APPLIED
Ethnomusicology

ARTICLE
Submitted date : 15th May 2025
Revised date : 18th June 2025
Accepted date : 30th June 2025

Correspondance Address:
Computer Science Department, Binus Online Learning, Universitas Bina Nusantara. Jl. K. H. Syahdan No. 9, Kemanggisan / Palmerah Jakarta Barat 11480, Indonesia.
E-mail: viga.hardjanto@binus.ac.id

Recognising Bonang Barung Gamelan Instrument Playing Technique Using Convolutional Neural Networks

Viga Laksa Hardjanto[1*], Wahyono[2]

[1]Computer Science Department, Binus Online Learning, Universitas Bina Nusantara, Jakarta, Indonesia

[2]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract:

The diversity of Indonesian culture is an interesting thing to be reviewed further. One of them is called karawitan, which uses gamelan instruments as its medium, with various playing techniques on each instrument. This can be found on the bonang barung instrument, which has at least thirteen different playing techniques or known as tabuhan techniques. However, it is not easy for beginners to learn karawitan, since many techniques must be learned on each instrument. Besides wanting to help beginners learn karawitan with the support of up-to-date system capabilities in processing data, this research is also expected to enrich research in the field of audio classification. Types of feature extractions, such as mel spectrogram and MFCC, were tested on the CNN architecture. In addition, the process of cleaning noise and levelling the loudness level of the raw data is applied with the aim of getting better audio quality. Apparently, at the best hyperparameter settings, it was found that the MFCC feature is better for audio data containing noise, which achieves an accuracy as good as 99%, while the Mel Spectrogram excels at noise-free audio data with an accuracy as good as 98%. Therefore, the end of this study shows that the MFCC and Mel Spectrogram features have their respective advantages.

Keywords: Gamelan; Mel Spectrogram; MFCC; CNN

Abstrak:

Keberagaman budaya Indonesia menjadi hal yang menarik untuk diulas dan ditinjau lebih lanjut. Salah satunya, ialah seni karawitan yang menggunakan alat musik gamelan sebagai medianya dengan beragam teknik permainan pada tiap instrumennya. Hal ini dapat ditemukan pada instrumen bonang barung yang memiliki sedikitnya tiga belas teknik tabuhan yang berbeda. Ternilai tidaklah mudah bagi pemula untuk mempelajari seni karawitan, mengingat begitu banyaknya teknik yang harus dipelajari pada tiap instrumennya. Sebagai upaya dalam membantu pemula mempelajari karawitan dengan dukungan keterbaruan kemampuan sistem dalam mengolah data, adanya penelitian ini juga diharapkan bisa memperkaya penelitian dibidang klasifikasi audio. Jenis esktraksi fitur seperti mel spektrogram dan MFCC diujikan terhadap arsitektur CNN. Selain itu, proses pembersihan noise dan penyamarataan level kenyaringan terhadap raw data diterapkan dengan tujuan untuk mendapatkan kualitas audio yang lebih baik. Rupanya, pada pengaturan hyperparameter terbaik, diperoleh bahwa fitur MFCC lebih unggul pada data audio yang memuat noise, yakni mencapai akurasi sebaik 99%, sedangkan Mel Spektrogram unggul pada data audio yang bersih dari noise dengan akurasi sebaik 98%. Oleh karena itu, akhir dari penelitian ini menunjukkan bahwa fitur MFCC dan Mel Spektrogram memiliki kenggulannya masing-masing.

Kata kunci: Gamelan; Mel Spektrogram; MFCC; CNN.

1. Introduction

The Javanese culture is the result of pure thoughts of the ancestors that are arranged in such a way that it becomes a beautiful unity to be seen, heard, and felt. As is the case with gamelan, which is a representation of Javanese culture in ensemble music, it has very complex rules for playing. Not only gamelan, but other types of ensemble music can be found in Indonesia, such as angklung from West Java, kolintang from North Sulawesi, and many more. Regarding gamelan, apart from Central Java, gamelan can also be found on the Island of the Gods, with some differences and variations. With a large number of instruments, it can be said that playing and learning gamelan ensemble music is a complex task. Moreover, there are different and diverse playing patterns or commonly referred to as playing patterns, on each instrument.

One of the instruments that has various playing techniques is bonang barung, which is also the object of this research. There are 13 bonang playing techniques according to Soeroso (1982). Regarding the problems faced in learning karawitan for beginners in the Student Activity Unit in the field of arts, most of them are considered difficult because the playing pattern of bonang barung seems the same when listened to directly. Therefore, as an effort to help beginners in learning karawitan with the support of updating the system's ability to process data, this research is also expected to enrich research in the field of audio classification by applying deep learning algorithms combined with the extraction of certain characteristics to classify several bonang barung playing techniques as learning media.

Audio classification has grown quite rapidly in the field of computer science. This is based on quite a lot of previous research in the scope around audio data processing. There are at least two things that need to be done to perform classification based on audio or voice files. The first step is featuring extraction from the audio file, followed by classification using machine learning algorithms. The idea behind this research in general is to create a model that has the ability to classify a dataset of bonang barung playing techniques.

