IJRCS

International Journal of Robotics and Control Systems
Vol. 5, No. 3, 2025, pp. 1900-1918
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

ASCEE

Person and Activity Recognition Based on Joint Motion Features Using Deep Learning with Drone Camera

Riky Tri Yunardi [a,b,1,*], Tri Arief Sardjono [b,c,2], Ronny Mardiyanto [b,3]

[a] Instrumentation and Control Engineering Technology, Department of Engineering, Faculty of Vocational, Universitas Airlangga, Surabaya 60115, Indonesia
[b] Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia
[c] Department of Biomedical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia
[1] rikytriyunardi@vokasi.unair.ac.id; [2] sardjono@bme.its.ac.id; [3] ronny@elect-eng.its.ac.id
* Corresponding Author

ARTICLE INFO

Article history
Received May 22, 2025
Revised August 16, 2025
Accepted September 15, 2025

Keywords
Person Recognition;
Activity Recognition;
Joint Motion Features;
Deep Learning Architectures;
Drone

ABSTRACT

The increasing demand for drone-based surveillance systems has raised significant concerns about advancements in person and activity recognition based on joint motion features within visual monitoring frameworks. This study contributes to developing deep learning models that improve surveillance systems by using RGB video data recorded by drone cameras. In this study, a framework for person and activity recognition based on 120 datasets is proposed, from drone camera-recorded videos of 10 subjects, each performing six movements: walking, running, jogging, boxing, waving, and clapping. Joint motion features, including joint positions and joint angles, were extracted and processed as one-dimensional series data. The 1D-CNN, LeNet, AlexNet, and AlexNet-LSTM architectures were developed and evaluated for classification tasks. Evaluation results show that AlexNet-LSTM outperformed the other models in person recognition, achieving a classification accuracy of 0.8544, a precision of 0.9161, a recall of 0.8575, and an F1-score of 0.8332, while AlexNet delivered superior performance in activity recognition with an accuracy of 0.8571, a precision of 0.8442, a recall of 0.8599, and an F1-score of 0.8463. The relatively small dataset size used likely favors simpler architectures like AlexNet. These findings highlight the effectiveness of joint motion features for person identification and emphasize the suitability of simpler classifier architectures for activity classification when working with small datasets.

1. Introduction

The implementation of visual control technologies within modern society has brought about substantial transformations in everyday human activities. One important use of this technology is spotting suspicious human actions using video surveillance [1]-[4]. The growing demand for effective surveillance and public safety has driven the development of intelligent monitoring systems, typically based on fixed-cameras [5]-[8]. However, traditional CCTV surveillance, which relies on continuous human monitoring, has become increasingly impractical and inefficient in large-scale or dynamic

http://dx.doi.org/10.31763/ijrcs.v5i3.1949

environments. To resolve this problem, an RGB drone camera has been introduced, but the low-resolution image in person and activity recognition are challenges in this study.

Recent advancements have explored the use of aerial video systems for human activity recognition. Some studies have employed alternative platforms such as helium balloons to simulate drone-like camera perspectives [9]-[11], while others, for example, Mishra et al. [12] introduced drone-recorded datasets for search and rescue applications that require accurate finding and tracking of humans, which is a very difficult and demanding task. Srivastava et al. [13] further expanded this area by proposing a dataset of eight human activities captured by drones flying at altitudes ranging from two to ten meters, which makes the captured body structures appear small. Despite these advancements, several problems persist in drone-based video capture, including sensitivity to lighting variations, image resolution limitations, fluctuating object distances, and drone-induced vibration artifacts. As a result, robust and effective feature extraction methods are crucial for accurate motion detection from aerial footage.

Human motion recognition remains a complex research area, particularly due to the intricate coordination of body joints that govern movement. Visual recognition systems commonly utilize variations in joint coordination across sequences of frames to identify human activities [14]-[18]. Grimmer et al. [19] emphasized the usefulness of lower limb joints in recognizing distinct gait patterns, especially during stair-climbing activities, by examining the kinematic movements of the hip, knee, and ankle. These joint dynamics provide important information for biomechanical evaluations for enhancing recognition performance. However, their study focuses on the lower body joints. In this study, sequences of key body joint positions (pose keypoints) were extracted from RGB images, instead of classical biomechanical features, to provide data on the sequence of body poses during movement.

Ahad et al. [20] proposed the extraction of kinematic posture features from skeletal joint positions to improve activity recognition accuracy. This approach involves measuring linear and angular joint features based on bone segment connections visible in video data, making it suitable for implementing pose estimation algorithms. For instance, digital cameras are used to capture front and side views of the pelvis, hip, knee, and ankle angles [21], [22]. Subsequent pose estimation performed using tools such as OpenPose [23], [24] and MediaPipe [25]-[28], enables the extraction of temporal joint movement sequences, which can be used to construct discriminative motion patterns for both activity and identity recognition.

Temporal feature sequences play a pivotal role in deep learning-based human motion analysis, as their effectiveness depends on the specific task and model architecture [29]-[32]. Nam et al. [33] employed convolutional neural networks (CNNs) trained on temporal joint motion features derived from coordinate sequences and depth maps. Similarly, Wang et al. [34] introduced skeletal edge movement features, consisting of rotation angles and movement distances between joints, which were then classified using neural networks. These approaches highlight the importance of both angular and spatial motion cues in achieving accurate classification results.

To overcome the shortcomings of earlier methods, recent research has employed deep learning architectures such as 1D-CNNs, LeNet, and AlexNet to improve the accuracy of person and activity recognition based on joint motion features [35]-[38]. The 1D-CNN architecture, a variation of the traditional CNN, is particularly popular due to its low complexity and high computational efficiency, which are essential for processing limited datasets such as joint motion data [39]. Furthermore, the Long Short-Term Memory (LSTM) architecture has been identified as an effective model for learning long-term dependencies in sequential data, particularly suited for time series classification tasks, especially when combined with CNNs for spatiotemporal feature extraction [40]-[42]. LSTMs are used to capture temporal data about joint motion features, which are crucial for distinguishing activities such as walking, running, or waving.

