ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika

ISSN(p): 2338-8323 | ISSN(e): 2459-9638 | Vol. 13 | No. 1 | Halaman 84 - 99
DOI : http://dx.doi.org/10.26760/elkomika.v13i1.84 January 2025

Audio Conversion for Music Genre Classification Using Short-Time Fourier Transform and Inception V3

Article History:
Received
30 September 2024
Revised
11 December 2024
Accepted
24 January 2025

DEWI ROSMALA, MOHAMMAD NOER FADHILAH

Informatics, Institut Teknologi Nasional Bandung, West Java, Indonesia
Email: d_rosmala@itenas.ac.id

ABSTRAK

Penelitian ini mengkaji tentang perkembangan genre musik dan aplikasi teknologi dalam pengenalan genre musik melalui pendekatan MIR (Music Information Retrieval). Pelabelan genre musik secara otomatis diharapkan dapat membantu, mengurangi, dan menekan peran manusia dalam hal pelabelan genre musik. Penelitian ini mengusulkan penggunaan Mel Spectrogram sebagai representasi audio dalam domain frekuensi serta Convolutional Neural Network (CNN), khususnya arsitektur Inception V3. CNN dipilih karena kemampuannya untuk mengenali pola yang kompleks dan hirarkis, yang sesuai dengan fitur-fitur musik yang direpresentasikan dalam spektogram. Diterapkan teknik pembelajaran transfer dan fine-tuning model yang dilatih pada dataset yang besar, yang memungkinkan untuk meningkatkan akurasi. Penelitian ini menggunakan dataset 1000 file audio dalam format .wav, dengan masing-masing genre diwakili oleh 100 file, untuk mengevaluasi kinerja dan keefektifan metode yang diusulkan dalam konteks klasifikasi genre musik.

***Kata kunci**: Mel Spectrogram, CNN Inception V3*

ABSTRACT

This research examines the development of music genres and technological applications in music genre recognition through the MIR (Music Information Retrieval) approach. Automatic music genre labeling is expected to help, reduce, and suppress the role of humans in terms of music genre labeling. This research proposes the use of Mel Spectrogram as an audio representation in the frequency domain as well as Convolutional Neural Network (CNN), specifically the Inception V3 architecture. CNN was chosen for its ability to recognize complex and hierarchical patterns, which corresponds to the musical features represented in the spectrogram. Transfer learning techniques and fine-tuning of models trained on large datasets were applied, which allowed to improve accuracy. This study uses a dataset of 1000 audio files in .wav format, with each genre represented by 100 files, to evaluate the performance and effectiveness of the proposed method in the context of music genre classification.

***Keywords**: Mel Spectrogram, CNN Inception V3*

1. INTRODUCTION

Machine Learning is a technique that is often used, especially deep learning to recognize, classify, etc. There are various techniques and architectures to handle some of these cases, with the CNN method of the Inception V3 architecture being a popular choice for classifying objects based on images. Inception V3 is an image recognition model that has been proven to achieve over 78.1 accuracy on the ImageNet dataset. This model has been a major development focus by many researchers over the years. The research and developments were inspired by the paper "Rethinking Inception Architectures for Computer Vision" **(Szegedy, 2016)**

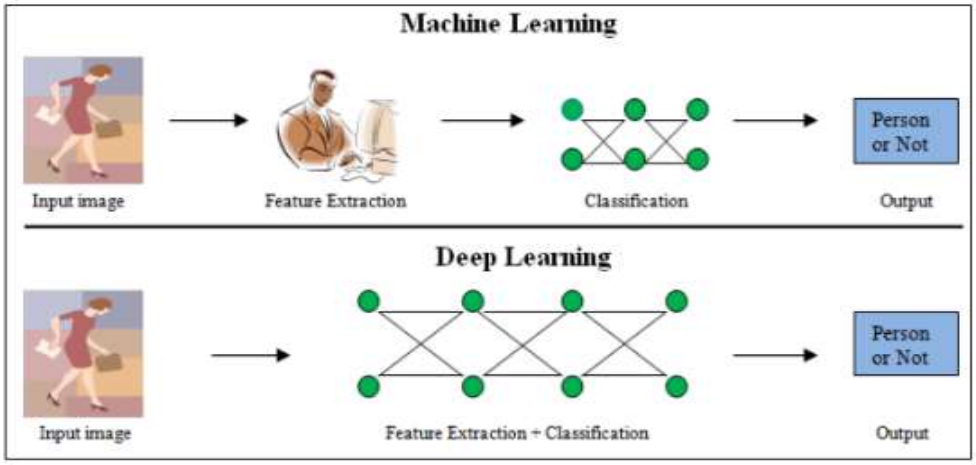


Figure 1. Difference between Deep Learning and Machine Learning

On the other hand, to classify audio with the Inception V3 architecture, the audio must first be converted into an image or spectrum (Mel Spectrogram) which will then be converted into a matrix (pixel values) to be used as input.

