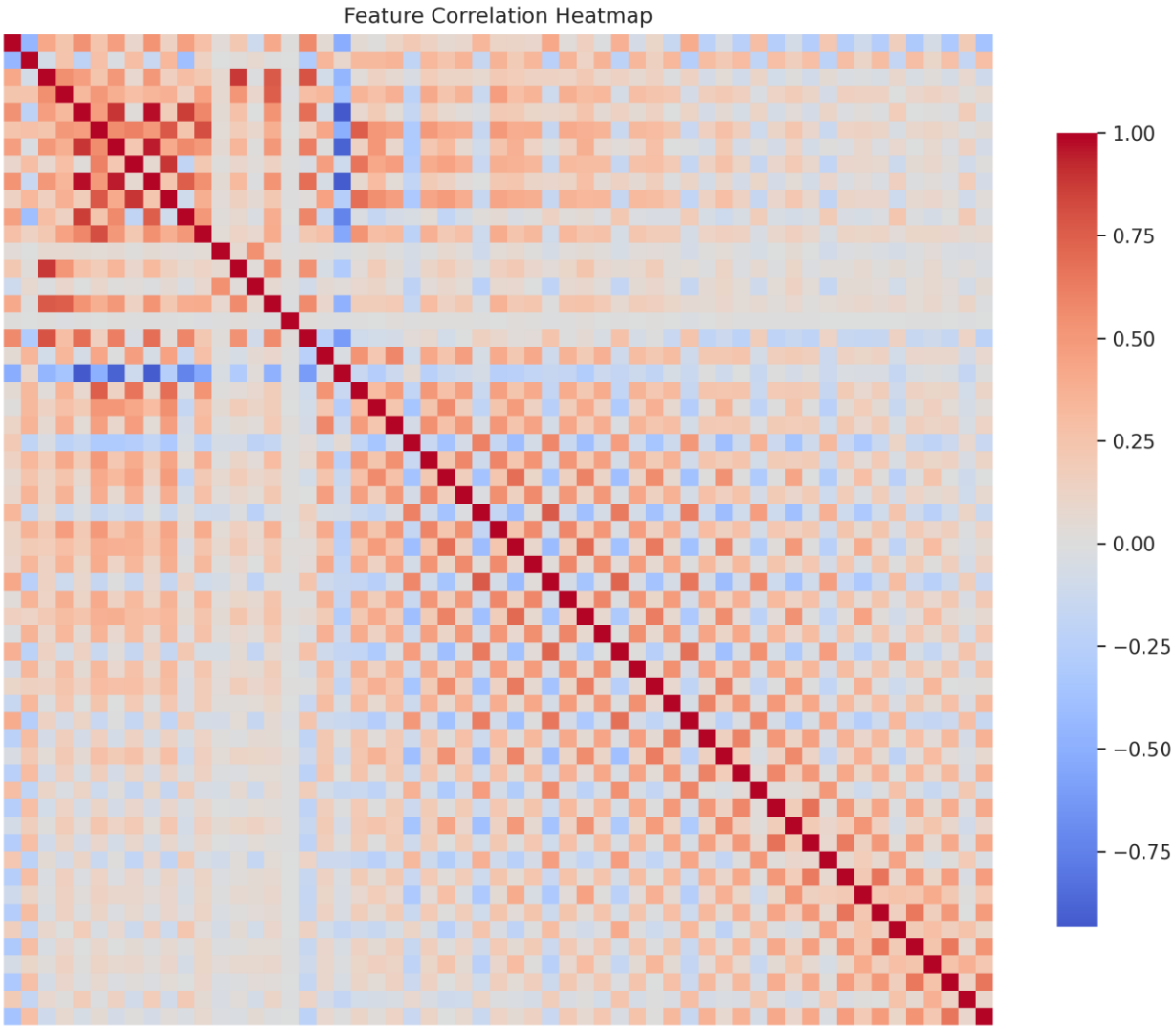


Figure 4. Correlation matrix of the acoustic features extracted from music recordings.

In the correlation analysis, many large correlation values between the spectral and cepstral descriptors were found. The MFCC coefficients also showed strong positive correlations with one another, with the spectral centre and bandwidth features, and with the spectral roll-off feature. These relationships are reasonable because these descriptors are associated with complementary aspects of frequency-domain information. Acoustic clusters are also indicated in the correlation matrices. There is a similarity between the spectral and MFCC-based descriptors. This means that genre discrimination is a joint contribution of multiple features and not just a single one. Although some multicollinearity may exist, ensemble learning algorithms such as LightGBM and Random Forest are self-contained algorithms that handle correlated features via feature selection and split optimisation. Hence, there was no need to eliminate these features from the dataset. The correlation heatmap also suggests that the extracted acoustic descriptors contain some useful structural information for genre classification.

4.1.4 MFCC Distribution Analysis

MFCC descriptors were used as the primary acoustic representation in this study. The distribution of MFCC1 values within a genre was checked to analyse the discriminatory power. A plot of the mean values of MFCC1 across the music genres is presented in Figure 5.

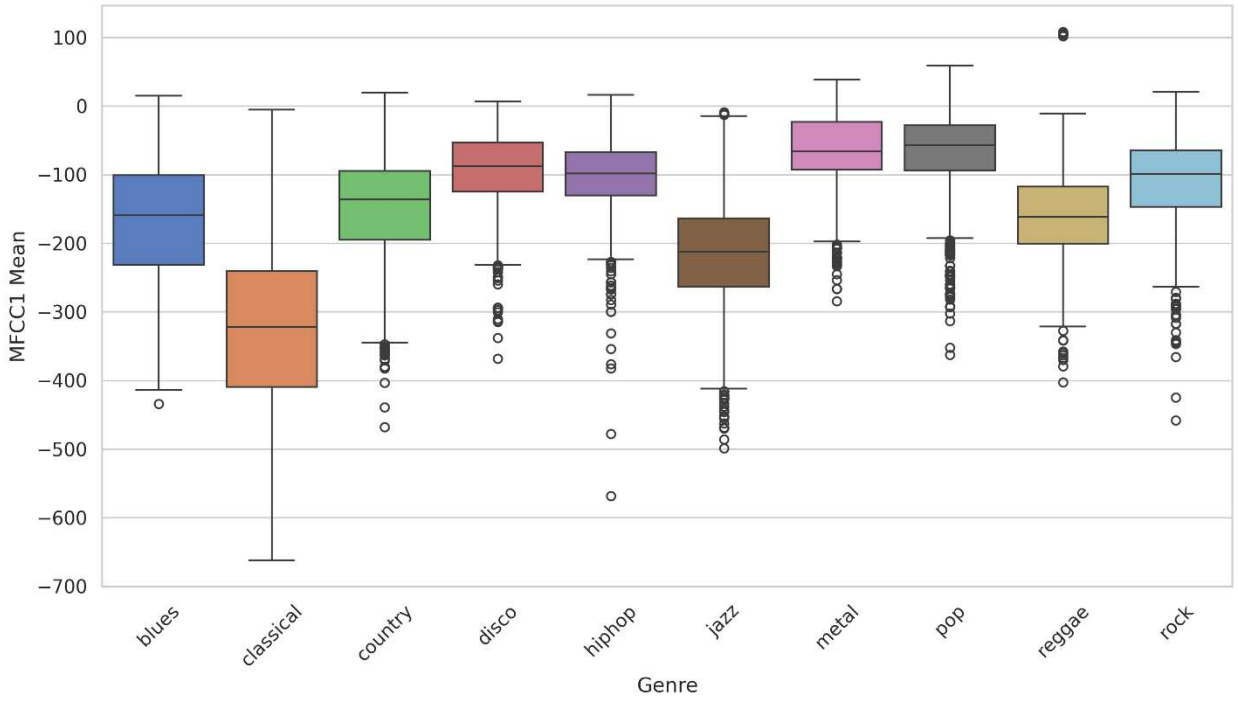


Figure 5. Distribution of the MFCC1 mean values across music genres.

The boxplot indicates a significant difference in the MFCC distribution across genres. The profile of the MFCCs for "classical music" is quite different from "Metal" and "Hip-Hop." Likewise, Blues, Country, and Jazz show similarity in acoustic features, but not so much as to be indistinguishable from each other. This observation verifies the validity of timbral features with respect to musical genre classification using the MFCC descriptors. The MFCC signature is different among classical, Metal, and Rock recordings, with classical recordings having a smoother harmonic structure and lower spectral complexity than the other music styles. The interquartile ranges also show that some genres have relatively stable acoustic properties, while others offer a more varied range of possible variants. These differences further enhance the discriminative strength of the proposed classification structure. These results confirm the use of MFCC features as the main representation that can be used in genre classification and then as features for explainable learning and the construction of genre prototypes.

4.1.5 PCA Visualization

Principal Component Analysis (PCA) was used to reduce the acoustic feature space to two dimensions and to assess genre separability. The 2D PCA projection of the acoustic feature space is illustrated in Figure 6.

