Machine Learning Analysis of Music Based on Music Information Retrieval Tasks

Folorunso, S. O.[1*], Banjo, O. O.[1], Awotunde, J. B.[2], Ayo, F. E.[1]

[1]Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria
[2]Department of Computer Sciences, University of Ilorin, Nigeria
*Corresponding author's email: sakinat.folorunso@oouagoiwoye.edu.ng
doi: https://doi.org/10.21467/proceedings.157.3

ABSTRACT

Music Information Retrieval (MIR) methods extracts from music high-level information like classification, musical feature extraction, song similarity and tonality. Musical genre is one of the orthodox methods of describing musical content and a significant part of MIR. At present, few MIR research has been done on Nigerian songs. So, this paper proposed to build a genre classification model based on Mel Spectrogram of audio songs. The process first converts ORIN audio dataset to Mel Spectrogram and extract numerical information from it using the Histogram of Oriented Gradient (HOG) and apply machine learning (ML) models to accurately categorize the songs into different genres of Apala, Fuji, Juju, Highlife and Waka. Support Vector Machine (SVM) with 4 different kernels, with 10- cross validation method were applied and assessed based on Accuracy and Receiver operating characteristics (ROC).

Keywords: Music Information Retrieval, Machine Learning, Support Vector Machine, Music genre, Histogram of Oriented Gradient.

1 Introduction

Music is universal and people create and listen to it. MIR research 1field is primarily tagged with the mining of useful features from music. It also concerned with indexing and the building of different search and retrieval methods like music recommendation systems, user interfaces for browsing large music collections as recommended or content-based search (Downie, 2003). Consequently, MIR goals to making available huge stock of music available to everyone (Downie, 2003). So, various illustrations of music-related topics like video clips, songwriters, composers, albums, performers, consumer and objects like music pieces, etc. will be considered. Considering the importance of music in our society, there is a relatively non-existent MIR research for the Nigerian community. In the work of (Folorunso, Afolabi, & Owodeyi, 2021), the authors presented a new music dataset named ORIN consisting mainly Nigerian music of five different genre. XGBoost ML model achieved superior recall rate of 81.64% compared to SVM, k-NN and Random Forest. (Yandre, Luiz, & Carlos, 2017) proposed a genre classification model based on spectrograms on three publicly available music datasets. The authors used Local Binary Patterns (LBP), Local Phase Quantization (LPQ), Gabor filters feature extraction and compared with a Convolutional Neural Network (CNN). The resultant features were learned on SVM which achieved a good accuracy of 92%.

2 Materials and Methods

This study proposed to convert ORIN (Folorunso, Afolabi, & Owodeyi, 2021) audio dataset to Mel Spectrogram (Song image) in png format. This data are majorly Nigerian songs of different genres (Apala, Fuji, Juju, Highlife and Waka) (Lasisi, 2012; Omojola, 2006) and the dataset distribution is shown in Table 1.

Proceedings DOI: 10.21467/proceedings.157; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-970666-0-3

Table 1: *ORIN Dataset Distribution*

Genre	Size
Apala	100
Fuji	99
Juju	120
Highlife	120
Waka	39
Total	**478**

The Histogram of Oriented Gradient (HOG) (Dalal & Triggs, 2005) technique is used to extract numerical information from the images and ML models is applied to accurately categorize the songs into different genres of Apala, Fuji, Juju, Highlife and Waka. Support Vector Machine (SVM) with 4 different kernels: Linear, Polynomial, Radial Basis Function (RBF) and Sigmoid with 10-cross validation method were applied and assessed based on Accuracy and Receiver operating characteristics (ROC). The research methodology adopted is shown in Figure 1.