The objective of this study is to propose a deep learning model designed for individual and activity recognition from video data captured by a drone-mounted camera. The proposed novel model combines the strengths of deep CNNs and LSTM networks to effectively classify both person identity and activity type. The dataset used in this research consists of video recordings in which a single subject performs a specific activity, with each frame providing motion data derived from joint position coordinates and joint angles. The raw video data faces potential degradation because of its low resolution. Consequently, it is processed using pose estimation to emphasize the detection of structural motion patterns and to acquire features based on distance and joint angles for input into the classifier.

These features, extracted as sequential data representing body movement patterns, serve as input for the classification model. The study contributes by developing and evaluating hybrid AlexNet-LSTM architecture and an evaluation of its classification performance in comparison to other leading deep learning models, including 1D-CNN, LeNet, and standard AlexNet. The effectiveness of the proposed method is measured based on accuracy, precision, recall, and F1-score, with the goal of improving recognition reliability in low-resolution, drone-based video settings.

2. Method

The proposed method for person and activity recognition based on low-resolution RGB video data captured from a drone-mounted camera consists of five main stages: (1) video data acquisition from the drone camera, (2) joint detection and feature extraction, (3) feature selection for person recognition and activity recognition, (4) classification using deep learning models, and (5) evaluation of the classification results to assess model performance. The overall framework of the proposed person and activity recognition method is depicted in Fig. 1.

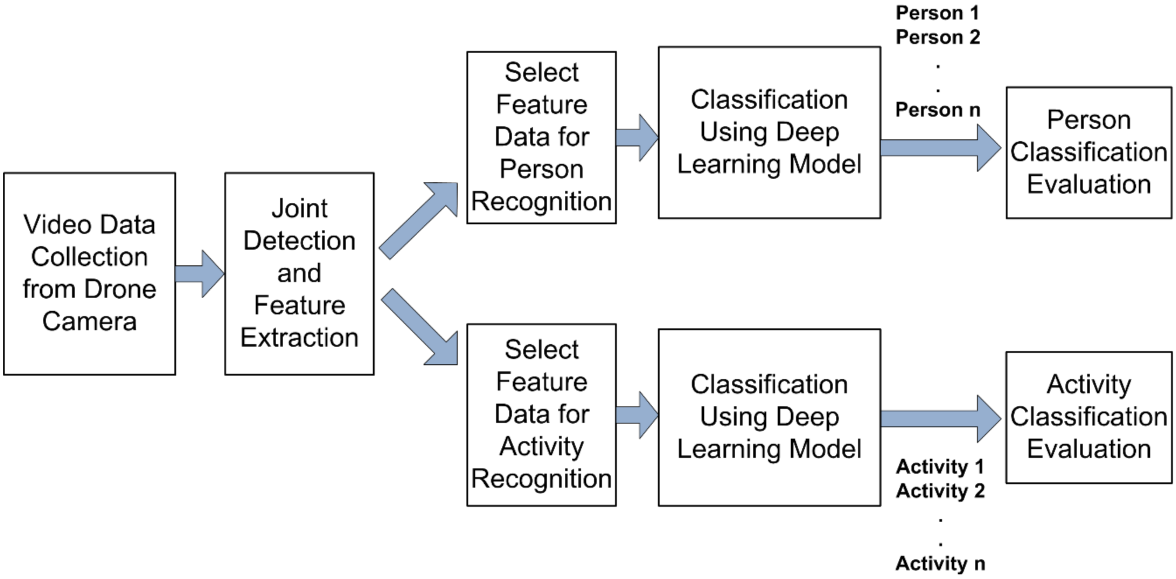

Video Data Collection from Drone Camera
Joint Detection and Feature Extraction
Select Feature Data for Person Recognition
Classification Using Deep Learning Model
Person 1
Person 2
.
.
Person n
Person Classification Evaluation
Select Feature Data for Activity Recognition
Classification Using Deep Learning Model
Activity 1
Activity 2
.
.
Activity n
Activity Classification Evaluation

Fig. 1. Overview of the method developed for person and activity recognition

2.1. Video Data Collection Using RGB Camera on a Drone

Low-cost drone navigation in unstructured environments using mounted cameras as the primary sensor has been used previously [43], [44]. The video data collection for human movement in this study was conducted using a drone-mounted camera. The dataset utilized in this study is following GitHub repository in https://github.com/yunardi-89/Drone-ITS. The recorded video sequences, each lasting between two to three seconds, captured various movements performed by 10 subjects (8 males and 2 females) aged between 18 and 21 years, with varying body heights, as depicted in Fig. 2. A low-cost drone (Eachine E88 Pro) was positioned outdoors at a height of two meters above the ground. To maintain a stable altitude and consistent camera orientation, considering the limitations of both the equipment and the environment, the drone was mounted on a two-meter-high pole. In collecting video data for the proposed model application, limitations regarding lighting conditions, occlusion, or variations in drone height were not taken into account. The videos were recorded in 1920 × 1080-pixel resolution at 20 frames per second (fps), transmitted to a smartphone via Wi-Fi FPV, and saved in MP4 format.

