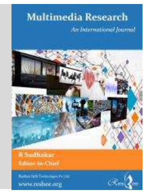
Vol.8 No.1 Jan 2025

Multimedia Research

Received 11 November, Revised 21 November, Accepted 4 January



Understanding the Decision-Making Process of Music Genre Classification with Explainable Algorithms

Msge Desta A.
Artificial Intelligence and Data Science
Bahirdar Institute of Technology
Bahirdar, Ethiopia
misge711@gmail.com

Chalachew Mulu C.
Artificial Intelligence and Data Science
Bahirdar Institute of Technology
Bahirdar, Ethiopia
tolerantfull@gmail.com

Abdulkerim Mohammed Y.
Artificial Intelligence and Data Science
Bahirdar Institute of Technology
Bahirdar, Ethiopia
abdulkerimm@gmail.com

Abstract: Music genre classification is a fundamental task in music information retrieval, but the lack of interpretability in many machine learning models poses a significant challenge. To address this issue, we propose using a Random Forest Classifier (RFC), a simple and explainable artificial intelligence (XAI) technique, for music genre classification. Our objective is to develop a model that can identify the genre of a given music sample based on its audio features, while also providing insights into which features are most important for each genre. By doing so, we aim to gain a deeper understanding of the underlying patterns in music data and enable the development of more effective music-related applications. In this study, we combine two datasets and employ XAI techniques to gain insights into the decision-making process of the model while addressing the issue of class imbalance in the dataset. Overall, our approach provides a comprehensive and effective solution to music genre classification using XAI tools and methods.

Keywords: Class Imbalance, Explainable Machine Learning, Interpretable Models, Music Genre Classification, Random Forest Classifier

Impact statement: Music genre classification is essential in a variety of music-related applications. It provides a better user experience and personalized music recommendations. However, the current music genre classification system has flaws that limit its usefulness in assuring fair royalties, delivering accurate classifications, and generating significant data insights. To address the problems of the current system, we suggest a new approach to music genre classification with explainable algorithms. Our strategy can address the inadequacies of the current classification system by providing more accurate genre classification, resulting in a better music listening experience. In addition, the interpretable model can provide useful insights into music data patterns, which can improve music research. The proposed technique assists both music enthusiasts and experts by providing an accurate music genre classification system. Overall, this technique has the potential to increase the efficiency of the music industry and assist music specialists in making informed decisions.

Nomenclature

Abbreviations	Descriptions
AUC / ROC	Area Under the Curve / Receiver Operating Characteristics
CNN	Convolutional Neural Network
DT	Decision Tree
GTZAN	GTZAN Dataset
KNN	K-Nearest Neighbors
LR	Logistic Regression
LIME	Local Interpretable Model-agnostic Explanations
MIR	Music Information Retrieval
MABD	Music Audio Benchmark Dataset
MFCC	Mel-Frequency Cepstral Coefficients
RFC	Random Forest Classifier
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence
ML	Machine learning
DL	Deep Learning

https://doi.org/10.46253/j.mr.v8i1.a1

1. Introduction

Music genre classification is a key task in MIR that involves automatically classifying music into different genres based on audio features [1]. ML algorithms have been frequently utilized for this task because they can learn complicated patterns in audio inputs and identify genres accurately. However, the understanding of these algorithms is frequently limited, making it difficult to grasp how they arrive at their predictions. RF is a frequently used ML method that can be used to classify music genres [3]. We used RFC to train a non-linear function that maps audio features to output genre labels while also providing insight into which features are most essential for each genre [4]. This approach allows us to not only produce correct genre predictions but also acquire a deeper understanding of the underlying patterns in the music data, making it a powerful tool for music genre classification.

XAI can provide insights into the model's decision-making process, assisting us in understanding why it made a particular classification [5]. XAI can also determine which elements are most significant in determining the genre, hence boosting the accuracy of music categorization systems.

XAI enhances the model's accountability and transparency while avoiding the ethical issues that come with utilizing black-box models [6]. Thus, we suggest an RFC model using XAI for this purpose, which is crucial for enhancing the precision and transparency of ML algorithms in music genre classification.

The paper is organized as follows: Section 2 presents a literature review on music genre classification and explainable AI techniques. Section 3 describes the methodology, including the dataset, preprocessing, feature engineering, and the classification models used. Section 4 details the experiments, covering model training, optimization, and evaluation. In Section 5, we discuss the results and provide insights into model performance and interpretability. Finally, Section 6 concludes with a summary of findings and suggestions for future research.

2. Literature Review

The classification of musical genres has long been a divisive area of study in the field of MIR. Recent advancements in ML, DL, and XAI techniques have made musical genre classification more accurate and understandable. This review will cover the most recent research on ML-based music genre classification using XAI.

