 ISSN:2302-8149
e-ISSN:2540-9719

Deep Learning Approach for Music Genre Classification using Multi-Feature Audio Representations

[1]Nurul Asanah*, [2]Irfan Pratama
[1,2]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Mercu Buana Yogyakarta
[1,2]Jl. Jembatan Merah, No. 84.C. Gejayan, Yogyakarta, Indonesia
*e-mail: 211210030@student.mercubuana-yogya.ac.id

(***received*:** 31 May 2025, ***revised*:** 7 June 2025, ***accepted*:** 13 June 2025)

Abstract

Automatic music genre classification is critical for enhancing user experience in streaming platforms and recommendation systems. This study proposes a Convolutional Neural Network (CNN)-based approach using the GTZAN dataset, which contains ten music genres. The original 30-second audio tracks were segmented into overlapping 3-second chunks, then preprocessed and converted into three feature representations: Mel-Spectrogram, Chroma, and Spectral Contrast. CNN model consisting of four convolutional layers with increasing filters (32–256). The model was trained over 13 epochs using the Adam optimizer. The proposed model achieved 91% accuracy, outperforming previous approaches based on single-feature extraction. The integration of diverse spectral and harmonic features enabled the model to better distinguish between similar genres and improved its generalization. This method offers practical value for real-time music classification, automatic tagging, and intelligent audio indexing in music streaming services and digital libraries.

Keywords: audio features, chroma, CNN, melspectogram, music information retrieval, spectral contrast

1 Introduction

In the current digital age, advancements in technology have significantly altered the processes of music creation, distribution, and consumption. The conventional models of music ownership are progressively being replaced by streaming services like Spotify and Apple Music, which have come to dominate the industry by providing users with immediate access to extensive libraries of songs through any device with internet connectivity [1]. Music, Frequently regarded as a medium for expressing emotions, is defined in the Indonesian Dictionary (KBBI) as a structured arrangement of tones or sounds characterized by rhytm, melody, and harmony [2].

The categorization of music into genres, originating from latin term genus, meaning type or kind, is crucial for the organization of musical content. Nevertheless, this classfication is inherently subjective and is shaped by individual preferences, cultural influences, and personal listening experiences [3]. As the amount of digital music data expands, the need for automated genre classification has become increasingly important for streaming platforms, recommendation algorithms, and content management system. Relying on manual classification by human experts is often inefficient and prone to inconsistency, highlight the demand for intelligent system that can perform genre recognition with precision and objectivity.

Several studies have explored genre classification using deep learning. Palve et al. [[4]] employed CNNs with MFCC features, achieving 85.74% accuracy, while Ilyasa [5] used Mel-Spectrograms with CNNs, obtaining 81.7%. However, these studies typically used single-feature inputs, which limit the model's capacity to capture the full complexity of musical signals. Spectral features such as Chroma and Spectral Contrast have shown promise in enhancing genre recognition but are rarely combined systematically in prior work [6], [7].

Considering these limitations, this study adopts a CNN architecture trained on a combination of Mel-Spectrogram, Chroma, and Spectral Contrast to capture both harmonic and spectral characteristics. The decision to combine these features is based on their complementary strengths: Mel-Spectrogram represents perceptual frequency patterns, Chroma encodes harmonic content, and

 ISSN:2302-8149
e-ISSN:2540-9719

Spectral Contrast highlights timbral variations. This study aims to develop a multi-feature CNN-based music genre classification model that improves accuracy and generalization across diverse genres using the GTZAN dataset.

2 Literature Review

In the domain of Music Information Retrieval (MIR), the automatic classification of music genres has emerged as a significant area of research. MIR is concerned with extracting valuable information from audio data to facilitate applications such as music recommendation, emotion detection, and genre identification [8]. A crucial factor for effective genre classification is the process of feature selection, which enables models to discern pertinent audio patterns while reducing extraneous noise [9].

Tzanctakis and Cook [10] were pioneers in the field of music genre classication, utilizing the GTZAN dataset and employing conventional machine leaming techniques like K-Nearest Neighbors (KNN) and Gaussian Mixture Models (GMM) with Mel-Frequency Cepstral Coefficients (MFCCs) as input features. Although their achieved accuracy of 61% was relatively modest, their research provided a foundational framework that propelled future developments within the discipline.

The advent of deep learning has led to the rise of Convolutional Neural Networks (CNNs) as particularly effective instruments for audio classification tasks. Their capability to leam hierarchical representations renders them exceptionally adept at recognizing intricate auditory patterns without the need for manual feature extraction [11], [12]. A notable study employing CNNs alongside MFCCs on the GTZAN dataset reported an impressive accuracy of 85.74%, underscoring the superiority of CNNs over traditional classification methods [4]. Additional research contrasting various techniques, including Support Vector Machines (SVM), KNN, and feedforward neural networks, has consistently reaffirmed the reliable efficacy of CNNs in genre recognition tasks [13].

