Vol. 11, No. 1, April 2026, Page. 91-104
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Classification of Human and AI-Generated Indonesian Pop Songs Based on Spectrogram Using Convolutional Neural Network

Faris Nur Tsani[*1], Wahyu Syaifullah Jauharis Saputra[2], Alfan Rizaldy Pratama[3]

[1,2,3]Data Science, Faculty of Computer Science, UPN "Veteran" Jawa Timur, Indonesia
Email: [1]22083010076@student.upnjatim.ac.id, [2]wahyu.s.j.saputra.if@upnjatim.ac.id,
[3]alfan.fasilkom@upnjatim.ac.id

Abstrak

Pesatnya perkembangan teknologi kecerdasan buatan (AI) memicu lonjakan produksi lagu generatif yang menyerupai karya manusia, sehingga menghadirkan tantangan signifikan terhadap orisinalitas dan hak cipta musik. Penelitian ini bertujuan mengklasifikasikan lagu Pop Indonesia kategori *human-generated* dan *AI-generated* menggunakan pendekatan *Convolutional Neural Network* (CNN) berbasis arsitektur ResNet-18. Dataset terdiri dari 100 lagu berformat MP3 yang terbagi seimbang antara karya manusia dan karya AI dari platform Suno dan Udio. Data audio diproses melalui teknik segmentasi *overlapping window* berdurasi 10 detik dengan *overlap* 5 detik, kemudian diekstraksi menjadi citra spektrogram *Short-Time Fourier Transform* (STFT). Total data yang dihasilkan mencapai 4.282 segmen audio. Hasil pelatihan selama 100 epoch menunjukkan bahwa model mencapai konvergensi dengan *train accuracy* 100% dan *validation accuracy* 95,09%. Pada tahap pengujian menggunakan data yang belum pernah dilihat sebelumnya, model menunjukkan performa unggul dengan tingkat akurasi 93,01%. Temuan ini mengonfirmasi bahwa penggunaan representasi spektogram dalam arsitektur CNN mampu menangkap perbedaan fitur frekuensi dan temporal secara efektif untuk mengidentifikasi musik berbasis AI pada genre Pop Indonesia.

Kata kunci: AI-Generated Music, Convolutional Neural Network, Klasifikasi Audio, Lagu Pop Indonesia, STFT Spektogram.

Abstract

The rapid evolution of artificial intelligence (AI) has led to a surge in generative music that mimics human creations, posing significant challenges to originality and copyright. This study aims to classify human-generated and AI-generated Indonesian pop songs using a Convolutional Neural Network (CNN) approach with the ResNet-18 architecture. The dataset comprises 100 MP3 tracks, balanced between human-made works and AI-generated tracks from Suno and Udio platforms. Audio data were processed using a 10-second overlapping window segmentation technique with a 5-second overlap and extracted into Short-Time Fourier Transform (STFT) spectrogram images. A total of 4,282 audio segments were produced. Training results over 100 epochs reached convergence with 100% training accuracy and 95.09% validation accuracy. In the testing phase, the model demonstrated robust performance, achieving an accuracy of 93.01%. These findings confirm that spectrogram representations within a CNN architecture effectively capture distinct frequency and temporal features to identify AI-generated music within the Indonesian pop genre.

Keywords: *AI-Generated Music, Audio Classification, Convolutional Neural Network, Indonesian Pop Songs, STFT Spectrogram.*

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. INTRODUCTION

The rapid development of artificial intelligence (AI) technology has spread to various creative industries, including the music industry, which is currently undergoing significant transformation. This phenomenon is reflected in a report by Deezer, which recorded a surge in songs generated entirely by AI, from around 10,000 songs per day in January 2025 to more than 30,000 songs per day in September 2025 [1]. AI-based generative platforms are now capable of automatically generating musical compositions with audio quality that is increasingly difficult to distinguish from human-made works. This has a significant impact, especially on the Indonesian pop genre, which is one of the most widely listened to and loved genres of music by the Indonesian public [2]. Pop music has clear chord

progressions and consistent harmonic structures, which are usually tied to one type of Tonnetz, so there is no ambiguity, making this genre increasingly popular with many people [3]. Although these innovations open up new opportunities in the creative process, the emergence of AI-based music also raises concerns, particularly regarding copyright and originality[4]. AI music now poses a growing problem for artists and music labels. Lawsuits are currently being filed against several AI companies. [5]. Therefore, reliable methods for distinguishing between human generated and AI generated music are increasingly needed.

To address these authentication challenges, deep learning approaches, specifically Convolutional Neural Networks (CNNs), have emerged as the standard for digital signal analysis. CNNs demonstrate superior performance in classifying complex patterns when audio signals are represented as spectrograms, which capture both temporal and frequency information simultaneously [6], [7], [8]. Previous research highlights the efficacy of deep learning models in extracting hierarchical representations from either spectrogram-based inputs or raw audio signals [9], [10].

Several studies have reported high classification accuracy under controlled environments. For instance, Afchar et al. achieved an accuracy of 99.8% using CNN-based models, yet noted that performance declined sharply when faced with audio manipulations such as pitch shifting or noise addition [11]. This suggests a reliance on superficial spectral artifacts rather than deep musical structures. Similarly, while traditional machine learning methods like Support Vector Machines (SVM) have achieved F1-scores of 96.9%, they often struggle with generalization when applied to unseen AI generation systems [12]. Recent explorations into architectures like ResNet-18 indicate that while these models offer a better balance of robustness, they still face difficulties in capturing high-level musical semantics, such as cohesive harmony and rhythm [13].

