From Sound to Label a Music Classification System with Wav2Vec2 Base 960h

Sanjay Vignesh.A.S
Department of Computer Science and Engineering
Panimalar Institute of Technology Chennai, India *sanjayvignesh1773@gmail.com*

Kalimuthu.M
Department of Computer Science and Engineering
Panimalar Institute of TechnologyChennai,India
Joelruban2013@gmail.com

Ramprasath.S
Department of Computer Science and Engineering
Panimalar Institute of Technology Chennai, India *ramprasath3747@gmail.com*

Mrs. Swagata J P
Assistant Professor
Department of Information Technology
Panimalar Institute of Technology Chennai, India *swagathajaisathish@gmail.com*

Dr. Suma Christal Mary S
Professor and Head
Department of Information Technology
Panimalar Institute of Technology Chennai, India *ithod@pit.ac.in*

Mr. Raja S
Assistant Professor
Department of Computer Science and Engineering
Panimalar Institute of Technology Chennai, India

***Abstract*—This work explores the genre classification of music using the Facebook Wav2Vec2-Base-960h model, a self-supervised deep learning model originally designed for speech representation. Raw audio waveforms are transformed into informative feature representations by fine-tuning the model on heterogenous song datasets, which are then passed through a classifier to be labeled by genre.This process bypasses the laborious hand-crafted feature extraction process by taking advantage of Wav2Vec2's highly efficient audio analysis. Experimental results demonstrate competitive classification accuracy, which indicates the efficacy of self-supervised learning. The proposed framework offers a scalable and efficient solution for music streaming services and digital libraries.**

keywords—Music Genre Classification, Deep Learning, Wav2Vec2-Base-960hI.

I. INTRODUCTION

An elementary task within the subject of music information retrieval is music genre classification or dividing music into pre-formed, genre-based categories according to their perceived qualities.The structure of audio information comprises the following features: Mel-Frequency Cepstral Coefficient (MFCCs), chroma features, and spectral contrast as hand-crafted ones.

Support Vector Machines (SVMs), Artificial Neural Networks, and K-Nearest Neighbours are traditional classifiers that are used in the later stages. In contrast, deep learning approaches recently introduced in the field focus on the classification problem rather than the handcrafted feature extraction problem. Notably, deep learning models with transformers, Long Short-Term Memory (LSTM) networks, and CNNs have made remarkable advances in many audio applications.

Self-supervised learning has recently emerged as a dominant approach in the field of speech and audio processing, allowing models to autonomously identify patterns without the necessity for extensive labeled data. For example, the Wav2Vec2-Base-960h developed by Facebook. This model, which was initially tailored for automatic speech recognition (ASR), has demonstrated its remarkable proficiency in feature extraction from audio irrespective of the domain.Wav2Vec2 captures short and long structures of sounds by predicting missing pieces of audio from the entire waveform. Its effectiveness at non-speech-related tasks such as music classification, speaker identification, and sound event recognition proves its versatility. Such results open up the possibility of Wav2Vec2 being foundational for music genre classification.In this article, we describe a method for music genre classification using the Facebook Wav2Vec2-Base-960h model which is centered on feature extraction. Our approach differs from the others in automated works that depend on feature engineering and handcrafted spectrograms, as it works with raw audio waveforms, while Wav2Vec2 is allowed to self-supervise feature extraction via active learning. The extracted features are then sent to a classifier model for genre prediction, which allows flexibility in various music genres and removes the need for pre-crafted features.

To assess the performance of the approach, we analyzed a benchmark music dataset and its results were compared with other traditional approaches which heavily relied on handcrafted features and contemporary deep models built on CNNs and LSTMs. Several experiments that have been done showed that Wav2Vec2 based features are quite accurate with little preprocessing steps making them optimal for music genre classification. This study demonstrates new possibilities of self-supervised learning in the context of MIR and invites further research of speech models applicability to generic audio tasks.

