Music Genre Classification using Convolutional Neural Networks and Machine Learning Classifiers

Islam Anwar
Faculty of Science
UPEI Cairo Campus
Cairo, Egypt
Islam.Anwar@uofcanada.edu.eg

Omar Eltouny
Faculty of Science
UPEI Cairo Campus
Cairo, Egypt
Omar.Eltouni@uofcanada.edu.eg

Ahmed Elsheikh
Dean of Mathematics and Computational Sciences
UPEI Cairo Campus
Cairo, Egypt
Ahmed.Elsheikh@uofcanada.edu.eg

Eslam Amer
Faculty of Science
Misr International University
Cairo, Egypt

***Abstract*—Music has become a vital aspect of our lives as time and technology have moved forward. With the emergence of growing audio streaming services, building effective music recommendation systems and music genre classification has been a difficult challenge. With the recent growth in the popularity of machine learning, we propose a deep learning neural network model for music genre classification in this paper. Spectrograms are used to extract features such as Mel-Frequency Cepstral Coefficients and others. Various algorithms are deployed including ensemble random forest and support vector machines for comparison with the proposed model. We experimentally illustrated that our proposed neural network provided better accuracy than our supervised classifiers. The deep learning model accuracy is at 85.5% in our evaluation. According to our experiments with real sound data and other relatable work, our model is effective and reliable.**

***Index Terms*—Music genre classification, Deep learning, Random forest, Mel-frequency cepstral coefficients, Support-vector machines**

I. Introduction

The recent rise of music streaming services has been accompanied by the accelerating emergence of machine learning and artificial intelligence [1]. Today, applications in the multimedia industry (Spotify, Apple Music, etc.) are continuously striving to enhance their music recommendation systems [2]. Followed by unique algorithm designs to classify music to attract customers as well as maintain existing customers [3].

Music Information Retrieval (MIR) is an advanced field that primarily focuses on retrieving relevant information from music. The growth of machine learning and artificial intelligence has allowed recent traditional methods to be replaced by more accurate models [4]. During the recent research performed on MIR, of the 104 papers, approximately half displayed quantitative results, and over 80% of those papers reported mean values of performance measures [5]. This shows that very few researchers are utilizing other evaluation techniques to gauge the variability of results and model performances. One of the main sub-problems of MIR is music genre classification [6]. Historically, genres of music were easily discernible according to their rhythm, tempo, the richness of sound, and other factors including harmony and acoustics.

Although there has been much research performed on music genre classification in recent years, few have been able to obtain consistent and positive results. However, the continuous advancement of machine learning models, ranging from improved hardware capabilities to the availability of well-structured data. These benefits have drastically eased the difficulty of proposing new and improved models for music genre classification [4].

Music genre classification is based on analyzing sound. Sound can be represented as audio waves consisting of several parameters including pitch, frequency, amplitude, etc. Music Analysis is performed based on audio signatures that showcase certain characteristics such as acoustics, harmony, and energy. Music is categorized by genre, which is a collection of music that follows similar audio signatures and rhythmic patterns [7].

Although music genre classification has been researched in detail over recent years, insufficient attention has been given to the fine-tuning of hyperparameters used in supervised classifiers. This aspect may be the difference between a highly accurate model and a poor model. Adjustment of hyperparameters is vitally important to optimize our classifiers to accurately predict music genres [1] [8]. Although there may be complications with choosing optimal hyperparameters, especially with deep learning models. This is because we are classifying music into genres with unknown hypotheses functions. Thus, increasing the cost of finding an optimal model and its optimal hyperparameters [6].

This paper aims to perform music genre classification by building various machine learning and deep learning models to classify music using only their audio signals [9]. Also, to examine the performances of these models alongside current models. Important factors such as combinations of handcrafted features and hyperparameter optimization will be discussed in further detail later.

The rest of the paper is organized as follows. Section 2

describes recent existing work regarding the task of music genre classification. Section 3 provides an overview of the proposed methodology of choice, diving into feature extraction techniques and other preprocessing steps. Results and discussions are provided in Section 4 including suggestions for future research work. Section 5 provides a conclusion to the research.