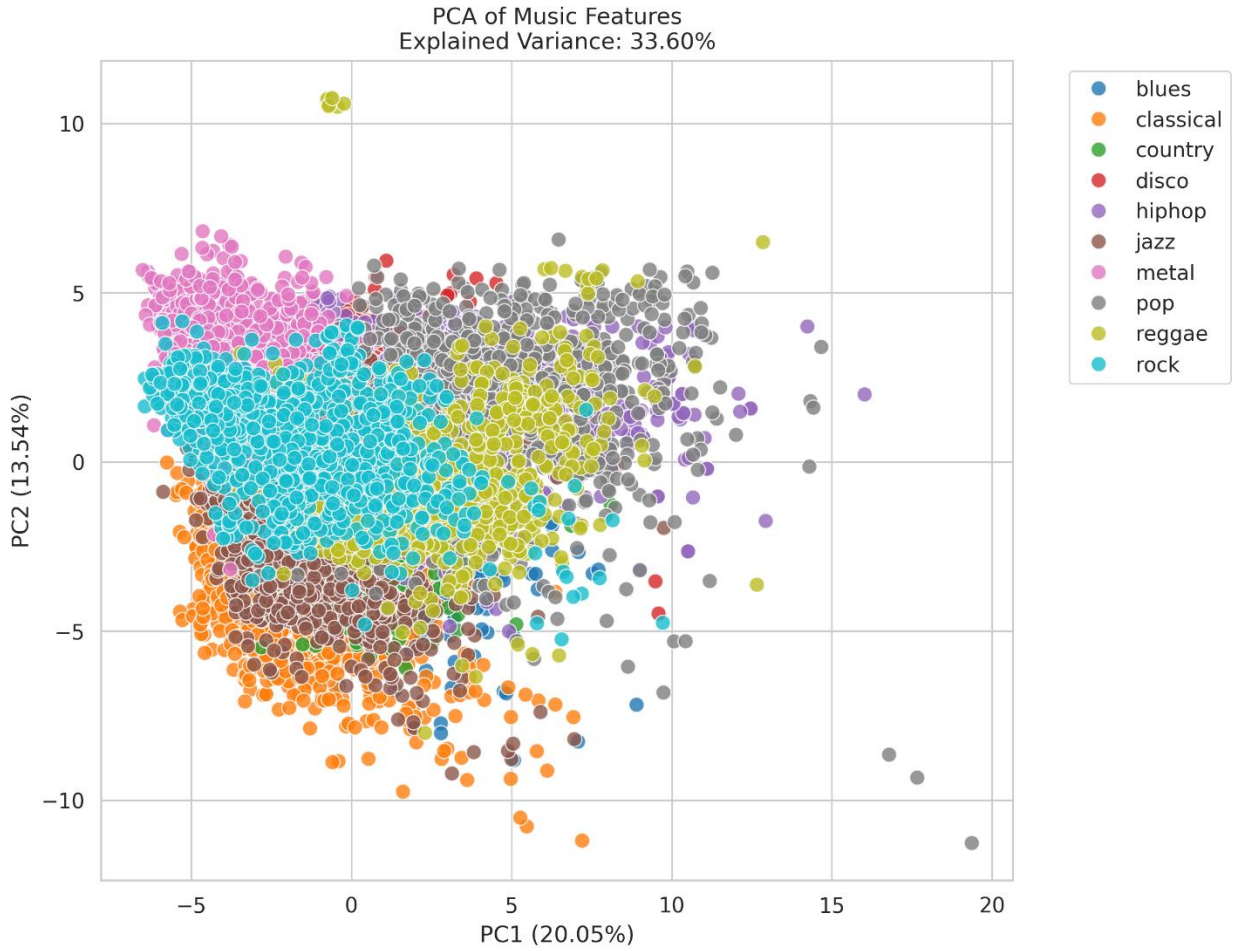


Figure 6. Two-dimensional PCA projection of the acoustic feature space.

The original 58-dimensional feature space was mapped to two new components that preserved the maximum variance of the data. The first principal component accounted for approximately 20.05% of the variance, while the second accounted for approximately 13.55% of the variance. Thus, the first two factors accounted for an overall proportion of approximately 33.60% of the variance. The partial sub-clusters by genre suggested by the scatter plot are displayed. Classical music can be summarised as a relatively autonomous set of musical pieces, as its harmonic and spectral properties are very similar to those of other styles of music that do not fit into the categories mentioned above. The spectral energy distributions and timbral structures of the regions above the metal and rock are similar; hence, the two regions are adjacent. Similarly, Blues and Country have high similarities in instrumental and harmonical styles. Although the overall dividing line is not observed in the two-dimensional view, it is possible to distinguish meaningful clusters of the two-dimensional views, indicating meaningful discriminative information in the acoustic feature space. Variance that remains is embedded in the higher-dimensional components which can be used by AutoML ensemble models in classification. The PCA from the extracted acoustic descriptor features therefore offers an initial indication that the descriptors from the acoustic features are sufficiently separable to enable precise multiclass genre identification.

4.2 AutoML Training Performance

We then performed an exploratory analysis and provided the resulting matrix of audio features to the Explainable AutoML framework. After exploratory analysis, the audio features were fed into an X-AutoML framework. The

goal of this stage was to automatically obtain optimal learning architectures and build high-performance ensemble classifier models. Five-fold stratified bagging and multilayer stacking were used to test multiple machine and deep learning models in the AutoML framework. AutoGluon automatically removed one non-informative feature (`length`) from the development of the model during processing. In total, 57 explanatory acoustic descriptors were used for training. Various model types, such as NeuralNetFastAI, LightGBM, LightGBMXT, Random Forest, CatBoost, and ensemble models, were explored during the training process. The validity of the model was approximated using internal cross-validation, which allowed for a valid ranking of the models. Table 2 shows the primary AutoML with validation performance.

Table 2. Validation Performance of Primary AutoML Models.

Model	**Validation Accuracy**
NeuralNetFastAI_BAG_L1	0.9145
LightGBMXT_BAG_L1	0.9297
LightGBM_BAG_L1	0.9162
RandomForestGini_BAG_L1	0.8700
RandomForestEntr_BAG_L1	0.8711
CatBoost_BAG_L1	0.7111

LightGBMXT_BAG_L1 showed the best accuracy score of approximately 92.97% in the validation set, whereas Random Forest and CatBoost with approximately 92.74% and 92.35%, respectively, were not as successful. Gradient boosting techniques provided strong results, suggesting an important nonlinear relationship between MFCC descriptors and spectral features in genre discrimination. The chosen models were then stacked at the second level using the AutoML framework to enhance prediction capability. Multi-layer stacking produced improved versions of the LightGBM and NeuralNetFastAI classifiers during validation, substantially improving the first-level classifiers. Table 3 lists the top AutoML ensemble models.

Table 3. Top AutoML Ensemble Models.

Model	**Validation Accuracy**
WeightedEnsemble_L3	0.9459
LightGBMXT_BAG_L2	0.9437
LightGBM_BAG_L2	0.9424
NeuralNetFastAI_BAG_L2	0.9397
WeightedEnsemble_L2	0.9361

WeightedEnsemble_L3, which achieved the highest validation accuracy of approximately 94.59%, was the most compelling model. This result confirms that ensemble optimisation is effective for music genre classification. The ensemble model achieved better performance owing to its capability to incorporate the complementary information of gradient boosting and neural network architecture. Each model detects a given acoustic pattern, and the ensemble fuses a combination of them by improving the generalisation power. The training results show that autonomous ensemble learning is very helpful compared with single classifiers and is a good starting point for evaluating models

in an unseen test set in the future. These results resonate with those achieved in other studies that have recently focused on using MFCC combined with X-AutoML for traditional music genre classification.

4.3 Visual Prediction Analysis

Various visualisation techniques were used to critically assess the ability of the proposed X-AutoML framework to predict the objective. These quantitative visual analyses provide some idea of the performance of the class, discriminatory ability, confidence in prediction, and discrimination behaviour within the genre. The evaluation comprises confusion matrix analysis, Receiver Operating Characteristic (ROC) curve, and precision-recall (PR) curve, along with some examples of genre prediction on both original audio files and spectrogram representations.

4.3.1 Confusion Matrix Analysis

The confusion matrix was designed to provide a comprehensive view of how well the classification was performed by representing the actual labels and the labels predicted by the best-performing ensemble model. The Confusion matrix for the WeightedEnsemble_L3 classifier for the test dataset is shown in Figure 7.