A significant gap remains in current literature, as most approaches classify music at the song level or rely on short, isolated excerpts that fail to represent a song's complete structural arc (intro, verse, chorus, and bridge). While overlapping window techniques are sometimes used to maintain temporal continuity, they are typically relegated to data augmentation strategies rather than serving as a core representation of the full musical structure [14].

To overcome these limitations, this study proposes a classification approach based on full-song segmentation using overlapping windows. By dividing each song into multiple interconnected segments across its entire duration, the model is empowered to learn more detailed temporal and spectral variations that define the complete musical structure. This methodology utilizes the ResNet-18 architecture, utilizing residual connections that are capable of maintaining gradient stability and capturing complex frequency patterns [15]. This research contributes to the field by implementing full-song segmentation to preserve musical integrity and enhance feature representation, while simultaneously applying deep residual learning to sharpen the model's ability to distinguish the subtle nuances between human-generated and AI-generated Indonesian pop music.

2. RESEARCH METHODS

This study formulates the problem as a binary classification task, namely distinguishing between Indonesian pop songs generated by humans (human-generated) and songs generated by artificial intelligence (AI-generated) based on audio signal characteristics. The objective of this study is to build a classification model capable of identifying the source of song creation by grouping each audio signal into one of these two classes. The classification process is carried out through a deep learning approach by utilizing visual representations of audio signals in the form of spectrograms, so that the frequency and temporal characteristics of the audio can be analyzed more effectively. The research method is essential to achieve results that are in line with the predetermined stages. The research stages are shown in Figure 1.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

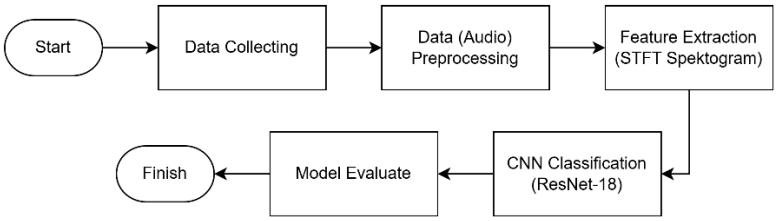


Figure 1. Research process

Figure 1 presents a research flowchart that illustrates the main stages carried out in this study. A more detailed description of each stage is provided in the following subsections.

2.1. Data Collecting

The data collection stage in this study was carried out by preparing two main data groups, namely Indonesian pop songs produced by humans (human-generated) and Indonesian pop songs produced by artificial intelligence (AI-generated). The dataset used in this study consisted of 100 MP3 audio files, comprising 50 human-generated and 50 AI-generated songs. Each song was labeled according to its source to support the classification model training and evaluation process. The selection of a balanced amount of data between classes aimed to minimize potential bias during the model learning process.

Human-generated song data is obtained by downloading songs from the singer's or record label's official online music platform. Official sources were selected to ensure the legality, authenticity, and adequate audio recording quality so that the data could be used as a valid reference in the study. After the download process, the audio content was extracted from the video and saved in high-quality MP3 format. Meanwhile, AI-generated song data was collected through the AI song generator platforms Suno and Udio, both through song creation using prompts designed by the author and songs published by other users. The prompts covered aspects such as tempo, mood, arrangement, and lyrics, resulting in diverse and representative data for classification purposes.

2.2. Data Preprocessing

The data preprocessing stage aims to purify and standardize audio signals before they are used in the next stage of analysis. This process is carried out to ensure that each audio sample has consistent quality, structure, and duration, as well as to minimize irrelevant parts of the signal. In this study, data preprocessing was carried out through three main stages, namely audio resampling, silence removal, and audio trimming.

Audio resampling is performed to equalize the sample rate of all audio files to the minimum sample rate value in the dataset, so that the spectral representation between files is consistent. Next, silence removal is applied to eliminate silent sections at the beginning, middle, or end of the audio that do not contain meaningful acoustic information.

In the silence removal process, signal energy represented by RMS values is used to distinguish between silent sections and musical sections that still contain meaningful acoustic information, as is commonly applied in audio detection and segmentation [16]. An RMS threshold of 25 dB is used as the threshold in the silence removal process to eliminate silent segments that do not contain meaningful acoustic information. This threshold is chosen as a conservative value that is low enough to preserve low-dynamic music sections, but high enough to eliminate silent segments that do not contain meaningful acoustic information.

The audio trimming process is performed by cutting the audio signal into 10-second segments using an overlapping window scheme with a 5-second overlap. The overlapping window-based segmentation approach is commonly used in music audio analysis to maintain temporal continuity and increase the number of training samples without losing musical context [9], [10]. Medium-length segments were chosen to balance musical structure completeness and model learning efficiency, as segments that are too short may lose harmonic and rhythmic patterns, while segments that are too long may reduce the amount of training data available [17].

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

For songs that are not multiples of 10 seconds in length, a final segment is added from the last 10 seconds of the song so that all audio information is represented. These three stages are carried out to ensure the uniformity of the audio data used in the feature extraction and model training processes.

2.3. Feature Extraction

Feature extraction in this study is the process of converting raw audio signals into visual representations that can be processed by image-based models before model training using data. Therefore, transformation into the frequency domain is required through methods such as Short-Time Fourier Transform (STFT) in order to describe frequency variations over time. STFT is a Fourier-related transformation used to determine the frequency of sinusoidal curves and the phase content of local parts of the signal due to changes over time [18].