II. Related Work

There has been extensive recent research performed in music genre classification. A. Elbir et al. [3] used Short Time Fourier Transform (STFT) based music genre classification. One of the most effective time-frequency analysis methods used to obtain sinusoidal frequencies and changes in signal over some time. By generating spectrograms of the music samples using differing window types, window sizes, and overlap ratios for feature extraction. Traditional Machine Learning (ML) techniques such as Support-Vector Machines (SVM) with different kernels and Random Forest (RF) are employed to gauge the performance of these varying features [3]. The most impressive results were obtained with polynomial SVM with an accuracy of 71.3% using a window size of 512, an overlap ratio of 256 as well as parzen window type [3].

J. Ramirez et al. [4] further explored this field by proposing a bundle of supervised and Deep Learning (DL) algorithms including Decision Trees (DT), Naive Bayes (NB), SVM, Fully-Connected Neural Networks (FCNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) applied to the audio set dataset. This dataset provides pre-trained features extracted using a Convolutional Neural Network (CNN) model that maps each sample to a corresponding value, known as audio embedding, thus allowing data to be fed directly into models for evaluation [4]. Results indicate that Neural Networks (NN) showed a significant gap in performance in comparison to traditional ML classifiers ranging from 50-80% accuracy for classifying different music genres [4].

S. Chillara et al. [7] contributed further with the use of DL algorithms, CNN was employed to convert the sound waves to spectrograms to predict the labels. Furthermore, content-based algorithms like Logistic Regression (LR) and Artificial Neural Networks (ANN) were used with hand-made features. Various network architectures were used and compared including Feed-Forward Neural Networks (FFNN) as well as Convolutional Recurrent Neural Networks (CRNN) while maintaining vital layers such as pooling layers and spectrogram generation using strict parameter values. Feature extraction was performed by retrieving a collection of time and frequency domain features such as central moments, tempo, and Mel-Frequency Cepstral Coefficients (MFCC) [7]. CNN outperformed all other spectrogram-based models, as well as feature-based models obtaining an accuracy of 88.5% [7].

A. Elbir et al. [10] performed his research on the GTZAN dataset, suggesting a different feature extraction methodology focusing on extracting time and frequency domain features such as spectral centroid, spectral contrast, spectral bandwidth, etc. In addition, retrieving statistical descriptors like mean, median, and standard deviation for each of the extracted features. An array of ML models are proposed including NB, RF, and CNN [10]. Results showed that SVM had the strongest performance among all models with an accuracy of approximately 72.5% with no significant improvement in results upon changing the window type or size [10].

H. Bahuleyan et al. [11] also conducted two classes of models: DL models that rely on the spectrogram, and hand-made features within time and frequency domain features. This was performed on the audio set dataset, utilizing a CNN model to label the digital signals based solely on the obtained spectrograms. On the other hand, hand-made features such as spectral bandwidth, chroma features, and MFCC were fed into an array of ML models including XGBoost (XGB), RF, and LR [11]. Among the best results were for the VGG-16 model and the ensemble method of VGG-16 CNN and XGB that used spectrograms to predict the music genre. It obtained better performance across three evaluation metrics including an accuracy of 65% [11]. On the other hand, feature-engineered models such as LR lagged due to their linear nature.

Table 1 illustrates the comparison between various authors and their techniques towards music genre classification.

TABLE I: Previous Work Results

Work	Dataset	Approach	Accuracy
A. Elbir et al. [3]	GTZAN	Polynomial SVM	0.713
J. Ramirez et al. [4]	Audio Set	Neural Networks	-
S. Chillara et al. [7]	FMA	CNN	0.885
A. Elbir et al. [10]	GTZAN	SVM	0.725
H. Bahuleyan et al. [11]	Audio Set	VGG-16 CNN	0.650

All previous research has failed to consider the effect of adjusting model hyperparameters to obtain optimized solutions. They were primarily focused on evaluating which algorithms obtained the best accuracies from DL or supervised ML classifiers. This is regardless of whether models are deep learning models or feature-based models. Moreover, another area that was not considered is the amount of data. Previous works' used the original datasets, which consist of low amounts of data, especially with DL models. However, the aspect of data augmentation is not discussed or implemented. This may have had a significant effect on the performance of DL models and could entail skewness of results. This paper aims to focus on optimizing hyperparameters to use our proposed model and classifiers to their maximum potential.

