Automated Audio Genre Classification using Ensembled Techniques

Parth Rohilla
ECED
Thapar Institute of Engineering and Technology
Patiala, India
parth05rohilla@gmail.com

Rudra Chandra Gupta
CSED
Thapar Institute of Engineering and Technology
Patiala, India
rudragupta1997@gmail.com

Prashant Singh Rana
CSED
Thapar Institute of Engineering and Technology
Patiala, India
psrana@gmail.com

Harpreet Singh
CSED
Thapar Institute of Engineering and Technology
Patiala, India
akalharpreet@gmail.com

Divya Khanna
CSED
Thapar Institute of Engineering and Technology
Patiala, India
divyakhanna1991@gmail.com

Neeraj Singh
CSED
Thapar Institute of Engineering and Technology
Patiala, India
neeraj.kumar@thapar.edu

***Abstract*—Music Genre classification which comes under the area of Music Information Retrieval (MIR) has been an area of interest among researchers. A music genre is characterized by various features related to instrumentation, rhythmic structure, and form of members. To identify the genre of a given audio file has been a big challenge for the MIR community. This work describes a novel approach for classifying music into different genres. The feature vector for various audio files is obtained and various machine learning models are trained and their performance matrix is computed accordingly. An ensemble model is proposed to improve the performance of the usual classifiers which yields an average accuracy of 85%. Further, K-Fold cross validation has been performed to check the consistency of the proposed ensemble model. The superiority of the proposed ensemble model is validated using topsis analysis and a score of 0.97 is obtained.**

***Index Terms*—Audio Signals, Ensemble model, Machine learning models, Music Genre Classification, Music Information Retrieval, Topsis analysis**

I. Introduction

Musical genres are predefined categories on the basis of an audio's rhythmic structure, instrumentation, and texture. They are used to categorize the music present everywhere. This categorization is used for organizing of music corpus thereby improving the process of content-based searching. Further, this can be used in song recommendation systems. People are good at kind characterization as examined in where it is demonstrated that people can predict the genre of an audio after listening to only 250 milliseconds of the audio [1] but Audio Genre classification may sometimes be subjected to interpretation and it is often the case that two or more genres might be a little similar in their features. This finding proposes that people can judge genre utilizing just the melodic surface without developing any larger amount theoretic portrayals as has been argued in. An improvement in the procedures for programmed genre classification would be an important expansion in the field of Music Information Retrieval.

In the present work, the aim is to propose an ensemble model for music genre classification with the goal to attain better performance than the existing machine learning models. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions [2].

A subset of the GTZAN Audio Dataset is taken and features have been extracted from all the audio files. Based on those features, various machine learning models are trained and hyperparameter tuning has been done. After testing out various combinations, an ensemble model has been proposed. It is validated that the proposed ensemble model performs better than the individual models using topsis analysis. Various techniques like feature selection and cross validation have been applied to achieve maximum accuracy in the prediction of a genre.

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.

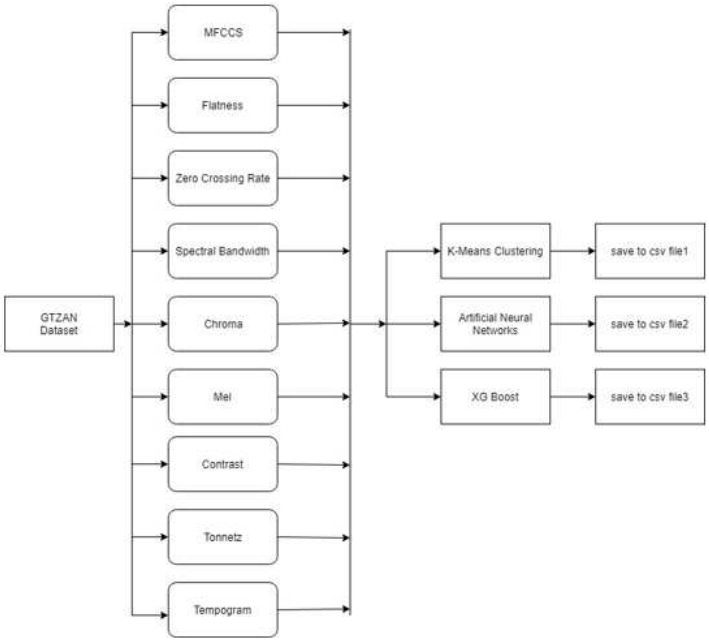

Fig. 1. Flowchart of Methodology

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 Celso [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 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%.