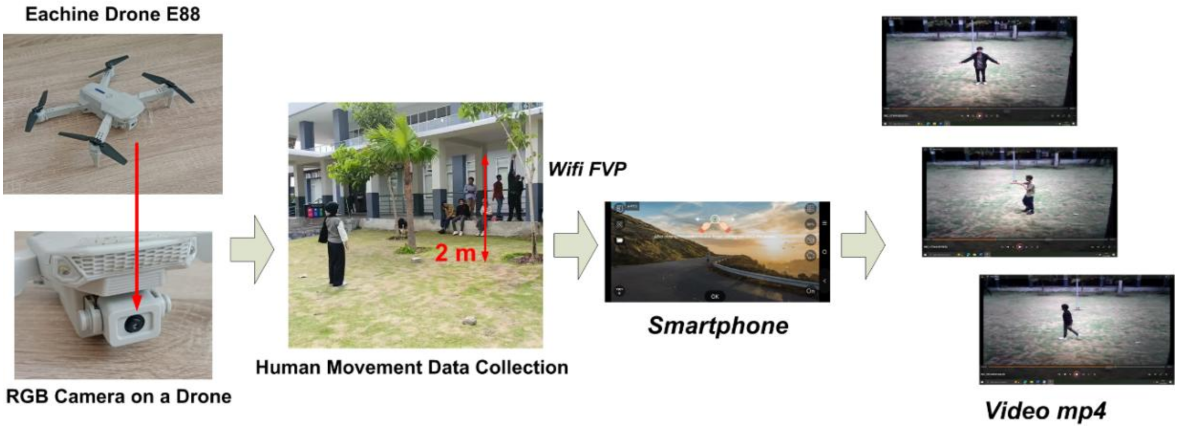


Fig. 2. Video data collection using a drone-mounted camera

2.2. Joint Detection and Feature Extraction

Each movement is detected and tracked using a video dataset captured by a drone-mounted camera, where each frame contains an individual performing one of six different movements: walking, running, jogging, boxing, waving, and clapping. To extract joint movement features, joint detection must first be performed. This is achieved using a pose estimation algorithm implemented through the MediaPipe and OpenCV platforms. Fig. 3 illustrates a sample of the pose estimation outcomes applied to the six movement types analyzed in this study. The calculation of joint angles and distances between body joints is highly dependent on the accuracy of joint points. Small errors in position detection can lead to large deviations in the calculation of angles or movements. A quantitative evaluation of joint point detection accuracy is necessary to ensure the reliability of pose estimation-based systems. Therefore, datasets with a maximum acceptable error are selected, as demonstrated in previous research [26].

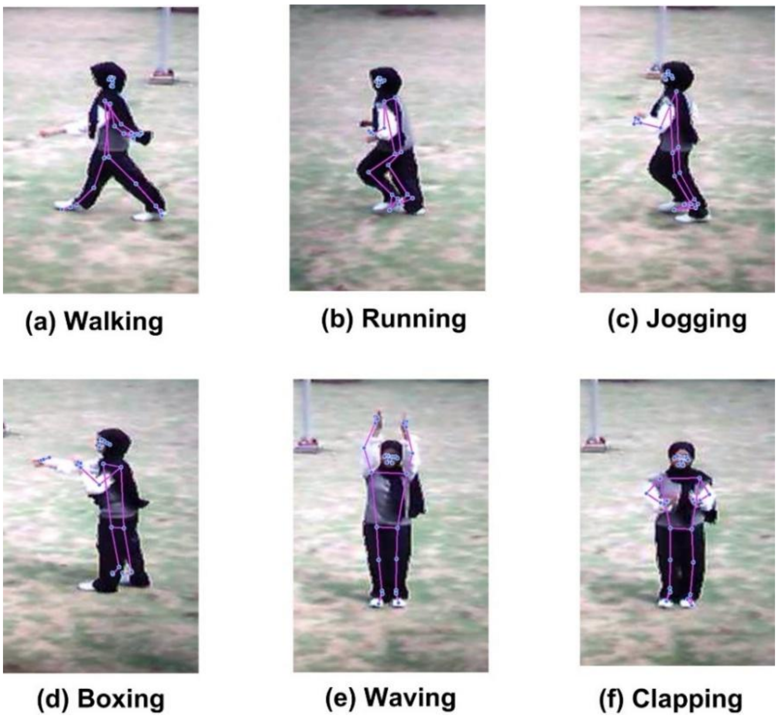


Fig. 3. Pose estimation for joint detection in the six movements

The pose estimation process aims to determine the joint coordinates that form a 2D skeletal structure for each sequence of image frames [45]. These joint points represent specific anatomical locations corresponding to key body joints. The process generates landmark points that are essential for identifying joint movements. In this study, feature extraction for person and activity recognition is based on five selected landmark points: the shoulder, wrist, waist, knee, and left ankle. Variable identifiers for each joint point, expressed as $J_{p_{(x,y,z)}}$, are provided in Table 1. Tracking the general movement of the body's joints depends on the choice of these joints. It is possible to monitor the human body's movement patterns during activities by using these reference points. These reference joint points are vectors that are used to extract joint information from each movement.

Table 1. Selected landmark joints and their variables

Landmark	Variables
Left Shoulder Joint	$J_{1(x,y,z)}$
Left Wrist Joint	$J_{2(x,y,z)}$
Left Hip Joint	$J_{3(x,y,z)}$
Left Knee Joint	$J_{4(x,y,z)}$
Left Ankle Joint	$J_{5(x,y,z)}$

Next, the feature pattern utilized in this study includes joint positions and angles, as shown in Fig. 4. When a person moves, body parts move and change position, particularly in the swing and joint angles. These changes in joint position allow for the identification of a movement by obtaining its features. The sequence of position data, as features, can describe the characteristics of an activity based on the movements of selected joints. A person's movement features are extracted from each image frame using joint landmark points in vector form from the skeleton model. These features are organized into sequential data, which consist of the distance of the knee joint position$|S_k|$, ankle joint position $|S_a|$, wrist joint position $|S_w|$, hip joint angle θ_h, knee joint angle θ_k, and wrist joint angle θ_w. Each distance and angle are calculated using equations (1) to (6).

$$|S_k| = \sqrt{\left(J_{3(x,y,z)} - J_{4(x,y,z)}\right)^2} \tag{1}$$

$$|S_a| = \sqrt{\left(J_{3(x,y,z)} - J_{5(x,y,z)}\right)^2} \tag{2}$$

$$|S_w| = \sqrt{\left(J_{3(x,y,z)} - J_{2(x,y,z)}\right)^2} \tag{3}$$

$$\theta_h = \cos^{-1}\left(\frac{\overrightarrow{J_{3(x,y,z)}} \cdot \overrightarrow{J_{4(x,y,z)}}}{\left|\overrightarrow{J_{3(x,y,z)}}\right| \cdot \left|\overrightarrow{J_{4(x,y,z)}}\right|}\right) \tag{4}$$

$$\theta_k = \cos^{-1}\left(\frac{\overrightarrow{J_{4(x,y,z)}} \cdot \overrightarrow{J_{5(x,y,z)}}}{\left|\overrightarrow{J_{4(x,y,z)}}\right| \cdot \left|\overrightarrow{J_{5(x,y,z)}}\right|}\right) \tag{5}$$

$$\theta_w = \cos^{-1}\left(\frac{\overrightarrow{J_{1(x,y,z)}} \cdot \overrightarrow{J_{2(x,y,z)}}}{\left|\overrightarrow{J_{1(x,y,z)}}\right| \cdot \left|\overrightarrow{J_{2(x,y,z)}}\right|}\right) \tag{6}$$