Metode CNN Inception V3 merupakan salah satu pendekatan dalam deep learning yang digunakan untuk melakukan proses klasifikasi. CNN Inception V3 dianggap sebagai deep learning karena memiliki arsitektur jaringan yang kompleks. CNN Inception V3 memiliki struktur kerja yang terdiri dari input, output, dan proses klasifikasi. Selain itu, CNN Inception V3 juga melakukan proses ekstraksi fitur dengan menggunakan berbagai macam hidden layer, antara lain convolution, pooling, activation (ReLU), softmax, dan fully connected layer. CNN Inception V3 bekerja secara terstruktur untuk mengorganisasikan objek, sehingga output dari satu lapisan konvolusi dapat digunakan sebagai input untuk lapisan konvolusi berikutnya. **(Arrofiqoh, 2018)**

The CNN Inception V3 method is one of the deep learning approaches used to carry out the classification process. CNN Inception V3 is considered deep learning because it has a complex network architecture. CNN Inception V3 has a work structure consisting of input, output, and classification process. Apart from that, CNN Inception V3 also carries out the feature extraction process using various hidden layers, including convolution, pooling, activation (ReLU), softmax, and fully connected layers. CNN Inception V3 works in a structured way to organize objects, so that the output of one convolution layer can be used as input for the next convolution layer.

The problem in classifying this music genre is the lack of data in the dataset used for research, due to the limited spectrum image dataset (mel spectrogram) available. This is a concern

because deep learning methods require large datasets to function well. If the dataset used is too small, the model's ability to recognize complex patterns is limited, which can lead to underfitting. According to (**Shu, 2019**) models such as VGG-16, VGG-19, Inception V3, and InceptionResNet V2 can still be used effectively on limited datasets. This can be achieved by selecting the right function and making modifications such as data addition, transfer learning, dropout, and fine-tuning. These approaches help produce high-quality models and reduce the risk of underfitting even with limited datasets.

This research aims to measure the accuracy of the CNN method with the Inception V3 architecture in color classification in a music genre classification system. This research uses an automatically extracted spectrum dataset (mel spectrogram). This dataset consists of 1000 images which have been divided into training, validation and testing folders.

2. METHODOLOGY

2.1 Mel Spectrogram

The word 'mel' itself comes from the word melody, which means the perception of notes that are judged by the listener to be the same distance from each other. This is often a perceptual scale of pitches that rise to pitch separations sound similarly far off to the audience. (**Robert, 2024**)The reference point between this scale and normal frequency measurements is determined by setting a 1000 mel tone converted to a 1000hz tone. 40dB above the listener's threshold. Above about 500hz, larger intervals are judged by the listener to produce the same increase in pitch. **(Minh Tuan Nguyen, 2022)**

In 1937, Stevens, Volkmann, and Newmann proposed pitch units that began in such a way that equally spaced tones sounded equally distant to the listener **(Stevens, 1937**)**.** This is called the Mel scale. The mel spectrogram is a variation of the spectrogram commonly used in speech processing and machine learning tasks. Mel spectrograms are similar to spectrograms in that they show the frequency content of an audio signal over time, but at different frequencies.

In the modern world, when music tracks are growing rapidly both online and offline, genre classification is crucial. We must appropriately index these to improve access to them. Automatic music genre classification is important to obtain music from a large collection **(S Vishnupriya, 2018)**

With time on the x-axis and frequency on the y-axis, a spectrogram is a two-dimensional representation of a signal. The magnitude of a certain frequency inside a specified time interval is quantified using a colormap. **(Bahuleyan, 2018**)

Massive experiments demonstrate that the Mel spectrum is better suited to human auditory perception, exhibiting a linear distribution below 1000 Hz and logarithmic growth above 1000 Hz. We use this point to derive the Log-Mel spectrum static. (**Hao Meng, 2019**)

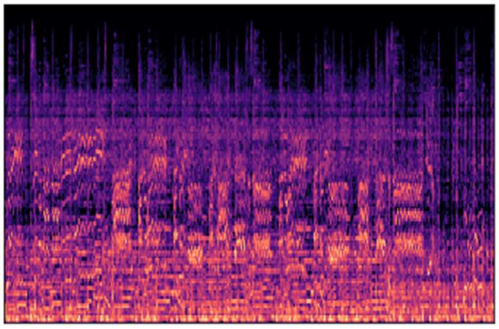


Figure 2. Example of Mel-Spectrogram

2.2 STFT (*Short-Time Fourier Transform*)

In this research, a flow diagram is needed for a general overview as follows. Short-Time Fourier Transform is a signal analysis method used to analyze how the amplitude and frequency of a signal vary or change with time. The basic principle of STFT (Short-Time Fourier Transform) is to divide the signal into shorter time segments and apply the Fourier Transform to each time segment.