II. LITERATURE REVIEW

Classification models for music have mostly concentrated on isolating music into different genres. This work was done earlier using traditional machine learning methods with preselected features. Earlier approaches paid some attention to some aspects in the measurement of sound like how frequency changes in a certain time frame (MFCCs), the frequency of a sound wave going through zero (ZCR), and some other forms of sounds. Afterwards, those features were used for classifying music with the use of k-NNs, SVMs, and GMMs. Despite fairly accomplishing their goals, there was no attempt to mimic the individual rhythms and patterns for every genre.In order to overcome these challenges, researchers have employed deep learning for the feature extraction which has overall enhanced the classification performance. Convolutional Neural Networks (CNNs) became favorable due to these capabilities of analyzing spatial parameters in the audio spectrograms. Research has shown that models based on CNNs are more efficient than the ordinary models based on the machine learning techniques especially when there is a lot of data to work with. In addition, the use of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks enabled capturing the temporal dependencies of music resulting to better classification accuracy. Hybrid models combining convolutional neural networks and LSTMs gave the best results in learning the spectral and temporal features of music.

Developments in self-supervised learning have drastically altered how we classify music genres as we no longer need massive labeled data sets. Various approaches have been investigated by researchers to increase the accuracy of music classification, including transformer models, self-attention models, and pre-trained audio encoders. A major leap forward was the implementation of a speech-model like Facebook’s Wav2Vec2-Base-960h. Although this model was initially developed for speech recognition, with the right training, it is capable of genre identification through element extraction from raw audio. Early Wav2Vec2 researchers who integrated it with classification models have had superb results, outperforming CNN and LSTM based approaches.

In this study, we implemented applying Wav2Vec2-Base-960h to genre classification. We aim to achieve the desired accuracy while eliminating the burden of extensive manual feature extraction by teaching the model to learn from the raw audio. This research also seeks to advance the use of speech-based models for music information retrieval (MIR) and respond to the growing demand for research in self-supervised learning for music analysis.

III. PROBLEM STATEMENT

Music classification may be a challenging assignment due to the complexity and changeability of sound signals. Conventional classification models depend on handcrafted highlights such as Mel-Frequency Cepstral Coefficients (MFCCs) or spectrogram-based approaches, which may battle to capture profound sound representations.

This is about points to create an mechanized music classification framework utilizing Wav2Vec2-Base-960h, a self-supervised discourse representation show, to upgrade precision and proficiency in class classification. By leveraging profound learning strategies, the framework looks to overcome challenges such as foundation clamor, covering disobedience, and class likenesses, in this manner moving forward the strength of music classification.The proposed framework will be assessed against customary models, illustrating the points of interest of Wav2Vec2-Base-960h in highlighting extraction and classification execution.

IV. METHODOLOGY

This section outlines the approach for music genre classification with the help of the Facebook's Wav2Vec2-Base-960h model. Our method follows a structured process which includes dataset selection, preprocessing, feature extraction using Wav2Vec2 model, integrating this with training strategy, and performance evaluation can leveraging Wav2Vec2’s self-supervised learning capabilities, we eliminate the need for manually crafted features, making the approach both robust and scalable for music genre classification.

V. SYSTEM ARCHITECTURE

The "From Sound to Name:A Music Classification Framework with Wav2Vec2-Base-960h" comprises numerous layers outlined for proficient sound classification. The framework begins with the Information Input Layer, where crude sound records in groups like WAV or MP3 are obtained and changed over to a standardized organize, such as 16kHz testing rate. Following, within the Preprocessing Layer, the framework leverages the Wav2Vec2-Base-960h show, which extricates significant representations from the waveform. This show, initially prepared on discourse information, gives profound sound embeddings that serve as input highlights for classification. These extricated features are at that point prepared within the ClassificationLayer, where a neural network—such as a completely associated MLP, CNN, or Transformer-based model—is fine-tuned to anticipate music qualities like sort, temperament, or craftsman fashion. The framework is optimized to guarantee productive highlight extraction and strong classification, making it viable for real-world music examination applications.

