SELECTED TOPICS OF MUSIC AND ACOUSTIC ENGINEERING

Project Report: Music Genre Classification

Politecnico di Milano
Music and Acoustic Engineering
Master Degree of Science

Written by:
Ipek Ceren BAYRAM
Sebastian GOMEZ MARTINEZ
Matéo VITALONE

Academic year: 2024-2025
Supervisors: Julio Jose CARABIAS ORTIS

Contents

1 Project Motivations

In the context of the Selected Topics in Music and Acoustic Engineering course, the following project involves developing a machine learning pipeline that automates music genre classification. Using the FMA-small dataset, a curated database of 8,000 30-second audio clips uniformly distributed across eight musical genres—the overarching goal is to create an algorithm that predicts the genre label of a given audio recording with a sufficiently accurate model.

Music genre classification remains an unresolved issue due to the fact that audio signals are extremely complex and multi-dimensional. In order to address this, our work investigates the applicability and effectiveness of various audio features that capture various musical dimensions such as rhythm (e.g., tempo, beat patterns), harmony (e.g., chroma-based features), and timbre (e.g., MFCCs, spectral properties). We investigate the way each of these features, individually and in combination, supports the discriminative ability of the classification system.

During the project, we experimented with a range of feature extraction techniques and machine learning models—from classical classifiers to deep learning approaches—and concentrated specifically on reproducibility and quantitative analysis. The performance of the ultimate system is evaluated using typical multi-class classification metrics, providing information about both the strengths and weaknesses of the adopted methodology.

Prelude

During the process of extracting the dataset, we repeatedly encountered an error for three particular data files: `108925.mp3`, `099134.mp3`, and `133297.mp3`. Opening them individually revealed that they were corrupted, which is why they were not being used.

2 Methodology Used

2.1 Dataset Acquisition and Preprocessing

The dataset employed in this project is the **FMA-small** (Free Music Archive) collection, which contains 8,000 audio tracks, each 30 seconds long, evenly distributed across 8 music genres.

To retrieve the metadata, we downloaded and extracted the official metadata file `tracks.csv`, which contains detailed information about each audio file, including its associated genre. From this file, we filtered the entries belonging exclusively to the `small` subset.

Using the track IDs from the metadata, we located the corresponding MP3 file paths and retained only those that matched valid entries and had a known `genre_top` label.

And to reconstruct the database, we built a `DataFrame` that stores both the file path and the genre label for each valid track. This structure allowed easy integration with subsequent processing steps.

2.2 Feature Extraction

2.2.1 Handcrafted Feature Extraction Pipeline

To represent each audio track numerically, we implemented the custom function `extract_features(path)` using the Librosa library. This function computes a variety of audio descriptors grouped into three musical dimensions:

2.2.2 Extended Feature Extraction

As a complementary method to the core feature extraction pipeline, an expanded set of handcrafted audio features was also computed in order to enrich the input representation provided to the traditional classifiers. These additional features aimed to capture a wider range of perceptual and signal-level properties of the audio tracks, thereby providing the models with a more comprehensive description of the musical content.

The extended feature set included the following descriptors. First, 13 Mel-frequency cepstral coefficients (MFCCs) were computed for each audio track, with both the mean and standard deviation of each coefficient calculated across time. This captured the spectral envelope and timbral characteristics of the signal in a compact and perceptually motivated manner.

Second, chroma features were extracted to represent the energy distribution across the 12 pitch classes (C, C #, D, etc.), again using both the mean and standard deviation to summarize temporal variations. Spectral contrast was also included, computed across 7 frequency bands; both the mean and standard deviation of each band's contrast were used to characterize the richness and balance between spectral peaks and valleys in the audio.

In addition to these core features, several spectral and temporal descriptors were added. Spectral centroid, spectral rolloff, and Root Mean Square (RMS) energy were each computed and summarized using their mean and standard deviation, providing further insight into the brightness, bandwidth, and energy profile of each track. The zero-crossing rate, which reflects the rate of sign changes in the time-domain signal and is often related to noisiness or percussive content, was also computed with its mean and standard deviation. Finally, the global tempo of each track was estimated in beats per minute (BPM), providing an overall rhythmic descriptor.

All features in this extended set were normalized using the StandardScaler, ensuring that each feature had zero mean and unit variance, thereby making them suitable for use with traditional machine learning algorithms. This normalization step was performed in a manner consistent with the initial feature pipeline to maintain methodological coherence across the experiments.

2.2.3 Rhythmic Features

- **Tempo (BPM):** Estimated from the onset strength envelope to capture the perceived speed of the music.
- **Beat Histogram:** A histogram of onset strength values (8 bins), providing a coarse rhythmic intensity profile.
- **Root Mean Square Energy (RMSE):** Used to quantify the average power or loudness of the signal. It is calculated as the square root of the mean of the squares of the signal's values.

2.2.4 Harmonic Features

- **Chroma Features:** Represent the energy distribution across the 12 pitch classes (e.g., C, C#, D), with both mean and standard deviation calculated over time.
- **Tonnetz (tonal centroid features):** Conceptual lattice or graph used to represent relationships between musical pitches, particularly in terms of harmonic closeness.

2.2.5 Timbral Features

- **MFCCs (Mel-frequency Cepstral Coefficients):** 13 coefficients describing spectral envelope characteristics, along with their mean and standard deviation.
- **Spectral Contrast:** Describes timbral richness across frequency bands, computed via mean and standard deviation.
- **Zero-Crossing Rate:** How frequently the signal waveform crosses the zero amplitude axis (changes sign) per unit of time or per frame.
- **Spectral bandwidth:** Measures the spread of the frequency spectrum around the spectral centroid. It essentially quantifies how wide a range of frequencies the signal occupies.

- **Spectral centroid:** The weighted average of the frequencies present in a signal, with the magnitudes of those frequencies serving as the weights.

These features were selected to capture key perceptual dimensions that are known to influence genre categorization.

2.3 Dataset Construction and Preparation

2.3.1 Dataset Construction

Using the function `prepare_X_y(df)`, we transformed the dataset into:

- **X:** A feature matrix where each row corresponds to an audio sample and each column to an extracted feature.
- **y:** An array of genre labels, encoded numerically using `LabelEncoder` from scikit-learn.

2.3.2 Data Splitting Strategy

We took 70% of the data to train the model. This would provide a sufficient number of samples for the model to learn. Additionally, the remaining 30% of samples would be enough to observe the real performance of the model. This splitting is suitable for our feature-based models. However, for the neural network based models, 30% of the test set is divided into two, where they were used as validation and test. The validation set is also required for those models to tune the hyperparameters. Finally, to ensure robust and unbiased evaluation, the dataset was split using stratified sampling, preserving the relative proportions of each genre.

Dataset Split	**Proportion**
Training set	70%
Test set	30%

Table 1: Dataset split proportions used in SVM and Random Forest Classifier

Dataset Split	**Proportion**
Training set	70%
Test set	15%
Validation set	15%

Table 2: Dataset split proportions used in MLP and CNN

This strategy ensures that genre distributions are consistent across splits and reduces the risk of overfitting.

2.3.3 Feature Scaling

Before training the models, all feature values were standardized using the `StandardScaler` transformation. This preprocessing step ensured that each feature had zero mean and unit variance, a condition that is particularly important for distance-based models such as Support Vector Machines (SVM), where differences in feature scales can significantly affect the computation of distances and the resulting decision boundaries.