2. Literature Review

Before discussing more related to the research conducted, here are some studies that have been done related to the classification of audio data. Research conducted by Bhat and R. (2020) discusses music classification based on genre using a convolutional neural network; the dataset used is GTZAN. The GTZAN dataset is a standard dataset commonly used for music classification. The types of feature extraction used in this research are zero crossing and MFCC. This research produces 98% training accuracy and 73% validation accuracy. Other research related to audio data classification was conducted by Vishnupriya and Meenakshi, (2018) who used convolutional neural network as the classification model. The dataset used in this study is the Million Song Dataset (MSD). The approach used to perform feature extraction is with MFCC and Mel Spectrogram features. The results obtained in this study on the learning process are as good as 76%, using MFCC, and get 47% accuracy when using Mel Spectrogram.

The research conducted by Cheng, Chang, and Kuo (2020) is in line with what was done by Vishnupriya and Meenakshi (2018), who used CNN to perform classification and MFCC as a feature. There is a layer added that aims to prevent overfitting, namely the dropout layer. As a result, this study obtained accuracy as good as 83.3%. The use of mel frequency cepstral coefficients as a feature extraction was also carried out by Nagawade and Ratnaparkhe (2017), the sound samples used were taken from the Electronic Music Studio, University of Iowa. There are several classes of instruments that are classified with more than one playing technique, including arco, pizzicato, vibrato, and non-vibrato. By using MFCC and K-NN as models to perform the classification task, the recognition accuracy for cello, piano, and trumpet is 91.66% and for flute and violin is 83.33%.

Audio spectrograms were used by Pelchat and Gelowitz (2019) for feature extraction. The dataset used to conduct this research is the Music Library (1880 songs). Classification was done with a convolutional neural network. The results of this study show accuracy as good as 62% on training and 67% test accuracy. Recent research was conducted by Puppala et al., (2021) who used MFCC as feature extraction and CNN as a model for classification. The composition of the division between train data and test data is 60% to 40%. The results are quite good, achieving accuracy as good as 97% on train data and 74% on test data.

Similar to some previous studies, this research aims to create a convolutional neural network model that is able to classify data on bonang barung playing techniques. Furthermore, there is a phenomenon that has not been discussed in more depth in previous studies. Although there has been no more detailed research related to handling audio noise according to Harshavardhan & Mahesh (2022), the phenomenon of the presence of noise is likely to affect the performance of the audio classification model so this research proposes a method used to get better audio quality when listened to, that is, by applying noise removal and audio compressors. In this case, Adobe Audition software can be used to apply both methods. The dataset used for this research is data taken by direct recording of the gamelan instrument bonang barung. Feature extraction used, namely Mel Spectrogram and MFCC. Basically, audio files consist of files that contain noise and files that are clean from noise will be extracted into Mel Spectrogram and MFCC features which will then be used in the CNN model for classification.

3. Methods

The research conducted will go through several stages including literature study, data acquisition process, pre-processing, feature extraction process, deep learning model building, and the last step is testing and evaluation. More details in Figure 1 show the stages of research that will be carried out.

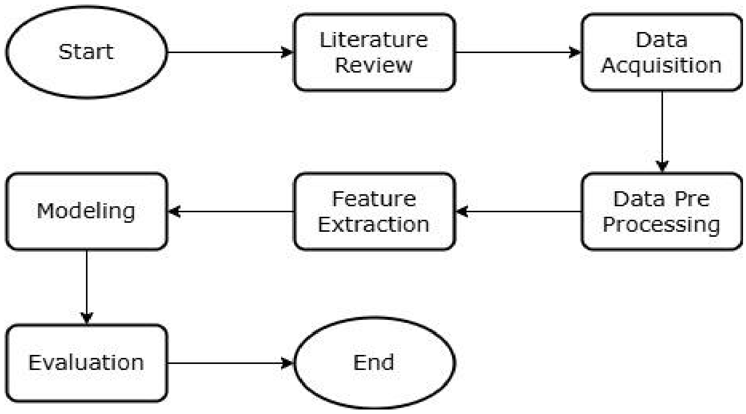


Figure 1: Flow chart of the research process

The process begins with data acquisition using the bonang barung instrument. Then raw data processing is done in Adobe Audition software. The audio data is stored in .wav format so that the details of the audio information are maintained. After obtaining raw audio data in .wav format, the dataset is stored on Google Drive because the author utilizes Google Colab for the implementation of the program code the consideration that Google Colab provides a GPU to carry out the training process so that it can save time on the data training process. In this case, there are three classes of bonang barung tabuhan techniques: *mipil cegatan nggembyang (nggembyang), mipil lamba mlampah,* and *mipil rangkep mlampah* with the data used in each class totaling 50 audio files with a duration of 30 seconds each audio track. To increase and lighten the feature extraction process, each audio track is split into several segments.

The dataset used to conduct the research was recorded directly using Adobe Audition software on the gamelan instrument bonang barung. The recording process can be done repeatedly with the aim of getting maximum results on the raw audio file data. The output format of the audio file is .wav, which is quite good when compared to .mp3 because the audio results are not so compressed. The pre-processing stage is divided into two, namely pre-processing in Adobe Audition software and pre-processing in the program or also called the feature extraction stage. The hardware used for the recording process is a dynamic microphone and an audio interface with a sample rate of 48000 Hz.

There are various ways to get better quality audio, some of which include applying noise removal and audio compressors. Good quality audio is assumed to be audio that is comfortable when listened to, so that it does not cause pain in the ears.

The concept of noise removal is to reduce the initial audio noise to the noise floor so that unnecessary frequencies can be reduced or even eliminated. In the goal of reducing audio noise, the thing that needs to be done is to capture the noise print/noise floor, which is done to get the background noise from the entire audio data. After getting the noise floor, the next step is to select all audio data and then apply noise reduction along the audio track.

