IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 3, March, 2025, pp. 840-850
Journal homepage: http://iieta.org/journals/mmep

Classification of Indonesian Music Genres Using Transfer Learning with ResNet-50 and Mel-Frequency Cepstral Coefficient Feature Extraction

Yudha Alif Auliya*, Dwiretno Istiyadi Swasono, Perdana Putro Harwanto

Faculty of Computer Science, University of Jember, Jember 68121, Indonesia

Corresponding Author Email: yudha.alif@unej.ac.id

https://doi.org/10.18280/mmep.120310

Received: 6 January 2025
Revised: 4 March 2025
Accepted: 10 March 2025
Available online: 31 March 2025

Keywords:
deep learning, genre classification, CNNs, MFCCs, ResNet-50

ABSTRACT

Indonesian music features diverse genres such as pop, dangdut, keroncong, and campursari, each with distinct characteristics and audiences. The increasing digitalization of music necessitates automated systems for accurate genre classification. This study develops a deep learning-based classification system using Convolutional Neural Networks (CNNs), specifically ResNet-50 and VGG-16, with Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction. MFCCs effectively capture music's frequency spectrum, making them suitable for genre classification. Experimental results show that ResNet-50 outperforms VGG-16, achieving 99% accuracy, 98% precision, 98% recall, and a 98% F1-score with an 80:20 data split. ResNet-50's residual connections enable better feature learning and mitigate gradient vanishing, leading to superior performance. VGG-16 also performed well but exhibited slightly lower accuracy due to its deeper structure without residual connections. The study emphasizes the impact of dataset size, showing that a larger training set improves generalization. However, limitations such as the relatively small dataset and the exclusive use of MFCCs may affect performance on unseen data. Future research should explore larger datasets, additional Indonesian genres, and hybrid feature extraction techniques.

1. INTRODUCTION

The rapid proliferation of digital music platforms has presented substantial obstacles in the organization and retrieval of music, particularly in the context of Indonesia's diverse musical traditions. Music genre classification, a critical component of music information retrieval, is essential for the organization of large-scale digital music archives and the improvement of recommendation systems [1]. The intricate blend of traditional and modern musical elements in Indonesia's rich musical heritage, which includes genres such as pop, dangdut, keroncong, and campursari, presents unique challenges for automated classification systems.

The audio signal processing and classification tasks have shown promising results as a result of recent advancements in artificial intelligence (AI), particularly deep learning frameworks. CNNs have emerged as a potent technique, demonstrating exceptional performance in a variety of signal processing applications, such as audio classification [2]. CNNs are particularly effective in the classification of music genres because they autonomously learn hierarchical features from raw audio data, thereby eliminating the necessity for manual feature extraction [3]. Convolutional layers and pooling layers are indispensable components of a CNN. The pooling layers down sample the feature maps to reduce computational complexity, while the convolutional layers employ learnable weights to identify complex patterns in the data. These components are organized in a deep network architecture, which allows CNNs to effectively process and categorize intricate audio signals [4].

MFCCs are one of the most frequently employed feature extraction techniques in music classification. They have consistently achieved accuracy rates exceeding 90% in high-quality audio recordings [5]. MFCCs are particularly well-suited for music genre classification tasks due to their capacity to replicate human auditory perception through Mel-scale attributes, which is the reason for their effectiveness [6]. Diverse methodologies for music genre classification have been investigated in numerous prior investigations. A number of studies have examined the use of MFCC-based feature extraction in machine learning methods, including Gaussian Processes, Deep Neural Networks, and Metric Learning. These studies have achieved accuracy rates of 78.3%, 80.5%, and 76.7%, respectively [7]. MFCC and Zero-Crossing Rate (ZCR) features have been combined with Support Vector Machines (SVM) in other studies, resulting in an accuracy of 83% [8]. Accuracy rates of 66.20%, 78.28%, and 80.14% have been reported for deep learning approaches that utilize CNN architectures, including MobileNetV2, VGG-16, and ResNet-50, with MFCC features.

Nevertheless, there is a substantial gap in the application of these methods to Indonesian music genres. The majority of prior research has primarily concentrated on Western music datasets, such as GTZAN and FMA, which do not possess the

unique characteristics of Indonesian traditional and contemporary genres [9]. Indonesian music frequently displays distinctive rhythmic patterns, instrumentations, and tonal structures that are inadequately represented in benchmark datasets that are frequently employed. As a result, models that have been trained on Western music data may encounter difficulty in accurately classifying Indonesian music, resulting in suboptimal performance in real-world applications. Furthermore, the majority of current research has employed generic deep learning architectures without optimizing them for the intricacies of Indonesian music classification. While some research has explored CNN-based models, there is a dearth of comparative studies that assess multiple CNN architectures specifically for Indonesian music genres.

The advancement of music genre classification research is facilitated by the integration of audio feature extraction methods and deep learning techniques. A significant study on the GTZAN dataset utilized the CNN-based MusicRecNet model in conjunction with a variety of feature extraction techniques, such as MFCC, Spectral Centroid, and Chroma Short-Time Fourier Transform (STFT), to achieve an average accuracy of 81.8% [3]. This study made a substantial contribution to the advancement of music recommendation systems by facilitating the implementation of more personalized music recommendations that are tailored to the preferences of the user. Additionally, it offered valuable insights into the classification of music genres through the use of machine learning. The objective of this study is to create a deep learning-based classification system for Indonesian music genres, as there is a dearth of research on Indonesian music genres and the limitations of existing studies. This research aims to improve genre classification performance and make a significant contribution to the broader field of music information retrieval by utilizing CNN architectures and MFCC features [10].

A substantial study employed CNN algorithms for audio processing in music genre classification, utilizing Librosa to convert original audio files into suitable Mel spectrograms. The study achieved an accuracy of 84% [11]. This demonstrated the effectiveness of combining spectral analysis with CNN for genre classification tasks. In a comprehensive review, Abdul and Al-Talabani underscored the importance of feature extraction techniques, particularly MFCC [12]. Their research emphasized that MFCC has become a primary feature extraction method in audio-centric machine learning applications, thereby demonstrating its widespread use in the field. Previous research conducted a comparative study that employed CNN and MFCC for feature extraction across a variety of architectures, thereby further advancing the field. Their research evaluated MobileNetV2, VGGNet16, and ResNet50, resulting in accuracies of 66.20%, 78.28%, and 80.14%, respectively. This analysis offers valuable insights into the efficacy of different deep learning architectures in the classification of music genres [13].

There is a dearth of research that compares the effectiveness of different CNN architectures specifically for classifying Indonesian music genres, despite the extensive exploration of various CNN architectures. This study conducts a comparative analysis of two prominent CNN architectures, ResNet-50 and VGG-16, in conjunction with MFCC feature extraction, to classify Indonesian music genres. This study endeavors to identify the most effective training parameters and data partitioning strategies to improve classification performance by assessing the efficacy of ResNet-50 and VGG-16 architectures in conjunction with MFCC feature extraction. This research is essential for the enhancement of music recommendation systems that are specifically designed for Indonesian music, as well as for the advancement of the broader fields of audio signal processing and pattern recognition. Providing a foundation for future advancements in AI-driven music analysis, the proposed model addresses both technical difficulties and cultural considerations in music genre classification.

