Tempo Recognition Of Kendhang Instruments Using Hybrid Feature Extraction

Muljono[1*], Pulung Nurtantio Andono[1], Sari Ayu Wulandari[2], Harun Al Azies[1], and Muhammad Naufal[1]

[1] Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang 50131, Indonesia
[2] Department of Electrical Engineering, Universitas Dian Nuswantoro, Semarang 50131, Indonesia
[*] Corresponding author. E-mail: muljono@dsn.dinus.ac.id

Received: Jan. 16, 2023; Accepted: Jun. 14, 2023

This article is the result of research on Gamelan instruments that examines from a technological perspective what is rarely done nowadays, through kendhang tempo recognition by proposing three classification modeling schemes. The proposed scheme is a new approach to kendhang tempo classification, using kendhang sound converted to image-based features via Mel spectrogram, then features are extracted from the image with Visual Geometry Group (VGG)-19 before incorporating the method K-Nearest Neighbour (K-NN) as a classification method. Based on the experimental results that have been obtained, modeling using the 3rd scheme, namely two-phase feature extraction from the Mel spectrogram image as the first phase and the second phase of VGG-19 with classification using K-NN has an advantage in accuracy (99.6%) of implementing Kendhang tempo recognition correctly and the average achievement of the fastest training processing time was 3.37 seconds compared to the 1st scheme with an accuracy of 94% and an average model training process time of 16.4 seconds and the 2nd scheme with a model accuracy of 98% and the average time to complete the model training process the longest is 6228.6 seconds.

Keywords: Features Extraction, K-Nearest Neighbour, Mel spectrogram, Sound Recognition, VGG-19

http://dx.doi.org/10.6180/jase.202403_27(3).0004

1. Introduction

The Gamelan musical instrument is a traditional art instrument of Javanese culture, Indonesia, which in its development is used to accompany the performance of wayang art as well as a performance of Javanese tribal customs. Gamelan consists of several instruments which are played simultaneously. Gamelan performances are traditional musical ensembles that typically feature music produced from a combination of gamelan musical instruments such as gongs, kenong, kempul, saron, bonang, kendhang, and others and can produce harmonious musical compositions [1] All gamelan instruments are played by striking or dragging the sound source to obtain a signal having a variety of frequencies. Kendhang or drum is a gamelan instrument, one of the main functions of which is to set the rhythm or tempo [2]. The important role of kendhang makes the player of kendhang always placed as the leader of Gamelan performance. The kendhang produces sound by being beaten or struck by hand without any tools on the membrane. When struck or beaten, this membrane then vibrates. It is this vibration of the drum membrane that produces the sound [2].

Kendhang musical instruments are usually played by professional gamelan players who have long been immersed in Javanese culture and most kendhang are played according to the sensitivity of the kendhang player, which means that the standardization of the resulting tones will be different for each player, so that if played by one person

with another person, the rhythm or the resulting tempo will be different, thus affecting other instruments in gamelan performances. Therefore, we need a system that can recognize the sound of this kendhang musical instrument, or in other words, synthesize the Kendhang signal to convert the signal into a reference signal that can automatically recognize the tone of this kendhang as a key standardization that will be useful for other instruments as they become an integrated ensemble in gamelan performances.

Research related to gamelan instruments is an interesting topic because so far research on gamelan or its instruments has mostly been done from an artistic or musical point of view, while research on gamelan from a technological view has rarely been conducted [3, 4]. Research on gamelan from a technological point of view is still little done, but recently several researchers such as Sari, Wulandari, and Suprapto [5], Dewi [6], Tjahyanto [7, 8] who conducted research on Gamelan from a technological perspective. In this research scheme, the researcher uses the sound of the kendhang as a tempo detection tool, the difference in frequency of each note coming out of the kendhang is an important key in the classification of the tempo of this musical instrument.

Sound samples of musical instruments, including kendhangs, have several characteristics or attributes that represent these musical instruments, so the sound or audio data of kendhangs that should be classified as a tempo type must first be extracted. The extraction method used in this study is the Mel-spectrogram. Using the Mel-spectrogram as feature extraction provides efficient results for working on audio in image-based features applied in classification [9]. Moreover, the Mel spectrogram is also considered to have better performance in terms of image-based features, which is in line with the framework of this study which extracts audio datasets into image features or forms of spectrogram and then performs modeling based on feature extraction. results of these images with the Mel spectrogram [10]. Besides using Mel-spectrogram as feature extraction, the scheme of this study will also extract informative features from Mel spectrogram images produced using Visual Geometry Group (VGG)-19 architecture, the selection of VGG-19 is based on the fact that this architecture is more sensitive and reliable in image extraction features (Image-Based Features) [10] and can help models train faster and more efficiently [9] In principle, VGG itself is a CNN architectural model that shows good performance for image classification [11], but experimental results show that the resulting model is not accurate enough in image classification, so the VGG model can be combined with other classification algorithms to improve performance. In this paper, kendhang's tempo recognition process uses the K-Nearest Neighbour (K-NN) to perform the classification because this method is not sensitive to outliers, does not require data input assumptions, and has good performance [12].

Based on some of the above issues and research done previously This study presents a novelty by proposing three schemes of methods used to recognize voices or speech recognition that are used to identify the tempo of the Kendhang instrument using the Mel-spectrogram and VGG-19 feature extraction schemes, followed by the recognition process K-NN to do the classification. The purpose of this article is to detect the tempo of kendhang from the sound it produces, so the accuracy of detecting the sound of kendhang is its main focus because the process that goes through obtaining a subset of features extracted from gamelan recordings using the Mel and VGG-19 spectrogram. The selected feature subsets were then validated using the cross-validation technique and classified using the K-Nearest Neighbour (K-NN) classification method. The remainder of this paper is divided into the following sections. Section 2 describes a review of the literature related to research on gamelan and the methods or algorithms used to identify gamelan sounds, Section 3 covers the methodology used, including a description of the concept of classification and feature extraction algorithms as well as data sources and research flow. Section 4 describes the experiments performed and the results obtained. Finally, Section 5 concludes the study.

2. Related work

In speech or sound recognition, many approaches have been applied to estimate or recognize voices, for example, the study of musical instrument detection from input audio signals initiated by Chakraborty and Parekh [13], this study aims to obtain the efficiency of several features with several approaches of machine learning method such as support vector machines (SVM), K-NN and artificial neural networks (ANN), but the best results of the various experiments performed are the cepstral coefficient (CC) With artificial neural networks (ANN) being the best experimental setup compared to the others, another alternative to avoid tuning can use the K-NN method. Additionally, Tran and Lundgren [14] detected the sound of a broken drill by converting the audio signal of the drill into an image signal in the form of a Mel spectrogram and scalogram images. Classification of these images will be performed using SVM and K-NN machine learning methods. The results of this study indicate that the proposed Mel spectrogram and scalogram image inputs with the proposed

K-NN and SVM methods are the best models with the same accuracy value of 80.25%.

Other research related to sound recognition was conducted by Nugroho who conducted experiments to recognize the voices of various ethnic groups in Indonesia. The experimental proposal of this research is to perform data augmentation to overcome the small number of datasets and feature extraction using the MFCC method, while the approach used is a seven-layer DNN which has model performance with 99.76% accuracy and 0.05 loss [15]. Meanwhile, other applications of the K-NN algorithm in musical instrument sound recognition have been made, one of which is research using Google-Net as a feature extraction method, this study compares the performance of the SVM method with K-NN in the classification of musical instrument sounds in experiments with sixteen types of musical instruments. the results of this study indicate that the good performance of the two methods is the SVM method [16]. Moreover, research by Jeyalakshmi compared the performance of the two classification methods of K-NN and HMM in speech recognition of four different musical instruments, namely the flute, guitar, violin, and piano. The feature extraction methods used in this study are MFCC, PLP, and RASTA-PLP. The results of musical instrument recognition using K-NN have the highest average accuracy when feature extraction is performed with MFCC, which is 85.5% for the same note, while for different grades it is also highest with MFCC of 75% [17]. Meanwhile, in another study using a dataset consisting of 1284 music samples from sixteen types of musical instruments, Prabavathy's research conducted sound detection of musical instruments using AlexNet as a detection method. feature extraction, with the K-NN method used, the obtained accuracy of 98.16% [18].