DT-based algorithms are among the most extensively used XAI algorithms for music genre classification. Beats per minute, spectral properties, and rhythmic patterns have all been utilized to categorize distinct musical genres using DT algorithms such as C4.5 and ID3 are examples of how music genres were identified using DT algorithms based on the tempo and rhythmic patterns of the song [7].

Another interesting XAI method for music genre classification is rule-based algorithms. The patterns and rules specific to each genre were discovered utilizing rule-based algorithms such as association rule mining and decision rule learning. For illustration, [8] employed association rule mining to identify repeating patterns in musical features and label musical genres based on these patterns.

Music genre classification has also been employed in DL algorithms such as CNNs and RNNs. These algorithms can learn complicated encodings of musical elements and record long-term connections between them. However, these models are frequently referred to as "black-box" models, and it can be difficult to understand how they make judgments. To address this issue, XAI techniques such as saliency maps and attention mechanisms have been used to show the features most important for classification [9].

Poonia *et al.* [10] used traditional ML and DL approaches for comparing the performance of the classifier. They used the GTZAN dataset for their model by extracting signal waveform translated into a spectrogram image as an input. The authors used four classification models; KNN, SVM, and RF for traditional ML and CNN for DL. They trained their model using a 3-second clip of the song. They experimented on the four models and CNN provided better performance with a score of 99.8% for training and 80.0% for testing than the other model.

The transfer learning approach to classify music was also investigated by [11]. The authors used the GTZAN dataset (2000-2001) for 10 genres with each genre containing 100 WAV files and having a 30-second duration they generalized that the idea of transfer learning can perform the classification process faster and with more successful results than humans with an average testing accuracy of 95.8%.

Xu *et al.* [12] developed a music classifier using a multilayer classifier based on SVM. They have used four labels for the classification; classic, jazz, pop, and rock. Feature selection techniques like beat spectrum, zero crossing rate, spectrum power, and MFCC were used. In the data collection, they have used CDs and the Internet as their data source. Their experiment has found that SVM provides better results with an error rate of 6.86% than other models.

Cheng & Kuo [13] built music genre classification based on the visual Mel spectrum. In the paper GTZAN dataset for 10 genres with 100 songs in each genre which have 10-second duration was used. In their study, 70% dataset was used for training, 20% for testing, and 10% for validation. Results show that music classification with Mel spectrum has great performance with an average test result of 97.93%.

Viswanathan [14] conducted a study on music genre classification using ML techniques for grouping similar music as one. They used the GTZAN dataset for 10 genres having 100 music files in each and a 30-second duration is used. The author used the Mel-frequency cepstral coefficient for audio feature extraction. 80% of the dataset for training and 20% of the dataset for testing and results show that the proposed traditional ML approach, KNN provided a test accuracy of 70%.

Finally, XAI algorithms have improved the precision and interpretability of music genre classification models [15]. DL techniques, DT-based algorithms, rule-based algorithms, and hybrid models have all shown promising results when categorizing musical genres. More research is needed, however, to develop XAI algorithms that can handle the great unpredictability and complexity of musical data and provide a better understanding of the model's decision-making process [16].

2.1 Review Table

Table 1 contains the details of the dataset and methodology used in the existing manuscript. GTZAN was the most commonly used dataset and some custom datasets are also used by the author.

Table 1. *Review of music genre classification*

Author	Methodology	Dataset	Year
Poonia et al. [10]	Traditional ML (KNN, SVM, Random Forest) and Deep Learning (CNN); compared classifier performance.	GTZAN	2022
Xu et al. [12]	Multilayer classifier with Support Vector Machine (SVM); used beat spectrum, zero crossing rate, spectrum power, and MFCC for feature selection	Custom dataset from CDs and Internet sources	2003
Cheng & Kuo [13]	Music genre classification using visual Mel spectrum	GTZAN	2022
Viswanathan [14]	Traditional ML approach with KNN for grouping similar music; used Mel-frequency cepstral coefficient for audio feature extraction	GTZAN	2016
Song et al. [11]	Transfer learning for genre classification; transfer of learned features for improved performance	GTZAN	2017
Arjannikov & Zhang [8]	Rule-based genre classification using association rule mining and decision rule learning	Custom dataset	2014

3. Methodology

This section describes the materials and methods used in the proposed study, which aims to improve music categorization and analysis using the XAI technique, which allows users to understand the inner workings of ML models, making it easier to interpret and explain their classification. In the following sections, we present a detailed step-by-step approach for creating a music genre classification model using an explainable ML algorithm technique.

