Texture Feature and Mel-Spectrogram Analysis for Music Sound Classification

M. E. ElAlami[1], S. M. K. Tobar[2], S. M. Khater[3], Eman. A. Esmaeil[4]
Computer Science Department-Faculty of Specific Education, Mansoura University, Mansoura, Egypt[1, 3, 4]
Musical Education Department-Faculty of Specific Education, Mansoura University, Mansoura, Egypt[2]

***Abstract*—The categorization of music has received substantial interest in the management of large-scale databases. However, the sound of music classification (MC) is poorly interesting, making it a big challenge. For this reason, this paper has proposed a new robust combining method based on texture feature with Mel-spectrogram to classify Arabic music sound. A music audio dataset consisting of 404 sound recordings for different four classes of Arabic music sounds has been collected. The collected data became available for free on the Kaggle website. Firstly, music sound is transformed into a Mel spectrogram, and then several texture features are extracted from these Mel spectrogram images. A two-dimensional Haar wavelet is applied to each Mel-spectrogram image, and Local Binary Patterns (LBP), Gray Level Co-occurrence Matrix (GLCM), and Histogram of Oriented Gradient (HOG) are utilized for feature extraction. K-nearest neighbors (KNN), random forest (RF), decision tree (DT), logistic regression (LR), AdaBoost, extreme gradient boosting (XGB), and support vector machine (SVM) classifiers were utilized in a comparative analysis of Machine Learning (ML) algorithms. Two different datasets have been employed in order to evaluate the effectiveness of our approach: the collected dataset that the authors had gathered and the global GTZAN dataset. Our method demonstrates superior performance with a five-fold cross-validation. The experimental findings indicated that the XGB exhibited a high accuracy with an average performance of 97.80% for accuracy, 97.72% for F1-Score, 97.75% for recall, and 97.81% for precision.**

***Keywords*—Mel-spectrograms; ML; texture features; MC**

I. INTRODUCTION

Music is considered an inseparable part of our culture and tradition [1]. The advent of online social networks and cloud technology have led to a massive rise in the demand for online data storage and data sharing services [2, 3]. Music Information Retrieval (MIR) systems have gained huge popularity in recent years and are used in many fields, such as musical similarity and genre categorization, music emotion identification, music source separation, acoustic descriptions of music, and music transcription [4]. Music classification (MC) has emerged as an important area for digital music services such as Tidal, SoundCloud, and Apple Music and is used for classifying and overseeing extensive musical datasets [5, 6]. Regarding this, musical sound classification is an intriguing area challenge in the field [7]. Among the several methods for representing the contents of an audio clip, extracting distinguishing features is the most widely employed. However, due to the subjectivity associated with the concept of musical genre, as well as the enormous variety of music genres, strong feature extraction has proven difficult. In the Arab world, Arabic music is an essential component of global music, but Western music dominates the field. A machine learning approach is extensively used in music information retrieval applications [7, 8, 9, 10]. As well, texture features have a high capacity for extracting features of musical patterns [11, 12].

In related works, Western music using the GTZAN dataset dominates the field, and Arabic music is not yet equivalent to them. So, Arabic musical instruments must be moved out of the country and promoted for it in the works. Therefore, we hope that this work will contribute to solving this problem and overcoming the absence of a dataset based on Arabic music. Therefore, this paper presents a new robust approach based on ML techniques with texture features and Mel-spectrogram for Arabic music sound classification using a newly collected dataset in favour of this work and became available free for use.

Our contribution can be summarized as follows:

- A music audio dataset consisting of 440 sound recordings for different four classes of Arabic music sounds has been collected. The collected data became available for free at: (https://www.kaggle.com/datasets/emanatyaesmaeil/zekrayati-dataset).
- Gathered and annotated a large corpus of Arabic music clips to cover the lack of a dataset for Arabic music. Although the GTZAN dataset is a benchmark for MGC, it has limitations such as mislabeling, distortions, and replicas (Strum, [13]).
- A new robust feature extraction approach is presented for music signal classification using Mel-spectrogram images. A two-dimensional Haar wavelet is applied to each image and texture features (GLCM, HOG, and LBP) are extracted from all wavelet transform sub-bands.
- Comparative analysis to examine the efficiency of most various machine learning algorithms in Arabic music sound classification and then determine which algorithm is better for this type of data.
- The best accuracy had resulted compared to previous studies using global GTZAN dataset.

The general structure of this paper is as follows: Section II presents the related studies, Section III covers the materials and methodologies, and Section IV provides the results and

*Corresponding Author.

discussion. Discussion is given in Section V. Finally, Section VI covers conclusion and future work.

II. Related Work

Recently, there have been a lot of studies related to music classification. [14]. These works have been supported by recent advancements in machine learning (ML) and deep learning (DL) methodologies. This section provides a thorough overview of many ML and DL applications in the music industry and examination of the prospects for AI in this domain as show in Table I.

In summary, western music dominates the field. Most of the literature focuses on it by using the GTZAN dataset. Arabic music needs to move out of the country, so it is hoped that this paper will make a small contribution to this goal and benefit from the ML approach that has proven its effectiveness in music classification.

TABLE I. Previous Studies Related to ML And DL Approach

REF.	Year	Task	Algorithm	Dataset	Accuracy
[15]	**2022**	Dissecting the Nigerian music genre.	Timbral texture feature using SVM, XGB, RF, and K-NN	ORIN dataset	XGB=0.82, SVM= 0.74, RF=0.71 and K-NN=0.51
[16]	**2024**	Classification of Musical Genres	14 audio features in total when using XGB	GTZAN Dataset	Accuracy=81%
[17]	**2023**	Classification of musical genres	CNN-based mel-frequency cepstral coefficients (MFCC)	GTZAN	Accuracy=85%
[18]	**2024**	Classification of Music Categories	MFCC + STFT + CNN	GTZAN and Extended-Ballroom datasets	Accuracy of dataset1=95.71 Accuracy of dataset2=95.20
[19]	**2023**	categorization of music genres	hybrid model for wavelet + spectrogram analysis	Ballroom and GTZAN datasets	Accuracy of dataset1=81% Accuracy of dataset2=71%
[20]	**2021**	categorization of music genres	MEL-Spectrogram based on logs and Transfer Learning	GTZAN dataset	The best accuracy is Resnet34=97%
[21]	**2023**	Automated Genre Classification of Music	MFCC and CNN	GTZAN dataset	Accuracy=83%
[22]	**2022**	Identification of musical genres	CNN with Mel-spectrograms: The Best Feature	GTZAN dataset	Accuracy=91%
[23]	**2023**	Suggested Music Track	DCNN and Mel-spectrograms	JUNO, GTZAN, and FMA-Small datasets	Dataset1 Accuracy=63%, Dataset2 Accuracy=78% and Dataset3 Accuracy=89%
[24]	**2023**	Categorization of Music Genres	Ideal model with CRNN and Mel-spectrograms	FMA-Small dataset	Accuracy=90%
[25]	**2024**	Indian Category of Musical Instruments	K-NN, RF, RNN, XGB, LR, DT, and SVM in an MFCC	Gathered 1177 audio samples in total with six classes from different online sources.	RNN has best accuracy=0.9872
[26]	**2020**	Categorization of Music Genres	feature extraction from metadata using SVM, K-NN, and NB	Spotify music dataset	SVM=80%, K-NN=77.18% and NB=76.08%
[27]	**2020**	Identification of Music Genres	CNN and the Mel Spectrum	GTZAN dataset	Accuracy=84%
[28]	**2021**	categorization of music	Spectrograms with many DNN models; the best model is ResNet50.	The datasets FMA, GTZAN_4, and EMA	Dataset1 Accuracy=80.14%, Dataset2 Accuracy=81.09% and Dataset3 Accuracy =77.03%
[29]	**2022**	Identification of Music Genres	CNN, LSTM, and MLP in an MFCC	GTZAN dataset	CNN = 70.42%; LSTM = 61.50%; and MLP = 63.28%
[30]	**2020**	Identification of Music Genres	combined (FC1, FC2, FC3, FC4) with SVM	Spotify Music Dataset	Accuracy of FC1 and FC2=80%
[31]	**2022**	Bangla music's classification by genre	Feature Scaling Method plus PCA utilizing NN, RF, K-NN, and SVM	Bangla Music Dataset	SVM-RBF=68.77%, K-NN=61.32%, RF=69.05% and NN=77.68%

III. MATERIALS AND METHODS

The flow diagram of the suggested model is illustrated in Fig. 1. The following three subsections will describe each of these stages in detail.

A. Mel-Spectrogram Production Stage

The Mel-spectrogram is resulted through the following steps:

1) Pre-emphasizing audio which improves clarity and reduces volume.

2) Blocking frames that render every audio frame end-to-end, maintaining audio continuity.

3) Introducing a window function to enhance the role of audio framing and prevent audio discontinuity caused by sampling and quantization.

4) Fast Fourier Transform which transforms the audio from the time domain to the frequency domain.

5) Map the FFT produce to the Mel scale; multiply it by the total number of triangular bandpass filters.