While research related to Gamelan instruments in terms of technology is still rarely conducted, recently there have been several researchers like what was done by Wulandari, who detected the appearance of gamelan musical instruments. Using spectral features extracted using superimposed short-term Fourier transform (STFT) by examining the effect of window length. The proposed method shows that by setting the window length and setting the appropriate dynamic threshold parameter, it can produce an F-measure greater than 0.80 for some methods. Moreover, Tjahyanto using four types of instruments in Gamelan namely demung, saron, peking, and bonang, this research uses the principal component method and spectrum-based feature sets as feature extraction used for the sound classification of gamelan instruments with the SVM method on RBF kernel. The experiments performed show that spectrum-based feature sets have a higher average F-size than appearance-based features. While the recognition of specific tones for the musical instrument saron (83.79%) is higher than the recognition of specific tones for the musical instrument demung (63.89%) [8].

The related experimental results that have been described above have similarities with this study in terms of the extraction method, namely the Mel spectrogram and the type of dataset in the form of a recorded audio signal. However, there are fundamental differences between some of the above studies and the experiments conducted in this study, namely the musical instruments used in this case the kendhang, the number of tempo classes, and the differences in the extraction process characteristics before the classification process.

3. Methodology

This section will describe the research framework used in this study. In general, the research framework of this study is divided into four stages, data collection (including distribution of training data and test data), pre-processing, modeling, and final model evaluation. The modeling step of this study is divided into three diagrams, each diagram has steps that include feature extraction and the stage of recognition of tempo kendhang sounds through a classification process. In more detail, Fig. 1 below explains the steps performed in this study.

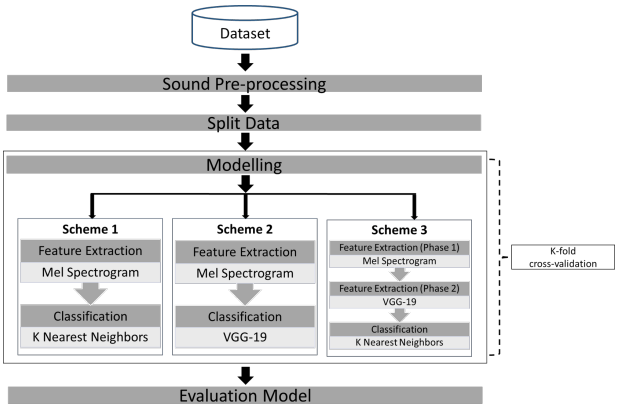

Fig. 1. Research Framework for Classification of Tempo Kendhang.

Based on the description of the proposed research framework in Fig. 1, the steps of this study can be explained through the following steps.

3.1. Dataset and Data Acquisition

The dataset used in this study is the experimental result of the original Kendhang instrument which is part of a

set of Gamelan instruments. The audio recording sample dataset in this experiment is a personal dataset consisting of 120 audio-isolated pitch samples in the data format in the extension .wav which is the standard audio file format, with 40 audios per time class. The sample rate is 44.1 kHz with a hop length of 512 and a Fast Fourier Transform (FFT) of 2048. Fig. 2 below illustrates the dataset of this study, namely the .wav file of the kendhang sound in each tempo class. Based on Fig. 2, it can be seen that each kendhang tempo class has a different size.

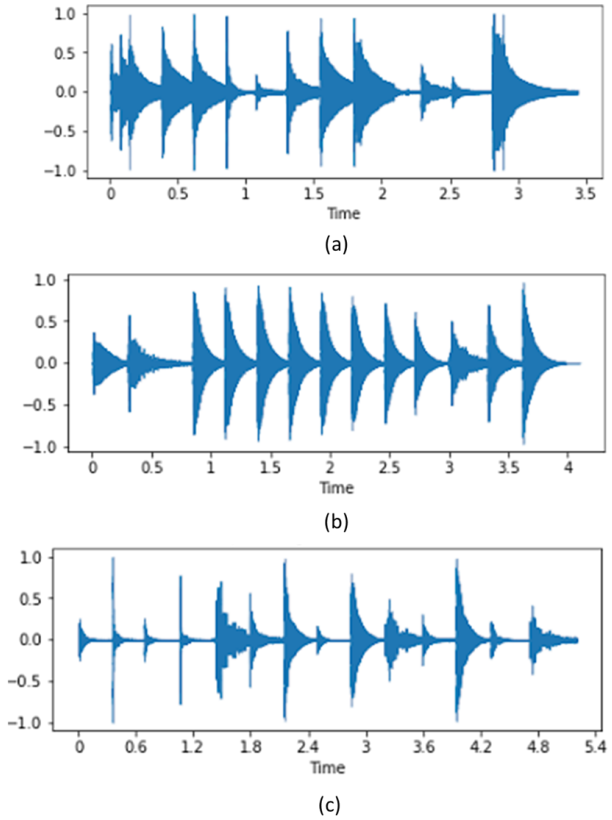

Fig. 2. Visualization of the "slow" kendhang tempo audio dataset (a), "medium" kendhang tempo types (b), "fast" kendhang tempo types (c).

Kendhang sound is recorded with a sampling rate of 44.1 kHz. The process for the kendhang to produce sound samples is to beat/knock each kendhang. In total, this study obtained about 120 isolated tone samples, with tempo levels divided into three, namely "fast", "medium" and "slow" tempos which were then used as classes or labels for each sample. of tone. This labelling is carried out by professionals involved in the performing arts of gamelan.

In carrying out the kendhang sound recording process, the equipment used in this study for the recording included a soundproof recording studio room and recording software such as Adobe Audition, external sound card M-Audio, and microphones AKG C1000S. The people involved in the recording are professional kendhangs and the technical recording team. In the recording process, everything related to the recording technique will be recorded, for example, the distance between the microphone and the kendhang instrument, the sampling frequency, and the associated sound signals, the room used, the tools used (hardware and software), and the rhythm of the kendhang played. This is done so that later if we want to re-record, we can get the same results.

Kendhang sound recording is done with a sampling rate of 44.1 kHz (which can then be up-sampled or down-sampled as needed), with a precision resolution of 16 bits, mono channel type, the duration of the kendhang accompaniment recording is saved as a storage file in the form of .wav. In total, there are 120 wav files of selected drum sound recordings consisting of 40 "slow" tempo/rhythm kendhang sounds, 40 "medium" tempo/rhythm kendhang sounds and 40 "fast" tempo/rhythm kendhang sounds.

3.2. Sound pre-processing

Before going through the pre-processing stage, the dataset obtained will go through a framing process, process cutting the kendhang tone recording is cut according to a predetermined time. The pre-processing step in this study involves an augmentation process. The augmentation process performed on the audio data is to add several variations including noise addition which is a method to add a random value (noise) that is input in the audio data, the added value is 0.005 in each audio file.