The Short Time Fourier Transform (STFT) is a general linear transform. This transformation uses a sliding window with invariant length, which is defined in Equation 1 as,

$$STFT(t,f) = \int -\infty\infty s(\tau)h * (\tau - t)e - j2\pi f\tau d \quad (1)$$

where t and f are time and frequency respectively. S(t) is considered a signal; h(t) denotes the window function, and "$*$" stands for conjugate.

2.3 *Inception* V3

Deep Learning is a sub-field of machine learning that abuses different adaptable and stackable manufactured neural systems arrangements that learn progressive layer of concepts. **(Saha, 2024)**

Deep learning is commonly utilized for sound classification in numerous distinctive spaces. A commonplace approach is to change over the sound record into an picture, such a spectrogram, and utilize a profound neural organize to handle that picture. **(Boddapati, 2017**)

Inception V3 is a Convolutional Neural Network (CNN) architecture that is capable of recognizing images, including image classification. It integrates multiple convolutions with different kernel sizes to efficiently extract features. Convolutonal layer is the center square of a CNN comprising of lters (or bits) to identify dierent sorts of highlights from the input and pass them forward. **(Shu, 2019**) Additionally, Inception V3 uses regulation and reduction techniques to prevent overfitting. This deep and complex architecture has been developed. **(Kapa, 2022)** *Inception V3 is a pre-trained model for classification tasks and is an improvement over Inception V2 with improvements in the use of smaller convolutions and a reduced grid size to reduce the required computation*. **(Murni, 2023)**

A Convolutional Neural Organize (CNN) may be a deep learning engineering which is able to memorize picture classification from crude picture information, utilizing representation

learning. Its architecture is closely resembling to the network design of the neurons within the human brain, being motivated by the organization of the visual cortex. **(Nicholson, 2019)**

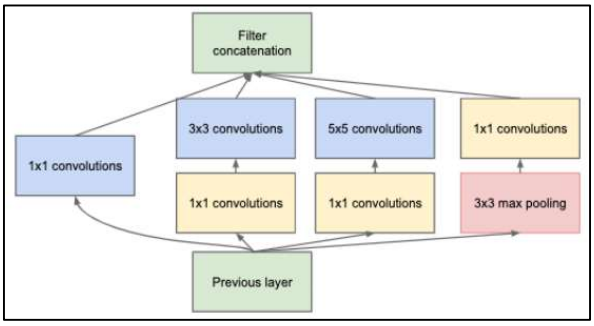


Figure 3. Inception V3 Model (Andrew & Santoso, 2022)

2.4 Diagram Block

In this research, a block diagram is needed for a general overview as follows:

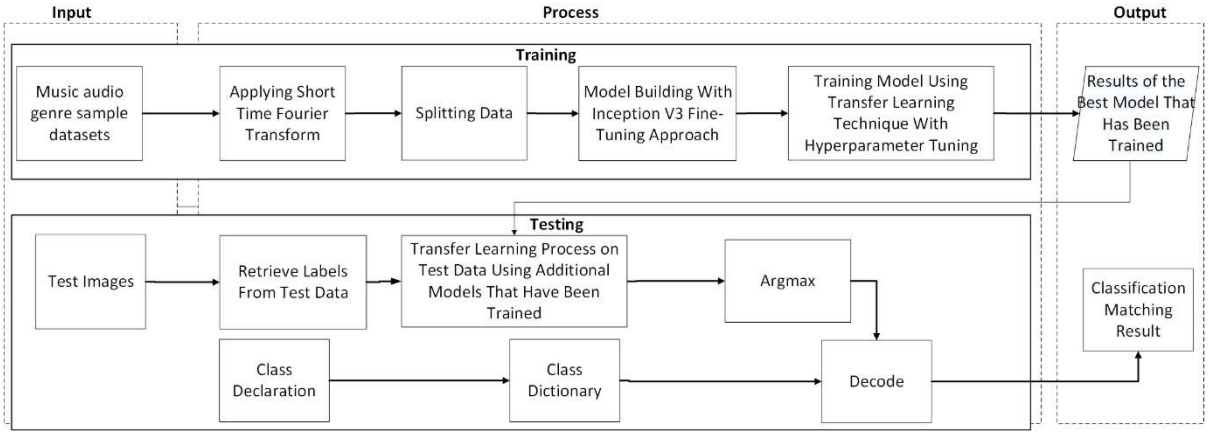


Figure 4. Block Diagram

Figure 4 shows a block diagram of the Prototype Network method, which consists of two stages: training and testing. The following is an explanation of the training steps in the system block diagram:

1. Input: The dataset contains images along with labels for each image, specifically images from each genre which are categorized into 10 different classes.