Recent research emphasizes the importance of combining multiple audio features to improve classification performance. For example, integrating MFCC with Spectral Contrast or using Mel-Spectrogram alone has shown higher accuracy than single-feature models [5], [6]. Spectrograms have been found to outperform MFCC in certain contexts, achieving up to 76% accuracy versus 58% for MFCC [7]. Moreover, short-duration music segments often yield better accuracy than full-length clips when processed by CNN models [14].

Mel-Spectrogram, Chroma, and Spectral Contrast are particularly powerful in capturing different dimensions of musical information. While Mel-Spectrogram provides perceptually relevant frequency analysis [15], Chroma emphasizes harmonic structure [16], and Spectral Contrast captures variations between harmonic and percussive elements [17]. Together, these features offer a more comprehensive representation than MFCCS alone, and their integration has shown promise in improving genre classification performance across studies.

While previous studies have examined these features individually or in limited combinations, this research distinguishes itself by systematically integrating all three Mel-Spectrogram, Chroma, and Spectral Contrast into a unified CNN framework. This comprehensive multi-feature approach, combined with audio chunking and a deep CNN architecture, enables more accurate and generalizable genre classification compared to prior methods.

3 Research Method

The research workflow is summarized in Figure 1**,** illustrating each step of the methodology used in this study.

 ISSN:2302-8149
e-ISSN:2540-9719

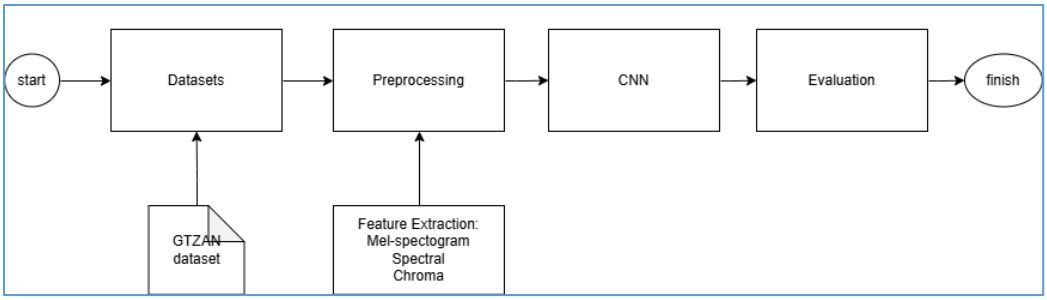


Figure 1. Research procedures

1. Datasets
   The GTZAN dataset used in this study includes 10 different music genres. Each genre is represented by 100 audio clips, each lasting 30 seconds. All audio file are provided in the .wav format. Figure 2 below is displays of the GTZAN dataset sourced from Kaggle, showing the directory structure and organization of audio files across different genre categories.

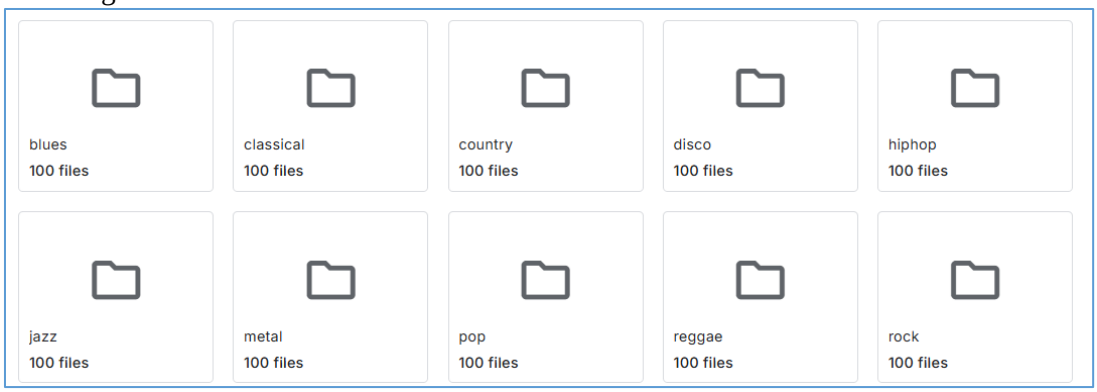


Figure 2. Dataset

   To enhance data volume and model generalization, the original 30-second tracks were segmented into 3-second overlapping chunks with a stride of 1.5 seconds, following best practices in audio data augmentation. This process ecpanded the dataset significantly, resulting in 15.162 training samples and 3.791 test samples.
2. Preprocessing
   The preprocssing and feature extraction stage plays a crucial role in transforming raw audio into structured data suitable for CNN-based classification. The steps involved are as follows:
   a. Data Segmentation
      Each 30-second audio file was divided into 3-second overlapping chunks with a 1.5 second stride, producing multiple segments per track. This technique increased the number of training samples and helped the model capture a wider range of temporal patterns.
   b. Standardization of Audio Format
      All audio chunks were resampled to 22.050 Hz, converted to mono-channel, and normalized to auniform amplitude range to ensure consistency in signal properties.
   c. Feature Extraction
      From each 3-second chunk, three types of spectral features were exracted using librora library :
      - Mel-Spectogram : Represent frequency content on a perceptual (Mel) scale, ideal for capturing timbre
      - Chroma : Hightlights harmoninc structure by capturng energy distribution across 12 pitch classess
      - Spectral Contrast : Measures the difference betwwen spectral peaks and valleys, useful for identifying percussive vs harmonic content