The audio compressor is used to obtain audio with a relatively flatter hardness level based on the threshold, which refers to the recording results at the data acquisition stage. Figures 2 to 3 show the process of applying the audio compressor. The threshold used is -18 dB with a ratio of 4:1, meaning that the audio compressor will activate when the audio amplitude is above -18 dB and will suppress the audio amplitude to a quarter of the peak amplitude value. For example, when it is found that the peak amplitude reaches -6 dB then the output amplitude will only reach -15 dB, this applies to the next multiple.

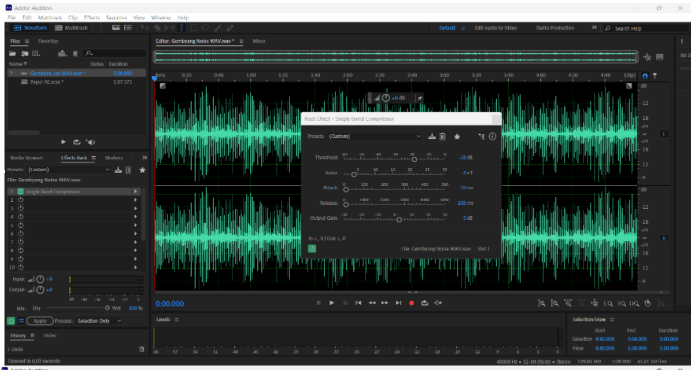


Figure 2: The process of applying audio compressor

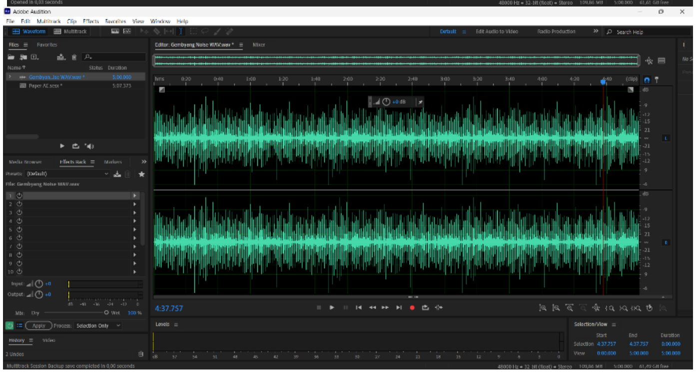


Figure 3: The Condition of audio after compressor is applied

There are two parts of audio processed in this research, audio containing noise and audio that is clean of noise. In addition, there are two audio segment lengths used, three and five second audio segments on each audio track. The schematic of the feature extraction process is depicted in Figure 4 to Figure 5.

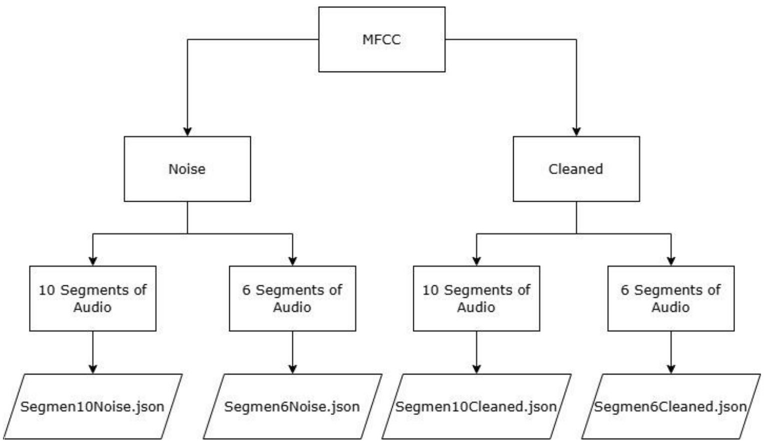
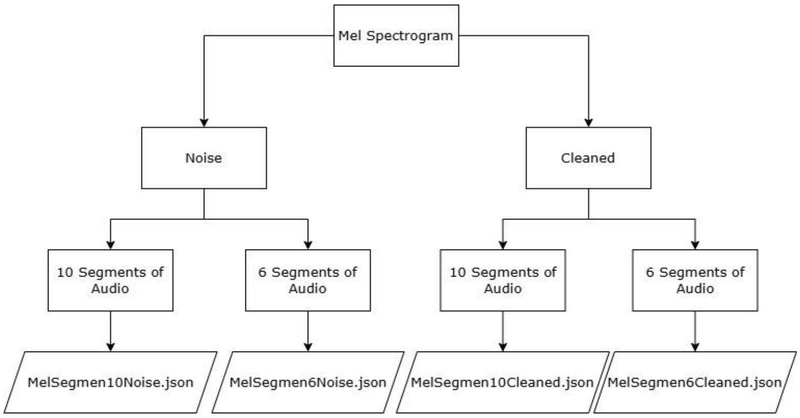


Figure 4: Schematic of file storing after MFCC feature extraction process

Figure 5: Schematic of file storing after Mel Spectrogram feature extraction process



The feature extractions to be used are MFCC and Mel Spectrogram. MFCC was chosen because its working system is very similar to the human hearing system. The human hearing system does not give the same response for the entire frequency range, so with MFCC, we hope to get the highest recognition accuracy Nagawade and Ratnaparkhe, (2017). Meanwhile, Mel Spectrogram was chosen because Mel Spectrogram applications are very commonly used for audio data classification, music genre recognition systems, and music instrument classification (Velardo, 2021).