2. Process: The process begins with pre-processing the raw data, applying a Short-Time Fourier Transform to the audio broken down into 30-second chunks, producing a mel spectrogram of the converted audio. Then, data separation will be necessary because the data is divided into 3 parts, namely training 80, validation 9, and testing 10. Next, model transformation using the Inpception V3 fine-tuning approach, and the next step is training the model using transfer learning techniques with hyperparameter tuning.

3. Output: Training will produce the best model from each iteration.

The following is an explanation of the testing steps in the system block diagram:

1. Input: The dataset contains images along with labels for each image, specifically images from each genre which are categorized into 10 different classes.

2. Process: The process begins by taking labels from the testing data, then the transfer learning process begins by adding models from the models that have been trained in the training process used, next is ArgMax. Before proceeding to the next step, namely decode, the process will initiate the class first, then proceed to the class dictionary, and then decode.

3. Output: Classification results are generated.

2.5 Classification System Flowchart

In Figure 4 which visualizes how the whole system works, there are several processes and sub-processes. Figure 5 is a system classification flow diagram.

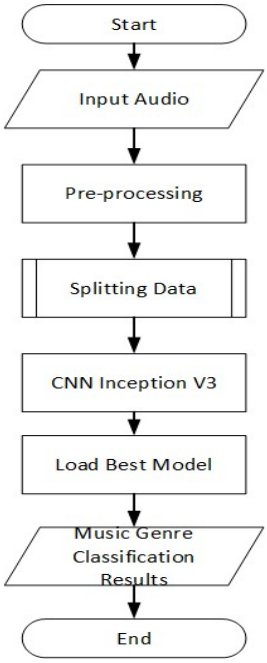


Figure 5. Classification System Flowchart

The following is an explanation of the flow diagram of the entire classification system

1. Input: The input is an image resulting from the audio conversion into a spectrogram that will be classified.

2. Preprocessing: At this stage, the process of changing and adjusting the image size to 299x299x3 is carried out, then converted into tensor values and normalized with a distance of 0 to 1.

3. Splitting Data: At this stage, the number of images for each class is equalized with 80 for training, then 9 for validation (equated to 9 because there is 1 error file in the jazz class in the jazz00054.wav file), and 9 for testing.

4. CNN Inception V3: At this stage, the feature extraction process is carried out using the Inception V3 method with transfer learning techniques and a fine-tuning approach. Additionally, layers 0 to 248 are frozen and the last 62 layers (249-311) are retrained.

5. Load Best Model: Load the best trained model to simulate testing.

3. RESULTS

3.1 Dataset Usage

The dataset used was collected from kaggle GTZAN Dataset Music Genre Classification. This dataset consists of a total of 1000 song samples which are divided into 10 classes, namely blues, classical, country, disco, hip-hop, jazz, metal, pop reggae, rock. This dataset distribution highlights limitations due to the need for more data, as deep learning requires larger datasets.

Table 1. Dataset Spliting

No.	Class	Training	Validation	Test
1	Blues	80	9	10
2	Classical	80	9	10
3	Country	80	9	10
4	Disco	80	9	10
5	Hiphop	80	9	10
6	Jazz	80	9	10
7	Metal	80	9	10
8	Pop	80	9	10
9	Reggae	80	9	10
10	Rock	80	9	10
	Total	800	90	100