To avoid introducing bias or data leakage into the modeling pipeline, the scaler was fit exclusively on the training set. This means that the mean and variance used for scaling were computed solely from the training data. Once fitted, the same transformation parameters were consistently applied to both the validation and test sets. By adhering to this practice, we ensured that information from the validation and test sets did not inadvertently influence the model during training, thereby preserving the integrity of the evaluation process.

2.4 Classification Models

2.4.1 Random Forest Classifier

A Random Forest classifier was trained as a baseline model for music genre classification. The ensemble was configured with 300 decision trees (`n_estimators=300`), and a fixed random seed (`random_state=42`) was set to ensure reproducibility across experimental runs. Model training was performed using the scaled handcrafted feature matrix, and evaluation was conducted on a held-out test set to assess generalization performance.

2.4.2 Support Vector Machine (SVM) with RBF Kernel

A second classifier was trained using a Support Vector Machine (SVM) with a radial basis function (RBF) kernel. The model was configured with the default setting `gamma='scale'`, which allows automatic adjustment of the kernel's influence. Importantly, the SVM was trained on exactly the same scaled handcrafted features used for the Random Forest, ensuring that both models operated on an identical input space.

2.4.3 Multilayer Perceptron (MLP) on Handcrafted Features

In addition to the ensemble-based and margin-based classifiers previously described, we trained a Multilayer Perceptron (MLP) neural network on the same set of handcrafted audio features that were used as input for both the Random Forest and SVM models. The MLP provided an additional deep learning baseline for evaluating the predictive power of these engineered features.

Unlike Random Forest and SVM, which are non-neural models, the MLP architecture consisted of a fully connected feedforward neural network. Training was performed using one-hot encoded labels, generated via the `to_categorical` function, which allowed the network to optimize predictions using a softmax output layer. Moreover, the MLP training pipeline incorporate callbacks to dynamically control the training process and improve the model's generalization.

The architecture of the MLP included four sequential dense layers. The first dense layer consisted of 128 units with ReLU activation, followed by a BatchNormalization layer to stabilize training and a Dropout layer with a rate of 0.5 to mitigate overfitting. The second dense layer contained 64 units, also with ReLU activation, followed by BatchNormalization and Dropout with a rate of 0.4. The third dense layer included 32 units, again with ReLU activation, BatchNormalization, and Dropout at a rate of 0.3. Finally, the network concluded with a dense output layer consisting of 8 units with softmax activation, corresponding to the eight target music genres.

Training was conducted using the **Adam optimizer**, which provided adaptive learning rate adjustment during optimization. The loss function used was categorical cross-entropy, suitable for multi-class classification tasks with one-hot encoded labels. The network was trained with a batch size of 200 for up to 100 epochs. Two callbacks were employed to enhance training efficiency and model robustness: `EarlyStopping`, configured with a patience of 10 epochs, halted training if the validation loss did not improve within this window and automatically restored the best-performing weights; and `ReduceLROnPlateau`, which reduced the learning rate by a factor of 0.5 when the validation loss plateaued, allowing the model to converge more effectively.

2.4.4 One Dimensional - CNN

In addition to feature-based classifiers, we implemented a 1D Convolutional Neural Network (CNN) to process the raw features of the audio tracks.

CNN architecture: The 1D CNN consisted of four layers with one increasing filter count (64 and 128) and max-pooling. After flattening the feature maps, two dense layers, one with 128 units and ReLU activation was used, followed by a dropout layer with a rate of 0.5. The final output layer was a softmax Dense classifier with 8 units, corresponding to the eight music genres.

Training configuration:
The network was trained using the Adam optimizer and sparse categorical cross-entropy loss. Training was performed with a batch size of 32 for up to 150 epochs. An `EarlyStopping` callback was employed with a patience of 15 epochs, ensuring that training would halt early if no improvement was observed in the validation set, with the best model weights automatically restored.

2.5 Evaluation Metrics and Visualization

To comprehensively assess the performance of the classifiers developed in this study, a consistent set of evaluation metrics was employed across all models. The primary metric used was **accuracy**, which measures the overall correctness of the model's predictions by calculating the proportion of correctly classified samples out of the total number of predictions.

In addition to overall accuracy, more granular performance indicators were computed to provide a deeper understanding of the models' behavior across different genres. **Precision** was calculated for each genre as the proportion of correct predictions among all predictions assigned to that genre, thus quantifying the model's ability to avoid false positives. **Recall** was also computed per genre, representing the proportion of actual genre instances that were correctly predicted, and thereby indicating the model's sensitivity to each genre class. To balance precision and recall, the **F1-score** was calculated for each genre as the harmonic mean of these two metrics, offering a single value that reflects the trade-off between false positives and false negatives.

Furthermore, the classification results were visualized using a **confusion matrix**, which provided detailed insight into which genres were most frequently misclassified and highlighted common confusions between genre pairs. This visualization complemented the quantitative metrics by enabling an intuitive and interpretable understanding of the classification performance.

Overall, this combination of metrics was well-suited for the multi-class classification task addressed in this project, enabling a nuanced and rigorous analysis of each model's ability to perform genre classification on the FMA-small dataset.

2.6 Genre-Specific Visualization and Interpretation

To better understand the audio characteristics of each genre, we conducted a quick visual inspection by plotting the Mel Spectrogram of single audio tracks across all genres present in the dataset (Fig. 2.1).

- **Hip-Hop:** The energy is concentrated in the lower frequencies, highlighting elements like kicks and bass lines. A clear rhythmic structure is also visible in the time domain.

- **Pop:** Energy is concentrated in the low-to-mid frequency range. There is a clear repetition over time, indicating a structured and catchy rhythmic pattern typical of pop music.

- **Folk:** Energy is distributed relatively evenly across the frequency spectrum, likely due to the presence of acoustic instruments and prominent vocals. The structure appears organic, with no strongly emphasized rhythmic or tonal components.

- **Experimental:** There is a lack of organization in both time and frequency domains. No dominant rhythmic or tonal patterns are present, reflecting the genre's abstract nature.

- **Rock:** The energy is distributed across the entire frequency range without a specific concentration in one area. In the time domain, it lacks a consistently dominant rhythmic pattern but still maintains an energetic and dense texture.

- **International:** This genre shows a wide frequency distribution. Some parts display periodicity, though it is not consistently maintained throughout the track.

- **Electronic:** The spectrogram reveals recurring vertical lines across a wide frequency range, indicating a strong and consistent rhythmic structure.