Next, the feature pattern for a single body movement of the subject includes joint positions and angles, which are organized into sequential data comprising $|S_k|$, $|S_a|$, $|S_w|$, θ_h, θ_k, and θ_w for every 30 image frames f, as follows:

$$Feature = \begin{bmatrix} |S_k|_{f=1}, & |S_k|_{f=2}, & |S_k|_{f=3}, & \cdots & |S_k|_{f=30}, \\ |S_a|_{f=1}, & |S_a|_{f=2}, & |S_a|_{f=3}, & \cdots & |S_a|_{f=30,} \\ |S_w|_{f=1}, & |S_w|_{f=2}, & |S_w|_{f=3}, & \cdots & |S_w|_{f=30}, \\ \theta_{h_{f=1}}, & \theta_{h_{f=2}}, & \theta_{h_{f=3}}, & \cdots & \theta_{h_{f=30}}, \\ \theta_{k_{f=1}}, & \theta_{k_{f=2}}, & \theta_{k_{f=3}}, & \cdots & \theta_{k_{f=30}}, \\ \theta_{w_{f=1}}, & \theta_{w_{f=2}}, & \theta_{w_{f=3}}, & \cdots & \theta_{w_{f=30}} \end{bmatrix} \tag{7}$$

2.3. Feature Data Selection for Person and Activity Recognition

The feature data selection in this study is categorized into two main tasks: person recognition and activity recognition. To recognize person based on body joint movements, the model leverages motion features that characterize a person's walking pattern, as depicted in Fig. 3 (a). The walking pattern is a relatively consistent activity performed in a repetitive manner and without object interaction. For activity recognition, it utilizes motion feature data derived from five distinct actions: running, jogging, boxing, waving, and clapping, as illustrated in Fig. 3 (b) through Fig. 3 (f).

2.4. Classification Using Deep Learning Models

This research contributes by evaluating the performance of deep learning models in classification tasks for person and activity recognition, based on features extracted from 1D time-series data. The deep learning classification models proposed in this study, 1D-CNN, LeNet, AlexNet, and AlexNet-LSTM, were developed and executed on a computer system using the Python programming language within the Jupyter Lab environment. The reference of joint positions and angles Fig. 4.

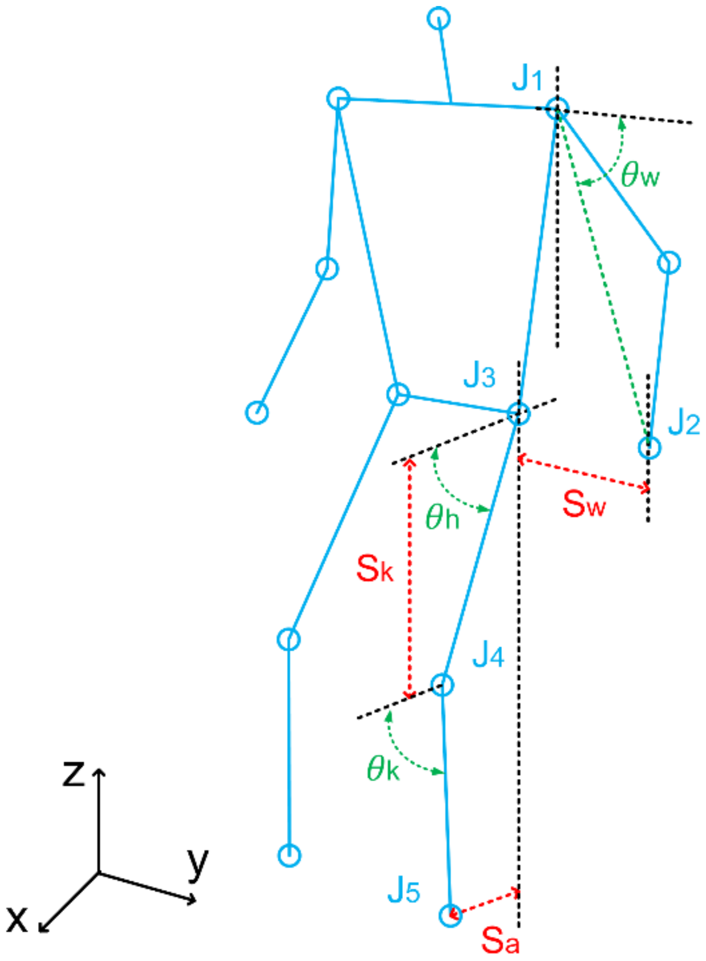


Fig. 4. The reference of joint positions and angles

2.4.1. 1D-CNN Classifier Model

In this study, the 1D-CNN classification model is designed to process one-dimensional sequential inputs, such as time-series data. Its architecture operates similarly to standard convolutional neural networks, applying convolutional filters directly to the input sequence [46]. In the context of person and activity classification, this model transforms numerical features into a structure optimized for a 1D-CNN, which is more effective in handling temporal data sequences [47], [48]. The architecture of the 1D-CNN classifier model, as shown in Fig. 5, consists of an initial Conv1D layer with 32 filters and a kernel size of 179, which extracts features from the input data. This kernel size allows the network to summarize the temporal structure early on, thus allowing for a reduction in the dimensionality of the sequence before deeper layers. A pooling layer is then employed to perform dimensionality reduction on the extracted features. To enhance feature extraction, a further Conv1D layer with 64 convolutional filters and a kernel size of 3 is applied to further process the data and accelerate training through a reduction in model complexity of parameters. In the final stage, the architecture incorporates a dense layer activated by ReLU, succeeded by a Softmax-activated layer to classify features derived from joint motion data.