3.1 Dataset

Combining datasets can produce a wider, more varied pool of data, which can lead to more accurate results and a deeper understanding of musical genres [17]. Because of this, we use the GTZAN dataset in conjunction with the MABD. GTZAN dataset has 10 genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock). Each genre has 100 .wav format audio files with a 30-second duration [1]. MABD has 9 genres (blues, folk country, electronic, hip-hop, jazz, pop, rock, alternative, and funk soul R&B) Each genre has a .mp3 format audio file with an 11-second duration [18]. Here, except for the last two genres (alternative and funk soul R&B) of MABD, we have used the other 7 genres with the GTZAN dataset in combination.

3.2 Data Preprocessing

We began the data preprocessing stage by dividing the dataset into three-second intervals to increase its size. We then separated the labels and features from the dataset to prepare it for ML. Because ML models can only interpret numerical inputs, we utilized Label Encoder to convert the labels' categorical text values to numeric values. Likewise, we used Standard Scaler to normalize the features so that they all had the same standard values of a mean of zero and a standard deviation of one. This phase was critical to avoid having one particular attribute dominate the model training and testing processes [19].

However, upon examining the dataset using the Matplotlib tool, we identified an unbalanced class distribution. To overcome this issue, we utilized SMOTE to increase minority classes and balance the dataset. **SMOTE** is an oversampling strategy used to balance unbalanced datasets by interpolating feature values across minority class samples and their KNN. It generates synthetic samples for the minority class [20]. A bar chart illustrated in Fig. 1 shows a skew in the dataset favoring some music genres over others and a balanced dataset after SMOTE was applied.

Fig. 1. *Comparison of music genre distributions: the left graph illustrates an imbalanced dataset, with rock containing more songs than other genres, while the right graph shows a balanced dataset, where each genre has an equal number of songs.*

To overcome this issue, we utilized SMOTE to increase minority classes and balance the dataset. Smote**,** an oversampling strategy used to balance unbalanced datasets by interpolating feature values across minority class samples and their KNN, generates synthetic samples for the minority class [20]. The second figure displays a visualization of the balanced dataset produced after SMOTE was applied.

3.3 Feature Engineering

In this phase, we applied Librosa to extract the necessary features from the music data. The algorithm's performance is determined by its capacity to extract features in the time, frequency, and time-frequency domains, which are all chosen based on the properties of the audio data. All of these characteristics are required for the method to function properly [21].

The feature extraction step includes converting the audio signal input into a set of features that can represent the audio signal. The major types of features utilized in musical genre classification include time-domain, frequency-domain, and time-frequency domain features. Time-domain features describe the signal's temporal qualities, whereas frequency-domain features describe the signal's spectrum characteristics, and time-frequency features combine information about the signal's temporal and spectral characteristics. The suitable feature set for music genre classification is determined by the application and the classification method used, as different algorithms require different features or combinations of features [22].

The use of a combination of feature sets that include time-domain, frequency-domain, and time-frequency domain features can significantly improve the accuracy of the classification model [23]. Due to that, we have used the combination of time domain, frequency domain, and time-frequency domain features. To summarize, we extracted the following feature in Table 2 from the combined audio dataset.

Table 2. *List of time-domain, frequency domain, and time-frequency domain features extracted from the Librosa*

Features Extracted in the Librosa library	
Mel Frequency Cepstral Coefficient	Tomogram
Power Spectrum Chromatogram	Polynomial Feature
Chroma Energy Normalized	Spectral Centroid
Constant-Q Chromatogram	Spectral Contrast
Mel Scale Spectrogram	Spectral Bandwidth
Flatness	Spectral Rolloff
Zero Crossing Rate	Tonal Centroid Feature
Root Mean Square	-

3.4 Model Building

In this stage, we used preprocessed data and feature-engineered features to train an ML model for music genre classification. To achieve this, we investigated several methods such as RFC, DT, and LR. Each model was chosen for its potential to provide insights into model interpretability while handling the complexities of genre classification. Although DL models like CNNs are widely used for audio classification tasks due to their ability to capture complex patterns, they are often considered "black-box" models [9]. This lack of interpretability makes it challenging to understand their internal decision processes, which is a primary objective of this study. Moreover, CNNs typically require larger datasets and higher computational resources, which may limit their feasibility in certain MIR applications. By using RFC,

combined with explainable AI techniques (LIME and SHAP), we achieve an interpretable and accurate model suited for music genre classification while maintaining transparency in the classification process.
To properly analyze the model's performance, we separated the dataset into two subsets: a training set that comprised 80% of the data and a testing set that comprised 20%. We then trained the models using default parameters and showed the accuracy results of each algorithm. Moreover, we adjust the model's parameters using Bayesian optimization methods to enhance the model's performance.