The results of STFT transformation are represented in the form of a spectrogram, which is a visual representation of a digital audio signal, making it easier for computers to understand. [6]. Spectrograms were chosen because they are raw, i.e., a visual representation of the amplitude of the sound spectrum over time without undergoing further transformations such as cepstral information extraction [19]. In deep learning-based research, particularly using Convolutional Neural Networks (CNN) with ResNet-18 architecture, spectrograms are very important because they can be treated like images. This allows the network to automatically extract complex features from audio data, thereby improving the effectiveness of the classification and analysis processes.

2.4. Modelling

The modeling stage in this study uses a deep learning model, namely Convolutional Neural Network (CNN) with ResNet-18 architecture. The model training and testing processes use a processed dataset, and the model performance is evaluated through various measurement metrics. The CNN structure is inspired by the way neurons work in the human and animal brain, similar to conventional neural networks, so it has the main advantage of being able to automatically identify relevant features without requiring manual feature engineering by humans. [20]. The main concept of CNN lies in its convolution operation, whereby an image will be extracted for each feature to form several patterns that will be easier to classify. [21].

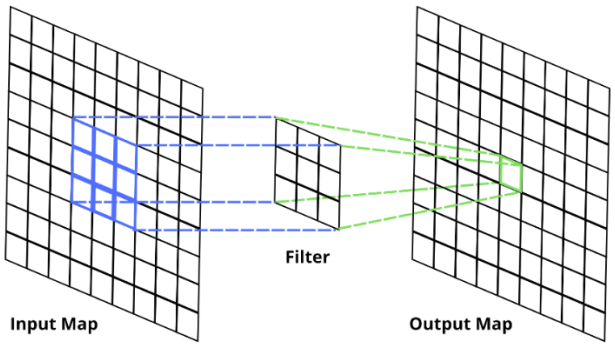


Figure 2. Convolutional Layer

The convolution layer shown in Figure 2 is a process that produces feature maps representing important patterns in the data, such as edges, textures, or specific frequencies in audio. This process allows CNNs to automatically learn features relevant to classification tasks. Mathematically, the activation map calculation can be written as follows:

$$A(i,j) = \sum_{m=0}^{M-1} \left(\sum_{N}^{N-1} I(i+m, j+n) \cdot K(m,n) + b \right) \qquad (1)$$

$A(i,j)$ = Values on the activation map, I = Input Image, $K = M \times N$ Kernel, b = Bias parameter

Each filter has the ability to detect specific features, so with many filters, CNNs can build layered representations from simple (e.g., lines or corners) to complex (e.g., object shapes).

The Convolutional Neural Network (CNN) model with ResNet-18 architecture is an important milestone in the development of deep learning because it successfully overcomes the vanishing gradient problem and accuracy degradation in very deep networks [22]. Unlike conventional CNNs, ResNet introduces the concept of residual learning through a structure called skip connection or shortcut connection, which is an additional path that directly passes through one or more convolutional layers and adds the results to the output of the previous layer. With this approach, the network can "skip" irrelevant layers, thereby facilitating the learning process of identity functions and accelerating model convergence. Systematically, residual blocks are expressed as:

$$y = f(x, \{W_i\}) + x \qquad (2)$$

x = Residual block input, $\{W_i\}$ = Weight collection from layers within the residual block, $f(x, \{W_i\})$, = Nonlinear transformation, y = Residual block output.

This concept enables much deeper network training with better performance compared to the "plain CNN" architecture without skip connections.

The ResNet-18 model is one of the early variants of the ResNet family, which has 18 layers (including convolutional and fully-connected layers) and uses a simple residual block (two 3×3 convolutional layers + skip connection) without a bottleneck.

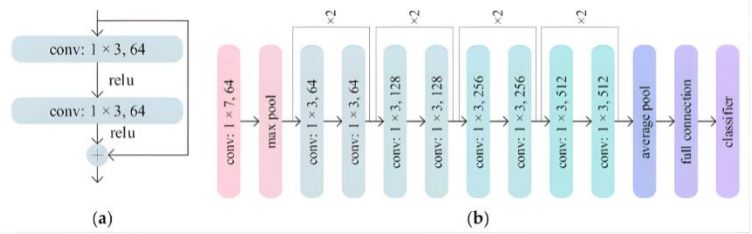


Figure 3. ResNet-18 Structure

Figure 3 shows the two main components of the ResNet-18 architecture. The first part shows the basic structure of the residual block, which consists of two convolution layers with a 3×3 kernel size with ReLU activation and a shortcut connection that combines the original input with the convolution result through a summation operation. This mechanism helps the network learn the differences between the input and the expected output, making the training process more efficient and stable. The second part shows the entire ResNet-18 architecture, which begins with a 7×7 convolution layer and max pooling, followed by four stages of residual blocks with a gradually increasing number of filters (64, 128, 256, and 512), and ends with global average pooling, a fully connected layer, and a classification layer.

During the training phase, the Convolutional Neural Network (CNN) model with ResNet-18 architecture was trained from scratch without using pre-trained weights. The final classification layer was modified by adding Dropout of 0.5 and a fully connected layer with two neurons to support the binary classification task. The model was trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32, as well as the Cross-Entropy Loss function as the objective function. To improve training stability, a ReduceLROnPlateau learning rate scheduler was used, which automatically reduced the learning rate by 50% if the validation loss did not improve for three consecutive epochs. The training process was carried out for 100 epochs until the model reached convergence.