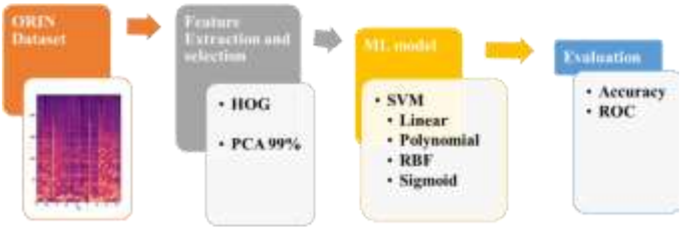


Figure 1: *Research Methodology*

HOG technique was used to extract numeric attributes from the images of the ORIN audio, Principal Component Analysis (PCA) was used to select 99% optimum features (from 46, 657 to 431) to perform MIR task of music genre classification. The resultant feature vectors were learned on Support Vector Machine (SVM) (Chang & Lin, 2011) ML model with 4 different kernels using LIBSVM library. 10- cross validation method was adopted on Waikato Experiment for Knowledge Analysis (WEKA) platform (Frank, Hall, & Witten, 2016). This analysis was inspired by efficient humans' ability at performing this MIR tasks will be the basis for comparison. Some of the other MIR tasks that could be performed for ML modelling is shown in Figure 2.

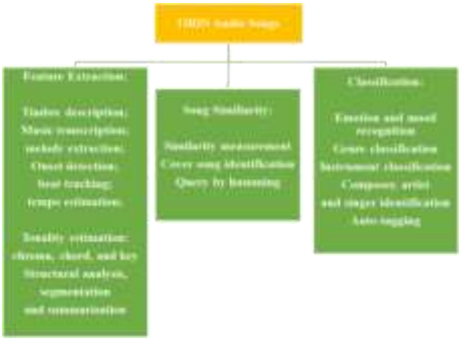


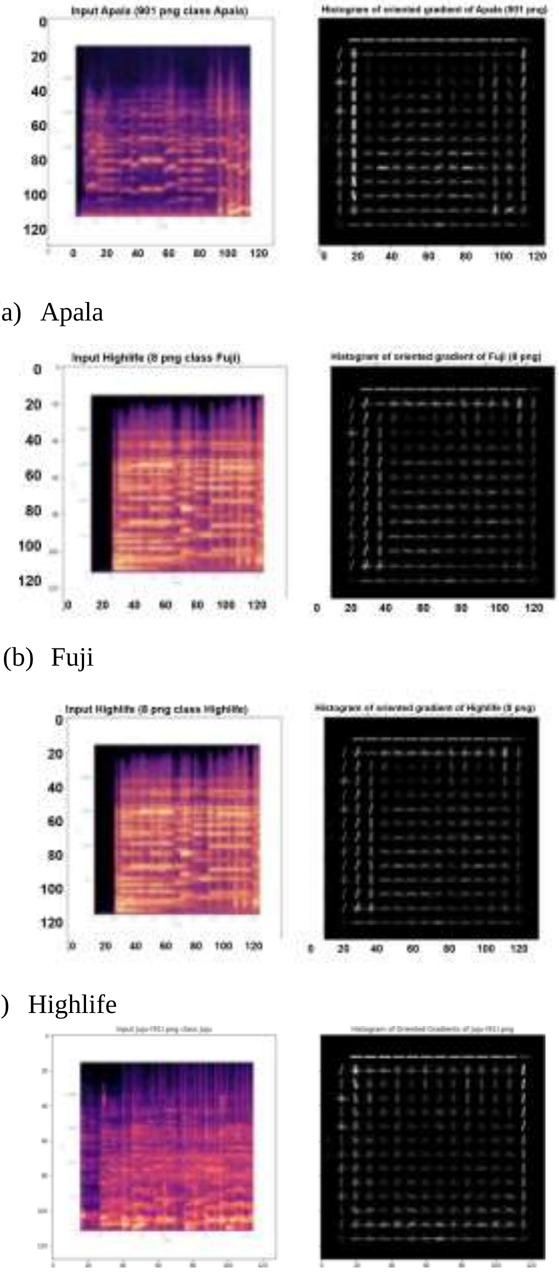
Figure 2: *Project MIR Tasks*

3 Feature Extraction

Mel-spectrogram is a visual representation of audio signals. Every audio sector of a song can be represented as a Mel-spectrogram (De Benito-Gorron, Lozano-Diez, Toledano, & Gonzalez-Rodriguez, 2019). This image is a time : frequency graph where the y- axis is frequency which uses the Mel-frequency scale (Stevens, Volkmann, & Newman, 1937), a log-based perceptual representation of the spectrum. The Mel-spectrogram is converted based on the calculation of the short-time Fourier transform (STFT) spectrogram. The frequency bins of the STFT are then converted to the Mel scale via a Mel-filter bank. In this study, hamming windows of 32 ms with 20 ms shifts and 128 Mel-filters was used. The modulus M of the obtained Mel-spectrograms has been converted to decibels by equation (1).