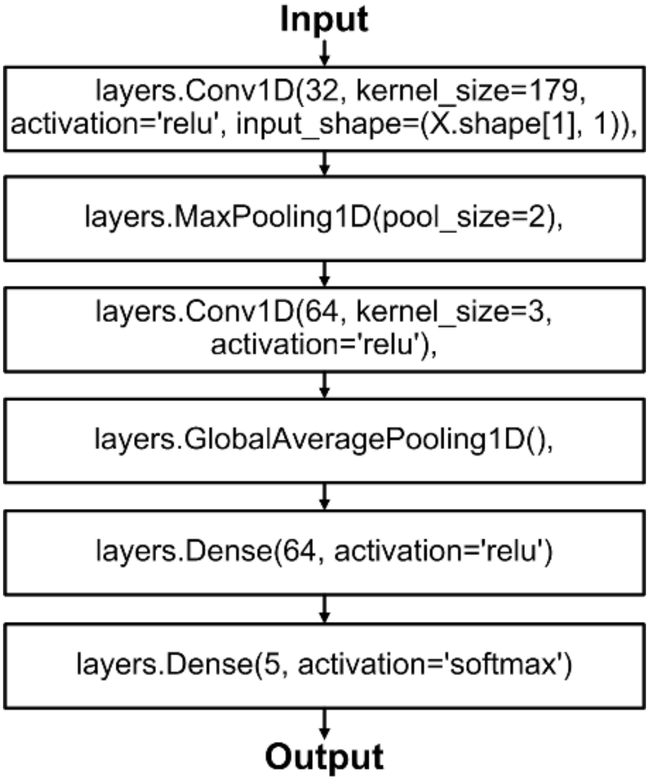
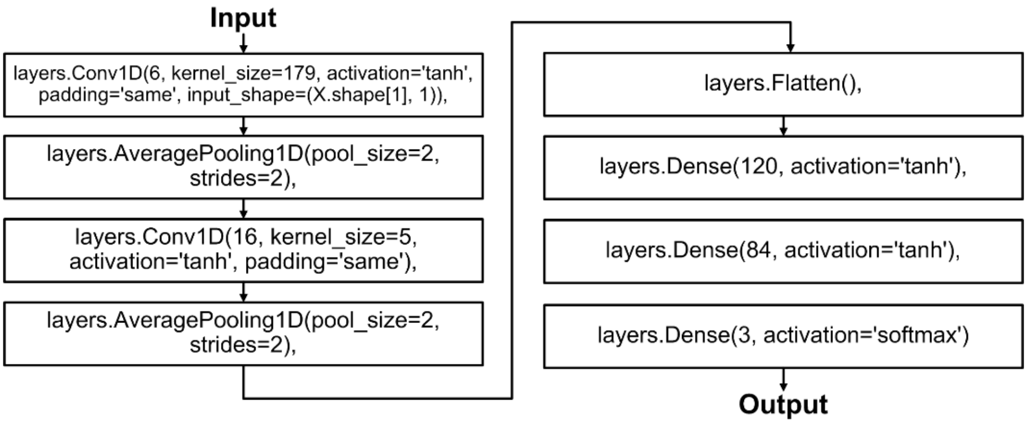

Input
layers.Conv1D(32, kernel_size=179, activation='relu', input_shape=(X.shape[1], 1)),
layers.MaxPooling1D(pool_size=2),
layers.Conv1D(64, kernel_size=3, activation='relu'),
layers.GlobalAveragePooling1D(),
layers.Dense(64, activation='relu')
layers.Dense(5, activation='softmax')
Output

Fig. 5. 1D-CNN classifier model architecture

2.4.2. LeNet Classifier Model

The LeNet classifier model is a convolutional neural network (CNN) architecture designed for classification tasks. Developed by Yann LeCun, the original LeNet architecture was created for handwritten digit recognition [49]. The model consists of convolutional layers that perform feature extraction, followed by pooling layers to reduce feature dimensionality [50], [51]. For this study, the LeNet architecture is adapted into a 1D-CNN structure to process numerical data, utilizing one-dimensional convolutional layers (Conv1D) to extract patterns from the input features. The architecture of the LeNet classifier model is shown in Fig. 6. LeNet comprises five layers, including three convolutional layers with tanh activation functions and average pooling layers to decrease the complexity of the feature maps before passing the data to the fully connected layers. The output is then processed through two fully connected layers, followed by a Softmax activation layer for classification.

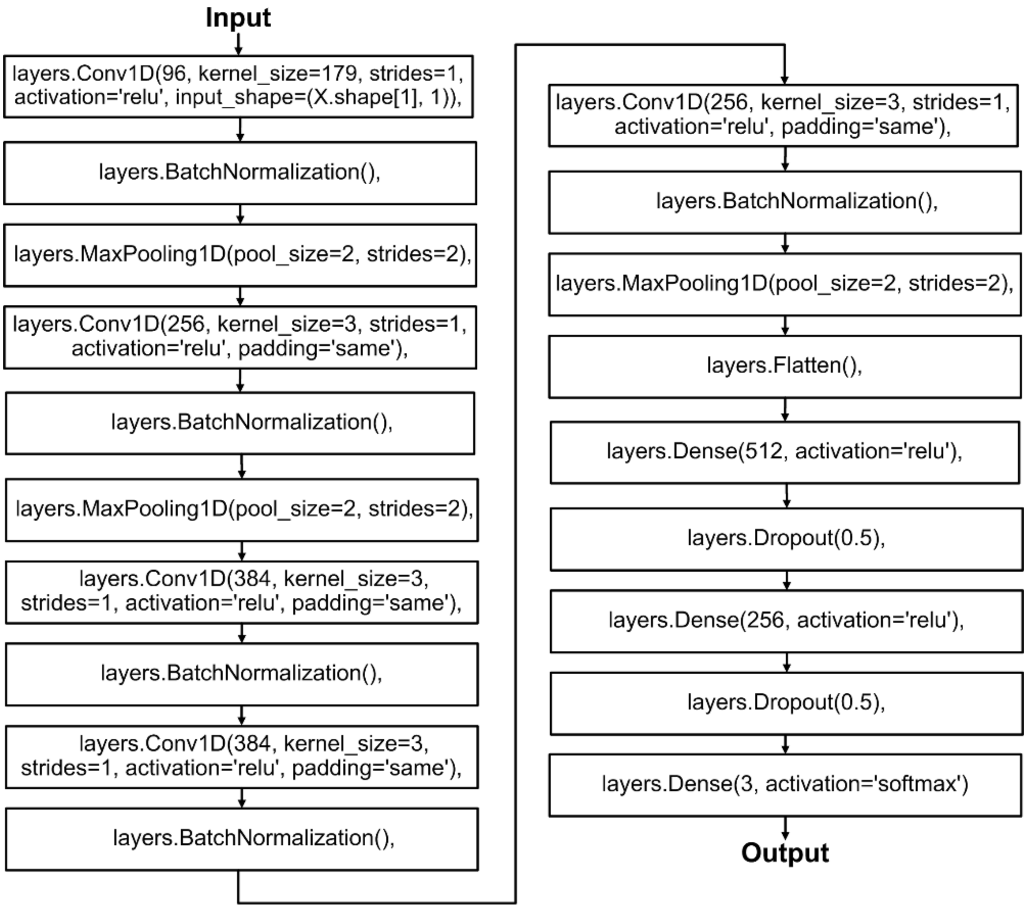


```
layers.Conv1D(6, kernel_size=179, activation='tanh', padding='same', input_shape=(X.shape[1], 1)),
layers.AveragePooling1D(pool_size=2, strides=2),
layers.Conv1D(16, kernel_size=5, activation='tanh', padding='same'),
layers.AveragePooling1D(pool_size=2, strides=2),
layers.Flatten(),
layers.Dense(120, activation='tanh'),
layers.Dense(84, activation='tanh'),
layers.Dense(3, activation='softmax')
```