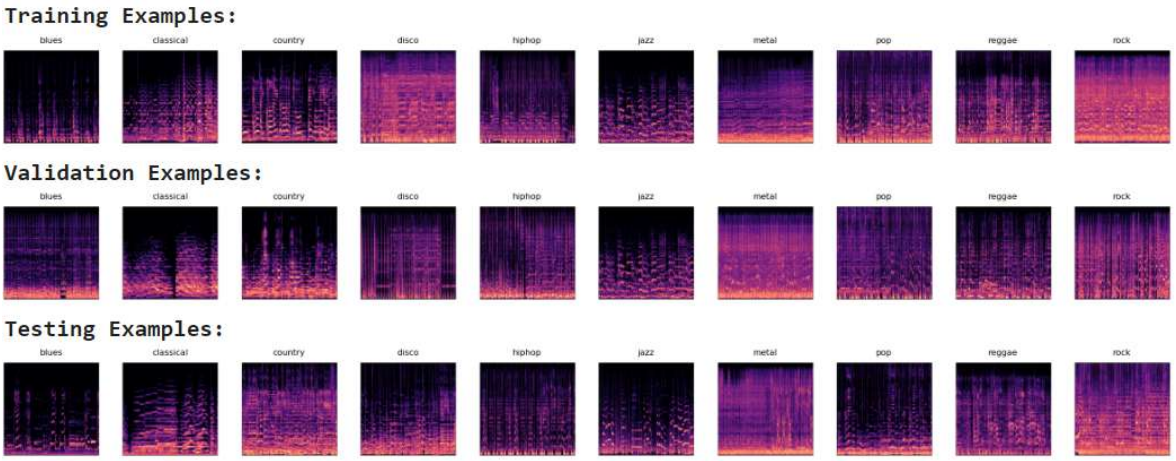


Figure 6. Training, Validation, Testing Dataset Samples

3.2 System Performance Testing

System performance testing is carried out to measure the results. The dataset used consists of mel spectrograms from each music genre. In testing the CNN Inception V3 model for 10 epochs with several parameter settings such as SGD optimizer, batch size 32, and learning rate 0.000001, the best accuracy value was obtained at 76% with a loss value of 0.9604. Figure 10 shows a graph between training and validation accuracy and training and validation loss. The test results are presented in Table 2.

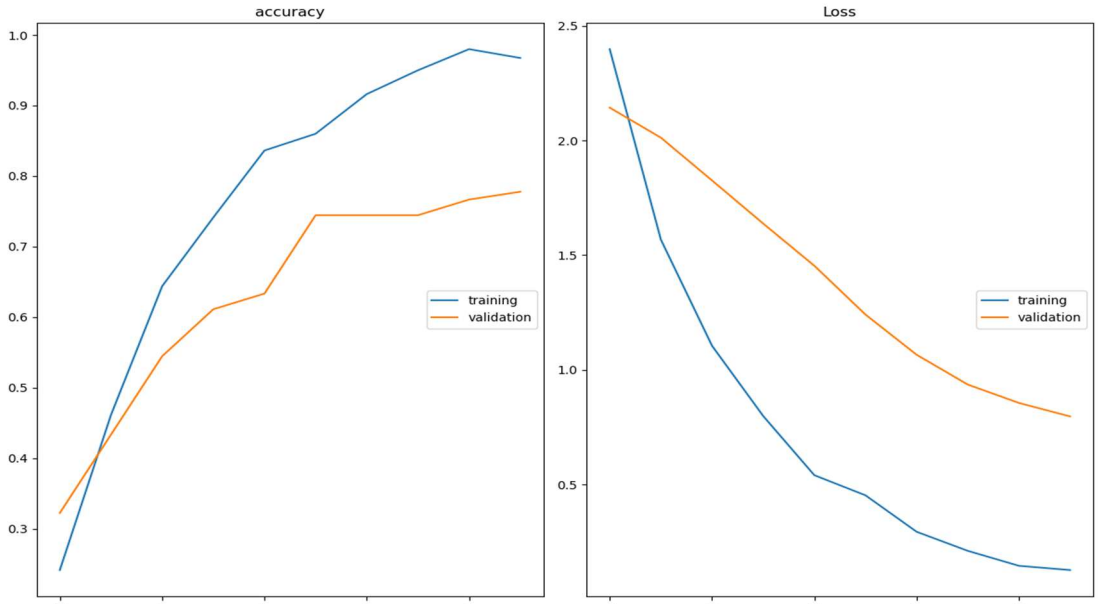


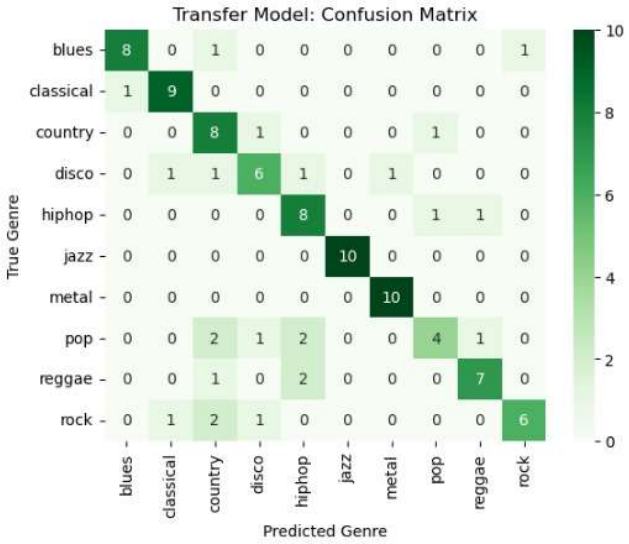
Figure 7. Graph of Accuracy and Loss Results in Training & Validation (10 epoch)