III. Proposed Work

The primary objective is to perform music genre classification using DL techniques and traditional ML classifiers. This is performed by first obtaining our data set in its raw format as shown in Fig. 1 of our proposed model. Regarding feature extraction, various features such as MFCC, means, and variances of several features including spectral centroid, spectral bandwidth, roll-off, and zero crossing rate will be extracted.

Upon extraction, data preparation techniques are employed such as treating missing values and detecting and treatment of possible outliers in our data. Furthermore, we will encode our genre labels to be numerical rather than categorical. Data normalization and data splitting will be performed to prepare the audio data to be fed to our chosen classifiers and DL models. Traditional ML classifiers such as SVM and RF. Our training phase will use default hyperparameters. We will use a validation set to optimize our models with a split of 60-20-20 for the training, validation, and test sets respectively.

Regarding model selection, we opted for CNN as our DL model because of its effectiveness in music genre classification. Furthermore, we chose SVM and RF among some of our supervised ML classifiers due to their flexibility in dealing with a variety of data structures to yield potentially high accuracies while optimizing hyperparameters. There is high variability like their hyperparameters, which is why they are promising candidates. Table 2 illustrates the most important hyperparameters chosen for our CNN model. This is important to mention because how there is a multitude of hyperparameters to adjust. Therefore, it is important to maintain a sensible scope of a few important hyperparameters.

We constructed our DL model and supervised ML classifiers then performed training as well as hyperparameter optimization to maximize results. Finally, deploying our models and classifiers on the test set and compare results.

Hyperparameters were set based on trial and error. The activation function was set to relu to serve as an enhanced logistic classifier function taking into account negative axes. Furthermore, a dropout rate was deployed to regularize an overfit model on the validation set. Concerning our supervised classifiers, their respective hyperparameters were set based on applying grid search to obtain the highest possible accuracy score in the validation set.

TABLE II: Proposed Hyperparameters

Hyperparameter	**Value**
Optimizer	Adam
Batch Size	128
Number Epochs	50
Activation Function	Relu
Dropout Rate	0.3
Loss Measure	Binary Crossentropy

IV. Results and Discussion

A. Dataset

The dataset of choice to conduct our study was the GTZAN dataset. This dataset is a wide collection of 1000 observations, which consist of extracts of a musical piece of thirty seconds duration [9]. The dataset is balanced between 10 different genres with 100 observations per genre.

For our DL model, we used the raw audio files to extract our MFCCs and other features to utilize the power of our convolutional network. With respect to our supervised classifiers, we used the attached .csv formatted dataset that already includes various features such as means and variances of spectral centroid, spectral bandwidth, and a collection of MFCCs per sample.

B. Evaluation Metrics

Regarding our evaluation metrics, we chose a collection of classification metrics to assess our proposed model alongside our supervised classifiers: RF and SVM. These metrics are the F1-score, precision, recall, accuracy, and mean squared error (Equations 1-5).

$$F1 - \text{Score} = \frac{2 \times \text{ Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{1}$$

$$Precision = \frac{\text{TruePositives } (TP)}{TP + \text{ FalsePositives } (FP)} \tag{2}$$

$$Recall = \frac{TP}{TP + \text{ FalseNegatives } (FN)} \tag{3}$$

$$Accuracy = \frac{TP + \text{ TrueNegatives } (TN)}{TP + TN + FP + FN} \tag{4}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{5}$$

Concerning some of our supervised classifiers, we simulated the training phase and optimized hyperparameters on the validation set to obtain optimal results. Concerning our RF classifier, we fixed the depth to 45 and the criterion to gini. The number of estimators will be optimized and compared later with our other algorithms. As for SVM, we set the C-regularizer to 0.1 and the gamma type to auto. Furthermore, we will run differing kernel types for comparison.