The MFCC feature extraction workflow is divided into several parts, namely pre-emphasis, windowing, application of Fast Fourier Transform, mel filter bank, log energy, and ends by applying inverse discrete Fourier transform. More details of the MFCC feature extraction flow can be seen in Figures 6 to Figure 7.

Figure 6: Stages of feature extraction in general

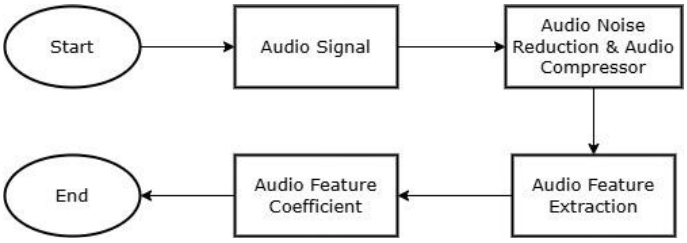
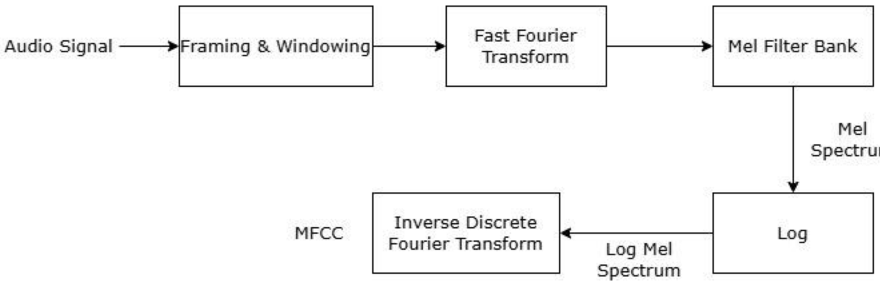


Figure 7: MFCC feature extraction process

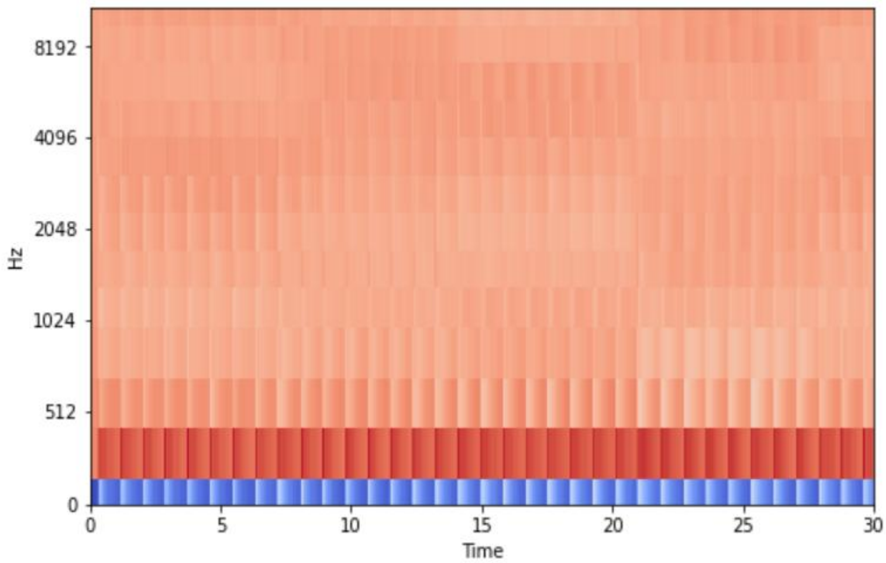


The initial stage is pre-emphasis, which aims to increase the energy of high frequencies, followed by the implementation of the Hamming window, which breaks the data into frames with a smaller duration and less change. In this step, the voice sample is divided into several frames. Each frame generally consists of 441 samples. Hamming window is applied to attenuate both ends of the signal towards zero. The next step is to apply the FFT to each frame. The concept Fast Fourier transform is used to convert the time domain to the frequency domain. Next is to map the spectrum into the mel scale. Mel filter banks are useful for filtering out information on the power spectrum. From this process, the energy bank filter is obtained, and then the logarithmic value is taken. The next step is to apply the inverse discrete Fourier transform to return the log energy from the frequency domain to the time domain. After the inverse discrete Fourier transform, the result is Mel frequency Cepstrum Coefficients (MFCC).

The Mel Spectrogram feature extraction stage is generally divided into three stages, starting from the short-time Fourier transform extraction stage, then converting the amplitude into decibels, and ending with converting the frequency to the mel scale. Converting frequency to the mel scale can be done simply with just a few steps to get the audio representation in the mel scale. First, choose the number of mel bands that is tailored to the problem, so there is no specific benchmark for the exact number of mel bands. Second, create mel filter banks, and the last step is to apply the mel filter banks to the original audio spectrogram (in Hz).

Below is a representation of the MFCC and Mel Spectrogram feature extraction results on the audio data of the nggembyang tabuhan technique, as shown in Figures 8 to 9.