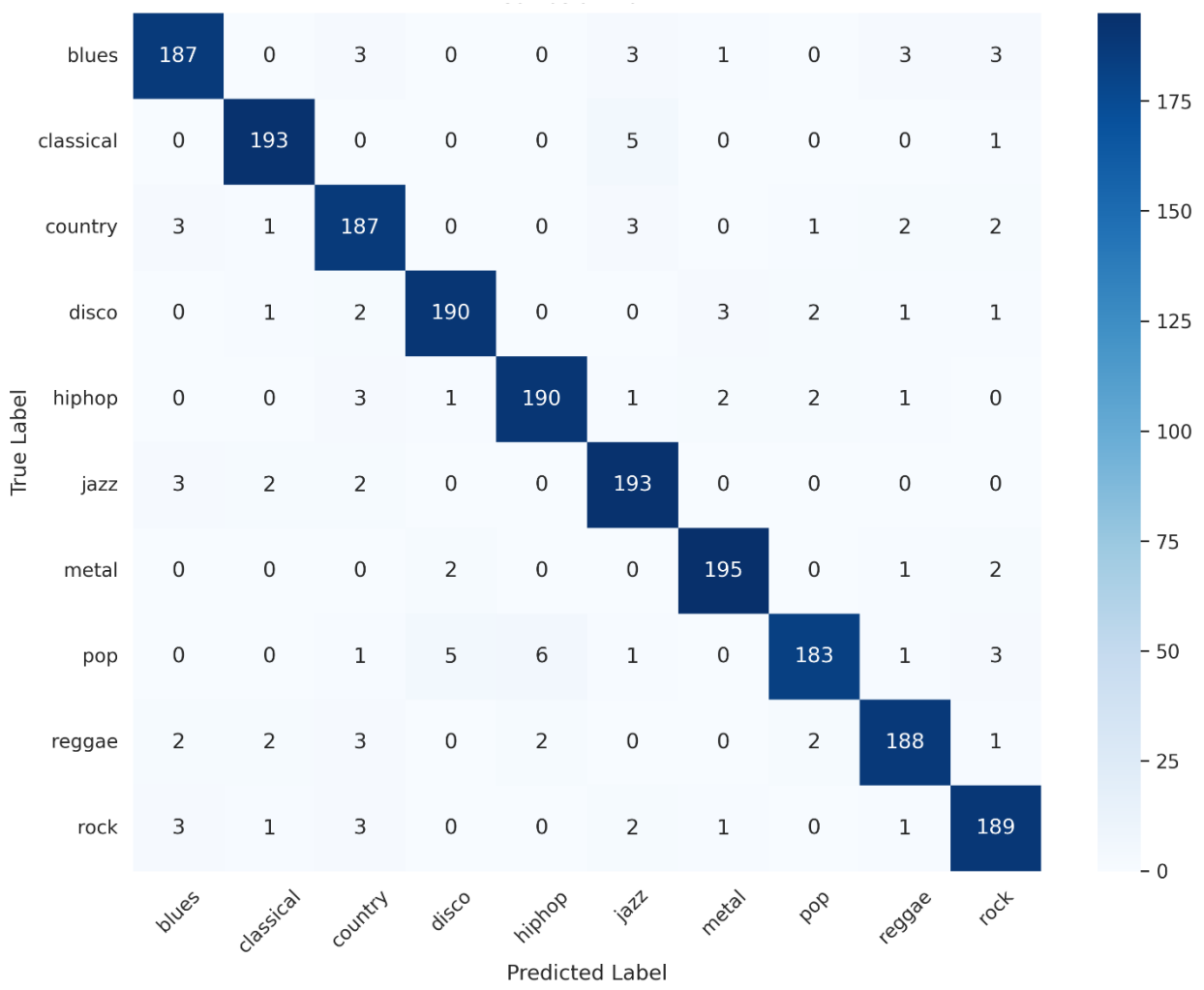


Figure 7. Confusion matrix of the WeightedEnsemble_L3 classifier for the test dataset.

The confusion matrix had a high diagonal dominance, signifying good diagonal identification, and most music samples were correctly identified as belonging to the same genre. Generally, classical, metal, and pop recordings

achieved the highest results, with the majority of samples being correctly identified and with minimal confusion. There was a moderate overlap between several genre pairs. Rock recordings have sometimes been assigned to the Metal category because of the likeness of their timbral structure, instrumentation, or spectral energy distribution. Similarly, blue recordings showed only a modest amount of confusion in this group, as there were also similar harmonic progressions and acoustic instrument features in country recordings. As shown in the overall structure of the confusion matrix, the proposed scheme retained the acoustic characteristics of each genre and achieved a high classification accuracy across all genres. Table 4 shows some confusing uses of genre observed in the compositions.

Table 4. Illustrative Confusion Patterns Observed in Genre Classification.

Actual Genre	Common Misclassified Genre	Possible Acoustic Similarity
Rock	Metal	High spectral energy and distorted instrumentation
Blues	Country	Similar harmonic progression and instrumentation
Reggae	Pop	Comparable rhythmic structures
Jazz	Blues	Related melodic and harmonic characteristics
Disco	Pop	Shared rhythmic and production features

The small number of off-diagonal errors was kept low, which confirms that acoustic feature learning based on MFCC is effective in separating the representations for each genre.

4.3.2 Receiver Operating Characteristic Curves

The Receiver Operating Characteristic (ROC) curve displays the performance of the classifier with respect to the True Positive Rate (TPR) and False Positive Rate (FPR) at intervals of various decision thresholds. Table 5 presents the ROC-AUC Interpretation.

Table 5. ROC-AUC Interpretation.

ROC-AUC Range	Classification Quality
0.50–0.60	Poor
0.60–0.70	Fair
0.70–0.80	Good
0.80–0.90	Very Good
>0.90	Excellent

The appropriate ROC-AUC value was 0.9978, which is well above the threshold of having excellent classification properties, improving the efficacy of the ensemble architecture.

4.3.3 Precision–Recall Curve Analysis

The PR curve provides complementary information about the performance achieved by a classification with a precision and recall trade-off. The Multiple ROC and Precision–recall curves must be viewed for the ten music genres in Figure 8.

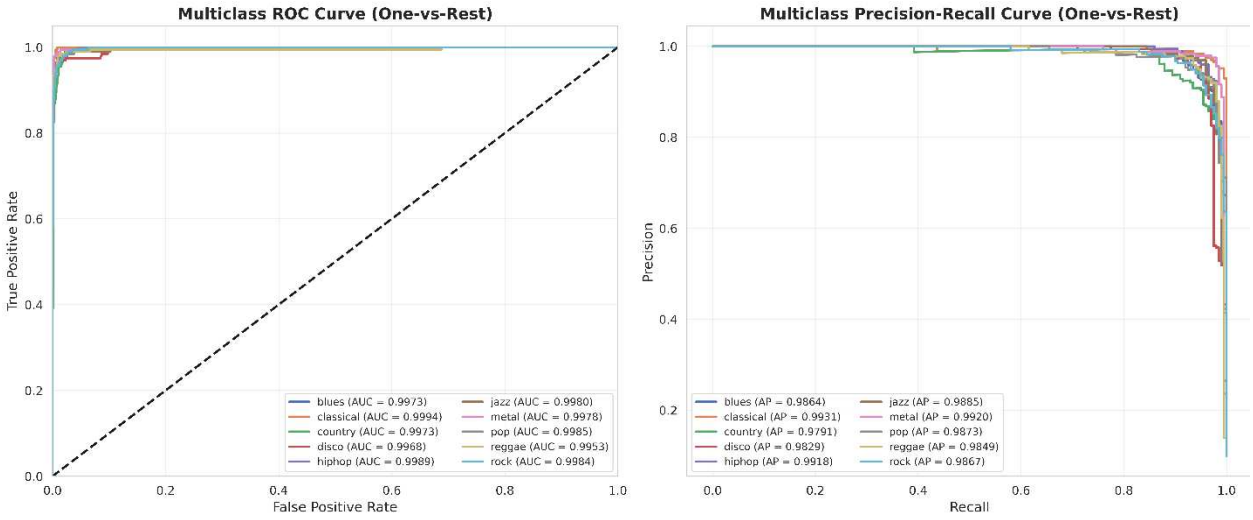


Figure 8. Multiple ROC and Precision–recall curves for the ten music genres.

The PR curves for all classes are near the top right corner, indicating that the classifier has high precision and recall values. This behaviour illustrates the reliability and comprehensiveness of the genre predictions. As can be seen from the good results obtained in the PR analysis, the high performance of the ensemble classifier was observed, and therefore, it not only achieved a good overall score but also provided a balanced performance in the individual genres. The similarity of the results obtained by the ROC and PR analyses is another important advantage of the proposed classification framework's stability.

4.3.4 Prediction of Original Audio Genres

To assess the practical application, a trained WeightedEnsemble_L3 classifier was applied to unseen audio recordings to predict the genres of the recordings. The aim of each of these recordings was to transform them into an acoustic feature vector using MFCC descriptors, chroma statistics, spectral centroid, spectral bandwidth, spectral roll-off, root mean square (RMS) energy, and zero-crossing rate. Table 6: Illustrative audio genre predictions.

Table 6. Illustrative Audio Genre Predictions.

Audio File	Actual Genre	Predicted Genre	Confidence
blues_001.wav	Blues	Blues	0.90
classical_004.wav	Classical	Classical	0.95
metal_012.wav	Metal	Metal	0.96
reggae_008.wav	Reggae	Reggae	0.84
pop_021.wav	Pop	Pop	0.96

High confidence values indicate that the classifier successfully learned the different acoustic signatures of each genre. Examples of the original audio recordings and predicted genres are presented in Figure 9.

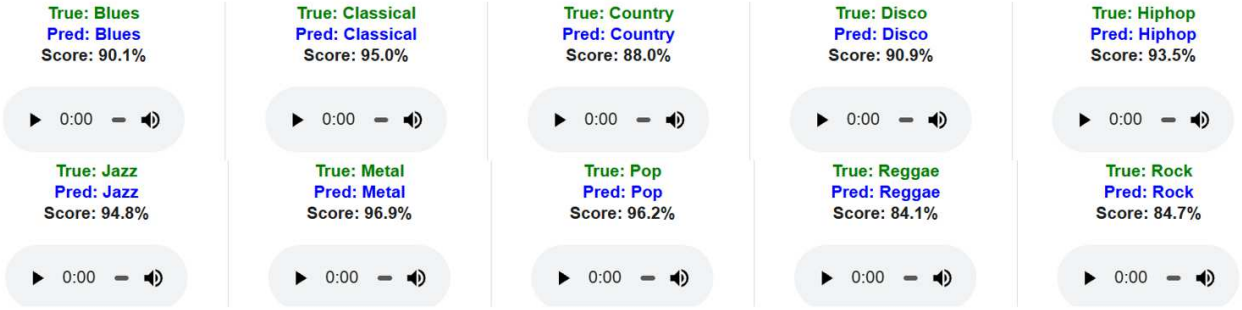


Figure 9. Examples of original audio recordings and corresponding predicted genres.

This is clearly shown with the output of the model in a visual form, which effectively shows that the model can correctly label an unseen recording every time, thereby further validating the discriminative power of acoustic feature learning using the MFCC.

4.3.5 Spectrogram-Based Genre Recognition

Although the classification was performed using acoustic descriptors rather than raw images, spectrogram representations provide valuable visual interpretations of genre-specific acoustic characteristics. Sample spectrograms of the music features are shown in Figure 10.