2. METHODOLOGY

2.1 Dataset

The custom dataset of 400 Indonesian music audio samples, which span four distinct genres: Indonesian pop, dangdut, campursari, and keroncong, is employed in this study. Each genre is represented by 100 samples. The dataset was compiled from a variety of digital music platforms, such as YouTube and online music repositories, to guarantee a fair distribution across all classes. The Waveform Audio File (WAV) format is the industry standard for uncompressed audio, and each audio sample has a duration of 30 seconds and a sampling rate of 22,050 Hz [14]. This guarantees that the audio quality is sufficient for feature-based analysis. The training dataset is more representative due to the balanced genre distribution, which enables a thorough examination of the characteristics of each genre. MFCCs are a technique that is frequently employed in the recognition of speech and audio patterns. This technique is used to extract features. The model can capture high-level feature patterns in the audio domain by utilizing a transfer learning approach that employs the ResNet-50 and VGG-16 architecture to improve classification performance.

2.2 MFCC feature extraction

The primary feature extraction technique in this study was MFCC. MFCC is extensively employed in a variety of audio processing applications, with a particular emphasis on speech and voice signal analysis, such as gender classification, speech recognition, and speaker recognition. The MFCC extraction process is comprised of a series of sequential steps, such as signal framing, power spectrum computation, applications of a Mel filter bank to the power spectra, logarithmic transformation of the filter bank outputs, and the application of the Discrete Cosine Transform (DCT) to obtain cepstral coefficients. The transformation of raw audio signals into a compact representation that captures essential spectral characteristics is facilitated by these steps. The MFCC computation process is illustrated in Figure 1 [12].

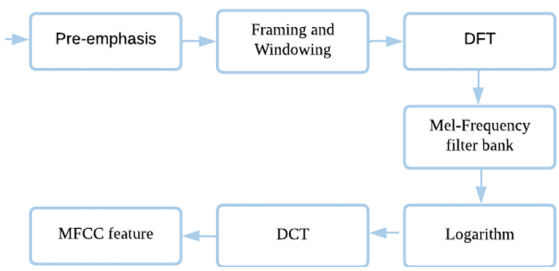


Figure 1. MFCC processed

2.2.1 Pre-emphasis

Pre-emphasis is a commonly used pre-processing technique in signal processing to mitigate high-frequency attenuation during signal production. As the initial step in MFCC adaptation, pre-emphasis applies a high-pass filter with a coefficient of [1, 0.97]. This filtering process redistributes energy across frequencies and influences the overall signal energy [12].

2.2.2 Framing and windowing

Dividing the signal into multiple frames helps partition the raw signal into smaller, more stationary segments. In speech analysis, a 20–30 ms interval is considered a Quasi-Stationary Segment (QSS), with glottal closures occurring approximately every 20 ms. Vowel sounds typically last between 40–80 ms. Spectral measurements are commonly performed in 20 ms frames with a 10 ms overlap to capture temporal characteristics. Each frame is processed using windowing functions, such as Hanning and Hamming, to reduce signal distortion and mitigate edge effects during DFT computation [12] as shown in Figure 2.

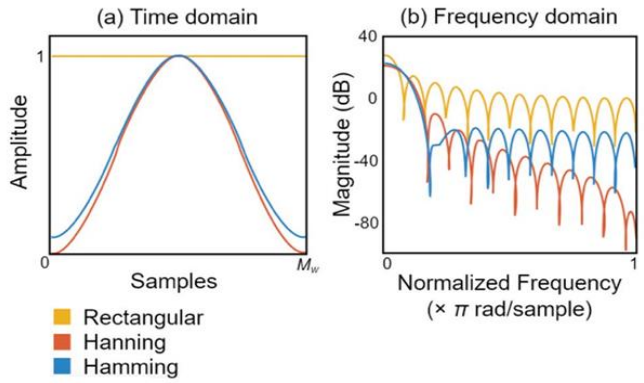


Figure 2. The rectangular Hanning and Hamming windows in the time and frequency domains

2.2.3 Power spectrum

A power spectrum represents the distribution of power across the frequency components of a signal. The Fast Fourier Transform (FFT) is an efficient algorithm developed in 1965 to convert signals from the time domain to the frequency domain, enabling the extraction of critical signal attributes. This transformation facilitates comprehensive signal analysis by revealing frequency-domain characteristics. The FFT is applied to each frame of N samples, converting them from the time domain to the frequency domain. It is particularly useful for analyzing sound wave responses and modeling the convolution of vocal fold vibrations. The power spectrum for each frame is computed using the following Eq. (1).

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{2\pi jnK}{N}}, k = 1,2,3 \ldots N-1 \quad (1)$$

2.2.4 Mel-frequency bank

The Mel band-pass filter is a filter bank constructed based on pitch perception. The Mel filter was initially created for speech analysis and, akin to the human ear's perception of speech, aims to extract a non-linear representation of the speech signal. The Mel filter bank comprises 40 triangular filters [12]. The transfer function (TF) of each m-th filter can be calculated using Eq. (2).

$$\begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \quad (2)$$

where, $f(m)$ is the centre frequency of the triangular filter and $\sum_{m}^{M-1} Hm\,(k) = 1$.

The Mel scale to the response frequency and vice versa is computed by Eqs. (3) and (4).

$$m = 2595 \log 10 \left(1 + \frac{f}{700}\right) \quad (3)$$

$$f = 700(10^{m} 2595 - 1) \quad (4)$$

2.2.5 DCT

A DCT represents a finite sequence of data points as a summation of cosine functions oscillating at various frequencies. The DCT was introduced by Nasir Ahmed in 1972. In the MFCC process, the DCT is utilized on the Mel filter bank to extract the most significant coefficients or to delineate the correlation in the log spectral magnitudes from the filter bank [12]. The DCT is calculated using the following Eq. (5).

$$X(k) = \sum_{n=0}^{N-1} x_n * \cos\left({}^{2\pi jnk}/_{N}\right), k = 1,2,3 \ldots N-1 \quad (5)$$

2.3 CNN

The name of the CNN, a prominent deep learning model, is derived from the convolution operation that is applied to matrices [15]. CNNs are constructed using Artificial Neural Networks (ANNs) that replicate the functionality of the human brain. They also include specialized layers, such as convolutional and fully connected layers with learnable parameters, as well as non-parametric layers like activation and pooling layers [16]. The CNN architecture is comprised of two primary sections, as illustrated in Figure 3. The input layer, convolutional layers, and pooling layers comprise the initial section, which is responsible for feature extraction. The output layer and fully connected layers comprise the second section, which concentrates on classification [17]. Following the final pooling layer, the fully connected layer receives the extracted features for classification.