$MdB = 20 \log10(1 + M)$ (1)

The Mel-spectrogram and HOG representations for this study are illustrated by Figure 2.

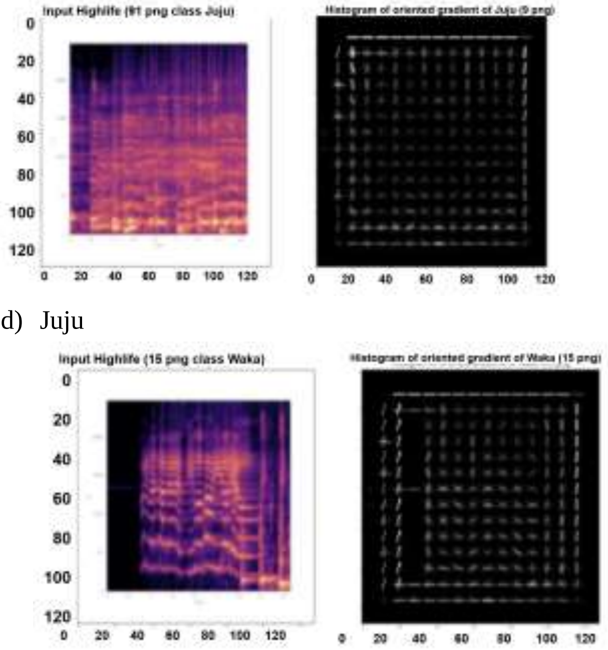


(a) Apala

(b) Fuji

(c) Highlife

j



(d) Juju

(e) Waka

Figure 2: *The spectrogram and HOG for ORIN dataset where γ, r and d are parameters for the kernel.*

4 ML Model

This section briefly describes the ML model SVM with its four different kernels. SVMs (Cortes & Vapnik, 1995) are suitable methods for classification task. SVM models the training data to predict the label values of the test data given only the variables of the test data. So, given a set of instance- label pairs (xi, yj); where x is set of variables and y is set of labels. The main requirement finds solution to the optimization problem in equation (2) while the different kernels are presented by equations (3) to (6).

$$SVM = \min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \qquad (2)$$

$$\text{Subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, i = 1, \cdots, l$$

Where C is the regularization parameter, x is set of training instances,

$$Linear\ Kernel = K(x_i, x_j) = x_t^T x_j \quad (3)$$

$$Polynomial\ Kernel = K(x_i, x_j) = (\gamma x_t^T x_j)^d, \gamma > 0 \quad (4)$$

$$Radial\ Basis\ Function\ (RBF) = K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0 \quad (5)$$

$$Sigmoid = K(x_i, x_j) = \tanh(\gamma x_t^T x_j + r) \quad (6)$$

Where $\gamma, r, and\ d$ are parameters for the kernel.

5 Model Evaluation

The ML model SVM deployed in this study uses accuracy and ROC metrics to evaluate the model performance. The metrics are presented by equations (7) to (10). Accuracy shows the percentage of correctly classified samples while ROC is the tradeoff between the TPR and FPR (Folorunso *et al.*, 2022). The Mel-spectrogram image was extracted from the audio using the librosa library (McFee, *et al.*, 2019). 224 Mel filter banks was used to generate the Mel spectrograms. Parameters of 2048 sample size of hanning window and hop-length of 512 samples (≈ 32 ms) were used in the study on scikit-learn library (Pedregosa, *et al.*, 2011).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$TPR = \frac{TP}{TP+FN} \quad (8)$$

$$FPR = \frac{FP}{FP+TN} \quad (9)$$

$$ROC = \frac{1+(TPR-FPR)}{2} \quad (10)$$

Where True Positive = TP, True Negative = TN, False Positive = FP, False Negative = FN, True Positive Rate (TPR), False Positive Rate (FPR)

6 Result and Discussion

This analysis inspired by efficient humans' ability at performing this MIR tasks will be the basis for comparison. The experimental results obtained showed that the proposed method achieve a good accuracy and Receiver operating characteristics (ROC).

