

Deep learning for audio signal-based tempo classification scenarios

Muljono¹, Pulung Nurtantio Andono¹, Sari Ayu Wulandari², Harun Al Azies¹, Muhammad Naufal¹

¹Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

²Faculty of Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia

Article Info

Article history:

Received May 15, 2023

Revised Oct 31, 2023

Accepted Nov 15, 2023

Keywords:

Convolutional neural network

Mel spectrogram

Melfrequency cepstral coefficients

Tempo recognition

ABSTRACT

This article explains how to determine the tempo of the kendhang, an Indonesian traditional melodic instrument. This research presents novelty as technological research related to gamelan instruments, which has rarely been achieved thus far, through the introduction of kendhang tempo types through the sounds produced, with the hope of creating an automatic system that can recognize the kendhang tempo during a gamelan performance. The testing in this work will categorize the tempo of kendhang into three categories: slow, medium, and fast, utilizing one of the two scenario models proposed, mel frequency cepstral coefficients (MFCC) and convolutional neural network (CNN) in the first scenario, and mel spectrogram and CNN in the second. Kendhang's original audio data, which was captured in real time and later enhanced, makes up the data set. The model 1 scenario, which entails feature extraction using MFCC and classification using the CNN classification approach, is the best scenario in this research, based on the experimental results. When compared to the other suggested modeling scenarios, model 1 has a level of 97%, an average accuracy, and a gain value of 96.67%, making it a solid assistant in terms of kendhang's good tempo recognition accuracy.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Muljono

Faculty of Computer Science, Universitas Dian Nuswantoro

Semarang 50131, Indonesia

Email: muljono@dsn.dinus.ac.id

1. INTRODUCTION

In principle, any musical instrument that produces sound and can be arranged in some way by musicians can be called a musical instrument, including a traditional Javanese musical instrument, namely the Javanese gamelan. The important acoustic parameters of Javanese gamelan are the pitch, frequency, tempo, and direction of sound played during gamelan performances [1]. This parameter is strongly influenced by the acoustic properties of each instrument, the position of the musician on the instrument played, the arrangement of the instruments on stage, and the acoustic characteristics of the room and the stage where the instrumental ensemble is played [2]. In a gamelan performance, several kinds of musical instruments are played simultaneously in the same notation. Each instrument in gamelan has its characteristics and creates a specific melody, which cannot be replaced or duplicated by another. One of the instruments in gamelan is the kendhang. Kendhang itself in gamelan performances is the main musical instrument to set the tempo of gamelan performances. Recognizing the sound of musical instruments, including in this case the sound of kendhang, is a fundamental problem in audio signal processing. The separation of two or more signals is very important in finding the characteristics of the signals. In addition, the separation of identical instruments can also be used for sound recognition from musical instruments [3]. Although it is currently difficult to define

the differences between each signal for various musical instruments, musical signals can be categorized with considerable specificity [4]. As a result, several projects and automatic classification methods have been set up forward recently. For example, to extract audio signals using mel frequency cepstral coefficients (MFCC) and classifiers like a k-nearest neighbors and support vector machines (SVM), use the recognition dataset of musical instruments prevalent in the Indian subcontinent, such as the harmonium, flute, monochord (ektara), wooden drum (dhol), tawala, and violin. To achieve the best classification performance, the results of this experiment use a classification algorithm that SVM with radial basis function (RBF). In the test set of the obtained data set, very high accuracy (97%) was attained [5]. In another study, experiments on three music databases with various characteristics—the western music collection (ISMIR 2004 database), the latin American music collection (ISMIR database LMD data), and the collection of African ethnic music—were conducted using the convolutional neural network (CNN) method in comparison to the SVM classifier. This experiment demonstrates that CNN is superior to other classifiers in several situations, making it a highly intriguing choice for music genre identification [6]. Another study, which classified input in the form of the mel spectrum from the audio signal of traditional Chinese musical instruments, in general, using an 8-layer CNN, achieved a classification accuracy of 99.3%. In addition, this study carried out additional trials in which the features were retrieved using the Res-Net model, followed by the classification of all the instruments using the SVM method, with an accuracy of 99.9% [7].

Although technological study on gamelan instruments is still hardly ever done, there have recently been several researchers, such as Sari *et al.* [4], who have noticed the emergence of gamelan musical instruments, evaluating the impact of window length on spectral characteristics retrieved using superimposed short-term fourier transform (STFT). The suggested approach demonstrates that it can generate an F-measure greater than 0.80 for some methods by adjusting the window length and selecting the proper dynamic threshold parameter [4]. Additionally, Tjahyanto *et al.* [8] employs the principal component approach and spectrum-based feature sets as feature extraction for the sound classification of gamelan instruments with the SVM method on RBF kernel [8]. These four categories of gamelan instruments are demung, saron, peking, and bonang. The results of the tests indicate that spectrum-based feature sets have an average F-size that is greater than appearance-based features. While the recognition of distinct tones for the musical instrument demung (63.89%) is lower than that for the saron (83.79%), the former is higher [8].

In terms of the dataset in the form of recorded audio signals and the kind of extraction procedure, the relevant experimental findings mentioned above are comparable to those of this work. The kendhang was the musical instrument utilized in this study, there were fewer tempo classes, and different characteristics of the extraction technique were used before the classification process, compared to the experiments conducted in the studies mentioned above. In this article, the researcher proposes several kendhang sound recognitions as a classification of kendhang tempo types. Just like audio data in general, of course, the audio data of the kendhang sound contains attributes that represent the tempo of the kendhang, so the audio recording data of the kendhang sound must be extracted to obtain sound object information. Several sound feature extraction methods can be used, but in this study, the feature extraction schemes used include MFCC and mel spectrogram. The MFCC itself is a sound signal feature extraction process that can recover the important information contained in the signal, produce data as minimal as possible without losing important information, and adapt the human sense of hearing to perceive sound [9]. Meanwhile, this research will also use a feature extraction scheme using the mel spectrogram method. The mel spectrogram is the result of feature mapping taken using the MFCC method, which will be classified and included in the classification method [10]. In addition, after the feature extraction process, the kendhang tempo recognition process in this study uses a CNN to perform the classification, this method is used for the classification process because it is very efficient to represent special models that allow extracting various features from the sound extraction process [11].

Reviewing some of the issues described above and the research that has been done before. This research presents a novelty as research related to gamelan instruments from a technological point of view, which until now has rarely been done through the introduction of kendhang tempo with the CNN method from the extraction of kendhang sound features using MFCC and mel spectrogram methods. Recognizing the types of kendhang tempos through the sounds produced by the kendhang is the goal to be achieved in this research, so it is hoped that an automatic system will be created that can recognize kendhang tempos when a gamelan performance is put on stage, so accuracy in recognizing that type of tempo is important. The rest of this document is divided into the following sections. Section 2 is an elaboration of research related to sound recognition on various musical instruments, including traditional gamelan musical instruments, as well as a comparison of classification methods that have been used. Section 3 contains the research design, including the datasets and analysis techniques used. The presentation of the experimental results can be found in section 4 of this article and ends with the results that are concluded in section 5.

2. EXPERIMENTAL SETUP

2.1. Research framework

This experimental study aims to create two classification models for detecting tempo types of kendhang, a traditional Javanese percussion instrument. The kendhang tempo classification research framework, shown in Figure 1, was applied in this study. The research is separated into various steps, beginning with the collection of kendhang audio data and breaking it into two sections, training data and test data. The MFCC approach is used in model 1 to extract audio features from kendhang data, and the results of this extraction are processed with the use of a CNN for the classification of kendhang tempo types.