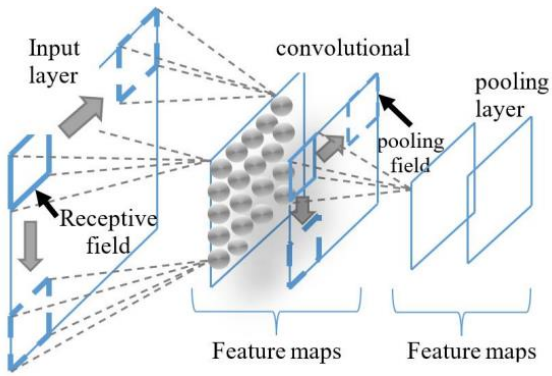


Figure 3. Structure CNN

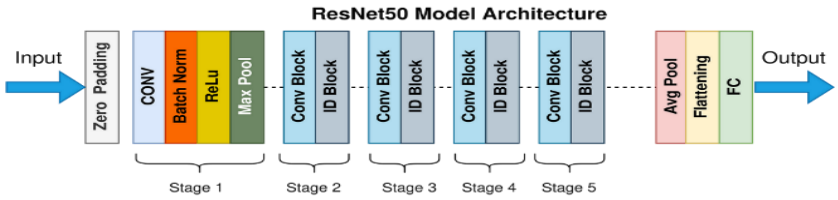


Figure 4. ResNet-50 architecture

2.4 ResNet-50

The ResNet-50 architecture is employed in this study to improve the training of deep neural networks and address the vanishing gradient issue. This architecture employs a deep residual learning framework. This section delineates the architectural components and specific implementations that are used to categorize Indonesian music genres.

The input module, four residual blocks (arranged with 3, 4, 6, and 3 layers, respectively), and an output module comprise the core architecture of ResNet-50. To enhance model adaptability and ensure training stability, each layer incorporates batch normalization and ReLU activation functions [18, 19]. The residual blocks of the network are its defining feature, as they utilize skip connections to mitigate gradient degradation during backpropagation.

In the initial convolution layer, a 7×7 convolutional filter with a stride of 2 is employed, followed by batch normalization and ReLU activation. Subsequently, a 3×3 max pooling layer with a stride of 2 reduces spatial dimensions while maintaining essential features. The architecture's foundation is composed of four sequential residual stages, each of which contains numerous bottleneck blocks. A three-layer architecture is employed by these bottleneck blocks: a 1×1 convolution for dimensionality reduction, a 3×3 convolution for feature extraction, and a subsequent 1×1 convolution for dimensionality restoration [19]. The ResNet-50 is illustrated in Figure 4.

2.5 VGG-16

The VGG-16 architecture employed in this study employs a deep convolutional neural network framework to execute classification tasks. This section delineates the architectural components and specific implementations that are employed to classify Indonesian music genres. Thirteen convolutional layers are systematically arranged in five blocks to form the core structure of VGG-16 [16]. Subsequently, three fully connected layers are employed for classification purposes. To introduce non-linearity, each convolutional layer utilizes ReLU activation functions. This is further enhanced by max-pooling layers, which effectively reduce spatial dimensions while preserving essential features [8]. This architecture is distinguished by its consistent implementation of 3×3 convolutional filters throughout the network, with each convolutional layer utilizing the ReLU activation function. A max-pooling layer with 2×2 kernel sizes is applied after each convolutional layer [20].

The architectural organization commences with the initial block, which contains two convolutional layers, each of which is equipped with 64 filters. Subsequently, a max-pooling operation is implemented. The second block retains a dual-layer structure that is like the first, but the filter count is increased to 128. The network then advances to more intricate blocks, with the third block implementing three convolutional layers with 256 filters, followed by the fourth and fifth blocks, each of which contains three layers with 512 filters [21]. Each block concludes with a max-pooling layer that employs a 2×2 window and a stride of 2, thereby expanding the capacity of the feature channel and systematically reducing spatial dimensions. The VGG-16 is depicted in Figure 5.

2.6 Confusion matrix

The confusion matrix is one of the most frequently employed tools in machine learning, and performance measurement is essential for evaluating classification methods [22]. It facilitates the computation of a variety of performance metrics by methodically organizing predicted and actual class labels in a tabular format [23]. The confusion matrix is composed of four primary outcomes: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Some of the most critical performance metrics that can be derived from these include: The first Eq. (6) calculates the accuracy (overall correctness of predictions), the second Eq. (7) calculates the precision (ratio of correct positive predictions), the third Eq. (8) calculates the recall (ability to identify positive instances), and the fourth Eq. (9) calculates the F1 Score (harmonic balance between precision and recall) [2, 3].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

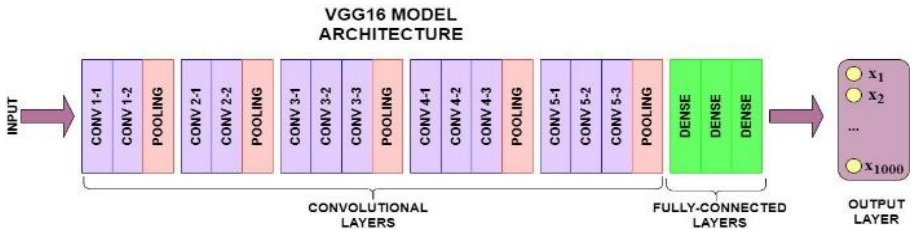


Figure 5. VGG-16 architecture [21]

3. RESULTS AND DISCUSSION

3.1 Dataset preparation

The dataset utilized in this study consists of Indonesian music samples from four distinct genres: dangdut, pop, campursari, and keroncong. Each audio sample was standardized to a 30-second duration to ensure consistency in the feature extraction process. The dataset compilation focused on maintaining representative characteristics of each genre while minimizing potential bias in the subsequent analysis.

3.2 MFCC implementation

The MFCC extraction process was implemented through several sequential stages to obtain optimal spectral representations of the audio signals. Initially, a pre-emphasis filter was applied to the audio signals using the following equation:

$$y[n] = x[n] - \alpha * x[n-1] \quad (10)$$

where, α = 0.97, to emphasize higher frequencies and improve the overall signal-to-noise ratio. Figure 6 demonstrates the signal waveform before and after pre-emphasis application.