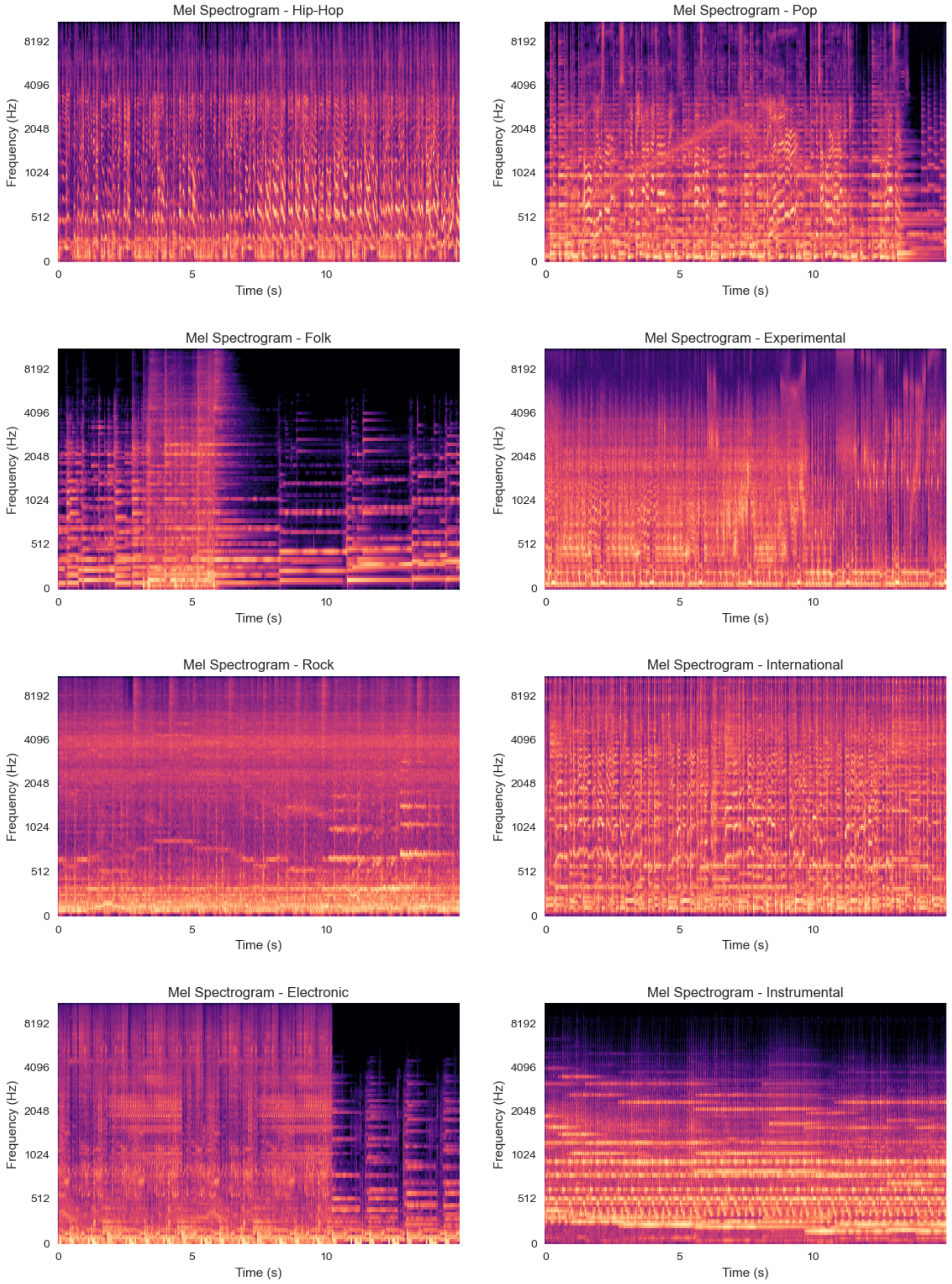


Figure 2.1: Mel spectrograms from a single audio track for each genre

- **Instrumental:** The spectrogram shows a high energy concentration in the low-frequency range. The temporal progression is smooth and less abrupt, indicating gradual transitions.

Although some music genres show characteristic patterns, distinguishing them based solely on visuals of audio representations remains challenging. The objective of this study is, therefore to extract relevant audio features and develop effective classification models capable of automatically recognizing musical genres from audio signals.

2.7 Reproducibility and Data Persistence

Ensuring the reproducibility of machine learning experiments was a key consideration throughout this project. To this end, several methodological choices and technical procedures were adopted to guarantee that results could be reliably reproduced and that training workflows could be efficiently optimized.

First, all random seeds controlling data splits, model initialization, and training procedures were fixed. This deterministic setup ensured that data partitions and model behaviors remained consistent across multiple experimental runs. Additionally, the same stratified data splitting strategy was applied systematically: 70% of the data was allocated to the training set, 15% to the validation set, and the remaining 15% to the test set. This approach preserved the genre distribution across splits, thereby reducing variance due to sampling artifacts and ensuring fair comparisons between models. All evaluation metrics were computed consistently across experiments, following the same protocol for each classifier and neural network, further reinforcing methodological rigor.

To optimize training workflows and enable efficient reuse of intermediate results, a comprehensive data persistence strategy was implemented. Extracted features (`X`) and their corresponding genre labels (`y`) were saved to disk using `numpy`'s `save` and `load` functions. This allowed subsequent training sessions to bypass the computationally intensive extraction stage. Furthermore, the `LabelEncoder` object used to transform genre labels into integer codes was serialized with `pickle`, ensuring consistent decoding of genre labels across all stages of model development and evaluation.

In addition, feature matrices corresponding to the training, validation, and test splits were stored as separate files, enabling repeated model training and testing to proceed from a well-defined and stable data state. This practice significantly reduced total computation time and allowed for rapid experimentation, model tuning, and evaluation without the need to re-extract features or reprocess the entire dataset.

Collectively, these measures ensured both the reproducibility of experimental outcomes and the robustness of the machine learning pipeline, while facilitating efficient iteration and experimentation throughout the project.

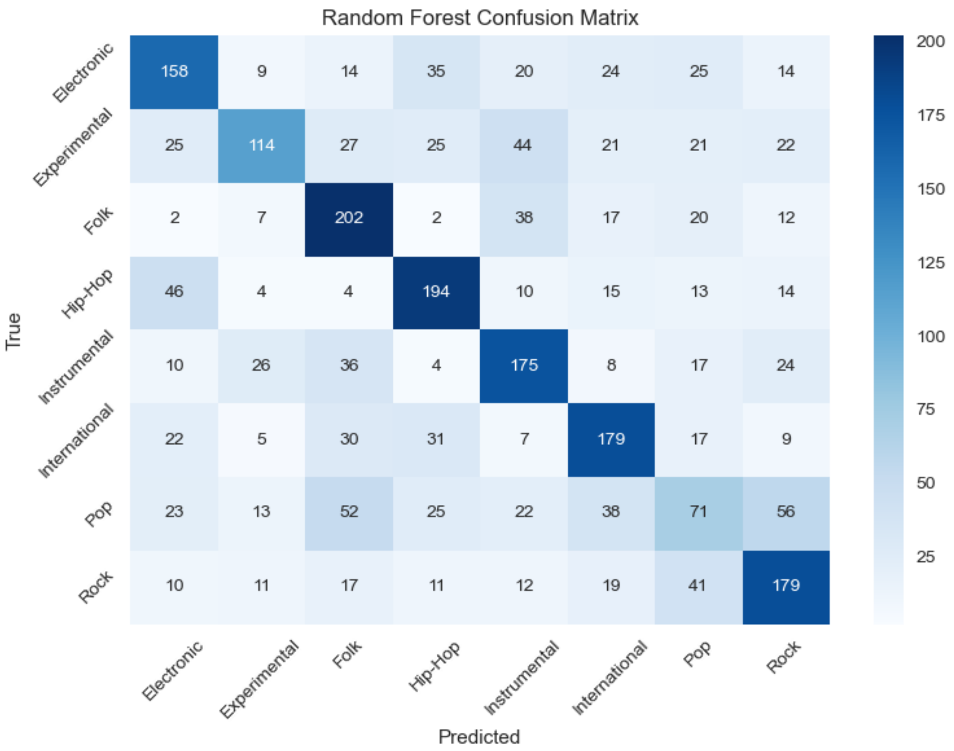
3 Evaluation of the model

3.1 Random Forest Classifier

The Random Forest Classifier was implemented as explained in part 2.4.1. We first plotted the Confusion Matrix (Fig. 3.1).

From this figure, we can see some good diagonal dominance, as it correctly classifies a good proportion of examples from most genres (deeper blue diagonal cells).

The genres Folk (202), Hip-Hop (194), International (179), and Rock (179) are classified with the highest accuracy, as indicated by the strong diagonal values and relatively low off-diagonal confusion. The genres Electronic (158) and Instrumental (175) also exhibit good results, but with more confusion with other classes.