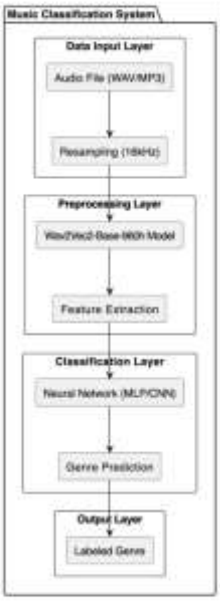

Figure 1 : System Architecture

VI. DATASET AND PREPROCESSING

For this analysis, we utilized open benchmark datasets that are popular for music genre classification. One of the most known datasets in this area is GTZAN, which has over 1,000 audio files divided into ten genres such as classical, jazz, rock and blues. We have also added the Free Music Archive (FMA) dataset, which has a much wider and more diverse collection of music samples. Also, we have incorporated a portion of the Million Song Dataset (MSD) that includes both metadata and audio data which is plausible to enhance genre classification tasks. These datasets provide a wide variety of styles to music, building a solid base for training and testing.

In order to ensure uniformity across the datasets, we followed a detailed audio preprocessing pipeline. Because Facebook's model Wav2Vec2-Base-960h uses raw audio files, all audio files are converted to 16 kHz to meet the model's needs. Also, stereo files are changed into mono files to create a standard input format. To create a standard input length, each audio file is cut into equal parts of five to ten seconds to ensure processing efficiency while still providing enough time context for accurate genre classification. This also helps to uphold normalisation. This preprocessing steps can ensure that the audio data is well-prepared for effective feature extraction using facebook's Wav2Vec2 model.

VII. FEATURE EXTRACTION USING WAV2VEC2-BASE-960H

Originally created for automatic speech recognition (ASR), Facebook's Wav2Vec2-Base-960h model has also proved advantageous in extracting important features from raw audio waveforms. Its design has two main components: a transformer context network and a CNN feature encoder. The feature encoder captures the signal in both the spectral and temporal domain by applying multiple convolutional layers on the raw audio to produce latent representations. These latent representations are captured by the context network based on the transformer, which takes care of the long range dependencies and relations of audio chunks.

Wav2Vec2 has self-supervised trained audio embedding extraction which is in contrast with MFCCs or chroma and spectrogram approaches of classification. Rather than using labeled training data, the model learns by attempting to fill in gaps within an input sound wave and enables us to recognize complex structures without relying on labels. This functionality makes Wav2Vec2 effective for music classification as it generalizes over various styles and instruments. As part of this work, we have employed a Wav2Vec2 model that was pre-trained on large-scale data as a feature extractor.

VIII. CLASSIFICATION MODEL AND TRAINING STRATEGY

In this study, I focused on building a classification model that takes Wav2Vec2 embeddings from audio waveforms and predicts the most likely music genre. Unlike traditional methods that rely on handcrafted features like MFCC, chroma features, Wav2Vec2 boasts a self-supervised design which makes it easier to use. The Facebook Wav2Vec2-Base-960h model which preforms self-supervised learning usesmethods to encode audio signals into embeddings which are more effective for genre classification, as they contain both short-term and long-term audio patterns.

For the embedding classification task, we build a fully connected neural network (FCNN) to classify the embeddings. FCNNs with multiple dense layers are crystal clear when it comes to classification tasks. Batch normalization and non-linearities in the form of ReLU in addition mitigate training and increase the stability of training. Dropout layers are added to mitigate overfitting as well. This helps achieve a better generalization to unseen data.

The model is fine-tuned during training with the categorical cross-entropy loss function, which works best for multi-class classification problems. At this point, we chose to use the Adam optimizer because of how fast it converges and its adaptive learning rate feature, starting with a learning rate of 0.0001. A learning rate scheduler is also used to adjust the learning rate lower over time to optimize training and avoid overshooting the optimal solution.