TABLE I
COMPONENTS OF FEATURE VECTOR

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

III. METHODOLOGY

The workflow of the present work is shown in Fig. 1. The extracted feature vector is fed into the various machine learning models. Further, an ensemble model is proposed and the performance is computed accordingly.

A. Dataset

The dataset used for the present work is the GTZAN Audio dataset. The dataset consists of 10 music genres namely Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae and Rock with each genre having 100 audio clips. The audio clips are all 22050 Hz Mono 16 bit audio files in .au format. For the purpose of Audio Genre Classification, five of the more common genres have been taken (Rock, Pop, Metal, Country and Jazz) [13] from the dataset.

B. Feature Extraction and Description

The features have been grouped under categories as described in Table I. The extraction of the features has been done in python using the open source library librosa [14]. The extracted feature vector contains 193 features as shown in Table II. Short descriptions of the features groups are given below.

TABLE II
FEATURE TABLE

AUDIO NO.	F1	F2	F3	F4	F5	—	F189	F190	F191	F192	F193	LABEL
1	-113.571	121.5718	-19.1681	42.36642	-6.36466	—	0.009556	0.010512	-0.02046	0.001493	-0.00643	pop
2	-207.502	123.9913	8.955128	35.87765	2.907321	—	0.018907	0.070679	0.014551	0.009352	-0.00866	pop
.												
.						—						
98	-20.4705	53.68523	5.986029	10.14361	17.07146	—	0.010361	0.024009	-0.03452	0.004169	-0.00781	classical
99	-58.9489	68.86537	-8.46514	3.622923	5.078615	—	-0.01272	0.001894	0.026377	0.004	-0.00349	classical
.												
.						—						
357	-108.521	69.97168	14.8811	45.51458	-5.37834	—	-0.0032	0.007805	0.022346	-0.00182	0.001521	rock
358	-226.288	78.31248	7.799703	53.84219	-1.16246	—	0.003306	-0.00244	0.070924	0.006931	0.000406	rock
.												
.						—						
547	-100.384	104.6881	-57.2479	56.5685	-5.5517	—	-0.00487	0.029969	-0.01257	0.002505	0.003393	metal
548	-93.5559	89.86496	-55.8847	51.63797	-5.57456	—	0.009625	0.040979	0.057395	-0.01102	0.006867	metal
.												
.						—						
695	-111.547	85.55908	3.526411	16.37183	2.21108	—	0.03479	-0.01803	0.044246	0.000682	0.014066	hip-hop
696	-63.5241	79.02744	42.74856	16.09584	15.27049	—	0.027467	0.035715	-0.04266	-0.00558	0.013449	hip-hop

TABLE III
MACHINE LEARNING MODELS

Model	**Method**	**Required Package**	**Tuning Parameters**
Random Forest	RandomForestClassifier	sklearn	Number of Estimators : 100
Linear SVM	SVC	sklearn	max iter = 1000
Poly SVM	SVC	sklearn	Degree : 3
Naive Bayes	GaussianNB	sklearn	alpha=1.0
Gradient Boosting	GradientBoostingClassifier	sklearn	Learning Rate : 0.08
Logistic RRegression	LogisticRegression	sklearn	class weight=None
K Nearest Neighbors	KNeighborsClassifier	sklearn	number of neighbors : 10
LDA	LinearDiscriminantAnalysis	sklearn	alpha=auto
QDA	QuadraticDiscriminantAnalysis	sklearn	tol=0.0001
Decision Tree	DecisionTreeClassifier	sklearn	criterion=gini
ANN	ANN	keras	Number of Layers : 3, Number of Neurons : 20 Hidden Activation : ReLu, Final Layer Activation : Softmax

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 [3]. This feature group is a large part of the final feature vector (40). 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.

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

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

2) Chroma: In music context, the term chroma feature closely relates to the 12 different pitch classes. These are also referred to as pitch class profiles and are a powerful tool for analyzing music whose pitch can be meaningfully characterized. The important property of chroma features is that they capture harmonic and melodic characteristics of music.

3) Mel: Mel spectrogram is a time frequency representation of a sound [3]. It is sampled into a number of points around equally spaced times t_i and frequency f_i 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 is the difference between parts or different instrument sounds. Seven contrast features were extracted from each audio to make up the files feature vector.