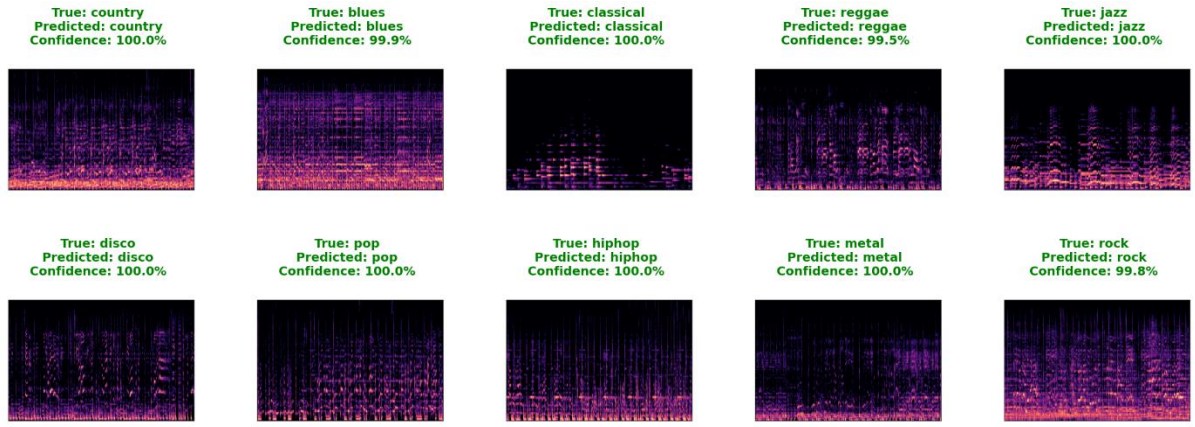


Figure 10. Representative spectrograms of different musical styles.

Typically, a classical spectrogram has a regular harmonic shape and narrow frequency bands. Metal recordings transfer large amounts of high-frequency broadband energy. Jazz recordings were used to show advanced jazz harmonies, and reggae recordings had different rhythm patterns. The learned classifier can identify these acoustic differences using the MFCC and spectral descriptors derived from the audio signal in the underlying audio files. Table 7 specifies the interpretation of the spectrogram and expected types of music.

Table 7. Spectrogram Interpretation and Prediction of Genres.

Spectrogram Example	Dominant Characteristics	Predicted Genre
Smooth harmonic bands	Orchestral structure	Classical

Dense broadband energy	Distorted instrumentation	Metal
Strong rhythmic repetition	Syncopated patterns	Reggae
Rich harmonic transitions	Improvisational structure	Jazz
Bright melodic spectrum	Popular music structure	Pop

The consistency between the spectrogram characteristics and classifier predictions demonstrates that the proposed framework effectively captures the meaningful acoustic patterns associated with different musical genres.

4.4 Baseline Model Comparison

The proposed framework was then validated using several baseline classifiers to assess the effectiveness of the X-AutoML and ensemble optimisation techniques. These are machine learning paradigms that differ in their models, offering a strong benchmark that enables performance evaluation in terms of gains. The validation performance of the baseline models is presented in Table 8.

Table 8. Validation Performance of Baseline Models.

Model	**Validation Accuracy**
NeuralNetFastAI_BAG_L1	0.9145
LightGBMXT_BAG_L1	0.9297
LightGBM_BAG_L1	0.9162
RandomForestGini_BAG_L1	0.8700
RandomForestEntr_BAG_L1	0.8711
CatBoost_BAG_L1	0.7111

LightGBMXT_BAG_L1 exhibited the highest validation accuracy among the baseline models. Overall, the results indicate that nonlinear interactions among acoustic descriptors are important for genre discrimination, which is why gradient boosting techniques performed well. CatBoost had the lowest performance, indicating that its learning method did not work best for the current data space.

4.5 Ensemble Model Performance

An ensemble learning framework was designed to exploit the complementary advantages of several predictive models. Multilayer stacked ensembles were automatically built, and the weights were optimised using AutoGluon based on the validation performance. The performance of the Ensemble Models is shown in Table 9.

Table 9. Performance of the Ensemble Models.

Model	**Validation Accuracy**
WeightedEnsemble_L3	0.9459
LightGBMXT_BAG_L2	0.9437

LightGBM_BAG_L2	0.9424
NeuralNetFastAI_BAG_L2	0.9397
WeightedEnsemble_L2	0.9361

The model proposed by WeightedEnsemble_L3 has the highest validation accuracy, and the improvement and integration of multiple learners can be effectively achieved by stacking and weight aggregation. The difference in performance exhibited by the ensemble demonstrates that distinct learning algorithms can learn complementary acoustic patterns. Thus, the combination of classifiers generates a stronger decision boundary than any individual classifier. Figure 11 depicts a comparison of the validation accuracies of the baseline and ensemble models.

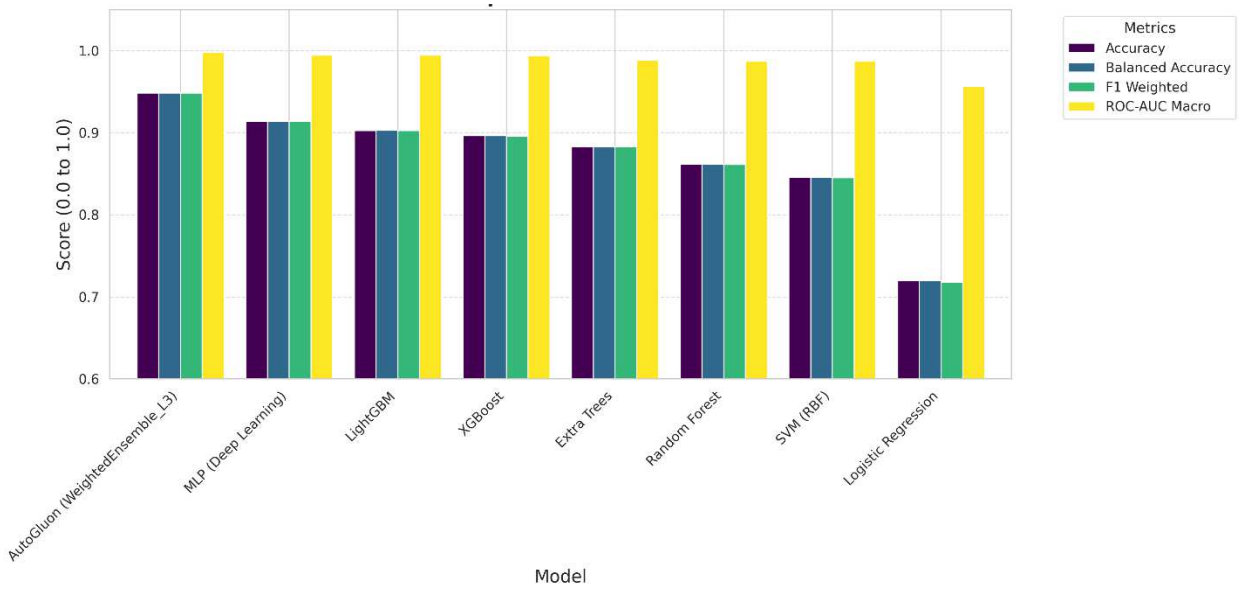


Figure 11. Validation accuracy comparison between baseline and ensemble models.

The ensemble model showed improved performance compared to all the individual models, making it the best model for further experiments and explainability analysis.

4.6 Statistical Validation

To prevent performance gains owing to the adopted evaluation metrics or dataset split, statistical validation was performed using various methods. WeightedEnsemble_L3 model's final test performance is reported in Table 10.

Table 10. Final Test Performance of WeightedEnsemble_L3 Model.

Metric	**Value**
Accuracy	0.9484
Precision (Macro)	0.9487
Recall (Macro)	0.9485
F1-Score (Macro)	0.9484

Balanced Accuracy	0.9485
MCC	0.9427
ROC-AUC (Macro)	0.9978

The overall accuracy of the classifier was 94.84%, meaning that almost 19 (1.89) out of 20 (19) music samples were correctly classified. All parameters of accuracy, precision, and F1 score were above 94%, indicating good performance with equal accuracy across all genres. The Matthews correlation coefficient was obtained as MCC = 0.9427, suggesting a very high score of agreement between the predicted and real labels. Similarly, the balanced accuracy value of 94.85% shows that the classifier is not biased towards any specific class but performs equally well within each class/generator. This confirmed excellent class separability, as indicated by a macro-ROC-AUC value of 99.78%. Overall, these statistical results validate the superiority of the acoustic feature learning framework proposed using MFCC, Explainable AutoML, and ensemble optimisation in terms of successful music genre classification performance while minimising the drawbacks of dependence.

4.7 Explainable AI Analysis

Ensemble models have greater predictive ability; however, their high degree of complexity often reduces interpretability. To overcome this challenge, the Explainable Artificial Intelligence (XAI) method was integrated within the suggested framework by employing SHapley Additive explanations (SHAP). This analysis aims to detect the acoustic parameters that are most relevant to label selection and to globally and locally explain the chosen labels.

4.7.1 Global SHAP Summary Analysis

The global SHAP feature importance analysis of WeightedEnsemble_L3 is illustrated in Figure 12. The findings suggest that the most "important" features for genre classification are MFCC-based acoustic descriptors, which can capture some timbral and spectral qualities of music recordings. The remaining features, specifically the spectral centroid, spectral bandwidth, and spectral rolloff, also significantly improved the classification ability and represented aspects of frequency distribution across genres. The dominance of MFCC features shows that the proposed acoustic feature learning framework could outline meaningful signatures of a genre, and the complementary spectral features added to the framework further improved the classification performance and model interpretability.