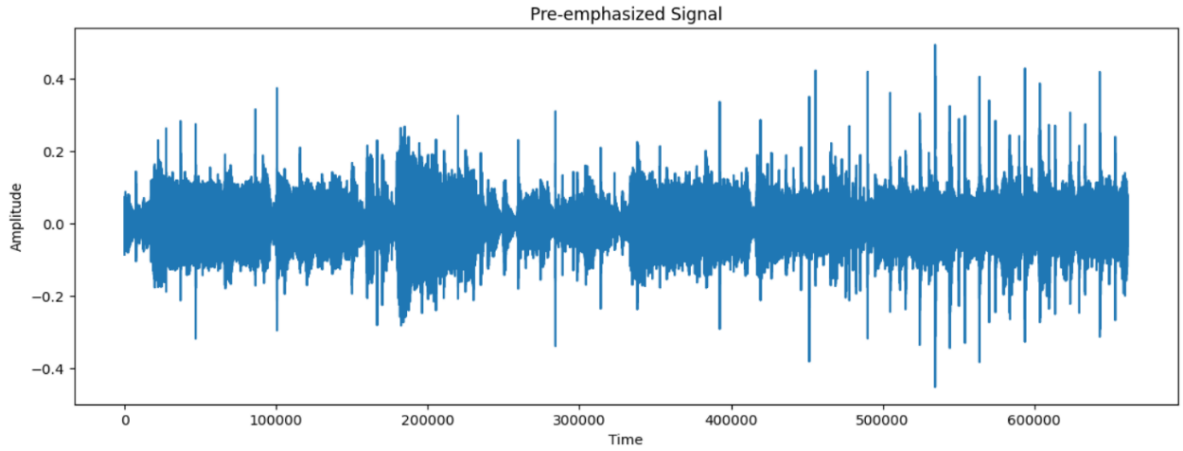
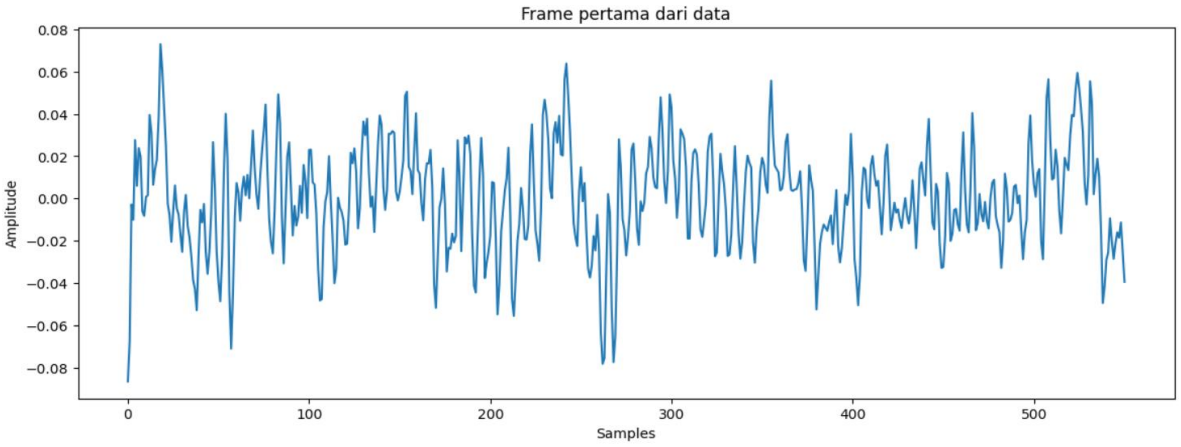
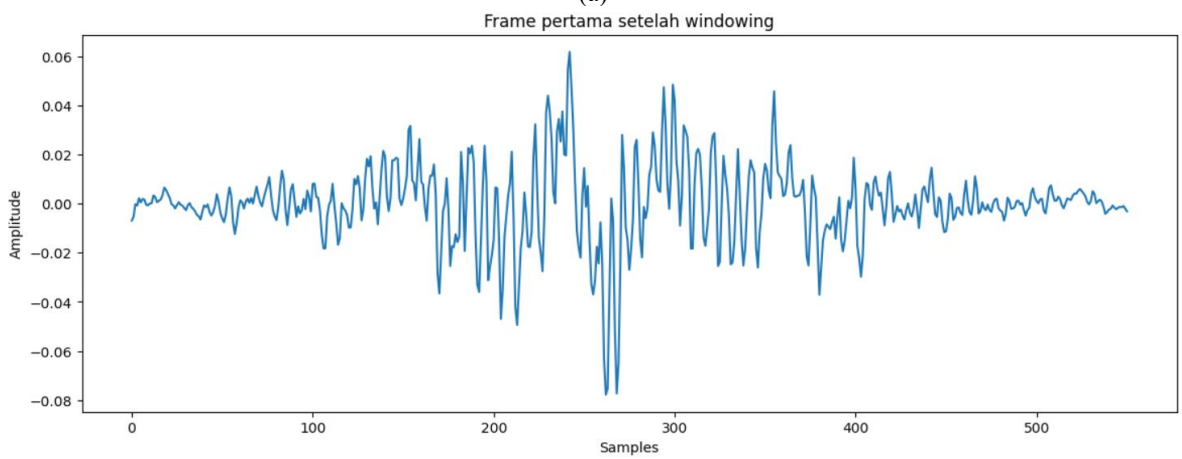


Figure 6. Audio signal after pre-emphasis process



(a)



(b)