 ISSN:2302-8149
e-ISSN:2540-9719

d. Feature stacking and resizing
These feature were stacked along the channel dimension, resulting in a 128x128x3 array for each chunk, analogous to a 3-channel image. This format is optimized for CNN input compatibility.

e. Label Encoding Dataset Splitting
Each chunk was labeled based on its genre and one hot encoded. The dataset was then split into 80% training and 20% testing useing stratified sampling to maintain class balance.

3. CNN

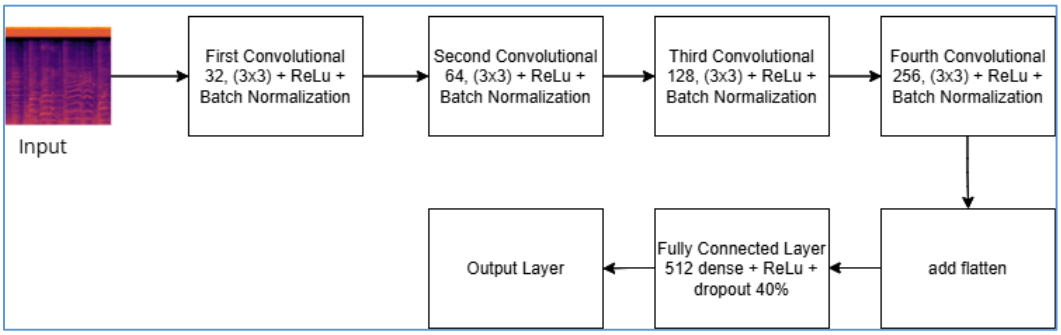


Figure 3. CNN architecture

The classification model used in this study is based on a Convolutional Neural Network (CNN), Figure 3 illustrates the architecture of the Convolutional Neural Network (CNN) used in this study, designed to process 3-channel audio feature images (Mel-Spectrogram, Chroma, Spectral Contrast) with dimensions 128 × 128 × 3. The CNN architecture was built using the TensorFlow Keras library and consists of the following components:

a. Convolutional and Pooling Layers
Many machine learning libraries implement cross-correlation but call it convolution [20]. The architecture starts with multiple conv121tional layers aimed at capturing spatial features from the input. The initial Conv2D layer applies 32 filters with a 3x3 kernel and uses a ReLU activation function, employing 'same' padding to maintain the input dimensions. Following this, BatchNormalization, which normalizes the output and accelerates the training process, and a MaxPooling2D layer to downsample the feature maps, reducing the computational load. This block is repeated three more times with increasing filter sizes: 64, 128, and 256, respectively. Each of these blocks continues to learn more complex features at deeper levels, enabling the model to build a hierarchical understanding of the audio input.

b. Flatten Layer
After the convolutional feature extraction is completed, the multi-dimensional output from the final pooling layer is transformed into a one-dimensional array through the Flatten() operation. This conversion is necessary as it allows the data to be fed into the fully connected layers, which then process the features to make the final classification. The Flatten() operation essentially prepares the data by simplifying the complex feature Slap into a form that can be handled by the dense layers for decision-making

c. Fully Connected Layer (Dense)
The flattened output is then forwarded through a dense (fully connected) layer that contains 512 neurons, with a ReLU activation function applied to each neuron. This layer plays a critical role in learning complex, non-linear combinations of the high-level features that have been extracted by the previous convolutional and pooling layers. Following this, a Dropout layer with a rate of 0.4 is introduced to help mitigate overfitting. By randomly deactivating 40% of the neurons during the training process, the Dropout layer forces the model to rely on different sets of neurons, promoting better generalization to new, unseen data.

 ISSN:2302-8149
e-ISSN:2540-9719

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 150, 150, 32)	320
batch_normalization (BatchNormalization)	(None, 150, 150, 32)	128
max_pooling2d (MaxPooling2D)	(None, 75, 75, 32)	0
conv2d_1 (Conv2D)	(None, 75, 75, 64)	18,496
batch_normalization_1 (BatchNormalization)	(None, 75, 75, 64)	256
max_pooling2d_1 (MaxPooling2D)	(None, 37, 37, 64)	0
conv2d_2 (Conv2D)	(None, 37, 37, 128)	73,856
batch_normalization_2 (BatchNormalization)	(None, 37, 37, 128)	512
max_pooling2d_2 (MaxPooling2D)	(None, 18, 18, 128)	0
conv2d_3 (Conv2D)	(None, 18, 18, 256)	295,168
batch_normalization_3 (BatchNormalization)	(None, 18, 18, 256)	1,024
max_pooling2d_3 (MaxPooling2D)	(None, 9, 9, 256)	0
flatten (Flatten)	(None, 20736)	0
dense (Dense)	(None, 512)	10,617,344
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 10)	5,130

Total params: 11,012,234 (42.01 MB)
Trainable params: 11,011,274 (42.00 MB)
Non-trainable params: 960 (3.75 KB)

Figure 4. Model summary

d. Output Layer

The final layer of the network is a dense layer where the number of neurons corresponds to the number of target classes. This layer uses a softmax activation function, which is crucial for generating a probability distribution over all possible classes. By doing so, it enables the model to output the probability for each genre, allowing it to predict the genre with the highest likelihood for a given audio sample. The softmax function ensures that the sum of all probabilities equals one, providing a clear indication of which class is most likely based on the model's learned features.