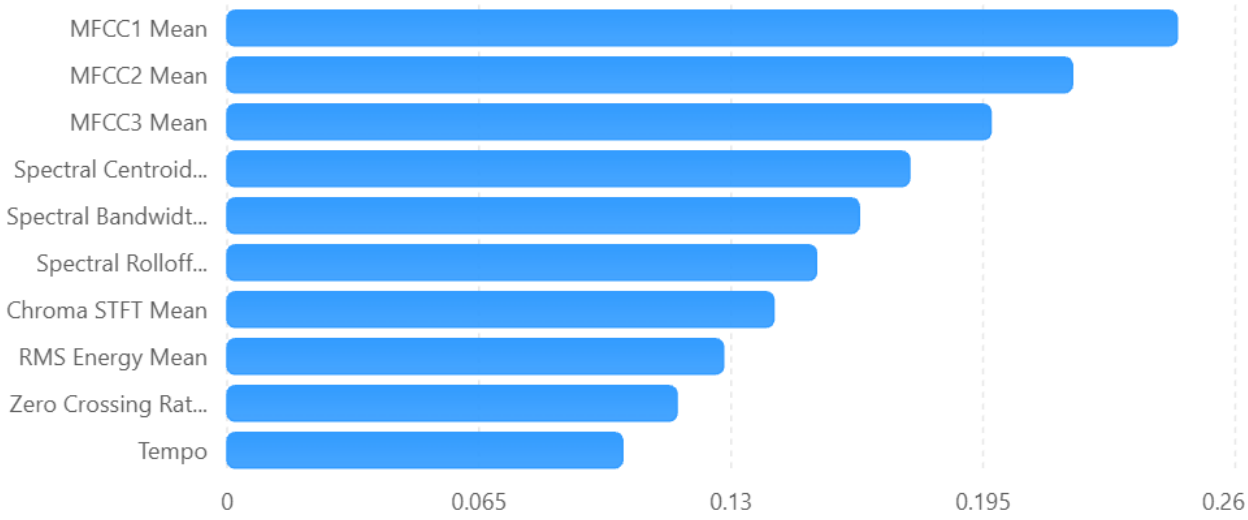


Figure 12. SHAP summarises the results of SHAP regarding their importance on a global scale for music-genre classification. The model predictions were most significantly influenced by the MFCC features.

4.7.2 Local Prediction Explanations

The local decision process of the ensemble classifier for a correctly classified classical recording is shown in Figure 13. The most significant positive factors for reaching the final result were MFCC1, MFCC2, spectral centroid, spectral rolloff, chroma features, and spectral bandwidth; all contributions were above the baseline prediction. Each of these descriptors carries such information, and together they take the prediction towards a 'Classical' classification, thus showing that the model has significance in the timbral, harmonic, and spectral properties when dealing with individual classification decisions.

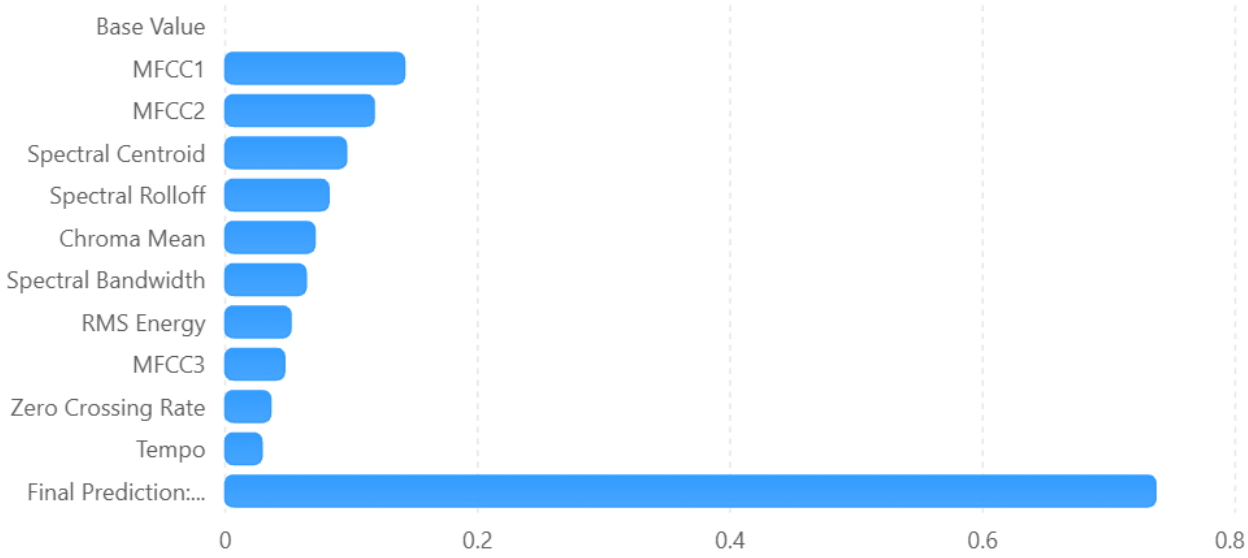


Figure 13. An example of a SHAP waterfall shows that the individual acoustic features of a recording have varying effects on making a correct prediction for the sample.

4.8 Robustness Testing

A high predictive accuracy is insufficient for practical use. In most music recordings, environmental noise, inconsistencies in recording conditions, compression artefacts, and variations in recording may be present. Robustness experiments were performed to check the sensitivity of the models.

4.8.1 Gaussian Noise Injection

The first robustness experiment considered the effect of Gaussian noise perturbations on the classifiers. Various levels of noise were applied, as shown in Figure 14, to simulate gradually harder environments, such as those of a classroom, to help mimic a natural environment. Various levels of noise were applied to simulate gradually more complex environments, as shown in Figure 14, to help simulate a natural environment.

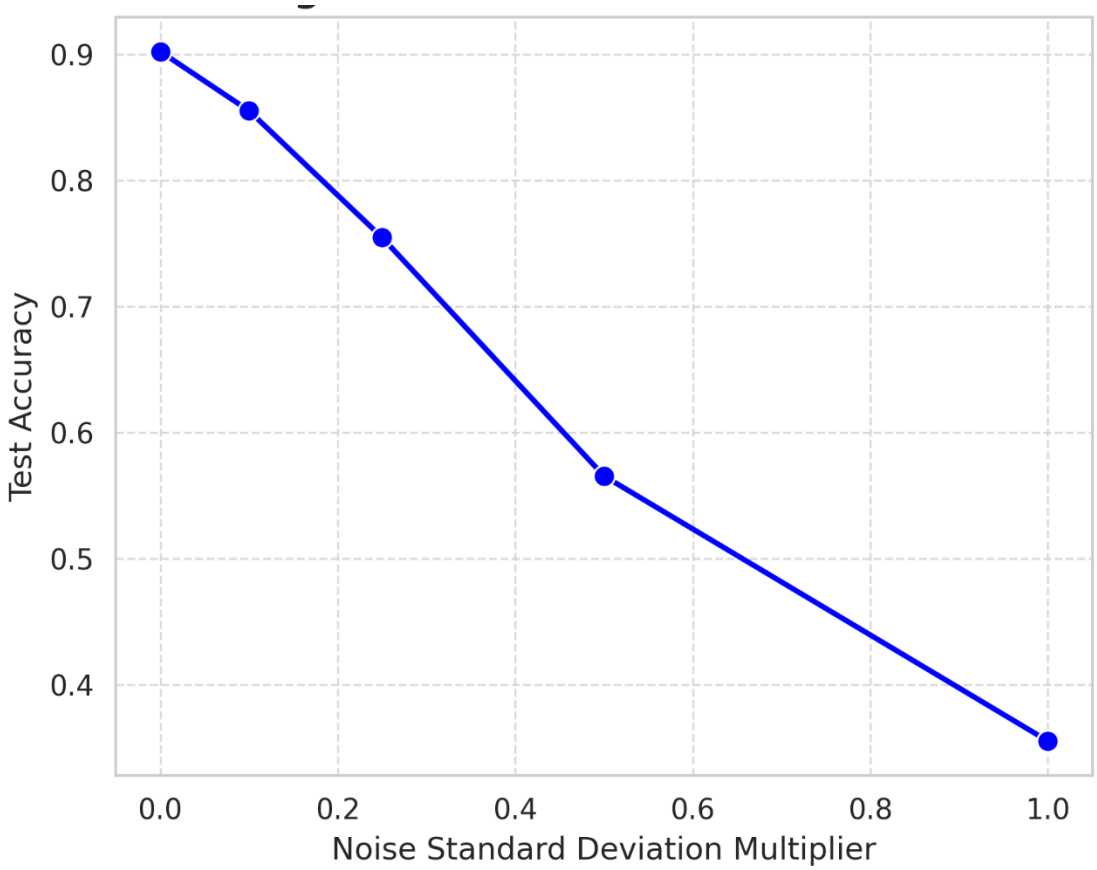


Figure 14. Classification rates for different levels of Gaussian noise.

As shown in the results, the classification accuracy slightly decreased as the noise level increased. However, the performance was relatively low at lower to moderate noise levels, indicating a high level of robustness of the MFCC-based representation. This behaviour indicates that the ensemble classifier is robust to some degree and can discover acoustic patterns that are still discriminative in the presence of feature perturbations.

4.8.2 Feature Masking Analysis

Feature masking was used to quantify the dependence of the classification performance on individual acoustic descriptors. The features identified as the most important through SHAP were systematically dropped from the features considered. The classification performance was then reanalysed at different masking stages for each, as shown in Figure 15.