5) Tonnetz: It is a representation of the tonal centroid features. Six features were extracted from each audio to make up the file feature vector.

C. Machine Learning Methods

Table III gives the description of the various machine learning models used for audio genre classification.

IV. PROPOSED ENSEMBLE MODEL

The workflow of the present work can be broken down into four phases which is represented in Fig. 2.

Phase I : The first phase includes identifying the five

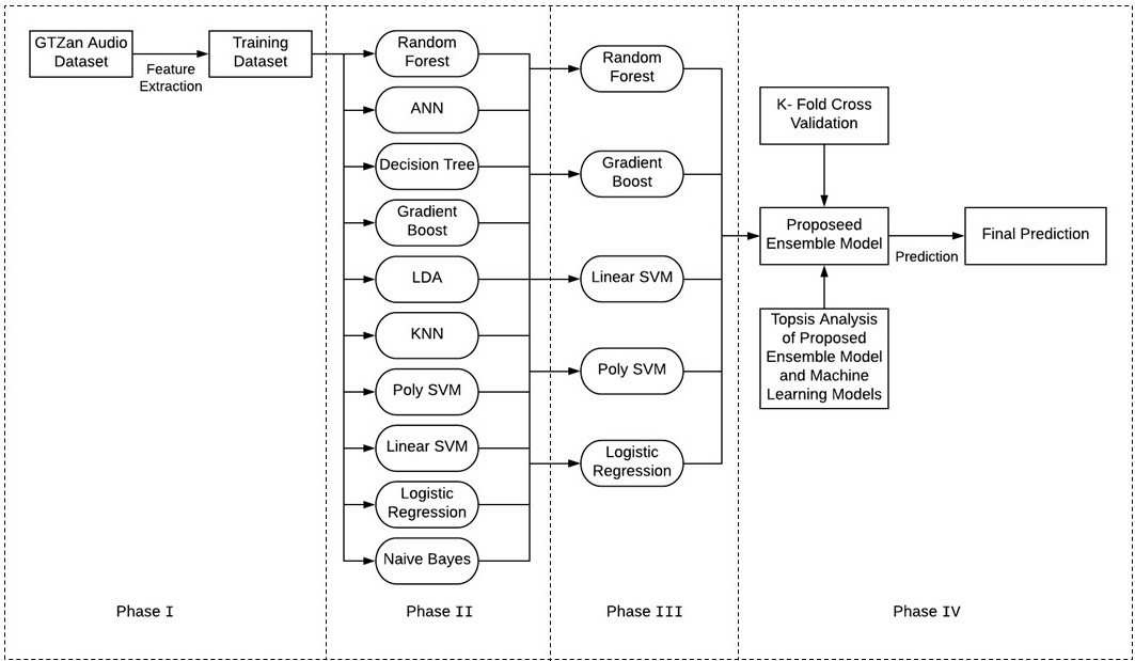

Fig. 2. Workflow of the Proposed Ensemble Model

most common genres namely Rock, Pop, Metal, Country and Jazz [13] from the GTZAN dataset. Features are then extracted from the audio files which has been discussed in Section III-B. Training and testing data is generated from the extracted features in the ratio of 80:20 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 various combinations of the machine learning models using soft and hard voting, the proposed ensemble model was obtained (based on model evaluation parameters discussed in Section V) which consisted of an ensemble of five machine learning models (Random Forest, Gradient Boost, Linear SVM, Poly SVM, Logistic Regression) with soft voting.

Phase IV : The ranking of various models is generated using Topsis Analysis (discussed in Section V-B). Further K-fold cross validation is done to check the consistency of the model (discussed in Section V-C).

V. Model Evaluation

Various parameters such as precision, recall and accuracy are calculated to evaluate the performance of the proposed ensemble model. The results are compiled in the form of table and have been shown below in Table IV. Repeated K-fold cross validation has been performed to test the robustness of the model.