The mpdel was compiled with:

- Optimizer: Adam (learning rate = 0.0001)
- Loss function: Categorical Crossentropy
- Metric: Accuracy

4. Evaluation

This study evaluated the performance of the proposed CNN model for music genre classification using various metrics for both overall and class-specific insights. The primary measure was accuracy, which reflects the proportion of correctly predicted genre labels. However, to address the limitations of accuracy, particularly in cases of class imbalance, additional metrics such as precision, recall, and F1-score were employed. Precision assesses the correctness of positive predictions, while recall indicates the model's ability to identify all relevant instances within each genre. The F1-score, the harmonic mean of precision and recall, offers a balanced assessment, particularly in uneven class distributions. Metrics were computed for each genre to evaluate model performance across categories. A confusion matrix was also created to illustrate misclassification patterns, revealing frequently confused genres. All evaluations were performed using the Scikit-learn library's classification_report and confusion_matrix functions on the test dataset post-training.

 ISSN:2302-8149
e-ISSN:2540-9719

4 Results and Analysis

1. Datasets

   The GTZAN dataset initially consists of 1,000 audio tracks that are evenly distributed among 10 different music genres. Following the segmentation process, the dataset was considerably expanded to enhance the diversity of training and minimize the chances of overfitting. This approach allowed the model to encounter a wider range of patterns within individual tracks, effectively capturing more distinct characteristics associated with each genre. Additionally, it contributed to reducing the possibility of overfitting by offering a greater number of training examples from the relatively small original dataset.

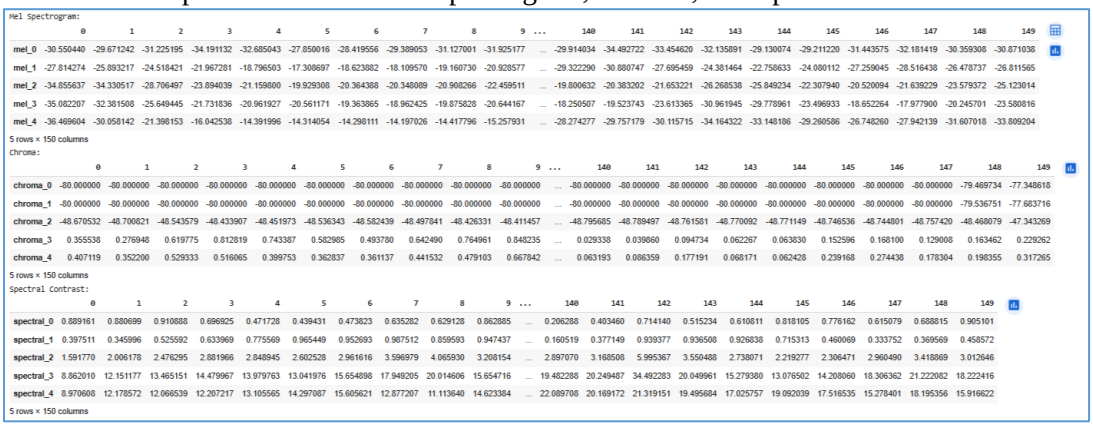
   Table 1 below presents the distribution of audio samples before and after segmentation, showing the increase in data volume resulting from chunking 30-second tracks into overlapping 3-second segments.

Table 1. Distribution of audio samples

Dataset Version	Samples
Original (30 second clips)	1.000 (100 per genres)
After chunking (3s, 1.5 overlap)	18.953 total - 15.162 training - 3.791 testing

2. Preprocessing

   The preprocessing and feature extraction procedures effectively organized the GTZAN dataset into a format suitable for genre classification using deep learning techniques. Each segment of audio was transformed into a feature matrix measuring 128 × 128 × 3, which included representations of Mel-Spectrogram, Chroma, and Spectral Contrast.