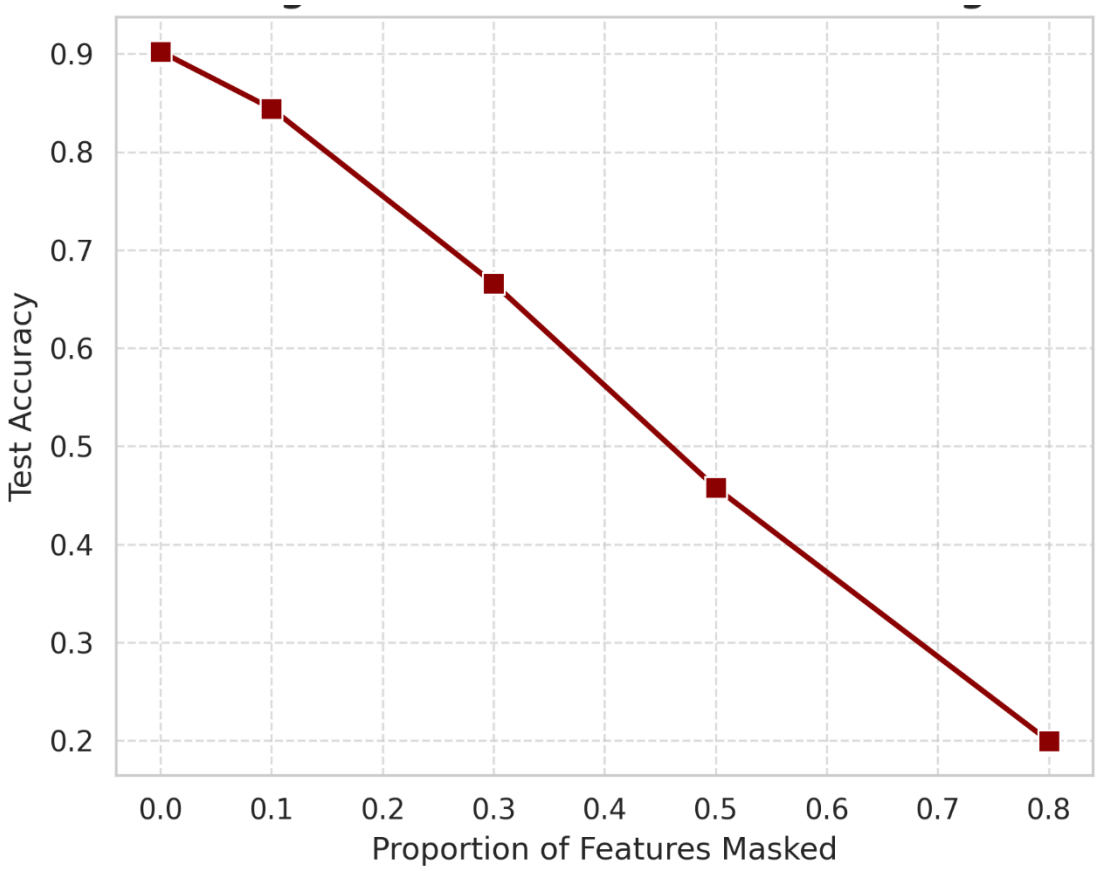


Figure 15. Performance of classifiers with progressive feature masking.

The results indicate a steady deterioration in the performance rather than a sudden failure. This is an observation of distributed acoustic information (no dominant descriptor) for achieving genre-recognition. The predictor continues to use complementary spectral and harmonic descriptors even after the removal of individual MFCC features because the predictor activity is based mostly on the information provided by the features. This behaviour in a distributed learning process makes the system more reliable and less prone to failures.

4.8.3 Random Seed Stability

The last robustness test was performed by running the experiment with different random initialisation seeds, as presented in Figure16.

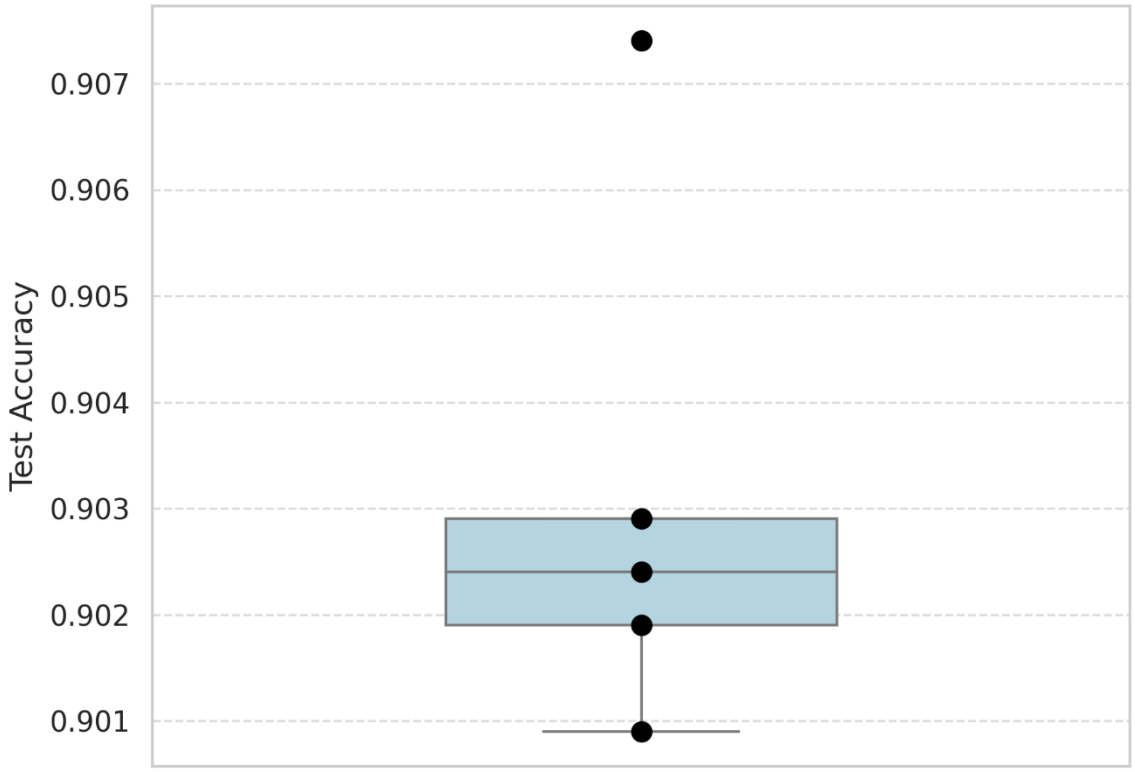


Figure 16. The accuracy of the classification for various random seeds is shown as a boxplot.

The obtained distribution indicated small variations, which showed the high stability and repeatability of the proposed solution. The narrow interquartile range shows that the model results were consistent across independent training runs, further supporting the robustness and reliability of the model results.

4.9 AI-Assisted Composition Guidance

The learned feature space prototype offers guidance for classical music composition using AI by capturing the representative acoustic features of classical music. The Classical prototype has a tempo of approximately 126 BPM, Chromo STFT Mean = 0.2645 indicating moderate harmonic intensity, a spectral centroid = 1351.3 Hz, a spectral bandwidth = 1519.5 Hz, and a spectral rolloff = 2475.8 Hz. The clearest timbral signatures are contained in the first, second, and fourth MFCCs, which are collectively the most distinctive parts of the genre envelope. These prototype values can be used as acoustic targets in generative music systems to produce compositions that are similar to the original classical music recordings.

4.10 Composition Profile: Classical

Table 11 summarises the classical genre prototype, which has a moderate harmonic intensity, smooth spectral properties, and a tempo of ~125.9 BPM. Its acoustic features are mostly characterised by Mfcc1, Mfcc2, and Mfcc4, and by the average value of 1351.3 of its spectral centroids and the 1519.5 average value of its bandwidth. These features form a representative acoustic plan that can guide the AI-assisted creation of classical music compositions in a consistent style and genre.

Table 11. Classical Genre Prototype Profiles.

Acoustic Attribute	**Prototype Value**
Tempo	125.9 BPM
Chroma STFT Mean	0.2645
Harmony Mean	-0.0002
Spectral Centroid	1351.3 Hz
Spectral Bandwidth	1519.5 Hz
Spectral Rolloff	2475.8 Hz
MFCC1 Mean	324.32
MFCC2 Mean	147.04
MFCC4 Mean	26.85

The acoustic prototype of the classical genre is shown in Figure 17, and all salient rhythmic, harmonic, spectral, and timbral characteristics extracted from the feature space are presented.