Table 2. Testing Results

No	Batch Size	Learning Rate	Epoch	Loss	Accuracy
1	32	0.000001	10	0,9604	76%
2	32	0.000001	15	1,0285	73%
3	32	0.000001	30	1,1152	65%

3.3 Inception V3 System Test Results

Figure 8 is a visualization of the confusion matrix representation of model performance carried out on test data to display actual and predicted data. The prediction results obtained by the model in 10 epochs are, in the blues genre class there are 8 correct predictions and 1 incorrect prediction, then in the classical genre there are 9 correct predictions and 2 incorrect predictions, then in the country genre there are 8 correct predictions and 6 incorrect predictions, then in the disco genre there are 6 correct predictions and 3 incorrect predictions, then in the hip-hop genre there are 8 correct predictions and 5 incorrect predictions. then in the jazz genre there are 10 correct predictions and 0 incorrect predictions, then in the metal genre there are 10 correct predictions and 1 incorrect prediction, then in the jazz genre there are 9 correct predictions and 1 incorrect prediction, then in the metal genre there are 10 correct predictions and 1 incorrect prediction, then in the pop genre there are 4 correct predictions and 2 incorrect predictions, then in the reggae genre there are 7 correct predictions and 2 incorrect predictions, and finally in the rock genre there are 6 correct predictions and 1 incorrect prediction.




```
Transfer Model
Training Accuracy: 99.875%
Train Loss: 0.0373

Validation Accuracy: 76.6667%
Test Loss: 0.7409

Testing Accuracy: 76.0%
Test Loss: 0.942
```

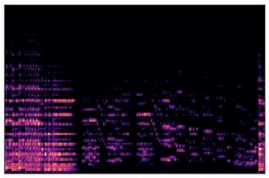
Figure 8. Confusion Matrix Test Results

3.4 System Performance Testing

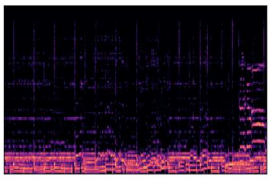
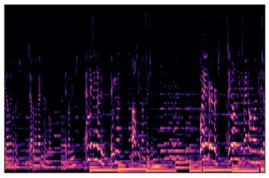
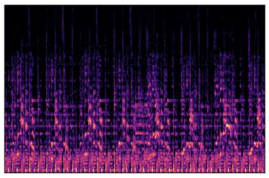
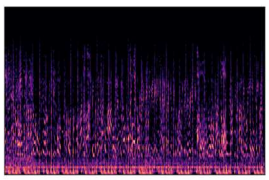
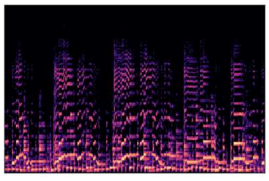
At the testing stage, classification prediction results are obtained. After training for 10 epochs, the best weights of the model are obtained. These weights are then loaded for testing as a final classification prediction result. Table 3 below shows the test results for data from each genre.

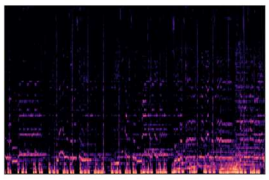
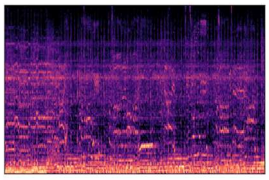
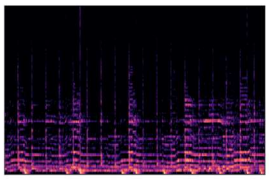
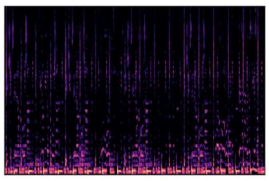
Table 3. Classification Test Results

No	Model	Target Class	Figure	Classification Result

1	*Inception V3*	Classical	`1/1 [==============================] - 1s`
			`Predicted genre: classical`
			`Actual genre: classical`
			`Filepath: classical/classical.00026.wav`
			 
				Classical

2	*Inception V3*	Blues	`1/1 [==============================] - 1s`
			`Predicted genre: blues`
			`Actual genre: blues`
			`Filepath: blues/blues.00047.wav`
				Blues

