Published online on: **http://jurnal.iaii.or.id**

Hyperparameter Optimization of CNN Classifier for Music Genre Classification

Rendra Soerkarta[1], Suhardi Aras[2]*, Ahmad Nur Aswad[3]
[1,2,3]Informatics Engineering Study Program, Muhammadiyah University of Sorong, Sorong, Indonesia
[1]rendrasoekarta@um-sorong.ac.id, [2]suhardi.aras@gmail.com, [3]ahmadnrwwd@gmail.com

Abstract

Playing music through a digital platform that has a large database of songs requires automated classification of music genres, highlighting the need to develop a model for music genre classification that is more efficient and accurate. This study evaluated the hyperparameters in the music genre classification process using the CNN on the GTZAN dataset with 30-second duration data optimized using the MFCC feature extraction. The model that is formed with a time of 3 (three) seconds classifies music genres in the first 3 seconds of music. This model has a high potential for error because the first 3 seconds of initial music is varied and cannot be used as a benchmark in determining music genres. This study performed hyperparameters on batch size, epoch, and split dataset variables with various scenarios. The highest accuracy result was obtained at 72% with a data split of 85%:15%, 32 batch size,s and 500 epochs.

Keywords: music genre classification; deep learning; GTZAN dataset

1. Introduction

The shift in the culture of playing music through digital platforms will require a machine that can automatically classify music according to the music genre without the need to play the song first. Thus, it will greatly reduce the time to search for music genres in a large music database. Since the audio platform provides millions of songs, having a system for indexing and labeling music genres is of high interest and enables the platform to recommend music genres according to the platform's users' profile. This system requires automation of music genre labeling on millions of songs from the database.

Music genres can be grouped based on the similarity of types or literature as well as the basis for their formation[1]. The work of classifying musical genres consists of 2 process stages, namely feature extraction and classification. In classifying music genres, several types of features are known, including magnitude-based features that can be identified from tone color, loudness level, pitch and sound, and musical harmony[2]. Sounds from one source can be distinguished and categorized by tone color[3]. Features included in this category are spectral decrease, spectral centroid, spectral flatness, spectral flux, spectral roll-off, spectral spread, spectral slope, and Mel Frequency Cepstral Coefficients (MFCC). Another category is the tempo-based features

which recognize music from the aspect of rhythm and tempo. This feature category is known as beats per minute, where beat histogram is used to visualize the audio signals, and the root mean square can measure signal intensity. In addition, there are also categories of chordal progression and pitch-based features [4].

Before the data is modeled, pre-processing needs to be done so that the raw data is easier to be adapted to the classification model used. Related research stated that the output of a feature used by giving a smaller value could represent the value of the entire audio signal. Giving a smaller value was intended to reduce the dimensions data [3]. In addition, pre-processing is carried out on raw audio data in the hope that the feature output will provide a meaningful representation understood by humans and machines [5].

Each music signal has its own characteristic acoustic features consisting of phase, frequency, amplitude, and time parameters. These features can be used to classify the type of music. The research conducted by Shah et al. [6] utilized many time domain and frequency domain elements to categorize music into several genres. To train Support Vector Machine (SVM), Random Forest, and Gradient, they extracted Spectral Centroid, Initial Intensity, Zero-Crossing Rate (ZCR), Tempo, Spectral Contrast, Spectral Bandwidth, Roll-off Contrast, and Flatness. The spectrogram was also retrieved and used

to train the Deep Convolution Neural Network (DCNN). Their classification results showed that Convolution Neural Network (CNN) outperformed machine learning algorithms.

Besides using CNN, research on music genre classification explores feature extraction by comparing the performance of feature extraction to obtain the best performance. Research comparing several feature extractions by dividing 30 seconds of audio into clips with a duration of three seconds and then classifying them using Visual Geometry Group-16 (VGG-16), CNN architecture, Short-Term Fourier Transform (STFT) feature extraction provided better results with an accuracy of 95,5% compared to Mel-Spectogram and MFCC which gave 90% and 86% accuracy results, respectively [7]. In another study, classification using CNN with a three-second music duration feature gave 72.4% better accuracy than a thirty-second music duration feature which was only 53.50%; the spectrogram feature showed increased accuracy but with an even greater number of epochs [3].

Another study using CNN with a duration of three seconds employing the Resnet architecture and Alexnet resulted in 54% and 42% of accuracy, respectively [8]. Meanwhile, in a study using audio feature extractions such as Root Mean Square Energy (RMSE), spectral centroid, spectral roll-off, spectral bandwidth, MFCC, Chroma STFT, zero crossing rate, perceptron, tempo, and harmony with a duration of thirty and ten seconds provided 58% and 81% of accuracy, respectively [9]. In an attempt to optimize the classification of music genres on neural network performance, maximum pooling was combined with average pooling to provide more neural network information. By passing one or several layers using shortcuts, these two methods were reported to increase the accuracy of music genre classification [10].

Meenakshi and Vishnupriya [11] used the GTZAN dataset to classify ten genres using feature extraction. The MFCC was the feature vector selected. By recording the general contours of the log-power spectrum on the Mel frequency scale, MFCC was able to encode the timbral characteristics of the musical signal. Mel spectrum, which had 128 coefficients, and MFCC, which had 13 coefficients, were the two types of feature vectors obtained. Feature vectors were collected and entered into the database. The database consisted of MFCC for genres defined using ten array sizes. The Mel Spectrum feature vector was 599x128x2, while the MFCC feature vector was 599x13x5. The data was then generalized by shuffling before being fed into the neural network. Two hundred song characteristics were used for testing, while the remaining 800 were used to train the model. The learning accuracy for the Mel Spec and MFCC feature vectors after training the CNN model was 76% and 47%, respectively. Another finding from the study suggested that MFCC took less time to assemble while Mel Spec took more time to train.

Exploration of neural network learning in the classification of music genres obtained using deep learning models gave more efficient results, achieving a 98% and 68% of training accuracy in 37 seconds and 36 seconds, respectively[12]. Research using the GTZAN dataset with a duration of three seconds has been conducted. Further, the research employed image data resulting from the spectrogram on the GTZAN dataset by comparing two CNN architectures. The results showed that VGG-16 architecture at 20 epochs performed better than the Resnet-50 architecture, with an accuracy of 60% [13].

2. Research Methods

2.1 Mel-Frequency Cepstral Coefficients

MFCC is known as an effective feature in speech recognition[14]; MFCC works similarly to how the ear perceives sound by distinguishing between sounds at high and low frequencies[15]. Human hearing is processed by the cochlea, which interprets the sound frequency. The pitch contained in the coefficient list is called MFCC[16]. The values on the pitch are arranged to imitate the workings of the cochlea. First, the sound frequency is calculated by the high and low. Then, the layer of the cochlea, which is in the form of small hairs, will vibrate according to the high and low voice. Nevertheless, the cochlea is difficult to pick up very low frequencies [14]. The feature extraction process using MFCC can be seen in the schematic Figure 1.
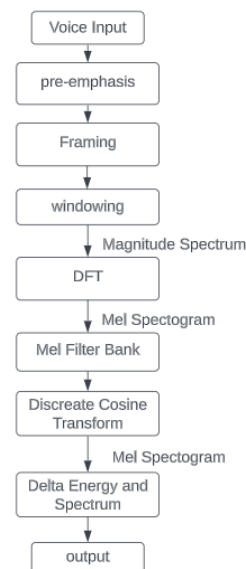

Figure 1. MFCC Feature Extraction Schematic

Initially, MFCC was used in audio or sound processing, along with the development of machine learning in the work of the Music Information Recognition (MIR) method. MFCC is known to be able to represent tones

effectively. The basic process in MFCC [17] is dividing the signal into frames, converting from Hertz to Mel Scale, taking the audio Mel representation algorithm, taking the logarithmic magnitude, and using the Discrete Cosine Transformation. Then, the result creates a spectrum above the Mel frequency as a reciprocal of time, thus producing MFCC as shown in Figure 2.
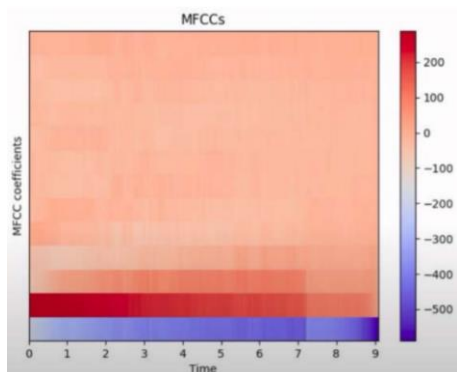

Figure 2. Visual Representation of MFCC

After the feature extraction process was carried out using MFCC, the audio data was forwarded to the CNN, as shown in Figure 3.
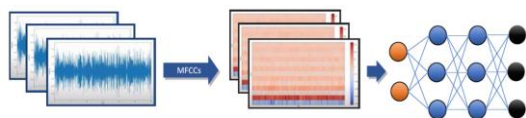

Figure 3. Pre-processing Pipeline

2.2 Dataset

Many research has been done on music genre classification to obtain the best classification model. The dataset used in music genre classification was GTZAN, which is the most frequently used dataset and consists of 1000 tracks with a duration of 30 seconds on each track. The number of music genres was ten genres, each containing 100 songs. The genres in this dataset were rock, reggae, pop, metal, jazz, hip-hop, disco, country, classic, and blues [18].

In addition to providing audio files, the GTZAN dataset provided 2 CSV files containing feature extraction results at a duration of thirty and three seconds containing 57 feature extraction values as shown in Table 1.

Table 1. Feature Extraction

Feature Extraction
Chroma SFTF
RMS
Spectral Centroid
Spectral Bandwidth
Spectral roll-off
Zero Crosing Rate
Harmony
Tempo
20 MFCC Coefficients [19]

This study utilized the GTZAN dataset to classify music genres using CNN, which previously extracted features using MFCC, and classified ten music genres using audio files that last for 30 seconds.

2.3 Convolutional Neural Network

A CNN consists of a convolutional layer, a pooling layer, and a fully connected layer[20]. The architecture of the three layers is depicted in Figure 4.
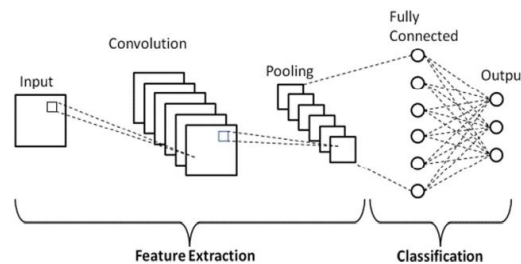

Figure 4. CNN Architecture

The function of the convolution layer is to study features, and the results obtained become the input to the next level with the activation function. The function of the pooling layer is to reduce the size of the input data by reducing the dimensions of each feature while maintaining the important features [21], resulting in a faster computation time. There are two known polling layers, namely the max-polling layer and the average-polling layer. In this study, the max-polling layer using the maximum value was performed to reduce the matrix data [22]. Figure 3 illustrates how the max-pooling works. The class classification process was assisted by the ability of the full connected layer. One way to reduce overfitting on a neural network is to use dropout layers[23]. Each training iteration on the dropout layer has the opportunity to kill neurons to obtain good results, and this can be adjusted with hyperparameters as needed. At the time of inference, all neurons will be activated, and the weight of each neuron will be multiplied by the appropriate probability to account for the dropout effect.

The dense layer is used to save computation time and maintain features with low complexity compared to other architectures that take longer processing time, such as GoogleNet and VGG [24]

Rectified Linear Unit (ReLU) is an activation function used in neural networks to introduce nonlinearity into models. This function is one of the most commonly used enable functions in deep learning. ReLU works by returning an input value for positive input and 0 for negative input. This function is simple and effective for solving vanishing gradient problems in deep neural networks. ReLU is computationally faster than other activation functions, such as Sigmoid and Tanh[22]. Adam's optimizer was chosen in the CNN model because it is more efficient in memory usage with fewer

computational resources needed. Adam's optimizer works by calculating adaptive learning rate (LR) on the parameters in the model[23].

2.4 Propose Method

Each audio clip can be represented as 30 seconds at 22050 samples per second, which is 661500 vector length. As noted in the preceding section, this would place a significant computational load on traditional machine learning techniques. According to the acoustic literature we reviewed, the MFCC feature is the most popular technique for encoding extended time domain waveforms, significantly reducing dimensions while still collecting the majority of the information. We first create smoothed framesets using a process that employs a 25 ms hamming window with a 10 ms overlap. The frame is then subjected to the Ftheier transform to obtain the frequency components. The mel scale, which models how humans can detect pitch differences generally below 1kHz and logarithmically above 1kHz, is then used to map those frequencies. This map divides the frequency into 20 squares by computing the coefficient of the triangle window based on the mel scale, multiplying the result by the frequency, and taking the log. Discrete cosine modifications are then used to embellish the frequency components. We ultimately only remember the first 13 of these 20 frequencies because higher frequencies are characteristics that have less of an effect on human perception and contain less information about the song. Then, each sample will have 2600 13 traits. For the trial setup, we further divided the MFCC feature into 4 pieces that were roughly equal in size, and we extracted the first 40 sections from each area. We frequently generate an MFCC feature length of 13 x 160 = 2080 to represent a 30 second audio recording for further studies.

The CNN model architecture used in this study consisted of five convolutional layers. The previously processed MFCC was used as the input, which was the extraction of 30 seconds of audio data and sent to the CNN model. The approach methods involved the application of a dense layer on each layer and 30% dropout to prevent overfitting. Optimizer used was Adam, while learning rate activation utilized Relu and Softmax. Specifically, the first dense layer had 1024 neurons with 40 input shapes while layers with values of 512, 256, and 64 had a 30% dropout. Relu was used as activation in each layer. Meanwhile, the output layer displayed the probability of ten genres through a fully connected layer, employing Softmax and Adam as the activation and optimizer, respectively. The genre that had the highest probability would become the label of the input given.
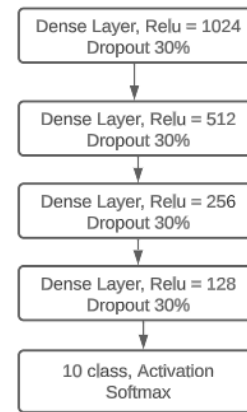

Figure 5. The CNN Architecture Used

2.5 Evaluation of Results

This study evaluated the results of the music genre classification accuracy by assessing the accuracy level of changes in the parameters of batch size, epoch, and data split. The scenarios taken to evaluate the training and validation results are shown in Table 2.

Table 2. Experiment Scenario

Parameter Type	Value
Split Data	70%:30%
	80%:20%
	85%:15%
	90%:10%
Epoch	300
	500
	800
Batch Size	16
	32
	64
Class	10

The accuracy value will generally increase according to the increase of epoch number but will stop at a certain point and remain constant or decrease [25]. In this study, three epoch numbers were used, i.e., 300, 500, and 800 epochs. Batch size is a measure of the amount of data that is trained in one iteration of the training process. The value of the batch size affects the stability and speed of the training time. Small batch size values require more training time compared to large batch size values and take longer to reach the convergence point. Although large batch size values will reach the convergence point faster, they have the potential to be overfitting with poor generalization[26]. In this study, the best batch size value was 32.

The scenario of dividing the training and testing data with several compositions is important because it affects the training results and the model formed. A large training value promises an accurate value because the amount of data being trained is large. However, this has the potential for overfitting [27]. With a large test value, a more reliable model

performance estimation can be obtained, yet the model has the potential to be inaccurate.

3. Results and Discussions

Based on the training results of the scenarios shown in Table 1, the batch sizes of 16 and 64 did not obtain good accuracy results (Tables 3 and 4). On the other hand, the batch size of 32 in each split data distribution scenario resulted in the best results (Table 5).

Table 3. Results with Batch Size of 16

Split Data	Batch Size	Epoch	Acc %
60%:40%	16	800	48
70%:30%	16	300	48
80%:20%	16	800	49
85%:15%	16	800	57
90%:10%	16	300	52

The 300th epoch in several scenarios with a batch size value of 16 had accuracy values that tended to remain the same or decrease. In another scenario, better accuracy results were obtained at the 800th epoch for split data of 85%:15% in accordance with the data in Table 2. The split data ratio of 70%:30% provided the best accuracy results for epoch 800 but not better than the accuracy obtained for epoch 500 with split datasets of 80%:20% and 85%:15%, as shown in Table 3.

Table 4. Results with Batch Size of 64

Split Data	Batch Size	Epoch	Acc %
60%:40%	64	500	51
70%:30%	64	800	52
80%:20%	64	800	52
85%:15%	64	500	51
90%:10%	64	800	50

From the scenario with a batch size value of 32, overall, the split dataset gave results above 50%. The split dataset 80%:20% had an accuracy of 62%, while the split dataset of 85%:15% had an accuracy value of 72%. Changing the split datasets further to 90%:10% did not provide better results. Instead, the accuracies decreased as shown in table 4.

Table 5. Results with Batch Size of 32

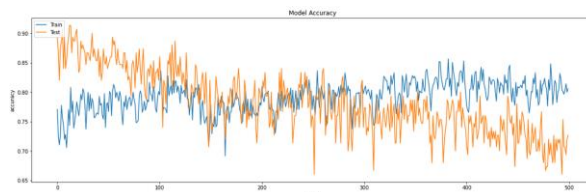| Split Data	Batch Size	Epoch	Acc %
60%:40%	32	300	51
70%:30%	32	800	55
80%:20%	32	500	62
85%:15%	32	500	72
90%:10%	32	300	55

Figure 6. Accuracy Model Graph

A graph of the accuracy model obtained from the training scenario of 85%:15% is illustrated in Figure 6, with an accuracy value of 72%.

Not much research has been done on the classification of music genres in the GTZAN dataset using 30 seconds of duration data compared to the use of 3 seconds of data. The time duration of the data affects the accuracy of results obtained, as seen in the comparison of several studies in the duration data comparison (Table 6).

Table 6. Accuracy Results on Data Usage Duration

Research	Time duration	Acc %
Ahmad[7]	3 seconds	95
Lau [19]	3 seconds	81.73
Zhang [10]	3 seconds	87,4
Vita [9]	30 seconds	58
Purnama [13]	30 seconds	60
Ndou [3]	30 seconds	53,5

The model formed with a duration of three seconds classifies the music genre in the first 3 seconds of the song. However, this model has a high error potential since the initial music varies greatly, particularly in the first 3 seconds. Thus, it cannot be used as a benchmark in determining the music genre.

Previous studies employed several feature extractions for the classification of musical genres, namely spectogram, Mel spectrogram, and RMSE [3] [9]. Accuracy results obtained using AlexNet, Resnet, and VGG16 did not give good accuracy results with the aforementioned feature extractions. Similarly, the CNN architecture with the spectogram extraction feature did not provide decent results, with an accuracy value of 53.5%. In contrast, our model with MFCC feature extraction resulted in a much better accuracy result, with an accuracy value of 72% as shown in Table 7.

Table 7. Comparison of Research Employing Time Duration of 30 Seconds

Research	Architecture	Extraction Feature	Acc %
Ndou [3]	CNN	Spectogram	53,5
Vita [9]	CNN	RMSE	58
Purnama [13]	Resnet	Spectogram	60
Purnama [13]	VGG-16	Spectogram	34
Current research	CNN	MFCC	72

4. Conclusion

Hyperparameters on variables can be used to optimize training results and the models formed. Batch size, epoch, and data division affect the accuracy obtained. In this study, several scenarios have been carried out with changes on the parameters of split data, batch size, and epoch used in the classification of music genres utilizing the CNN architecture with the addition of the MFCC feature extraction to the 30-second GTZAN dataset. Compared to other scenarios, the 500 epochs, a batch size of 32, and a dataset division of 85%:15%

parameter scenario gave the best results with an accuracy value of 72%.

References

[1] M. A. As Sarofi, I. Irhamah, and A. Mukarromah, "Identifikasi Genre Musik dengan Menggunakan Metode Random Forest," *Jurnal Sains dan Seni ITS*, vol. 9, no. 1, pp. 79–86, 2020, doi: 10.12962/j23373520.v9i1.51311.

[2] R. Ajoodha, R. Klein, and B. Rosman, "Single-labelled Music Genre Classification Using Content-Based Features."

[3] N. Ndou, R. Ajoodha, and A. Jadhav, "Music genre classification: A review of deep-learning and traditional machine-learning approaches," *2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings*, 2021, doi: 10.1109/IEMTRONICS52119.2021.9422487.

[4] M. Masekwameng and R. Ajoodha, "A Hybrid Method of Using Feedback from Users to Improve Music," 2022, [Online]. Available: https://ssrn.com/abstract=4332719

[5] T. Nkambule, "Classification of Music by Genre using Probabilistic Models and Deep Learning Models."

[6] M. Shah, N. Pujara, K. Mangaroliya, L. Gohil, T. Vyas, and S. Degadwala, "Music Genre Classification using Deep Learning," in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 2022, pp. 974–978. doi: 10.1109/ICCMC53470.2022.9753953.

[7] F. Ahmad and Sahil, "Music genre classification using spectral analysis techniques with hybrid convolution-recurrent neural network," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 1, pp. 149–154, 2019, doi: 10.35940/ijitee.A3956.119119.

[8] D. Xie, "AlexNet and ResNet for Music Genre Classification 1".

[9] Y. Vita Via, I. Yuniar Purbasari, and A. Putra Pratama, "Analisa Algoritma Convolution Neural Network (Cnn) Pada Klasifikasi Genre Musik Berdasar Durasi Waktu," *SCAN Jurnal Teknologi dan Informasi*, vol. 17, no. 1, pp. 35–41, 2022, [Online]. Available: http://ejournal.upnjatim.ac.id/index.php/scan/article/view/3251/2003

[10] W. Zhang, W. Lei, X. Xu, and X. Xing, "Improved music genre classification with convolutional neural networks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, pp. 3304–3308, 2016, doi: 10.21437/Interspeech.2016-1236.

[11] S. Vishnupriya and K. Meenakshi, "Automatic Music Genre Classification using Convolution Neural Network," in *2018 International Conference on Computer Communication and Informatics (ICCCI)*, 2018, pp. 1–4. doi: 10.1109/ICCCI.2018.8441340.

[12] W. Suo, "Efficient Music Genre Classification with Deep Convolutional Neural Networks," in *2022 5th International Conference on Data Science and Information Technology (DSIT)*, 2022, pp. 1–5. doi: 10.1109/DSIT55514.2022.9943952.

[13] N. Purnama, "Music Genre Recommendations Based on Spectrogram Analysis Using Convolutional Neural Network Algorithm with RESNET-50 and VGG-16 Architecture," *JISA(Jurnal Informatika dan Sains)*, vol. 5, no. 1, pp. 69–74, 2022, doi: 10.31326/jisa.v5i1.1270.

[14] F. M. Rammo and M. N. Al-Hamdani, "Detecting the Speaker Language Using CNN Deep Learning Algorithm," *Iraqi Journal for Computer Science and Mathematics*, vol. 3, no. 1, pp. 43–52, 2022, doi: 10.52866/ijcsm.2022.01.01.005.

[15] A. Akbar and A. Lawi, "Konferensi Nasional Ilmu Komputer (KONIK) 2021 Implementasi Algoritma Deep Artificial Neural Network Menggunakan Mel Frequency Cepstrum Coefficient Untuk Klasifikasi Audio Emosi Manusia", [Online]. Available: https://www.kaggle.com/ejlok1/toronto-

[16] K. Nugroho, Edy Winarno, Eri Zuliarso, and Sunardi, "Multi-Accent Speaker Detection Using Normalize Feature MFCC Neural Network Method," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 4, pp. 832–836, Aug. 2023, doi: 10.29207/resti.v7i4.4652.

[17] M. O. Faruk and S. Ghosh, "Developing Music Recommendation System by Integrating an MGC with Deep Learning Techniques," *Technology, Engineering & Mathematics (EPSTEM)*, vol. 19, 2022, [Online]. Available: www.isres.org

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

[19] D. S. Lau and R. Ajoodha, "Music Genre Classification: A Comparative Study Between Deep Learning and Traditional Machine Learning Approaches," *Lecture Notes in Networks and Systems*, vol. 217, no. 1433596, pp. 239–247, 2022, doi: 10.1007/978-981-16-2102-4_22.

[20] S. Aras and A. Setyanto, "Deep Learning Untuk Klasifikasi Motif Batik Papua Menggunakan EfficientNet dan Trasnfer Learning," vol. 8, no. 1, 2022.

[21] S. Aras, A. Setyanto, and Rismayani, "Classification of Papuan Batik Motifs Using Deep Learning and Data Augmentation," in *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, 2022, pp. 1–5. doi: 10.1109/ICORIS56080.2022.10031320.

[22] K. Eckle and J. Schmidt-Hieber, "A comparison of deep networks with ReLU activation function and linear spline-type methods," 2018.

[23] R. O. Ogundokun, R. Maskeliunas, S. Misra, and R. Damaševičius, "Improved CNN Based on Batch Normalization and Adam Optimizer," in *Computational Science and Its Applications – ICCSA 2022 Workshops*, O. Gervasi, B. Murgante, S. Misra, A. M. A. C. Rocha, and C. Garau, Eds., Cham: Springer International Publishing, 2022, pp. 593–604.

[24] A. Ameh Joseph, M. Abdullahi, S. B. Junaidu, H. Hassan Ibrahim, and H. Chiroma, "Improved multi-classification of breast cancer histopathological images using handcrafted features and deep neural network (dense layer)," *Intelligent Systems with Applications*, vol. 14, p. 200066, May 2022, doi: 10.1016/J.ISWA.2022.200066.

[25] Y. Martha, K. Damanik, K. Adi, and C. E. Widodo, "THE EFFECT OF EPOCH ON THE ACCURACY OF DETECTION OF LUNG CANCER," *International Journal of Innovative Research in Advanced Engineering*, vol. 7, no. 8, pp. 331–337, Aug. 2020, doi: 10.26562/ijirae.2020.v0708.004.

[26] R. Lin, "Analysis on the Selection of the Appropriate Batch Size in CNN Neural Network," in *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, 2022, pp. 106–109. doi: 10.1109/MLKE55170.2022.00026.

[27] M. Hu and Y.-H. F. Hu, "The Effects of Different Parameters on the Accuracy of Deep Learning Models for Predicting U.S. Citizen's Life Expectancy," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2021, pp. 105–109. doi: 10.1109/CSCI54926.2021.00016.