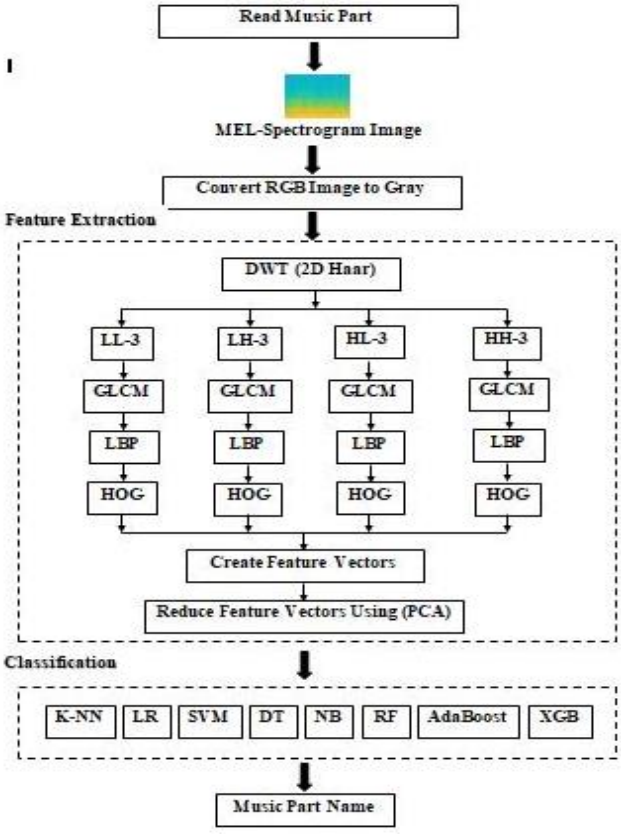


Fig. 1. The proposed system for lute signal classification.

Mel-spectrograms in this study are a two-dimensional representation of input signals. All audio signals are produced using the Short Time Fourier Transform (STFT) with Mel-frequency rather than normal frequency. The parameters used to generate the power Mel-spectrograms are stated in Table II, and Fig. 2 depicts various Mel-spectrograms for the dataset and shows the region of the image used for feature extraction as the black box, and the output is saved as a .png file with a size of 256*256. The output photos are transformed to grayscale before the textural features are extracted.

Parameter	Value
Audio Length (second)	12:91
Window Length (frames)	1024
Overlap Length (frames)	512
FFT Length (frames)	4096
Number bands (Filters)	64

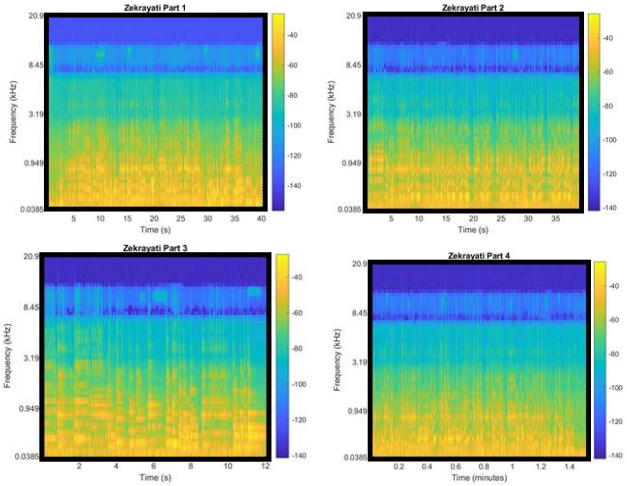


Fig. 2. Mel-spectrogram for some datasets.

B. Feature Extraction Stage

The primary phases for feature extraction are as follows:

Step 1: apply the discrete wavelet transforms to the Mel-spectrogram image to extract sub-bands.

Step 2: extracted all of (GLCM, HOG, and LBP) from all wavelet sub-bands and combine all features into one feature vector.

Step 3: reduce the final feature vector by using Principal component analysis (PCA).

1) Discrete Wavelet Transforms (DWT): It is a powerful image signal analysis tool. It has an effective analysis function and multi-resolution analysis capability, making it suited for the image signal analysis area [32]. The spatial domain (DWT or 2D_DWT) is resulted by first applying the output to the DWT along the vertical axis and then applying the horizontal axis to the (1D_DWT). Hence, (2D-DWT) contains four bands: (LL, LH, HL, and HH) bands [33, 34].

Eq. (1) symbolizes the transformation (DWT) of any signal, x(t).

$$x(t) = \sum a_{j.k} \Psi_{j.k}(t) \quad (1)$$

Where $a_{j.k}$ are called wavelet coefficients, $\Psi_{j.k}(t)$ is called the fundamental function. j is the scale and k is mother wavelet translated $\Psi(t)$.

The (2_D DWT) can be achieved by using Eq. (2) to apply DWT across rows and columns of a picture in both the (x and y) dimensions.

$$f(x.y) = \sum_{j.k} C_{j_0}(k.l)\varphi_{j.k.l}(x.y) + \sum_{s=H.V.D}\sum_{j=j_0}^{\infty} D_j^s[k.l]\Psi_{j.k.l}^s(x.y) \quad (2)$$

Where C_{j_0} is the approximation coefficient, $\varphi_{j.k.l}(x.y)$ is scaling function, D_j^s is set of detailed coefficients and $\Psi_{j.k.l}^s$ is wavelet function.

In this work, three level wavelet decomposition has been performed by using 'haar' mother wavelet function as shown in Fig. 3 and the sub bands of level three were used for extracting the features.

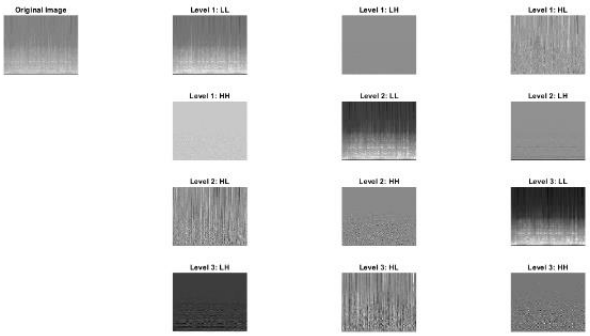


Fig. 3. Three-level wavelet decomposition.

2) GLCM Algorithm: The GLCM texture extraction approach has become more and more popular in recent years for picture classification and detection [35, 36]. It alludes to a widely used technique for characterizing texture through an examination of grayscale's spatial correlation features. It determines the frequency of occurrence for each piece of grayscale data it contains. The GLCM is a (L × L) counting matrix, where each element in the GLCM represents a potential combination of pixels, assuming the original image has (L) grayscale levels. Several studies have demonstrated that this approach is highly adaptable and stable in capturing detailed information such as direction, distance, and variation range between image pixel grayscales. The contrasts and patterns of texture features acquired using this method accurately characterize the properties of picture texture [37]. Some GLCM features [38] used in this work are briefly explained below.

- Contrast: This feature calculates the intensity of a pixel and its surrounding pixels over the entire image. The contrast function also calculates the color and brightness differences between each cellular object and other objects in the same field of view. It can be computed using the following equation.

$$Contrast = \sum_i \sum_j (i-j)^2 p(i.j) \quad (3)$$

For an image, $p(i.j)$ reflects the chance of a pair of pixels with gray level values(i and j)occurring in a specific space and direction.

- Correlation: It computes gray-level linear dependence among pixels at specified distances from one another. The(μ_i and μ_j)are the average of each row and column, and (σ_i and $\sigma_{j)}$ are, correspondingly, the standard deviations for each row and column.

$$Correlation = \sum_i \sum_j \frac{p(i.j)[(i-\mu_i)(j-\mu_j)]}{\sigma_i \sigma_j} \quad (4)$$

- Energy: It calculates regularity or pixel pair repetitions, as illustrated in the equation below. When a pixel pair is repeated multiple times, the energy characteristic returns a higher value.

$$Energy = \sum_i \sum_j p(i.j)^2 \quad (5)$$

- Homogeneity: it is refers to the consistency of element distribution along a GLCM's diagonal. When matrix elements are spread diagonally, homogeneity is high, as calculated by the equation below.

$$Homogeneity = \sum_i \sum_j \frac{p(i.j)}{1+|i-j|} \quad (6)$$

In this study, Mel-spectrogram image characteristics are extracted using the GLCM approach, first set the order of the grayscale co-generation matrix to 16 and selected 0°, 45°, 90° and 135° as the four directions of the grayscale co-generation matrix, The final eigenvalue co-generation matrix was calculated by averaging the four directional matrix eigenvalues. Finally, we retrieved four-dimensional GLCM features for each subband, yielding a final feature vector of 16 features.

3) LBP Algorithm: Local Binary Pattern is a well-known texture descriptor that has been successfully used in works made for several application domains, such as music genre detection [39, 40]. According to [41], it is uses the local neighborhood of a center pixel to find a local binary pattern. The feature vector, which characterizes the image's textural richness, corresponds to the histogram of local binary patterns present in all pixels. There are two basic parameters that can be adjusted to extract the LBP from an image. The first is the number of nearby pixels that will be considered for the central pixel, while the second is the distance between the central pixel and its neighbors. These values are referred to, in turn, as (P and R). Fig. 4 shows examples of Mel-spectrogram images, corresponding maps of LBP values, and the LBP histograms.