				
3	*Inception V3*	Country	1/1 [==============================] - 2s **Predicted genre:** country **Actual genre:** country **Filepath:** country/country.00089.wav 	Country
4	*Incpetion V3*	Disco	1/1 [==============================] - 1s **Predicted genre:** disco **Actual genre:** disco **Filepath:** disco/disco.00045.wav 	Disco
5	*Inception V3*	Hiphop	1/1 [==============================] - 1s **Predicted genre:** hiphop **Actual genre:** hiphop **Filepath:** hiphop/hiphop.00096.wav 	Hiphop
6	*Inception V3*	Jazz	1/1 [==============================] - 2s **Predicted genre:** jazz **Actual genre:** jazz **Filepath:** jazz/jazz.00006.wav 	Jazz
7	*Inception V3*	Pop	1/1 [==============================] - 1s **Predicted genre:** pop **Actual genre:** pop **Filepath:** pop/pop.00000.wav	Pop

				
8	*Inception V3*	Metal	1/1 [==============================] - 1s Predicted genre: metal Actual genre: metal Filepath: metal/metal.00001.wav 	Metal
9	*Inception V3*	Rock	1/1 [==============================] - 1s Predicted genre: rock Actual genre: rock Filepath: rock/rock.00009.wav 	Rock
10	*Incpetion V3*	Reggae	1/1 [==============================] - 1s Predicted genre: reggae Actual genre: reggae Filepath: reggae/reggae.00071.wav 	Reggae

3.4 Model Classification Results

The best accuracy was achieved by training the model using 10 epochs, batch size 32, SGD optimizer, and learning rate 0.00001 compared to other tests. The number of epochs and other parameters can affect the accuracy of the results. This test uses a transfer learning process with a fine-tuning approach, by freezing the first 248 layers of the Inception V3 architecture and retraining the last 63 layers. The choice of the number of layers to be frozen and retrained aims to maintain the general features that already exist in the initial layers so that the model can adapt to the classification of music genres.

There are 100 test data images, which are then divided into 10 of each class (genre).

1. Genre Blues

There were 8 correct predictions and 1 incorrect prediction, where in the classical genre there was 1 error.

$$Precision = \frac{8}{8+1+0} = 0{,}89$$

$$Recall = \frac{8}{8+2+0} = 0.80$$

$$f1 - score = 2x\frac{0.712}{1.69} = 0.44$$

2. Genre Classical

There were 9 correct predictions and 2 incorrect predictions, where in the rock genre there was 1 error, and in the disco class there was 1 error.

$$Precision = \frac{9}{9+2+0} = 0{,}81$$

$$Recall = \frac{9}{9+1+0} = 0.90$$

$$f1 - score = 2x\frac{0.729}{1.71} = 0.85$$

3. Genre Country

There are 8 correct predictions and 7 incorrect predictions, where in the blues genre there is 1 error, in the disco genre class 1 error, in the pop genre class there are 2 errors, in the reggae genre class there is 1 error, and in the rock genre class there are 2 errors.

$$Precision = \frac{8}{8+7+0} = 0{,}53$$

$$Recall = \frac{8}{8+2+0} = 0.80$$

$$f1 - score = 2x\frac{0.424}{1.33} = 0.63$$

4. Genre Disco

There were 6 correct predictions and 3 incorrect predictions, where in the country genre class there was 1 error, in the rock genre class there were 2 errors and in the pop genre class there was 1 error.

$$Precision = \frac{6}{6+3+0} = 0{,}66$$

$$Recall = \frac{6}{6+4+0} = 0.60$$

$$f1 - score = 2x\frac{0.396}{1.26} = 0.62$$

5. Genre Hiphop

   There were 8 correct predictions and 5 incorrect predictions, where in the disco genre there was 1 error, in the pop genre class 2 errors, in the reggae genre class there were 2 errors.

$$Precision = \frac{8}{8+5+0} = 0{,}61$$

$$Recall = \frac{8}{8+2+0} = 0.80$$

$$f1 - score = 2x\frac{0.488}{1.41} = 0.69$$

6. Genre Jazz

   There are 10 correct predictions and 0 incorrect predictions.

$$Precision = \frac{1}{10+0+0} = 1$$

$$Recall = \frac{10}{10+10+0} = 1$$

$$f1 - score = 2x\frac{1}{2} = 1$$

7. Genre Metal

   There were 10 correct predictions and 1 incorrect prediction, where in the disco genre there was 1 error.

$$Precision = \frac{10}{10+1+0} = 0{,}90$$

$$Recall = \frac{10}{10+0+0} = 1$$

$$f1 - score = 2x\frac{0.90}{1.90} = 0.94$$

8. Genre Pop

   There are 4 correct predictions and 2 incorrect predictions, where in the hiphop genre there is 1 error, and in the country genre there is 1 error.