Bayesian optimization is a mathematical framework that helps to optimize black-box functions by reducing the number of function evaluations required to identify the optimal parameters for a given problem. By using this technique, we were able to tune the hyper-parameters of the RFC model to achieve maximum accuracy [24].

After performing hyper-parameter tuning, we saw a considerable improvement in the performance of the RFC model for music genre classification. We discovered that RFC is an easy-to-understand and effective algorithm for music genres by modifying the hyper-parameters. The model takes into account a variety of input features and generates a clear output indicating the likelihood of each genre. The optimized hyper-parameters and their associated values are shown in Table 3 below.

Table 3. *Presents the optimized hyper-parameters for the RFC model achieved through Bayesian optimization. Key hyper-parameters include maximum depth, minimum samples per leaf, minimum samples required to split an internal node, and number of trees, which were adjusted to enhance accuracy.*

Parameter	Description	Value
max-depth	The maximum depth of the tree; limits how deep each tree can grow	None
min_samples_leaf	The minimum number of samples required in a leaf node	1
min_sample_split	The minimum number of samples required to split an internal node	2
n_estimators	Number of trees in the forest	500

In addition, we used methodologies to improve the transparency and the interpretability of the model. These strategies enabled us to better understand the decision-making process and evaluate the model [25]. This method was quite useful and valuable for accurately classifying and interpreting the prediction of musical genres. Our goal was not only to achieve high accuracy but also to understand how the model works on the classification task. As a result, we employed XAI techniques to explain RF mode.

3.5 Model Evaluation

After training and optimizing the model, the next critical step is to assess its performance. To accomplish that, we employed metrics including Accuracy, Precision, Recall, ROC curve, Confusion Matrix, and F1-score. These metrics allow us to analyze the model's performance based on many parameters such as the number of correctly and erroneously categorized cases, the number of true and false positives, and the number of true and false negatives.

Confusion Matrix is a highly useful tool for analyzing the model's performance. This matrix displays how many examples are correctly and incorrectly classified across different categories, providing a visual representation of how the model is working [5].

Overall, evaluating the performance of an ML model is a critical step toward improving its accuracy and efficacy. We can obtain important insights into the model's performance and make data-driven decisions to improve its overall efficacy by combining metrics and visualization tools.

3.6 Model Interpretation

The most important stage in our research was interpreting the model using XAI ML techniques, which offer transparency and interpretability in the decision-making process of the model [26]. We employed a variety of techniques to discover the most relevant traits and aspects in the classification process, allowing us to better understand how the model makes its predictions. XAI algorithms are becoming increasingly crucial in promoting openness and accountability in ML applications.

We can better understand how and why ML models create predictions by using XAI approaches. This is critical in musical genre categorization since the high variability and complexity of musical data make it difficult for human experts to understand the model's decision-making process. XAI tools such as feature importance analysis, summary plot, and force plot can offer insight into the model's decision-making process by emphasizing the most essential features and elements that contribute to the classification task [16].

XAI approaches enable fair and accurate music genre classification by recognizing and correcting biases, eliminating ethical difficulties associated with black-box models. They improve precision, openness, and accountability by producing accurate predictions, flagging areas for improvement, and providing insights into the decision-making process [27]. XAI approaches are needed for accurate and understandable music genre classification models. Ultimately, we used the following XAI strategies to increase transparency and explanation in the music genre classification processes:

Local Interpretable Model-Agnostic Explanations (LIME): LIME is a sophisticated approach that provides insight into classifier prediction by creating an interpretable model approximation. More specifically, LIME is used in our situation to bring attention to the audio features that have the most influence over the classifier's decision-making process [5].

SHapley Additive exPlanations (SHAP): SHAP is a measure of feature importance that considers the interactions between features, giving a more complete and accurate representation of the classifier's decision-making. This method is very helpful for developing credible explanations for the model's conclusions [25]. The proposed model has the following architecture Fig. 2.

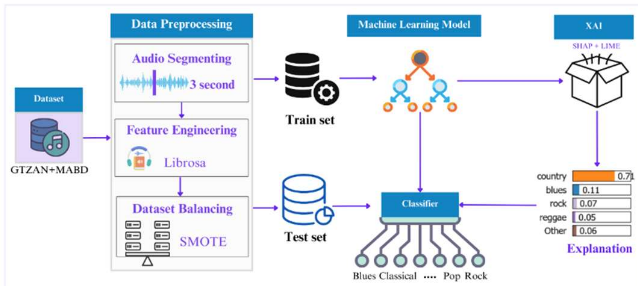


Fig. 2. *This figure illustrates the proposed Explainable ML architecture for music genre classification by combining two datasets: GTZAN and MABD. The combined dataset undergoes a preprocessing stage, which involves audio segmentation, feature engineering, and balancing. The preprocessed data is then used to train the model and the XAI explains the classification using SHAP and LIME.*

4. Experiments

The experiment was carried out on a Windows 10 platform using Python version 3.9.1 in a Jupyter Notebook environment. Our study includes comparing the performance of various ML models, such as the RFC, DT, and LR, in conjunction with XAI techniques. Our objective was to find the best model combinations for accuracy and transparency. To do this, we designed and carried out experiments that compared the performance of each model and combination, allowing us to discover the most effective strategies for achieving our goals.