$SVM = \min \imath 1$

$w^T w + C \sum \xi_i$ (2)

The higher the values of these metrics, the better. Table 2 showed the results obtained

w,b,ξ 2

i=1

based on the evaluation metrics for SVM and the 4 different kernels based on Accuracy and ROC with their

Subject to $y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i,$

$\xi_i \geq 0, i = 1, \cdots, l$

Where C is the regularization parameter, x is set of training instances, corresponding standard deviation values. This outcome shows that SVM with the linear kernel obtained the higher accuracy score of 53.59% with 7.19 standard deviation value. Also, for ROC metric, SVM with the linear kernel obtained the higher ROC score of 0.70 with 0.05 standard deviation value.

Table 2: *Result obtained by SVM*

Dataset	Linear	Polynomial	RBF	Sigmoid
Accuracy	**53.59**	38.69	43.12	47.89
	(7.19)	(6.13)	(5.80)	(6.24)
ROC	**0.70**	0.61	0.63	0.66
	(0.05)	(0.07)	(0.08)	(0.04)

7 Conclusion

This study proposed a Mel spectrogram-based method for classification of Nigerian music genre for ORIN audio dataset. The dataset was first changed to spectrogram, HOG was used to extract features and PCA was used to select features. Then, SVM with four different kernels were used to learn the resultant dataset and evaluated based on accuracy and ROC. Future direction for this work includes the combination of the Mel spectrogram with handcrafted features to improve the classification performance.

8 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in institutional affiliations.

How to Cite

Folorunso *et al*. (2024). Machine Learning Analysis of Music Based on Music Information Retrieval Tasks. *AIJR Proceedings*, 21-27. https://doi.org/10.21467/proceedings.157.3

References

[1] Chang, C. C., & Lin, C. J. (2011). LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 1-27.

[2] Cortes, C., & Vapnik, V. (1995). Support-vector network.*Machine Learning, 20*, 273 - 297.

[3] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)* (pp. 886 – 893). San Diego, Calif, USA: IEEE.

[4] De Benito-Gorron, D., Lozano-Diez, L., Toledano, D. T., & Gonzalez-Rodriguez, J. (2019). Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. *EURASIP Journal on Audio, Speech, and Music Processing , 9*, 1-19. doi:10.1186/s13636-019-0152-1.

[5] Downie, J. S. (2003). Music Information Retrieval. *Annual Review of Information Science and Technology, 37*, 295–340.

[6] Folorunso, S. O., Afolabi, S. A., & Owodeyi, A. B. (2021). Dissecting Genre of Nigerian Music with Machine Learning Models. *Journal of King Saud University- Computer and Information Sciences*, 1-24. doi:https://doi.org/10.1016/j.jksuci.2021.07.009.

[7] Folorunso, S. O., Awotunde, J. B., Adeboye, N. O., & Matiluko, O. E. (2022). Data Classification Model for COVID-19 Pandemic. In A. E. Hassanien, S.M. Elghamrawy, & I. Zelinka (Eds.), *Advances in Data Science and Intelligent Data Communication Technologies for COVID-19* (Vol. 378, pp. 93 - 118). Springer. doi:10.1007/978-3-030-77302-1_6.

[8] Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques* (Fourth Edition ed.). Morgan Kaufmann.

[9] Lasisi, S. A. (2012). Traditional music in Nigeria: example of Ayinla Omowura's music . *Developing Country Studies, 2*, 108-118.

[10] McFee, B., McVicar, M., Balke, S., Lostanlen, V., Thomé, C., Raffel, C., . . . Carr, C. J. (2019). Librosa. *librosa/librosa: 0.6.3.*,18-24. doi:10.5281/zenodo.2564164.

[11] Omojola, B. (2006). Popular music in western Nigeria : theme, style, and patronage system. *Ibadan:IFRA*.

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython.*Journal of Machine Learning Research, 12*, 2825--2830.

[13] Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am. , 8*(3), 185– 190. doi:10.1121/1.1915893.

[14] Yandre, M. C., Luiz, S. O., & Carlos, N. S. (2017). An evaluation of Convolutional Neural Networks for music classification using spectrograms. *Applied Soft Computing, 52,* 28–38. doi:10.1016/j.asoc.2016.12.024.