The next variation of this augmentation is the Pitch Shifting method, this method changes the pitch of the audio data without changing the speed or duration of the audio data randomly so in this method, the duration of the audio will remain the same and only the pitch will be changed, the added value is between 0.8 and 1. The last variation in the process of increasing the data is to add values of amplitude randomly, while the added values are between 1.5 and 3. From the results of this augmentation process, the dataset becomes 1200 audio files from the original 120 audio files. In addition, the audio data resulting from the augmentation process which is still a file with the *.wav extension is clipped to the sound signal with the same duration of five seconds for each audio file, this process is called a windowing process [19].

3.3. Split data

In this study, the number of datasets will be divided later. When the entire dataset has gone through the pre-

processing stage, the dataset will then be divided into training data and test data. The division method used is the stratified splitting method [20]. The stratified splitting method takes all the data, shuffles the data, and then divides the data into training and testing sets for each class. The ratio of the distribution of training and testing sets is 80:20.

3.4. Modeling

The modeling stage in this study is divided into three schemes, each scheme has stages which include feature extraction and sound recognition stage of kendhang tempo through a classification process. Before explaining each scheme in detail, this sub-chapter will generally describe the feature extraction method used.

3.4.1. Feature extraction

At this stage, it is a question of obtaining a digital signal value and this value will be used as a learning model. It is necessary to extract features in order for the characteristics to be studied for the classification task [21]. The feature extraction procedure on CNN VGG19 involves some operations spanning from convolution to pooling, to extract the significant features from the image. In particular, the feature extraction procedure on the VGG19 CNN begins with convolution, which is a 224x224x3 initial input image processed through a series of convolution layers comprised of small-size convolution filters. Each of these filters iterates over the input image, producing a new feature map. The number of filters in subsequent convolution layers can vary, gradually increasing the complexity of the retrieved features. Following that is activation, a post-convolution step in which each feature map is assigned a nonlinear activation function such as ReLU (Rectified Linear Unit) to introduce nonlinearity in the feature representation. The final step is pooling, which occurs after activation. The maximum pooling procedure is typically used in the pooling process. Clustering takes the largest value in a particular window such as 2×2 on each feature map to decrease the spatial dimension and extract the important characteristics. The stages of convolution, activation, and pooling are repeated numerous times to build a more complicated and abstract feature representation, resulting in a final feature of dimensions (7×7×512). several convolution blocks in the VGG19 architecture feature several convolution and grouping layers.

Based on the scheme of this study shown in Fig. 1. The feature extraction used in this study is the Mel spectrogram which is an algorithm used to calculate the spectrogram of the kendhang sound signal which has been processed before. The results of feature extraction with the Mel spectrogram will later visually represent the kendhang sound signal in terms of frequency and amplitude as a function of the time domain, as shown in Fig. 3.

Then, based on Scheme 3 in Fig. 1, the Mel spectrogram extraction results in Fig. 1 will be the feature extraction using the Visual Geometry Group (VGG-19) method with 19 lattice layers of neurons, this VGG-19 architecture as a model for extracting image features from the extract. Spectrogram Mel. The results of feature extraction from the Mel spectrogram using the VGG-19 architecture are shown in Fig. 4.

3.4.2. Classification scheme

As proposed in the research framework shown in Fig. 1, in this study, the modeling step consists of three schemes. The difference between each of these schemes lies in the differences in the feature extraction and classification methods used. The following explains each diagram at this modeling stage.

a. Classification with K-NN (Scheme 1)
   The kendhang tempo recognition algorithm in this study uses the k-Nearest Neighbors (k-NN) algorithm. In general, the k-NN algorithm measures the distance between the goal center point and a set of observation points in the dataset to assign the goal center point to the most common class among the "k" nearest neighbors surrounding it, this indicates that there is a flag that implies that there must be a measure of distance that can be calculated between samples using independent variables [22]. In this study, the consideration of the distance measure used is the Minkowski distance. Minkowski distance has another form of calculating distances similar to Euclidean and Manhattan, but the powers and roots of p used range from 1 to 2 [23]. The Minkowski distance between two points x and y is.

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \quad (1)$$

   where d is the minkowski distance between data at the center point (x) and data on attribute (y) for each i-th data, where n is the amount of data and parameter p is power. Fig. 5 below illustrates the k-NN classification for three tempo classes of kendhang, namely “fast”, “medium” and “slow” tempo which will be displayed in this study. In this study, the number of neighbors parameter or symbolized by k was used with a configuration of k values, namely 2, 3, and 4. In Fig. 5, the red color represents the "fast" tempo class, while the color yellow represents the "medium" tempo class,

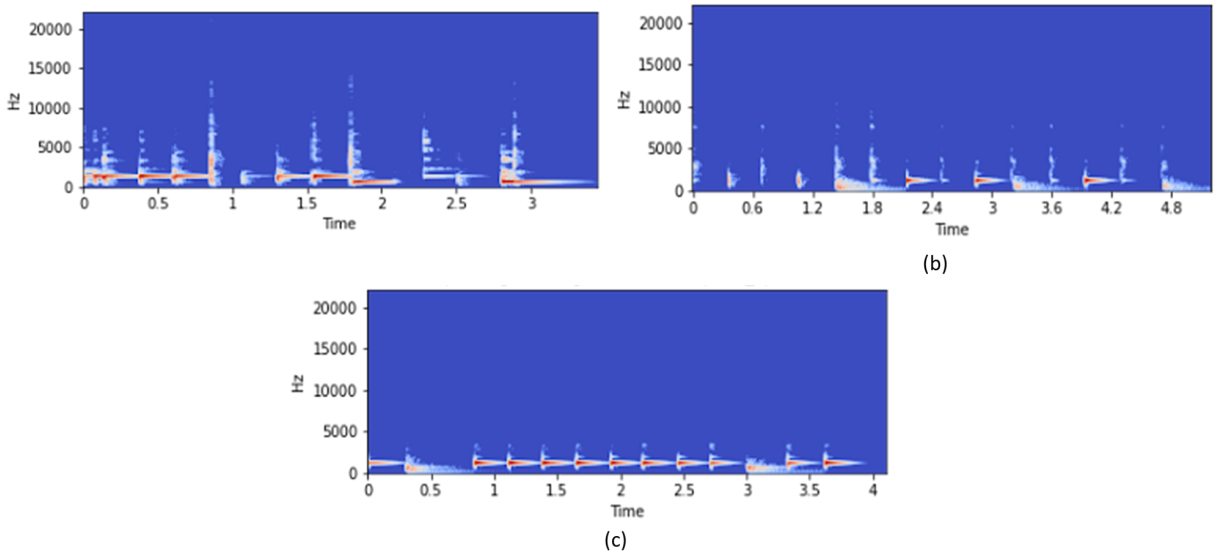

Fig. 3. Visualization of a sample of the results of feature extraction using the Mel Spectrogram for slow kendhang tempos (a), medium kendhang tempos (b), and fast kendhang tempos (c).

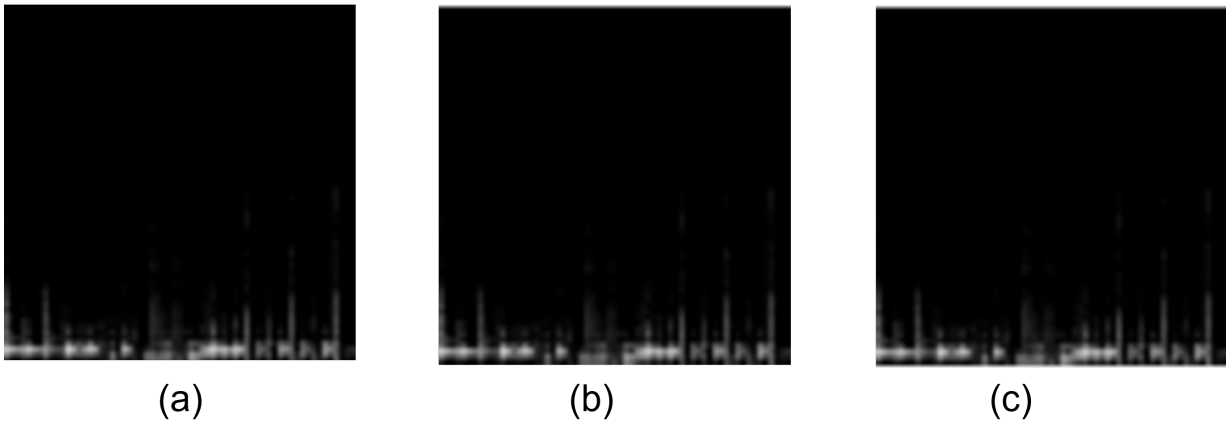

Fig. 4. Visualization of a sample of the results of feature extraction using the VGG-19 for slow kendhang tempos (a), medium kendhang tempos (b), and fast kendhang tempos (c).

and green the "slow" tempo class. While the blue dots are illustrations of the target points.