Mel Spectrogram:

	0	1	2	3	4	5	6	7	8	9	...	140	141	142	143	144	145	146	147	148	149
mel_0	-30.550440	-29.671242	-31.225195	-34.191132	-32.685043	-27.850016	-28.419556	-29.389053	-31.127001	-31.925177	...	-29.914034	-34.492722	-33.454620	-32.135891	-29.130074	-29.211220	-31.443575	-32.181419	-30.359308	-30.871038
mel_1	-27.814274	-25.893217	-24.518421	-21.967281	-18.796503	-17.308697	-18.623882	-18.109570	-19.160730	-20.928577	...	-29.322290	-30.880747	-27.695459	-24.381464	-22.758633	-24.080112	-27.259045	-28.516438	-26.478737	-26.811565
mel_2	-34.855637	-34.330517	-28.706497	-23.894039	-21.159800	-19.929308	-20.364388	-20.348089	-20.908266	-22.459511	...	-19.800632	-20.383202	-21.653221	-26.268538	-25.849234	-22.307940	-20.520094	-21.639229	-23.579372	-25.123014
mel_3	-35.082207	-32.381508	-25.649445	-21.731836	-20.961927	-20.561171	-19.363865	-18.962425	-19.875828	-20.644167	...	-18.250507	-19.523743	-23.613365	-30.961945	-29.778961	-23.496933	-18.652264	-17.977900	-20.245701	-23.580816
mel_4	-36.469604	-30.058142	-21.398153	-16.042538	-14.391996	-14.314054	-14.298111	-14.197026	-14.417796	-15.257931	...	-28.274277	-29.757179	-30.115715	-34.164322	-33.148186	-29.260586	-26.748260	-27.942139	-31.607018	-33.809204

5 rows × 150 columns

Chroma:

	0	1	2	3	4	5	6	7	8	9	...	140	141	142	143	144	145	146	147	148	149
chroma_0	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	...	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-79.469734	-77.348618
chroma_1	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	...	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-80.000000	-79.536751	-77.683716
chroma_2	-48.670532	-48.700821	-48.543579	-48.433907	-48.451973	-48.536343	-48.582439	-48.497841	-48.426331	-48.411457	...	-48.795685	-48.789497	-48.761581	-48.770092	-48.771149	-48.746536	-48.744801	-48.757420	-48.468079	-47.343269
chroma_3	0.355538	0.276948	0.619775	0.812819	0.743387	0.582985	0.493780	0.642490	0.764961	0.848235	...	0.029338	0.039860	0.094734	0.062267	0.063830	0.152596	0.168100	0.129008	0.163462	0.229262
chroma_4	0.407119	0.352200	0.529333	0.516065	0.399753	0.362837	0.361137	0.441532	0.479103	0.667842	...	0.063193	0.086359	0.177191	0.068171	0.062428	0.239168	0.274438	0.178304	0.198355	0.317265

5 rows × 150 columns

Spectral Contrast:

	0	1	2	3	4	5	6	7	8	9	...	140	141	142	143	144	145	146	147	148	149
spectral_0	0.889161	0.880699	0.910888	0.696925	0.471728	0.439431	0.473823	0.635282	0.629128	0.862885	...	0.206288	0.403460	0.714140	0.515234	0.610811	0.818105	0.776162	0.615079	0.688815	0.905101
spectral_1	0.397511	0.345996	0.525592	0.633969	0.775569	0.965449	0.952693	0.987512	0.859593	0.947437	...	0.160519	0.377149	0.939377	0.936508	0.926838	0.715313	0.460069	0.333752	0.369569	0.458572
spectral_2	1.591770	2.006178	2.476295	2.881966	2.848945	2.602528	2.961616	3.596979	4.065930	3.208154	...	2.897070	3.168508	5.995367	3.550488	2.738071	2.219277	2.306471	2.960490	3.418869	3.012646
spectral_3	8.862010	12.151177	13.465151	14.479967	13.979763	13.041976	15.654898	17.949205	20.014606	15.654716	...	19.482288	20.249487	34.492283	20.049961	15.279380	13.076502	14.208060	18.306362	21.222082	18.222416
spectral_4	8.970608	12.178572	12.066539	12.207217	13.105565	14.297087	15.605621	12.877207	11.113640	14.623384	...	22.089708	20.169172	21.319151	19.495684	17.025757	19.092039	17.516535	15.278401	18.195356	15.916622

5 rows × 150 columns

Figure 5. Top 5 of each extraction

Maintaining genre balance throughout this process ensured that the model was not biased toward any specific genre class. In contrast to earlier studies that relied solely on isolated features such as MFCC or Mel-Spectrogram, this research introduces a multi-channel feature map which notably improves classification performance. This methodology enhances the model's robustness and generalizability, as evidenced by the elevated validation accuracy observed in subsequent results.