With our dataset consisting of only 1000 observations, results of our different classifiers and proposed model may generate high skewness of results. This is due to the low amounts of data provided for our proposed model.

In terms of RF, its results varied amongst the different number of estimators across all 5 evaluation metrics. The f1-score was maximized at 0.71 for 200 estimators while the precision score was maximized for 500 estimators. Moreover, recall and accuracy were best at 0.73 and 0.7 respectively for 200 estimators. Overall, it would be appropriate to conclude that RF performed best at 200 estimators across most of the aforementioned metrics.

Concerning SVM, it was easy to determine the best kernel type. Linear SVM performed best across all evaluation metrics, which was unexpected. This may be due to the one-dimensional nature of sound data as opposed to a multi-dimensional nature, which would have made polynomial SVM more effective than the results indicate. XGB was also optimized alongside SVM and RF as an ensemble method with hard voting to compare with out proposed model.

Table III illustrates the accuracy metrics comparison between RF, SVM, and other classifiers such as XGB, MLP,

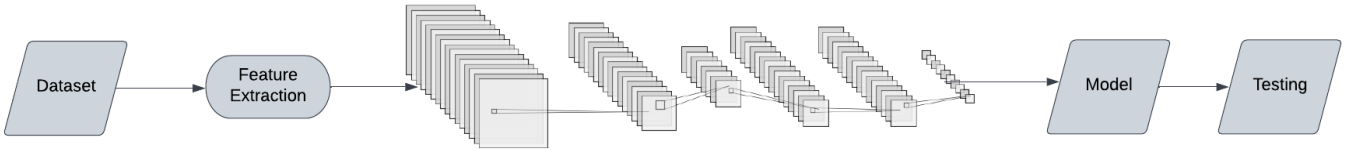

Fig. 1: Model Overview

TABLE III: Comparison of proposed algorithms

Classification Algorithm	F1-Score	Precision	Recall	Accuracy	MSE
RF (100 Estimators)	0.70	0.70	0.69	0.695	5.385
RF (150 Estimators)	0.68	0.69	0.69	0.685	5.130
RF (200 Estimators)	0.71	0.70	0.73	0.70	5.690
RF (250 Estimators)	0.68	0.69	0.69	0.685	5.120
RF (500 Estimators)	0.70	0.71	0.70	0.700	4.880
RF (750 Estimators)	0.67	0.68	0.68	0.675	4.790
RF (1000 Estimators)	0.68	0.69	0.69	0.685	4.815
SVM (Linear)	0.73	0.73	0.72	0.725	6.15
SVM (Poly)	0.31	0.62	0.31	0.310	21.94
SVM (RBF)	0.44	0.62	0.48	0.485	11.125
XGB	0.70	0.71	0.69	0.69	6.99
XGB + RF (200) + SVM (Linear)	0.71	0.72	0.71	0.715	6.425
Proposed Model	0.85	0.86	0.86	0.855	3.01
Best Algorithm	**0.85**	**0.86**	**0.86**	**0.855**	**3.01**

and others with varying hyperparameters. Furthermore, they are compared with our proposed model. Our proposed neural network outperformed all our supervised classifiers at any proposed hyperparameter with an accuracy of 0.855 and a minimization of the MSE at 3.01. In addition to our metrics, Fig. (2) supports the conclusion that our proposed neural network was the best algorithm since it was overfit by a small margin.