The illustration of the dotted line inside that forms a circle has a parameter value of $k = 2$, this indicates that the target point is classified as a yellow class or a "medium" tempo kendhang because there are three yellow dots in this area. Another illustration for the value of the parameter $k = 3$ shows that the target is classified in the green class or the kendhang tone with a "slow" tempo.

b. Classification with Visual Geometry Group (VGG)-19 (Scheme 2)

VGG is a CNN architectural model which has good performance for image classification, VGG model was first introduced in the ImageNet competition in 2014, it is a solution to the weakness of CNN, which has a long model training processing time. The VGG-19 architecture has similarities with the VGG-16 architecture, the difference lies only in the presence of 3 additional convolutional layers for the VGG-16 network [24]. With fewer datasets than ImageNet, this research will implement multiple layers into the pre-trained model before continuing with the classification process. The overall the process in scheme 2 is visualized in Fig. 6 as follows.

Fig. 6.a is the architecture of VGG-19 in general, then

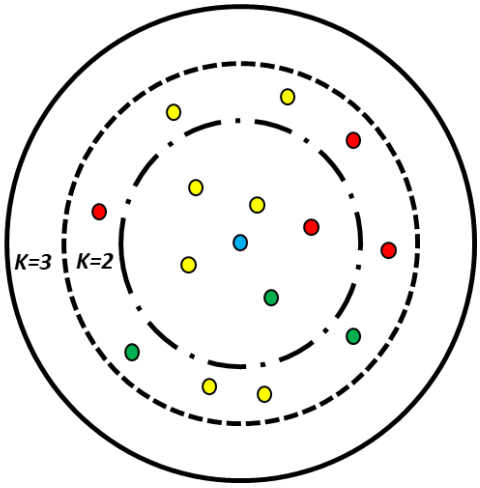

Fig. 5. Representation of the K-NN algorithm.

Fig. 6.b depicts the VGG-19 architectural process in this study where the ImageNet weight is transferred to form the kendhang dataset. In scheme 2, the last layer is replaced by fine-tuning. The purpose of fine-tuning is to train the pre-trained model on the top layer to recognize the features of the dataset we have [25, 26]. The fine-tuning configuration in this study is to avoid underfitting [26], namely by doing a freeze layer on the VGG architecture up to the fourth convolutional block, and adding a fine-tuning layer with the configuration as shown in Fig. 7. The fine-tuning process in this study was carried out by freezing 15 trainables in the first layer of a total of 26 trainable layers held by VGG19. Furthermore, by including max pooling before flattening, the pooling layer is responsible for minimizing the spatial size of the feature map matrix, which is the result of the previous convolution. Many dropout layers are built after max pooling is enabled. Dropout is a neural network regularization technique in which some neurons are selected at random and not used during training; these neurons are deleted at random. Dropping out helps to avoid overfitting and speeds up the learning process. In addition to dropout, batch normalization is performed. Batch normalization is a technique used in convolutional neural networks (CNN) to improve training stability and speed up model training. Batch normalization's major goal is to normalize each batch of data at each CNN layer.

c. Classification with K-NN + VGG-19 (Scheme 3)
Furthermore, scheme 3 is a configuration scheme of the various methods applied to scheme 1 and scheme 2. In principle, VGG shows good performance for image classification, but the experimental results show that the resulting model is not accurate enough in classifying images, so the VGG model can be combined with a classification algorithm to improve its performance [10]. In scheme 3, VGG-19 acts as a feature extraction, the feature extraction performed by VGG-19 is a feature extraction from the Mel Spectrogram extraction results as shown in Fig. 6c. In this study, the features extracted through the pre-training VGG-19 model are continued for the process classification with the K-NN method approach. In more detail, the process in Scheme 3 is as follows.
At the modeling stage, validation will be performed for all schemes using the K-fold or K-fold cross-validation method as shown in the research framework in Fig. 1. This validation is performed to determine the performance of the model offered for each scheme. K-fold validation is one of the cross-validation models that work by splitting data with a few k values and iterating/repeating up to k values. K-fold validation is a cross-validation model that works by dividing data with multiple k values and iterating/repeating up to k values, one of the important things to do with K-fold is to minimize fluctuations during the model training process, as well as provide stable output and helping to provide believable training errors [27].
This study uses a K-fold validation scenario with a k iteration value of 10 based on an experiment that was conducted by Yadav [28]. Technically, experiments will be carried out with a value of k = 10 to obtain the best precision value. The iteration occurs 10 times, while the training and testing variants use a combination of 10 parts.
The evaluation stage of this model is the stage to see the performance of the model built according to each schema. The confusion matrix is a commonly used tool in the evaluation of machine learning models [29]. This matrix represents the accuracy of the prediction with the actual data through row and column matrices. Fig. 8 below visualizes the general confusion matrix for classification with several classes, namely C_1, C_2, and C_n classes.
In the confusion matrix shown in Fig. 8, the value N_{ij} is the number of samples in class C_i but classified in class C_j. This confusion matrix is used to evaluate the performance of the classifier model in this study, namely accuracy, precision, recall. The model with good performance is the model with the maximum value on the three indicators.

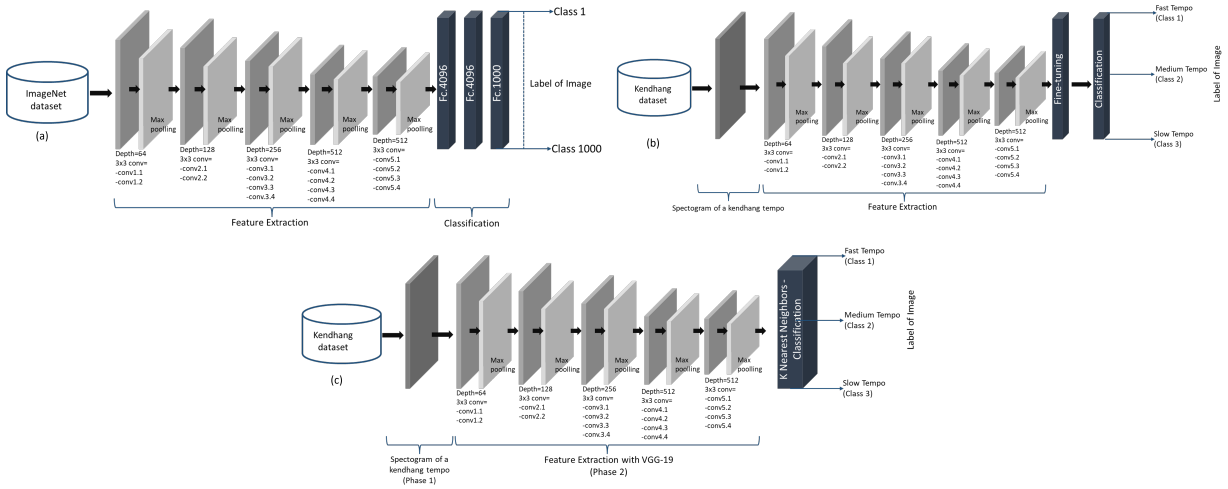

Fig. 6. VGG-19 architecture in general (a), VGG-19 architecture for tempo kendhang dataset classification (Scheme 2) (b), VGG-19 architecture for feature extraction and classification at the using the K-NN method (Scheme 3) (c).

