Audio Classification Using Ensembling Techniques

Soumil Jainer, *Student* , Prashant Singh Rana, *Mentor*

***Abstract*—Music genre classification, a vital component of Music Information Retrieval (MIR), continues to captivate researchers due to its relevance in content-based searching and recommendation systems. This paper introduces a comprehensive methodology for classifying music into distinct genres. Leveraging features extracted from audio files and employing various machine learning models, including Naive Bayes, Stochastic Gradient Descent, KNN, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, Neural Networks, and XGBoost, an ensemble model is proposed to enhance classification accuracy. The ensemble model demonstrates superior performance with an accuracy of 90 percent, outperforming individual models. Additionally, blending techniques and feature importance analysis, such as Permutation Importance, are explored to refine the classification process. The paper concludes with a discussion on the significance of the proposed approach and avenues for future research.**

***Index Terms*—Audio Signals, Ensemble Model, Machine Learning Models, Music Genre Classification, Music Information Retrieval, Feature Importance Analysis, Blending Techniques**

I. Introduction

MUsical genres serve as fundamental categories in the organization and retrieval of music collections, with classification based on characteristics such as rhythmic structure, instrumentation, and form. Despite human proficiency in genre identification, automated music genre classification poses significant challenges in the field of Music Information Retrieval (MIR). This study aims to address these challenges by proposing a robust approach to music genre classification using machine learning techniques.

The classification process begins with the extraction of various features from audio files, including waveplots, Fourier transforms, spectrograms, mel spectrograms, zero-crossing rate, harmonics, perceptual features, BPM, spectral centroid, spectral rolloff, mel-frequency cepstral coefficients, and chroma frequency. Exploratory Data Analysis (EDA) is conducted on the GTZAN dataset to understand feature distributions and relationships. Principal Component Analysis (PCA) is applied for dimensionality reduction and visualization.

Subsequently, the dataset is split into training and testing sets, and various machine learning models are trained, including Naive Bayes, Stochastic Gradient Descent, KNN, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, Neural Networks, and XGBoost. Model performance is evaluated using accuracy metrics, with XGBoost emerging as the top-performing model with an accuracy of 90%.

Ensemble techniques, including blending and weighted blending, are explored to further improve classification accuracy. Additionally, feature importance analysis using Permutation Importance is employed to identify the most informative features for classification.

The proposed methodology offers a comprehensive approach to music genre classification, demonstrating the potential for advancements in Music Information Retrieval. Through the integration of feature extraction, machine learning models, ensemble techniques, and feature importance analysis, this study contributes to the ongoing research efforts in automated music genre classification.

II. Literature review

The most significant contribution in the field of genre classification has been given by the creators of the GTZan dataset - Tzanetakis and Cook [1]. Till date, it is considered as the standard for audio genre classification. They used Gaussian Mixture Model (GMM) and achieved a highest accuracy of 61.0%.

Michael, Yang, and Kenny [3] investigated various machine learning algorithms including KNN, K Means, Multiclass SVM and Neural Networks for classification of genres. However, they relied completely on Mel Frequency Cestral Coefficients to characterize genres.

Tao Li et al. [4] used SVM and LDA for content based music genre classification on the GTZan dataset and custom dataset constructed by the author. They achieved the best accuracy of 78.5%.

Bergstra et al. [5] used decision stumps as classifiers on MIREX 2005 dataset achieving an accuracy of 82.34%. Pampalk et al. [6] achieved an accuracy of 82.3% on MIREX 2004 dataset using Neural Net and GMM as classifiers.

Carlos, Alessandro, and Arjun, Kamelia, Ali, and Raymond [9] used deep neural networks for the said classification and inferred that neural networks are comparable to classical models when the data is represented in a rich feature spaceCelso [7] proposed an ensemble approach using a combination of various classical machine learning models on a Latin music dataset. They also included feature selection and conducted various experiments related to feature selection using genetic feature paradigm

Tao Feng [8] used restricted Boltzmann Machine to build Deep Belief Neural Networks to perform a multiclass classification task of labeling music genres and compared it to that of vanilla neural networks.

Arjun, Kamelia, Ali, and Raymond [9] used deep neural networks for the said classification and inferred that neural networks are comparable to classical models when the data is represented in a rich feature space

Miguel [10] used deep learning approach in music genre classification. He used mel spectrograms as input to the convolution neural networks. However, the results were not at par with the ones computed from the conventional methods. Chaturanga [11] used SVM as a base learner in Adaboost

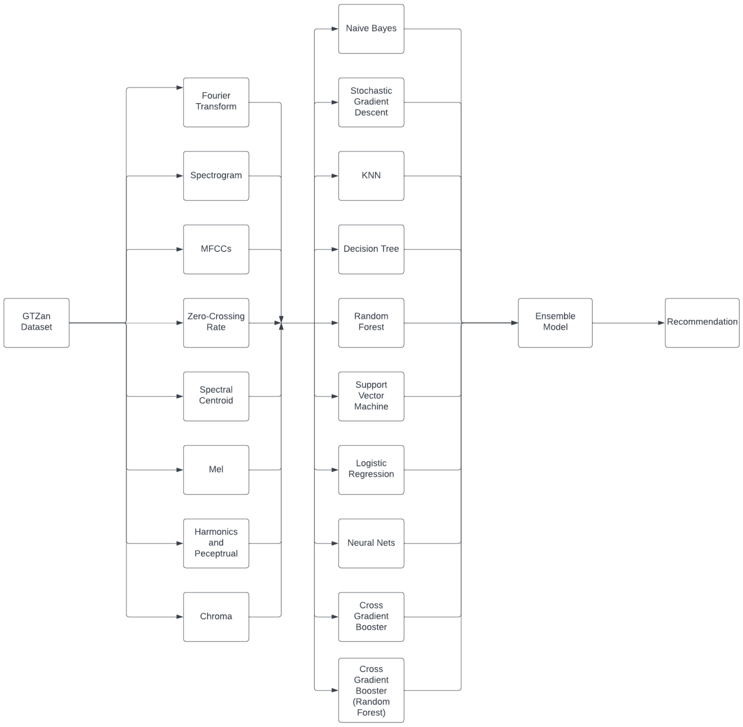


Fig. 1: Flowchart of Methodology

techniques to attain an accuracy of 81% on the GTZAN Audio dataset and 78% on ISMIR2004 Genre dataset. Chun Pui Tang [12] used Long Short Term Memory(LSTM) on GTZAN Audio dataset and obtained an highest accuracy 57.45%.

III. Methodology

A. Dataset

The GTZAN Audio dataset is utilized for this study, consisting of 10 music genres: Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, and Rock. Each genre comprises 100 audio clips, formatted as 22050 Hz Mono 16-bit audio files in .au format. For the purpose of audio genre classification, a subset of five common genres (Rock, Pop, Metal, Country, and Jazz) is selected from the dataset.

B. Feature Extraction and Description

As shown in figure, features are extracted using the open-source library librosa in Python, resulting in a feature vector containing 193 features. The features are grouped into categories as described below:

S.No.	Feature Group	Number Of Features
1	MFCCS	40
2	Chroma	12
3	Mel	128
4	Contrast	7
5	Tonnetz	6
	Total	193

TABLE I: Components Of Feature Vector

1) Mel-Frequency Cepstral Coefficients (MFCCs): Mel-frequency cepstral coefficients (MFCCS) are coefficients that represent short term power spectrum of a sound based on a linear cosine transform of a log power spectrum. This feature group is a large part of the final feature vector. MFCCS is derived as follows:

- The first step involves dividing the audio into several short frames. The aim of the step is to keep the audio signal constant.
- A periodogram estimate of the power spectrum is then calculated for each frame which represent the frequencies present in the short frames.
- Power spectra is then pushed into the mel filter bank and summing the collected energy in each filter.

$$M(f) = 1125 * ln(1 + f/700) \tag{1}$$

- The logarithm of filter bank energies is evaluated.
- The Discrete Cosine Transform is calculated.
- Keep first 40 DCD features.

2) Chroma: Chroma features relate to the 12 different pitch classes and capture harmonic and melodic characteristics of music.

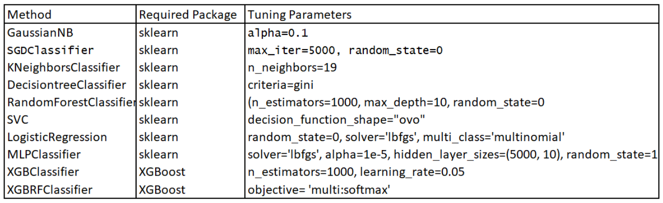
3) Mel : Mel spectrogram is a time frequency representation of a sound. It is sampled into a number of points around equally spaced times t; and frequency fi on a Mel frequency scale.

$$Mel = 2595 * log(1 + f/700) \tag{2}$$

128 features were extracted from each audio file making it an integral part of the final feature vector.

4) Contrast: Contrast features quantify the difference between parts or different instrument sounds, contributing seven features to the feature vector.

5) Tonnetz: Tonnetz represents the tonal centroid features and contributes six features to the feature vector.



Method	Required Package	Tuning Parameters
GaussianNB	sklearn	alpha=0.1
SGDClassifier	sklearn	max_iter=5000, random_state=0
KNeighborsClassifier	sklearn	n_neighbors=19
DecisiontreeClassifier	sklearn	criteria=gini
RandomForestClassifier	sklearn	(n_estimators=1000, max_depth=10, random_state=0
SVC	sklearn	decision_function_shape="ovo"
LogisticRegression	sklearn	random_state=0, solver='lbfgs', multi_class='multinomial'
MLPClassifier	sklearn	solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(5000, 10), random_state=1
XGBClassifier	XGBoost	n_estimators=1000, learning_rate=0.05
XGBRFClassifier	XGBoost	objective= 'multi:softmax'

Fig. 2: Machine Learning Models

C. Machine Learning Methods

Various machine learning models are employed for audio genre classification, including Naive Bayes, Stochastic Gradient Descent, KNN, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, Neural Networks, and XGBoost. Each model is trained and evaluated based on its performance in classifying audio files into their respective genres. Fig. 2 gives the description of the various models used.

IV. Proposed Ensemble Model

Phase I: The first phase includes identifying the five most common genres namely Rock, Pop, Metal, Country and Jazz from the GTZAN dataset. Features are then extracted from the audio files. Training and testing data is generated from the extracted features in the ratio of 70:30 respectively.

Phase II: Various machine learning algorithms are trained on the training set. Hyperparameter tuning of these individual algorithms is done to achieve best results on the test set. Table III gives a description of the various machine learning models along with their tuned hyperparameters.

Phase III: After running rigorous iterations of combinations of the machine learning models using soft and hard voting, the proposed ensemble model was obtained which consisted of an ensemble of five machine learning models (Random Forest, Cross Gradient Booster, SVM, KNN, Logistic Regression) with soft voting.

Phase IV: The ranking of various models is generated using Topsis Analysis. Further K-fold cross validation is done to check the consistency of the model.

V. Model Evaluation

Various parameters such as precision, recall and accuracy are calculated to evaluate the performance of the proposed ensemble model. Repeated K-fold cross validation has been performed to test the robustness of the model.

A. Model Evaluation Parameters

Model evaluation parameters are calculated using the confusion matrix.

1) Precision: Precision is the fraction of relevant instances among the retrieved instances. Precision is computed as:

$$Precision = TP/(TP + FP) \tag{3}$$

2) Recall: Recall is the fraction of relevant instances that have been retrieved over the total number of relevant instances. Recall is computed as:

$$Recall = TP/(TP + FN) \tag{4}$$

3) F1-Score: F-1 Score is the harmonic average of precision and recall. F-1 Score is computed as:

$$F1 - Score = 2 * Precision * Recall/(Precision + Recall) \tag{5}$$

4) Accuracy: Accuracy is the measure of correctness of the classifier. Accuracy is computed as:

$$Accuracy = (TP + TN)/TotalData \tag{6}$$

B. Topsis

Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is a decision analysis method which was developed in 1981 by Hwang and Yoon. Among numerous MCDM/MCDA methods developed to solve real-world decision problems. TOPSIS continues to work satisfactorily across different application areas [15]. It is based on the idea that the solution taken should be the closest to the positive best solution and farthest from the negative best solution. It compares to alternative solutions by assigning weights to the different criteria contained in them, normalizing their scores and then calculating the total score and rank for each alternative.

C. K-Fold Cross Validation

Estimating the accuracy of a classifier induced by supervised learning algorithms is important not only to predict its future prediction accuracy, but also for choosing a classifier from a given set (model selection), or combining classifiers [16]. To ensure that the proposed ensemble model is consistent with low bias and low variance, repeated K-fold Cross Validation is performed. In this present work, 10-fold Cross Validation is repeated for five times. The final average accuracy obtained after 5 runs is 76.35%.

VI. Result Analysis, Comparison and Discussion

Table gives the list of the machine learning models that are trained on the dataset along with their tuned hyperparameters. Feature extraction is done and the dataset is split into two parts - training dataset (comprising of 70 percent of the total dataset) and testing dataset (comprising of 30 percent of the total dataset). The models are trained on the training dataset and are further tested on the testing dataset. The proposed ensemble model is a combination of five models. The models

Sno.	Author	Classifier	Number Of Genres	Best Accuracy (in Percentage)
1.	Tzanetakis et al. [1]	Gaussian Mixture Model	10	61
2.	Michael et al. [3]	Neural Networks	4	96
3.	Tao Feng [8]	Deep Belief Neural Networks	4	63.75
4.	Miguel [10]	Convolution Neural Networks	10	58.73
5.	Chathuranga [11]	SVM in Adaboost	10	81
6.	Chun Pui Tang [12]	LSTM	10	57.45
7.	Present Work	Proposed Ensemble Model	10	82.5

TABLE II: Comparison With Existing Works On GTZAN

are evaluated on various parameters as mentioned in Section V. Topsis Analysis reveals that the proposed ensemble model outperforms many other machine learning models.

A problem which may occur while training is overfitting. To deal with the issues of overfitting, the model should be cross validated and if the resultant accuracy after various runs is consistent in all the runs, then the trained models are not overfitted. Overfitting is when a model models a data well and learns too much. The accuracy is validated by applying 10-fold cross validation five times.

An analysis of the proposed ensemble model with the existing works on GTZan is shown in Table II. The proposed ensemble model achieves higher accuracy than [1], [10], [8] however [3] achieves a higher accuracy that the proposed ensemble model using Neural Networks classifying audio's into four genres.

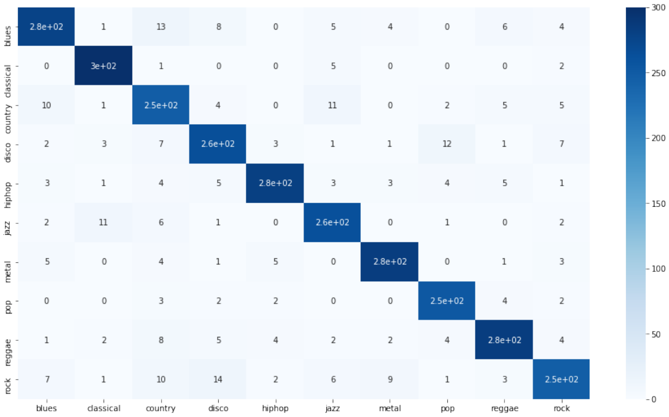


Fig. 3: Confusion Matrix

VII. Conclusion

In today's world, Music Genre Classification finds numerous applications in content based searching and recommendation systems. An ensemble model for Music Genre Classification is proposed which is created by soft voting among Random Forest, Linear SVM, Poly SVM, Logistic Regression and Gradient Boost which achieves an average accuracy of 82.5% percent. In future we intend to study different deep learning architectures like Artificial Neural Networks, Convolution Neural Net- works and stacked autoencoders for audio genre classification on a larger dataset and higher computational power.

VIII. References

[1] George Tzanetakis, Georg Essl, and Perry Cook. Automatic musical genre classification of audio signals. In Proc. of 2nd Annual Interna- tional Symposium on Music Information Retrieval, Indiana University Bloomington, Indiana, USA, 2001.

[2] Thomas G. Dietterich. Ensemble methods in machine learning. In Multiple Classifier Systems, 2000.

[3] Michael Haggblade, Yang Hong, and Kenny Kao. Music genre classifi- cation. Department of Computer Science, Stanford University, 2011.

[4] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content- based music genre classification. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 282-289. ACM, 2003.

[5] Emmanouil Benetos and Constantine Kotropoulos. A tensor-based approach for automatic music genre classification. In Signal Processing Conference, 2008 16th European, pages 1-4. IEEE, 2008.

[6] Elias Pampalk, Arthur Flexer, Gerhard Widmer, et al. Improvements of audio-based music similarity and genre classificaton. In ISMIR, volume 5, pages 634-637. London, UK, 2005.

[7] Carlos N Silla Jr, Alessandro L. Koerich, and Celso AA Kaestner.

A machine learning approach to automatic music genre classification.

Journal of the Brazilian Computer Society, 14(3):7-18, 2008.

[8] Tao Feng. Deep learning for music genre classification. private document, 2014.

[9] Arjun Raj Rajanna, Kamelia Aryafar, Ali Shokoufandeh, and Raymond Ptucha. Deep neural networks: A case study for music genre classifica- tion. In Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on, pages 655-660. IEEE, 2015.

[10] Miguel Flores Ruiz de Eguino. Deep music genre.

[11] YMD Chathuranga and KL Jayaratne, Automatic music genre classifica- tion of audio signals with machine learning approaches. GSTF Journal on Computing (JoC), 3(2), 2018.

[12] Chun Pui Tang, Ka Long Chui, Ying Kin Yu, Zhiliang Zeng, Kin Hong Wong, et al. Music genre classification using a hierarchical long short term memory (lstm) model. 2018.

[13] The Top Tens. music/. https://www.thetoptens.com/most-popular-types-of- I

[14] Librosa. https://librosa.github.io/librosa/.

[15] Majid Behzadian, S Khanmohammadi Otaghsara, Morteza Yazdani, and Joshua Ignatius. A state-of the-art survey of topsis applications. Expert Systems with Applications, 39(17): 13051-13069, 2012.

[16] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In ljcai, volume 14, pages 1137-1145. Montreal, Canada, 1995.