The main advantage of the ResNet-18 architecture lies in its ability to maintain high accuracy with relatively low complexity. Because it uses basic blocks without bottlenecks, this model is lighter than deeper ResNet variants such as ResNet-50 or ResNet-101, but remains effective in extracting important features from image data [22]. Therefore, ResNet-18 is often used as a backbone for various transfer learning and visual classification tasks in the field of digital image-based pattern recognition.

2.5. Model Evaluation

Model evaluation is an important step in determining how well a model can predict previously unseen data. The evaluation process aims to ensure that the developed model not only performs well on training data, but is also capable of generalizing to new data. Evaluation methods vary according to the type of task, whether classification, regression, or clustering. In classification tasks, evaluation is generally performed using a Confusion Matrix and its derivative metrics, such as accuracy, precision, recall, and F1-score.

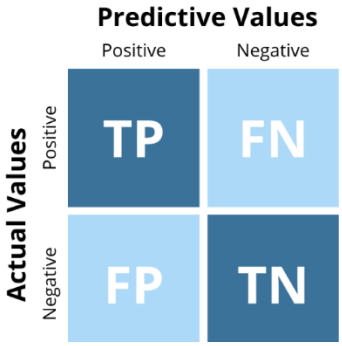


Figure 4. *Confusion Matrix*

A confusion matrix is a matrix representation used to evaluate the performance of a classification model by comparing the model's predictions with the actual values [23]. A confusion matrix has four main components, which can be seen in Figure 4. True Positive (TP) is when the model successfully predicts the positive class correctly. False Positive (FP) occurs when the model incorrectly classifies negative data as positive (known as a type I error). False Negative (FN) occurs when the model incorrectly predicts positive data as negative (type II error). Meanwhile, True Negative (TN) is when the model correctly identifies the negative class. From these four components of the confusion matrix, various evaluation metrics such as accuracy, precision, recall, and F1-score can be calculated to assess the model's performance in more detail.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

Accuracy is the most commonly used metric for evaluating classification models. Accuracy measures the percentage of correct predictions compared to the total number of predictions.

$$Precision = \frac{TP}{TP+FP} \qquad (4)$$

Precision is a measure of how accurately a model or other system predicts positive classes correctly, where high precision means that most of the positive predictions generated by the model are correct. It is calculated as the ratio of true positives to total positive predictions.

$$Recall = \frac{TP}{TP+FN} \qquad (5)$$

Recall measures the model's ability to correctly identify all true positive instances. This metric is important when we want to minimize false negatives (FN), which are missed positive cases.

$$F\text{-}1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (6)$$

F-1 Score is a combination of precision and recall in a single value. This metric is the harmonic mean of precision and recall, useful for balancing both aspects.

3. RESULT AND DISCUSSION

This section presents and discusses the experimental results of the proposed approach for classifying human-generated and AI-generated Indonesian pop songs. It begins with a description of the dataset and preprocessing steps, followed by the model evaluation results. The performance of the Convolutional Neural Network (CNN) based on spectrogram

representations is then analyzed to assess its effectiveness in distinguishing between the two classes.

3.1. Dataset

The data used in this study consisted of MP3 audio files totaling 100 Indonesian pop songs. The HUMAN-labeled songs were obtained through the official online music platforms of the singers or record labels, with a total of 50 songs. Official sources were selected to ensure the legality, authenticity, and adequate quality of the audio recordings.

Meanwhile, AI-labeled song data was collected from the AI song generator platforms Suno and Udio. AI-generated songs were obtained by creating songs using a variety of prompts designed by the author and utilizing songs that had been published by other users. The variety of prompts covered aspects such as tempo, mood, arrangement, and lyrics, resulting in diverse and representative data. Here is one of the prompts in generating AI songs.

```
indonesian pop, classic, Vivace, female vocals(young adult, wise,
friendly, brave, easy-going, Alto, indonesian),  gitar akustik, piano,
drum, comfortable, relaxed, funny, playful, laughter, criticism, jokes,
```

The number of AI-generated songs used in this study was adjusted to be balanced with human-generated data, namely 50 songs. A summary of the song data acquisition is presented in the following table.

Table 1. Summary of song acquisition as a data source

Label	Source		Count of Song	Total Duration
HUMAN	Official online music platform		50	3:27:40
AI	Suno	Prompt Penulis	6	2:47:03
		Published Song	15	
	Udio	Prompt Penulis	5	
		Published Song	24	

Based on the information in Table 1, it shows variations in song sources from AI labels that are not limited to a single source, with the aim of making the model more robust in recognizing data variations.

3.2. Result of Data Preprocessing and Feature Extraction

The data preprocessing stage produces a collection of standardized audio signals that are ready for use in feature extraction and model training. All data displays a sample rate value of 44.1 kHz and 48 kHz. Through the resampling process, all audio files have their sample rates standardized using the down sampling technique to the minimum value found in the dataset, which is 44.1 kHz, so that differences in spectral characteristics between songs caused by sample rate variations can be minimized.

The silence removal process successfully eliminated silent sections at the beginning, middle, and end of songs that did not contain significant musical information. The removal of these silent segments contributed to an increase in the consistency of the effective duration of the audio and reduced the potential for temporal bias due to differences in the length of silence between songs.

