Pencak Silat Movement Classification Using Convolutional Neural Network (CNN)

Vira Nur Rahmawati
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
07111950050005@mhs.its.ac.id

Eko Mulyanto Yuniarno
Department of Electrical Engineering
Department of Computer Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ekomulyanto.ee.its.ac.id

Supeno Mardi Susiki Nugroho
Department of Electrical Engineering
Department of Computer Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
mardi@ee.its.ac.id

***Abstract*— Pencak silat, besides from being useful for self-protection, also has many other benefits, such as increasing physical strength, maintaining posture, and maintaining cardiac health. Due to various reasons, such as busyness and the recent pandemic, practicing pencak silat is difficult to do in groups. Even when there is study material on pencak silat at school, it is difficult for the sports teacher to teach the movements directly. Pencak silat exercises that are practiced alone without a coach can cause injury if the movements are not correct. Therefore, this study builds a system to recognize pencak silat movements. The system was built using the bodypose-based CNN method. MediaPipe is used to extract the body pose because MediaPipe provides the best estimation accuracy. Then the bodypose is used for input to CNN to recognize the movement. This system uses CNN because because it has the ability to automatically extract the patterns and representations from input with more accurately result. The accuracy that be obtained reaches 77% when tested on data that has never been used. This model can be used as a starting point for creating a pencak silat classification system that has a better accuracy and more recognizable moves.**

Keywords—body pose, CNN, convolution, mediapipe, pencak silat,

I. INTRODUCTION

Pencak silat is a martial art that originates from Indonesia. pencak silat was created from the way Indonesian ancestors protected themselves and defended their lives from natural challenges. Pencak silat has many benefits, including increasing physical strength, maintaining ideal body weight, maintaining posture, and improving heart health. In addition, pencak silat is also useful for growing children's character, by growing courage, discipline, self-confidence, sportsmanship, and social skills.To get these various benefits, pencak silat training must be routinely carried out three to four times a week, either at school, at a pencak silat course, or at home. However, for various reasons, especially busyness, the four time a week practice is difficult to do, and practicing alone at home is not effective because there is no guide who can judge whether the movements are correct or not. In addition, if the movement is done incorrectly, it can cause injury. For this reason, it is necessary to have a system for detecting and classifying pencak silat movements.

Many researchers have carried out and explored human movement recognition systems. Human pose estimation is used to recognize standing, squatting, body exercise, and falling movements. It used mainly for people who live alone, so they can get help quickly in case of an accident [1]. Experiments on more movements have also been carried out, for 20 general activities. Pose estimation is performed using CNN regression and produces 14 joint keypoints. Classification is carried out using the CNN that implemented using caffe based on AlexNet. This method can achieve accuracy of 80.51% [2]. Gupta et al. perform recognition of several human activities, including sitting, standing, running, dancing, and laying. That study used open poses with 18 body keypoints to extract body poses. For classification, they compared several algorithms that are Multiple Logistic Regression, KNN, SVM, Decision Tree, and Random Forest. As a result, the Multiple Logistic Regression, SVM, and Random Forest algorithms give good results, which are above 80% [3].

As it develops, human movement recognition is also used in various fields, such as the arts and sports. In the arts, it can be used to create generative models for dance [4]. In the field of sport it is most widely used in yoga because it has many unique movements [5][6]. Research to recognize yoga postures was carried out by Kishore et al., the postures used were Ardha Chandrasana, Tadasana, Trikonasana, Veerabhadrasana, and Vrukshasana. In his research, he compared several deep learning architectures used for pose estimation. The architectures being compared are EpipolarPose, OpenPose, PoseNet, and MediaPipe. Mediapipe provides the best estimation accuracy, above 80% [7]. This makes MediaPipe widely used for pose estimation. The recognition of yoga movement that based on MediaPipe was carried out by Tanugraha et al. Classification was performed using LSTM with Adam Optimizer. The movements performed are T-Pose, Warrior Pose, and Tree pose in yoga. The research was carried out by recording movements with a camera, then continued with the detection of 33 keypoints from the human body posture. After that make movement classification with LSTM in real time and achieve accuracy of 91% [8].