	Electronic	Experimental	Folk	Hip-Hop	Instrumental	International	Pop	Rock
Electronic	158	9	14	35	20	24	25	14
Experimental	25	114	27	25	44	21	21	22
Folk	2	7	202	2	38	17	20	12
Hip-Hop	46	4	4	194	10	15	13	14
Instrumental	10	26	36	4	175	8	17	24
International	22	5	30	31	7	179	17	9
Pop	23	13	52	25	22	38	71	56
Rock	10	11	17	11	12	19	41	179

Figure 3.1: Confusion Matrix of the Random Forest Classifier.

For Electronic, 46 Electronic tracks were misclassified as Hip-Hop. This suggests some overlap in the features used (e.g., rhythm or spectral content).

Experimental has the lowest performance (114 correct out of 300), with substantial confusion, especially toward Instrumental (44) and other classes. This is expected as Experimental music is, by nature, less stereotypical and may have overlapping characteristics with other music genres.

Pop is the most heterogeneous in classification, as it's correctly classified 71 times. It presents confusion with Folk (52), International (38), and Rock (56), suggesting that Pop as a label is less clearly defined in this feature space, and the genres might share similar production features.

The classification report of the model execution is shown in Table 3.

Train Score: `0.9998`
Test Score: `0.5392`

Genre	Precision	Recall	F1-Score	Support
Electronic	0.54	0.53	0.54	299
Experimental	0.61	0.37	0.46	299
Folk	0.55	0.66	0.60	300
Hip-Hop	0.59	0.68	0.63	300
Instrumental	0.56	0.61	0.58	300
International	0.57	0.60	0.59	300
Pop	0.30	0.24	0.27	300
Rock	0.56	0.62	0.59	300
Accuracy			**0.54**	2398
Macro Avg	0.53	0.54	0.53	2398
Weighted Avg	0.53	0.54	0.53	2398

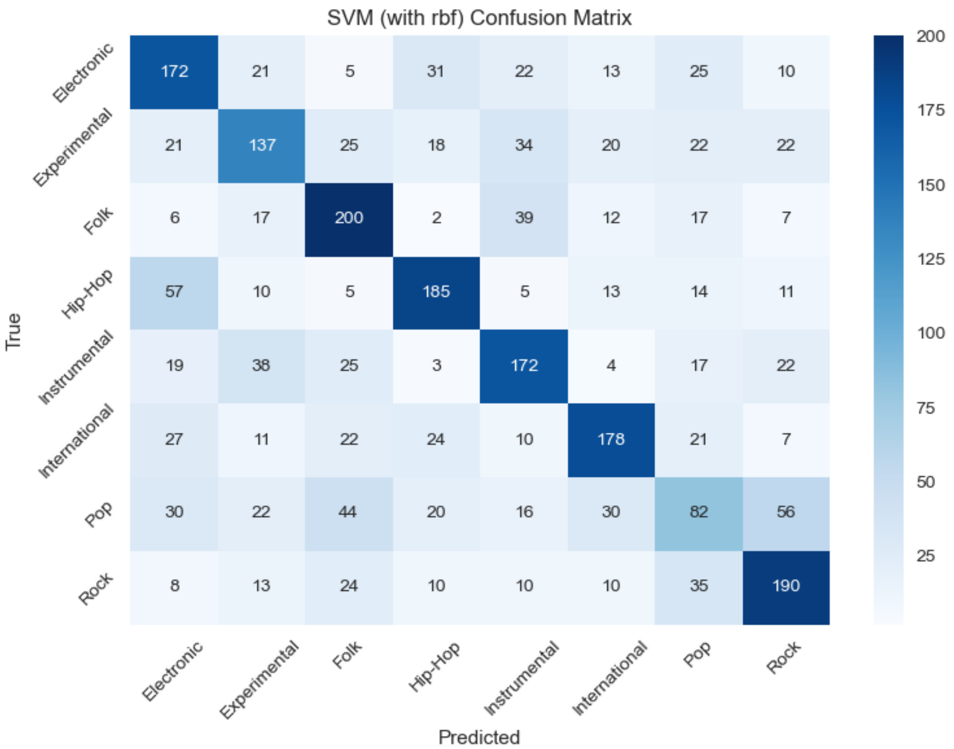
Table 3: Random Forest Classification Report

We can see that the Random Forest classifier achieved a training accuracy of 99.98%, indicating near-perfect learning on the training set, but only 53.92% accuracy on the test set, suggesting limited generalization. The overall macro-averaged F1-score was 0.54. Per-class performance varied substantially. Genres such as Hip-Hop (F1 = 0.63), Folk (F1 = 0.60), International (F1 = 0.59), and Rock (F1 = 0.59) were classified with relatively high precision and recall, consistent with a clear diagonal dominance observed in the confusion matrix. In contrast, Pop showed the weakest performance (F1 = 0.27, recall = 0.24), with a high level of confusion across several other genres, particularly Folk, Rock, and International. Experimental also suffered from lower recall (0.37) despite having moderately good precision (0.61), highlighting the model's difficulty in distinguishing genres with high intra-class variability.

While the Random Forest model captured clear patterns for certain well-defined genres, it showed significant signs of overfitting, as evidenced by the large gap between training and test performance. The model struggles to generalize when facing more heterogeneous or less clearly defined genres, such as Pop and Experimental, where audio features may overlap substantially with those of other classes. The confusion matrix highlights this behavior through the dispersal of predictions for these classes. Moreover, the use of tabular audio features alone may not provide sufficient discriminative power for complex genres where temporal and spectral dynamics are critical.

3.2 Support Vector Machine (SVM)

The SVM was implemented as explained in part 2.4.2. The confusion matrix can be seen in Fig. 3.2.



True \ Predicted	Electronic	Experimental	Folk	Hip-Hop	Instrumental	International	Pop	Rock
Electronic	172	21	5	31	22	13	25	10
Experimental	21	137	25	18	34	20	22	22
Folk	6	17	200	2	39	12	17	7
Hip-Hop	57	10	5	185	5	13	14	11
Instrumental	19	38	25	3	172	4	17	22
International	27	11	22	24	10	178	21	7
Pop	30	22	44	20	16	30	82	56
Rock	8	13	24	10	10	10	35	190

Figure 3.2: Confusion Matrix of the SVM with RBF kernel model.

The Confusion Matrix seems to present a good diagonal dominance, indicating a clear ability to classify several genres correctly. Genres such as Folk (200), Hip-Hop (185), Rock (190), Instrumental (172), and International (178) exhibit strong performance with well-defined diagonal cells. Electronic (172) and Experimental (137) also show acceptable classification results, but with more dispersion toward other classes.

Once again, Pop is the most confused genre, with only 82 correct predictions out of 300, and a broad spread of predictions across almost all other classes (especially toward Rock, International, and Folk), confirming the model's difficulty in capturing the diversity of this genre.

There is somewhat less confusion between Electronic and Hip-Hop than with Random Forest, which suggests that the SVM with RBF kernel captures the nonlinear boundaries between these two classes more effectively.

To nuance, SVM confirms the difficulty of classifying Pop and Experimental genres: the separation of other genres (Folk, Hip-Hop, Rock, International) is better balanced than with Random Forest, with slightly less cross-genre confusion.

The classification report for the SVM model is shown in Table 4.