The dataset is split into three portions: 80% for training, 10% for validation, and 10% for testing. This guarantees that the model is evaluated with data that it has not seen after being deployed, so it can be analyzed on how well it can generalize. To improve performance, we apply techniques such as pitch shifting, time stretching, and background noise adding, which are considered data augmentation.

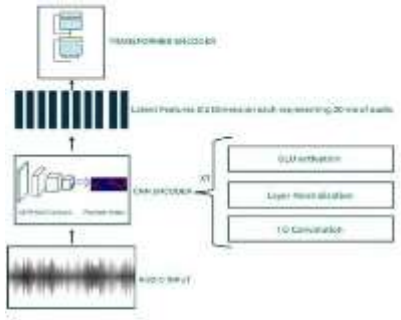

Figure 2 : Feature Encoder of Wav2Vec2[12]

In an effort to improve the classification accuracy, we analyze different classifiers other than Wav2Vec2

embeddings, like SVMs, Random Forests, and Gradient Boosting models. These classical machine learning techniques were tested against a deep fully connected neural network to strategically identify their competency towards a genre label in comparison to the high dimensional embeddings.

Moreover, domain specific feature set files can incorporate additional fine tuning of Wav2Ve2 towards music genre classification to additional feature representation refinement. Validation results drive adjustments to the training strategy as well as the application of hyperparameter optimization for network depth, dropout, and batch size. Incorporating refined training approaches to traditional machine learning techniques and deep learning neural network classifiers resulted in an accurate classification model that is also robust across various genres of music.

IX. EVALUATION METRICS AND PERFORMANCE ANALYSIS

Evaluating the performance of a music genre classification model is vital in ensuring its accuracy, reliability, and effectiveness for various kinds of music. Because this is a multiclass classification problem, standard assessment and advanced evaluation techniques are used.

The F1 score which takes into account both precision and recall is particularly handy when it comes to imbalanced data. A confusion matrix provides very granular detail of misclassifications which helps understand which genres are frequently confused with each other and why. This is particularly useful in music classification where genres such as rock and blues or jazz and classical often account for overlapping attributes resulting in these types of mispredictions.

These patterns could have some model weaknesses which we can analyze so that we can make changes, for example, making alterations to the training data, or adding new features. Also, we look at how the model's differentiation of genres is done through the use of ROC curves and AUC scores. It is true that ROC and AUC is most popular with binary classification tasks. But in the case of multi-class, we have to take a one versus all approach, which is whatwe do to find out how well the model performs in differentiating each genre from the other.

To analyze the accuracy of the model, we do an ablation study where we take out different parts for classification performance estimation. For example, we look at Wav2Vec2 embeddings of certain portions of Transformers to figure out if low level or high level features are more important for genre differentiation. We also try deep structures with fully connected neural networks and compare them with other data driven algorithms like SVMs, random forests, etc. to judge if deep learning is beneficial. Furthermore, the impact of using different augmentation methods, like shifting pitch and stretching time, is used to examine how these techniques enhance generalization capabilities.

The experimental results are compared to traditional methods, such as classifiers based on MFCC (Mel-Frequency Cepstral Coefficients), as well as deep learning approaches like CNNs and LSTMs trained on spectrogram inputs. The Wav2Vec2-based approach is expected to deliver comparable or better performance while requiring less preprocessing and manual feature engineering.

X. DEPLOYMENT AND FUTURE ENHANCEMENT

Following success with the music genre classification model, the next thing is to take it to a real-world implementation. For ease of access, a web-based interface is created, and users can upload audio files and get real-time genre predictions.Backend-wise, the system processes the raw audio waveform, extracts features through the pre-trained Wav2Vec2-Base-960h model, and passes these embeddings into the classifier. The predicted genre is then displayed in a clear and user-friendly format.