Research on the recognition of pencak silat movements has been developed. This research uses a Neural Network Teachable machine with the RNN algorithm, then integrated with the mobile application program using Flutter. The movements used are perfect stance, horse anchor, frog anchor, stance 1 and stance 2. The accuracy obtained reaches 70% when the object is in a stationary position, but when the object moves, the accuracy is only

How to cite: Rahmawati, V.N., Yuniarno, E.M., Nugroho, S.M.S. (2023). Pencak Silat Movement Classification Using Convolutional Neural Network (CNN). *JAREE (Journal on Advanced Research in Electrical Engineering), 7(2).*

up to 10% [9]. Therefore, to get better accuracy, this journal proposes a new approach to classification pencak silat movement using video as a data set. All parts of the video will be included in the classification process, not just a slice of an image. This study uses a Convolutional Neural Network (CNN) as a model architecture with MediaPipe as body pose estimation framework.

II. Methods

The proposed framework for classifying pencak silat movement is illustrated on figure 1. The method consisted of 4 stages that are dataset, pose extraction, classification, and result.

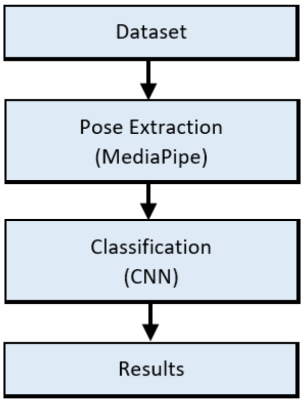


Fig. 1. Proposed framework

2.1. Dataset Preparation

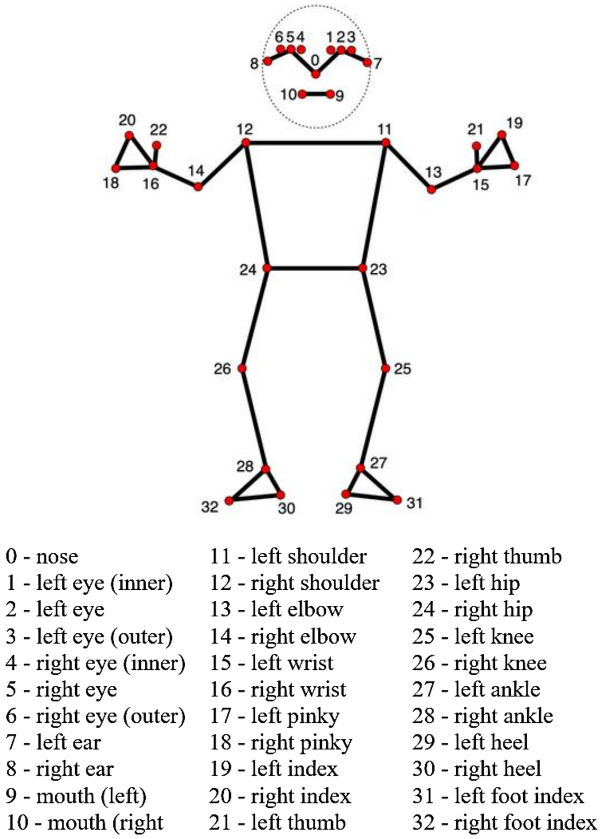
The dataset that used for this study is in the form of motion videos. That videos were taken directly and has never been used for any research. The participant for the dataset for this study is the children of extracurricular pencak silat members in Cemandi Elementary School. The the process of dataset collection is supervised by a coach. This is done with the intention to make sure that the movements are correct. The video was taken from the front of participant. The camera is considered as an opponent and the kick is directed at the camera. The distance between participant and camera is 2 meters and the camera height is 85 cm. The video frame shows the full body of the participant, include head, body, hand, until foot. The duration of the video is not same, depending on the movement. One video starts with both feet on the ground, then kicking, until it returns to its starting position, where both feet are on the ground. It takes about 1 to 2 seconds.