Fig. 6. LeNet classifier model architecture

2.4.3. AlexNet Classifier Model

The AlexNet classifier model is a deep learning neural network originally developed by Alex Krizhevsky for image classification tasks [52], [53]. The AlexNet architecture integrates a series of convolutional layers for extracting relevant features, alongside batch normalization layers that facilitate more stable and efficient training [54], [55]. In this study, the AlexNet model is adapted for sequential data by employing 1D convolutional layers and max pooling layers. These adaptations allow for dimensionality reduction of sequential data while preserving critical information. The processed data is subsequently passed through two fully connected layers, with dropout applied in deeper layers to mitigate overfitting. The classification process is completed by connecting the output to a Softmax activation layer. The architectural layout of the AlexNet model is illustrated in Fig. 7.

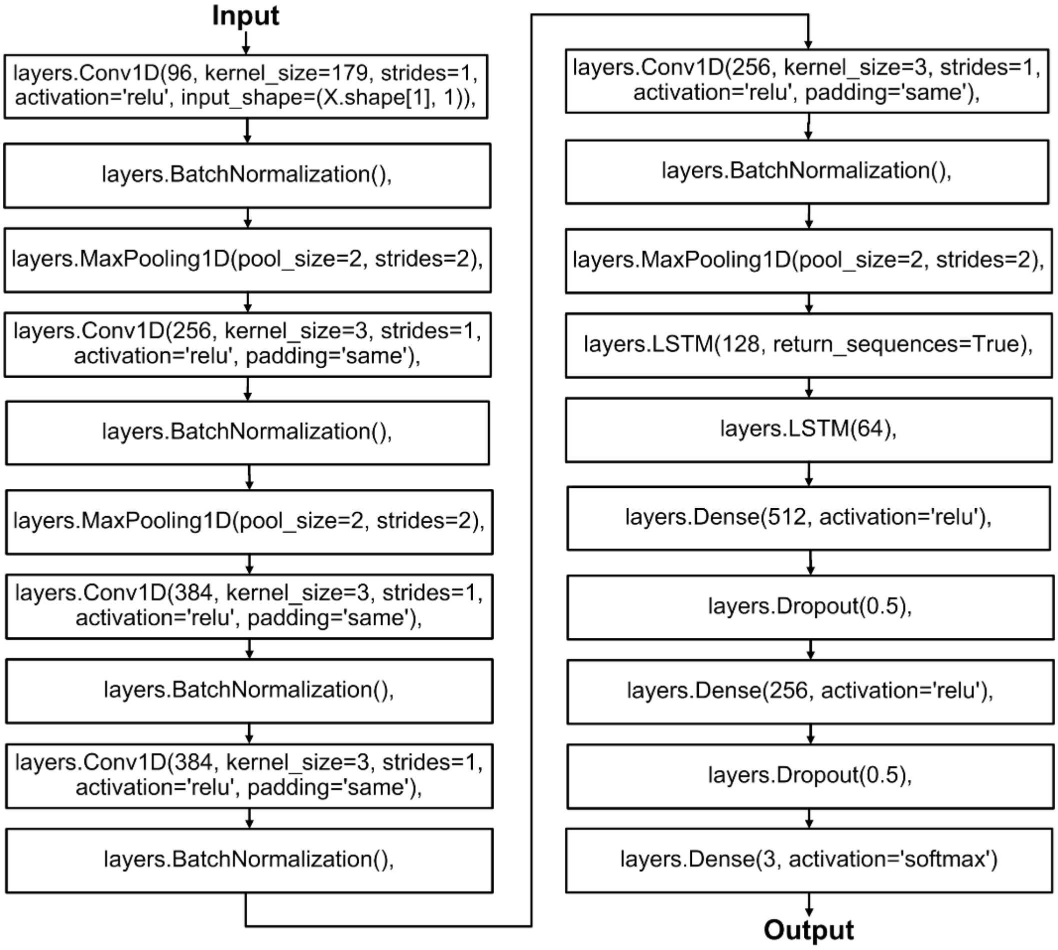


```
layers.Conv1D(96, kernel_size=179, strides=1, activation='relu', input_shape=(X.shape[1], 1)),
layers.BatchNormalization(),
layers.MaxPooling1D(pool_size=2, strides=2),
layers.Conv1D(256, kernel_size=3, strides=1, activation='relu', padding='same'),
layers.BatchNormalization(),
layers.MaxPooling1D(pool_size=2, strides=2),
layers.Conv1D(384, kernel_size=3, strides=1, activation='relu', padding='same'),
layers.BatchNormalization(),
layers.Conv1D(384, kernel_size=3, strides=1, activation='relu', padding='same'),
layers.BatchNormalization(),
layers.Conv1D(256, kernel_size=3, strides=1, activation='relu', padding='same'),
layers.BatchNormalization(),
layers.MaxPooling1D(pool_size=2, strides=2),
layers.Flatten(),
layers.Dense(512, activation='relu'),
layers.Dropout(0.5),
layers.Dense(256, activation='relu'),
layers.Dropout(0.5),
layers.Dense(3, activation='softmax')
```