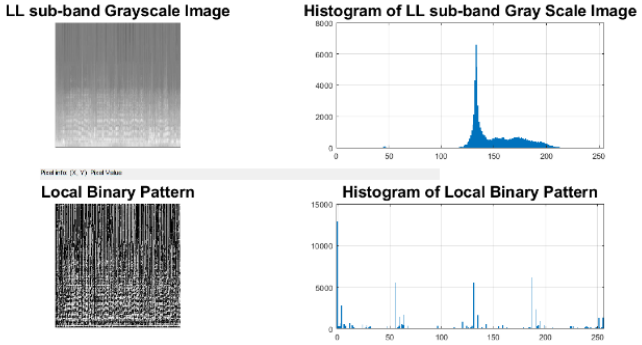


Fig. 4. Local binary patterns visualization.

In this study, 8 neighbors at a distance of 2 was used to extracted 59 features for each sub-bands and the final feature vector for this step was 236 feature value.

4) HOG Algorithm: HOG is a feature descriptor used to detect targets in image processing. It creates features by calculating the histogram of directional gradients in discrete parts of the image [42]. This method consists of two major processes [43]. The first is the histogram extraction, as the gradient of direction and magnitude is retrieved from each pixel in the input image. These steps are used to generate an angular histogram of gradients, which is then applied as an image texture feature vector. The vertical and horizontal components of the image I (i, j) are derivatives of pixel (i, j). They're computed as follows:

$$G_i(i.j) = I(i+1.j) - I(i-1.j) \quad (7)$$

$$G_j(i.j) = I(i.j+1) - I(i.j-1) \quad (8)$$

$$G(i.j) = \sqrt{G_i(i.j)^2 + G_j(i.j)^2} \quad (9)$$

$$\alpha_0(i.j) = \tan^{-1}\left[\frac{G_j(i.j)}{G_i(i.j)}\right].\alpha_0\epsilon\left[-\frac{\pi}{2}.\frac{\pi}{2}\right] \quad (10)$$

where $G_i(i.j)$, $G_j(i.j)$ are the derivatives in the horizontal and vertical directions at pixel (i ,j).

The second phase involves the generation of the HOG descriptor, which is built based on the gradient of the image. The whole image is split into blocks with sizes [2 2], [4 4], and [8 8]. The gradient direction range [-π/2, π/2] is calculated equally into nine direction intervals (bins). To provide a strong vector to brightness changes, the HOG feature values are normalized by segmenting each bin with the total of the histogram. Fig. 5 shows different block sizes of the sub-band HH of the Mel-spectrogram image.

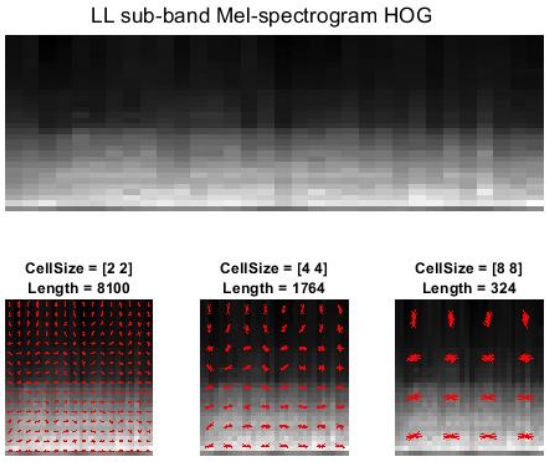


Fig. 5. HOG features of Mel-spectrogram image LL sub-band with different cell sizes.

In this paper the HOG Cell Size of [4 4] is used for extracted 1764 feature value for Each sub-bands of Mel-spectrogram image and the final feature vectored of this step was 7056 feature values.

5) Combining the features of the proposed system and PCA: This part describes a hybrid feature extraction technology. The characteristic of this method is the merger of characteristics derived from HOG, GLCM, WDT, and LBP. The suggested approach is distinguished by its expeditiousness in training the dataset and its demand for computer resources of moderate cost. At first, a Haar 2D wavelet is It is used for extraction 4 sub-bands from the Mel-spectrogram image and extract GLCM, HOG, and LBP for all sub-bands for training the Mel-spectrogram image to create a 440 x 7308 matrix of features. The PCA algorithm [44] is applied to the feature matrix in order to minimize the dimensions and select the most appropriate characteristics for each image.

C. Ml Algorithms

In this paper, many ML algorithms were used to classify the Mel-spectrogram image.

1) K-NN classifier: The K-NN technology is a basic data mining strategy where all samples are assigned to the same group in a feature space, and the algorithm has the same properties for both regression and classification [45]. This technique is considered effective for classification problems.

2) LR classifier: This Classifier [46] is a Strong statistic tool for developing resilient methods. It applies the linear regression principle to classification problems. It predicts dependent data by examining the connection between one or more pre-existing independent variables. The LR formula is represented by the following equation:

$$P = \frac{e^y}{1+e^y} \quad (11)$$

3) SVM classifier: SVM is a collection of supervised learning algorithms. These are commonly used for classification and regression tasks on both linear and nonlinear data [47]. This method finds a decision boundary among two classes in order to forecast labels using one or more feature vectors. It lacks a natural growth to several courses and performs slowly throughout training.

4) DT classifier: One of the most powerful categorization methods is the DT, which simulates decisions using a tree framework. [48]. Computing the data allows it to classify the dataset and assign values to each of its attributes. The decision tree follows a top-down approach. Information gain is a typical strategy for choosing a decision node in DT. The equation for information gain is as follows:

$$IG(D.A) = Entropy(D) - \sum_{v\in Values(A)} \frac{|D_v|}{|D|} \cdot Entropy(D_v) \quad (12)$$

5) NB classifier: NB is a class of supervised learning approaches which employ likely reasoning to anticipate the optimal result [49]. Using the Bayes theorem to it is easy to construct the classifier and the Gaussian normal distribution to forecast the class. The collection of probabilities for a certain

set of data is determined by counting the value and frequency of the value. The Bayesian formula is:

$$P(H|E) = \frac{P(E|H)\cdot P(H)}{P(E)} \quad (13)$$

6) RF classifier: The RF technique, which is based on Ho's method and was later developed and introduced to the literature by [50], is a collective learning technique that determines the output class by training a large number of decision trees and taking the mode or average of their results. It's a popular algorithm because of its high prediction performance, capacity to deal with imbalance issues, and ability to produce consistent results in a variety of applications.

7) AdaBoost classifier: The adaboost boosting algorithm is a well-known ensemble technique for binary classification [51]. The group moves toward AdaBoost trains and installs trees in a sequential manner. AdaBoost combines a series of weak classifiers to perform boosting. The goal of each iteration of the weak classifier is to correct samples that were incorrectly classified by the preceding weak classifier. AdaBoost uses an iterative approach to help bad classifiers get better by using the mistakes they have made.

8) XGB classifier: XGB is an additional ensemble ML method that tackles regression and classification issues by utilizing many decision trees [52]. To lessen overfitting and boost performance, it uses more regularized prediction models.

Table II illustrates the hyperparameter method for determining the optimal parameters for ML algorithms.

TABLE II. THE FINE HYPERPARAMETERS OPTIMIZATION

Model	Hyperparameters
K-NN	n_neighbors=5, Euclidean distance
LR	solver='linear'
DT	max_depth=100, criterion='entropy'
NB	var_smoothing=1e−04
RF	n_estimators=100, max_depth=50
AdaBoost	n_estimators=20, learning rate=0.5 3.3
XGB	n_estimators=100, learning rate=0.1
SVM	Kernel= RBF, C=3.3

IV. RESULTS AND DISCUSSION

A. Dataset

1) Collected data: In this paper, authors have collected a music audio dataset consisting of 440 recordings for different four classes of Arabic music sounds has been collected.

The collected data became available for free at: (https://www.kaggle.com/datasets/emanatyaesmaeil/zekrayati-dataset).

2) GTZAN dataset: GTZAN [53] was one of the first widely available datasets for MGC and is well-known within the scientific community. This dataset contains (1000) music clips from (10) western music genres. GTZAN's ten genres ('blues', 'classical', 'country', 'disco', 'hip-hop', 'jazz',' metal', 'pop', 'reggae', and 'rock') are evenly distributed, each featuring '100' clips. Each music clip is (30) seconds long. Table III shows the description of the dataset.

TABLE III. DESCRIPTION OF THE DATASET

Class Name	No of file	Minimum duration (s)	Maximum duration (s)
Zekrayati (P_1)	120	12	60
Zekrayati (P_2)	120	26	41
Zekrayati (P_3)	80	12	35
Zekrayati (P_4)	120	46	91

B. Performance Evaluation

The results related to ML models have been measured using the following indicators: accuracy, precision, recall, and F1-score. Equations (14 through 17) employ the confusion matrix to calculate these values. Where (TP=true_positive), (FP=false_positiv), (FN=false_negative), and (TN=true_negative) [54].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

$$Recall = \frac{TP}{TP+TN} \quad (15)$$

$$Precision = \frac{TP}{TP+FP} \quad (16)$$

$$F1-Score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (17)$$

C. K-Fold Cross Validation

In this method, epochs are split up into k groups at random, each with roughly the same amount of features. The remaining groups are used to test the learning, while the K-1 groups are used for training. The technique is repeated (k) times, every time with a different group of tests [55]. For performance evaluation, a 5-fold cross-validation procedure is used, and the result is calculated as the 5-fold average.