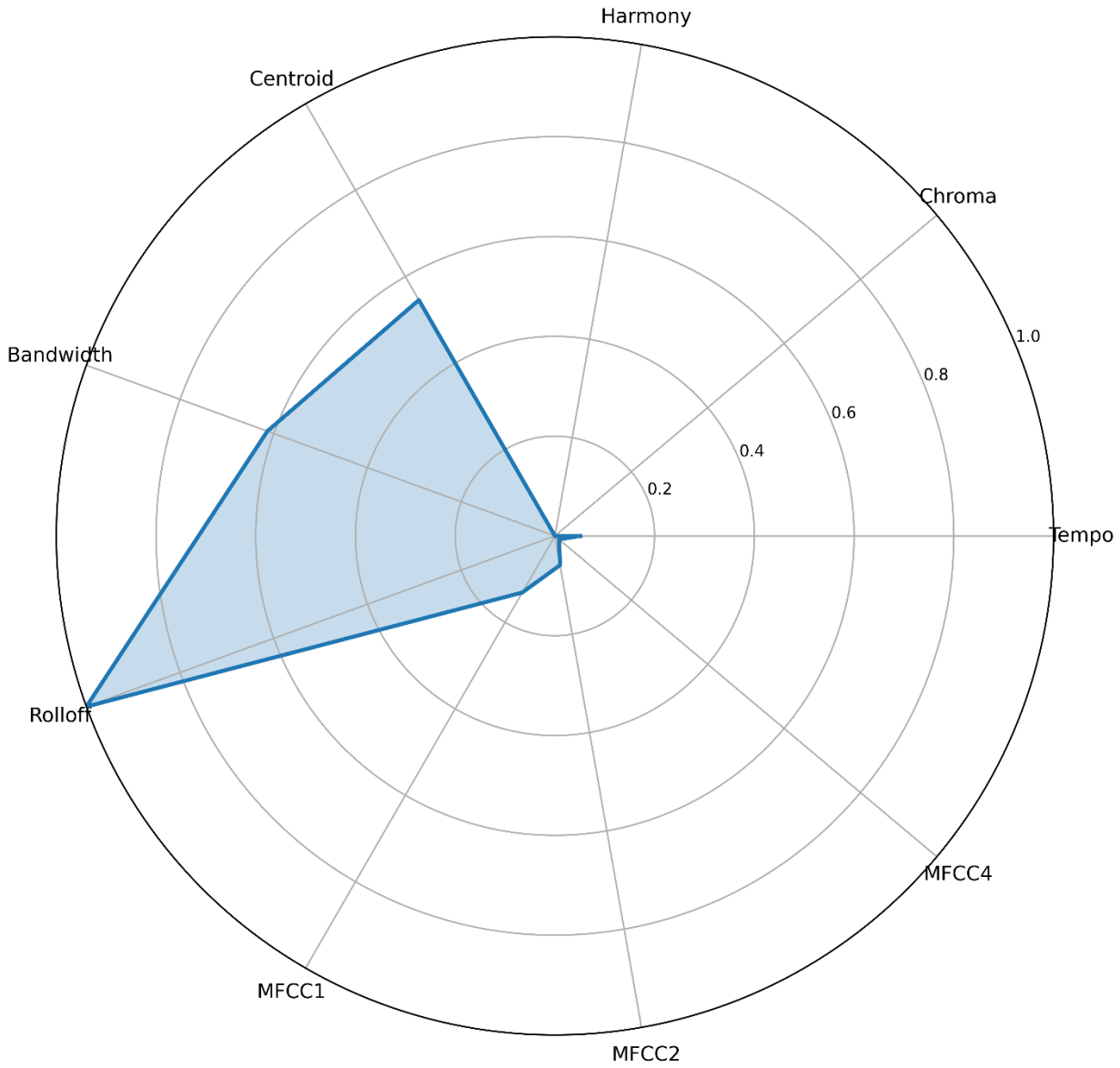


Figure 17. Radar plot of the classical acoustic prototype.

The composition profile, in turn, shows how classification knowledge can be made salient to create actionable guidance for creative AI systems. This ability is a substantial improvement over current genre recognition methods, such as GN, and is also indicative of the potential of explainable acoustic feature learning in music informatics and AI-aided composition.

5. Discussion

Based on the outcomes of this study, we see the potential of using automated machine learning combined with special acoustic feature engineering to address complex challenges in MIR. The trained framework successfully identified different traditional and ethnic music genres with 94.59% validation accuracy, which demonstrates the ability of the developed models to achieve high-performance results with music stream data, which is a poorly explored domain by traditionally developed AI applications. This performance represents a major leap over existing deep learning architectures, demonstrating that a combination of multiple model strengths produces better performance for complex audio data than any one-per-algorithm method of dealing with complex data.

5.1 Effectiveness of the Acoustic Feature Pipeline

The success of the classification system depends heavily on the comprehensiveness and accuracy of the extracted features. To capture a multi-dimensional "acoustic fingerprint" of each track, 40 Mel-Frequency Cepstral Coefficients and descriptors were combined, including spectral centroid, chroma features, and spectral rolloff. The spectral centroid and rolloff variables were significant for distinguishing the timbral "brightness" of the timbres between different genres with similar rhythmic structures. In addition, it was also possible to use ‘chroma features' to extract harmonic content and pitch distributions, which are vital for the recognition of the distinctive melodic characteristics of traditional music. These descriptors are used together, and the high degree of accuracy confirms that the combination of descriptors reflects a representation of musical identity with a degree of comprehensiveness and congruence with the human auditory system.

5.2 Advantages of Ensemble Optimization and AutoML

One of the key aspects of this study was the application of an automated machine learning workflow to optimise over 160 different models. The final Weighted Ensemble L2 exceeded all other classifiers by using the complementary power of numerous algorithms, including LightGBM and other neural designs. This automatic method removes the conventional constraint of manually tuning the hyperparameters, which hinders the scaling of an audio classification system. The final ensemble results indicate the reliability of the multilayer stacking method, which can be regarded as a standard method for future research on ethnic music recognition because of the difficulty of traditional model selection.

5.3 Bridging the Interpretability Gap

One of the biggest challenges in the current era of AI is that AI models are "black-boxes": they perform very well but have no transparency. In this study, the problem was solved by embedding SHapley Additive exPlanations in the classification workflow. The system offers a degree of transparency that is uncommon for deep learning, potentially highlighting the acoustic qualities, or the corresponding spectral coefficients or harmonic patterns, that were most useful for making a prediction for a genre. It is not only a technical diagnostic but also a means to build trust with users, especially musicians and researchers, who need to know the reasons for an AI's conclusion. These feature contributions are represented visually, which allows one to know that the model is training meaningful musical patterns and not noise or artefacts in the data.

5.4 Practical Utility of Genre Prototypes in Composition

Breaking away from mere classification, "genre prototypes" made the transition to composition with the help of AI. Each genre has an average value for each feature in every segment of the song. We determined the “acoustic center” of each genre by averaging the feature values across all segments of that genre. This enables a new method of generative support by measuring the Euclidean distance between a new composition and the targets. It offers a way to obtain a numerical approximation of how much a created segment reminds a listener of a typical example of the genre. The structured approach provides a roadmap for musicians interested in leveraging AI as a collaborative agent, transforming the role from something that labels music to something that actively shapes the creation of culturally congruent sound.

5.5 Cultural Implications and Future Directions

The results of this study have broader implications for the preservation and revitalisation of cultural heritage. This framework offers a reliable and clear way to structure music from the region and ethnic groups and would facilitate the digitisation and organisation of vast music archives that might otherwise be unavailable. However, it is crucial to recognise that computational measures do not adequately represent the musical and emotional richness of a piece. Further studies are needed to improve the models to incorporate the ‘perceptual gap' between the mathematical data of features and human aesthetic judgement. Furthermore, integrating a wider range of world-class music and testing

its performance and stability in an actual noisy recording context will be key to developing a truly global music AI assistant.

6. Conclusion

The research work proposed in this study introduced an integrated framework for the classification of music genre and composition through acoustic feature learning using MFCC-based feature extraction, optimising the ensemble using Explainable AutoML, and modelling of music genre prototypes. A complex array of spectral, harmonic, and timbral descriptors were extracted and studied from the GTZAN dataset to reveal the characteristics of a genre. The AutoML framework automatically investigated various machine learning architectures and created optimised ensemble models, and the best validation accuracy was achieved by WeightedEnsemble_L3 (94.59%). The results prove that MFCC features with ensemble learning are powerful and accurate for genre recognition through multiclass genre classification. Additionally, SHAP-based explainability highlighted the role played by each acoustic feature, which not only made the models more transparent but also enhanced the interpretability of the decisions made. More than just a classification approach, the provided ""genre prototype learning" innovations enabled the re-mapping of the learned acoustic representations into ""genre specific feature targets" which can be used for music generation based on AI and consistency evaluation. In summary, the proposed framework achieved good music recognition performance and tangible acoustic knowledge, making the mathematical model of music interpretable, which helps bridge the gap between music recognition and creative music generation. These results suggest that explainable ensemble learning is a viable starting point for the development of intelligent systems that can assist in music analysis, preservation, and AI-based music composition.

Author Contributions: Conceptualisation, T.Z. and Y.L.; methodology, T.Z.; software, T.Z.; validation, T.Z. and Y.L.; formal analysis, T.Z.; investigation, T.Z. and Y.L.; resources, T.Z. and Y.L.; data curation, T.Z.; writing—original draft preparation, T.Z.; writing—review and editing, T.Z. and Y.L.; visualisation, T.Z.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of this manuscript.

Funding: This research did not receive any external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset analysed in this study is the publicly available GTZAN Music Genre Classification Dataset. The dataset can be accessed through Kaggle at https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification. The processed features and analytical results generated during this study are available from the corresponding author upon reasonable request.

Acknowledgements: The authors acknowledge the publicly available GTZAN music dataset used in this study, as well as the computational resources and open-source software tools that supported the analysis.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] Y. Ding, "Evolution and Emerging Trends in Musical Information Retrieval: A Comprehensive Review and Future Prospects," *Highlights in Science, Engineering and Technology,* vol. 85, pp. 1041-1045, 03/13 2024, doi: 10.54097/nwj62j56.

[2] J. Mycka and J. Mańdziuk, "Artificial intelligence in music: recent trends and challenges," *Neural Computing and Applications,* vol. 37, no. 2, pp. 801-839, 2025/01/01 2025, doi: 10.1007/s00521-024-10555-x.

[3] M. A. Budyputra, A. Reyfanza, A. A. S. Gunawan, and M. E. Syahputra, "Systematic Literature Review of The Use of Music Information Retrieval in Music Genre Classification," *International Journal of Computer Science and Humanitarian AI,* vol. 2, no. 1, pp. 9-14, 02/20 2025, doi: 10.21512/ijcshai.v2i1.13019.

[4] L. Moysis *et al.*, "The Challenges of Music Deep Learning for Traditional Music," in *2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, 28-30 June 2023 2023, pp. 1-5, doi: 10.1109/MOCAST57943.2023.10176775.

[5] Y. Guo, "Research on music genre recognition method based on deep learning," *Molecular & Cellular Biomechanics,* vol. 21, no. 1, p. 373, 10/23 2024, doi: 10.62617/mcb.v21i1.373.

[6] J. Liang, "Harmonizing minds and machines: survey on transformative power of machine learning in music," (in English), *Frontiers in Neurorobotics,* Review vol. Volume 17 - 2023, 2023-November-10 2023, doi: 10.3389/fnbot.2023.1267561.

[7] P. Naidu K, B. J. Rao, K. Revathi, M. G. Gayathri, and G. S. Jayasri, "Classification of music genres using multimodal deep learning technique," in *E3S Web of Conferences*, 2025, vol. 616: EDP Sciences, p. 02012.

[8] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust Sound Event Classification Using Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 3, pp. 540-552, 2015, doi: 10.1109/TASLP.2015.2389618.

[9] P. Rawat, M. Bajaj, S. Vats, and V. Sharma, "A comprehensive study based on MFCC and spectrogram for audio classification," *Journal of Information and Optimization Sciences,* vol. 44, no. 6, pp. 1057-1074, 2023.