Figure 8: MFCC visualization on one of the audio tracks of nggembyang playing technique



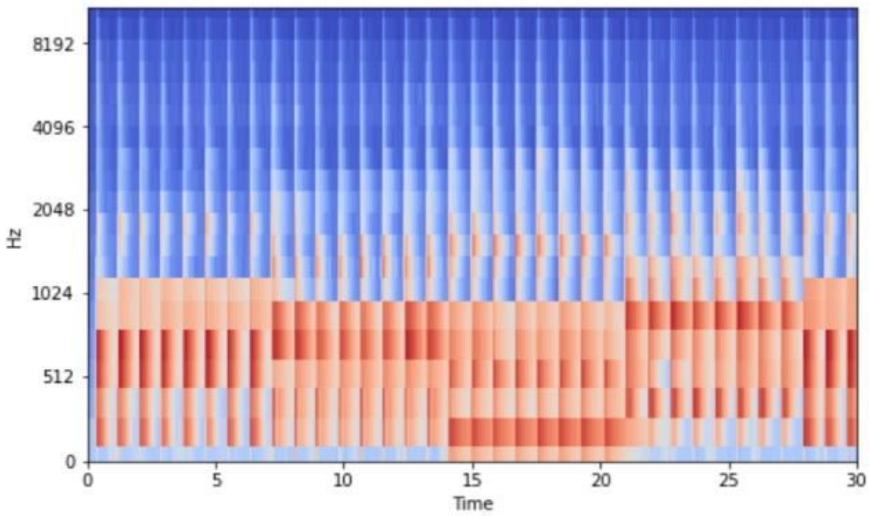


Figure 9: Mel Spectrogram visualization on one of the audio tracks of nggembyang playing technique

Convolutional neural networks perform well for two-dimensional feature classification. CNN also perform well on data with irreversible order or missing elements. Audio is one such data in irreversible order. If the order or elements are changed in the audio sequence, the new audio and the original audio will be different files. Cheng, Chang and Kuo (2020). This research proposes CNN as a model for performing classification tasks on audio data. The CNN architecture consists of three convolution layers, each using 32 filters, followed by a max-pooling layer before finally connecting to a fully connected layer. A dropout layer is used to minimize the possibility of overfitting during the training process. The softmax activation function is applied to the output layer while the other layers use the ReLU activation function. In addition, the Adam optimization function is used in the training process along with the sparse categorical cross entropy loss function. The basic CNN architecture used in this research can be seen in Figure 10.

Layer (type)	Output Shape	Param #
conv2d_8 (Conv2D)	(None, 128, 11, 32)	320
max_pooling2d_6 (MaxPooling2	(None, 64, 6, 32)	0
conv2d_9 (Conv2D)	(None, 62, 4, 32)	9248
max_pooling2d_7 (MaxPooling2	(None, 31, 2, 32)	0
conv2d_10 (Conv2D)	(None, 30, 1, 32)	4128
max_pooling2d_8 (MaxPooling2	(None, 15, 1, 32)	0
flatten (Flatten)	(None, 480)	0
dense (Dense)	(None, 64)	30784
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 3)	195

Total params: 44,675
Trainable params: 44,675
Non-trainable params: 0

Figure 10: Convolutional Neural Network Architecture

To evaluate the ability of the system to perform classification tasks, an evaluation matrix called accuracy is used. In the case of binary classification or two-class classification, there are four terms that represent the results of the classification process in the confusion matrix, namely true positive, true negative, false positive, and false negative (Anggreany, 2020). Furthermore, to see the broad picture of the classification model's ability, a confusion matrix is used where the output can be two or more classes (multiclass classification). Figure 11 is a representation of the multiclass classification confusion matrix.

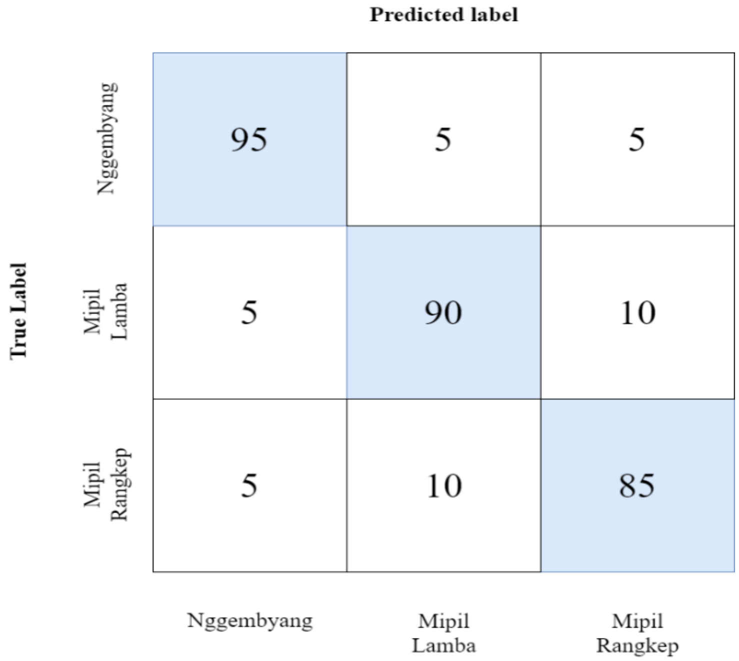


Figure 11: Example of a confusion matrix in the multiclass classification case

4. Results

This research focuses on audio classification using the two most commonly used feature extractions. Below are some tables showing the comparative results of training using MFCC and Mel Spectrogram features.

Table 1: Feature Comparison Results

Features	Noise Condition	Acc	Val Acc	Loss	Val Loss
MFCC	Noise	99%	99%	0,0123	0,1379
MFCC	Clean	93%	96%	0,1617	0,1017
Mel Spektogram	Noise	98%	97%	0,0526	0,0875
Mel Spektogram	Clean	94%	98%	0,1422	0,0522

Based on Table 1, audio data containing noise tends to perform better on both features. The MFCC feature achieved 99% accuracy while the Mel Spectrogram achieved 97%. Both are accuracies in the validation data. In the test data, the noise-loaded MFCC feature achieved 99% accuracy, while the Mel Spectrogram feature achieved accuracy as good as 98%. On the other hand, in audio data that is clean from noise, the Mel Spectrogram feature has a better performance, namely in the test data it produces an accuracy as good as 98% while the MFCC feature in the test data reaches an accuracy of 95% so that in the problem of classifying the Bonang Barung instrument audio dataset, the MFCC and Mel Spectrogram features have their respective advantages.