3. CNN

 ISSN:2302-8149
e-ISSN:2540-9719

```
Epoch 1/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1541s 3s/step - accuracy: 0.4730 - loss: 1.6861 - val_accuracy: 0.7199 - val_loss: 0.7838
Epoch 2/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1554s 3s/step - accuracy: 0.7273 - loss: 0.7838 - val_accuracy: 0.8172 - val_loss: 0.5551
Epoch 3/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1514s 3s/step - accuracy: 0.8370 - loss: 0.4811 - val_accuracy: 0.8285 - val_loss: 0.5145
Epoch 4/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1541s 3s/step - accuracy: 0.8976 - loss: 0.2912 - val_accuracy: 0.8351 - val_loss: 0.5102
Epoch 5/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1498s 3s/step - accuracy: 0.9364 - loss: 0.1834 - val_accuracy: 0.8992 - val_loss: 0.3041
Epoch 6/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1504s 3s/step - accuracy: 0.9568 - loss: 0.1298 - val_accuracy: 0.9217 - val_loss: 0.2314
Epoch 7/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1521s 3s/step - accuracy: 0.9719 - loss: 0.0927 - val_accuracy: 0.9016 - val_loss: 0.3108
Epoch 8/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1555s 3s/step - accuracy: 0.9764 - loss: 0.0786 - val_accuracy: 0.8963 - val_loss: 0.3280
Epoch 9/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1520s 3s/step - accuracy: 0.9769 - loss: 0.0689 - val_accuracy: 0.8963 - val_loss: 0.3335
Epoch 10/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1551s 3s/step - accuracy: 0.9825 - loss: 0.0547 - val_accuracy: 0.9261 - val_loss: 0.2368
Epoch 11/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1550s 3s/step - accuracy: 0.9796 - loss: 0.0642 - val_accuracy: 0.9114 - val_loss: 0.2720
Epoch 12/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1506s 3s/step - accuracy: 0.9879 - loss: 0.0397 - val_accuracy: 0.9137 - val_loss: 0.3241
Epoch 13/13
474/474 ━━━━━━━━━━━━━━━━━━━━ 1548s 3s/step - accuracy: 0.9845 - loss: 0.0441 - val_accuracy: 0.9111 - val_loss: 0.2825
```

Figure 6. Training with 13 epochs

The CNN architecture demonstrated effective learning during training. The model was trained for 13 epochs, achieving 91% accuarcy on test set.

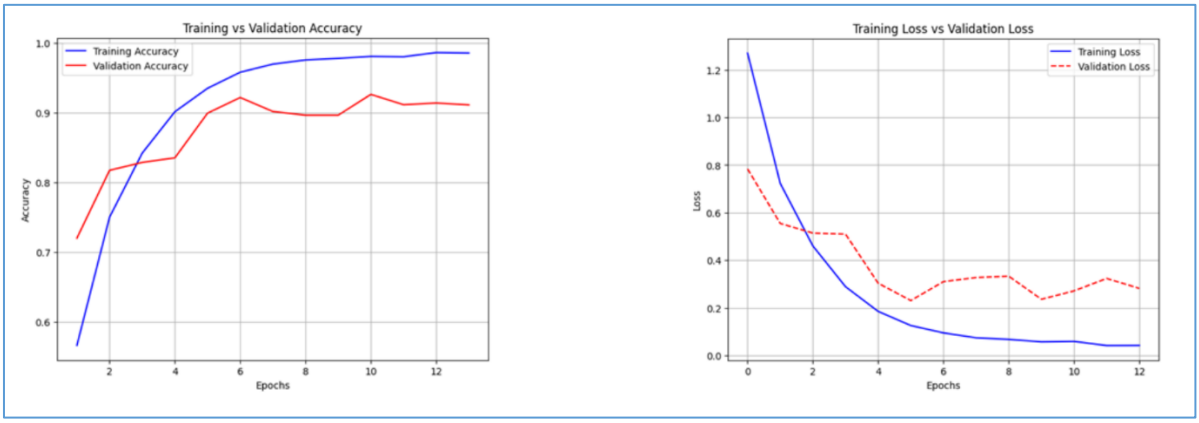


Figure 7. Graph accuracy vs loss

The training accuracy increased steadily, while the validation accuracy stabilized after epoch 6, indicating early signs of overfitting but still maintaining strong generalization. Loss curves also support this interpretation, with validation loss remaining low and stable despite fluctuations.

The CNN architecture effectively identified essential temporal and spectral characteristics across various channels, including mel, chroma, and contrast. Each layer enhanced the abstraction of features, enabling the model to differentiate between genres that exhibit intricate similarities in auditory properties.

4. Evaluation

   The classification report below summarizes the CNN model's precision, recall, and F1-score for each of the 10 music genres:

 ISSN:2302-8149
e-ISSN:2540-9719

	precision	recall	f1-score	support
blues	0.8974	0.9552	0.9254	357
classical	0.9842	0.9739	0.9790	383
country	0.8986	0.8817	0.8901	372
disco	0.9242	0.9130	0.9186	414
hiphop	0.8929	0.9390	0.9153	426
jazz	0.9407	0.9694	0.9549	360
metal	0.9291	0.9794	0.9536	388
pop	0.8981	0.8670	0.8823	376
reggae	0.9963	0.7330	0.8446	367
rock	0.7809	0.8908	0.8322	348
accuracy			0.9111	3791
macro avg	0.9142	0.9102	0.9096	3791
weighted avg	0.9150	0.9111	0.9105	3791

Figure 8. Classification report

The classification report above showcases how well the Convolutional Neural Network model effectiveness in categorizing diverse music genres, utilizing mel-spectrogram representations as input features. It provides detailed evaluation metrics such as precision, recall, F1-score, and support for each of the ten genre categories. Overall, the model achieves an impressive accuracy of 91%, meaning that it was able to assign the correct genre label in the vast majority of test instances.