$$Precision = \frac{4}{4+2+0} = 0{,}66$$

$$Recall = \frac{4}{4+6+0} = 0.40$$

$$f1 - score = 2x\frac{0.264}{1.06} = 0.49$$

9. Genre Reggae

There were 7 correct predictions and 2 incorrect predictions, where in the hip-hop genre there was 1 error, and in the pop genre class there was 1 error.

$$Precision = \frac{7}{7+2+0} = 0{,}77$$

$$Recall = \frac{7}{7+3+0} = 0.70$$

$$f1 - score = 2x\frac{0.539}{1.47} = 0.73$$

10. Genre Rock

There were 6 correct predictions and 1 incorrect prediction, where in the blues genre there was 1 error.

$$Precision = \frac{6}{6+1+0} = 0{,}85$$

$$Recall = \frac{6}{6+4+0} = 0.60$$

$$f1 - score = 2x\frac{0.51}{1.45} = 0.70$$

Table 4 shows the model classification results for each class (genre):

Table 4. Classification Model Result

Class	Precision	Recall	F1-Score
Classical	0.81	0.90	0.85
Blues	0.89	0.80	0.84
Country	0.53	0.80	0.63
Disco	0.66	0.60	0.62
Hiphop	0.61	0.80	0.69
Jazz	1.00	1.00	1.00
Pop	0.66	0.40	0.49
Metal	0.90	1.00	0.94
Rock	0.85	0.60	0.70
Reggae	0.77	0.70	0.73

Based on the table above shows that the jazz genre has the highest level of accuracy, it is because of the characteristics of the jazz genre itself which has elements of distinctive guitar instruments, saxophone, distinctive drum beats. On the other hand, the pop genre is a genre that gets the lowest accuracy value because the characteristics of pop itself are almost difficult to classify, because basically the pop genre comes from the word 'popular' which means it does not have identical characteristics and is attached to the genre itself.

4. CONCLUSIONS

From the research and writing above, it can be concluded that in this research, the implementation of the CNN Inception V3 architecture method with a fine-tuning approach using transfer learning techniques succeeded in providing quite good results aimed at classifying music genres. The classification results on CNN Inception V3 testing data recorded several correct predictions and several incorrect ones. In this test there are 10 classes, namely blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock. It is proven that the genre that is most difficult to study with this architecture is the pop genre, because if you examine it, the pop genre is an abbreviation for popular, or refers to music or songs that are currently popular and have less distinctive characteristics than other genres. From this research, using 1000 image datasets, with a batch size of 32, epoch 10, learning rate 0.000001, the SGD optimizer produces the best hyperparameter preset results by achieving training accuracy of 99.625%, validation 78.8889%, and testing 76,000%.

REFERENCES

Arti, Y. (2023). Face Spoofing Detection using Inception-v3 on RGB Modal and Depth Modal. *Jurnal Ilmu Komputer dan Informasi*, 47-57.

Arrofiqoh, H. (2018). Implementasi Metode Convolutional Neural Network Untuk Klasifikasi Tanaman Pada Citra Resolusi Tinggi. *GEOMATIKA*, *24*(2), 61.

Bahuleyan, H. (2018). Music Genre Classification using Machine Learning Techniques. 3.

Boddapati, V. e. (2017). Classifying environmental sound using image recognition networks. *Procedia Comput*, 112.

Hao Meng, T. Y. (2019). Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network. *IEEE Access*, *7*.

Kapa, M. R. (2022). Klasifikasi Citra Penyakit Leukemia Menggunakan Convolutional Neural Network Dengan Arsitektur Inception-V3. 129.

Minh Tuan Nguyen, W. W. (2022). Heart Sound Classification Using Deep Learning Techniques Based on Log-mel Spectrogram. *Circuits, Systems, and Signal Processing*.

Murni, A. (2023). Face Spoofing Detection using Inception-v3 on RGB Modal and Depth Modal. . *Jurnal Ilmu Komputer dan Informasi*, 47-57.

Nicholson, C. (2019). *A Beginner's Guide to LSTMs and Recurrent Neural Networks*. Diambil kembali dari https://pathmind.com/wiki/lstm

Robert, L. (2024). *Understanding the Mel Spectrogram*. Diambil kembali dari Analytics Vidhya: https://medium.com/analytics-vidhya/understanding-the-melspectrogram-fca2afa2ce53

S Vishnupriya, K. M. (2018). Automatic Music Genre Classification using Convolution Neural Network. *2018 International Conference on Computer Communication and Informatics (ICCCI)*.

Shu, M. (2019). Deep learning for image classification on very small datasets using transfer learning. *Semantic Scholar*, 3-4.

Stevens, S. S. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, *8*, 185–190.

Szegedy, V. I. (2016). *Rethinking the Inception Architecture for Computer Vision.*