Table 2. Changes in data duration after silence removal

Label	Duration Before Silence Removal (s)	Duration After Silence Removal (s)
AI	9803.93	9663.27
HUMAN	12431.48	12003.04

Based on Table 2, the silence removal process resulted in a reduction in the total audio duration in both data categories, AI-generated and human-generated songs. In AI-generated data, the total duration

of silence was 140 seconds, while in human-generated data, the total duration of silence was 428 seconds. Overall, the audio duration cut due to the silence removal process was 2.56% of the total data duration, indicating that the silent parts removed were relatively small compared to the overall audio content. This reduction in duration indicates that the silence removal process successfully removed silent parts that did not contain significant musical information, making the audio data more compact and focused on relevant acoustic content. Despite the reduction in duration, the proportion of the main musical structure is maintained, so this process does not remove important information needed for feature extraction. Thus, silence removal contributes to improving the consistency of audio data and supports the efficiency of the model training process in the next stage.

Furthermore, the audio trimming process with a 10-second overlapping window scheme and a 5-second overlap produces uniform and overlapping audio segments, thereby maintaining the temporal continuity of songs with continuous musical structures.

Table 3. Number of Song Segments Formed

Label	Duration (Song)	Total Segment
AI	9663.27	1907
HUMAN	12003.04	2375

Based on Table 3, the audio segmentation process using the overlapping window scheme produces a different number of song segments in each data category. AI-generated data produced 1,907 song segments from a total duration of 9,663.27 seconds, while human-generated data produced 2,375 song segments from a total duration of 12,003.04 seconds. This difference in the number of segments is in line with the difference in the total audio duration in each category, where a longer duration produces more pieces. These results show that the segmentation process runs consistently and proportionally to the length of the audio signal, ensuring that each part of the song is represented in the form of segments of uniform duration. Thus, overlapping window-based segmentation allows for an increase in the number of training samples while maintaining the temporal continuity of information that is important to support the classification model learning process in the next stage. However, before entering the audio data modeling stage, the segmented audio undergoes a spectrogram feature extraction process.

The result of the feature extraction stage is a spectrogram image that represents the distribution of frequency energy over time for each audio segment, as shown in Figure 5.

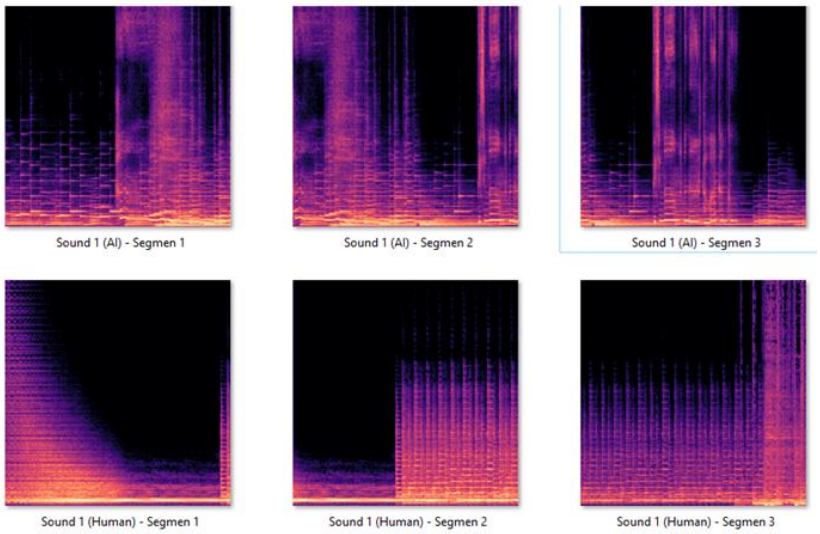


Figure 5. Spectrogram of feature extraction results

The resulting spectrograms numbered as many as the segmentations produced in the previous audio trimming stage, with a total of 4282 images. Thus, the combination of preprocessing and feature extraction results in the form of these images became the input data for training the CNN-based classification model in the next stage.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

3.3. Result of *Train and Validation* Model

The model training process was carried out using a ResNet18-based CNN model with input in the form of spectrogram images from each audio segment. Training was carried out for 100 epochs with data division of 80% for training data and 10% for validation data with data split details in the following table.

Table 4. Data splitting

Label		Train	Validation	Test
AI	Song	40	5	5
	Segment	1522	191	191
HUMAN	Song	40	5	5
	Segment	1899	237	238

Based on Table 4, the percentage distribution of data is not only based on the number of segments produced, but also takes into account the original song of each segment. This approach is considered relevant given that each song is cut into several segments through the segmentation process, so that the data distribution is proportional and still represents the entire song. Thus, the data distribution strategy applied is able to maintain the representativeness of the audio characteristics of each song and reduce the potential for bias that may arise from cutting one song into many segments.

During the training process, model performance is observed through changes in training accuracy and validation accuracy values at each epoch to evaluate the model's ability to learn patterns from data and maintain generalization capabilities.

Table 5. Model training and validation results

Epoch	Train		Validation	
	Accuracy	Loss	Accuracy	Loss
1	0.8220	0.3828	0.7874	0.6253
2	0.9491	0.1220	0.8388	0.3381
3	0.9743	0.0663	0.6939	1.6151
4	0.9947	0.0212	0.9042	0.2825
...	...	...	...	...
97	0.9991	0.0017	0.9486	0.1321
98	1.0000	0.0002	0.9463	0.1382
99	1.0000	0.0003	0.9439	0.1415
100	1.0000	0.0002	0.9509	0.1259

Based on Table 5 above, at epoch 100, the model shows stable performance with a training accuracy value of 100% and a validation accuracy of 95.09%, accompanied by a very low training loss value of 0.0002 and a validation loss of 0.1259. These results indicate that the model has reached convergence, where the learning process no longer experiences significant performance improvements. The high accuracy on the training data shows that the model is able to learn the characteristic audio patterns from the spectrogram images optimally, while the validation accuracy value, which remains high and relatively stable, indicates the model's good generalization ability to data that was not used during training. Although there is a difference between the training accuracy and validation accuracy, the difference is still within reasonable limits and does not indicate excessive overfitting.