There are three movements that will be classified in this study, that are straight kicks, circular kicks, and T kicks. The total number of datasets that are used in this study were 330 samples of the three classes. Each class has 110 samples. That 330 data will be used for training process. Then this study used more data for the testing process. Total data that used for training process was 100 samples video of the three classes. The data testing was new data that never be used for training process. Sample movement is shown in figure 3. The image was the capture image from video at specific time, from t=1 until t=n.

2.2. Pose Extraction

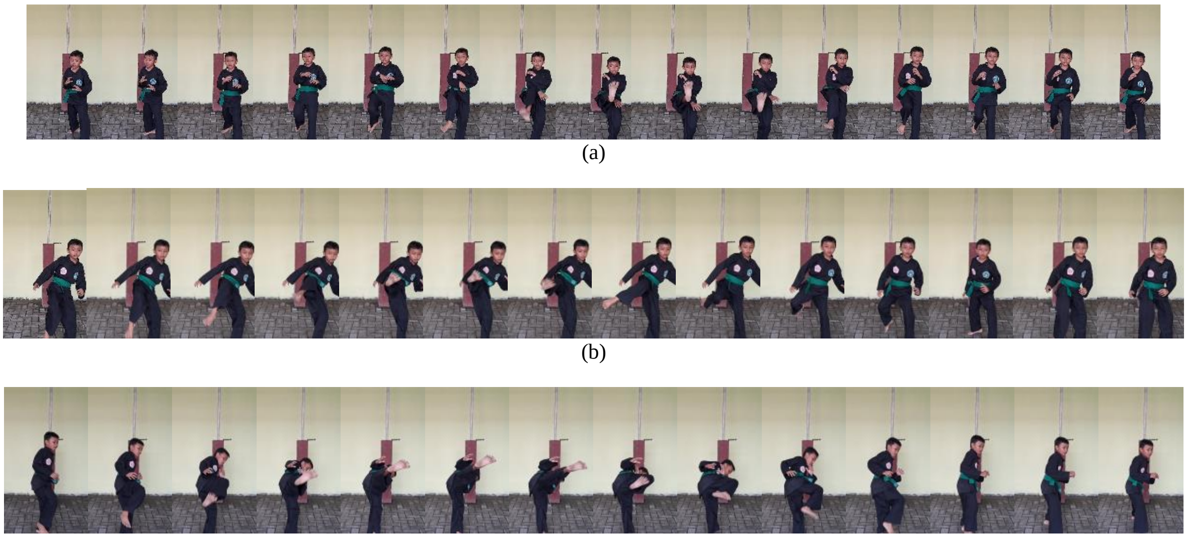
The datasets will be processed using pose extraction. Pose extraction obtains the pose of an articulated human body, which consists of joints and rigid parts using image-based observations [10]. The location of the joints of the human body were estimated, then given a marker called a keypoint. Keypoints will be connected with lines in such a way that a human body pose is formed. Then, the original image will be removed and only the body pose will remain. The location of the keypoint will change according to the movement, so it can be learned by CNN for the classification process.

This study use MediaPipe for the pose extraction process because MediaPipe provides the best estimation accuracy [7]. MediaPipe is a framework for building pipelines to perform inference over arbitrary sensory data. Then the pipeline can be built as a graph of modular components, including model inference, media processing algorithms and data transformations, etc [11]. For the pose extraction, the sensory data that used is video and then extracted 33 landmark/keypoint on the human body as shown if figure 2. Among the 33 MediaPipe keypoints, this study used all the keypoints, because when doing the movement, the position of all the keypoints were moved or changed.



0 - nose
1 - left eye (inner)
2 - left eye
3 - left eye (outer)
4 - right eye (inner)
5 - right eye
6 - right eye (outer)
7 - left ear
8 - right ear
9 - mouth (left)
10 - mouth (right
11 - left shoulder
12 - right shoulder
13 - left elbow
14 - right elbow
15 - left wrist
16 - right wrist
17 - left pinky
18 - right pinky
19 - left index
20 - right index
21 - left thumb
22 - right thumb
23 - left hip
24 - right hip
25 - left knee
26 - right knee
27 - left ankle
28 - right ankle
29 - left heel
30 - right heel
31 - left foot index
32 - right foot index

Fig. 2. Definition of Landmarks in MediaPipe Pose [12].