Train Score: `0.743`
Test Score: `0.5613`

Genre	Precision	Recall	F1-Score	Support
Electronic	0.54	0.56	0.55	299
Experimental	0.50	0.46	0.48	299
Folk	0.59	0.68	0.63	300
Hip-Hop	0.63	0.67	0.65	300
Instrumental	0.58	0.60	0.59	300
International	0.63	0.60	0.62	300
Pop	0.37	0.28	0.31	300
Rock	0.60	0.64	0.62	300
Accuracy			**0.56**	2398
Macro Avg	0.55	0.56	0.56	2398
Weighted Avg	0.55	0.56	0.56	2398

Table 4: SVM Classification Report (see screenshot at figure **??**)

As shown, the SVM classifier with RBF kernel achieved a training accuracy of 74.3% and a test accuracy of 56.13%, slightly outperforming the Random Forest baseline. The overall macro-averaged F1-score reached 0.56. Per-class performance shows that the SVM performed particularly well on Hip-Hop (F1 = 0.65), Folk (F1 = 0.63), Rock (F1 = 0.62), and International (F1 = 0.62), confirming the classifier's ability to distinguish these genres effectively. Instrumental and Electronic also achieved acceptable performance with F1-scores of 0.59 and 0.55, respectively. On the other hand, Experimental (F1 = 0.48) remained a challenging genre to classify, although it slightly improved compared to Random Forest. Pop remained the most problematic class (F1 = 0.31), showing extensive confusion across multiple other genres, as observed in the confusion matrix. Overall, the SVM produced a clearer separation of well-defined genres and a slightly more balanced classification compared to the Random Forest.

While the SVM model demonstrated better generalization than the Random Forest baseline and was able to model nonlinear boundaries between classes effectively, it still faced considerable challenges in classifying certain ambiguous genres such as Pop and Experimental. The confusion matrix highlighted that Pop tracks were widely misclassified across other genres such as International, Folk, and Rock, suggesting that the extracted tabular audio features may not sufficiently capture the diversity of this genre. Furthermore, despite the improved performance, Experimental music remained difficult to classify accurately, likely due to its inherently broad stylistic range. The smaller train-test accuracy gap observed with SVM also indicates less overfitting, suggesting that kernel-based approaches are better suited for this type of feature space. Nonetheless, both models show limitations in handling the subtleties of more heterogeneous genres.

3.3 Multi Layer Perception (MLP)

The MLP model was implemented as explained in part 2.4.3. The confusion matrix can be seen in Fig. 3.3.

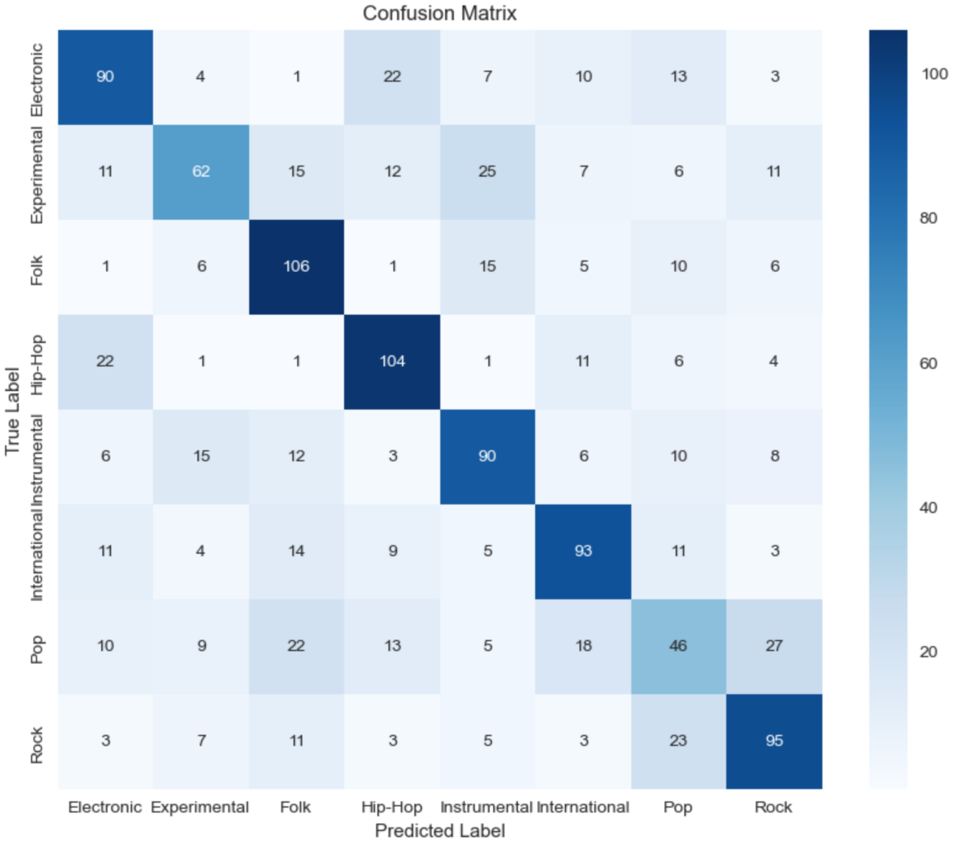


Figure 3.3: Confusion Matrix of the MLP model.

The matrix shows a clear ability of the model to correctly classify several genres, in particular, the model demonstrates strong performance on the Folk (106 correct predictions), Hip-Hop (104), Rock (95), International (93), and Instrumental (90) classes. These genres show relatively low confusion with other categories, suggesting that the MLP was able to capture the underlying discriminative patterns for these musical styles.

In contrast, the model exhibits more difficulty in distinguishing Experimental and Pop genres. Experimental music is often confused with International and Pop, which may reflect shared audio characteristics or inherent variability within the genre.
The Pop class presents the most significant challenge, with only 46 correct predictions and substantial confusion across multiple other genres, particularly Folk, International, and Rock. This result suggests that the audio features used by the MLP may not sufficiently capture the heterogeneity of Pop music or that the genre labels in the dataset lack a clear definition.

Globally, the MLP model achieves satisfactory results on well-separated genres, while struggling with more ambiguous or stylistically diverse categories. These outcomes highlight the need for more informative feature representations or alternative architectures to further improve classification performance on challenging genres.

The classification report for the MLP model is shown in Table 5.

Train Score: `0.6646`
Test Score: `0.5721`

Genre	Precision	Recall	F1-Score	Support
Electronic	0.58	0.60	0.59	150
Experimental	0.57	0.42	0.48	149
Folk	0.58	0.71	0.64	150
Hip-Hop	0.62	0.69	0.66	150
Instrumental	0.59	0.60	0.59	150
International	0.61	0.62	0.61	150
Pop	0.37	0.31	0.33	150
Rock	0.61	0.63	0.62	150
Accuracy			**0.57**	1199
Macro Avg	0.57	0.57	0.57	1199
Weighted Avg	0.57	0.57	0.57	1199

Table 5: MLP Classification Report (see screenshot at figure **??**)

The classification report (Table 5) confirms the observations drawn from the confusion matrix, showing that the MLP classifier achieves an overall accuracy of 57%. The macro- and weighted-average F1-scores are both equal to 0.57, indicating a relatively balanced performance across classes despite noticeable differences between genres.
The best results are obtained for Hip-Hop and Rock, with F1-scores of 0.66 and 0.62, respectively, followed closely by International (0.61) and Folk (0.64), which also exhibit high recall values, reflecting the model's ability to capture relevant patterns for these genres. In contrast, Experimental and Pop perform substantially worse, with F1-scores of 0.48 and 0.33, respectively.

Pop, in particular, suffers from both low precision (0.37) and recall (0.31), consistent with the confusion previously noted in the matrix analysis. These results indicate that while the MLP is effective on genres with well-defined acoustic features, its capacity to generalize is significantly reduced when dealing with more stylistically diverse or overlapping categories.