[10] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access,* vol. 11, pp. 106620-106649, 2023, doi: 10.1109/ACCESS.2023.3318015.

[11] M. Abebe, L. Endashew, J. Heikkonen, R. Kanth, and S. K. Mohapatra, "Decoding Cultural Music Classification with Machine Learning and Segment Length Analysis," *SN Computer Science,* vol. 7, no. 2, p. 146, 2026/01/30 2026, doi: 10.1007/s42979-026-04747-6.

[12] S. Prajuli, A. Karna, and O. Dhakl, "Music Genre Classification: A Comparative Analysis of Classical Machine Learning and Deep Learning Approaches," *arXiv preprint arXiv:2603.15440,* 2026.

[13] K. Mogonediwa, "Music Genre Classification: Training an AI model," *arXiv preprint arXiv:2405.15096,* 2024.

[14] A. Akman, Q. Sun, and B. W. Schuller, "Audio Explanation Synthesis with Generative Foundation Models," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6-11 April 2025 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10890370.

[15] S. A. Patil, G. Pradeepini, and T. R. Komati, "Novel mathematical model for the classification of music and rhythmic genre using deep neural network," *Journal of Big Data,* vol. 10, no. 1, p. 108, 2023/06/21 2023, doi: 10.1186/s40537-023-00789-2.

[16] T. Smith and G. Golcarenarenji, "Music Genre Recognition: An Explainable Approach," in *2025 9th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, 27-28 June 2025 2025, pp. 1-8, doi: 10.1109/ISAS66241.2025.11101724.

[17] A. E. Jenifer, K. S. Abirami, and M. Rajeshwari, "Enhanced Audio Signal Classification with Explainable AI: Deep Learning Approach in Time and Frequency Domain Analysis," *Procedia Computer Science,* vol. 258, pp. 2372-2381, 2025/01/01/ 2025, doi: https://doi.org/10.1016/j.procs.2025.04.500.

[18] T. Senatori, D. Nardone, M. Lo Giudice, and A. Salvini, "Explainable Instrument Classification: From MFCC Mean-Vector Models to CNNs on MFCC and Mel-Spectrograms with t-SNE and Grad-CAM Insights," *Information,* vol. 16, no. 10, p. 864, 2025. [Online]. Available: https://www.mdpi.com/2078-2489/16/10/864.

[19] L. Moysis *et al.*, "Music Deep Learning: Deep Learning Methods for Music Signal Processing—A Review of the State-of-the-Art," *IEEE Access,* vol. 11, pp. 17031-17052, 2023, doi: 10.1109/ACCESS.2023.3244620.

[20] Y. Zhang and T. Li, "Music genre classification with parallel convolutional neural networks and capuchin search algorithm," *Scientific Reports,* vol. 15, 03/20 2025, doi: 10.1038/s41598-025-90619-7.

[21] P. Naidu K, B. J. Rao, K. Revathi, M. G. Gayathri, and G. S. Jayasri, "Classification of Music Genres using Multimodal Deep Learning Technique," *E3S Web Conf.,* vol. 616, p. 02012, 2025. [Online]. Available: https://doi.org/10.1051/e3sconf/202561602012.

[22] X. Li, "Intelligent Classification and Recognition of Ethnic Music Styles Using Siamese Learning Algorithms," in *2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT)*, 5-6 Sept. 2025 2025, pp. 01-07, doi: 10.1109/ICICNCT66124.2025.11232611.

[23] C. Papaioannou, I. Valiantzas, T. Giannakopoulos, M. Kaliakatsos-Papakostas, and A. Potamianos, "A dataset for greek traditional and folk music: Lyra," *arXiv preprint arXiv:2211.11479,* 2022.

[24] N. Kamuni, M. Jindal, A. Soni, S. R. Mallreddy, and S. C. Macha, "Exploring Jukebox: A Novel Audio Representation for Music Genre Identification in MIR," in *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*, 3-4 May 2024 2024, pp. 1-6, doi: 10.1109/AIIoT58432.2024.10574541.

[25] V. Franzoni, "Cross-domain synergy: Leveraging image processing techniques for enhanced sound classification through spectrogram analysis using CNNs," *Journal of Autonomous Intelligence,* vol. 6, no. 3, p. 8, 2023.

[26] L. Comanducci, F. Antonacci, and A. Sarti, "Interpreting End-to-End Deep Learning Models for Speech Source Localization Using Layer-Wise Relevance Propagation," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, 26-30 Aug. 2024 2024, pp. 316-320, doi: 10.23919/EUSIPCO63174.2024.10715394.

[27] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, and W. Samek, "AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark," *Journal of the Franklin Institute,* vol. 361, no. 1, pp. 418-428, 2024/01/01/ 2024, doi: https://doi.org/10.1016/j.jfranklin.2023.11.038.

[28] V. Lyberatos, S. Kantarelis, E. Dervakos, and G. Stamou, "Challenges and Perspectives in Interpretable Music Auto-Tagging Using Perceptual Features," *IEEE Access,* vol. 13, pp. 60720-60732, 2025, doi: 10.1109/ACCESS.2025.3555741.

[29] D. A. MSGE, "UNDERSTANDING THE DECISION-MAKING PROCESS OF MUSIC GENRE CLASSIFICATION WITH EXPLAINABLE ALGORITHMS," *MULTIMEDIA RESEARCH Учредители: Resbee Publisher,* vol. 8, no. 1, pp. 1-10, 2025.

[30] L. Sharma, "Analyzing the performance of machine learning models in music genre classification," *Int J Res Appl Sci Eng Technol,* vol. 11, no. 6, pp. 4196-4203, 2023.

[31] B. Schuller, J. Dorfner, and G. Rigoll, "Determination of nonprototypical valence and arousal in popular music: features and performances," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 2010, no. 1, p. 735854, 2010.

[32] Z. Cataltepe, Y. Yaslan, and A. Sonmez, "Music genre classification using MIDI and audio features," *EURASIP Journal on Advances in Signal Processing,* vol. 2007, no. 1, p. 036409, 2007.

[33] J. Ramírez and M. J. Flores, "Machine learning for music genre: multifaceted review and experimentation with audioset," *Journal of Intelligent Information Systems,* vol. 55, no. 3, pp. 469-499, 2020/12/01 2020, doi: 10.1007/s10844-019-00582-9.

[34] Y. Sun *et al.*, "Automated Data Augmentation for Audio Classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 32, pp. 2716-2728, 2024, doi: 10.1109/TASLP.2024.3402049.

[35] D. Chi, "Audio Style Conversion Based on AutoML and Big Data Analysis," *International Journal of Advanced Computer Science & Applications,* vol. 15, no. 1, 2024.

[36] Y. Wang, J. Jiang, M. Zhang, and H. Zhang, "SHAP-LM: Optimized Music Genre Classification via SHAP-Guided Feature Selection and Lightweight Modeling," *IEEE Access,* vol. 14, pp. 27733-27742, 2026, doi: 10.1109/ACCESS.2026.3664687.

[37] D. T. Speckhard, K. Misiunas, S. Perel, T. Zhu, S. Carlile, and M. Slaney, "Neural architecture search for energy-efficient always-on audio machine learning," *Neural Computing and Applications,* vol. 35, no. 16, pp. 12133-12144, 2023/06/01 2023, doi: 10.1007/s00521-023-08345-y.

[38] A. Mendes, P. Trigo, and J. Paulo, "Hyperparameter Optimization for Impulsive Sound Classifiers," *Procedia Computer Science,* vol. 237, pp. 618-625, 2024/01/01/ 2024, doi: https://doi.org/10.1016/j.procs.2024.05.147.

[39] N. Bryan-Kinns, B. Zhang, S. Zhao, and B. Banar, "Exploring Variational Auto-encoder Architectures, Configurations, and Datasets for Generative Music Explainable AI," *Machine Intelligence Research,* vol. 21, no. 1, pp. 29-45, 2024/02/01 2024, doi: 10.1007/s11633-023-1457-1.

[40] S. Chakrabarty, P. Bishwas, M. Bandyopadhyay, and J. Sublime, "Can We Trust AI With Our Ears? A Cross-Domain Comparative Analysis of Explainability in Audio Intelligence," *IEEE Access,* vol. 13, pp. 179733-179758, 2025, doi: 10.1109/ACCESS.2025.3622161.

[41] Q. Wang and M. Jiu, "Deep Multi-Modal Kernel Map Network for Music Genre Classification," *Algorithms,* vol. 19, no. 6, p. 467, 2026. [Online]. Available: https://www.mdpi.com/1999-4893/19/6/467.

[42] D. Li and S. Du, "Investigation of a music genre data classification method based on an improved ECAPA-TDNN algorithm," *International Journal of Information and Communication Technology,* vol. 27, no. 13, pp. 38-60, 2026.

[43] Y. Sun, "Feature centric based deep learning approach for music mood recognition with HuBERT transformer model," *Scientific Reports,* vol. 15, no. 1, p. 45460, 2025/11/27 2025, doi: 10.1038/s41598-025-29200-1.

[44] L. Cheng, "The impact of generative AI on school music education: Challenges and recommendations," *Arts Education Policy Review,* vol. 126, no. 4, pp. 255-262, 2025/10/02 2025, doi: 10.1080/10632913.2025.2451373.

[45] T. Wang and J. Zhang, "Enhancing musical creativity through AI: a chain-mediation model of self-efficacy and emotional intelligence," (in English), *Frontiers in Psychology,* Original Research vol. Volume 16 - 2025, 2025-November-05 2025, doi: 10.3389/fpsyg.2025.1665609.

[46] Q. Zhang, "Application of generative AI in composition assistance and creative music teaching," *International Journal of Information and Communication Technology,* vol. 27, no. 19, pp. 70-91, 2026.