(a)

(b)

(c)

Fig. 3. Sample movement from t=1 to t=n, for: (a) Straight Kick, (b) Circular Kick, (c) T Kick

2.3. Classification

The next step is classification. This study using CNN model for doing classification because it has the ability to automatically extract the patterns and representations from input image with more accurately result [13-16]. CNN is an architectural development of MLP (Multi-Layer Perceptron). It has a deep network consisting of several layers. A commonly used type of CNN, which is similar to the multi-layer perceptron (MLP), consists of numerous convolution layers preceding sub-sampling (pooling) layers, while the ending layers are FC layers [17].

CNN is usually used to classify images. Video classification is done by extracting videos into images. To classify a set of image sequences, a method is needed to combine all the classification results. If the classification is done for each frame separately, then the model cannot see the entire video, and causes the prediction results to change quickly and fluctuate.

In the research conducted by Aruna regarding Human Activity Recognition using Single Frame CNN, a moving average was used to obtain the final classification results, and to obtain an accuracy of 95%.[26] The moving average is the average of all prediction results in each frame. This method takes frames from the video and makes predictions according to the activities in the training model. Then these predictions are averaged and the final output results are labeled according to the average predicted results.

The final predicted probability is calculated using moving average by equation (1). [26] Final Predicted probability as P_f, number of frame to be averaged as n, probability predicted for the current frame as P, probability predicted for the last frame as P_{-1}, probability predicted for the penultimate frame as P_{-2}, and probability predicted for the (n-1) frame as P_{-n+1}.

$$P_f = \frac{\sum_{i=-n+1}^{0} P_i}{n} \qquad (1)$$

The CNN model that used in this study is consist of three convolution layers and two fully connected layers. The dataset image that sized 128×128 with three layers is processed using convolution layer with 32 filters, 3×3 sized kernel, and ReLU activation function. Then processed with max pooling layer with a pool size 2×2. After that, it is processed using convolution layer again with 32 filters, 3×3 sized kernel, and ReLU activation function, then max pooling layer with a pool size 2×2. Before enter in to fully connected layer, it is processed using convolution layer with 32 filters, 3×3 sized kernel, and ReLU activation function again.

The output from the convolutional layer then being flatten to make into one dimension. After being one dimension, it goes into the fully connected layer with theReLU activation function twice, then arrives at the output layer with the softmax activation function. The last output is depended on the number of classes in the dataset, in this study are three that are straight kick, circular kick, and t kick. The model CNN is shown in figure 4.

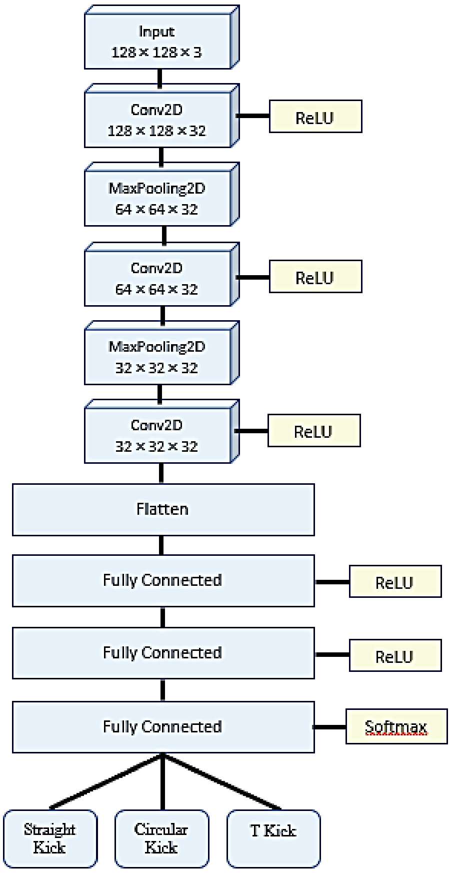


Fig. 4. Model CNN

III. Results and Discussion

The proposed framework in this study consisted of 4 stages, that are dataset, pose extraction, classification, and result. After dataset is collected, then the pose extraction process is carried out using MediaPipe. The results of the pose extraction are shown in Figure 6. The figure showed the body pose result of each frame that extract from video input. It shows the body pose from the beginning of movement (t=1) until the end of the movement (t=n) for three classes.