To further analyze the training dynamics of the MLP model, Figure 3.4 presents the evolution of training and validation accuracy and loss over 80 epochs. The curves show that the model undergoes a rapid initial improvement during the first 10 to 20 epochs, followed by a more gradual increase in accuracy and decrease in loss. The validation accuracy steadily improves and stabilizes around 0.57, which is consistent with the final accuracy reported in Table 5.

The validation loss exhibits a similar behavior, decreasing sharply during the initial epochs and then leveling off after approximately 50 epochs, indicating that the model reaches a plateau. Importantly, no substantial divergence between training and validation curves is observed, suggesting that the model does not suffer from severe overfitting despite the complexity of the classification task. These observations confirm that the MLP model achieves stable and robust learning behavior, but its capacity remains limited by the complexity and heterogeneity of certain musical genres, as previously highlighted in the classification report and confusion matrix analysis.

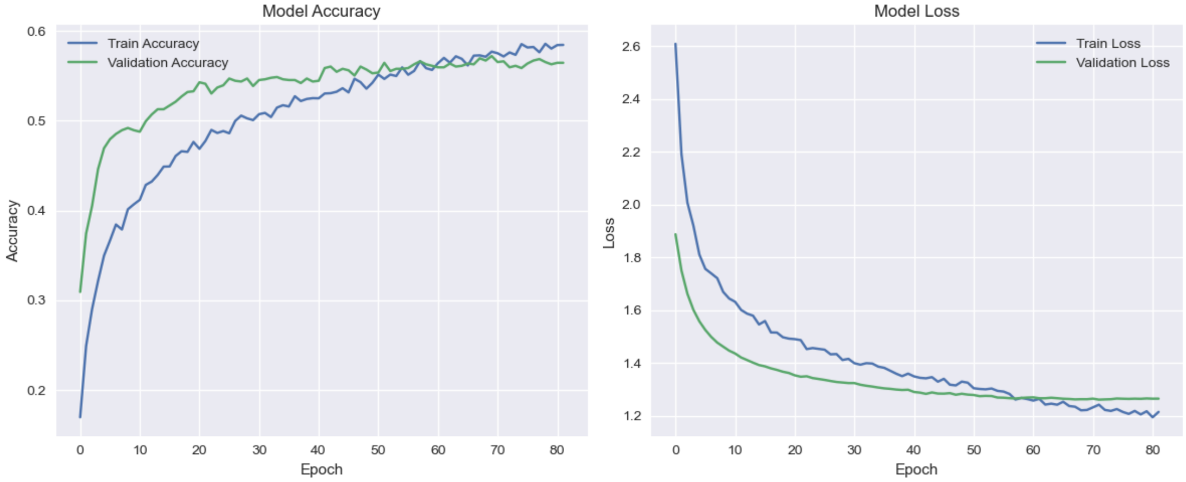


Figure 3.4: Accuracy and Loss vs. Epochs — MLP Model

3.4 1D - CNN

The MLP model was implemented as explained in part 2.4.4. The confusion matrix can be seen in Fig. 3.5.

The confusion matrix of the 1D-CNN model, presented in Figure 3.5, reveals that the model achieves a well-structured classification behavior across most musical genres. The diagonal dominance observed for the Folk (109 correct predictions), International (100), Rock (97), and Electronic (87) classes indicates that the 1D-CNN is capable of extracting relevant temporal patterns from the input representations. The Hip-Hop and Instrumental classes also exhibit good classification performance, with 93 and 80 correct predictions, respectively, although a moderate level of confusion persists, particularly between Hip-Hop and Electronic, and between Instrumental and other genres such as Folk and Experimental.

To contrast, Experimental and Pop continue to present struggles to be identified, with the Experimental genre showing a wider spread of predictions across several classes, and the Pop genre accomplishing only 43 correct predictions and a substantial number of misclassifications towards Rock, Folk, and International.

These results seem to be consistent with previous observations, where Pop and Experimental were systematically identified as ambiguous categories due to their greater intra-class variability and stylistic overlap with other genres.

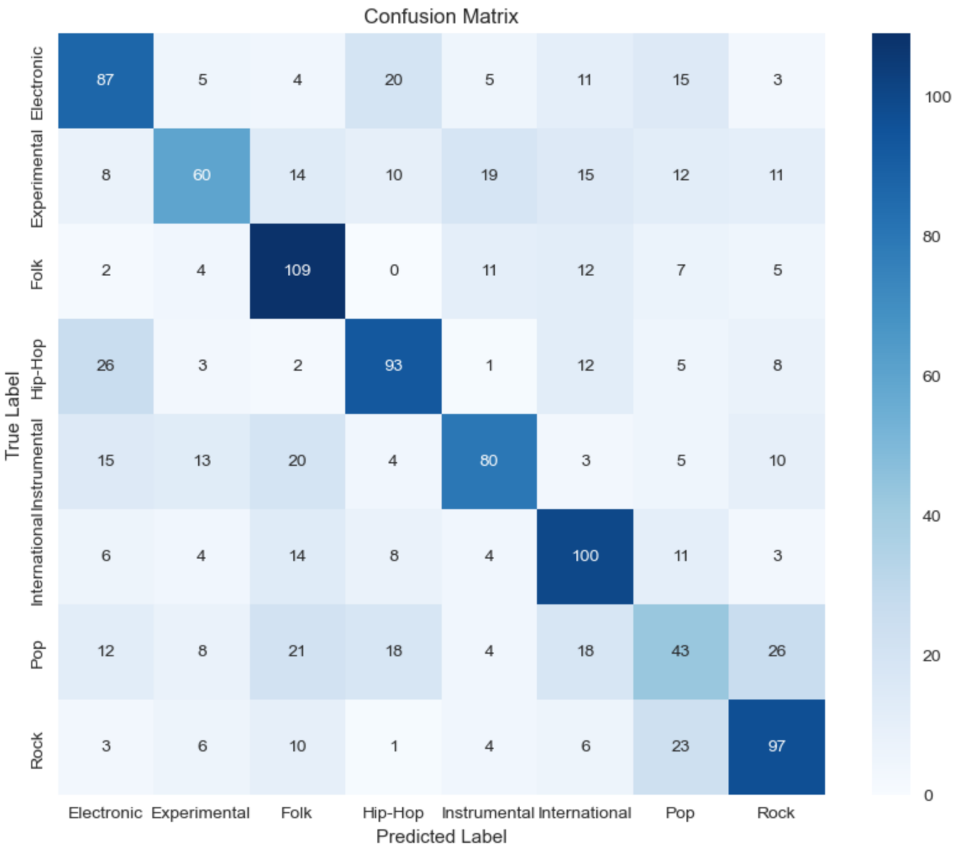


Figure 3.5: Confusion Matrix of the 1D-CNN model.

Overall, the 1D-CNN model demonstrates an improved capacity to separate well-defined genres compared to the MLP, while still facing similar limitations on more heterogeneous classes, highlighting the need for even richer temporal and spectral representations to further enhance classification performance.

The classification report for the 1D-CNN model is shown in Table 6.