Fig. 7. AlexNet classifier model architecture

2.4.4. AlexNet-LSTM Classifier Model

The AlexNet-LSTM model represents a hybrid architecture that merges the AlexNet and LSTM architecture. By combining the feature extraction power of CNN with the sequential learning capability of LSTM, the AlexNet-LSTM model demonstrates superior performance in sequential data classification when compared to traditional models [56], [57]. AlexNet is specifically designed for spatial feature extraction, while LSTM excels in processing sequential data [58]. The architecture of the AlexNet-LSTM classifier model is depicted in Fig. 8. As part of the model framework in this study, sequential data representing joint positions and angles are first passed through several AlexNet-based 1D convolutional layers to capture essential spatial features. Batch normalization and max pooling layers are then applied to accelerate training and reduce the dimensionality of the extracted features. Following the feature extraction phase, the output is passed to the LSTM layers, that learn the temporal dependencies present within the data.



```
layers.Conv1D(96, kernel_size=179, strides=1, activation='relu', input_shape=(X.shape[1], 1)),
layers.BatchNormalization(),
layers.MaxPooling1D(pool_size=2, strides=2),
layers.Conv1D(256, kernel_size=3, strides=1, activation='relu', padding='same'),
layers.BatchNormalization(),
layers.MaxPooling1D(pool_size=2, strides=2),
layers.Conv1D(384, kernel_size=3, strides=1, activation='relu', padding='same'),
layers.BatchNormalization(),
layers.Conv1D(384, kernel_size=3, strides=1, activation='relu', padding='same'),
layers.BatchNormalization(),
layers.Conv1D(256, kernel_size=3, strides=1, activation='relu', padding='same'),
layers.BatchNormalization(),
layers.MaxPooling1D(pool_size=2, strides=2),
layers.LSTM(128, return_sequences=True),
layers.LSTM(64),
layers.Dense(512, activation='relu'),
layers.Dropout(0.5),
layers.Dense(256, activation='relu'),
layers.Dropout(0.5),
layers.Dense(3, activation='softmax')
```

Fig. 8. AlexNet-LSTM classifier model architecture

2.5. Performance Evaluation of Deep Learning Classification Models

To evaluate the performance of the proposed deep learning models for person and activity recognition tasks, a series of experiments were conducted. The models evaluated include 1D-CNN, LeNet, AlexNet, and AlexNet-LSTM, all utilizing one-dimensional time-series data features extracted from the dataset. The evaluation process involved comparing the models' performance based on precision, recall, F1-score, and accuracy, which were computed using equations (8)-(11). These results were then compared to those obtained from other state-of-the-art (SOTA) classification models, as referenced in [59], [60]. To further analyze model performance, the multiclass confusion matrix was utilized to compute the values of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

3. Results and Discussion

To assess the effectiveness and contribution of deep learning architectures in person and activity recognition tasks based on the proposed joint position and angle features, this study conducted two separate experimental evaluations. The first experiment focused on evaluating model performance in recognizing individual identities, while the second focused on recognizing human activities. Both experiments employed four different deep learning architectures: 1D-CNN, LeNet, AlexNet, and AlexNet-LSTM. Model performance was evaluated using standard statistical metrics, including average accuracy, precision, recall, and F1-score, to measure their effectiveness in classifying both individual identities and the types of activities performed.

3.1. Classification Evaluation for Person Recognition Using Deep Learning

In the person recognition task, all four models were trained using motion features derived from walking patterns of ten distinct subjects. The data extraction process yielded a total of 120 datasets, which were subsequently divided into training and testing sets. Considering the relatively small dataset size, accuracy validation was performed using three different train-test split ratios: 70:30, 80:20, and 90:10. Each configuration was trained for 100 epochs to ensure a fair performance comparison. The optimal train-test ratio, based on overall classification performance, was selected as the final configuration for model evaluation. In this experiment, the primary goal is to evaluate model performance under various conditions, not to optimize the model. Therefore, it provides more diverse data and more testing data. Although the main validation focus was on 70:30, 80:20, and 90:10 splits, we also present results for 50:50 and 60:40 splits in Table 2 for completeness and to illustrate performance trends across a wider range of data distributions. The 50:50 or 60:40 ratio is used when the dataset size is limited or to evaluate the overall model performance. This ratio provides a larger test set size, allowing for a more stable and representative evaluation of pattern recognition within the dataset.

The performance comparison of the deep learning models in person recognition is presented in Table 2, highlighting the results across different split ratios. The analysis demonstrates that the AlexNet-LSTM architecture consistently outperformed other models, validating the effectiveness of combining convolutional and sequential learning for person identification tasks using the proposed joint position and angle features.

Validation of the model using a 70:30 train-test data split, 70% for training and 30% for testing, demonstrated superior performance in person recognition compared to other split configurations. This data split yielded higher accuracy, likely due to the relatively larger training portion, which contributed significantly to performance improvement. As presented in Table 2, the AlexNet-LSTM model achieved the highest classification performance for person recognition, with an accuracy of 0.8544, a precision of 0.9161, a recall of 0.8575, and an F1-score of 0.8332.