In the first experiment, we compared the performance of different audio features. We extracted the mean, variance, and standard deviation of one feature (MFCC) for the first, six features (spectrogram, Mel-spectrogram, MFCC, spectral centroid, zero-crossing rate, and RMS energy) for the second, and fifteen features depicted in Table 4 for the third experiment. We found that the combination of features provided the best result with an accuracy of 85.25% using RFC.

In the second experiment, we compared the performance of RFC, LR, and DT before and after balancing the dataset for the combined feature. as seen in Table 5.

Table 4. *illustrates the effect of data balance on the accuracy of three different models: RFC, DT, and LR.*

Model	**Accuracy in % Before and After balancing**	
	Before	After
RFC	82.14	89.57
LR	65.32	68.04
DT	69.82	69.90

As demonstrated in Table 4, data balancing helped all three models, with the RFC model witnessing the greatest improvement in accuracy. After data balancing, the accuracy of the RFC model increased from 82.14% to 89.57%. The DT and LR models significantly improved in accuracy following data balancing, with accuracy of 69.90% and 68.04%, respectively.

In the third experiment, we evaluated the performance of the models by optimizing parameters using a Bayesian optimizer on a balanced dataset. The results of this experiment are presented in Table 5.

Table 5. *The accuracy of RFC before optimization was 89.57 but after optimization, its accuracy improved to 90.02. The table shows the impact of optimization on the accuracy of models including DT and LR.*

Model	**Accuracy in (%) after optimization**
RFC	90.02
LR	68.71
DT	70.14

As demonstrated in Table 5, all three models benefited from optimization, with the RFC model gaining the greatest improvement in accuracy. The DT and LR models significantly improved in accuracy following optimization, with accuracy of 70.14% and 68.71%, respectively. These results highlight the effectiveness of optimization in increasing the performance of ML models.

As part of our last experiment, we did an in-depth analysis of the model's accuracy on the test set by utilizing the advanced tools provided by XAI, including LIME and SHAP.

These interpretive tools allowed us to develop a more sophisticated understanding of the factors influencing the performance of our model on the test set. Using this strategy, we were able to uncover any potential weaknesses or biases in the model and make informed modifications as a result. By combining XAI and the RFC model, we achieved the goal of creating a more accurate and reliable model with a higher degree of transparency.

5. Result and Discussion

The results showed that the RFC model is a valid and successful strategy for the task at hand. However, it is vital to analyze the model's apparent flaws and identify potential future study and development opportunities. By incorporating explainable ML methodologies into our study, we were able to gain a full understanding of the RFC model's strengths and limitations. Overall, we were able to conduct a complete and rigorous examination of the model's performance using this method.

To study and assess the accuracy of the music genre classification model, we used a variety of performance indicators. The confusion matrix, ROC curve, precision, recall, and f1-score were all included.

To provide a deep and comprehensive understanding of the model's categorization abilities, we began by showing the confusion matrix. This matrix provided a tabular representation of the model's true positive, false positive, true negative, and false negative classifications [28]. We learned how well the model could accurately identify and differentiate between different music genres by studying these individual components. This helped us to have a better understanding of the classifier's performance and make more accurate and informed decisions about further optimization or model modification.

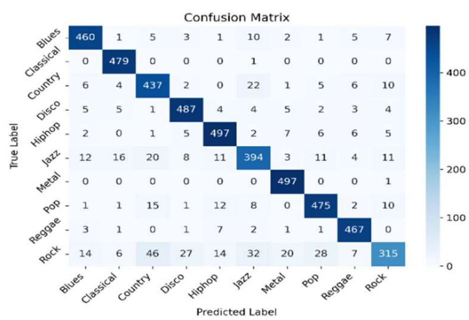


Fig. 3. *In this figure the rows reflect the true labels, while the columns indicate the predicted labels. The numbers in the cells show the number of cases categorized as having that specific combination of true and predicted labels.*

Following that, we visualized the ROC curve in Fig. 5, which allowed us to examine the model's capacity to distinguish between positive and negative classifications [29]. This provided us with information about the classifier's overall performance as well as its sensitivity to false positives and false negatives.