Vol. 11, No. 1, April 2026, Page. 91-104
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

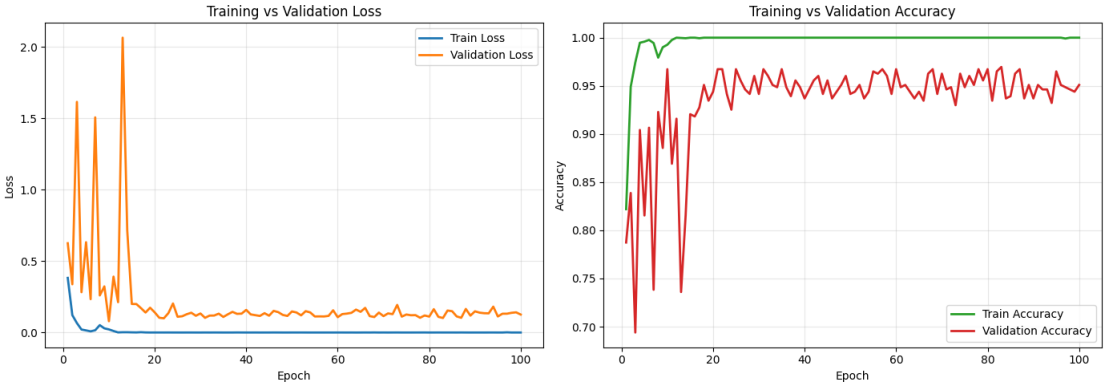


Figure 6. Accuracy and loss graphs for train vs. validation models

In Figure 6, based on the training vs. validation loss graph, it can be seen that the training loss value decreases very rapidly in the early epochs and approaches zero after the first few epochs. This shows that the model is able to learn patterns from the training data effectively and reach convergence in a relatively short time. Meanwhile, the validation loss shows considerable fluctuation in the early epochs, indicating that the model is still adjusting to the patterns in the validation data. After approximately 15–20 epochs, the validation loss value tends to stabilize at a lower range, indicating that the learning process has become consistent and the model is able to maintain its performance on data not used in training.

On the training vs. validation accuracy graph, training accuracy increases very quickly and reaches a value close to 100% from the early epochs, then remains stable until the end of training. This indicates that the model has sufficient capacity to learn feature representations from spectrogram-based training data. On the other hand, validation accuracy fluctuates in the early stages of training, but gradually increases and stabilizes in the range of 94–96% until the 100th epoch. The stability of validation accuracy in the final epochs indicates that the model not only memorizes the training data, but is also capable of generalizing well to the validation data.

Based on the training and validation results, it can be concluded that the CNN model with the ResNet18 architecture has reached convergence with stable performance. The high training accuracy value accompanied by consistent validation accuracy indicates that the spectrogram representation used is capable of capturing the important characteristics of human-generated and AI-generated Indonesian pop songs. Although there is a difference between the performance on the training data and the validation data, the difference is still within reasonable limits and does not indicate significant overfitting. Thus, these training and validation results confirm that the model built is feasible to proceed to the evaluation stage using test data to assess overall classification performance.

3.4. Model Evaluation

The model evaluation was conducted using unseen test data to assess the generalization performance of the proposed model. The evaluation results are presented using a confusion matrix, which provides a detailed distribution of classification outcomes, including true positives, false positives, true negatives, and false negatives. This analysis enables a deeper understanding of the model's performance in distinguishing between human-generated and AI-generated songs, as well as the types of misclassification that occur.

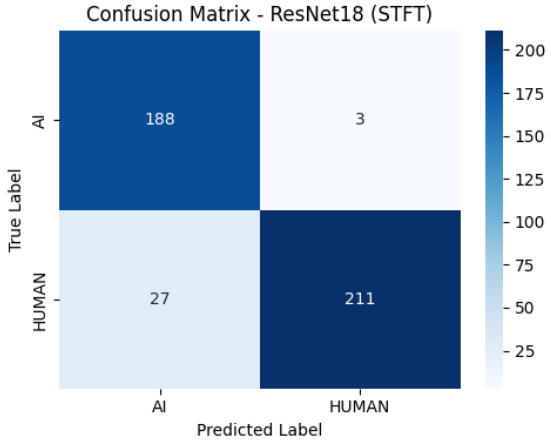


Figure 7. Confusion Matrix test model result

As shown in Figure 7, the confusion matrix illustrates that the model achieves excellent classification performance on test data without augmentation.. A total of 188 AI-generated data and 211 human-generated data were correctly classified. The classification errors that occurred were relatively small, namely 3 AI-generated data that were incorrectly predicted as human-generated and 27 human-generated data that were incorrectly predicted as AI-generated. This error distribution shows that the model has good discrimination capabilities for both classes, although there are still a number of ambiguous cases that are difficult to distinguish spectrally.