Figure 7. Signal audio after framing and windowing

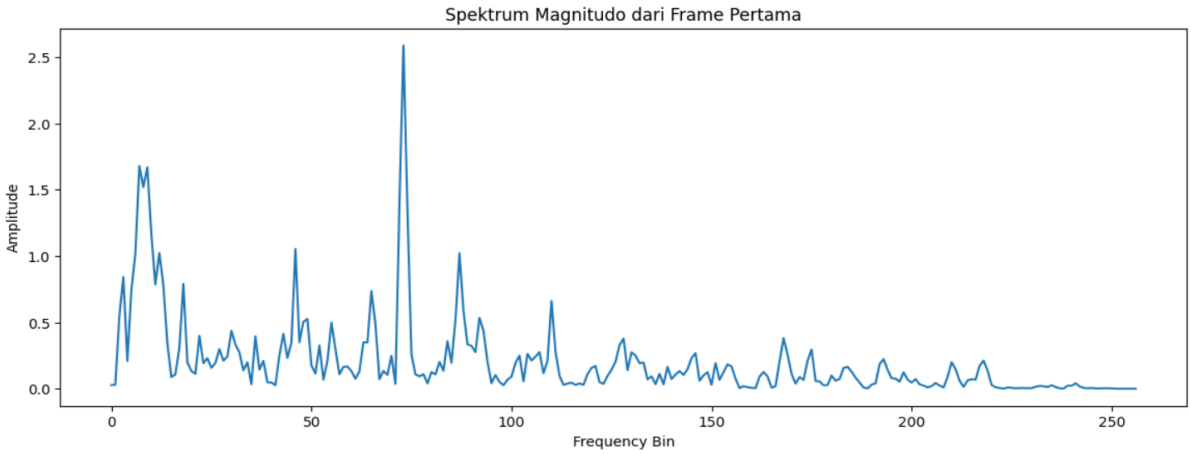


Figure 8. FFT spectrum visualization convert audio signal to frequency domain

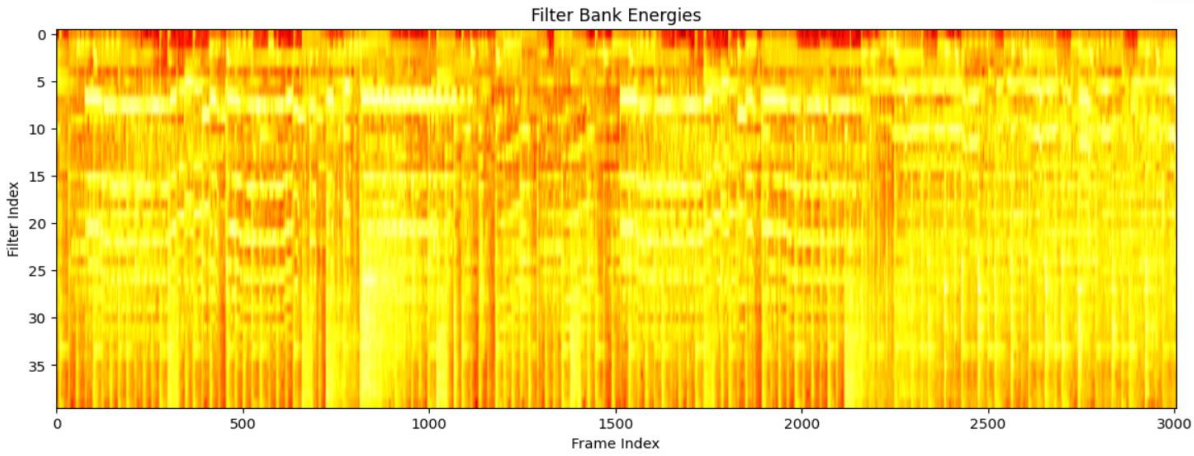


Figure 9. Mel filter bank output visualization

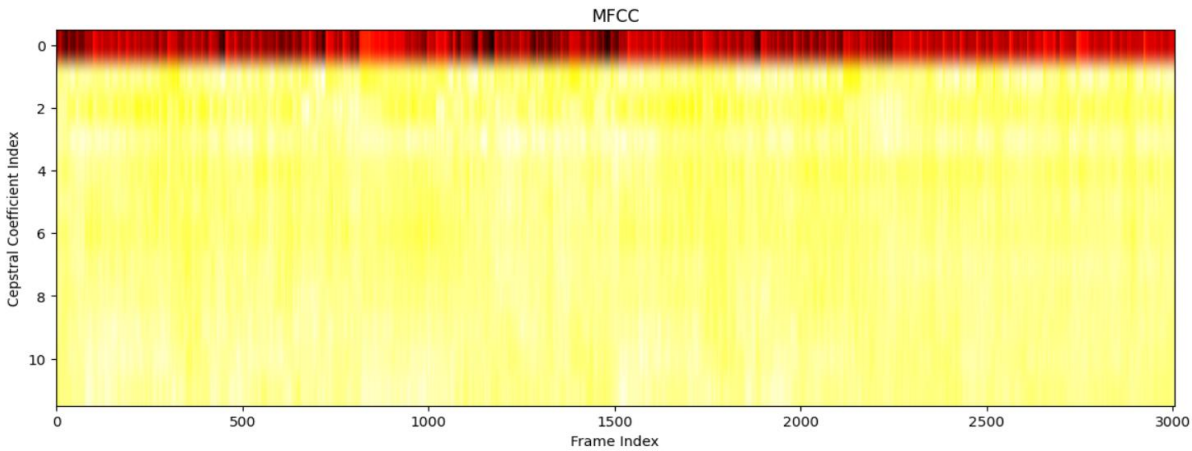


Figure 10. Visualisation feature extraction MFCC

The pre-emphasized signal was segmented into overlapping frames of 25ms with a 10ms hop length to capture temporal variations effectively. A Hamming window function was applied to each frame to minimize spectral leakage, as illustrated in Figure 7.

The windowed frames were transformed into the frequency domain using FFT, producing the power spectrum. This transformation revealed the frequency components present in each frame, as shown in Figure 8.

A set of triangular Mel-scale filter banks was applied to the power spectrum, mapping the frequencies to the Mel scale, which better approximates human auditory perception. Logarithmic compression was performed on the Mel-spectrum to match human perception of loudness. The resulting Mel-spectrum is visualized in Figure 9.

Finally, DCT was applied to obtain the final MFCC features, decorrelating the filterbank energies and producing a compact representation of the spectral envelope. The resulting MFCC coefficients are visualized in Figure 10.

In this research, the extraction features were carried out using the librosa library in each genre. The following Figure 11 is the visualization results of MFCC extraction features in each genre.

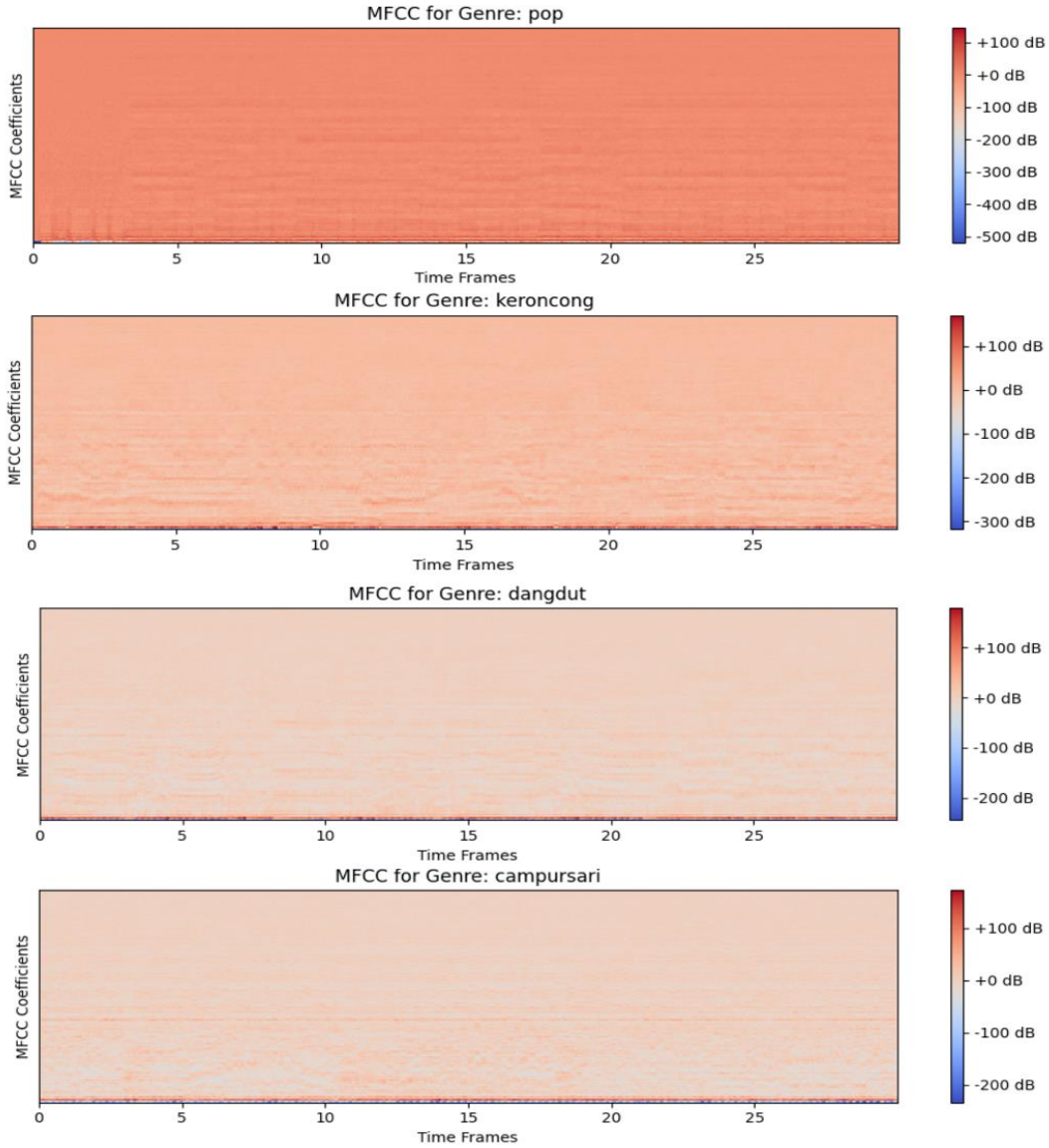


Figure 11. Result MFCC feature extraction

Table 1. Data splitting scenarios

Data Train	Data Test	Count Data
80%	20%	Data Train: 320 Data Test: 80
70%	30%	Data Train: 280 Data Test: 120
60%	40%	Data Train: 240 Data Test: 160

3.3 Train model

The Mel-Frequency Cepstral Coefficients (MFCC) feature extraction results were utilized to train two deep learning architectures: ResNet-50 and VGG-16. These architectures were both customized for the classification of Indonesian music genres. In order to enhance performance, the final fully connected layers were restructured to accommodate the genre classification task, and a Softmax activation function was implemented to generate class probabilities. Furthermore, in order to reduce overfitting, dropout regularization (rate = 0.5) was implemented prior to the final classification layer.