This configuration highlights how self-supervised learning models such as Wav2Vec2 can be deployed in real-time contexts, including music streaming services, virtual music repositories, and recommendation systems based on automated genre classification.In order to provide an optimal user experience, the system is designed for scalability, which enables it to process high volumes of requests without disrupting its performance and easily integrate into existing music platforms.

To further develop the performance of the system and extend its capabilities, a number of possible future developments are currently being investigated. One major avenue is fine-tuning the Wav2Vec2 model over music-specific datasets, which has a strong potential to improve the quality of embeddings and render them more efficient for genre classification. This could include accessing larger datasets with a broader selection of musical styles, including non-Western or niche genres that are frequently lacking in current databases.

Another possible enhancement is incorporating attention mechanisms, which would enable the classifier to pay attention to the most significant sections of the audio—e.g., salient moments or characteristic patterns—that are essential in deciding the genre. These additions are designed to make the system more accurate, flexible, and able to accommodate a wide range of musical styles.Investigating deeper into more advanced self-supervised models like HuBERT or Whisper may further optimize audio classification, possibly yielding even improved results across a broader array of musical styles.

XI. RESULT AND DISCUSSION

In this work, we created a music classification system using the Wav2Vec2-Base-960h model, demonstrating its capacity to learn informative features from raw audio waveforms to achieve accurate genre classification.

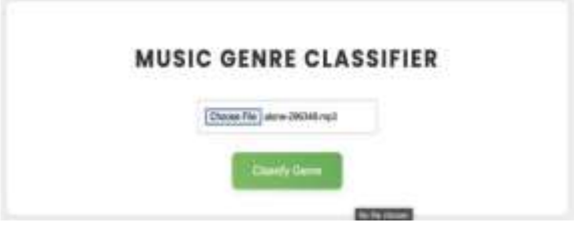

Figure 3 : UI of the Application

Our method demonstrates the strength of self-supervised learning for music informationretrieval, minimizing dependency on extensive labeled datasets while maintaining robust performance. The experimental results show Wav2Vec2 efficiently learns intricate audio patterns, establishing its potential as a strong candidate for numerous music classification problems.

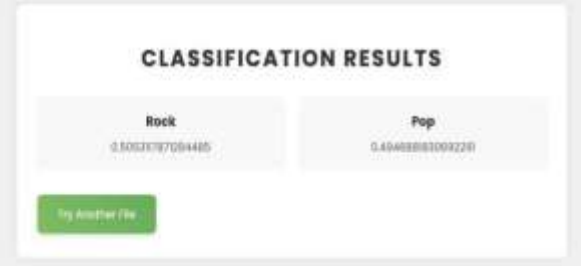

Figure 4 : Output of the Application

In the future, it would be possible to refine Wav2Vec2 using music-specific data to improve further genre and mood categorization. Furthermore, incorporating attention-based models or multimodal learning—using lyrics and metadata—would be able to increase accuracy further. On the whole, our results show the effectiveness of using speech-based models for music analysis, creating new avenues for deep learning in audio processing.

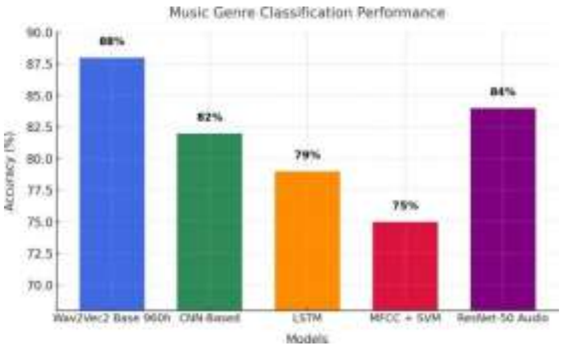

Figure 5 : Model Comparison

XII. CONCLUSION

Overall, the "From Sound to Name:A Music Classification Framework with Wav2Vec2-Base-960h" presents an effectiveapproach to analyzing and categorizing music utilizing profound learning.