D. Experiment 1: ML Methods with Proposed Dataset

In this paper, all samples in the proposed dataset converted to Mel-spectrogram images, then three levels of DWT with haar mother wavelet function were used to represent each Mel-spectrogram images, then chosen sub-band of level three for feature extracted by using GLCM, LBP, and HOG and reducing feature vector using PCA, and finally used eight machine learning techniques SVM, K-NN, DT, RF, LR, NB, XGB, and AdaBoost for classification Mel-spectrogram images dataset using 5 k- Fig. 6 depicts the splitting of Mel-spectrogram pictures into 5 k-folds for training and testing. Table IV and Fig. 7 show the level of accuracy ratings of models after applying 5 k-fold. Fig. 8 to Fig. 12 depicts the confusion matrix for various ML methods.

According to Table IV, the XGB model achieves high performance in the music classification when compared to other classifiers. The XGB achieved an average performance of 99.54% for accuracy, 99.44% for F1 score, 99.41% for recall, and 99.51% for precision.

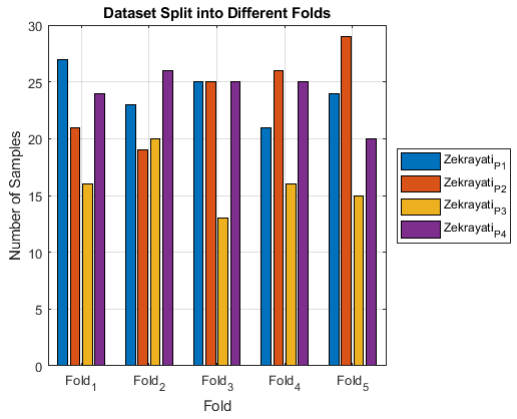


Fig. 6. The split of Mel-spectrogram images into train and test using 5 k-fold.

TABLE IV. SHOWS THE ACCURACY RESULTS OF VARIOUS ML MODELS

Model	Fold	Accuracy	F1-Score	Recall	Precision
K-NN	1	98.86	98.77	99.07	98.53
	2	97.73	97.52	97.60	97.49
	3	98.86	98.56	99.00	98.21
	4	97.73	97.94	97.85	98.15
	5	98.86	99.04	98.96	99.17
	Mean	**98.41**	**98.37**	**98.50**	**98.31**
LR	1	96.59	96.67	96.84	96.56
	2	95.45	95.57	95.39	95.81
	3	96.59	96.54	97.00	96.25
	4	97.73	97.94	97.85	98.15
	5	97.73	97.47	98.10	97.06
	Mean	**96.82**	**96.84**	**97.04**	**96.77**
SVM	1	98.86	98.94	98.81	99.11
	2	98.86	98.83	98.91	98.81
	3	100	100	100	100
	4	98.86	98.72	98.44	99.07
	5	100	100	100	100
	Mean	**99.32**	**99.30**	**99.23**	**99.40**
DT	1	86.36	86.20	86.62	86.14
	2	87.50	87.46	87.35	87.61
	3	86.36	85.50	86.15	85.08
	4	86.36	86.11	86.00	86.37
	5	87.50	86.35	86.55	87.01
	Mean	**86.82**	**86.32**	**86.53**	**86.44**
NB	1	88.64	88.35	88.48	88.73
	2	89.77	90.08	89.97	90.45
	3	88.64	88.19	88.15	88.27
	4	89.77	89.72	89.61	90.06
	5	88.64	88.65	87.68	90.43
	Mean	**89.09**	**89.00**	**88.78**	**89.59**
RF	1	98.86	98.88	98.81	99.00
	2	97.73	97.60	97.43	97.82
	3	97.73	96.94	96.15	98.08
	4	97.73	97.93	98.04	97.86
	5	98.86	98.86	98.96	98.81
	Mean	**98.18**	**98.04**	**97.88**	**98.31**
AdaBoost	1	98.86	98.71	98.96	98.53
	2	97.73	97.70	97.79	97.71
	3	100	100	100	100
	4	98.86	98.63	98.81	98.53
	5	100	100	100	100
	Mean	**99.09**	**99.01**	**99.11**	**98.95**
XGB	1	98.86	98.71	98.96	98.53
	2	100	100	100	100
	3	98.86	98.51	98.08	99.04
	4	100	100	100	100
	5	100	100	100	100
	Mean	**99.54**	**99.44**	**99.41**	**99.51**

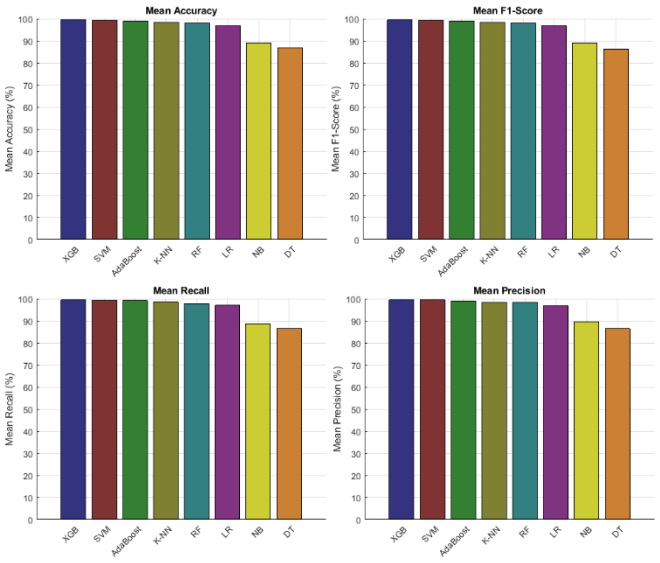


Fig. 7. Result for ML models.

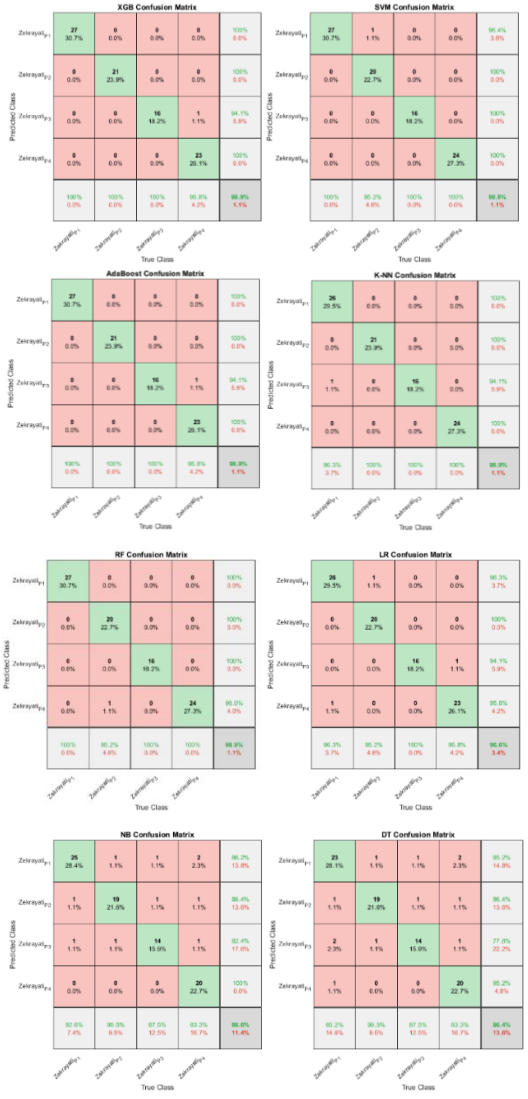


Fig. 8. Shows the confusion matrix of various ML models for fold-1.