Furthermore, there are three learning rate values to analyze which one has the best performance, starting from 0.01, 0.001, and 0.0001. On audio data containing noise and audio data clean of noise, the learning rate value of 0.001 shows the best results, reaching 99% on noise audio data and 96% on audio clean of noise. The results of testing three learning rate values using the Mel Spectrogram feature also show the best results at a learning rate of 0.001 with accuracy reaching 97% on noise audio data and 98% on clean audio data from noise.

In this study, several hyperparameters were tuned to get the best model performance. The first hyperparameter to be tuned is the number of filters, among the three values tested, the number of filters 32 has the best performance than the number of filters 3 and 10, meaning that as more filters, the model will get more information. In this study, two types of pooling layers commonly used for classification problems were tested, namely max-pooling and average-pooling. The results show that max-pooling performs better with an accuracy as good as 96%, this value is very clear on the MFCC feature type of audio noise data where average pooling only achieves an accuracy as good as 91%.

One approach to prevent overfit is the use of a dropout layer. Two values of dropout probability were tested, 0.3 (30%) and 0.5 (50%) respectively. The results show that for both types of audio data using MFCC features, the 50% dropout probability tends to be better than the 30% dropout probability. Furthermore, for audio data using the Mel Spectrogram feature, the opposite occurs. Dropout with a 30% probability performed better for the model on the validation data.

There are four epoch values tested, namely 5, 10, 15, 20. Basically, the learning process that is done repeatedly will result in better model performance. But in fact, more epochs do not always improve performance. The test results show that the epoch value with the most significant effect occurs at 10.

Furthermore, different audio sample lengths were also tested, with three-second and five-second samples in each segment. The results show that the convolutional neural network does not benefit from longer audio samples, as evidenced by the three-second audio samples that perform better on almost all data types used. The table below summarizes the most optimal parameter configuration used for the bonang barung audio classification dataset.

Table 2: Optimum Parameter

Parameter	Value
Learning Rate	0,001
Number of Filter	32
Pooling Layer	Max Pooling
Dropout Layer	50%
Length of Audio Segment	3 sec
Number of Epoch	10

With the optimal parameter configuration as in the table 2, the test data obtained 99% accuracy on the MFCC feature on audio data containing noise and 95% on the MFCC feature of audio data clean from noise. While the results obtained 98% on the Mel Spectrogram feature of noise audio data, and 98% on the Mel Spectrogram feature of audio data clean from noise. The confusion matrices for these types of data are shown in Figure 12 to Figure 15.