Fig. 5 shows the result of body pose at the beginning of straight kick and circular kick movements, all the 33 keypoints in the bodypose can be detected well. But at the T kick movement, some keypoints cannot be detected. The respondent kicked with the left foot, but the keypoint of the left foot could not be detected, from the knee to the sole of the foot. In addition, the keypoint of elbow of the left arm is also not detected.

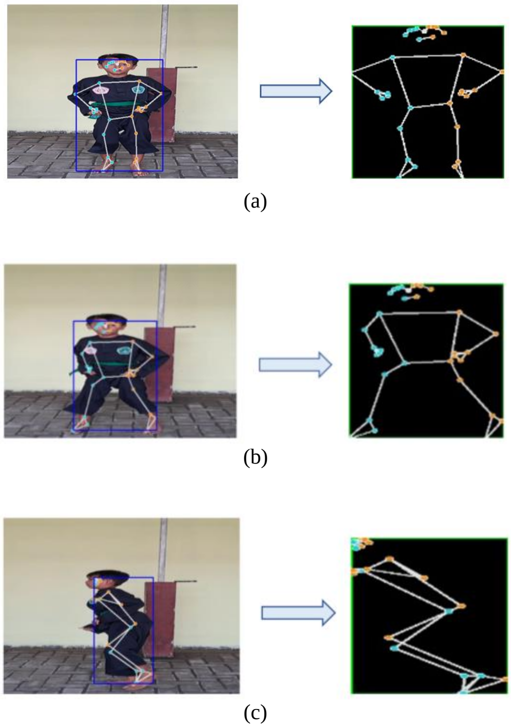


Fig. 5. The result of the pose extraction at the beginning of movement (a) Straight Kick and (b) Circular Kick, the keypoint in the body pose can be detected well. (c) T Kick, the keypoint in the body pose can not be detected well.

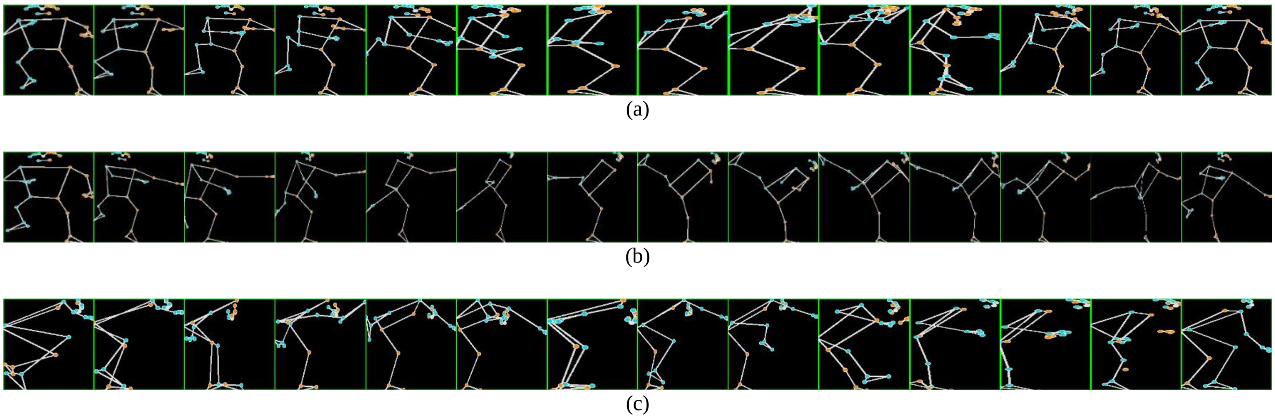
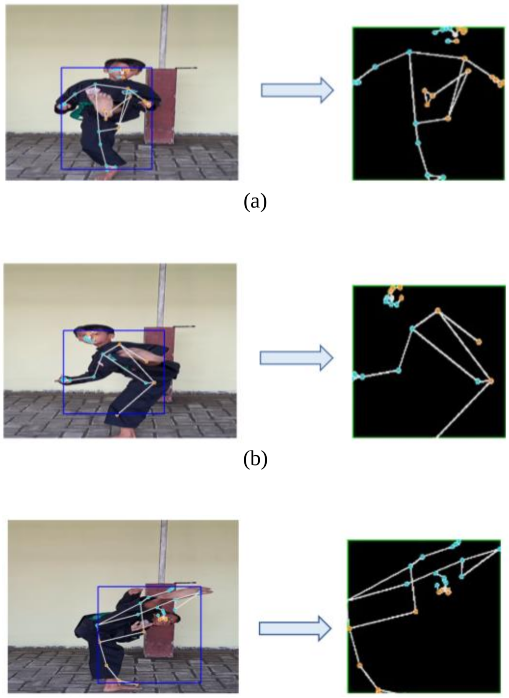