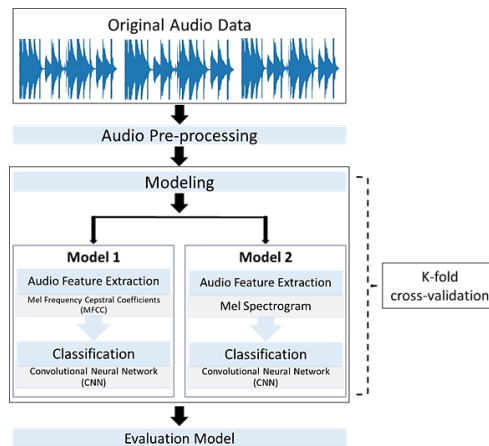


Figure 1. Research framework for classification of tempo kendhang

Model 2 employs a different technique for audio feature extraction, using the mel spectrogram. Following feature extraction, CNN was used to do classification. These two models were then put to the test to evaluate how effectively they could distinguish between different types of kendhang tempos. Criteria like accuracy, precision, recall, and F1-score are used to evaluate classification performance. This study framework aids in visualizing the research steps, from data collection to analysis of findings. The trial results will be published in a study report, along with conclusions stating which model is more effective at categorizing kendhang tempo types, as well as suggestions for further refinement and development.

2.2. Data acquisition

The kendhang sound recording was completed professionally in a soundproof recording studio environment. Collaboration between skilled kendhang players and the recording technical team is essential for achieving the best possible sound quality. Sophisticated devices and equipment are employed to record the sound of the kendhang. Adobe audition, a dependable audio processing software, is used to record, edit, and treat kendhang sounds. The Shure SM57 microphone is connected to the computer through an M-audio external sound card. The Shure SM57 microphone is a great choice for recording musical instruments because of its clean sound and excellent response to recorded sound sources. The recording specifications were also extensively considered; a 48 kHz sample rate was used to capture the kendhang sound, guaranteeing a high degree of aural clarity. To produce audio with the same quality as a CD, a 16-bit resolution is utilized. The kendhang instrument's features dictated that mono channel recording be used, and the resulting files were saved in the uncompressed *.wav format, which preserves the original audio quality. The kendhang sound is captured as accurately as possible thanks to the expert sound recording technique and top-notch gear, which also serve as a strong foundation for additional analysis like audio feature extraction or application in music research and development.

A total of 120 kendhang sound recordings in *.wav file format were used in this study. The recordings were divided into three tempo categories: slow, medium, and fast, each with 40 recordings. Technical notes were taken during the audio recording process about the following: the room, hardware (such as an external M-audio sound card and Shure SM57 microphone), sampling frequency, the distance between the microphone and the kendhang, and the rhythm of the kendhang sound when played. Figure 2 shows the kendhang tempo audio dataset, which is divided into three categories: fast in Figure 2(a), medium in Figure 2(b), and slow in Figure 2(c). This illustration includes two essential components of audio analysis: frequency and time. In this instance, frequency refers to the frequency spectrum of the recorded sound, and

the time axis shows how the sound has evolved. By displaying the frequency and timing characteristics of this dataset, the study was able to gain a better knowledge of the differences between the various types of kendhang tempos used in this experiment. This data is essential for training and testing models that classify kendhang tempos.

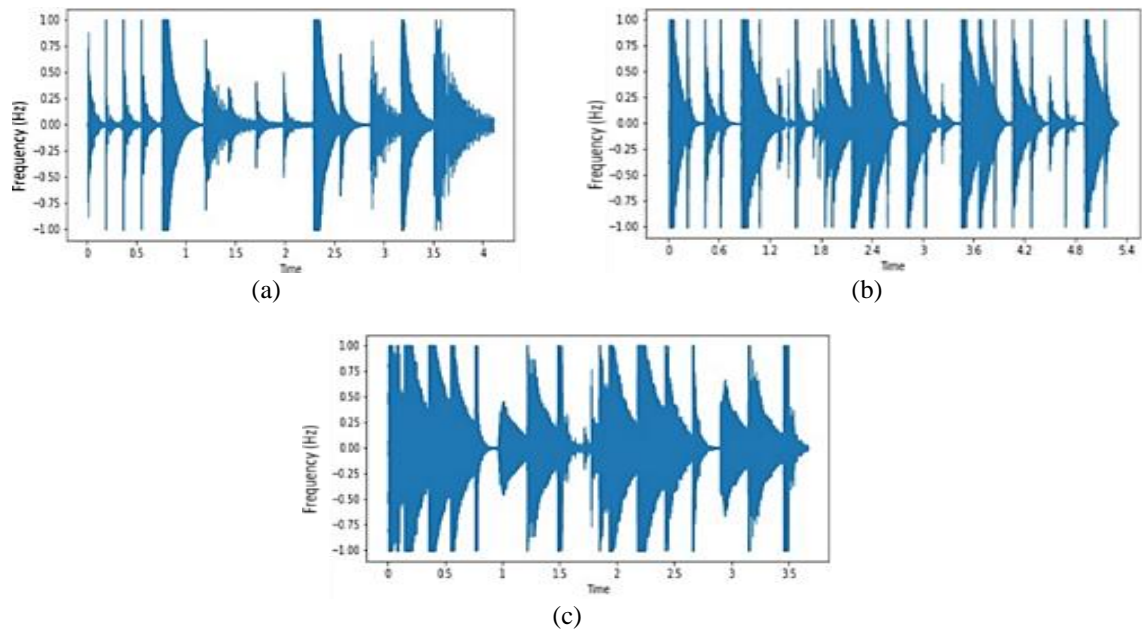


Figure 2. Visualization of the (a) fast kendhang tempo audio dataset, (b) medium kendhang tempo types, and (c) slow kendhang tempo types

2.3. Audio pre-processing step

Before proceeding to the pre-processing stage, the obtained dataset will go through a framing process, which is the process of cutting the kendhang tone recording at a predetermined time. The augmentation process is used in this study's pre-processing step. The audio data is augmented by adding several variations, including noise [12], which is a method for adding random values (noise) input to audio data, with an added value of 0.005 for each audio file. The pitch shifting method is the next variation of this augmentation; this method changes the pitch of the audio data without changing the speed or duration of the audio data at random [13], so that the audio duration remains constant and only the pitch is changed; the added value is between 0.8 and 1. The final variation in the data addition process is to randomly add up the amplitude values [14], with the added value ranging between 1.5 and 3. As a result of this augmentation process, the dataset has grown from 120 audio files to 1200 audio files. Furthermore, the audio data generated by the augmentation process, which is still a file with the *.wav extension, is clipped to the sound signal for the same duration of five seconds for each audio file, a process known as windowing [15]. The datasets in this study will be divided later. Following the pre-processing stage, the dataset will be divided into training and test data. For the division, the stratified splitting method was used. The stratified splitting method shuffles all of the data before dividing it into training and testing sets for each class [16]. The training and test sets are split in an 80:20 ratio.

2.4. Concept of a convolutional neural network

One of the deep learning strategies is the CNN [17]. The CNN is a derivation of the multilayer perceptron (MLP), a form of deep neural network [18]. CNN is intended to have an input (input) array with at least two dimensions. Whereas CNN and MLP both function in similar ways, CNN represents each neuron in two dimensions, while MLP only allows for one. A convolutional network also referred to as a CNN, is a particular kind of neural network with a topology resembling a grid that is used for data processing [19]. A CNN is a name given to a network that makes use of the convolution mathematical technique [20]. Convolution is a linear process in and of itself. A neural network that uses minimal convolution in one of its layers is said to be convolutional [21].

An illustration of the CNN architecture implemented in this study is shown in Figure 3. The following part will go over each proposed model's architecture. The existence of convolutional and pooling layer pairings is what makes CNN unique [22], [23]. The convolutional layer uses two-dimensional input data to parse sub-matrix filters (strides) to extract structured information. By fusing the data from the step submatrices into a single value, the pooling layer condenses the output of the convolution matrix. The CNN architecture also has some fully connected layers, with the classifiers located in the topmost (final) layer [23].

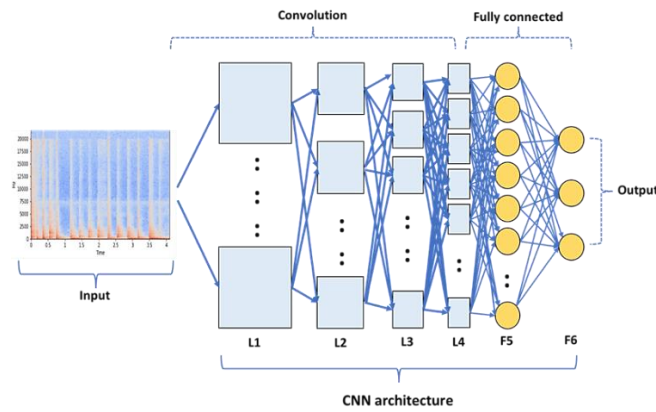


Figure 3. The typical architecture of a CNN

2.5. Concept of the proposed method for model 1

The experiment's proposed model 1 uses the feature extraction method known as MFCC to generate spectrograms of the input data for each kind of kendhang sound tempo [24]. This process converts the audio input into a time-frequency map, which provides valuable information on the acoustic characteristics of the kendhang. MFCC is used to transform the original audio data into a spectrogram, which is a graphic representation of the audio frequency spectrum across time. With the help of spectrograms, one can evaluate several aspects of audio, including rhythm, tempo, and other sound attributes, by understanding the way the frequency energy of kendhang sounds varies over time. The study's Figure 4 shows the spectrogram for each kind of kendhang tempo. Understanding the variations in the frequency spectrum between slow, medium, and fast tempos is made much easier with the help of this representation. These spectrograms will be used as input data in model 1's modeling methodology. This research uses advanced audio techniques (MFCC and spectrogram analysis) to handle kendhang data in a way that improves the detection of kendhang tempo types and allows for more accurate modeling. Figure 4 clarifies discrepancies that could be hard to notice in the raw audio data, providing a fuller understanding of the kendhang's sound qualities at different tempos.

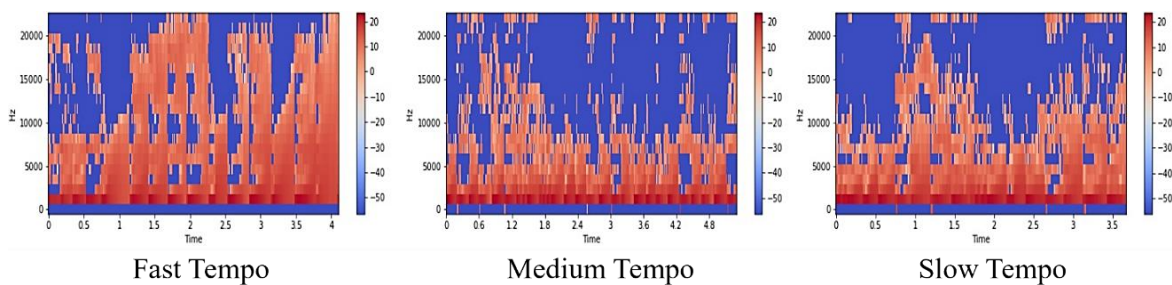


Figure 4. The spectrogram data for each type of tempo for model 1

Sound feature extraction using the MFCC is a well-known and popular technique. A sound signal is used as the method's input, and an MFCC feature is produced as the method's output. One sound signal will result in several feature vector lines since each frame creates one feature vector. An audio signal is shown on a time domain graph in digital signal processing. However, the sound wave will be transformed into a vector in the MFCC [25]. A vector of numbers is returned by MFCC. The vector is a property (aspect) of a signal

that is shown as a spectrogram and contains a map of the intensity (energy) of the gathered spectrum [26]. The classification approach utilized in this study, the CNN method, is used to classify the kendhang tempo type utilizing the vector produced by the feature extraction procedure. The overall model 1 process is shown as follows in Figure 5.

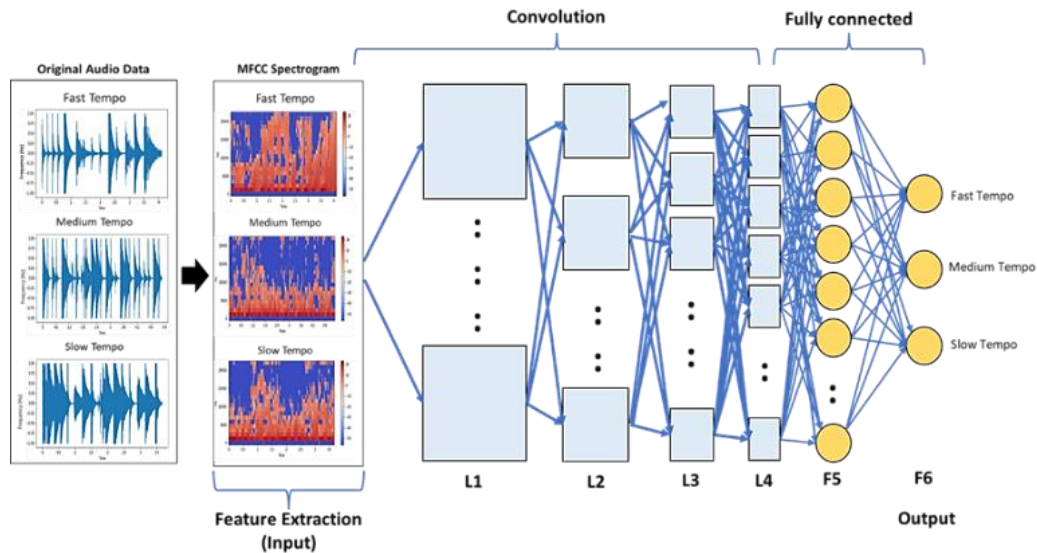


Figure 5. Architecture of model 1 for CNN classification and MFCC feature extraction

Figure 5 depicts the architectural plan of model 1 for this study. Initially, using the MFCC approach, features are retrieved from the original kendhang tempo data. The spectrogram produced by this feature extraction approach captures the essential components of the kendhang audio signal. This spectrogram is used as input in the next stage, which entails using a CNN for classification. In the CNN classification stage, the spectrogram obtained from the MFCC feature extraction method is used as input data. In this perspective, spectrograms are considered images. CNN is a form of neural network design that can recognize complex patterns in picture data. CNN will analyze and process this spectrogram to uncover patterns that distinguish between different kendhang tempos. The classification of kendhang tempo kinds is the outcome of the CNN classification step. As a result, model 1 employs the MFCC technique to convert kendhang audio data into a spectrogram, and CNN is then used to categorize the type of kendhang tempo based on this spectrogram. Figure 5 displays the workflow from MFCC feature extraction to the CNN classification process as a visual depiction of model 1's architectural design. Based on spectral analysis of the audio signal, this method can provide a greater knowledge of kendhang tempo types.

2.6. Concept of the proposed method for model 2

The feature extraction procedure in the proposed model 2 differs significantly from the feature extraction process in model 1 because it uses the mel spectrogram approach to extract features from the original kendhang tempo data. The output of feature extraction using the mel spectrogram approach is an audio signal's frequency spectrum, often known as a spectrogram because it changes over time. Melody is the abbreviation for it. This suggests that it is a measurement of a pitch-based perceptual scale. The mel spectrogram, which combines the mel scale and spectrogram, is a time-domain visual depiction of the frequency and amplitude of the sound [27]. Figure 6 displays the spectrogram data for each type of tempo.

The outputs of this feature extraction procedure are fed into the CNN approach, which is the classification technique employed in this study, to classify the various tempo kendhang kinds. The overall model 2 process is shown in Figure 7. The feature extraction process as input, as shown in Figure 7, which uses the mel spectrogram approach to extract the original kendhang tempo data, is the main difference between model 2's architectural design and model 1's. This process later produces results in the form of a kendhang tempo-type classification. Models 1 and 2 will both undertake validation processes at the modeling stage utilizing the K-fold or K-fold cross-validation approach, as shown in the research framework of Figure 1. Finding out how well the proposed models perform for each model is the goal of this validation.

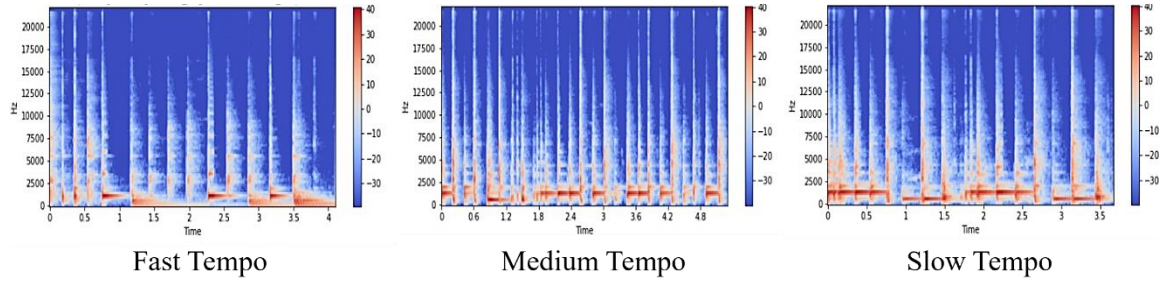


Figure 6. The spectrogram data for each type of tempo for model 2

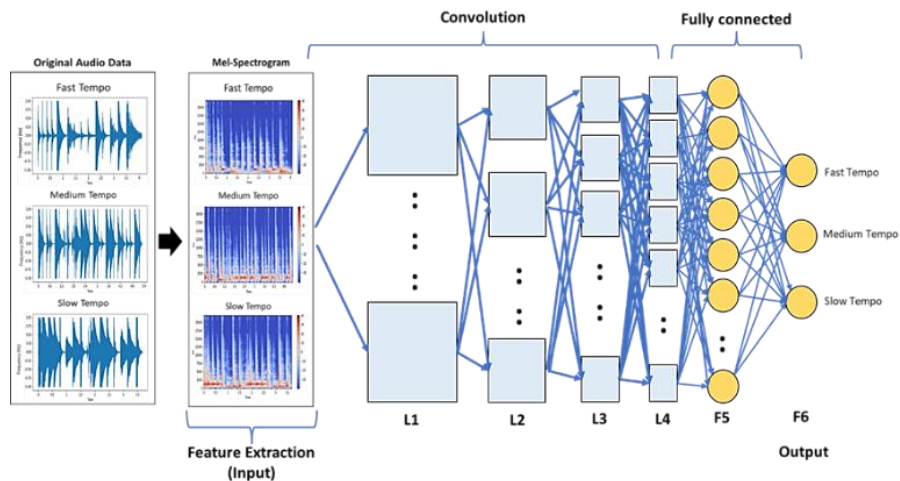


Figure 7. Architecture of model 2 for CNN classification and mel spectrogram feature extraction

A cross-validation model called K-fold validation separates data with multiple k values and iterates/repeats up to k values. One of the key things to do with K-fold is to minimize fluctuations during the process model training, as well as to provide stable output and support the provision of reliable training errors [28]. K-fold validation is a cross-validation model that works by dividing data by multiple k values and iterating through k values. Badža and Barjaktarović [29] employs a K-fold validation scenario with a k iteration value of 10. To achieve the best precision value, experiments will be performed using a value of k=10. The training and test variations each utilize a combination of 10 pieces, and the iteration happens ten times [29].

2.7. Evaluation model

The model evaluation stage is the last step in the research framework for this experiment, and it is described in this sub-chapter. This stage is important to evaluate the effectiveness of the models created using the confusion matrix approach by the corresponding model suggestions. The confusion matrix is a matrix with rows and columns that displays the predictive accuracy using actual data [30]. This confusion matrix is used to assess the accuracy, precision, and recall of the classifier model in this study. The model that performs well is the one that receives the highest score across all three metrics.

$$Accuracy = \frac{\sum_{i=1}^n N_{ii}}{\sum_{i=1}^n \sum_{j=1}^n N_{ij}} \tag{1}$$

$$Precision_i = \frac{N_{ii}}{\sum_{k=1}^n N_{ki}} \tag{2}$$

$$Recall_i = \frac{N_{ii}}{\sum_{k=1}^n N_{ik}} \tag{3}$$

The value of N_{ij} is formed in classes C_j and C_i , respectively, where N_{ij} is the number of samples. The ratio of the number of accurate forecasts to all predictions is shown by the accuracy value (1) [31]. The precision

value (2) is the sensitivity value or the accuracy value of the system between the information provided by the system to correctly display the data in a particular class. The recall value (3) is a value that indicates the level of success or specificity in correctly finding information about data of a particular class [32].

3. RESULTS AND DISCUSSION

3.1. Result of kendhang tempo recognition model with MFCC and CNN (model 1)

The feature extraction procedure employing the MFCC technique in model 1 results in a two-dimensional temporal frequency table with a size of 431×20 . The MFCC features recovered from 431-time frames of kendhang audio data are represented in this table. There are 20 retrieved MFCC per frame. To create a prediction model, this table is used to combine feature extraction data from several frames. The training dataset used by this model includes samples of different kendhang tempos. CNN algorithm was used to develop this predictive model. Figure 8 shows the CNN architecture utilized to construct the prediction model, which explains the topology of this neural network. In this context, the MFCC extraction's temporal frequency table is considered an image, and CNNs are a sort of neural network that is good at finding patterns in image data. To train the model on a training dataset, the CNN architecture is used, allowing the machine to learn from varied drum tempos. This algorithm, once trained, may be used to predict and classify different types of drum tempos using previously unseen data.

Model 1's performance can be rigorously evaluated using the K-fold validation scenario technique with 10 iterations. The data is separated into ten equal subset groups in this case. This procedure was repeated ten times during the trial, with one of the ten groups receiving validation data and the other nine groups receiving training data. This is carried out to guarantee a thorough evaluation of the model and its effective generalization to a range of test data sets. The experimental outcomes of model 1 are displayed in Table 1. Table 1 gives an in-depth view of model 1's performance across a range of iterations by providing data on accuracy, validation accuracy, loss, validation loss, and time of the experiment. This aids in the understanding of the model's classification accuracy and consistency for different drum tempo kinds by researchers. The information in this table enables a thorough evaluation of the model's functionality and aids researchers in determining whether the model produces accurate findings.

Layer (type)	Output Shape	Param #
conv2d_12 (Conv2D)	(None, 20, 431, 32)	832
max_pooling2d_12 (MaxPooling2D)	(None, 10, 215, 32)	0
dropout_16 (Dropout)	(None, 10, 215, 32)	0
conv2d_13 (Conv2D)	(None, 10, 215, 64)	18496
max_pooling2d_13 (MaxPooling2D)	(None, 5, 107, 64)	0
dropout_17 (Dropout)	(None, 5, 107, 64)	0
conv2d_14 (Conv2D)	(None, 5, 107, 64)	36928
max_pooling2d_14 (MaxPooling2D)	(None, 2, 53, 64)	0
dropout_18 (Dropout)	(None, 2, 53, 64)	0
flatten_4 (Flatten)	(None, 6704)	0
dense_8 (Dense)	(None, 128)	868480
dropout_19 (Dropout)	(None, 128)	0
dense_9 (Dense)	(None, 3)	387
Total params: 925,123		
Trainable params: 925,123		
Non-trainable params: 0		

Figure 8. Summary of the architecture for model 1

Table 1. Performance of model 1's classifier utilizing cross-validation and parameter configuration

N-Fold (K)	Accuracy	Validation accuracy	Loss	Validation loss	Time (minutes)
1	95.60	92.71	0.11	0.23	00:04:07.91
2	92.71	93.75	0.17	0.20	00:04:07.33
3	95.83	88.54	0.11	0.24	00:04:04.56
4	92.94	91.67	0.22	0.29	00:04:00.36
5	92.82	95.83	0.17	0.17	00:04:04.23
6*	97.45	94.79	0.07	0.13	00:04:00.64
7	95.25	91.67	0.16	0.22	00:03:22.26
8	91.78	86.46	0.25	0.35	00:03:22.29
9	90.97	94.79	0.22	0.20	00:03:28.28
10	92.48	90.62	0.20	0.22	00:03:27.11
Average	93.78	92.08	0.17	0.23	00:04:12.40

Note: *) Best performance

According to the experimental findings for model 1, the average model training process takes 4 minutes and 12 seconds, the average model accuracy is 93.78%, and the loss value, which measures the difference between the actual label and the predicted label, is 0.17. According to Table 1, the experimental validation method for the sixth iteration (k=6), with an accuracy rate of 97.45%, a time of four minutes, and a loss value of 0.07, has the best accuracy when estimating the tempo of kendhang. Figure 9 displays the summary outcomes of the model with k=6 that was run using the Adam optimizer, the 20-epoch hyperparameter, and the loss of sparse categorical cross-entropy.

According to the accuracy and length of the model learning process, model 1 k=6 cross-validation becomes a model in the suggested model 1 scheme, which is then utilized to create the confusion matrix using the test dataset as shown in Figure 10. The confusion matrix displays the model's performance in identifying or predicting the class of kendhang tempo. Model 1 can correctly identify 80 test types of "slow" kendhang tempo, or it can forecast the complete test dataset for the type of slow tempo kendhang in the "slow" class, according to the confusion matrix above. Model 1 can properly predict up to 75 sets of test data by recognizing the "medium" tempo type.

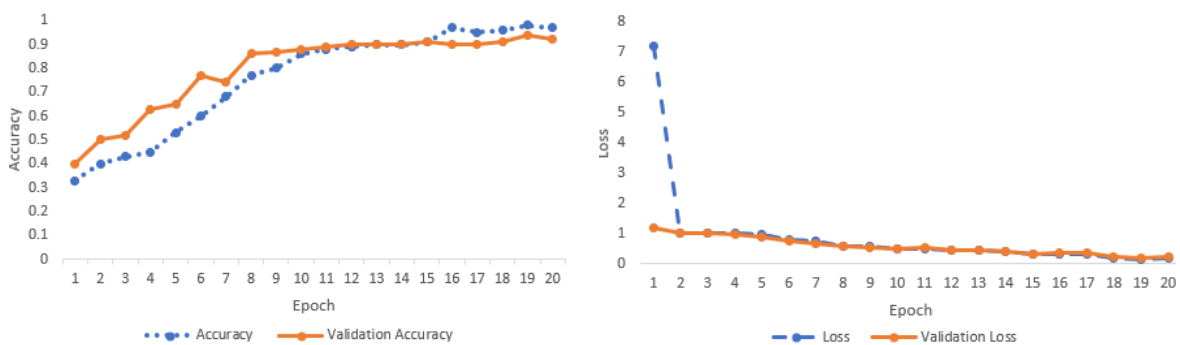


Figure 9. Performance of model 1 on training data with k=6 cross-validation

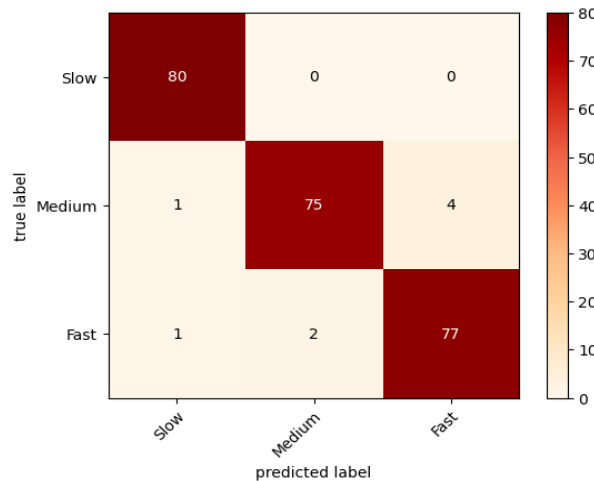


Figure 10. Confusion matrix of model 1

The set of test data must be included in the "medium" tempo type even though one set of test data is wrongly predicted to be of the "slow" tempo type and four sets of test data are incorrectly expected to be of the "fast" tempo type. The audio signal that was recorded as the "medium" tempo type is indicative of this prediction inaccuracy because it resembles other tempo kinds. Three test datasets are wrongly predicted to have this kind of "fast" tempo, with two of them being falsely classified as having a "medium" tempo and one of them being falsely classed as having a "slow" tempo. However, 77 accurate test datasets are classified as "fast" tempo kinds.

3.2. Result of kendhang tempo recognition model with mel spectrogram and convolutional neural network (model 2)

In the proposed model 2, the mel spectrogram technique is applied in the workflow to develop a feature extraction procedure. Following this extraction, a 431×128 two-dimensional time-frequency table is generated. In comparison to the kendhang tempo data, the table provides a more thorough description of auditory aspects in this context. This feature has 431-time frames, with 128 attributes per frame used to characterize audio quality. The data in this time-frequency table is used to develop a prediction model. Using a training dataset and the CNN approach, this model generates a prediction model that can categorize various types of kendhang tempos.

Figure 11 shows the specifics of the CNN utilized in model 2's prediction model design. CNNs are a suitable choice for understanding patterns in image data, and time-frequency tables are considered image data in this sense. Model 2 is designed to classify kendhang tempo types more correctly by employing features derived from kendhang tempo data with the mel spectrogram approach, and it employs the CNN method to discover patterns in audio data that may be difficult to distinguish. Figure 11 shows the model architecture, allowing for a better understanding of how time-frequency table features are used by the CNN for classification. The results of the K-fold validation scenario with an iteration value of 10 will therefore technically be carried out in 10 experimental iterations in order to provide the best estimate of the model in model 2. Table 2 presents the model 2 experimental results.

The experimental results of model 2 show that the average training session lasts 24 minutes and 30 seconds. The models' accuracy is 73.70% on average. Calculating the difference between the actual label and the projected label yields a loss value of 0.56. For the ninth iteration ($k=9$), the experimental validation process is shown in Table 2 with an accuracy rate of 93.98%. The model with $k=9$ has the best ability to predict kendhang tempo, with a loss value of 0.07. Figure 12 displays the summary outcomes of a 20-iteration run of a model with $k=9$ utilizing hyperparameters, sparse categorical cross-entropy reducers, and the Adam optimizer.

Layer (type)	Output Shape	Param #
conv2d_12 (Conv2D)	(None, 128, 431, 32)	832
max_pooling2d (MaxPooling2D)	(None, 64, 215, 32)	0
dropout (Dropout)	(None, 64, 215, 32)	0
conv2d_1 (Conv2D)	(None, 64, 215, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 32, 107, 64)	0
dropout_1 (Dropout)	(None, 32, 107, 64)	0
conv2d_2 (Conv2D)	(None, 32, 107, 64)	36928
max_pooling2d_2 (MaxPooling2D)	(None, 16, 53, 64)	0
dropout_2 (Dropout)	(None, 16, 53, 64)	0
flatten (Flatten)	(None, 54272)	0
dense (Dense)	(None, 128)	6946944
dropout_3 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 3)	387
Total params: 7,003,587		
Trainable params: 7,003,587		
Non-trainable params: 0		

Figure 11. Summary of the architecture for model 2

Table 2. Performance of model 2's classifier utilizing cross-validation and parameter configuration

N-Fold (K)	Accuracy	Validation Accuracy	Loss	Validation Loss	Time (Minutes)
1	87.15	82.29	0.4	0.51	00:24:02.11
2	70.37	80.21	0.68	0.53	00:24:00.54
3	83.45	93.75	0.4	0.24	00:23:56.24
4	91.9	94.79	0.22	0.17	00:23:49.39
5	69.68	85.42	0.75	0.46	00:23:55.88
6	40.15	30.21	1.08	1.09	00:27:03.10
7	87.96	90.62	0.31	0.24	00:24:32.62
8	34.03	33.33	1.1	1.1	00:24:14.34
9*	93.98	89.58	0.17	0.33	00:25:39.70
10	78.36	87.5	0.53	0.39	00:24:51.82
Average	73.70	76.77	0.56	0.51	00:24:30.47

Note: *) Best performance

Based on the length of the learning process and the outcomes of the performance evaluation, Model 2 with the $k=9$ cross-validation approach was determined to be the best model. This indicates that

Model 2 requires less training time in this arrangement and provides excellent accuracy. The test data set is used to test this model after the best model has been chosen. The test's outcomes are displayed as a confusion matrix, demonstrating how well the model categorizes the different kinds of kendhang tempos. The confusion matrix is visualized in Figure 13, demonstrating how effectively model 2 can identify the kind of kendhang tempo found in the test data. The number of accurate classifications (true positives), false positives (false positives), and false negatives (false negatives) are all represented in the confusion matrix. Metrics including accuracy, precision, recall, and F1-score, which give a general picture of the model's performance in categorizing different kinds of kendhang tempos, can be assessed from this confusion matrix.

For up to 73 data sets out of 80 data sets, which are the test data sets, the scheme 2 models can correctly predict the kendhang tempo in the "medium" class. Seven data sets were unpredictable, with five predictions in the "fast" class and two in the "slow" class. The scheme 2 model's precision in predicting the "medium" class kendhang tempo is 91.25%, expressed as a percentage. Model 2 correctly predicts the "slow" class tempo type test dataset with 83.75 percent accuracy or 67 of the 80 a testing data. While the rest were predicted incorrectly, that is, six data sets were classified into the "medium" class and the other seven were classified into the "fast" class. Meanwhile, model 2 has an 86.25% prediction accuracy in recognizing "fast" type kendhang tempos, with 69 out of 80 test data correctly classified. While the rest of the data was unpredictable, three of them were predicted incorrectly in the "slow" class and eight in the "medium" class.

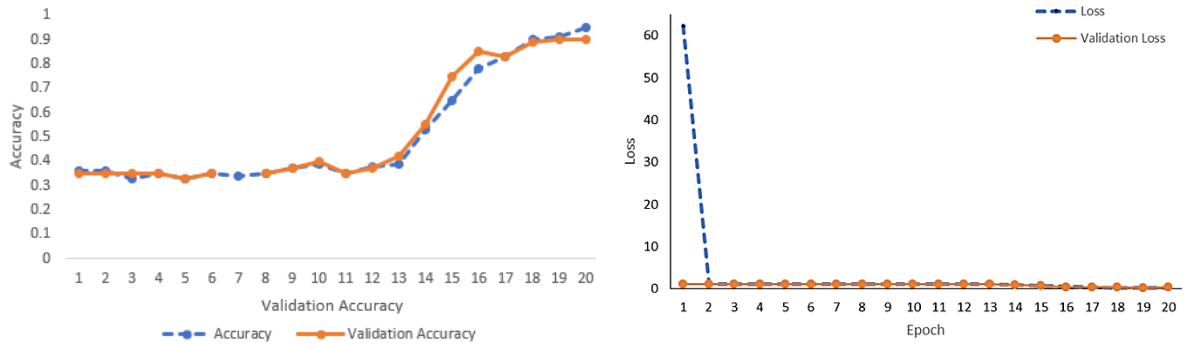


Figure 12. Performance of model 2 on training data with $k=9$ cross-validation

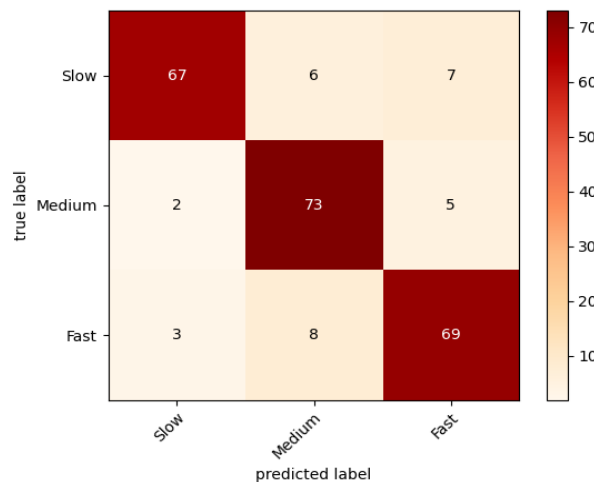


Figure 13. Confusion matrix of model 2

3.3. Evaluation model for each best model scheme

The best modeling outcomes from each research design were assessed and compared in this study to better understand the differences in model accuracy. Figure 14 depicts the variations in the accuracy of each model provided in this research as a result of this comparison. This diagram depicts the performance

differences between models 1 and 2. The epoch value in this experiment was set to 20, which specifies the number of iterations or training cycles completed by the model. In terms of accuracy, a comparison of models 1 and 2 will provide a clearer idea of which model is better at distinguishing kendhang tempo types.

Figure 14 illustrates the results of the hyperparameter adjustment for the epoch 1 experiment, which produced accuracy values of 33% for model 1 and 35% for model 2. Additionally, there is a contrast in the two models' accuracy differences. Model 1's average is greater than model 2's. The point with the greatest accuracy disparity between the two models, or 53.82%, is reached when the epoch value is set to 11. The accuracy of model 1 at epoch 11 was 91.78%, which was two times better than model 2's accuracy of 37.96%. The final accuracy value produced is directly proportional to the epoch value changing and increasing. An epoch value of 20 results in a final accuracy value of 97.45% for model 1 and 93.98% for model 2. In Figure 14, the accuracy value increases exponentially over 20 epochs for both models 1 and 2. Every scheme's best model is assessed for accuracy as well as how well it performs in terms of measurement precision and recall. Figure 11 depicts the proposed research model's precision and recall measures. Precision and recall are metrics used to assess system performance. Precision is the correspondence between the portion of the data retrieved and the information required. The precision value can also be referred to as the sensitivity value or the system accuracy value. It is based on the information provided by the system to display the precise type of tempo based on its type.

Figure 15 depicts the results of the data analysis. The average precision value for model 1 was 96.67%, with the highest precision value of 98% in the "slow" class kendhang tempo. This graph demonstrated that 98% of the correct signal type ("slow") tempo of the entire signal predicted "slow" tempo type. The precision values for the "medium" and "fast" tempo types have the same interpretation. Meanwhile, model 2's average precision is 87.3%, which is lower than model 1's average precision. Precision is 84% and 85% for "medium" and "fast" class kendhang tempos, respectively; these values are close together. This is due to the similarity of the signals from "medium" and "fast" tempos, which affects precision in each class. As shown in Figure 15, which depicts the recall variance of each model proposed in this study, the recall value is a value that indicates the level of success or specificity in determining the correct information about tempo class data. This value specifies how much of the signal is predicted for a given tempo when compared to the total signal for that tempo. On average, the recall value for model 1 is 96.67, which is higher than the recall value for model 2. As an example, the lowest recall value in model 2 for the "slow" tempo type is 84%. This value indicates that 84% of the signals predicted to be of the "slow" tempo type are of the "slow" tempo type, compared to 100% of the signals that are of the "slow" tempo type. If the data test for the "slow" tempo type includes up to 80 datasets, then 84% of the signals are predicted to be of the "slow" tempo type.