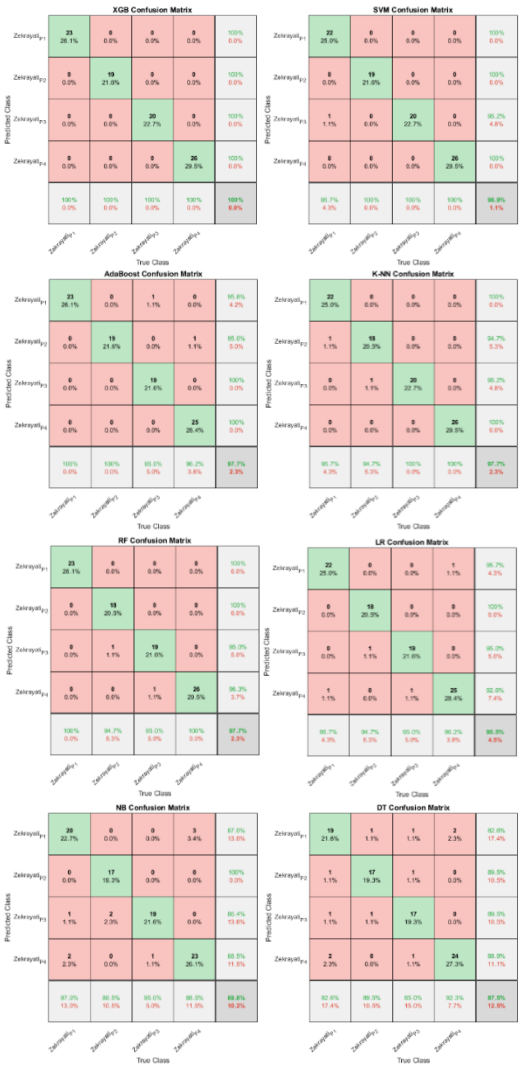
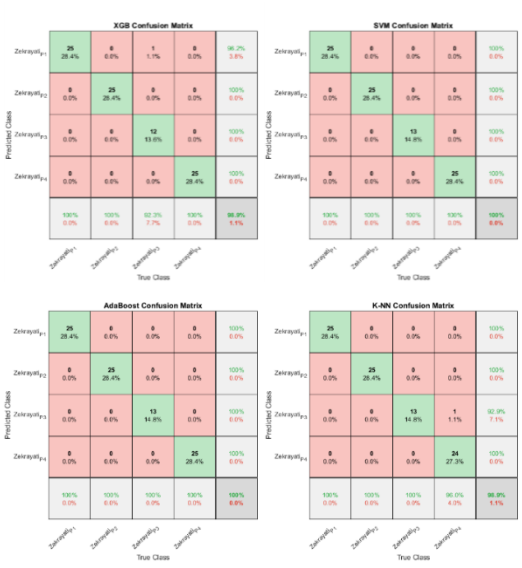


Fig. 9. Shows the confusion matrix of various ML models for fold-2.



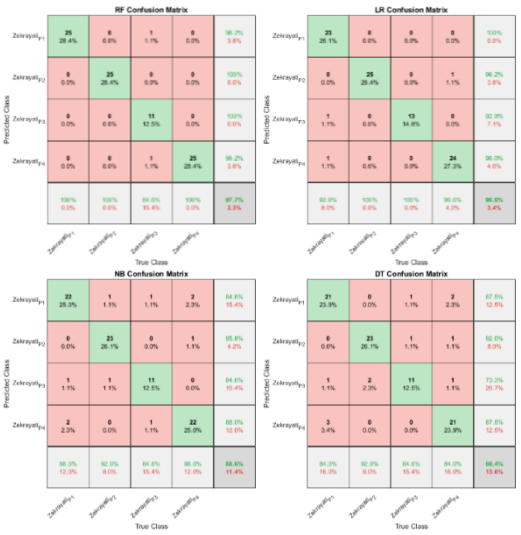


Fig. 10. Shows the confusion matrix of various ML models for fold-3.

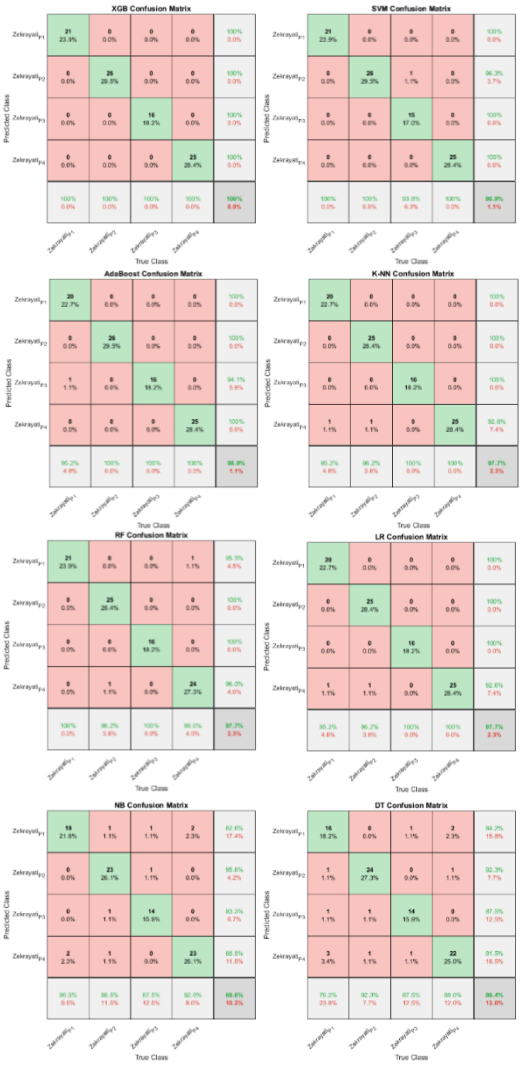


Fig. 11. Shows the confusion matrix of various ML models for fold-4.

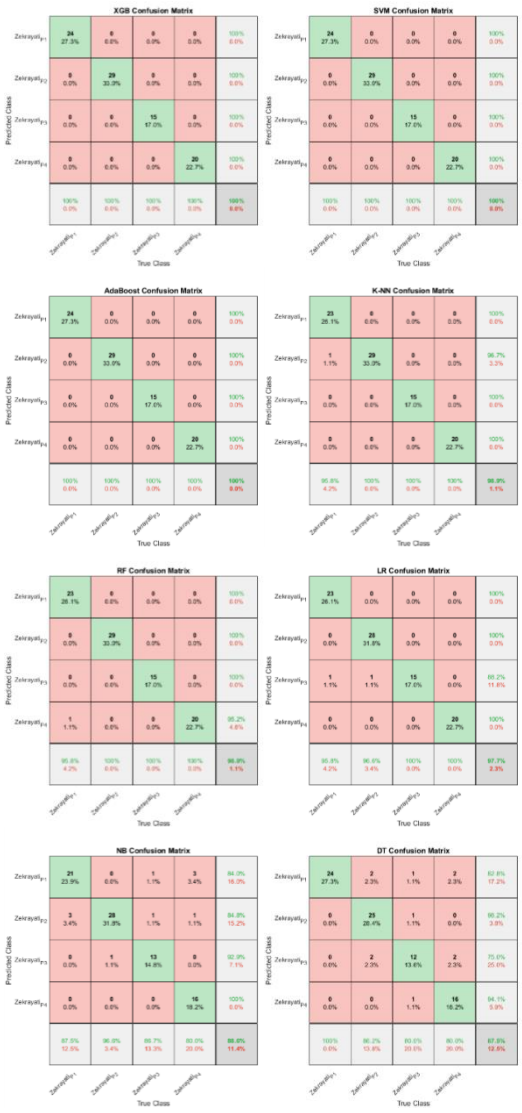


Fig. 12. Shows the confusion matrix of various ML models for fold-5.

E. Experiment 2: ML Methods with GTZAN Dataset

the approach suggested had been compared to state-of-the-art models that used the GTZAN dataset, comprising DL models, specifically convolutional neural networks, bottom-up broadcast neural networks (BBNN), deep unsupervised representation learning from acoustic data auDeep, and ML SVM for categorizing music genres using MEL-spectrogram images and the GTZAN dataset. The steps of the comparison process can be summarized as follows

Step 1: Convert all class to MEL spectrogram images using window length 1024 with overlap length 512, FFT length 4096 and number bands 64 Fig. 13 shown some MEL spectrogram images for GTZAN dataset.

Step 2: DWT is applied using three levels and calculates GLCM, HOG, and LBP for level 3 for each sub-band, and finally reduces the feature vector to 1000 samples and 100 features using PCA.

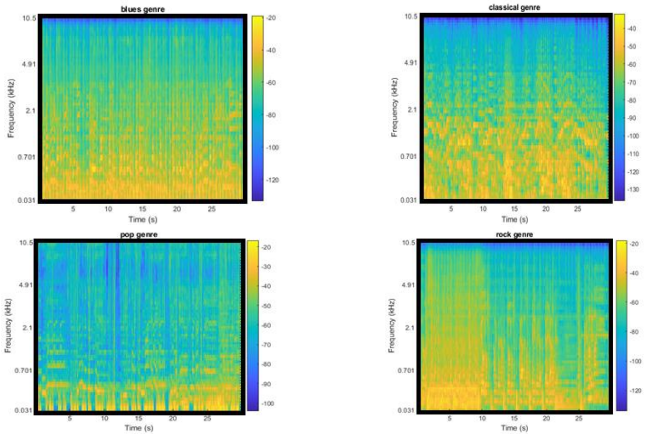


Fig. 13. Mel-spectrogram for some GTZAN dataset.

Step 3: Classification genres using ML models proposed in this paper and the model's performance using Accuracy, F1-Score, Recall and Precision using 5k-fold Fig. 14 shows 5 K-fold split of the GTZAN dataset 20% for testing and 80% for training. Fig. 15 shows the results of ML models for GTZAN dataset. Table V illustrates the accuracy scores and Fig. 16 to 20 shown confusion matrix's.

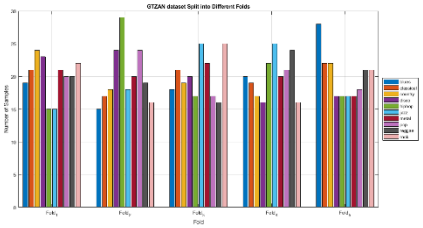


Fig. 14. Using five k-folds, GTZAN splits Mel-spectrogram pictures into train and test.