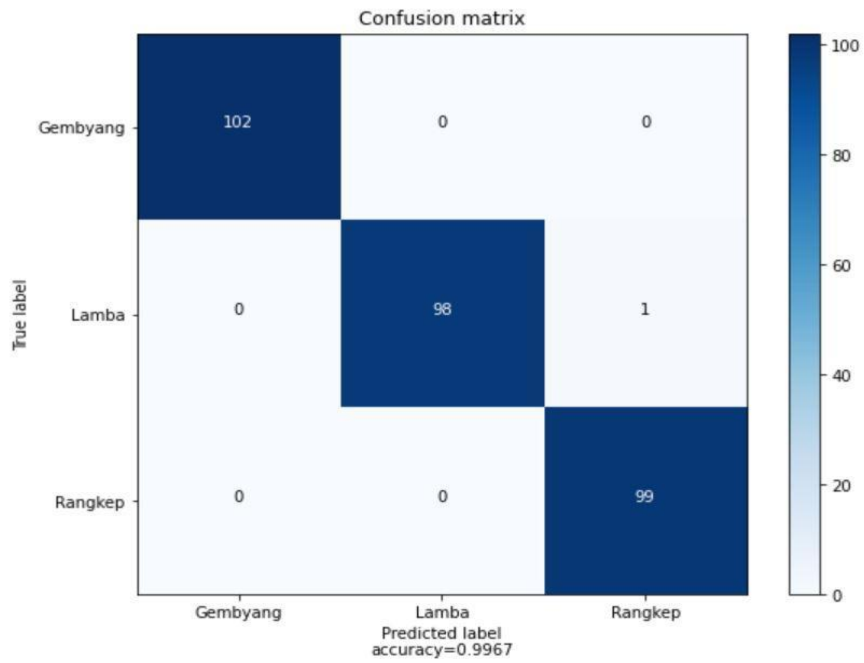
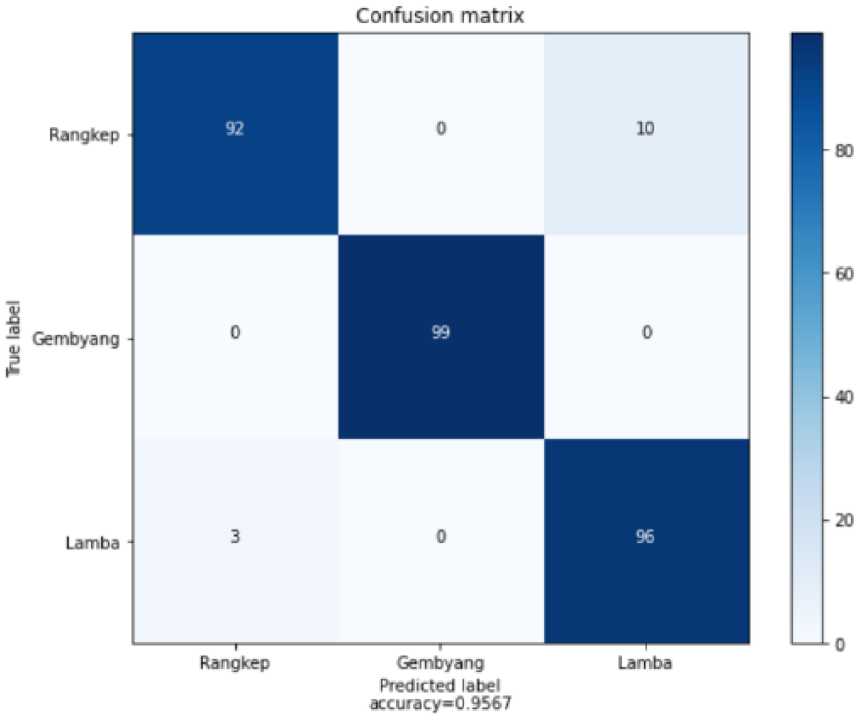
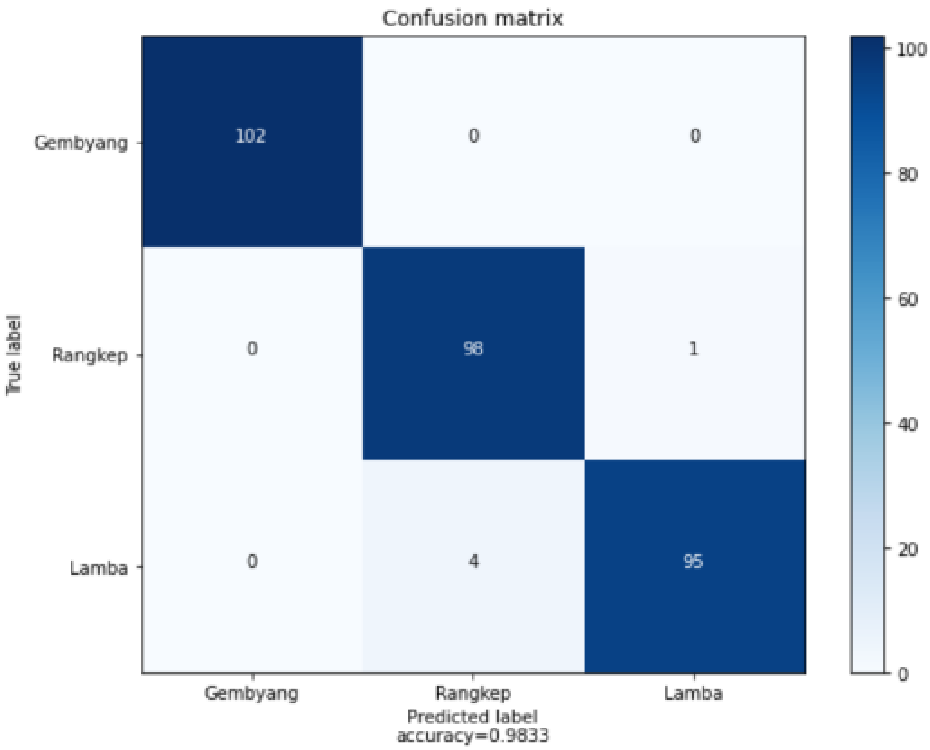


Figure 12: Confusion matrix on test data of noisy audio MFCC features

Figure 13: Confusion matrix on test data of audio MFCC features cleaned of noise	
Figure 14: Confusion matrix on test data of noisy audio Mel Spectrogram features	

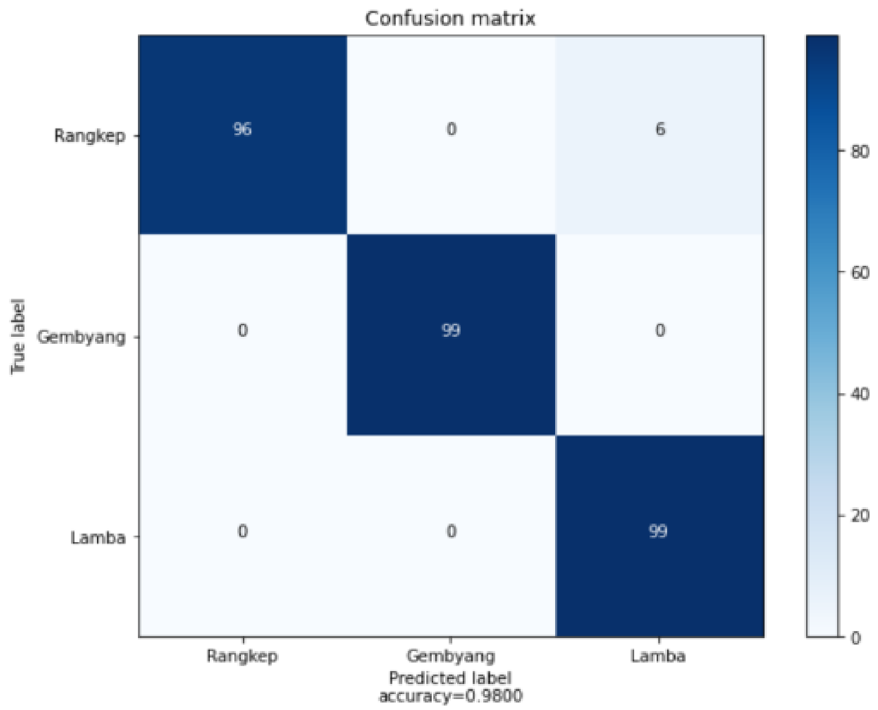


Figure 15: Confusion matrix on test data of audio Mel Spectrogram features cleaned of noise