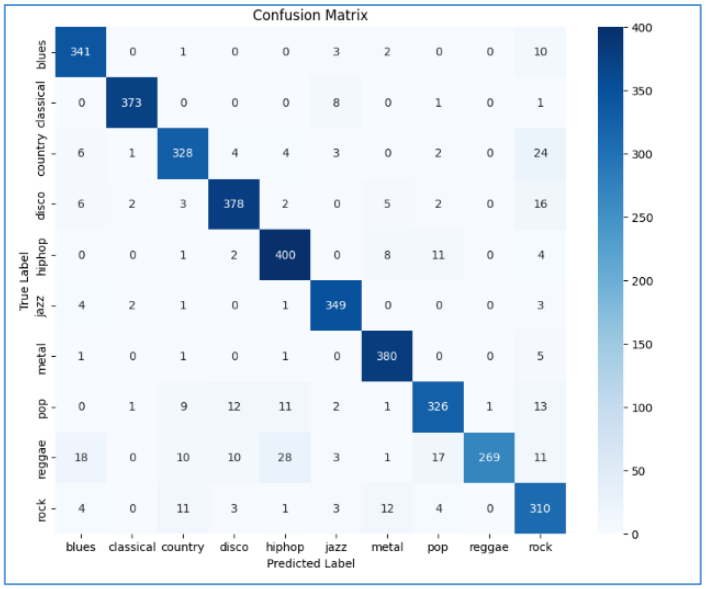


Figure 9. Confusion matrix

The figure above presents the confusion matrix resulting from a music genre classification model. Each row in the matrix corresponds to the true genre, and each column shows the predicted genre. The diagonal entries reflect the number of correct predictions for each genre, while the off-diagonal entries highlight the misclassifications. The model excels particularly in identifying genres such as hiphop (400 correct predictions), metal (380), classical (373), and disco (378), indicating strong learning of their distinctive features. However, there is some confusion between similar genres. For instance:

- Several reggae tracks are misclassified as hiphop and pop, possibly due to overlapping rhythm or vocal patterns.
- Country and rock also show moderate confusion with reggae and pop, suggesting shared acoustic characteristics.

Despite a few misclassifications, the matrix suggests that the CNN model achieves high overall accuracy and can distinguish between most music genres effectively.

 ISSN:2302-8149
e-ISSN:2540-9719

Comparision with Previous Study

To further validate the effectiveness of the proposed CNN model using Mel- Spectrogram, Chroma, and Spectral Contrast features, a comparative analysis was conducted with several previous studies that utilized the GTZAN dataset. The comparison focuses on the feature types, model architecture, and the resulting accuracy achieved in each study.

Table 2. Comparision with previous study

Research	Feature(s) used	Model type	Dataset	Accuracy
[10]	MFCC	GMM, KNN	GTZAN	61%
[4]	MFCC	CNN	GTZAN	85.74%
[5]	Mel-Spectogram	CNN	GTZAN	81.7%
[13]	MFCC	SVM, KNN, CNN	GTZAN	76%
[6]	MFCC, Spectral contrast	LSTM	GTZAN	82%
[18]	MFCC, Chroma, Mel-Spectogram	CNN, MLP	GTZAN	76%
[19]	STFT	CNN,nnet1	GTZAN	87.4%
This Study	Mel-Spectogram, Chroma, Spectral Contrast	CNN	GTZAN	91%

5 Conclusion

This research introduces a convolutional neural network (CNN) methodology for classifying music genres by utilizing a blend of Mel-Spectrogram, Chroma, and Spectral Contrast features. The combination of these features allowed the model to attain an impressive accuracy of 91% on the GTZAN dataset, demonstrating superior performance compared to earlier studies. The use of overlapping audio segmentation, combined with a deep and structured CNN architecture, addressed limitations in previous research related to low feature diversity, overfitting, and insufficient generalization.

This research contributes to the field by offering a systematic, multi-feature integration framework for CNN-based audio classification, which enhances the model's ability to distinguish between genres with similar acoustic traits. This framework not only improves classification accuracy but also increases robustness across various types of music content.

In practical terms, the findings of this study can be applied to real-world music streaming platforms, automated music tagging systems, and digital content management tools, where precise genre recognition is essential for enhancing user personalization and searchability. Future studies may consider examining a wider variety of datasets, implementing temporal modeling with hybrid architectures, or developing real-time genre detection capabilities for streaming services.

References

[1] X. Guo, "*The Evolution of the Music Industry in the Digital Age: From Records to Streaming*," *Journal of Sociology and Ethnology*, Vol. 5, No. 10, 2023, doi: 10.23977/jsoce.2023.051002.

[2] "Hasil Pencarian - KBBI VI Daring." Accessed: Apr. 25, 2025. [Online]. Available: https://kbbi.kemdikbud.go.id/entri/musik

[3] A. S. Pratama, "Klasifikasi Genre Musik Populer menggunakan Metode *Convolutional Neural Network* dengan Data *Augmentation*," 2021.