TABLE V. THE ACCURACY SCORES OF DIFFERENT ML MODELS FOR GTZAN

Model	Fold	Accuracy	F1-Score	Recall	Precision
K-NN	1	95.50	95.29	95.32	95.49
	2	95.00	94.66	94.67	94.87
	3	95.00	94.88	95.27	94.90
	4	95.00	94.69	94.97	95.02
	5	94.50	94.35	94.45	94.60
	Mean	**95.00**	**94.77**	**94.94**	**94.98**
LR	1	93.00	92.91	92.76	93.24
	2	92.50	92.02	92.37	91.92
	3	92.50	92.40	92.46	92.96
	4	92.50	92.42	92.98	92.92
	5	93.00	92.91	92.86	93.23
	Mean	**92.70**	**92.53**	**92.69**	**92.85**
SVM	1	97.50	97.39	97.53	97.30
	2	97.00	96.82	97.03	96.74
	3	97.50	97.53	97.72	97.46
	4	97.00	96.91	97.01	97.13
	5	97.50	97.49	97.30	97.82
	Mean	**97.30**	**97.23**	**97.32**	**97.29**
DT	1	82.50	82.70	82.67	83.43
	2	82.00	81.63	81.73	82.76
	3	82.50	82.01	82.50	82.38
	4	82.00	81.58	81.34	83.39
	5	83.00	81.94	82.52	83.48
	Mean	**82.40**	**81.97**	**82.15**	**83.09**
NB	1	86.50	86.31	86.49	87.05

	2	87.00	87.16	87.82	89.10
	3	87.00	86.53	86.55	88.04
	4	86.50	86.23	86.59	86.63
	5	87.50	86.65	86.37	88.30
	Mean	**86.90**	**86.58**	**86.76**	**87.82**
RF	1	96.50	96.20	96.04	96.47
	2	96.00	95.86	96.00	95.95
	3	96.00	95.86	95.96	95.95
	4	96.00	95.87	95.89	96.18
	5	96.00	96.02	95.94	96.14
	Mean	**96.10**	**95.96**	**95.97**	**96.14**
AdaBoost	1	97.00	96.88	96.92	96.98
	2	97.00	96.80	97.03	96.85
	3	96.50	96.33	96.34	96.49
	4	96.50	96.29	96.24	96.55
	5	97.00	96.99	97.18	96.98
	Mean	**96.80**	**96.66**	**96.74**	**96.77**
XGB	1	98.00	97.91	97.90	97.98
	2	97.50	97.35	97.38	97.47
	3	98.00	97.95	97.97	97.98
	4	97.50	97.45	97.64	97.49
	5	98.00	97.96	97.87	98.13
	Mean	**97.80**	**97.72**	**97.75**	**97.81**

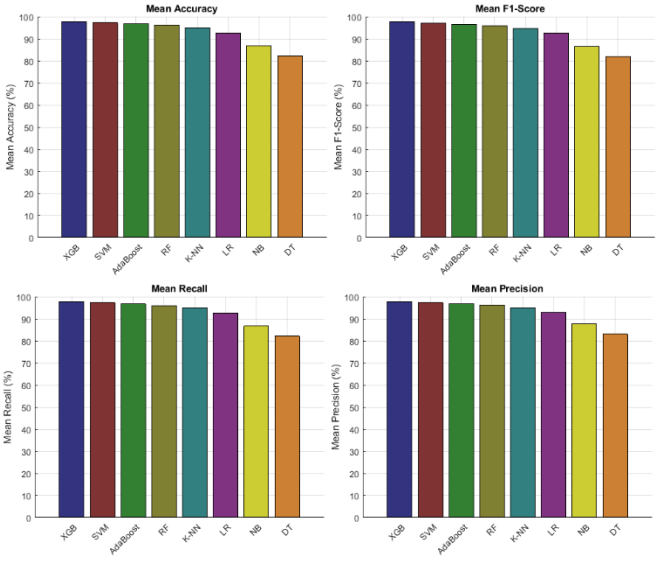
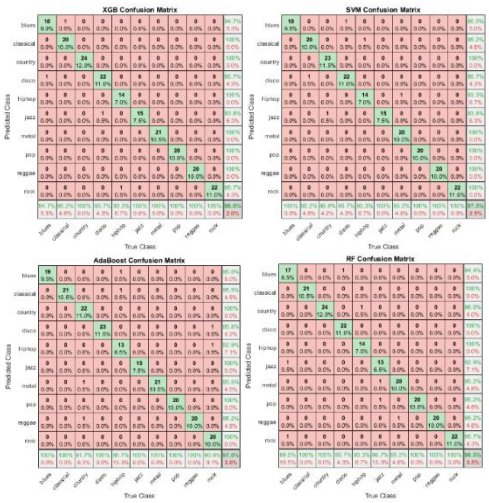


Fig. 15. Results of ML Models for GTZAN dataset.



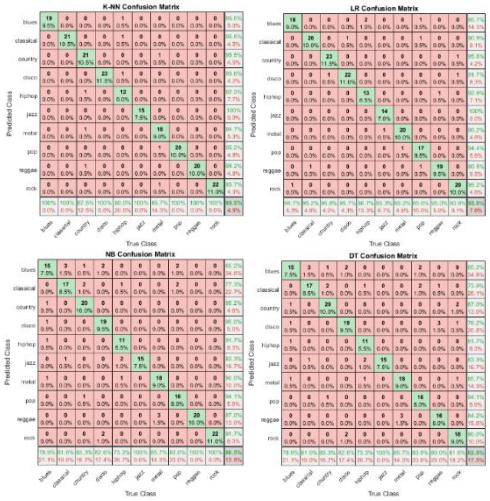


Fig. 16. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-1.

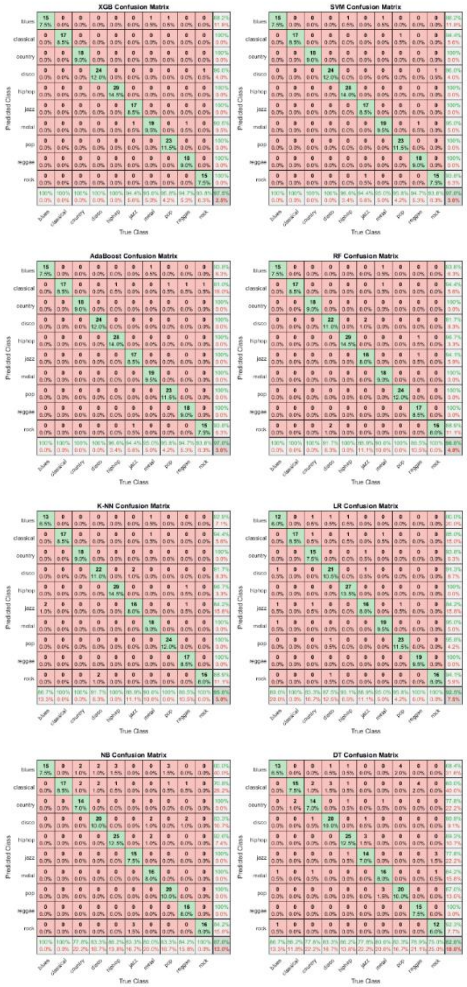


Fig. 17. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-2.

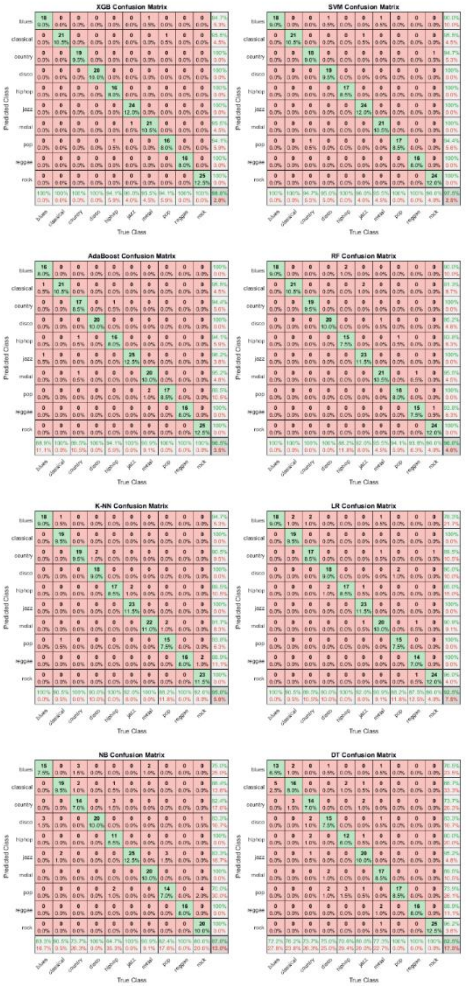


Fig. 18. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-3.

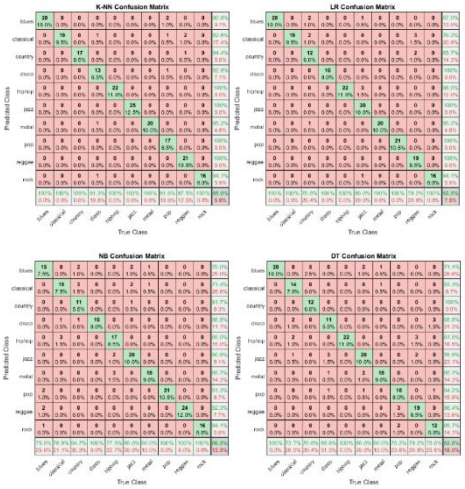


Fig. 19. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-4.