Fig. 6. The results of the pose extraction (a) Straight Kick, (b) Circular Kick, (c) T Kick

This happened because, when doing a straight kick and circular kick, the object faces the camera and the direction of the kick is forward, so the keypoints can be spread evenly on the x and y axes. But when doing the T kick, the object faces sideways and the direction of the kick is to the side. When facing sideways, the position of the right body keypoint and left body keypoints will be on the same x-axis, so that the keypoints are overlapping and difficult to detect.

At the peak of the movement or when the foot is in a position close to the opponent (in this study is the camera), the keypoint in bodypose is also not detected properly, several keypoints are overlapping, and there are keypoints that cannot be detected. At the result of pose extraction in fig 7, for the straight kick (fig 7a), the keypoint on the body pose can be detected well, as well as for the T kick (fig 7c), the keypoint on the feet can be detected well even though the keypoint on the hands cannot be detected. But for the circular kick (fig 7b), the keypoint in body pose cannot be detected properly. The keypoint of both legs cannot be detected, only one knee keypoint can be detected, that was the right knee, but it is incorrectly interpreted as the left knee.

This happens because when the foot is in an attacking position and close to the opponent, the position of other body parts will move according to the body's center of gravity so that it can stay balanced, causing the foot's keypoint position to be on the same x-axis as other body parts and making the keypoints are overlapping.


(a)

(b)

(c)

Fig. 7. The result of the pose extraction at the peak of movement (a) Straight Kick, the keypoint in the body pose can be detected well. (b) Circular Kick, the keypoint in the body pose can not be detected well. , (c) T Kick, the keypoint in the body pose can be detected well.

The pose extraction can have better performance by adding variations angle while taking the video. Especially angles that are not 0° at the x-axis, y-axis, and z-axis, such as 30°, 45°, or 60°. When the angel is not at the 0°, the keypoints can be evenly distributed throughout the x axis and the body pose can be detected well.

Moreover, the number of the keypoint that used could be reduces. There are several keypoints that may not be used, such as the keypoints number 1 – 10 which is located on the face and keypoints number 19 – 22, which is located on fingers. If there are fewer keypoints, then the possibility of overlapping keypoints will also be smaller, and the body pose can be detected well.

After the body pose was formed, the data is ready to be used for classification. The architecture that used for classification is CNN model which has been mentioned in method section. For the classification, the proposed model must be trained using data training. Data training that used in this study was