Classification errors in human-generated data predicted as AI-generated are thought to be related to audio production characteristics that resemble AI-generated songs. Some segments of these songs tend to have dominant digital instrumentation, high vocal compression levels, and relatively homogeneous harmonic and dynamic patterns. These characteristics have the potential to produce more uniform and less varied spectrogram patterns, making it difficult for image-based models to distinguish them from AI-generated songs, which also generally exhibit high spectral consistency.

Conversely, misclassification of the three AI-generated song segments predicted as human-generated is likely due to the increasing quality of AI generative music production. These songs exhibit more complex musical structures, natural chord progressions, and frequency energy distributions and temporal transitions that resemble the characteristics of human-made pop songs. These findings indicate that as generative technology advances, the spectral differences between human-generated and AI-generated songs are becoming increasingly subtle, ultimately contributing to classification errors. Nevertheless, overall, the model continues to demonstrate high accuracy and generalization capabilities without significant bias toward either class.

Table 6. Model evaluation matrix

Evaluation Matrix	**Score**
Accuracy	0.9301
Precision	0.9860
Recall	0.8866
F1-Score	0.9336

As shown in Table 6, the model achieved an accuracy value of 93.01%, indicating that the model has high classification performance in distinguishing between human-generated and AI-generated Indonesian pop songs overall. The precision value of 98.60% indicates that most of the positive predictions generated by the model have a very high accuracy rate, resulting in a relatively low positive prediction error. Meanwhile, the recall value of 88.66% shows that although most of the positive data was successfully identified, there were still a small number of positive samples that were not detected by the model. The combination of precision and recall values resulted in an F1-score of 93.36%, reflecting a good balance between prediction accuracy and completeness. These results confirm that the ResNet18-based CNN model with spectrogram representation is capable of providing reliable and stable classification performance on test data.

To further evaluate the effectiveness of the proposed model, a comparison with previous studies was conducted. Prior research has reported very high performance in detecting AI-generated music, with accuracy reaching up to 99.8% using CNN-based approaches under controlled experimental conditions [5]. In addition, traditional machine learning methods such as Support Vector Machine (SVM) have achieved high performance, with F1-scores up to 96.9% in distinguishing AI-generated music. Furthermore, deep learning architectures such as ResNet-18 have demonstrated competitive performance, achieving accuracy of up to 92.4% in in-domain evaluation scenarios [12], [24].

However, these studies employ different input representations and experimental settings, such as raw audio waveforms or high-level audio embeddings, whereas this study utilizes spectrogram-based representations to capture time–frequency characteristics of audio signals. The proposed model achieved an accuracy of 93.01% on the test dataset, which is comparable to prior deep learning approaches. Although some studies report higher performance, such comparisons are not strictly equivalent due to differences in input representation, dataset composition, and evaluation conditions.

Despite these results, several limitations should be noted. The dataset consists of 100 Indonesian pop songs, which may limit the generalization of the model to other genres and larger datasets. In addition, the AI-generated data were sourced from specific platforms, potentially restricting broader applicability. The preprocessing approach, particularly audio trimming with overlapping windows, was applied to preserve the full musical content of each song; however, it may introduce similarity between segments derived from the same audio. Furthermore, spectrogram-based feature extraction focuses on time–frequency representations and may not fully capture higher-level musical structures. The confusion matrix also shows that misclassifications occur more frequently in human-generated songs, indicating that certain spectral characteristics remain difficult to distinguish.

4. CONCLUSION

4.1. Conclusion

This study demonstrates that a Convolutional Neural Network (CNN) approach based on spectrogram representation using the ResNet-18 architecture is effective in classifying human-generated and AI-generated Indonesian pop songs. Through preprocessing, overlapping window-based segmentation, and spectrogram feature extraction, the model successfully captured distinguishing audio characteristics, achieving 93.01% accuracy, 98.60% precision, 88.66% recall, and an F1-score of 93.36% on the test dataset. The results indicate that the proposed approach performs well in identifying subtle differences between the two classes, although some misclassifications occur due to similar spectral patterns. Overall, the findings confirm that spectrogram-based CNN models provide a reliable approach for AI-generated music detection within the scope of Indonesian pop songs.

4.2. Suggestion

Based on the findings of this study, the following recommendations for future research are as follows:

1. Expanding the dataset by including a larger number of songs, a wider variety of music genres, and AI-generated content from more diverse platforms and generation models. This is important to improve the generalization capability of the model and ensure its applicability across different musical characteristics and real-world scenarios.
2. Exploring alternative feature representations and data processing strategies beyond spectrogram-based inputs, including the evaluation of different segmentation approaches to reduce similarity between overlapping segments. In addition, future work may investigate segment-level classification aggregation to enable song-level analysis, allowing the system to estimate the proportion or percentage of AI-generated characteristics within a full track.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

REFERENCES

[1] T. Ingham, "Nearly a third of all tracks uploaded to Deezer are now fully AI-generated," Music Business Worldwide. Accessed: Nov. 10, 2025. [Online]. Available: https://www.musicbusinessworldwide.com/nearly-a-third-of-all-tracks-uploaded-to-deezer-are-now-fully-ai-generated-says-platform/

[2] E. Darlene, "Identitas Kultural Musik Pop Indonesia dalam Konteks Seni Urban," *Tonika J. Penelit. Dan Pengkaj. Seni*, vol. 7, no. 1, pp. 49–61, May 2024, doi: 10.37368/tonika.v7i1.573.

[3] L. Bigo and M. Andreatta, "A GEOMETRIC MODEL FOR THE ANALYSIS OF POP MUSIC," no. Sonus-special issue on modelling in musical analysis, pp. 36–48, 2014.