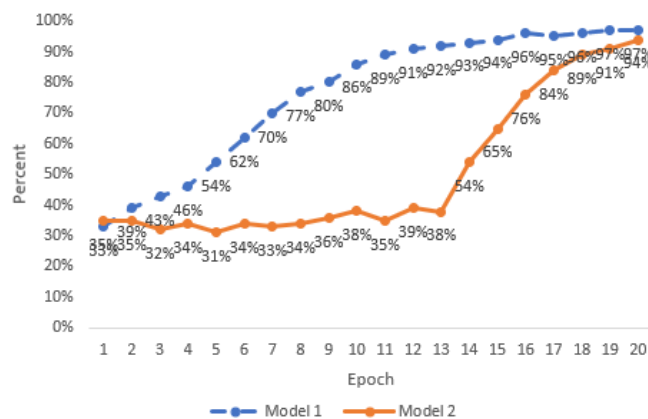


Figure 14. Comparison of the accuracy of each proposed model

Model 1 has the highest accuracy score of 97% based on comparisons of the accuracy, precision, and recall values of the suggested models. Aside from that, model 1's average precision and recall are approximately 96.67%. In simpler terms, model 1 exhibits outstanding precision and memory together with a high degree of accuracy in drum tempo classification. To put it briefly, the model in scenario 1 uses MFCC to extract features from kendhang audio signals, while CNN is used for CNN classification. According to the evaluation results, model 1 is a great fit for this task and effectively identifies the different types of drum tempos with good memory, precision, and accuracy.

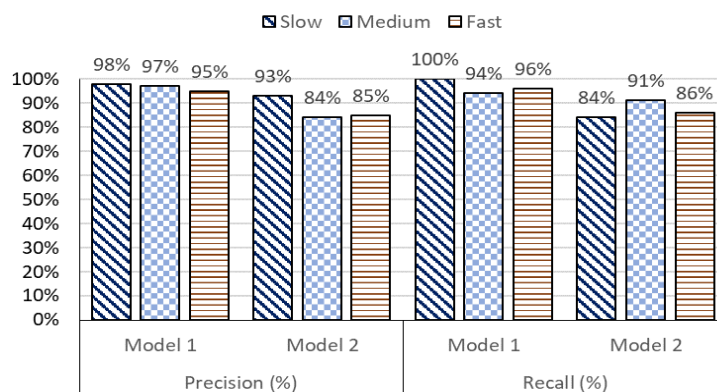


Figure 15. Comparison of performance evaluation of the best schematic models (precision and recall values)

4. CONCLUSION

This article describes the CNN application to distinguish between three different kendhang tempos. This study takes a different approach from other studies in that it focuses on understanding the tempo of traditional Javanese instruments, which are rarely played. Other findings reported in the scientific literature, on the other hand, attempt to identify the sounds of instruments used in modern music. This study classifies the kendhang tempos into three categories: slow, medium, and fast. Our strategy is solely based on the kendhang instrument's speed. The number of audio files increased from 120 to 1200 in total. A spectrogram of the audio sample, which is the output of feature extraction using the MFCC for model 1 and the mel spectrogram for model 2, is provided by CNN as input. Eighty percent of the whole dataset is used to train the architecture, while the remaining twenty percent is used to test it. To enhance the performance of the kendhang tempo recognition system, modeling makes use of model 1, specifically feature extraction using MFCC and classification utilizing the CNN classification approach after feature extraction. In comparison to other proposed modeling systems, the proposed model 1 performs well in terms of accuracy for good kendhang tempo recognition, with a 97% rate and with an average precision and recall value of 96.67%. The results of this study serve as a technological contribution to the study of gamelan musical instruments and as a reference point for the development of kendhang tempo musical instruments.

ACKNOWLEDGEMENTS

The authors of this study express sincere gratitude to the Ministry of Education, Culture, Research, and Technology. This article is the research result funded by the Ministry under Hibah Penelitian Dasar Unggulan Perguruan Tinggi for two years 2021-2022 (Grant: 7/061031/PB/SP2H/AK.04/2022). Furthermore, the authors would also like to thank Universitas Dian Nuswantoro for the continuous support in completing this study.





REFERENCES

- [1] J. Becker, *Traditional Music in Modern Java. Gamelan in a Changing Society*. Honolulu, USA: University of Hawaii Press, 2023, doi: 10.2307/768214.
- [2] S. Gokulkumar, P. R. Thyla, L. Prabhu, and S. Sathish, "Measuring methods of acoustic properties and influence of physical parameters on natural fibers: a review," *Journal of Natural Fibers*, vol. 17, no. 12, pp. 1719–1738, 2020, doi: 10.1080/15440478.2019.1598913.
- [3] A. R. Hermawan, E. M. Yuniarno, and D. P. Wulandari, "Gamelan demung music transcription based on STFT using deep learning," *JAREE (Journal on Advanced Research in Electrical Engineering)*, vol. 6, no. 2, pp. 68–74, 2022, doi: 10.12962/jaree.v6i2.276.
- [4] D. K. Sari, D. P. Wulandari, and Y. K. Suprpto, "Training performance of recurrent neural network using RTRL and BPTT for gamelan onset detection," *Journal of Physics: Conference Series*, vol. 1201, no. 1, pp. 1–9, 2019, doi: 10.1088/1742-6596/1201/1/012046.
- [5] H. Anuz, A. K. M. Masum, S. Abujar, and S. A. Hossain, "Musical Instrument classification based on machine learning algorithm," *Lecture Notes in Networks and Systems*, vol. 164, pp. 57–67, 2021, doi: 10.1007/978-981-15-9774-9_6.
- [6] Y. M. G. Costa, L. S. Oliveira, and C. N. Silla, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied Soft Computing Journal*, vol. 52, pp. 28–38, 2017, doi: 10.1016/j.asoc.2016.12.024.
- [7] R. Li and Q. Zhang, "Audio recognition of Chinese traditional instruments based on machine learning," *Cognitive Computation and Systems*, vol. 4, no. 2, pp. 108–115, 2022, doi: 10.1049/ccs2.12047.
- [8] A. Tjahyanto, D. P. Wulandari, Y. K. Suprpto, and M. H. Purnomo, "Gamelan instrument sound recognition using spectral and facial features of the first harmonic frequency," *Acoustical Science and Technology*, vol. 36, no. 1, pp. 12–23, 2015, doi: 10.1250/ast.36.12.