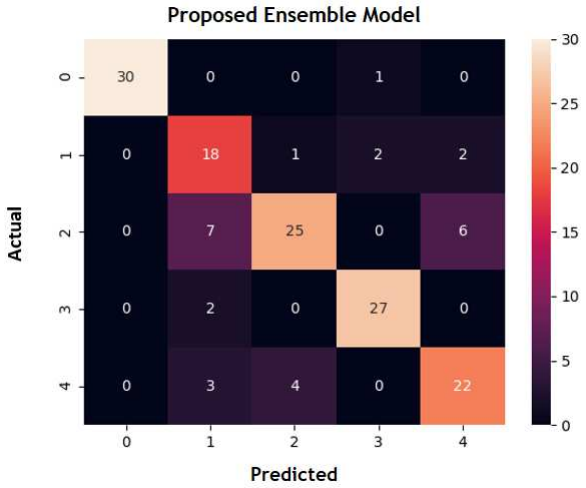

Fig. 3. Confusion Matrix of Proposed Ensemble Model

A. Model Evaluation Parameters

Model evaluation parameters are calculated using the confusion matrix. The confusion matrix for the proposed ensemble model is given in Fig. 3.

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.

TABLE IV
EVALUATION PARAMETERS OF MACHINE LEARNING MODELS

Sno.	Algorithm	Precision	Recall	F1 Score	Accuracy
1	Random Forest	0.81	0.77	0.78	0.76
2	Linear SVM	0.76	0.76	0.76	0.76
3	Poly SVM	0.74	0.75	0.74	0.74
4	Naive Bayes(G)	0.71	0.65	0.65	0.64
5	Gradient Boost	0.79	0.76	0.77	0.76
6	Logistic Regression	0.74	0.73	0.73	0.72
7	KNN	0.68	0.65	0.66	0.65
8	LDA	0.68	0.67	0.66	0.66
9	ANN	0.74	0.71	0.72	0.71
10	Decision Tree	0.70	0.66	0.67	0.66
11	Proposed Ensemble Model	0.84	0.82	082	0.82

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:

$$F-1Score = 2 * Precision * Recall/(Precision + Recall) \tag{5}$$

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. Table V gives the result of topsis analysis done on the machine learning models and the proposed ensemble model.

TABLE V
TOPSIS

Sno.	Algorithm	Rank	Score
1	Ensemble Model	1	0.97
2	Random Forest	2	0.87
3	Gradient Boost	3	0.85
4	Linear SVM	4	0.82
5	Poly SVM	5	0.78
6	Logistic Regression	6	0.75
7	ANN	7	0.72
8	Decision Tree	8	0.61
9	LDA	9	0.6
10	KNN	10	0.58
11	Naive Bayes(G)	11	0.58

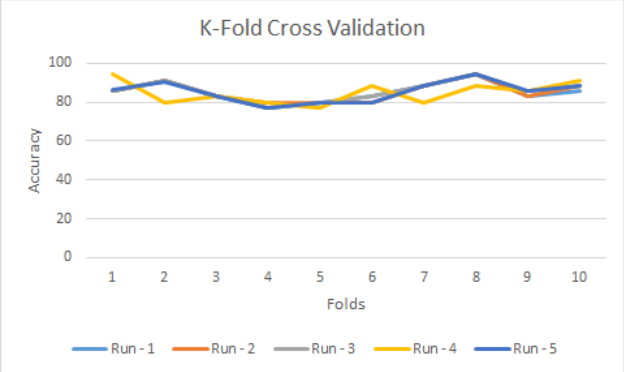

Fig. 4. Graph for K-Fold Cross Validation

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 results obtained are plotted against accuracy as shown in Fig. 4 and the lines are overlapping which signifies the proposed ensemble model is robust. The final average accuracy obtained after 5 runs is 85.25%.

VI. RESULT ANALYSIS, COMPARISON AND DISCUSSION

Table III 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 80% of the total dataset) and testing dataset (comprising of 20% 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 are evaluated on various parameters as mentioned in Section V. Topsis Analysis reveals that the proposed ensemble model outperforms other machine learning models.

TABLE VI
COMPARISON WITH EXISTING WORKS ON GTZAN

Sno.	Author	Classifier	Number of Genres	Best Accuracy
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**	**5**	**82.5%**

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. Fig. 4 shows the results after applying K-fold cross validation.

An analysis of the proposed ensemble model with the existing works on GTZan is shown in Table VI. 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.

VII. CONCLUSION AND FUTURE WORK

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 85.25%.

In future we intend to study different deep learning architectures like Artificial Neural Networks, Convolution Neural Networks and stacked autoencoders for audio genre classification on a larger dataset and higher computational power.

REFERENCES

[1] George Tzanetakis, Georg Essl, and Perry Cook. Automatic musical genre classification of audio signals. In *Proc. of 2nd Annual International 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 classification. *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 classification. 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 classification 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. https://www.thetoptens.com/most-popular-types-of-music/.

[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 *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.