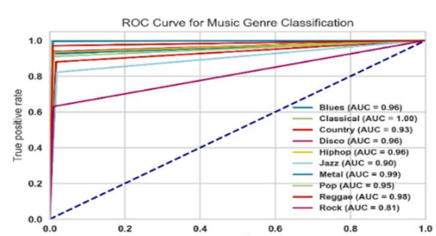


Fig. 4. *In this figure, the x-axis represents the false positive rate (FPR), while the y-axis represents the true positive rate (TPR). The line over the dotted line represents the ROC curve, while the area under each curve represents the AUC.*

Furthermore, we used classification reports, as depicted in Fig. 6, to demonstrate the model's precision, recall, and f1-score. These measurements were critical in understanding the classifier's ability to reliably categorize music genres.

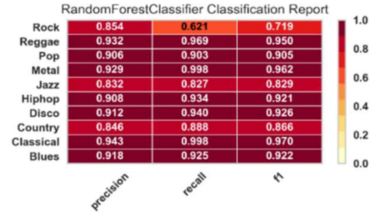


Fig. 5. *Presents the precision, recall, and f1-score for the music genre classification task. The report evaluates the model's performance for each genre label, including the model's ability to correctly classify each genre.*

Then, we used XAI approaches to acquire a better understanding of the model's predictions. SHAP was specifically employed to provide insights into the features that the model depended on for classification. To evaluate the results, we used summary and force plots. Similarly, we used LIME, which gave us feature importance graphs, allowing us to examine the value of the features that the classifier relied on for accurate predictions [25].

Using these XAI techniques, we were able to increase our understanding of the model's predictions and make informed decisions about the model to improve its classification accuracy. We choose RFC to learn more about the accuracy of the predictions.

SHAP offers a useful summary graphic for the top features in prediction, allowing for the identification of the most essential features that lead to correct genre prediction. As a result, this information proved useful in model development and identifying significant musical features that are highly related to distinct genres, as shown in Fig. 7.

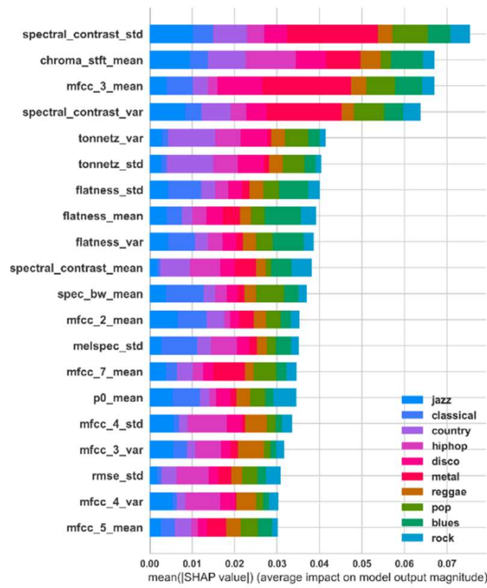


Fig. 6. *The figure shows the summary plot of the data that uses rectangular bars to display the SHAP values of different features. The bars are arranged horizontally and show the contribution of each feature to the final prediction model. The plot provides a clear visualization of each feature's contribution and significance in predicting the music genre.*

Then we used force plots to display the significance of each variable in predicting the genre for a selected sample of test sets to better examine the model's performance. This method is seen in the image below, and it allows for a more detailed analysis of the model's correctness.

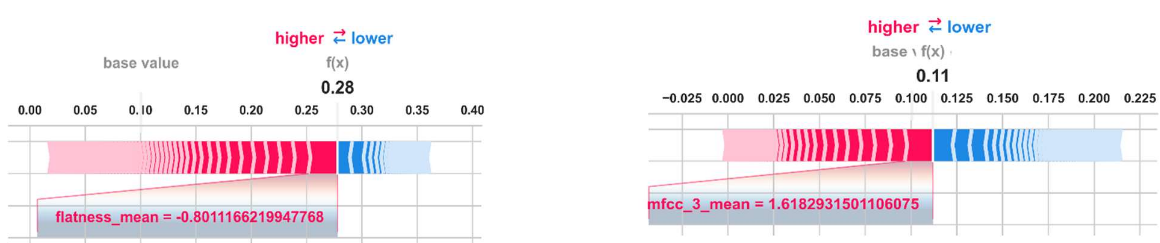


Fig. 7. *Shows the impact of each feature on the prediction of the model, including its direction and magnitude. The plot is color-coded, highlighting the positive contribution (red) and negative contribution (blue) of each feature to the prediction. The base value, which is the actual value of the feature, and the expected value (f(x)) are also included in the plot.*

Afterward, we used LIME to understand the model prediction pattern. In this section, we develop explanations for a sample of occurrences and show their predictions. This enabled the identification of biases and provided insights into how different features contributed to the prediction. The diagram below summarizes the prediction probabilities of the model from the given sets of features.