TABLE IV: Comparison with previous work

Research	Dataset	Algorithm	Accuracy
A. Elbir et al. [3]	GTZAN	Polynomial SVM	0.713
J. Ramirez et al. [4]	Audio Set	Neural Networks	-
S. Chillara et al. [7]	FMA	CNN	0.885
A. Elbir et al. [10]	GTZAN	SVM	0.725
H. Bahuleyan et al. [11]	Audio Set	VGG-16 CNN	0.650
Proposed Work	**GTZAN**	**CNN**	**0.855**

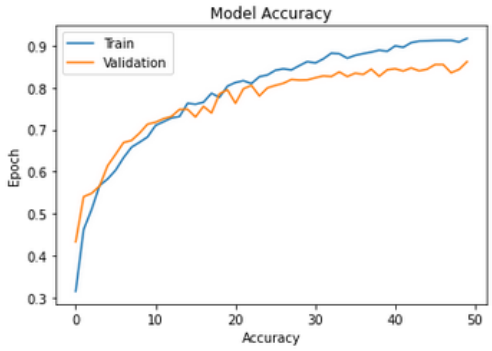

Fig. 2: Proposed model accuracy

V. Conclusion

Until the recent emergence of ML and DL was applied to MIR tasks, building effective algorithms for music genre classification has been at the forefront of research for a multitude of applications including audio streaming services, signal analysis, and more. However, accurate and effective models were few and far between.

In this paper, we examined the effects of optimizing hyperparameters on various supervised ML classifiers as well as a proposed DL model. In particular, to choose a collection of hyperparameters that would yield significant improvement across various evaluation metrics. We found that some of our supervised classifiers such as random forest after adjusting hyperparameters did not majorly impact results. However, other classifiers such as SVM was strongly affected by adjusting kernel type as it pertains to the dimensionality nature of the data. We have proposed an effective DL model that examines audio spectrograms, extracts unique MFCCs, and classifies each audio file according to genre type. Experiments have shown that our proposed model yields an accuracy of 85.5%. Moving forward, it would be advisable for future work to attend to the low amount of data by performing data augmentation techniques to further enhance the reliability of our proposed network.

TABLE V: Comparing best model with previous work

Work	**Dataset**	**Approach**	**Accuracy**
H. Bahuleyan et al. [11]	Audio Set	VGG-16 CNN	0.650
S. Chillara et al. [7]	FMA	CNN	0.885
Proposed Model	**GTZAN**	**CNN**	**0.855**

REFERENCES

[1] M. Chaudhury, A. Karami, and M. A. Ghazanfar, "Large-scale music genre analysis and classification using machine learning with apache spark," *Electronics*, vol. 11, no. 16, p. 2567, 2022.

[2] D. S. Lau and R. Ajoodha, "Music genre classification: A comparative study between deep learning and traditional machine learning approaches," in *Proceedings of Sixth International Congress on Information and Communication Technology*, pp. 239–247, Springer, 2022.

[3] A. Elbir, H. O. İlhan, G. Serbes, and N. Aydın, "Short time fourier transform based music genre classification," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pp. 1–4, IEEE, 2018.

[4] 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.

[5] A. Flexer, "Statistical evaluation of music information retrieval experiments," *Journal of New Music Research*, vol. 35, no. 2, pp. 113–120, 2006.

[6] Y. Xu and W. Zhou, "A deep music genres classification model based on cnn with squeeze & excitation block," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 332–338, IEEE, 2020.

[7] S. Chillara, A. Kavitha, S. A. Neginhal, S. Haldia, and K. Vidyullatha, "Music genre classification using machine learning algorithms: a comparison," *Int Res J Eng Technol*, vol. 6, no. 5, pp. 851–858, 2019.

[8] R. Singhal, S. Srivatsan, and P. Panda, "Classification of music genres using feature selection and hyperparameter tuning," *Journal of Artificial Intelligence and Capsule Networks*, vol. 4, no. 3, pp. 167–178, 2022.

[9] N. Ndou, R. Ajoodha, and A. Jadhav, "Music genre classification: A review of deep-learning and traditional machine-learning approaches," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–6, IEEE, 2021.

[10] A. Elbir, H. B. Çam, M. E. Iyican, B. Öztürk, and N. Aydin, "Music genre classification and recommendation by using machine learning techniques," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–5, IEEE, 2018.

[11] H. Bahuleyan, "Music genre classification using machine learning techniques," *arXiv preprint arXiv:1804.01149*, 2018.