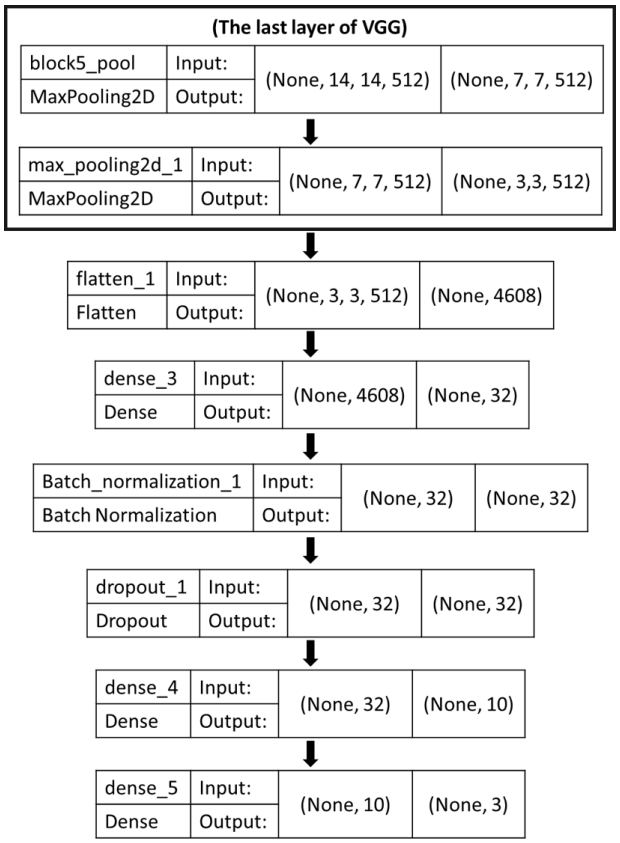

Fig. 7. Fine-tuning of VGG in the research of kendhang tempo classification.

3.5. Evaluation model

The evaluation stage of this model is the stage to see the performance of the model built according to each schema. The confusion matrix is a commonly used tool in the evaluation of machine learning models [29]. This matrix represents the accuracy of the prediction with the actual data through row and column matrices. Fig. 8 below visualizes the general confusion matrix for classification with several classes, namely C_1, C_2, and C_n classes.

In the confusion matrix shown in Fig. 8, the value N_{ij} is the number of samples in class C_i but classified in class C_j. This confusion matrix is used to evaluate the performance of the classifier model in this study, namely accuracy, precision, recall. The model with good performance is the model with the maximum value on the three indicators.

		Predicted		
		C_1	$\cdots\ C_j\ \cdots$	C_n
Actual	C_1	N_{11}	N_{1j}	N_{1n}
	$\vdots$ C_j $\vdots$	N_{i1}	$\cdots\ N_{ij}\ \cdots$	N_{in}
	C_n	N_{nj}	N_{nj}	N_{nn}

Fig. 8. Confusion matrix for classification with multi-class.

The accuracy value (2) represents the ratio between the number of correct predictions and the total number of predictions [29].

$$Accuracy = \sum_{i=1}^{n} N_{ii} / \sum_{i=1}^{n} \sum_{j=1}^{n} N_{ij} \quad (2)$$

The precision value (3) 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 par-

ticular class [29].

$$Precision_i = N_{ii} / \sum_{i=1}^{n} \sum_{k=1}^{n} N_{ki} \quad (3)$$

The recall value (4) is a value that indicates the level of success or specificity in correctly finding information about data of a particular class [29].

$$Recall_i = N_{ii} / \sum_{k=1}^{n} N_{ik} \quad (4)$$

F1-Score is the harmonic mean of precision and recall. The best F1-Score is 1.0 and the worst is 0. Representationally, if the F1-Score has a good score, it indicates that our classification model has good precision and recall [30].

$$F - score_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (5)$$

4. Result discussions

4.1. Experiment results

The purpose of this study is to detect the tempo of kendhang from the sound it produces, so the accuracy of detecting the sound of kendhang is the main goal. In this subchapter, the experimental results will be presented for the three model schemes that were described in the previous sub-chapter.

4.1.1. Extract features using Mel spectrogram and classify using K-NN classifiers (scheme 1)

In this first scheme, the researcher extracted the original audio dataset following the augmentation, namely 12,000 datasets with the same class composition for each kendhang tempo (400 "fast" tempos, 400 tempos "medium" and 400 "fast" tempos) in a Mel spectrogram image with an image size of 224 × 224 × 3. The dataset is split by the stratified splitting method as explained in the previous chapter, the split data composition is 80% (960 data sets) for the training set and 20% (240 data sets) for the test set. Fig. 9 shows the appearance of the Mel spectrogram image with an image size of 224×224×3 results in this sub-chapter.

Moreover, according to the flow of scheme 1, the results of the performed feature extraction will be used to build a predictive model using a training dataset using the K-NN method with different parameter configurations. In this case the parameter k or the number of neighbors. The k parameters used are k = 2, 3, and 4, along with the K-fold validation scenario with an iteration value of 10, and the distance used in the K-NN implementation is the Minkowski distance. Therefore, technically, 10 iterations of experiments will be performed to obtain the best estimate of the model in this scheme 1. Fig. 10 shows the

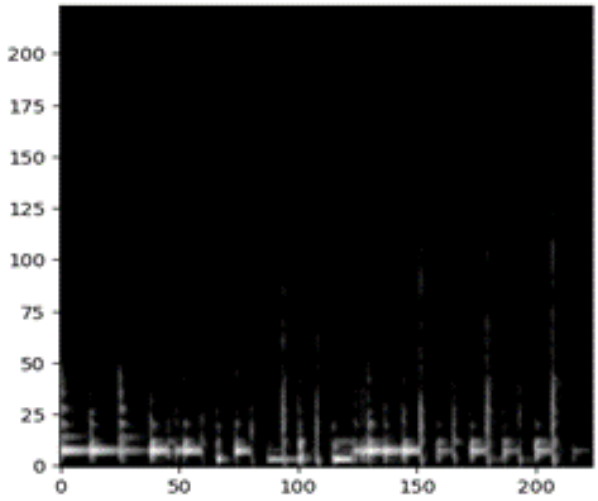

Fig. 9. Sample of the Mel spectrogram image size (224×224×3) augmentation results of the Kendhang tempo dataset.

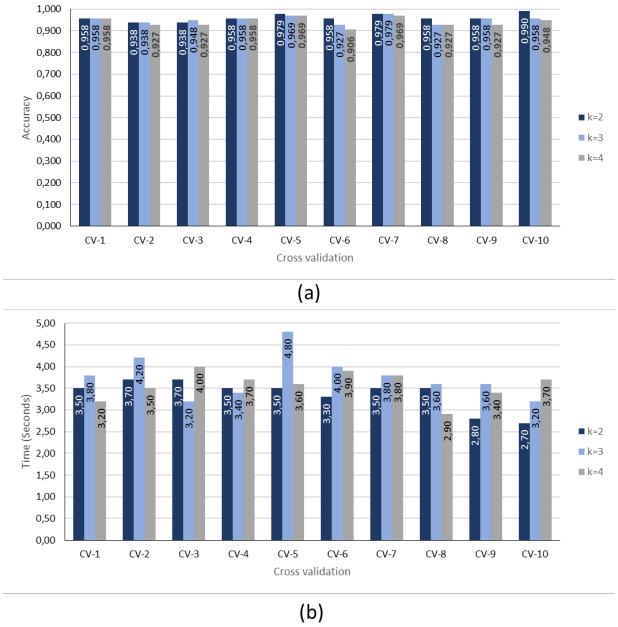

Fig. 10. Classifier performance in Scheme 1 using parameter configuration and cross-validation, accuracy value (a) model training process time (b).

experimental results in scheme 1 for different parameter configurations and cross-validation.

Based on Fig. 10a, K-NN with parameter k=2 in the third iteration (CV-3) is the configuration with the best accuracy (96.9%) when classifying kendhang tempo, this is also aligned where K-NN with parameter k=2 has the best overall average accuracy (93.33%) when classifying

kendhang's tempo compared to K-NN with other k parameters when they are visualized as a function of the model training process time (Fig. 10b).

The average training process time for the model is respectively 16.4 seconds for parameter k=2, 17 seconds for parameter k=3 and 17.1 seconds for parameter k=4, in other words, K-NN with parameter k=2 is the best model in scheme 1 in terms of accuracy and duration of the model learning process which is then used to construct the confusion matrix using the test dataset as shown in Table 1.

The purpose of this study is to detect the type of kendhang tempo based on its sound so that the accuracy of detecting the sound of kendhang is the main focus. From Table 1, 240 data items are included in the test data used to compile the confusion matrix. Based on the confusion matrix with the best model in Scheme 1, namely K-NN with parameter value k=2, it can predict the kendhang tempo in the "slow" class of no less than 82 data sets, then predict the tempo of "medium" kendhang up to 83 sets of data, and predict the tempo of kendhang in "fast" classes up to 75 sets of data. The results of the prediction model are compared with the actual data, so the results of scheme 1 based on Table 1 show that the classification model has accuracy in classifying "slow" kendhang tempos in the "slow" class reaching 95% or there are 76 datasets of "slow" kendhang tempos classified correctly and the rest are classified in the "medium" class. Meanwhile, the model in scheme 1 has the accuracy of classifying kendhang tempo into "medium" and "fast" classes respectively 94% while the others are not classified correctly or in other words, are classified in another tempo class of kendhang.

4.1.2. *Extract Features Using Mel Spectrogram and Classify Using VGG-19 (Scheme 2)*

In contrast to Scheme 1, in this second scheme, the results of the Mel spectrogram process which is an image with an image size of 224 × 224 × 3 (Fig. 9) will be classified with one of the CNN architectural models, namely VGG-19, where VGG-19 works by implementing multiple layers in the pre-training model as a feature extraction step (Fig. 4) before proceeding to the classification process, this process is depicted in more detail in Fig. 6b.

Additionally, according to the flow in Scheme 2 (Fig. 6b), after going through the fine-tuning stages with the fine-tuning configuration in this study, from these results, will be used to build a predictive model using a training dataset with various parameter configurations used in this case using a learning rate value of 0.0001, an epoch of 10, the standard loss function is cross-entropy and using Adaptive Moment Estimation (ADAM) as the optimizer like the experiment conducted by Salem, choosing the Adam optimizer as the solution to improve performance and minimize overfitting [31]. Fig. 11 shows the experimental results in Scheme 2 for different parameter configurations and cross-validation.

As it is known that VGG-19 is a solution to the weakness of CNN which has a long model training process time, based on Fig. 11, it is known that the average model training process time in Scheme 2 is 103.81 minutes, the experiment on CV-2 has the longest processing time compared to other experiments.

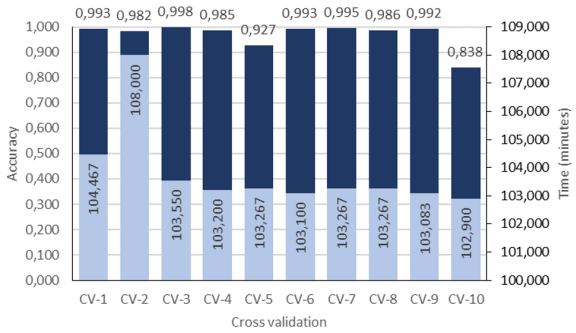

Fig. 11. Classifier performance in scheme 2 using parameter configuration and cross-validation.

In contrast to VGG-19 on CV-3 which has a model training process time fastest and is the model setup with the best accuracy (99.8%) when classifying drum tempo types in this Scheme 2, where the test results, the training dataset on the CV-3 for epochs 1-10 is shown in Fig. 12.

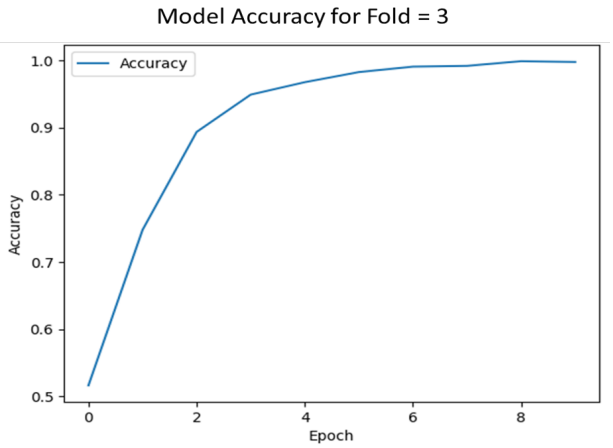

Fig. 12. The results of the performance testing of the training dataset on CV-3 for epoch 1-10.

Therefore, the VGG model on CV-3 is then used to construct a confusion matrix using a test dataset as shown in Table 2.

Table 1. Confusion matrix of scheme 1 (% accuracy).

		Predicted		
		Slow Tempo	Medium Tempo	Fast Tempo
Actual	Slow Tempo	76 (95%)	4	0
	Medium Tempo	5	75 (94%)	0
	Fast Tempo	1	4	75 (94%)

Table 2. Confusion matrix of scheme 2 (% accuracy).

		Predicted		
		Slow Tempo	Medium Tempo	Fast Tempo
Actual	Slow Tempo	78 (97.5%)	2	0
	Medium Tempo	0	80 (100%)	0
	Fast Tempo	0	3	77 (96.3%)

The best classification model built on Schema 2 can correctly classify 80 datasets into the "medium" class (100% accuracy). Meanwhile, two sets of data out of 80 sets of data are not correctly classified into the "medium" class kendhang tempo so the accuracy of the model in scheme 2 in the classification of "medium" class kendhang tempo " is 97.5% and the "fast" class kendhang tempo is 96.3% or there are three data sets that are not correctly classified out of a total of 80 test data sets for the class "fast" of slow tempo.

4.1.3. Extract Features Using Mel Spectrogram with VGG-19 and Classify Using K-NN Classifiers (Scheme 3)

Additionally, Scheme 3 is a layout scheme of the different methods applied in Scheme 1 and Scheme 2. VGG shows good performance for image classification, but in Scheme 3 as depicted in Fig. 6c, VGG-19 is only used for feature extraction from the results Mel-spectrogram, so the VGG-19 extraction results will be continued for the classification process using the K-NN method approach. In this scheme, with the same input size, namely the Mel spectrogram image with an image size of 224 × 224 × 3 which will be extracted using VGG-19, the convolutional layer for the lattice VGG-19 in the last layer is achieved by a flattening process which is an operation that transforms the matrix into a one-dimensional vector, Fig. 13 shows the flow of this process.

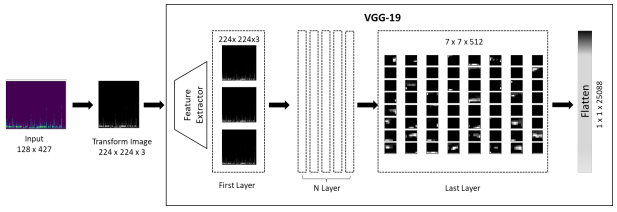

Fig. 13. VGG-19 process flow in Scheme 3.

The flattening process transforms the feature map that was obtained from the previous layer into a vector map [32], so that it can be used to build predictive models using training datasets using the K-NN method with the same configuration of parameters, namely parameters k = 2, 3, and 4, and K-fold validation with an iteration value is 10. Fig. 14 shows the experimental results in Scheme 3 for different parameter configurations and cross-validation.

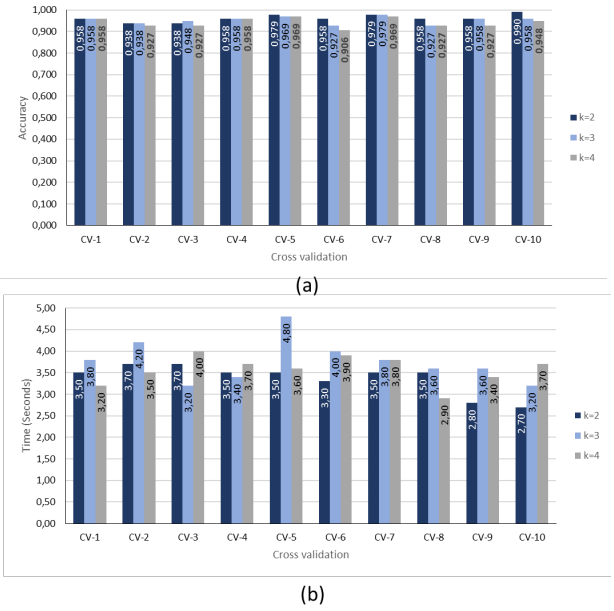

Fig. 14. Classifier performance in Scheme 3 using parameter configuration and cross-validation, accuracy value (a) model training process time (b).

In terms of accuracy and model training processing time in Fig. 14, the model at k=2 and CV-10 is the model with the highest accuracy rate (99%) compared to the other model configurations and the fastest pattern formation processing time (2.7 seconds) when classifying kendhang tempo. Overall, a similar condition is found in the model with k=2 which has the best overall average accuracy (96.1%) when

classifying velocity and based on the model training process time (Fig. 14b), the average time of the model training process at k=2 is the fastest, namely 2.7 seconds. So, in this scheme 3, K-NN with parameter k=2 is the best model to classify the tempo of kendhang and will proceed to prepare the confusion matrix using a test dataset (Table 3).

Table 3 shows the confusion matrix from the test results of the best model in Scheme 3, this model correctly classifies the kendhang tempo for the "medium" class and the kendhang tempo for the "slow" class, this is put highlighted by training 80 sets of data for each class ranked by its class or by its actual data (100% accuracy). Meanwhile, there is a drum tempo dataset for the "fast" class, out of 80 datasets used as a practice dataset, it is not properly classified in the "fast" class, so the accuracy obtained is 98.8%.

4.2. Discussion

The best test results of each scheme in this study will be evaluated and compared with each other as mentioned in Table 4. The best model of each scheme performs differently in terms of the assessment measures of accuracy, precision, recall, and f1-score. The best model in Scheme 3, namely features extraction using Mel-spectrogram + VGG-19 and classification using K-NN classifiers with parameter k = 2, yields the best precision and recall values compared to the two models in scheme 1 and 2, except for the "slow" tempo class for the precision value. This indicates that the model in Scheme 3 is the model that has the best performance compared to the models in Scheme 1 and Scheme 2, this evidence is reinforced by the F1-Score value for the model in Scheme 3 which is the highest compared to models in other schemes, this indicates that the model This classification has good precision and recall. This further explains why the model in Scheme 3 achieves the best results.

Additionally, for an overall performance comparison aimed at the accuracy value, the model in Scheme 3 achieves the highest accuracy value of 0.996 (99.6%) with an average the achievement of the fastest model training process duration of 3.37 seconds. In conclusion, the model in Scheme 3, namely the Mel spectrogram with VGG-19 transfer learning, was found to help improve the performance of K-NN both in terms of computation time and performance in the classification of the tempo of the kendhang sound.

5. Conclusions

The findings of this study have several contributions, this contribution is related to the development of new data on the sound of traditional Indonesian musical instruments, especially for the introduction of the tempo of a traditional Javanese musical instrument, namely the Kendhang, which is one of the contributions to the study of gamelan musical instruments from a technological point of view and provides a source of data on the introduction of gamelan tempo instruments.

The modeling uses the 3rd scheme, namely with two-phase feature extraction, the first phase is the feature extraction step based on Mel-spectrogram images, then in the second phase, these results are extracted using VGG-19, and the final results of feature extraction in the second phase are then classified using the K-NN classification method to improve the performance of the Kendhang tempo recognition system. The model proposed in Scheme 3 has an advantage in performance accuracy of good tempo recognition of Kendhang, which is 99.6%, and has the fastest average training time compared to other schemes proposed modeling.

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* 2021-2022. Furthermore, the authors would also like to thank Universitas Dian Nuswantoro for the continuous support in completing this study.

References

[1] J. M. Arms, P. D. Candidate, and C. Musicology, (2017) *"Presented at the Thirteenth Annual Graduate Student Research Symposium"*:

[2] M. Perlman and C. L. Krumhansl, (1996) *"An Experimental Study of Internal Interval Standards in Javanese and Western Musicians"* **Music Perception *14***: 95–116. DOI: 10.2307/40285714.

[3] A. M. Syarif, K. Hastuti, and P. N. Andono, (2022) *"Traditional Javanese Membranophone Percussion Play Formalization for Virtual Orchestra Automation"* **Proceedings - 2022 IEEE International Conference on Cybernetics and Computational Intelligence, CyberneticsCom 2022**: 381–386. DOI: 10 . 1109 / CYBERNETICSCOM55287.2022.9865292.

[4] A. Wintarti, D. Juniati, and I. N. Wulandari, (2018) *"Classification of Gamelan Tones Based on Fractal Analysis"* **IOP Conference Series: Materials Science and Engineering *288***: 012022. DOI: 10.1088/1757-899X/288/1/012022.

Table 3. Confusion matrix of scheme 3 (% accuracy).

		Predicted		
		Slow Tempo	Medium Tempo	Fast Tempo
Actual	Slow Tempo	80 (100%)	0	0
	Medium Tempo	0	80 (100%)	0
	Fast Tempo	1	0	79 (98.8%)

Table 4. Comparison of the evaluation results of the three proposed model schemes.

Scheme	Average duration of the training process	Accuracy	Precision			Recall			F1-Score		
			Class of Tempo			Class of Tempo			Class of Tempo		
			Slow	Medium	Fast	Slow	Medium	Fast	Slow	Medium	Fast
Scheme 1	16.4 second	0.940	0.930	0.900	1*	0.950	0.940	0.940	0.940	0.920	0.970
Scheme 2	6228.6 second	0.980	1*	0.940	1*	0.970	1*	0.960	0.990*	0.970	0.980
Scheme 3	3.37 second	0.996*	0.990	1*	1*	1*	1*	0.990*	0.990*	1*	0.990*

*: Best performance.