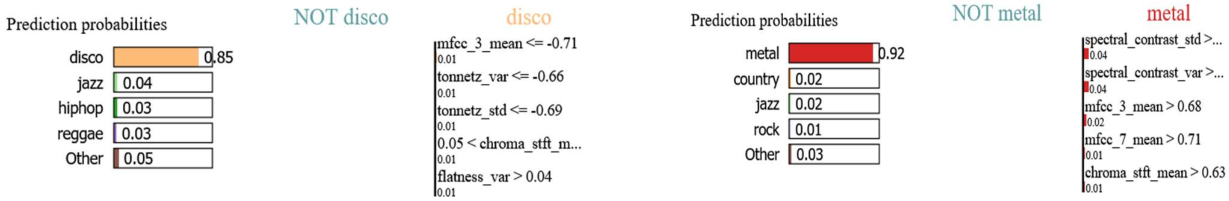


Fig. 8. *shows the local explanations provided by LIME for two instances. Each plot depicts the contribution of the top features to the predicted label for the relevant instance. The features are organized by contribution, with the most important feature at the top. The length of the bars denotes the size of the contribution, while the color of the bars indicates the type of music genre associated with the feature.*

The proposed method demonstrates several notable advantages. By integrating the RFC with explainable AI techniques (LIME and SHAP), the model achieves both accuracy and transparency in music genre classification. This combination allows users to understand the importance of specific features in the classification process, enhancing the interpretability of the model's predictions and making it well-suited for applications where interpretability is crucial.

However, the approach also has certain limitations. The RFC, while interpretable, may not match the performance levels of more complex DL models, which could impact accuracy in highly nuanced classification tasks. Additionally, the reliance on LIME and SHAP explanations introduces extra computational overhead, which could be a limitation when applied to larger datasets. Moreover, as the model is trained on a specific set of genres from combined datasets, its scalability to new genres or larger datasets may require extensive reconfiguration and parameter tuning.

6. Conclusion

Music genre classification is a critical task in MIR, especially with the increasing demand for personalized music recommendations and content-based retrieval systems. This study introduces an interpretable model that combines the advantages of the RFC with Explainable AI (XAI) techniques, specifically LIME and SHAP. Achieving 90.02% accuracy, the model ensures high performance while maintaining transparency in decision-making. LIME and SHAP explanations provide a comprehensive understanding of how various audio features contribute to genre classification, improving model interpretability and fostering user trust in predictions. These tools also help identify relevant features and trends in the data, ensuring that the model's outputs are meaningful and reliable. However, static visualizations from LIME and SHAP can be complex without context. Future work could incorporate interactive visualizations, allowing users to explore data and predictions in real-time to enhance usability. This approach has significant implications for MIR tasks that require interpretable models, addressing the growing demand for transparency in ML. Extending this research to integrate explainable AI with advanced DL models could further enhance accuracy while preserving interpretability. Applying the method to larger, more diverse datasets could improve scalability and reliability across a broader range of genres. Developing real-time, interactive tools for exploring feature contributions and predictions could also enhance user engagement and the utility of explainable AI in music classification.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] G. Tzanetakis, S. Member, and P. Cook, “Musical Genre Classification of Audio Signals”, vol. 10, no. 5, pp. 293–302, 2002. doi: 10.1109/TSA.2002.800560.

[2] C. Molnar, B. Bischl, and G. Casalicchio, “Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges”, PKDD/ECML Work., no. 01, pp. 1–15, 2020. doi: 10.1007/978-3-030-65965-3_28.