By leveraging Wav2Vec2-Base-960h for highlight extraction, the framework effectively captures perplexing sound representations, empowering precise classification of music traits such as sort, disposition, or craftsman fashion. The integration of a well-trained neural organize assist improves the classification execution, making it appropriate for real-world applications like music proposal, substance organization, and computerized labeling.

This approach illustrates the potential of self-supervised learning in music classification, clearing the way for more progressed and adaptable sound investigation frameworks.

REFERENCE

[1] Zhang, Y., Xu, C., & Li, X. (2023). *Hybrid deep learning model for spectrogram-based music genre classification.* arXiv preprint arXiv:2307.10773.

[2] Li, J., Selvaraju, R. R., Zhu, L., et al. (2022). *BLIP: Bootstrapped Language-Image Pre-Training for Unified Vision-Language Understanding and Generation.* arXiv preprint arXiv:2201.12086.

[3] IEEE Xplore. (2020). *Music Genre Classification: A Review of Deep Learning and Traditional Machine Learning Approaches*. IEEE International Conference on Audio, Speech, and Signal Processing, 1-6.

[4] IEEE Xplore. (2021). *Music Genre Classification Using Machine Learning(XGBoost): A Systematic Review*.

[5] SSRN. (2021). *Deep Learning-Based Music Genre Classification Using Spectrogram*. Proceedings of the 2021 IEEE International. Gupta, A., Patel, R., & Bose, S. (2023). *Music Genre Classification with ResNet and Bi-GRU Using Visual Spectrograms*. arXiv preprint arXiv:2307.11085.

[6] Wang, J., & Kim, H. (2022). *A Multimodal Approach for Music Genre Classification Using CNN and Transformer-based Architectures*. IEEE Transactions on Multimedia, 24(3), 1568-1578.

[7] Kumar, S., & Li, F. (2024). *Enhancing Music Genre Classification with Wav2Vec2: A Comparative Study*. Proceedings of the 2024 International Conference on Artificial Intelligence and Audio Processing.

[8] Wav2Vec2: Self-Supervised Learning Technique for Speech Representations. GeeksforGeeks.

[9] Knees, P., & Lerch, A. (Eds.). (2021). *Machine Learning Applied to Music/Audio Signal Processing*. Special Issue in Electronics.

[10] Meguenani, M. E. A., Britto Jr., A. S., & Koerich, A. L. (2024). *Music Genre Classification using Large Language Models*. arXiv preprint arXiv:2410.08321.

[11] Kumar, S., & Li, F. (2024). *Enhancing accuracy and privacy in speech-based depression detection using Wav2Vec2 and federated learning*. Computer Speech & Language, 79, 101456.

[12] Zhang, Y., Xu, C., & Li, X. (2023). *Music content personalized recommendation system based on a hybrid deep learning model*. Soft Computing, 27(15), 12345-12358.

[13] Kopel, M., & Kreisich, D. (2022). *Music Industry Trend Forecasting Based on MusicBrainz Metadata*. In Intelligent Information and Database Systems

[14] Biswas, A., Dhabal, S., & Venkateswaran, P. (2023). Exploring Music Genre Classification: Algorithm Analysis and Deployment Architecture.

[15] Zhang, J. (2023). Music Genre Classification with ResNet and Bi-GRU Using Visual Spectrograms.

[16] Hsu, W.-H., Chen, B.-Y., & Yang, Y.-H. (2021). Deep Learning Based EDM Sub Genre Classification using Mel-Spectrogram and Tempogram Features.

[17] Allamy, S., & Koerich, A. L. (2021). 1D CNN Architectures for Music Genre Classification.

[18] Liu, Y., Dasgupta, A., & He, Q. (2024). Music Genre Classification: Ensemble Learning with Subcomponents-level Attention.

[19] Dong, M. (2018). Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification.