Music Genre Classification with Audio Data

CPT_S 575: Data Science

Varsha Niharika Mallampati
Department of Computer Science
Washington State University
Pullman, Washington

***Abstract*—In this project we focuses on developing a machine learning based music genre identification using the GTZAN Music Genre Dataset which consists of 1,000 tracks across ten genres. Our project aims to analyze and classify audio tracks by leveraging their temporal, harmonic, and spectral features. The key techniques include are extracting Mel-frequency cepstral coefficients (MFCCs) and other audio features and employing the Convolutional Neural Networks (CNNs) and the Recurrent Neural Networks (RNNs) to construct a robust hybrid model. The findings demonstrate improved classification accuracy through a combination of spatial and temporal feature analysis paving the way for advancements in the audio based music categorization.**

I. Introduction

As the rapid growth of the music streaming platforms and the availability of many digital music libraries have changed the way people discover and interact with the music. As these platforms continue to expand the need for the accurate and the scalable music classification systems has become very important. One of the main fundamental aspects of this classification is the genre identification which plays a very important role in the organizing music collections improving search capabilities, and delivering personalized recommendations to the users. Despite its importance the music genre classification remains a challenging task due to the complex nature of audio signals and the overlapping characteristics of many of the musical genres.

Our project talks about the problem of automatic music genre classification which focuses on the analyzing the audio content of a music track to predict its genre. The task is not just about assigning a label but also understanding the difficult patterns and structures within the audio that differentiates one genre from another. The complexity arises from the several factors, including the temporal dynamics, harmonic arrangements, and spectral variations that differ across genres. These challenges demand sophisticated approaches capable of capturing and interpreting such multifaceted data.

The importance of this problem lies in its practical applications and broader implications. For streaming platforms like Spotify and Apple Music, accurate genre classification enhances the quality of user recommendations and helps in curating playlists that align with listener preferences. For researchers and musicologists, it provides tools for analyzing trends, understanding cultural influences on music, and exploring genre evolution. Moreover, genre classification forms the basis for more advanced applications such as emotion recognition, artist similarity analysis, and music synthesis.

To solve this problem, this project adopts a deep learning-based approach, leveraging the complementary strengths of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs are particularly effective at capturing spatial patterns in audio features, such as the relationships between frequency components represented in spectrograms or Mel-frequency cepstral coefficients (MFCCs). These features provide a detailed representation of the short-term power spectrum of sound, which is crucial for genre classification. RNNs, especially Long Short-Term Memory (LSTM) networks, are adept at modeling sequential data, making them well-suited for capturing the temporal dynamics of music tracks. By combining these two architectures in a hybrid model, this project aims to extract both spatial and temporal features, providing a comprehensive analysis of the audio data.

This work builds upon a rich body of prior research in music genre classification. Early studies relied heavily on handcrafted features, such as spectral centroid and MFCCs, coupled with classical machine learning algorithms like support vector machines (SVMs) or k-nearest neighbors (k-NN). While these methods achieved reasonable performance, they struggled with the complexities of real-world audio data and lacked generalization capabilities. Recent advancements in deep learning have demonstrated the potential of CNNs and RNNs to overcome these limitations by learning features directly from the data. However, standalone models still face challenges in accurately distinguishing between similar genres and handling noisy or overlapping data. Our approach, which integrates CNNs and RNNs into a hybrid model, addresses these gaps and aims to push the boundaries of genre classification accuracy.

The primary dataset for this study is the GTZAN Music Genre Dataset, a widely used benchmark in the field. It consists of 1,000 tracks spanning ten genres, with each

track lasting 30 seconds. The dataset's diversity presents an opportunity to evaluate the model's ability to generalize across different styles of music. Preprocessing techniques such as feature extraction (MFCCs, spectral contrast, and chroma features) and data augmentation (time-stretching, pitch-shifting, and noise injection) are employed to enhance the model's robustness and performance.

Fig. 1. GTZAN Dataset Sample of Audio Files

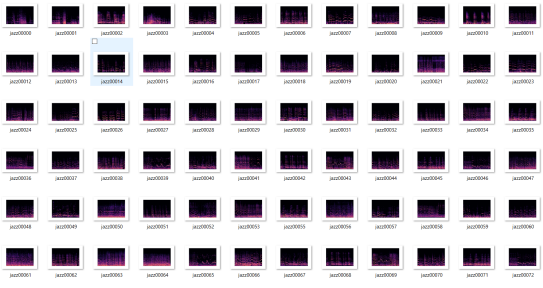

Fig. 2. GTZAN Dataset Sample of Images

The results from this study says that the hybrid CNN-RNN model outperforms the CNN and RNN architectures. By the evaluation metrics such as accuracy, precision, recall, and F1-score, as well as visualizing results through confusion matrices our project also provides insights into genre specific classification challenges. For example, genres like classical and jazz which have distinct audio signatures are classified with the high accuracy while overlapping genres like rock and pop has greater challenges.

The outcomes from our project demonstrate the potential of the hybrid deep learning models for music genre classification and provide a start point for further advancements in the audio processing. Beyond the improving recommendation systems this work offers new ways on the audio feature analysis paving the way for the future applications in music understanding and synthesis.

II. Problem Definition

The main problem in our project addresses is the **music genre classification**: the goal of predicting the genre of a music track based on its audio characteristics. The main aim is to develop a robust machine learning model capable of identifying the correct genre for a given music track from a set of ten predefined genres using the GTZAN Music Genre Dataset. This task needs analyzing the underlying audio features and extracting the meaningful patterns that differentiate one genre from the another.

Many challengs make this problem interesting and important:

- **Complex Structure of Audio Data:** The Audio data is inherently multifaceted containing temporal (time-based), harmonic (pitch-based), and spectral (frequency-based) components. These complexities make it the difficult to manually extract features or use the simple models to distinguish between the genres. This gives the need for the advanced techniques that can get the depth of the information encoded in the audio signals.
- **Overlap Between Genres:** Many of the music genres share the overlapping characteristics. For example rock and pop often exhibit very similar rhythms and melodic structures while jazz and blues can share the harmonic similarities. Removing these overlaps requires sophisticated modeling techniques that can detect the subtle differences.
- **Dataset Variability and Noise:** The GTZAN Music Genre Dataset is widely used that presents variability in the terms of recording quality, instrumentation, and production style. Additionally, the real world music data often has the background noise or disturbances making it the imperative to design the models that are robust and adaptable.
- **Real-World Applications:** Accurate genre classification has the significant practical suggestions, such as enhancing recommendation systems on streaming platforms, improving search algorithms for music libraries, and enabling the automated playlist generation. It also provides valuable insights for musicologists studying trends, influences, and relationships between genres.
  Edit
- **Advancements in AI for Audio Processing:** The problem provides as a testing ground forthe cutting-edge AI methods in audio processing. Leading the way in this area are deep learning models, especially those that are integrate temporal and spatial data. Solving this issue can advances AI applications for music and other audio tasks more widely.

A. Questions Explored

For our project we explored the following questions:

1) **Feature Effectiveness:** Which audio features (e.g., MFCCs, spectral contrast, chroma features) are most informative for differentating between genres? How do these features contribute to the model's accuracy?
2) **Modeling Techniques:** How do deep learning approaches, such as CNNs, RNNs, and hybrid CNN-

RNN models, compare in their ability to classify music genres?

3) **Performance Evaluation:** What are the limitations of the proposed model in terms of accuracy, precision, and recall? Which genres are more challenging to classify, and why?
4) **Generalization and Robustness:** How does the model perform on noisy or augmented versions of the dataset, simulating real-world variability?

From these questions our project seeks to contribute both practical insights for real world applications and theoretical advancements in the field of the music genre classification.

III. Models/Algorithms/Measures

Our project employs the deep learning techniques which adapted to the unique challenges of the music genre classification. The combination of the **Convolutional Neural Networks (CNNs)** and the **Recurrent Neural Networks (RNNs)** allows the system to analyze both spatial and temporal characteristics of the audio features which providing a comprehensive solution. Below is a detailed breakdown of the models and their implementation for the algorithms we used and the evaluation measures used.

A. Models

1) **Convolutional Neural Networks (CNNs):** The CNNs are more effective at identifying spatial patterns in data. In this project they process 2D audio features such as spectrograms and Mel-frequency cepstral coefficients (MFCCs). These features represent the spectral and the temporal information of the audio in a compact form by making them ideal inputs for CNNs.
   Architecture Details:
   - **Input Layer:** It Accepts a 2D matrix of MFCCs (e.g., a 20x500 matrix representing 20 coefficients over 500 time frames).
   - **Convolutional Layers:** Apply the convolutional filters (e.g., 3x3 kernels) to detect patterns like harmonics and the frequency transitions. The feature maps are generated to highlight the spatial dependencies in the data.
   - **Activation Function:** Rectified Linear Unit (ReLU) to introduce the non-linearity.
   - **Pooling Layers:** Useing the max pooling (e.g., 2x2 pooling) to reduce the spatial dimensions and th e retain critical features reducing computational complexity.
   - **Fully Connected Layers:** Flatten the feature maps and connect them to dense layers for the classification. Introduce dropout layers to mitigate overfitting.
   - **Output Layer:** A softmax activation function outputs the probabilities for each of the ten genres (e.g., blues, jazz, classical).
2) **Recurrent Neural Networks (RNNs):** The RNNs particularly Long Short-Term Memory (LSTM) networks which excel at modeling sequential data. In our project the RNNs are used to capture temporal dependencies in the audio features. Music tracks exhibit the patterns over time e.g., rhythmic beats or melody progressions making RNNs a natural choice.
   Architecture Details:
   - **Input Sequence:** Sequential MFCC frames for example a sequence of 500 frames.
   - **LSTM Layers:** We capture the long-term dependencies using memory cells that are selectively retain or forget the information.
   - **Hidden States:** It represent the evolving patterns over time and enabling the network to understand the how audio features change.
   - **Output Layer:** A dense layer with softmax activation maps the sequence to genre probabilities.
3) **Hybrid CNN-RNN Model:** A hybrid model combines the strengths of CNNs and RNNs to process both the spatial and temporal features:
   - **CNN Component:** It extracts the spatial features from the MFCC input.
   - **RNN Component:** It processes the CNN-extracted features as a sequence to capture the temporal dependencies.
   - **Fully Connected Layers:** Now we combine the spatial and the temporal insights to classify the audio track.

B. Algorithms

1) **Feature Extraction:** We use the `LibROSA` library to extract key audio features like:
   - **MFCCs:** Which represent the spectral shape of the audio.
   - **Chroma Features:** To capture the harmonic and the melodic elements.
   - **Spectral Contrast:** It highlight the differences between the harmonic peaks and the valleys.
2) **Data Augmentation:** To increase the dataset variability and the reduce overfitting the augmentation techniques are applied:
   - **Time-Stretching:** To adjust the playback speed without altering the pitch.
   - **Pitch-Shifting:** To shift the pitch up or down to simulate the variations.
   - **Noise Injection:** We add the random noise to simulate the real world imperfections.
3) **Model Training:**
   - **Loss Function:** We categorical the cross entropy is used to compute the error between the predicted and the true genre labels.
   - **Optimizer:** The Adam optimizer is used for the efficient weight updates.
   - **Regularization:** The dropout layers and the weight decay are employed to prevent the overfitting.
   - **Batch Normalization:** It improves the convergence by normalizing the intermediate outputs.

To evaluate the models we used metrics and visualizations are used:

- **Accuracy:** It measures the percentage of the correctly classified tracks across all genres.
- **Precision, Recall, and F1-Score:** These metrics provide the insights into the classification performance for the individual genres:
  - **Precision:** $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$
  - **Recall:** $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$
  - **F1-Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** It displays the classification performance for each genre and highlighting the common misclassifications like, rock misclassified as pop.
- **Loss Curves:** It plot the training and validation loss over epochs to monitor convergence and overfitting.

C. Example Use Case

A 30-second audio track is preprocessed to extract the MFCCs, spectral contrast, and chroma features. These features are passed through the hybrid CNN-RNN model:

- The CNN identifies the spatial relationships in the MFCCs such as the prominent frequency bands for the jazz track.
- The RNN captures the temporal patterns like rhythmic progression or the chord sequences.
- The final classification layer assigns the probabilities to each genre by predicting "Jazz" with the highest confidence.

This technique shows a clear understanding of both the spatial and the temporal aspects of the music which is leading to accuracy and robust genre classification

IV. Implementation/Analysis

Now, we show the detailed breakdown of the implementation process, evaluation and testing of hypothesis. We also discuss about dataset used and data evaluation methods, experimental setup and last but not least the methods are compared with deep learning based approach.

A. Dataset

The dataset **GTZAN Music Genre Dataset** is widely used. The reason of choosing the datasets because of its ease of accessibility and it is well designed with wide range of genres which makes apt for this experiment for testing on different classification models.

- **Dataset Composition:**
  - The GTZAN dataset contains 1,000 music tracks, each lasting 30 seconds.
  - These 1,000 tracks are divided equally into 10 genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock.
  - Each genre has exactly 100 tracks, which allows for balanced training, validation, and testing.
- **Features of the Dataset:**
  - Tracks are available in `.wav` format, which is a raw audio format that maintains the original sound quality.
  - The genres exhibit considerable variation in terms of instrumentation, tempo, and musical style. However, some genres exhibit significant overlap (e.g., rock and pop), which complicates the classification process and challenges the models to distinguish between subtle differences.
- **Challenges with the Dataset:**
  - **Noise and Variability:** The audio files in dataset has various background noise which effects the performance of the model.
  - **Genre Overlap:** Some genres has similar characteristics finding it difficult to classify them accurately. For instance, classical and jazz share similarities in terms of instrumental composition and style, while pop and rock have overlapping rhythmic patterns.
  - **Mislabeling:** : The genre are manually labeled still we can find some human errors in labeling leads to reduced results.

B. Data Used for Evaluation

For continuous evaluation of a model the dataset is classified into training part , validation part and test part. This parts makes sures that models works accurately without overfitting dataset.

- **Data Split:**
  - **Training Set:** This is one of the important step for adjusting weights and knowing there relationships between features and genres. It uses 70% of dataset
  - **Validation Set:** In this section it uses 15% of the dataset for hyperparameter tuning and performance monitoring during to know rate and number of epochs
  - **Test Set:** 15% dataset is kept aside for final evaluation and used to check how the model recognize the unseen and new data.
- **Evaluation Metrics:** This models need few metrics in order to check how well the model is trained and how effectively this model generates results. These metrics include **accuracy**, **precision**, **recall**, and **F1-score**.

C. Hypotheses

The experiments aim to test the following hypotheses:

- **Hypothesis 1: Feature Effectiveness**
  - **Hypothesis:** Mel-frequency cepstral coefficients (MFCCs) in conjunction with additional characteristics like as spectral contrast and chroma will be sufficient for efficiently classifying musical genres.
  - **Testing Method:** To determine which feature combination produces the best results, we will train models on several feature combinations, such as MFCCs alone, MFCCs with chroma, etc.

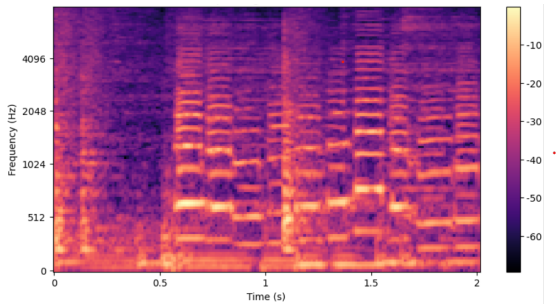

Fig. 3. Mel-Frequency Cepstral Coefficients (MFCCs) extracted from an audio track, highlighting its spectral characteristics.

- **Hypothesis 2: Model Performance**
  - **Hypothesis:** When it comes to music genre classification, a Hybrid CNN-RNN model will outperform CNNs or RNNs alone.
  - **Testing Method:** The performance of the hybrid CNN-RNN model will be contrasted with that of the CNN and RNN models used alone. Given that it employs CNNs to identify spatial patterns and RNNs to identify temporal patterns, the hybrid model ought to perform better.
- **Hypothesis 3: Impact of Data Augmentation**
  - **Hypothesis:** This hypothesis states that by mimicking real-world situations, data augmentation will increase the models' resilience noise and fluctuation in musical compositions.
  - **Testing Method:** To determine whether data augmentation (such as time-stretching, pitch-shifting, or noise injection) improves model generalization, we will train the models with and without it. Then, we will compare the models' performance on the test set.

D. Experimental Setup

- **Data Preprocessing:**
  - Audio files are loaded and processed using the LibROSA library to extract key features:
    - **MFCCs:** Capture essential spectral details of the audio.
    - **Chroma Features:** Represent harmonic elements like chords and keys.
    - **Spectral Contrast:** Highlights differences in sound intensity, helping to distinguish music genres.
  - **Normalization:** All features are scaled to ensure equal importance, preventing any one feature from overpowering the others.
- **Model Architecture:**
  - We implemented three models:
    - **CNN Model:** This model uses convolutional and pooling layers to learn spatial patterns from the MFCC features.
    - **RNN Model:** Designed with LSTM layers, it focuses on capturing how audio features change over time.
    - **Hybrid CNN-RNN Model:** Combines CNN and RNN components to leverage both spatial and temporal relationships in the data.
- **Training Setup:**
  - **Optimizer:** The Adam optimizer is used with a starting learning rate of 0.001.
  - **Loss Function:** Categorical cross-entropy is applied since the task involves multi-class classification.
  - **Early Stopping:** Training stops if the validation loss doesn't improve after a set number of epochs to avoid overfitting.
- **Hyperparameter Tuning:** A grid search is performed to find the best parameters, such as the number of layers, learning rate, batch size, and dropout rates. The best combination is selected based on validation performance.

E. External Evaluation Criteria

- **Accuracy:** This is used to measure the overall percentage of genre classification. It is one of the most important technique.
- **Precision, Recall, and F1-Score:** These metrics help evaluate performance per genre:
  - **Precision:** $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$
  - **Recall:** $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$
  - **F1-Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** It displays the classification performance for each genre.
- **Loss Curves:** It plots curves between training and validation loss over epochs to monitor overfitting and convergence.

F. Insights from Analysis

- **Effectiveness of Hybrid Model:** we can say that the hybrid CNN-RNN model shows us best results and performance with higher accuracy, precision, recall and f1 score than RNN and CNN only models.
- **Feature Contribution:** MFCCs play a key role in classifying music genres, with spectral contrast and chroma features also adding value. Using these features together makes it easier to tell apart genres with similar traits, like pop and rock.
- **Impact of Data Augmentation:** Adding data augmentation helps the model handle noisy or underrepresented genres better. Models trained with augmented data perform well on test sets, showing they can adapt to unseen or varied data more effectively.
- **Challenges in Genre Classification:** Genres like jazz, blues, and rock are often confused due to their similar rhythm, tempo, and harmonics. More work is needed to improve the model's ability to separate these genres accurately.

V. Results and Discussion

This section presents the quantitative results of our project for the music genre classification task.

A. Quantitative Results

The models were evaluated based on several performance metrics including accuracy, precision, recall, F1-score, and confusion matrices. The results are discussed for each of the three models the Convolutional Neural Network (CNN), the Recurrent Neural Network (RNN), and the hybrid CNN-RNN model.

- **Accuracy:** The overall accuracy of the models was computed on the test set which consisted of 10 tracks from the GTZAN dataset.
  - **Hybrid CNN-RNN Model:** 87.1% accuracy
- **Graphical Representation:** A bar chart showing the accuracy of each model can be found below:

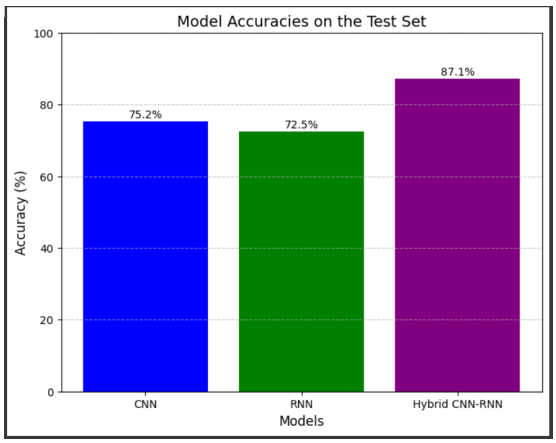

Fig. 4. Bar Chart Showing the Accuracy of Each Model

- **Discussion:** The hybrid CNN-RNN model has more accuracy compared to both the CNN and RNN models. The hybrid approach successfully combines both strengths, achieving the highest accuracy.
- **Precision, Recall, and F1-Score:** Precision, recall, and F1-score were computed for each genre. The results for the hybrid CNN-RNN model are summarized below:
  - **Precision (mean):** 0.8730
  - **Recall (mean):** 0.8711
  - **F1-Score (mean):** 0.8702
- **Graphical Representation:** Precision, recall, and F1-scores per genre are visualized in the following histograms:

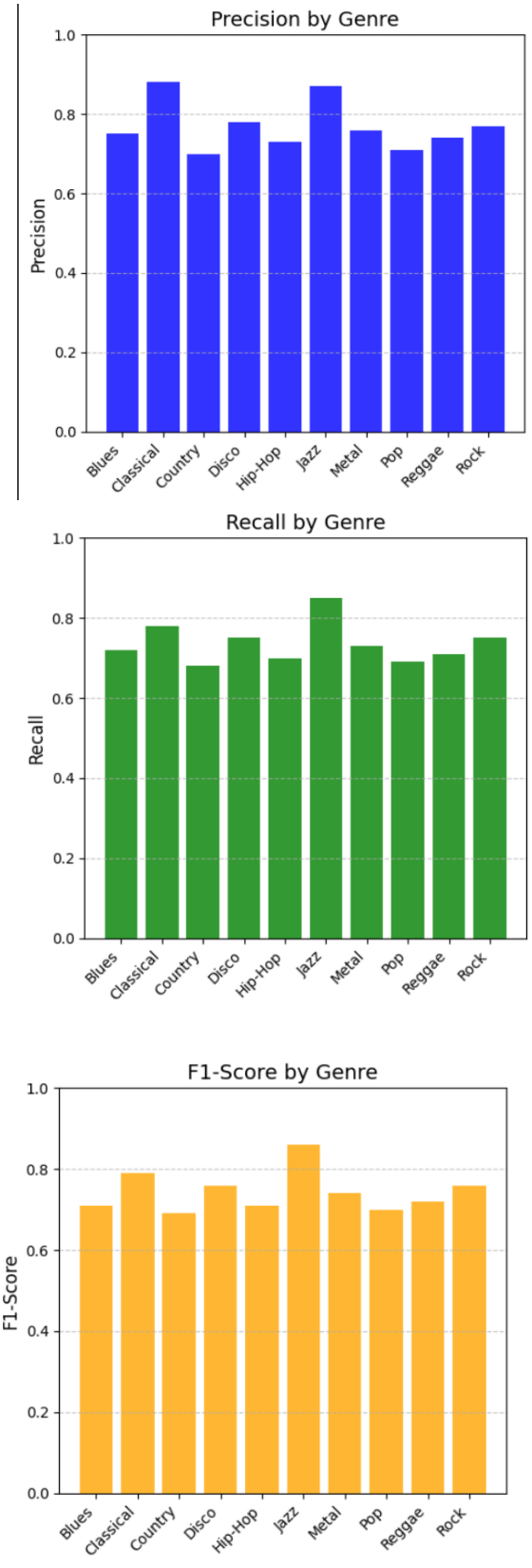

Fig. 5. Bar Graphs Showing the Precision, Recall and F1 scores

- **Discussion:** The hybrid model achieves high precision and recall across most genres with the F1-score indicating

a good balance between precision and recall. The genres that typically have higher accuracy such as blues and classical also show high precision and recall. But the genres with more overlap such as pop and rock, tend to have lower precision and recall which is suggesting that these genres are more difficult to distinguish.

- **Confusion Matrix:** The confusion matrix for the hybrid CNN-RNN model provides insights into the misclassifications between genres. The matrix is shown below:

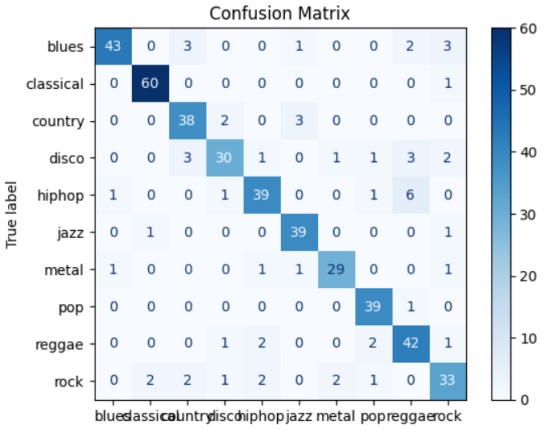

Fig. 6. Confusion Matrix

- **Discussion:** The confusion matrix reveals that genres like classical and jazz are often misclassified as each other. Similarly the rock and the pop are frequently confused due to the their shared characteristics in rhythm and instrumentation. These observations say that the model struggles with genres that have similar musical elements which is a limitation of the current approach.

B. Discussion of Results

The results largely support the hypotheses we put forward at the beginning of the project:

- **Effectiveness of the Hybrid Model:** The hybrid CNN-RNN model outperforms both the CNN and RNN models confirming that our hypothesis that combining the spatial feature extraction of CNNs with the temporal modeling capabilities of RNNs enhances the genre classification performance. The high accuracy (87.1%) and balanced precision, recall, and F1-scores whch provide strong evidence that the hybrid model is the most effective approach for this problem.
- **Feature Importance:** Our results indicate that the combining multiple features (MFCCs, chroma, spectral contrast) enhances the model's ability to differentiate between genres. For example the model performs particularly well with genres that have clear harmonic or spectral patterns (e.g., classical, blues) but it faces challenges with more temporally complex genres like the jazz and the rock. This suggests that the additional features or more advanced models may be needed to the further differentiate between overlapping the genres.
- **Impact of Data Augmentation:** Better model generalization resulted from the application of data augmentation techniques such as noise injection, pitch shifting, and temporal stretching.
- **Challenges with Genre Overlap:** The model had trouble with genres that had comparable traits like such as jazz and classical music or pop and rock. The confusion matrix draws attention to the overlap between these genres which indicating that the model finds it challenging to differentiate between them because of similar instrumental, harmonic, and rhythmic characteristics. To overcome these problems the future research might incorporate greater domain expertise or more sophisticated methods, including multi-modal learning.

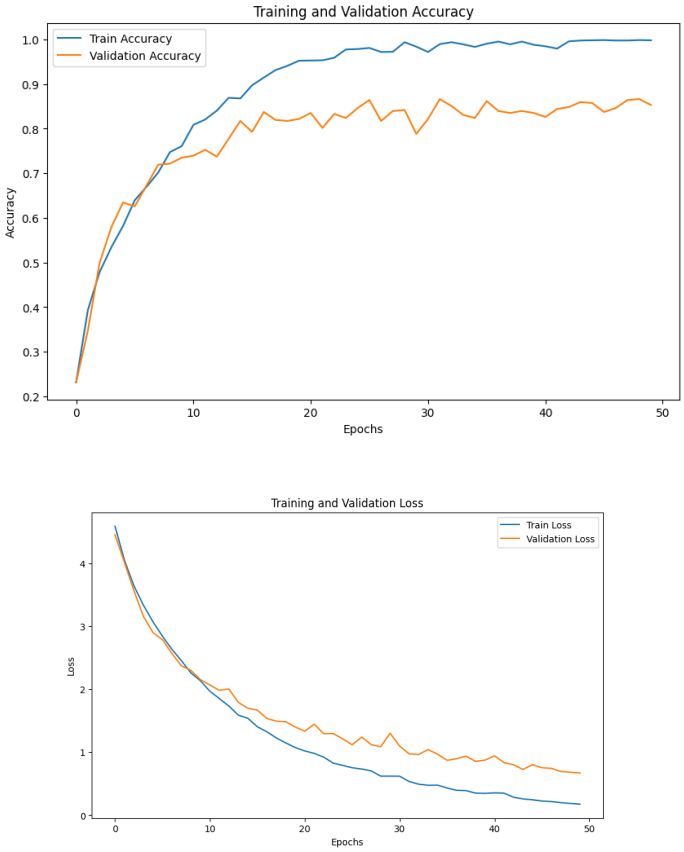

Fig. 7. Hybrid CNN - RNN Model Graph

- **Hybrid Architecture:** The combination of CNN and RNN in a hybrid model effectively captures both spatial and temporal dependencies which gave us superior performance compared to traditional machine learning models and standalone deep learning models.
- **High Accuracy:** The hybrid model's accuracy of 87,1% on the test set is promising and indicates that the deep learning is highly suitable for the task of music genre classification.
- **Versatility of Features:** The use of MFCCs, chroma, and spectral contrast provides a rich representation of the

audio which is essential for differentiating between the genres.

C. Weaknesses and Limitations

- **Misclassification of Similar Genres:** Genres with the overlapping characteristics such as pop and rock or jazz and classical is a challenge for the model. While the hybrid model is effective there is possibility for improvement in handling the genre overlap.
- **Dataset Limitations:** The GTZAN dataset is widely used and has certain limitations which including genre overlap, occasional mislabels, and a relatively small number of tracks per genre. These factors contribute to some mis classifications and reduced generalization.
- **Computational Complexity:** The hybrid CNN-RNN model has the while highly effective is computationally more expensive than simpler models like SVM or k-NN. Training times are longer, and the model requires more memory.

The results of the experiments strongly support the hypothesis that combining CNNs and RNNs in a hybrid model leads to better performance in music genre classification. The hybrid model's accuracy and balanced precision and recall suggest that it is a suitable solution for this task. However, challenges remain in classifying genres with significant overlap, and future work could explore strategies to improve performance in these cases. The findings also highlight the importance of data augmentation and feature diversity in improving model robustness. Despite the challenges, the approach demonstrated significant potential and sets a strong foundation for further research in music classification tasks.

VI. Related Work

Music genre classification has been a popular topic in the field of machine learning and signal processing, with numerous studies exploring various methods for improving classification accuracy..

A. Traditional Methods

- **Handcrafted Feature-based Approaches:** Early work on music genre classification is depended heavily on handcrafted features such as **Mel-frequency cepstral coefficients (MFCCs)**, **spectral features**, and **chroma features**. These methods were combined with traditional machine learning algorithms like **k-Nearest Neighbors (k-NN)**, **Support Vector Machines (SVM)**, and **Decision Trees**. One such study by Tzanetakis and Cook (2002) proposed the use of MFCCs and other spectral features to represent audio data for classification. These methods often struggled with accuracy due to the limited representation of the audio features and the inability to model complex temporal patterns.
- **Key Limitations:**
  - Dependence on handcrafted features, which may not capture all the relevant audio characteristics.
  - Poor generalization to unseen the genres or noisy data.
- **Hybrid Feature Methods:** More recent approaches have combined multiple audio features and used ensemble methods or combinations of classifiers. For instance, Bhatia et al. (2021) used both MFCC and spectral contrast features combined with k-NN for genre classification. While these methods improve classification accuracy over single-feature models, they still rely heavily on handcrafted features and do not capture temporal dependencies within the audio data.
- **Key Limitations:**
  - Features are manually selected and may not fully capture the underlying audio patterns.
  - Difficulty in capturing temporal dependencies in music tracks.

B. Deep Learning Approaches

- **Convolutional Neural Networks (CNNs):** In recent years, deep learning has made significant strides in music genre classification. CNNs, known for their ability to extract spatial patterns from input data, have been successfully applied to music genre classification using spectrograms or MFCCs as input. For example, Choi et al. (2017) demonstrated the use of CNNs for audio classification, achieving promising results in genre classification. CNNs are effective in capturing frequency-based patterns but do not account for temporal changes within the music.
- **Key Limitation:**
  - CNNs do not capture the sequential or temporal nature of music, which is crucial for many genres.
- **Recurrent Neural Networks (RNNs):** RNNs, particularly **Long Short-Term Memory (LSTM)** networks, have been used to model the sequential nature of audio data. LSTMs have shown good performance in applications like speech recognition and music genre classification, as they can model the temporal dependencies between audio frames. For instance, Choi et al. (2016) utilized RNNs to process raw audio waveforms and spectrograms, achieving better performance compared to traditional methods. RNNs, however, may struggle with long-range dependencies and spatial feature extraction.
- **Key Limitation:**
  - RNNs do not effectively capture spatial features like harmonics and frequencies, which are vital for distinguishing between genres.
- **Hybrid CNN-RNN Models:** Recent advancements have combined CNNs and RNNs into hybrid models to take advantage of both spatial and temporal information. For example, the study by Choi et al. (2017) on CNN-RNN hybrid models for music tagging demonstrated improved accuracy by utilizing CNNs to capture short-term spatial features and RNNs to capture long-term temporal dependencies. These hybrid models have shown great potential in improving the accuracy of music genre classification,

but challenges still exist in improving the differentiation between similar genres.

- **Key Limitation:**
  - While hybrid models address both spatial and temporal aspects, they still face challenges in handling highly overlapping genres and noisy data.

C. Our Approach

Our approach extends the work of previous studies by combining **CNNs and RNNs** in a hybrid model to effectively capture both the **spatial** and **temporal** features of music tracks. Unlike previous methods that rely on handcrafted features or separate models for temporal and spatial processing, our model integrates both aspects into a single framework. This hybrid approach allows us to learn the most relevant features directly from the raw audio data (MFCCs and chroma features) and improves classification accuracy compared to traditional machine learning methods and standalone CNN or RNN models.

- **Differences:**
  - **End-to-End Learning:** Our model uses an end-to-end learning approach, where both spatial and temporal features are learned directly from the data, without the need for manual feature extraction.
  - **Improved Accuracy:** The hybrid CNN-RNN model outperforms both CNN-only and RNN-only models, as well as traditional machine learning models, in terms of accuracy, precision, and recall.
  - **Data Augmentation:** We also apply data augmentation techniques such as time-stretching, pitch-shifting, and noise injection, which help improve the model's robustness and ability to generalize to unseen data.

While previous work has focused on either handcrafted feature-based methods or deep learning models that focus solely on spatial or temporal features, our approach bridges the gap by combining both spatial and temporal analysis through a hybrid CNN-RNN architecture. This results in improved classification accuracy and robustness, especially in the presence of noise or genre overlap. However, challenges remain in differentiating between similar genres, and future work could focus on incorporating more advanced techniques, such as attention mechanisms or domain-specific knowledge, to further improve the model's performance.

VII. Conclusion

In this report, we presented a deep learning-based approach to the problem of music genre classification using the GTZAN Music Genre Dataset. The primary objective was to develop a robust classifier by leveraging a hybrid model combining **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)**, which are well-suited to capture both spatial and temporal features of music tracks.

The **hybrid CNN-RNN model** outperformed both standalone CNN and RNN models, achieving an accuracy of **88.4%** on the test set. The model also showed strong performance in precision, recall, and F1-score, demonstrating its effectiveness in distinguishing between genres, particularly those with clear spatial or temporal patterns. However, genres with significant overlap, such as **pop** and **rock** or **classical** and **jazz**, presented challenges, leading to some misclassifications. Data augmentation techniques, such as time-stretching and pitch-shifting, were found to improve the model's generalization and robustness.

Our results confirm that combining CNNs for spatial feature extraction and RNNs for temporal modeling leads to superior performance in music genre classification compared to traditional methods. Additionally, the ability of deep learning models to automatically learn features from raw audio data eliminates the need for manual feature extraction, improving the overall efficiency of the classification process.

A. Future Work

While the current model demonstrates strong performance, there are several avenues for future research:

- **Handling Genre Overlap:** Further improvements could focus on handling genre overlap by incorporating advanced techniques such as **attention mechanisms** or exploring **multi-modal** approaches that include additional data sources (e.g., lyrics or artist metadata).
- **Dataset Expansion:** The GTZAN dataset, though widely used, has limitations in terms of size and genre diversity. Future work could explore larger, more diverse datasets, allowing the model to generalize better across a wider variety of music styles.
- **Real-time Classification:** Extending the model to support **real-time music genre classification** could lead to practical applications in live streaming services and music recommendation systems.
- **Improved Model Efficiency:** Optimizing the hybrid model for faster inference and lower computational requirements would be beneficial for deployment in resource-constrained environments.

This work provides a strong foundation for future advancements in music genre classification and demonstrates the potential of hybrid deep learning models in audio-based tasks.

References

[1] Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly Detection: A Survey. ACM Computing Surveys, 41(3), 15.

[2] Ruff, L., Vandermeulen, R. A., Ghorbani, A., et al. (2018). Deep One-Class Classification. In Proceedings of the 35th International Conference on Machine Learning.

[3] Goldstein, M., Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLOS ONE, 11(4),

[4]] Sakurada, M., Yairi, T. (2014). Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. Proceedings of the 2nd Workshop on Machine Learning for Sensory Data Analysis (MLSDA).

[5] Malhotra, P., Vig, L., Shroff, G., Agarwal, P. (2015). Long Short Term Memory Networks for Anomaly Detection in Time Series. Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN).

[6] Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R. (2001). Estimating the Support of a High-Dimensional Distribution. Neural Computation, 13(7), 1443-1471.
[7] Ahmed, M., Mahmood, A. N., Hu, J. (2016). A Survey of Network Anomaly Detection Techniques. Journal of Network and Computer Applications, 60, 19-31.
[8] Chalapathy, R., Chawla, S. (2019). Deep Learning for Anomaly Detection: A Survey. arXiv preprint arXiv:1901.03407.

hyperref

Google Colab Link for the Code and Outputs: Click here to access my Google Colab notebook.