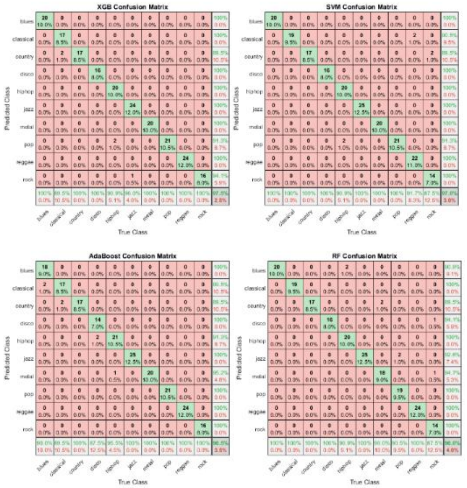


Fig. 20. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-3.

According to Table V, the XGB model achieves high performance in the music classification when compared to other classifiers. The XGB achieved an average performance of 97.80% for accuracy, 97.72% for F1 score, 97.75% for recall, and 97.81% for precision. Table VI shown comparison accuracy with the GTZAN dataset using XGB classifier.

V. DISCUSSION

One of the main objectives of this work is to evaluate the performance of proposed ML models using two different dataset: the collected dataset that the authors had gathered and the global GTZAN dataset.

For the collected dataset , the XGB model achieves high performance in the music classification when compared to other classifiers. The XGB achieved an average performance of 99.54% for accuracy, 99.44% for F1 score, 99.41% for recall, and 99.51% for precision.

By using GTZAN , According to Table VI, the XGB model achieves high performance in the music classification when compared to other classifiers. The XGB achieved an average performance of 97.80% for accuracy, 97.72% for F1 score, 97.75% for recall, and 97.81% for precision. Table VI shown comparison accuracy with GTZAN dataset using XGB classifier.

TABLE VI. COMPARISON ACCURACY WITH GTZAN DATASET

Reference	Feature	Model	Accuracy (%)
Liu, Caifeng, et al. [56]	MEL spectrogram	BBNN network	93.9
Nanni et al. [57]	MEL spectrogram	SVM	90.9
Ghildiyal et al. [58]	MEL spectrogram	CNN	91.00
Nakashika et al. [59]	MEL spectrogram + GLCM	CNN	72.00
Yang et al. [60]	MEL spectrogram	CNN	90.7
Freitag et al. [61]	MEL spectrogram	AuDeep	85.4
Proposed Method using XGB	MEL spectrogram + GLCM + HOG + LBP	XGB	97.80

VI. CONCLUSION AND FUTURE WORK

ML techniques are beneficial for classification tasks, especially music genre classification, in which music is classified into different genres concerning its features. The objective of this paper is the classification of musical sound using ML techniques with texture features and Mel-Spectrogram. In the methodology, the audio data was transformed into a Mel-spectrogram, then texture features were applied to extract the audio features, and finally, a classification task was carried out using six ML classifiers. We performed a complete comparison of six ML classifiers in this study. By using two different datasets, the experimental findings indicated that the XGB exhibited a high accuracy with an average performance of 97.80% for accuracy, 97.72% for F1 score, 97.75% for recall, and 97.81% for precision. Comparing the reviewed related works mainly implemented using various ML and DL algorithms, our method obtained higher accuracy on automatic classification for music.

Future enhancements: Current work processes audio files that are 30 to 90 seconds long. More research should be conducted to handle audio of any length.as well as , Implementing music genre classification for other audio formats can be investigated, as the established ML models perform well for the (.WAV) format, but there are many other formats available, including MP3, FLAC, and others.

Finally, in future work, the authors can combine CNN approaches with texture features to enhance computational efficiency, minimize processing time, and identify music subgenres.

REFERENCES

[1] A. Kumar, A. Rajpal and D. Rathore, “Genre classification using feature extraction and deep learning techniques,” in 2018 10th Int. Conf. on Knowledge and Systems Engineering (KSE), pp. 175– 180, 2018. doi:10.1109/kse.2018.8573325.

[2] K.-K. R. Choo, M. M. Kermani, R. Azarderakhsh, and M. Govindarasu, “Emerging embedded and cyber physical system security challenges and Innovations,” *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 3, pp. 235–236, May 2017. doi:10.1109/tdsc.2017.2664183.

[3] B. Koziel, R. Azarderakhsh, and M. Mozaffari-Kermani, “Low-resource and fast binary Edwards curves cryptography,” *Lecture Notes in Computer Science*, pp. 347–369, 2015. doi:10.1007/978-3-319-26617-6_19.

[4] L. Liu, “Lute acoustic quality evaluation and note recognition based on the Softmax regression BP neural network,” *Mathematical Problems in Engineering*, vol. 2022, pp. 1–7, Apr. 2022. doi:10.1155/2022/1978746.

[5] L. Almazaydeh, S. Atiewi, A. Al Tawil, and K. Elleithy, “Arabic music genre classification using deep convolutional neural networks (cnns),” *Computers, Materials & Continua*, vol. 72, no. 3, pp. 5443–5458, 2022. doi:10.32604/cmc.2022.025526.

[6] F. Ahmed, P. P. Paul, and M. Gavrilova, “Music genre classification using a gradient-based local texture descriptor,” *Smart Innovation, Systems and Technologies*, pp. 455–464, 2016. doi:10.1007/978-3-319-39627-9_40.

[7] A. S. Girsang, A. S. Manalu, and K.-W. Huang, “Feature selection for musical genre classification using a genetic algorithm,” *Advances in Science, Technology and Engineering Systems Journal*, vol. 4, no. 2, pp. 162–169, 2019. doi:10.25046/aj040221.

[8] S. Prabavathy, V. Rathikarani, and P. Dhanalakshmi, “Musical Instrument Sound classification using GoogleNet with SVM and KNN Model,” *Lecture Notes in Networks and Systems*, vol. 300, pp. 230–240, Sep. 2021. doi:10.1007/978-3-030-84760-9_21.

[9] 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, Aug. 2022. doi:10.3390/electronics11162567.

[10] X. Mu, “Implementation of music genre classifier using KNN algorithm,” *Highlights in Science, Engineering and Technology*, vol. 34, pp. 149–154, Feb. 2023. doi:10.54097/hset.v34i.5439.

[11] Y. M. Costa, L. S. Oliveira, A. L. Koerich, and F. Gouyon, “Comparing textural features for music genre classification,” *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, Jun. 2012. doi:10.1109/ijcnn.2012.6252626.

[12] L. Nanni, Y. Costa, and B. Sheryl , “Set of texture descriptors for music genre classification,” In proceeding of the 22nd WSCG International Conference on Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic, 2014..

[13] B. L. Sturm, “An analysis of the GTZAN Music Genre Dataset,” Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies, Nov. 2012. doi:10.1145/2390848.2390851.

[14] A. Yadav, S. Gaikwad, T. Kuigade, and A. Patil, “MUSIC CHORD PREDICTION USING MACHINE LEARNING,” *International*

Research Journal of Modernization in Engineering Technology and Science, vol. 5, no. 12, pp. 194–199, 2023. doi:10.56726/IRJMETS46945.

[15] S. O. Folorunso, S. A. Afolabi, and A. B. Owodeyi, "Dissecting the genre of Nigerian music with Machine Learning Models," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6266–6279, Sep. 2022. doi:10.1016/j.jksuci.2021.07.009.

[16] D. D. Himabindu, K. Avaneesh, M. A. Mudiraj, S. M. Reddy, and M. S. Varma, "Music genre classification using XGB Boost," *Springer Proceedings in Mathematics & Statistics*, pp. 269–276, 2024. doi:10.1007/978-3-031-51167-7_26.

[17] A. Bawitlung and S. K. Dash, "Genre classification in music using Convolutional Neural Networks," *Lecture Notes in Computer Science*, pp. 397–409, Oct. 2023. doi:10.1007/978-981-99-7339-2_33.

[18] T. Li, "Optimizing the configuration of deep learning models for music genre classification," *Heliyon*, vol. 10, no. 2, Jan. 2024. doi:10.1016/j.heliyon.2024.e24892.

[19] K. K. Jena, S. K. Bhoi, S. Mohapatra, and S. Bakshi, "A hybrid deep learning approach for classification of music genres using wavelet and Spectrogram analysis," *Neural Computing and Applications*, vol. 35, no. 15, pp. 11223–11248, Jan. 2023. doi:10.1007/s00521-023-08294-6.

[20] J. Mehta, D. Gandhi, G. Thakur, and P. Kanani, "Music genre classification using transfer learning on log-based Mel Spectrogram," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021. doi:10.1109/iccmc51019.2021.9418035.

[21] M. Kiran Kumar et al., "Automated music genre classification through Deep Learning Techniques," *E3S Web of Conferences*, vol. 430, p. 01033, 2023. doi:10.1051/e3sconf/202343001033.