Transfer learning was employed to initialize both architectures with pre-trained weights from ImageNet during the training process. Fine-tuning was performed with the Adam optimizer (learning rate = 0.0001, weight decay = 0.0001), categorical cross-entropy loss function, and a batch size of 32, utilizing standardized hyperparameters. Training was conducted over a period of 20 epochs, with early terminations implemented to prevent overfitting. Data augmentation techniques, including pitch shifting, noise injection, and time stretching, were implemented on the input audio signals to improve model generalization. The model's performance was comprehensively assessed across varying data distributions using three train-test split configurations: 80:20, 70:30, and 60:40. This method facilitated a comprehensive evaluation of the effectiveness of classification in relation to the quantity of training data. Table 1 provides a comprehensive summary of the dataset split configurations.

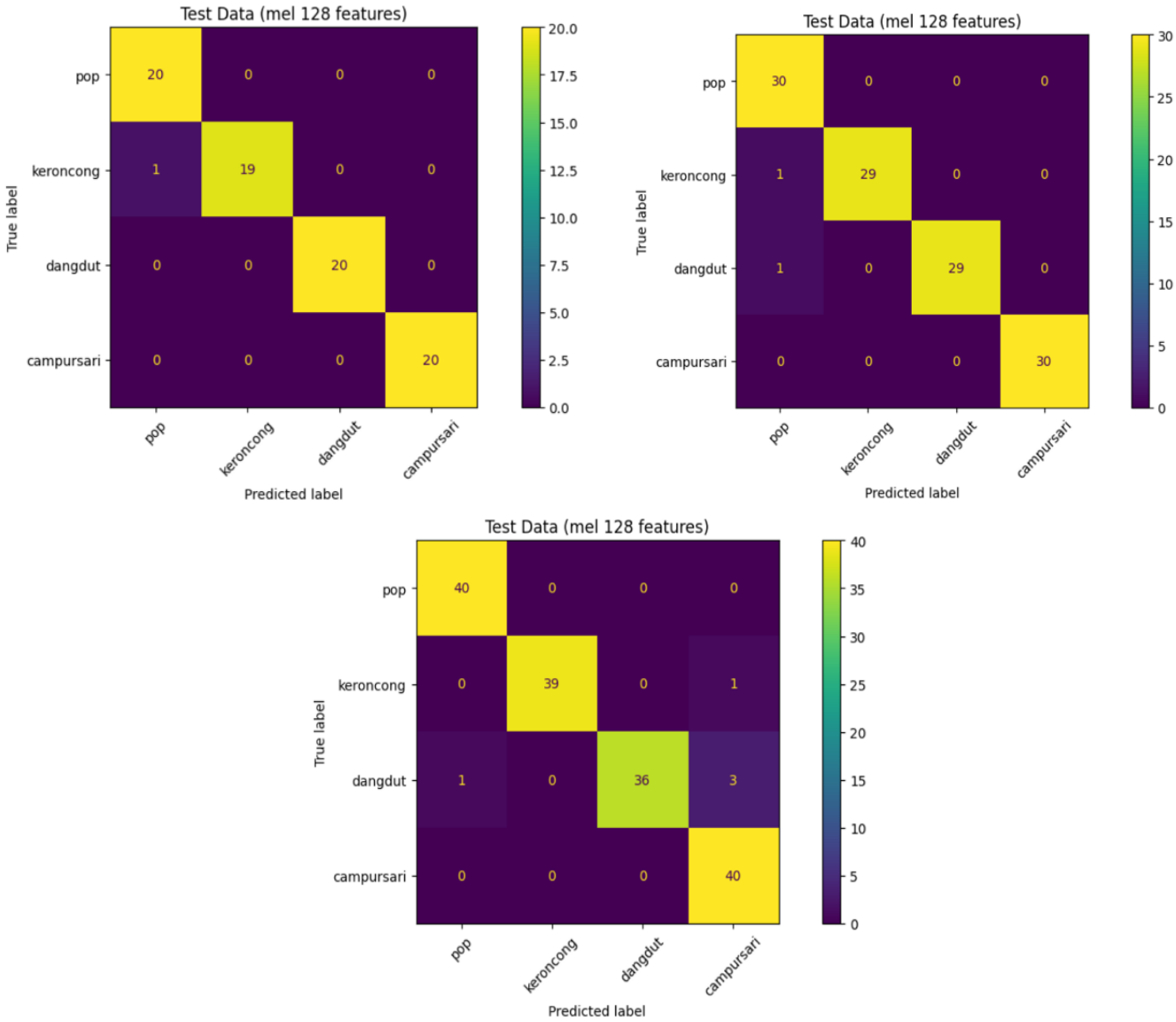
3.4 Classification result

The models' classification performance was comprehensively assessed across various dataset split scenarios using confusion matrices. In-depth insights into the classification behavior of the models under various training-testing distributions are provided by these confusion matrices, which are based on a total dataset of 400 samples. ResNet-50

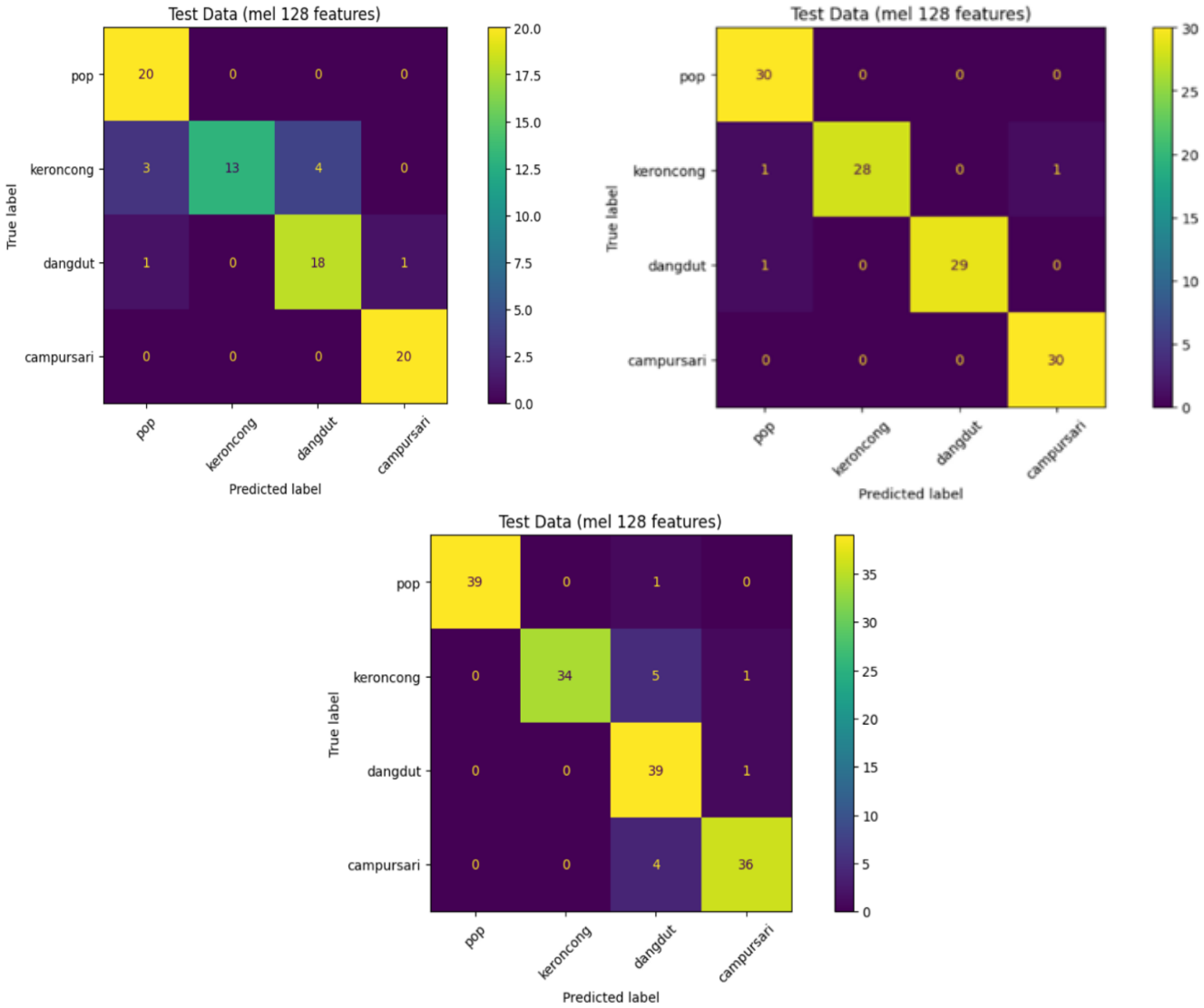
and VGG-16 both exhibited exceptional performance in the classification of Indonesian music genres. Figure 12 illustrates the classification outcomes for each splitting scenario. ResNet-50 demonstrated exceptional performance, achieving an F1-score of 98%, a recall of 98%, a precision of 98%, and an accuracy of 99%. In the interim, VGG-16 also demonstrated satisfactory performance, achieving a 97% F1-score, 97% recall, 96% precision, and 97% accuracy. ResNet-50 outperformed VGG-16 in this classification task, even though both models performed strongly. Detailed performance metrics for each music genre classification are presented in Table 2.

ResNet-50's architectural advantages are the reason for its superiority over VGG-16. ResNet-50 utilizes residual connections to enable the training of deeper networks without encountering the vanishing gradient problem. During backpropagation, these shortcut connections within residual blocks enable the model to more effectively capture complex features and enhance training stability by allowing gradients to flow back to earlier layers. In contrast, VGG-16 has a more complex architecture that lacks residual connections, which makes it susceptible to training challenges in deeper layers as a result of gradient degradation. Consequently, it achieves a slightly lower performance than ResNet-50. Additionally, the performance of the model was examined in relation to the proportion of training data. Performance decreased as the quantity of training data decreased, as evidenced by the three dataset split configurations (80:20, 70:30, and 60:40). Compared to the 80:20 and 70:30 configurations, both ResNet-50 and VGG-16 experienced a performance degradation when the 60:40 train-test split was implemented.

This is consistent with the fundamental principle of deep learning, which posits that a well-generalized model is contingent upon the availability of a larger dataset. The model's capacity to generalize effectively to unseen test data is diminished, and the risk of overfitting to the training set is elevated. This is because the model has fewer samples to learn meaningful patterns when the training data is restricted. ResNet-50 is a more effective architecture for Indonesian music genre classification than VGG-16, as evidenced by these evaluations. This is primarily due to its capacity to handle deep feature extraction through residual connections. Nevertheless, the model's efficacy is contingent upon the availability of an adequate amount of training data. To optimize performance, it is imperative to ensure that deep learning models have a sufficiently large training set.



(a) Confusion matrix model ResNet-50 with splitting data 80:20; 70:30; 60:40



(b) Confusion matrix model VGG-16 with splitting data 80:20; 70:30; 60:40

Figure 12. Confusion matrix result classification model ResNet-50 and VGG-16

Table 2. Performance metrics for each music genre classification

Model	Scenario Splitting Data	Genre	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
ResNet-50	80:20	Pop	95%	100%	98%	99%
		Dangdut	100%	95%	97%	99%
		Campursari	100%	100%	100%	99%
		Keroncong	100%	100%	100%	99%
	70:30	Pop	94%	100%	97%	98%
		Dangdut	100%	97%	98%	98%
		Campursari	100%	100%	100%	98%
		Keroncong	100%	97%	98%	98%
	60:40	Pop	91%	100%	95%	96%
		Dangdut	100%	97%	99%	96%
		Campursari	98%	100%	99%	96%
		Keroncong	100%	90%	95%	96%
VGG-16	80:20	Pop	95%	100%	91%	89%
		Dangdut	100%	65%	79%	89%
		Campursari	83%	100%	91%	89%
		Keroncong	82%	90%	86%	89%
	70:30	Pop	94%	100%	97%	97%
		Dangdut	100%	93%	97%	97%
		Campursari	100%	97%	98%	97%
		Keroncong	94%	100%	97%	97%
	60:40	Pop	95%	90%	92%	93%
		Dangdut	100%	85%	92%	93%
		Campursari	100%	97%	99%	93%
		Keroncong	80%	97%	99%	93%

4. CONCLUSION

The classification of Indonesian music genres was examined in this study using CNN with ResNet-50 and VGG-16 architectures, with MFCC being employed for feature extraction. According to the experimental findings, both deep learning architectures were highly effective in the classification of Indonesian music genres. ResNet-50 demonstrated exceptional performance, particularly in the 80:20 data split configuration, achieving 99% accuracy, 98% precision, 98% recall, and 98% F1-score. Similarly, VGG-16 demonstrated exceptional performance, achieving a peak accuracy of 97% with a 70:30 split ratio, as well as a 97% F1-score, 96% precision, and 97% recall. ResNet-50's superiority over VGG-16 can be attributed to its residual connections, which improve the training efficiency in deeper networks and mitigate the vanishing gradient problem. These connections facilitate improved gradient flow during backpropagation, thereby enabling ResNet-50 to learn complex features more effectively than VGG-16. In contrast, VGG-16 is more susceptible to performance degradation in deeper layers due to its lack of residual connections, despite its strong performance. Additionally, the research demonstrated that the quantity of training data has a substantial impact on the performance of the model. A higher proportion of training data led to improved generalization and accuracy, as evidenced by the 80:20, 70:30, and 60:40 data split configurations. The model's classifier effectiveness is diminished, and the risk of overfitting is elevated when the training data is diminished, as it has fewer examples to learn from. This emphasizes the significance of possessing a dataset that is both diverse and of a sufficient size to achieve the best possible performance of a deep learning model. Nevertheless, it is imperative to recognize several constraints. The research employed a dataset that was relatively modest in size, consisting of 100 audio files per genre, each lasting 30 seconds. The generalization capability of the models may be influenced by the limited dataset size when classifying new, unseen music. Additionally, the models were restricted from utilizing other spectral and temporal audio features that could improve classification performance, as MFCC was the sole feature extraction method employed in this study.