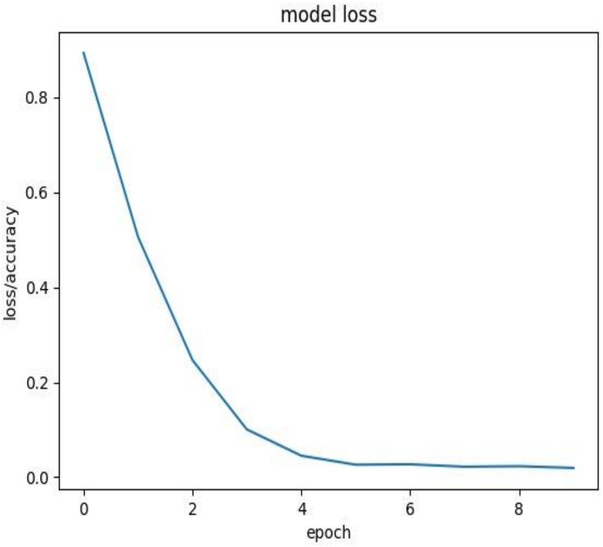


Fig. 8. Training process

Figure 8 shows the loss training process of CNN model. The CNN model achieved good performance. From the first iteration, it can obtain high accuracy, and doesn’t need too much iteration to reach the best accuration. In the first iteration, the CNN model obtained an accuracy of 0.60, then the second iteration is 0.83, and continues to increase until the 9th iteration. At the 9th iteration, the CNN obtained the best accuracy 0.99, but decreased to 0.98 in the 10th iteration. Therefore, the training process was stopped at the 10th iteration to avoid overfitting.

After the training process, the trained CNN model is tested using data testing. Total data testing that used for testing process was 100 video. Data testing was never used in the training process. The result of testing process is our CNN model can classify 77 data correctly. That’s make our CNN model achieved total accuracy 77%. The detail result is shown at confusion matrix in fig 9..

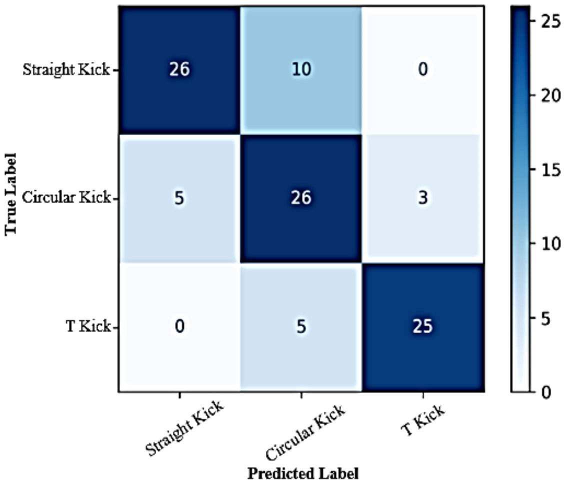


Fig. 9. Confusion matrix

For Straight Kick, the data tested was 36, 26 were correct and 10 were incorrectly labeled as Circular Kick. Also for the T Kick, the T Kick data tested was 30, 25 were correct and 5 were labeled incorrectly as Circular Kick. For the Circular Kick itself, the data tested for Circular Kick was 34, 26 were correct, 5 were incorrectly labeled as Straight Kick and 3 were incorrectly labeled as T Kick. From these results, it can be seen that all errors must be related to circular kicks, whether circular kicks are labeled as other movements or other movements are labeled as circular kicks.

All error result were related to Circular Kick because the body pose of Circular Kick could not be extracted well. At the result of pose extraction, could be seen that the keypoint of the right foot, which is the foot that used to attack, could not be detected. So the generated body pose do not represent the circular kick movement. That make the CNN model cannot extract the right feature of the movement and cannot learn well.

IV. Conclusion

This study proposed model for pencak silat movement classification using Convolutional Neural Network (CNN). The movements that are classified include straight kick, circular kick, and T kick movement. This study use dataset that is in video form. The video will be processed using MediaPipe as the pose extraction framework to take the body pose. This study use MediaPipe for the pose extraction process because MediaPipe provides the best estimation accuracy. The body pose then used for classifying using CNN model. The CNN model was used because because it has the ability to automatically extract the patterns and representations from input with more accurately result. This CNN model can achive total accuracy 77%. This accuracy is already good, but still can be improved by by revise the pose extraction. The pose extraxtion can have better performance by adding variations angle of video capture so that the keypoints can be evenly distributed throughout the x axis and the body pose can be detected well. Moreover, we can reduce the number of the keypoint that is used. There are several keypoints that may not be used, such as the keypoints on the face and fingers. If there are fewer keypoints, then the possibility of overlapping keypoints will also be smaller.