[4] L. A. Chanan, "Analisis prinsip penggunaan wajar (Fair Use) pada lagu buatan AI di Platform Spotify berdasarkan Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta," undergraduate, Universitas Islam Negeri Maulana Malik Ibrahim, 2025. Accessed: Dec. 15, 2025. [Online]. Available: http://etheses.uin-malang.ac.id/74572/

[5] D. Afchar, G. Meseguer-Brocal, and R. Hennequin, "AI-Generated Music Detection and its Challenges," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India: IEEE, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10890655.

[6] A. Parlys, A. A. Zahra, and A. Hidayatno, "PENGGOLONGAN LAGU BERDASARKAN SPEKTOGRAM DENGAN CONVOLUTION NEURAL NETWORK," *Transient J. Ilm. Tek. Elektro*, vol. 7, no. 1, pp. 28–33, Mar. 2018, doi: 10.14710/transient.v7i1.28-33.

[7] S. N. Utama, J. E. W. Prakasa, and W. Hariyanto, "Klasifikasi Irama Bacaan Al-Qur'an Menggunakan Algoritma CNN," vol. 7, no. 1, 2025.

[8] S. Ghosh, A. Ghosh, M. Bhutiani, and G. P. Mondal, "Interpretable Deep Learning for Music Genre Classification Using Spectrogram Analysis," in *2025 International Conference on Computing, Intelligence, and Application (CIACON)*, Durgapur, India: IEEE, Jul. 2025, pp. 1–6. doi: 10.1109/CIACON65473.2025.11189702.

[9] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 2392–2396. doi: 10.1109/ICASSP.2017.7952585.

[10] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy: IEEE, May 2014, pp. 6964–6968. doi: 10.1109/ICASSP.2014.6854950.

[11] J. Wilson, A. Rao, and A. Taranum, "An Enhanced CNN Architecture for Music Genre Classification on the GTZAN Dataset," *2025 Third Int. Conf. Netw. Multimed. Inf. Technol. NMITCON*, pp. 01–08, Oct. 2025, doi: DOI:%2010.1109/NMITCON65824.2025.11188194.

[12] Y. Li, Q. Sun, H. Li, L. Specia, and B. W. Schuller, "Explainable detection of machine generated music and early systematic evaluation," vol. Article number: 13757 (2026), Apr. 2026, doi: https://doi.org/10.1038/s41598-026-42133-7.

[13] M. Ashraf, F. Abid, M. O. Raza, J. Rasheed, S. Alsubai, and T. Asuroglu, "Music genre classification with modified residual learning and dual neural network," *PLOS One*, vol. 20, no. 10, p. e0333808, Oct. 2025, doi: 10.1371/journal.pone.0333808.

[14] A. Rojas, G. J. Dolecek, J. M. De La Rosa, and G. Liñán-Cembrano, "Increasing the Accuracy of Spectrogram-based Spectrum Sensing Trained by a Deep Learning Network Using a Resnet-18 Model," in *2024 39th Conference on Design of Circuits and Integrated Systems (DCIS)*, Catania, Italy: IEEE, Nov. 2024, pp. 1–4. doi: 10.1109/DCIS62603.2024.10769191.

[15] Z. Chen, H. Wang, C.-H. Yeh, and X. Liu, "Classify Respiratory Abnormality in Lung Sounds Using STFT and a Fine-Tuned ResNet18 Network," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Taipei, Taiwan: IEEE, Oct. 2022, pp. 233–237. doi: 10.1109/BioCAS54905.2022.9948614.

[16] T. Virtanen, M. D. Plumbley, and D. Ellis, Eds., *Computational Analysis of Sound Scenes and Events*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-63450-0.

[17] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002, doi: 10.1109/TSA.2002.800560.

[18] A. H. Prawironegoro and B. H. Prasetio, "PENERAPAN SHORT TIME FOURIER TRANSFORM PADA KOEFISIEN CEPSTRAL MEL UNTUK MENGHITUNG JUMLAH PEMBICARA," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 1, no. 1, Jan. 2017.

[19] R. A. N. Diaz, N. L. G. P. Suwirmayanti, and K. Budiarta, "PERBANDINGAN KUALITAS PENGENALAN SUARA UNTUK EKSTRAKSI FITUR MENGGUNAKAN MFCC DAN SPECTRAL," *Naratif J. Nas. Ris. Apl. Dan Tek. Inform.*, vol. 6, no. 1, pp. 58–63, Jul. 2024, doi: 10.53580/naratif.v6i1.281.

[20] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *Proc. 32nd Int. Conf. Mach. Learn. PMLR 37448-456 2015*, vol. 37, pp. 448–456, Jun. 2015.

[21] M. Arsal, B. Agus Wardijono, and D. Anggraini, "Face Recognition Untuk Akses Pegawai Bank Menggunakan Deep Learning Dengan Metode CNN," *J. Nas. Teknol. Dan Sist. Inf.*, vol. 6, no. 1, pp. 55–63, Jun. 2020, doi: 10.25077/TEKNOSI.v6i1.2020.55-63.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[23] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technologies*, vol. 12, no. 7, p. 113, Jul. 2024, doi: 10.3390/technologies12070113.

[24] L. C. Vila, B. L. T. Sturm, L. Casini, and D. Dalmazzo, "The AI Music Arms Race: On the Detection of AI-Generated Music," vol. 8(1), pp. 179–194, June 25, doi: DOI:%20https://doi.org/10.5334/tis%20mir.254.