To more accurately represent the intricacies of traditional and contemporary Indonesian music, we suggest that the dataset be expanded in terms of the sample size per genre and the diversity of Indonesian music genres for future research. Furthermore, investigating alternative or hybrid feature extraction techniques, including Spectral Contrast, Mel-Spectrograms, or Chroma features, could improve the accuracy of classification and increase the robustness of the model. Finally, the state of Indonesian music genre classification could be further advanced by experimenting with other advanced deep learning architectures, such as Vision Transformers (ViTs) or EfficientNet. Future research can facilitate the advancement of music information retrieval and automated music analysis by addressing these aspects, which will result in the development of more accurate, generalized, and comprehensive deep learning models for Indonesian music genre classification.

REFERENCES

[1] Mahanta, S.K., Khilji, A.F.U.R., Pakray, P. (2021). Deep neural network for musical instrument recognition using MFCCs. Computación y Sistemas, 25(2): 351-360. https://doi.org/10.13053/CyS-25-2-3946

[2] Cheng, Y.H., Kuo, C.N. (2022). Machine learning for music genre classification using visual Mel spectrum. Mathematics, 10(23): 4427. https://doi.org/10.3390/math10234427

[3] Elbir, A., Aydin, N. (2020). Music genre classification and music recommendation by using deep learning. Electronics Letters, 56(12): 627-629. https://doi.org/10.1049/el.2019.4202

[4] Auliya, Y.A., Fadah, I., Baihaqi, Y., Awwaliyah, I.N. (2024). Green bean classification: Fully convolutional neural network with Adam optimization. Mathematical Modelling of Engineering Problems, 11(6): 1641-1648. https://doi.org/10.18280/mmep.110626

[5] Abbas, A.K., Abed, K.W., Abd, O.I., Al Mashhadany, Y., Jasim, A.H. (2021). High performance of solar panel based on new cooling and cleaning technique. Indonesian Journal of Electrical Engineering and Computer Science, 24(2): 803-814. https://doi.org/10.11591/ijeecs.v24.i2.pp803-814

[6] Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y., Feng, L. (2020). Deep attention based music genre classification. Neurocomputing, 372: 84-91. https://doi.org/10.1016/j.neucom.2019.09.054

[7] Yehezkiel, S.Y., Suyanto, Y. (2022). Music genre identification using SVM and MFCC feature extraction. Indonesian Journal of Electronics and Instrumentation Systems, 12(2): 115-122. https://doi.org/10.22146/ijeis.70898

[8] da Silva, A.C.M., Coelho, M.A.N., Neto, R.F. (2020). A music classification model based on metric learning applied to MP3 audio files. Expert Systems with Applications, 144: 113071. https://doi.org/10.1016/j.eswa.2019.113071

[9] Juwita, S.R., Endah, S.N. (2019). Classification of Indonesian music using the convolutional neural network method. In 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS): Semarang, Indonesia, pp. 1-5. https://doi.org/10.1109/ICICoS48119.2019.8982470

[10] Ndou, N., Ajoodha, R., Jadhav, A. (2021). Music genre classification: A review of deep-learning and traditional machine-learning approaches. In 2021 IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, Canada, pp. 1-6. https://doi.org/10.1109/IEMTRONICS52119.2021.9422487

[11] Cheng, Y.H., Chang, P.C., Kuo, C.N. (2020). Convolutional neural networks approach for music genre classification. In 2020 International Symposium on Computer, Consumer and Control (IS3C), Taichung City, Taiwan, pp. 399-403. https://doi.org/10.1109/IS3C50286.2020.00109

[12] Abdul, Z.K., Al-Talabani, A.K. (2022). Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 10: 122136-122158. https://doi.org/10.1109/ACCESS.2022.3223444

[13] Seo, W., Cho, S.H., Teisseyre, P., Lee, J. (2023). A short survey and comparison of CNN-based music genre classification using multiple spectral features. IEEE Access, 12: 245-257. https://doi.org/10.1109/ACCESS.2023.3346883

[14] Li, J., Han, L., Li, X., Zhu, J., Yuan, B., Gou, Z. (2022). An evaluation of deep neural network models for music classification using spectrograms. Multimedia Tools and Applications, 81(4): 4621-4647. https://doi.org/10.1007/s11042-020-10465-9

[15] Oleiwi, Z.C., AlShemmary, E.N., Al-Augby, S. (2024). Developing hybrid CNN-GRU arrhythmia prediction models using fast Fourier transform on imbalanced ECG datasets. Mathematical Modelling of Engineering Problems, 11(2): 413-429. https://doi.org/10.18280/mmep.110213

[16] Tao, J., Gu, Y., Sun, J., Bie, Y., Wang, H. (2021). Research on VGG16 convolutional neural network feature classification algorithm based on Transfer Learning. In 2021 2nd China International SAR Symposium (CISS), Shanghai, China, pp. 1-3. https://doi.org/10.23919/CISS51089.2021.9652277

[17] Badr, B.E., Altawil, I., Almomani, M., Al-Saadi, M., Alkhurainej, M. (2023). Fault diagnosis of three-phase induction motors using convolutional neural networks. Mathematical Modelling of Engineering Problems, 10(5): 1727-1736. https://doi.org/10.18280/mmep.100523

[18] Wang, G., Yu, H., Sui, Y. (2021). Research on maize disease recognition method based on improved ResNet50. Mobile Information Systems, 2021(1): 9110866. https://doi.org/10.1155/2021/9110866

[19] Moussafir, M., Chaibi, H., Saadane, R., Chehri, A., Rharras, A.E., Jeon, G. (2022). Design of efficient techniques for tomato leaf disease detection using genetic algorithm-based and deep neural networks. Plant and Soil, 479(1): 251-266. https://doi.org/10.1007/s11104-022-05513-2

[20] Das, P.P., Acharjee, A., Jannat, M.E. (2019). Double coated VGG16 architecture: An enhanced approach for genre classification of spectrographic representation of musical pieces. In 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 1-5. https://doi.org/10.1109/ICCIT48885.2019.9038339

[21] Deshpande, A., Pardhi, J. (2021). Automated detection of diabetic retinopathy using VGG-16 architecture. International Research Journal of Engineering and Technology, 8(3): 2936-2940.

[22] Grandini, M., Bagli, E., Visani, G. (2020). Metrics for multi-class classification: An overview. arXiv preprint arXiv:2008.05756. http://arxiv.org/abs/2008.05756.

[23] Meliboev, A., Alikhanov, J., Kim, W. (2022). Performance evaluation of deep learning based network intrusion detection system across multiple balanced and imbalanced datasets. Electronics, 11(4): 515. https://doi.org/10.3390/electronics11040515