References

[1] Kim, J.-W.; Choi, J.-Y.; Ha, E.-J.; Choi, J.-H. Human Pose Estimation Using MediaPipe Pose and Optimization Method Based on a Humanoid Model. Appl. Sci. **2023**, 13, 2700.

[2] Bearman, A., Dong, C., “Human Pose Estimation and Activity Classification Using Convolutional Neural Networks” http://cs231n.stanford.edu/reports/2015/pdfs/cdong-paper.pdf

[3] A. Gupta, K. Gupta, K. Gupta and K. Gupta, "Human Activity Recognition Using Pose Estimation and Machine Learning Algorithm," 2021 International Semantic Intelligence Conference (ISIC), New Delhi, India

[4] Zaman, Lukman, Sampeno, dan Hariadi. (2019). Analisis Kinerja LSTM dan GRU sebagai Model Generatif untuk Tari Remo. JNTETI, Vol. 8, No. 2, Mei 2019

[5] Asshidiqy, R.A., Setiawan, Sasongko. (2022). Penerapan Metode Posenet untuk Deteksi Ketepatan Pose Yoga. JoYSC Vol. 4 No. 1, ISSN 2714-7150 E-ISSN 2714-8912

[6] Upadhyay, A.; Basha, N.K.; Ananthakrishnan, B. Deep Learning-Based Yoga Posture Recognition Using the Y_PN-MSSD Model for Yoga Practitioners. Healthcare **2023**, 11, 609.

[7] Kishore DM, Bindu S, Manjunath NK. Estimation of yoga poses using machine learning techniques. Int J Yoga 2022;15:137-43.

[8] Tanugraha, F.D., Pratikno, Musayyanah, dan Kusumawati. (2022). Pengenalan Gerakan Olahraga Berbasis (Long Short-Term Memory) menggunakan Mediapipe. ***JAIIT*** *(Journal of Advances in Information and Industrial Technology) Vol. 4, No. 1 Mei 2022, ISSN 2723-4371, E-ISSN 2723-5912*

[9] Taruna, I.P.J., Fredlina, dan Sudiatmika. (2022). Pengenalan Gerakan Sikap Dasar Pencak Silat Bakti Negara Berbasis Aplikasi *Mobile* menggunakan *Neural Network*. ISSN:2477-0043 ISSN ONLINE:2460-7908.

[10] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang and C. Yang, "The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation," in IEEE Access, vol. 8, pp. 133330-133348, 2020, doi: 10.1109/ACCESS.2020.3010248.

[11] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, Matthias Grundmann: MediaPipe: A Framework for Building Perception Pipelines. CoRR abs/1906.08172 (2019)

[12] Mediapipe Pose. Available online: https://developers.google.com/mediapipe/solutions/vision/pose_landmarker/ (accessed on 21 June 2023).

[13] D. Dai, "An Introduction of CNN: Models and Training on Neural Network Models," 2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR), Shanghai, China, 2021, pp. 135-138, doi: 10.1109/ICBAR55169.2021.00037.

[14] S. Tripathi and R. Kumar, "Image Classification using small Convolutional Neural Network," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 483-487, doi: 10.1109/CONFLUENCE.2019.8776982.

[15] A. Singh, S. Agarwal, P. Nagrath, A. Saxena and N. Thakur, "Human Pose Estimation Using Convolutional Neural Networks," 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 2019, pp. 946-952, doi: 10.1109/AICAI.2019.8701267.

[16] A. S. Dileep, N. S. S., S. S., F. K. and S. S., "Suspicious Human Activity Recognition using 2D Pose Estimation and Convolutional Neural Network," 2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India, 2022, pp. 19-23, doi: 10.1109/WiSPNET54241.2022.9767152.

[17] Alzubaidi, L., Zhang, J., Humaidi, A.J. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[18] Aruna, V. & Deepthi, Aruna & Leelavathi, R.. (2022). Human Activity Recognition Using Single Frame CNN. 10.1007/978-981-19-4831-2_17.