- [9] S. Ali, S. Tanweer, S. Khalid, and N. Rao, "Mel frequency cepstral coefficient: a review," in *Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020, 27-28 February 2020, Jamia Hamdard, New Delhi, India*, 2021, pp. 1–10, doi: 10.4108/eai.27-2-2020.2303173.
- [10] T. Tran and J. Lundgren, "Drill fault diagnosis based on the scalogram and MEL spectrogram of sound signals using artificial intelligence," *IEEE Access*, vol. 8, pp. 203655–203666, 2020, doi: 10.1109/ACCESS.2020.3036769.
- [11] A. A. Hidayat, T. W. Cenggoro, and B. Pardamean, "Convolutional neural networks for scops owl sound classification," *Procedia Computer Science*, vol. 179, pp. 81–87, 2021, doi: 10.1016/j.procs.2020.12.010.
- [12] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *Journal of Neuroscience Methods*, vol. 346, pp. 1–55, 2020, doi: 10.1016/j.jneumeth.2020.108885.
- [13] W. Bian, J. Wang, B. Zhuang, J. Yang, S. Wang, and J. Xiao, "Audio-based music classification with DenseNet and data augmentation," in *PRICAI 2019: Trends in Artificial Intelligence*, Cham: Springer, 2019, pp. 56–65, doi: 10.1007/978-3-030-29894-4_5.
- [14] O. George, R. Smith, P. Madiraju, N. Yahyasoltani, and S. I. Ahamed, "Data augmentation strategies for EEG-based motor imagery decoding," *Heliyon*, vol. 8, no. 8, pp. 1–14, 2022, doi: 10.1016/j.heliyon.2022.e10240.
- [15] S. P. Dewi, A. L. Prasasti, and B. Irawan, "The study of baby crying analysis using MFCC and LFCC in different classification methods," in *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, 2019, pp. 18–23, doi: 10.1109/ICSIGSYS.2019.8811070.
- [16] M. Bhagat and B. Bakariya, "Implementation of logistic regression on diabetic dataset using train-test-split, K-Fold and stratified K-Fold approach," *National Academy Science Letters*, vol. 45, no. 5, pp. 401–404, 2022, doi: 10.1007/s40009-022-01131-9.
- [17] O. Ghorbanzadeh, T. Blaschke, K. Gholamnia, S. Meena, D. Tiede, and J. Aryal, "Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection," *Remote Sensing*, vol. 11, no. 2, pp. 1–21, 2019, doi: 10.3390/rs11020196.
- [18] S. Kolagati, T. Priyadarshini, and V. M. A. Rajam, "Exposing deepfakes using a deep multilayer perceptron – convolutional neural network model," *International Journal of Information Management Data Insights*, vol. 2, no. 1, pp. 1–8, 2022, doi: 10.1016/j.ijime.2021.100054.
- [19] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, "Fundamental concepts of convolutional neural network," *Intelligent Systems Reference Library*, vol. 172, pp. 519–567, 2019, doi: 10.1007/978-3-030-32644-9_36.
- [20] M. Sargül, B. M. Ozyildirim, and M. Avci, "Differential convolutional neural network," *Neural Networks*, vol. 116, pp. 279–287, 2019, doi: 10.1016/j.neunet.2019.04.025.
- [21] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6105–6114.
- [22] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 53, p. 74, Dec. 2021, doi: 10.1186/s40537-021-00444-8.
- [23] S. Jin, X. Wang, L. Du, and D. He, "Evaluation and modeling of automotive transmission whine noise quality based on MFCC and CNN," *Applied Acoustics*, vol. 172, pp. 1–11, 2021, doi: 10.1016/j.apacoust.2020.107562.
- [24] E. Franti, I. Ispas, and M. Dascalu, "Testing the Universal baby language hypothesis-automatic infant speech recognition with CNNs," in *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, 2018, pp. 1–4, doi: 10.1109/TSP.2018.8441412.
- [25] M. Yildirim, "Automatic classification and diagnosis of heart valve diseases using heart sounds with MFCC and proposed deep model," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 24, 2022, doi: 10.1002/cpe.7232.
- [26] M. D. Pawar and R. D. Kokate, "Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients," *Multimedia Tools and Applications*, vol. 80, no. 10, pp. 15563–15587, 2021, doi: 10.1007/s11042-020-10329-2.
- [27] T. Zhang, G. Feng, J. Liang, and T. An, "Acoustic scene classification based on Mel spectrogram decomposition and model merging," *Applied Acoustics*, vol. 182, p. 108258, 2021, doi: 10.1016/j.apacoust.2021.108258.
- [28] Z. Mushtaq, S. F. Su, and Q. V. Tran, "Spectral images based environmental sound classification using CNN with meaningful data augmentation," *Applied Acoustics*, vol. 172, pp. 1–15, 2021, doi: 10.1016/j.apacoust.2020.107581.
- [29] M. M. Badža and M. C. Barjaktarović, "Classification of brain tumors from mri images using a convolutional neural network," *Applied Sciences*, vol. 10, no. 6, pp. 1–13, 2020, doi: 10.3390/app10061999.
- [30] H. Xu, W. Zeng, X. Zeng, and G. G. Yen, "An evolutionary algorithm based on Minkowski distance for many-objective optimization," *IEEE Transactions on Cybernetics*, vol. 49, no. 11, pp. 3968–3979, 2019, doi: 10.1109/TCYB.2018.2856208.
- [31] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Information Sciences*, vol. 340–341, pp. 250–261, 2016, doi: 10.1016/j.ins.2016.01.033.
- [32] A. H. Villacis, S. Badruddoza, A. K. Mishra, and J. Mayorga, "The role of recall periods when predicting food insecurity: A machine learning application in Nigeria," *Global Food Security*, vol. 36, pp. 1–29, 2023, doi: 10.1016/j.gfs.2023.100671.

BIOGRAPHIES OF AUTHORS






Muljono     holds a Doctor of Electrical Engineering degree from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia in 2016. He joined an internship program at School of Media Science, Tokyo University of Technology Japan in 2014. He received his Magister of Computer (Informatics) from STTIBI Jakarta, Indonesia in 2001 and he received his B.Sc. (Mathematics) from Universitas Diponegoro (UNDIP) in 1996. He is currently an associate professor at Department of Informatics Engineering in Dian Nuswantoro University, Semarang, Indonesia. His research includes artificial intelligence, machine learning, data mining, data science, and natural language processing. He has published over 90 papers in international journals and conferences. He can be contacted at email: muljono@dsn.dinus.ac.id.






Pulung Nurtantio Andono    holds a Doctor of Electrical Engineering degree from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia in 2014. He received his Magister of Computer (Informatics) from Universitas Dian Nuswantoro, Indonesia in 2009 and he received his B.Sc. (Informatics) from Universitas Trisakti Jakarta, Indonesia in 2006. He is currently a professor at Department of Informatics Engineering in Dian Nuswantoro University, Semarang, Indonesia. His research includes artificial intelligence, computer vision and image processing. He has published over 80 papers in international journals and conferences. He can be contacted at email: pulung.nurtantio.andono@dsn.dinus.ac.id






Sari Ayu Wulandari    holds a Bachelor of Engineering (B.Eng.) from Universitas Diponegoro (UNDIP) in 2007 and Master of Engineering (M.Eng.) from Universitas Gadjah Mada (UGM) in 2013. She is currently lecturing with the Department of Electrical and Biomedic Engineering at Dian Nuswantoro University, Semarang, Indonesia. She is a member of the Indonesia society of engineers and the institute of electrical and electronics engineers (IEEE). His research areas of interest include biomedic, artificial intelligence, and digital signal processing. She can be contacted at email: sari.wulandari@dsn.dinus.ac.id.



Harun Al Azies    holds a Master of Statistics degree (M.Stat.) from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2022. He received his B.Sc. in Statistics from PGRI Adi Buana University, Surabaya, Indonesia, in 2021 through a cross-track program after obtaining an associate degree (A.Md.) in Business Statistics from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2017. He is currently a junior lecturer in Department of Informatics Engineering at Dian Nuswantoro University, Semarang, Indonesia. His research includes applied statistics, machine learning, data science, and material informatics. He has published more than 30 papers in international journals and conferences. He can be contacted at email: harun.alazies@dsn.dinus.ac.id.



Muhammad Naufal    holds a Bachelor's degree (S.Tr.T.) in Informatics Engineering at Harapan Bersama Polytechnic, Tegal, Indonesia, in 2017. He was continuing his Master in Informatics Engineering at Dian Nuswantoro University, Semarang, Indonesia, with a Magister of Computer (M.Kom.) degree in 2022. He is currently a junior lecturer in Department of Informatics Engineering at Dian Nuswantoro University, Semarang, Indonesia. His research interest is in computer vision, and he has published several of them in scientific journals and conference proceedings. He can be contacted at email: m.naufal@dsn.dinus.ac.id