[4] S. Palve, S. Dubey, M. Dhanait, N. Purswani, K. P. Birla 1234 Student, and K. K. Wagh, "*Music Genre Classification using Convolutional Neural Networks (CNN)*," *International Journal of Research and Analytical Reviews*, 2023, Accessed: May 06, 2025. [Online]. Available: www.ijrar.org

 ISSN:2302-8149
e-ISSN:2540-9719

[5] A. N. Ilyasa, "Pengembangan Aplikasi Klasifikasi Sepuluh Genre Musik," *KALBISIANA Jurnal Sains, Bisnis dan Teknologi*, Vol. 10, No. 4, pp. 388–394, Dec. 2024, doi: 10.53008/KALBISIANA.V10I4.633.

[6] M. N. Farid, ¶,½í, A. F. Rahman, H. Wicaksono, and I. T. Kalimantan, "Analisis Pengaruh Kombinasi Fitur Spektral terhadap Tingkat Akurasi *Speech Emotion Recognition*," *Jurnal Sistim Informasi dan Teknologi*, Vol. 5, No. 2, pp. 120–129, Jun. 2023, doi: 10.37034/JSISFOTEK.V5I2.234.

[7] S. M. Fardhani, Y. Wihardi, and E. Piantari, "Klasifikasi *Genre* Musik dengan *Mel Frequency Cepstral Coefficient* dan *Spektogram* menggunakan *Convolutional Neural Network*," *Jurnal Aplikasi dan Teori Ilmu Komputer*, Vol. 4, No. 1, pp. 26–34, 2021, doi: 10.17509/JATIKOM.V4I1.41465.

[8] J. A. R, "Rancang Bangun Aplikasi *Musicmoo* dengan Metode *Mir (Music Information Retrieval*) pada Modul *Mood, Genre Recognition*, dan Tempo *Estimation*," Institut Teknologi Sepuluh Nopember , Surabaya, 2017. Accessed: Apr. 27, 2025. [Online]. Available: https://repository.its.ac.id/3713/2/5113100021-Undergraduate-Theses.pdf

[9] L. A. A. R. P. Putri, "Seleksi Fitur dalam Klasifikasi Genre Musik", Accessed: Apr. 27, 2025. [Online]. Available: https://ojs.unud.ac.id/index.php/jik/article/view/39772/24169

[10] G. Tzanetakis and P. Cook, "*Musical Genre Classification of Audio Signals*," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, pp. 293–302, Jul. 2002, doi: 10.1109/TSA.2002.800560.

[11] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "*A Survey of Audio Classification using Deep Learning*," *IEEE Access*, Vol. 11, pp. 106620–106649, 2023, doi: 10.1109/ACCESS.2023.3318015.

[12] Y. Cui and F. Wang, "*Research on Audio Recognition based on the Deep Neural Network in Music Teaching*," *Comput Intell Neurosci*, Vol. 2022, 2022, doi: 10.1155/2022/7055624.

[13] J. Dias, V. Pillai, H. Deshmukh, and A. Shah, "*Music Genre Classification & Recommendation System using CNN*," *SSRN Electronic Journal*, Apr. 2022, doi: 10.2139/SSRN.4111849.

[14] Y. V. Via, I. Y. Purbasari, and A. P. Pratama, "Analisa *Algoritma Convolution Neural Network* (CNN) pada Klasifikasi Genre Musik berdasar Durasi Waktu," *Scan : Jurnal Teknologi Informasi dan Komunikasi*, Vol. 17, No. 1, pp. 35–41, Feb. 2022, doi: 10.33005/scan.v17i1.3251.

[15] D. Joshi, J. Pareek, and P. Ambatkar, "*Comparative Study of Mfcc and Mel Spectrogram for Raga Classification using CNN*," *Indian J Sci Technol*, Vol. 16, No. 11, pp. 816–822, Mar. 2023, doi: 10.17485/IJST/V16I11.1809.

[16] F. Korzeniowski and G. Widmer, "*Feature Learning For Chord Recognition: The Deep Chroma Extractor*", Accessed: May 06, 2025. [Online]. Available: https://rodrigob.github.io/are

[17] C.-H. Lee, J.-L. Shih, K.-M. Yu, and J.-M. Su, "Automatic Music Genre Classification Using Modulation Spectral Contrast Feature".

[18] X. Li, F. Li, Z. P. Lu, and Z. yue Yang, "*Music Genre Classification: A Comprehensive Study on Feature Fusion with CNN and MLP Architectures*," *Applied and Computational Engineering*, Vol. 132, No. 1, pp. 159–166, Jan. 2025, doi: 10.54254/2755-2721/2024.20632.

[19] W. Zhang, W. Lei, X. Xu, and X. Xing, "*Improved Music Genre Classification with Convolutional Neural Networks*," 2016, doi: 10.21437/Interspeech.2016-1236.