[22] V. Phulmante, A. Bidkar, Y. Mundada, and P. Kulkarni, "Recognition of music genres using deep learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 9, no. 5, pp. 936–942, 2022.

[23] T. Yin, "Music track recommendation using Deep-CNN and Mel Spectrograms," *Mobile Networks and Applications*, Jul. 2023. doi:10.1007/s11036-023-02170-2.

[24] P. Ghosh, S. Mahapatra, S. Jana, and R. Kr. Jha, "A study on music genre classification using machine learning," *International Journal of Engineering Business and Social Science*, vol. 1, no. 04, pp. 308–320, Apr. 2023. doi:10.58451/ijebss.v1i04.55.

[25] S. Chikkamath et al., "Indian music instrument classification using Deep Learning on embedded platforms," *Lecture Notes in Networks and Systems*, pp. 301–313, 2024. doi:10.1007/978-981-99-9442-7_26.

[26] D. R. Ignatius Moses Setiadi et al., "Comparison of SVM, KNN, and Nb classifier for genre music classification based on metadata," *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, vol. 475, pp. 12–16, Sep. 2020. doi:10.1109/isemantic50169.2020.9234199.

[27] Y.-H. Cheng, P.-C. Chang, and C.-N. Kuo, "Convolutional neural networks approach for music genre classification," in *2020 International Symposium on Computer, Consumer and Control (IS3C)*, Nov. 2020. doi:10.1109/is3c50286.2020.00109.

[28] J. Li et al., "An evaluation of deep neural network models for music classification using spectrograms," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 4621–4647, Feb. 2021. doi:10.1007/s11042-020-10465-9.

[29] M. Preetham, J. B. Panga, J. Andrew, K. Raimond, and H. Dang, "Classification of music genres based on Mel-frequency cepstrum coefficients using deep learning models," *Lecture Notes in Electrical Engineering*, vol. 905, pp. 891–907, 2022. doi:10.1007/978-981-19-2177-3_83.

[30] D. R. Ignatius Moses Setiadi et al., "Effect of feature selection on the accuracy of music genre classification using SVM Classifier," in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2020. doi:10.1109/isemantic50169.2020.9234222.

[31] T. Ahmed, M. A. Alam, R. R. Paul, Md. T. Hasan, and R. Rab, "Machine learning and deep learning techniques for genre classification of Bangla Music," in *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, Feb. 2022. doi:10.1109/icaeee54957.2022.9836434.

[32] Q. Zhang, W. Lu, R. Wang, and G. Li, "Digital image splicing detection based on Markov features in block DWT domain," *Multimedia Tools and Applications*, vol. 77, no. 23, pp. 31239–31260, Jun. 2018. doi:10.1007/s11042-018-6230-z.

[33] N. K. Naik, P. K. Sethy, A. G. Devi, and S. K. Behera, "Few-shot learning convolutional neural network for primitive Indian paddy grain identification using 2D-DWT injection and Grey Wolf optimizer algorithm," *Journal of Agriculture and Food Research*, vol. 15, p. 100929, Mar. 2024. doi:10.1016/j.jafr.2023.100929.

[34] O. Gheyath and D. Q. Zeebaree, " The Applications of Discrete Wavelet Transform in Image Processing: A Review ," *Journal of Soft Computing and Data Mining*, vol. 1, no. 2, pp. 31–43, 2020.

[35] H. Shayeste and B. M. Asl, "Automatic seizure detection based on gray level co-occurrence matrix of STFT imaged-EEG," *Biomedical Signal Processing and Control*, vol. 79, p. 104109, Jan. 2023. doi:10.1016/j.bspc.2022.104109.

[36] R. Anand, T. Shanthi, R. S. Sabeenian, and S. Veni, "GLCM feature-based texture image classification using machine learning algorithms," *EAI/Springer Innovations in Communication and Computing*, pp. 103–125, Oct. 2022. doi:10.1007/978-3-031-20541-5_5.

[37] M. Lv et al., "Sound recognition method for white feather broilers based on spectrogram features and the Fusion Classification Model," *Measurement*, vol. 222, p. 113696, Nov. 2023. doi:10.1016/j.measurement.2023.113696.

[38] M. H. Daneshvari, E. Nourmohammadi, M. Ameri, and B. Mojaradi, "Efficient LBP-GLCM texture analysis for asphalt pavement raveling detection using extreme gradient boost," *Construction and Building Materials*, vol. 401, p. 132731, Oct. 2023. doi:10.1016/j.conbuildmat.2023.132731.

[39] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, F. Gouyon, and J. G. Martins, "Music genre classification using LBP textural features," *Signal Processing*, vol. 92, no. 11, pp. 2723–2737, Nov. 2012. doi:10.1016/j.sigpro.2012.04.023.

[40] A. E. Salazar, "CLBP texture descriptor in multipartite complex network configuration for music genre classification," *Procedia Computer Science*, vol. 222, pp. 331–340, 2023. doi:10.1016/j.procs.2023.08.172.

[41] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002. doi:10.1109/tpami.2002.1017623.

[42] C. Zhu, W. Zhao, and H. Lian, "Image recognition and classification with hog based on nonlinear support Tensor Machine," *Multimedia Tools and Applications*, vol. 82, no. 13, pp. 20119–20138, Dec. 2022. doi:10.1007/s11042-022-14320-x.

[43] J. N. Hasoon et al., "Covid-19 anomaly detection and classification method based on supervised machine learning of chest X-ray images," *Results in Physics*, vol. 31, p. 105045, Dec. 2021. doi:10.1016/j.rinp.2021.105045.

[44] E. M. Senan and M. E. Jadhav, "Diagnosis of dermoscopy images for the detection of skin lesions using SVM and KNN," *Advances in Intelligent Systems and Computing*, vol. 1404, pp. 125–134, 2022. doi:10.1007/978-981-16-4538-9_13.

[45] M. N. Sikder and F. A. Batarseh, "Outlier detection using AI: A survey," *AI Assurance*, pp. 231–291, 2023. doi:10.1016/b978-0-32-391919-7.00020-2.

[46] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic regression model optimization and case analysis," in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Oct. 2019. doi:10.1109/iccsnt47585.2019.8962457.

[47] M. Awad and R. Khanna, "Support Vector Machines for classification," *Efficient Learning Machines*, pp. 39–66, 2015. doi:10.1007/978-1-4302-5990-9_3.

[48] B. Charbuty and A. Abdulazeez, "Classification based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021. doi:10.38094/jastt20165.

[49] A. Zolnierek and B. Rubacha, "The empirical study of the naive bayes classifier in the case of Markov Chain Recognition Task," *Advances in Soft Computing*, vol. 30, pp. 329–336, 2005. doi:10.1007/3-540-32390-2_38.

[50] A. Börekci and O. Sevli, "A classification study for Turkish folk music makam recognition using machine learning with data augmentation techniques," *Neural Computing and Applications*, vol. 36, no. 4, pp. 1621–1639, Nov. 2023. doi:10.1007/s00521-023-09177-6.

[51] R. Wang, "AdaBoost for feature selection, classification and its relation with SVM, a review," *Physics Procedia*, vol. 25, pp. 800–807, 2012. doi:10.1016/j.phpro.2012.03.160.

[52] X. Shi, Y. D. Wong, M. Z.-F. Li, C. Palanisamy, and C. Chai, "A feature learning approach based on XGBoost for driving assessment and risk prediction," *Accident Analysis & Prevention*, vol. 129, pp. 170–179, Aug. 2019. doi:10.1016/j.aap.2019.05.005.

[53] https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification.

[54] M. H. Daneshvari, B. Mojaradi, M. Ameri, and E. Nourmohammadi, "Hybrid texture analysis of 2D images for detecting asphalt pavement bleeding and raveling using tree-based ensemble methods," *Alexandria Engineering Journal*, vol. 107, pp. 150–164, Nov. 2024. doi:10.1016/j.aej.2024.07.028.

[55] B. Oltu, M. F. Akşahin, and S. Kibaroğlu, "A novel Electroencephalography based approach for alzheimer's disease and mild cognitive impairment detection," *Biomedical Signal Processing and Control*, vol. 63, p. 102223, Jan. 2021. doi:10.1016/j.bspc.2020.102223.

[56] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7313–7331, 2020. doi:10.1007/s11042-020-09643-6.

[57] L. Nanni, Y. M. G. Costa, D. R. Lucio, C. N. Silla, and S. Brahnam, "Combining visual and acoustic features for audio classification tasks," *Pattern Recognition Letters*, vol. 88, pp. 49–56, 2017. doi:10.1016/j.patrec.2017.01.013.

[58] A. Ghildiyal, K. Singh, and S. Sharma, "Music genre classification using machine learning," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020. doi:10.1109/iceca49313.2020.9297444.

[59] T. Nakashika, C. Garcia, and T. Takiguchi, "Local-feature-map integration using convolutional neural networks for music genre classification," *Interspeech 2012*, pp. 1752–1755, 2012. doi:10.21437/interspeech.2012-478.

[60] H. Yang and W.-Q. Zhang, "Music genre classification using duplicated convolutional layers in neural networks," *Interspeech 2019*, pp. 3382–3386, Sep. 2019. doi:10.21437/interspeech.2019-1298.

[61] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2018.