5. Discussions

The test results show that the two types of feature extraction, that is, MFCC and Mel Spectrogram, have different performance characteristics depending on the condition of the audio data. MFCC shows excellence on audio data containing noise, with accuracy reaching 99% on both validation and test data. This is in line with the nature of MFCC, which is designed to resemble the human hearing system that is able to filter out important information even in unclean sound conditions. In contrast, Mel Spectrogram excels when used on clean audio data, with the best accuracy of 98%. This difference suggests that the selection of extraction features should be tailored to the quality conditions of the audio data in practical implementation.

The effect of hyperparameter tuning also plays an important role in improving the performance of CNN models. A learning rate of 0.001 proved to be the optimal value for both types of feature extraction, as it provided stability in the model learning process without experiencing significant error spikes. In addition, the use of 32 filters and max pooling layer type provides more informative and efficient results in capturing important patterns from two-dimensional data such as audio spectra. The use of 50% dropout in MFCC and 30% in Mel Spectrogram also shows that the dropout probability is not universal, but must be set contextually according to the type of feature used.

Interestingly, the duration of the audio segments has a significant effect on the classification accuracy. CNN model is better able to recognize audio patterns in three-second segments compared to five-second segments, possibly because shorter inputs make it easier for the model to identify discriminative features without being distracted by irrelevant information. It is an important insight for the development of audio-based classification systems that the quality of audio segmentation has a significant impact on model performance. Overall, this research not only emphasizes the importance of feature selection, but also how the entire pipeline from pre-processing, segmentation, to model tuning should be designed in an integrated and contextual approach.

6. Conclusions

This research successfully shows that the classification approach of bonang barung tabuhan techniques using Convolutional Neural Network can be done effectively through the utilization of MFCC and Mel Spectrogram extraction features. Each feature has its own strengths that depend on the audio quality characteristics used. MFCC proved to be more robust in dealing with data with noise, while Mel Spectrogram showed superior performance on clean data. This shows that there is no absolute best feature, rather they are complementary in different contexts.

Testing various configurations of the model parameters provided important insights that optimal performance can be achieved through the right combination of settings, such as the learning rate, number of filters, and use of dropouts. In addition, the choice of audio segment duration was also shown to significantly affect the model's performance, with three-second segments showing more stable and accurate results than five-second segments.

Overall, this study not only offers a technical solution to support karawitan learning, but also opens up opportunities for the development of local culture-based audio classification systems that are adaptive to data conditions. In the future, similar approaches can be extended to other gamelan instruments or used in sound-based pattern recognition systems for the preservation of Indonesian culture.

7. References

Anggreany, M. S. (2020). *Confusion Matrix*. https://socs.binus.ac.id/2020/11/01/confusion-matrix/

Bhat, R. S., B., R. R., & R., M. K. (2020). Music Genre Classification. *International Journal Of Engineering And Computer Science*, *7*(1), 8–13. https://doi.org/10.18535/ijecs/v4i10.38

Cheng, Y. H., Chang, P. C., & Kuo, C. N. (2020). Convolutional neural networks approach for music genre classification. *Proceedings - 2020 International Symposium on Computer, Consumer and Control, IS3C 2020*, 399–403. https://doi.org/10.1109/IS3C50286.2020.00109

Harshavardhan, K. S., & Mahesh. (2022). Urban sound classification using ANN. *International Interdisciplinary Humanitarian Conference for Sustainability, IIHC 2022 - Proceedings*, 1475–1480. https://doi.org/10.1109/IIHC55949.2022.10060146

Nagawade, M. S., & Ratnaparkhe, V. R. (2017). Musical instrument identification using MFCC. *RTEICT 2017 - 2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, Proceedings*, *2018-Janua*, 2198–2202. https://doi.org/10.1109/RTEICT.2017.8256990

Pelchat, N., & Gelowitz, C. M. (2019). Neural Network Music Genre Classification. *Canadian Journal of Electrical and Computer Engineering*, *43*(3), 170–173. https://doi.org/10.1109/CJECE.2020.2970144

Puppala, L. K., Muvva, S. S. R., Chinige, S. R., & Rajendran, P. S. (2021). A Novel Music Genre Classification Using Convolutional Neural Network. *Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021*, 1246–1249. https://doi.org/10.1109/ICCES51350.2021.9489022

Soeroso. (1982). *Bagaimana Bermain Gamelan*.

Velardo, V. (2021). *Mel Spectrograms Explained Easily*. https://github.com/musikalkemist/AudioSignalProcessingForML/tree/master/17 - Mel Spectrogram Explained Easily

Vishnupriya, S., & Meenakshi, K. (2018). Automatic Music Genre Classification using Convolution Neural Network. *2018 International Conference on Computer Communication and Informatics, ICCCI 2018*, 9–12. https://doi.org/10.1109/ICCCI.2018.8441340