[3] L. Breiman, "Random Forests. Machine Learning", Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001. doi: https://doi.org/10.1023/A:1010933404324.
[4] R. Singhal, S. Srivatsan, and P. Panda, "Classification of Music Genres using Feature Selection and Hyperparameter Tuning", J. Artif. Intell. Capsul. Networks, vol. 4, no. 3, pp. 167–178, 2012. doi: https://doi.org/10.36548/jaicn.2022.3.003.
[5] M. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You ? Explaining the Predictions of Any Classifier", in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., 1998, pp. 445–453. doi: 10.5555/645527.657469.
[6] M. Brent, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The Ethics of Algorithms: Mapping the Debate", Big Data Soc., vol. 3, no. 2, 2016. doi: 10.1177/2053951716679679, 2016.
[7] G. M. Bressan, B. C. de Azevedo, and E. A. S. Lizzi, "A decision tree approach for the musical genres classification", Appl. Math. Inf. Sci., vol. 11, no. 6, pp. 1703–1713, 2017. doi: http://dx.doi.org/10.18576/amis/110617.
[8] T. Arjannikov and J. Z. Zhang, "An Association-based Approach to Genre Classification in Music", 15th Int. Soc. Music Inf. Retr. Conf. (ISMIR 2014), pp. 96–100, 2014.
[9] W. Zhang, "Music Genre Classification Based on Deep Learning", Mob. Inf. Syst., vol. 2022, pp. 1–11, 2022. doi: 10.1155/2022/2376888.
[10] S. Poonia, C. Verma, and N. Malik, "Music Genre Classifi cation using Machine Learning A Comparative Study", IITM J. Manag. IT, vol. 13, no. 1, pp. 15–21, 2022 [Online].
[11] G. Song, Z. Wang, F. Han, and S. Ding, "Transfer learning for music genre classification", In Intelligence Science I: Second IFIP TC 12 International Conference, ICIS 2017, Shanghai, China, October 25-28, 2017, Proceedings 2 (pp. 183-190). Springer International Publishing, 2017.
[12] C. Xu, N. C., X. S. Maddage, F. Cao, and Q. Tian, "Musical genre classification using support vector machines", 2003 IEEE Int. Conf. Acoust. Speech, Signal Process. 2003. Proceedings. (ICASSP '03)., vol.5, pp.V–429, 2003. doi: 10.1109/icassp.2003.1199998.
[13] Y. H. Cheng and C. N. Kuo, "Machine Learning for Music Genre Classification Using Visual Mel Spectrum", Mathematics, vol. 10, no. 23, pp. 4427, 2022. doi: 10.3390/math10234427.
[14] A. P. Viswanathan, "Music Genre Classification", Int. J. Eng. Comput. Sci., vol. 11, no. 5, pp. 255–259, 2016. doi: 10.18535/ijecs/v4i10.38.
[15] D. Afchar, R. Hennequin, E. V Epure, and M. Moussallam, "Explainability in Music Recommender Systems", AI Mag. 2022, vol. 1, no. 1, pp. 1–23, 2022.
[16] N. Burkart and M. F. Huber, "A Survey on the Explainability of Supervised Machine Learning", J. Artif. Intell. Res. X, vol. 70, pp. 245.347, 2020. doi: 10.48550/arXiv.2011.07876.
[17] N. Karunakaran, "A Scalable Hybrid Classifier for Music Genre Classification using Machine Learning Concepts and Spark", vol. 11, no. 1. IEEE, 2018. doi: 10.1109/ICoIAS.2018.8494161.
[18] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst, "A Benchmark Dataset for Audio Classification and Clustering", International Conference on Music Information Retrieval (ISMIR), 2005. https://www-ai.cs.tu-dortmund.de/audio.html (accessed Apr. 04, 2023).
[19] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. Moh, "The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance", MDPI, vol. 6, no. 2, 2021. doi: https://doi.org/10.3390/data 6020011.
[20] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique", Journal Artif. Intell. Res., vol. 16, pp.321–357,2002,doi: https://doi.org/10.1613/jair.953.
[21] S. Mahanta, "Audio Feature Extraction", Devopedia, 2021. https://devopedia.org/audio-feature-extraction (accessed Mar. 19, 2023).
[22] M. Schedl, E. Gómez, and J. Urbano, "Music Information Retrieval : Recent Developments and Applications", Found. Trends R© Inf., vol. 8, no. 2, pp. 127–261, 2014. doi: 10.1561/1500000042.
[23] S. Tiwari and R. Verma, "Music Genre Classification," New Delhi, 2018. [Online]. Available: file:///C:/Users/toler/Downloads/FinalReport_MT19019.pdf
[24] B. J. Snoek, H. Larochelle, And R. P. Adams, "Practical Bayesian Optimization Of Machine Learning," Adv. Neural Inf. Process. Syst., Pp. 1–12, 2012, Doi: 10.48550/Arxiv.1206.2944.
[25] S. M. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions Scott", Adv. Neural Inf. Process. Syst., pp. 4765–4774, 2017. doi: https://doi.org/10.48550/arXiv.1705.07874.
[26] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, and R. Chatila, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", Information fusion, Vol. 58, pp.82-115, 2020.
[27] K. Choi, M. Sandler, and G. Fazekas, "Explaining Deep Convolutional Neural Networks on Music Classification," ArXiv, vol. abs/1607.0, 2016. doi: https://doi.org/10.48550/arXiv.1607.02444.
[28] Y. Padgaonkar, J. Gole, and B. Tekwani, "Music Genre Classification using Machine Learning", Int. Res. J. Eng. Technol., vol. 09, no. 04, pp. 139–143, 2022.
[29] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve", Radiology, vol. 143, no. 1, pp. 29–36, 1982. doi: 10.1148/radiology.143.1.7063747.