Train Score: 0.7007
Test Score: 0.5580 0.5580

Genre	**Precision**	**Recall**	**F1-Score**	**Support**
Electronic	0.55	0.58	0.56	150
Experimental	0.58	0.40	0.48	149
Folk	0.56	0.73	0.63	150
Hip-Hop	0.60	0.62	0.61	150
Instrumental	0.62	0.53	0.58	150
International	0.56	0.67	0.61	150
Pop	0.36	0.29	0.32	150
Rock	0.60	0.65	0.32	150
Accuracy			**0.56**	1199
Macro Avg	0.55	0.56	0.55	1199
Weighted Avg	0.55	0.56	0.55	1199

Table 6: 1D-CNN Classification Report

As shown in Table 6, the classification report for the 1D-CNN model further confirms the patterns observed in the confusion matrix. The overall accuracy reaches 56 %, with macro and weighted average F1-scores of 0.55, indicating a balanced yet modestly improved performance relative to the MLP classifier. The model achieves its best results on genres such as Folk (F1 = 0.63), Hip-Hop (F1 = 0.61), and International (F1 = 0.61), which show both high recall and satisfactory precision, suggesting that the 1D-CNN is capable of effectively capturing sequential information for genres with relatively stable temporal and spectral structures. Rock and Instrumental also yield consistent performance, with F1-scores of 0.60 and 0.58, respectively.

As usual, Experimental and Pop remain the least accurately classified genres. Experimental music obtains an F1-score of 0.48, reflecting a persistent imbalance between its precision and recall, while Pop reaches only 0.32, the lowest score across all categories. The low recall for Pop (0.29) indicates that a large proportion of Pop samples are misclassified, supporting previous observations of its overlap with multiple other genres. These results highlight the continued difficulty in modeling complex or stylistically fluid classes using relatively shallow temporal representations, and motivate the exploration of more expressive architectures capable of learning richer time-frequency features. Overall, the 1D-CNN demonstrates incremental improvements in genre separation compared to traditional dense models, while still showing clear limitations for ambiguous categories.

Figure 3.6 provides additional insights into the training behavior of the 1D-CNN model by displaying the evolution of both accuracy and loss over the course of 30 training epochs. The training curves show a steady increase in accuracy and a consistent decrease in loss, indicating effective optimization on the training set. However, the validation curves diverge noticeably after the first few epochs.

While the validation accuracy initially improves and stabilizes around 0.56, it shows limited progression beyond epoch 10. Simultaneously, the validation loss *"plateau's"* and subsequently begins to increase slightly, suggesting the onset of overfitting. This divergence between training and validation performance confirms that the model continues to adapt excessively to the training distribution without yielding further generalization benefits. These findings are consistent with the moderate classification accuracy observed in Table 6, and reinforce the notion that deeper regularization strategies or more expressive feature representations may be required to enhance the model's ability to generalize, particularly for underperforming or ambiguous genres.

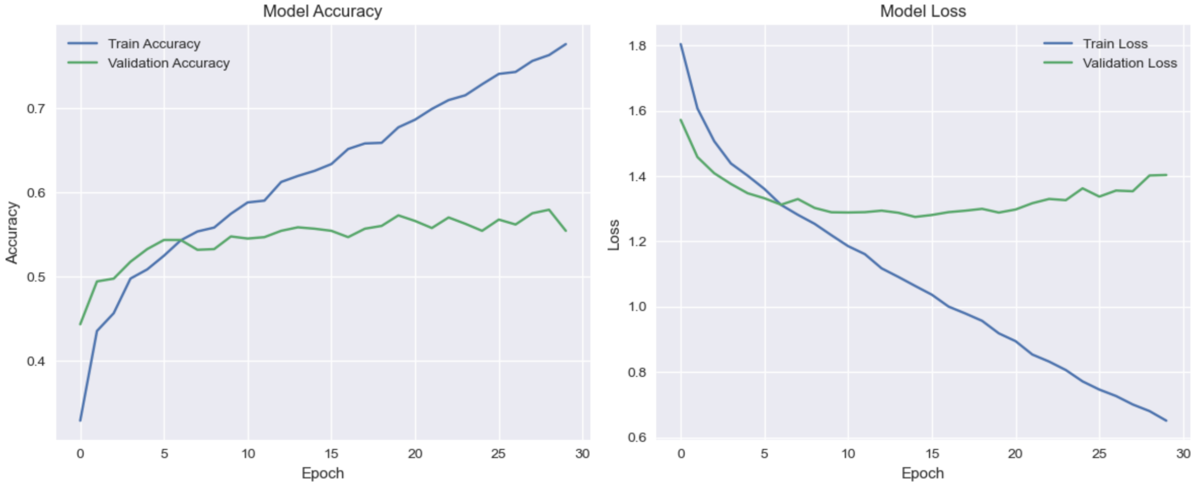


Figure 3.6: Accuracy and Loss vs. Epochs — 1D-CNN Model

3.5 Comparison with other studies' results

As part of this project, we performed a bibliographic review aimed at comparing our models with fine-tuned approaches described in the literature. This allowed us to verify whether our implementation followed similar methodological choices and to critically analyze variations in reported accuracy scores, with the goal of identifying potential reasons why our models may exhibit stronger or weaker performance relative to prior work.

3.5.1 Random Forest

When comparing our Random Forest results (Table 3) with those reported in the reference study (see at report 5.2), a clear difference in accuracy is observed: 54% in our case versus 67.5% in the reference.

However, this discrepancy must be interpreted cautiously. The reference study was conducted on the smaller GTZAN dataset from Kaggle* with a slightly different set of genres (10 genres) and fewer features per class, while our model was evaluated on a larger and more heterogeneous dataset (2398 samples) covering eight genres, including more challenging classes such as Pop and Experimental. Additionally, differences in feature extraction, dataset balance, and genre definitions likely contribute to the observed performance gap. Overall, while our Random Forest model performs adequately given the dataset complexity, more advanced architectures appear necessary to close the performance gap with respect to state-of-the-art results.

3.5.2 SVM with RBF

When comparing our SVM with RBF Kernel results to those reported in the literature [3], several key differences emerge. Firstly, it is important to notice that the reported research focused its study using the GTZAN dataset from Kaggle as it also did for the Random Forest comparison seen above.

Overall accuracy reported in the reference work (5.2) is substantially higher (74%) compared to our model (54.88%). However, it is important to note that the datasets used are not directly comparable. The bibliographic study reports results on a much smaller test set (200 samples) with a different set of ten musical genres, while our evaluation was performed on a larger and more diverse dataset (2398 samples) covering eight genres.

Moreover, the genre distribution and classification difficulty appear to differ (see classification report at 5.3). For example, the referenced study includes genres such as blues, classical, country, and disco, which exhibit relatively high F1-scores, suggesting that the corresponding audio characteristics are well captured by the SVM. In contrast, our dataset includes genres such as Pop and Experimental, which are consistently challenging to classify, as reflected in their lower F1-scores (0.31 and 0.48, respectively). This likely contributes to the lower overall accuracy observed in our results.

Another important distinction lies in the dataset balance and genre overlap. The small sample size and higher per-genre support homogeneity in the reference study may have favored better generalization on their specific task. In our case, larger genre variability, a higher number of ambiguous samples, and greater intra-class variability likely increased the complexity of the classification task.

To sum-up, although both models employ similar SVM-based architectures, the differences in dataset composition, genre selection, and evaluation methodology must be carefully considered when comparing the reported performances. These observations underscore the importance of dataset characteristics and genre definitions in shaping model performance on music classification tasks.

* *The GTZAN dataset is a widely used dataset for evaluation in machine learning research for music genre classification, but comprises 10 genres with only 100 audio files (each 30 seconds long).*

4 Conclusion

Among the models evaluated, the SVM with RBF kernel and the 1D-CNN demonstrated the most balanced and robust performance, achieving the highest generalization capacity across well-defined genres.