[5] D. K. Sari, D. P. Wulandari, and Y. K. Suprapto, (2019) *"Training Performance of Recurrent Neural Network using RTRL and BPTT for Gamelan Onset Detection"* **Journal of Physics: Conference Series *1201***: 012046. DOI: 10.1088/1742-6596/1201/1/012046.

[6] J. Mantik, I. A. Mirah, C. Dewi, I. Gede, A. Gunadi, and G. Indrawan, (2022) *"Gamelan Rindik Classification Based On Mood Using K-Nearest Neigbor Method"* **Jurnal Mantik *6***: 1693–1702. DOI: 10.35335/MANTIK.V6I2.2592.

[7] A. Tjahyanto, Y. K. Suprapto, and D. P. Wulandari, (2013) *"Spectral-based Features Ranking for Gamelan Instruments Identification using Filter Techniques"* **TELKOMNIKA (Telecommunication Computing Electronics and Control) *11***: 95–106. DOI: 10.12928/TELKOMNIKA.V11I1.895.

[8] A. Tjahyanto, D. P. Wulandari, Y. K. Suprapto, and M. H. Purnomo, (2015) *"Gamelan instrument sound recognition using spectral and facial features of the first harmonic frequency"* **Acoustical Science and Technology *36***: 12–23. DOI: 10.1250/AST.36.12.

[9] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Koňir, (2019) *"Audio-visual emotion fusion (AVEF): A deep efficient weighted approach"* **Information Fusion *46***: 184–192. DOI: 10.1016/J.INFFUS.2018.06.003.

[10] A. Bansal and N. K. Garg, (2022) *"Environmental Sound Classification: A descriptive review of the literature"* **Intelligent Systems with Applications *16***: 200115. DOI: 10.1016/J.ISWA.2022.200115.

[11] M. Bansal, M. Kumar, M. Sachdeva, and A. Mittal, (2021) *"Transfer learning for image classification using VGG19: Caltech-101 image data set"* **Journal of Ambient Intelligence and Humanized Computing *14***: 3609–3620. DOI: 10.1007/S12652-021-03488-Z/TABLES/8.

[12] P. N. Andono, G. F. Shidik, D. P. Prabowo, D. Pergiwati, and R. A. Pramunendar, *"Bird Voice Classification Based on Combination Feature Extraction and Reduction Dimension with the K-Nearest Neighbor"* **International Journal of Intelligent Engineering and Systems *15***: 2022. DOI: 10.22266/ijies2022.0228.24.

[13] S. S. Chakraborty and R. Parekh, (2018) *"Improved musical instrument classification using cepstral coefficients and neural networks"* **Methodologies and Application Issues of Contemporary Computing Framework**: 123–138. DOI: 10.1007/978-981-13-2345-4_10/COVER.

[14] T. Tran and J. Lundgren, (2020) *"Drill fault diagnosis based on the scalogram and MEL spectrogram of sound signals using artificial intelligence"* **IEEE Access *8***: 203655–203666. DOI: 10.1109/ACCESS.2020.3036769.

[15] K. Nugroho, E. Noersasongko, Purwanto, Muljono, and D. R. I. M. Setiadi, (2022) *"Enhanced Indonesian Ethnic Speaker Recognition using Data Augmentation Deep Neural Network"* **Journal of King Saud University - Computer and Information Sciences *34***: 4375–4384. DOI: 10.1016/J.JKSUCI.2021.04.002.

[16] S. Prabavathy, V. Rathikarani, and P. Dhanalakshmi, (2022) *"Musical Instrument Sound Classification Using GoogleNet with SVM and kNN Model"* **Lecture Notes in Networks and Systems *300 LNNS***: 230–240. DOI: 10.1007/978-3-030-84760-9_21/COVER.

[17] C. Jeyalakshmi, B. Murugeshwari, and M. Karthick, (2019) *"HMM and K-NN based automatic musical instrument recognition"* **Proceedings of the International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2018**: 350–355. DOI: 10.1109/I-SMAC.2018.8653725.

[18] S. Prabavathy, V. Rathikarani, and P. Dhanalakshmi, (2020) *"Musical Instruments Classification using Pre-Trained Model"*:

[19] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F. M. Smolle-Jüttner, H. Olschewski, and F. Pernkopf, (2020) *"Multi-channel lung sound classification with convolutional recurrent neural networks"* **Computers in Biology and Medicine *122***: 103831. DOI: 10.1016/J.COMPBIOMED.2020.103831.

[20] T. Pronk, D. Molenaar, R. W. Wiers, and J. Murre, (2022) *"Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment"* **Psychonomic Bulletin and Review *29***: 44–54. DOI: 10.3758/S13423-021-01948-3/FIGURES/1.

[21] A. K. Aggarwal, *"Learning Texture Features from GLCM for Classification of Brain Tumor MRI Images using Random Forest Classifier"*: DOI: 10.37394/232014.2022.18.8.

[22] S. Das and U. R. Jena, (2017) *"Texture classification using combination of LBP and GLRLM features along with KNN and multiclass SVM classification"* **2nd International Conference on Communication, Control and Intelligent Systems, CCIS 2016**: 115–119. DOI: 10.1109/CCINTELS.2016.7878212.

[23] H. Xu, W. Zeng, X. Zeng, G. Y. .-. I. transactions on, and undefined 2018, *"An evolutionary algorithm based on Minkowski distance for many-objective optimization"* **ieeexplore.ieee.org**:

[24] M. Mateen, J. Wen, Nasrullah, S. Song, and Z. Huang, (2018) *"Fundus Image Classification Using VGG-19 Architecture with PCA and SVD"* **Symmetry 2019, Vol. 11, Page 1 *11***: 1. DOI: 10.3390/SYM11010001.

[25] , (2022) *"Chest X-ray Images Analysis with Deep Convolutional Neural Networks (CNN) for COVID-19 Detection"* **EAI/Springer Innovations in Communication and Computing**: 403–423. DOI: 10.1007/978-3-030-72752-9_21/COVER.

[26] S. Chauhan, M. Singh, and A. K. Aggarwal, (2021) *"Data Science and Data Analytics: Artificial Intelligence and Machine Learning Integrated Based Approach"* **Data Science and Data Analytics**: 3–18. DOI: 10.1201/9781003111290-1-2.

[27] K. Pal and B. V. Patel, (2020) *"Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques"* **Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020**: 83–87. DOI: 10.1109/ICCMC48092.2020.ICCMC-00016.

[28] S. Yadav and S. Shukla, (2016) *"Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification"* **Proceedings - 6th International Advanced Computing Conference, IACC 2016**: 78–83. DOI: 10.1109/IACC.2016.25.

[29] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, (2016) *"An improved method to construct basic probability assignment based on the confusion matrix for classification problem"* **Information Sciences *340-341***: 250–261. DOI: 10.1016/J.INS.2016.01.033.

[30] R. O. Alabi, M. Elmusrati, I. Sawazaki-Calone, L. P. Kowalski, C. Haglund, R. D. Coletta, A. A. Mäkitie, T. Salo, A. Almangush, and I. Leivo, (2020) *"Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer"* **International Journal of Medical Informatics *136***: 104068. DOI: 10.1016/J.IJMEDINF.2019.104068.

[31] H. Salem, K. R. Negm, M. Y. Shams, and O. M. Elzeki, (2022) *"Recognition of Ocular Disease Based Optimized VGG-Net Models"* **Studies in Computational Intelligence *1005***: 93–111. DOI: 10.1007/978-3-030-91103-4_6/COVER.

[32] F. Li, H. Tang, S. Shang, K. Mathiak, and F. Cong, (2020) *"Classification of Heart Sounds Using Convolutional Neural Network"* **Applied Sciences 2020, Vol. 10, Page 3956 *10***: 3956. DOI: 10.3390/APP10113956.