The 1D-CNN showed an improved ability to capture sequential audio patterns, while the SVM effectively modeled non-linear class boundaries, outperforming the Random Forest and MLP in terms of balanced genre classification. However, all models exhibited consistent limitations when handling ambiguous or stylistically heterogeneous genres, particularly Pop and Experimental.

These findings highlight the need to integrate richer time-frequency representations and more expressive architectures, such as deeper CNNs or hybrid models combining spectro-temporal features, to further enhance classification performance on complex musical datasets.

5 Appendices

5.1 Models Bloopers

5.1.1 CNN Architecture for Genre Classification

We designed and tested a two-dimensional Convolutional Neural Network (2D CNN) using the Keras API with `TensorFlow` as backend. The architecture was developed to process Mel-spectrogram representations of audio tracks and to learn hierarchical patterns from the time-frequency domain. However, during preliminary experiments, the model failed to achieve competitive results relative to other approaches explored in this study, and further efforts were reoriented toward alternative architectures.

The model takes as input Mel-spectrograms of shape `(64, 128, 1)`, representing the time-frequency content of each audio track. The architecture is composed of three `Conv2D` blocks with 16, 16, and 32 filters, respectively, each followed by `BatchNormalization` and `MaxPooling2D` layers to promote feature stability and downsampling. After the convolutional feature extraction, a `Flatten` layer connects to a fully connected `Dense` layer with 64 `ReLU` units and a `Dropout` rate of 50 %, providing regularization. The final classification layer is a `Dense` softmax layer whose size matches the number of target genres. Additionally, `L2` regularization (factor = 0.001) is applied to all convolutional and dense layers to further enhance generalization.

Parameter	**Value**
Optimizer	Adam (learning rate = 0.0005)
Loss function	Categorical cross-entropy
Epochs	< 100
Batch size	16
Validation split	20% of training data

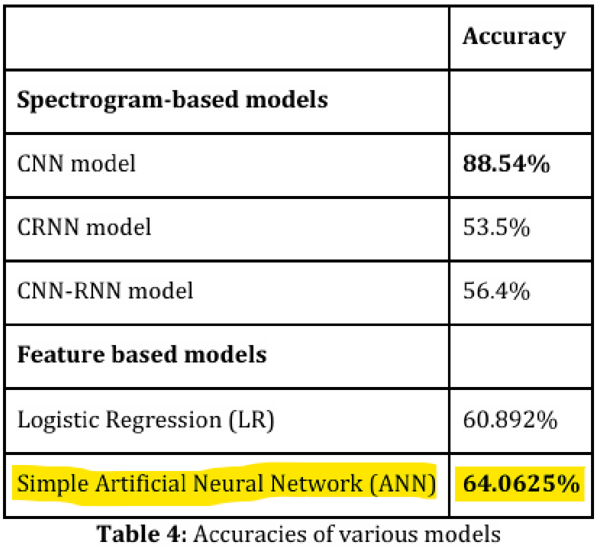
Table 7: Training configuration

Training Strategy and Callbacks While trying to improve model performance and prevent overfitting, the following callbacks were used:

- `EarlyStopping`
- `ModelCheckpoint`
- `TqdmCallback`

5.2 References's results

5.2.1 Reference 1.



	Accuracy
Spectrogram-based models	
CNN model	**88.54%**
CRNN model	53.5%
CNN-RNN model	56.4%
Feature based models	
Logistic Regression (LR)	60.892%
Simple Artificial Neural Network (ANN)	**64.0625%**

Table 4: Accuracies of various models

Figure 5.1: Models's precision from [1].

5.2.2 Reference 3.

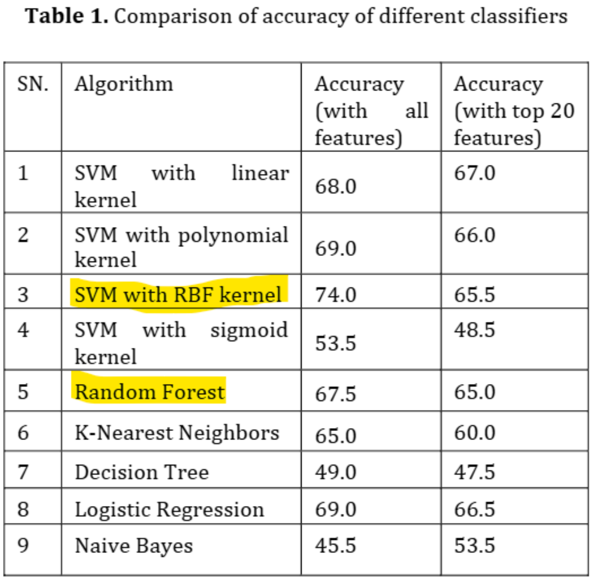


Table 1. Comparison of accuracy of different classifiers

SN.	Algorithm	Accuracy (with all features)	Accuracy (with top 20 features)
1	SVM with linear kernel	68.0	67.0
2	SVM with polynomial kernel	69.0	66.0
3	SVM with RBF kernel	74.0	65.5
4	SVM with sigmoid kernel	53.5	48.5
5	Random Forest	67.5	65.0
6	K-Nearest Neighbors	65.0	60.0
7	Decision Tree	49.0	47.5
8	Logistic Regression	69.0	66.5
9	Naive Bayes	45.5	53.5

Figure 5.2: Accuracy Score of SVM with Kernel and Random Forest models from [3].

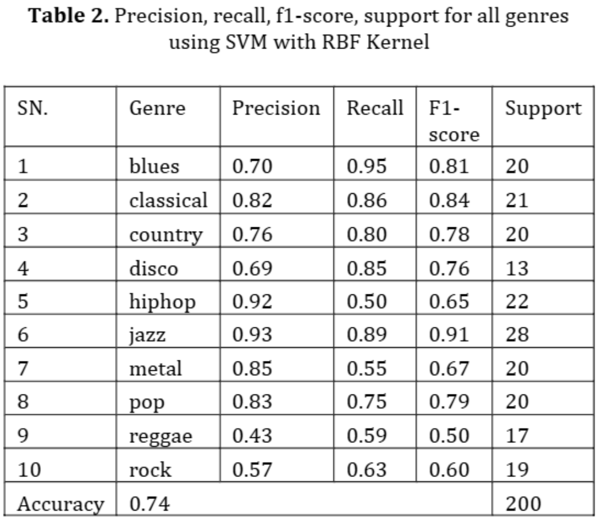


Table 2. Precision, recall, f1-score, support for all genres using SVM with RBF Kernel

SN.	Genre	Precision	Recall	F1-score	Support
1	blues	0.70	0.95	0.81	20
2	classical	0.82	0.86	0.84	21
3	country	0.76	0.80	0.78	20
4	disco	0.69	0.85	0.76	13
5	hiphop	0.92	0.50	0.65	22
6	jazz	0.93	0.89	0.91	28
7	metal	0.85	0.55	0.67	20
8	pop	0.83	0.75	0.79	20
9	reggae	0.43	0.59	0.50	17
10	rock	0.57	0.63	0.60	19
Accuracy	0.74				200

Figure 5.3: Evaluation of 10 genres - SVM with RBF kernel model from [3].

6 Bibliography

References

[1] Guo, Gu, and Liu, "Music Genre Classification via Machine Learning," rep.

[2] "Music Genres Classification and Recommendation System" Jo-Chen Ma, Apr 20, 2023, `https://medium.com/@jochenma/music-genres-classification-and-recommendation-system-b0ae1f2fdb82`

[3] Snigdha Chillara, Kavitha A.S., Shwetha A. Neginhal, Shreya Haldia, and Vidyullatha K.S. 2019. Music Genre Classification using Machine Learning Algorithms: A comparison.