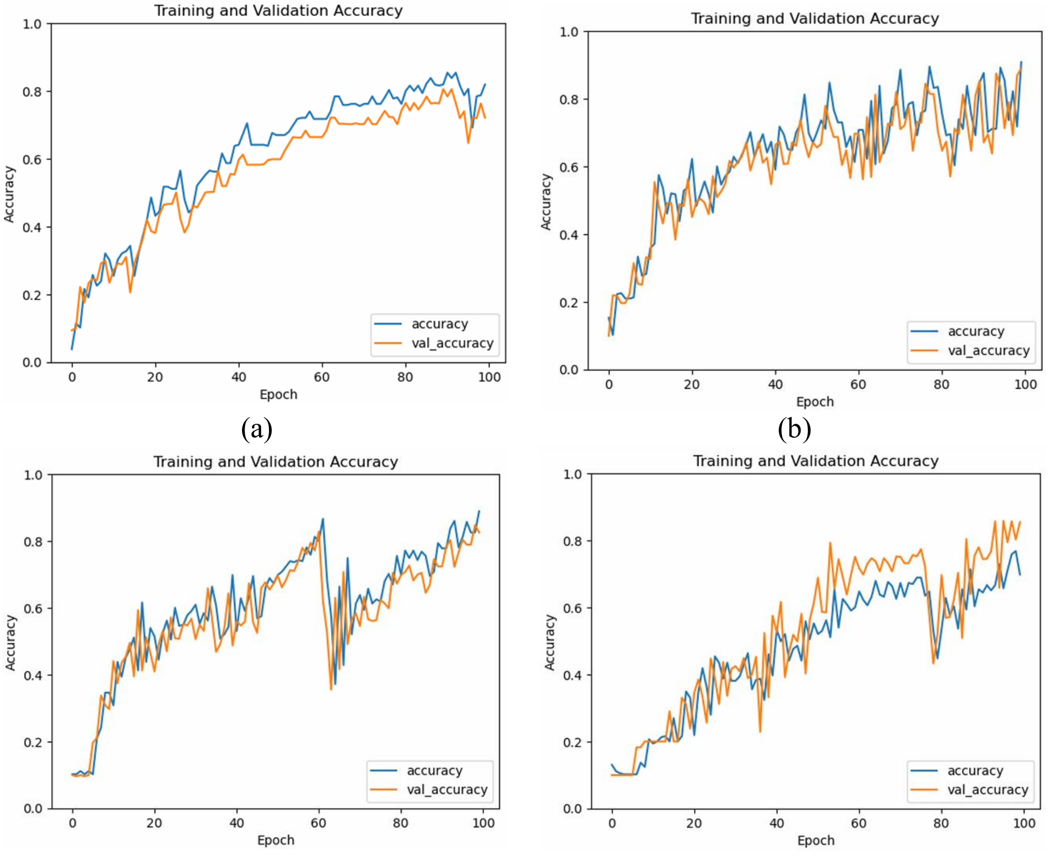
The 70:30 ratio in Table 2 can be considered a balanced approach, providing sufficient data for training while maintaining the validity of the performance evaluation. While a ratio like 90:10 may seem advantageous from a training perspective, its use can increase the risk of overfitting and decrease the reliability of the model evaluation. The evaluation results indicate that AlexNet outperforms other models in activity recognition tasks. Several architectural aspects and data characteristics explain this superiority. First, AlexNet has a relatively deep but not overly complex structure, making it suitable for use on moderately sized activity datasets without a high risk of overfitting. Furthermore, the large filters in the early layers of AlexNet allow the model to effectively capture global spatial patterns, which is crucial when feature-based activity representation is used.

Table 2. The performance of classifier models for person recognition

Data Split Train: Test	Classifier Models	Accuracy	Precision	Recall	F1-Score
50:50	1D-CNN	0.5448	0.5385	0.5537	0.5028
	LeNet	0.6438	0.5176	0.6417	0.5395
	AlexNet	0.4210	0.3022	0.4233	0.3426
	AlexNet-LSTM	0.7105	0.5557	0.7000	0.6049
60:40	1D-CNN	0.7825	0.8710	0.7895	0.7520
	LeNet	0.8190	0.8794	0.8270	0.8082
	AlexNet	0.8381	0.8905	0.8405	0.8298
	AlexNet-LSTM	0.8571	0.8783	0.8544	0.8470
70:30	1D-CNN	0.7211	0.7258	0.7290	0.6738
	LeNet	0.8898	0.8477	0.9000	0.8646
	AlexNet	0.8259	0.7991	0.8350	0.7969
	AlexNet-LSTM	**0.8544**	**0.9161**	**0.8575**	**0.8332**
80:20	1D-CNN	0.7702	0.7326	0.7755	0.7357
	LeNet	0.7476	0.7075	0.7542	0.7155
	AlexNet	0.7940	0.7851	0.7993	0.7482
	AlexNet-LSTM	0.6583	0.5651	0.6672	0.6023
90:10	1D-CNN	0.7492	0.7892	0.7499	0.7403
	LeNet	0.7090	0.7620	0.7107	0.6858
	AlexNet	0.6328	0.6083	0.6344	0.5826
	AlexNet-LSTM	0.7365	0.8444	0.7375	0.6973

Fig. 9 illustrates a comparative analysis of classification accuracy for person recognition using four models: 1D-CNN, LeNet, AlexNet, and AlexNet-LSTM, all evaluated under the 70:30 train-test split. The accuracy trend indicates that all models exhibited improved performance as the number of training epochs increased, demonstrating a direct correlation between epoch progression and accuracy. The graph shows that on small datasets, AlexNet-LSTM tends to experience a large accuracy gap between training and validation, while AlexNet is relatively more stable. Among these models, AlexNet-LSTM consistently outperformed the others, achieving a training accuracy of 0.7510, which stabilized at epoch 60 out of 100 total epochs. However, a slight performance drops between epoch 80 and 87 suggests the onset of overfitting. Overall, the training trend confirms the model's ability to effectively learn and recognize individuals in the validation set.

Table 3 shows the evaluation report of the AlexNet-LSTM classification model, which was trained and tested using a 70% training and 30% testing data split, in terms of accuracy, precision, recall, and F1-score for each Id individual label class fed into the model. The overall classification accuracy, obtained using consistent joint position and angle features across all Id individual labels, reached 0.85. Utilizing only these features enabled the AlexNet-LSTM model to minimize classification ambiguity. In general, the model achieved a precision of 1.00 across most person labels. However, in this case, the precision for Id individual 4 and Id individual 9 was relatively lower, recorded at 0.62 and 0.54, respectively. These evaluation results indicate that the selected body joint features are effective in enhancing the performance of deep learning-based classification, particularly in addressing challenges associated with a limited dataset in person recognition tasks. The feature pattern partially represents a single subject's body movement, indicating that the sequential data is temporally one-dimensional. This observation aligns with the design of a hybrid AlexNet and LSTM model to distinguish each feature pattern. This finding emphasizes the possible use of spatiotemporal analysis to capture the specific characteristics of a person's movement.



(a) (b)

(c) (d)

Fig. 9. Comparison of model accuracy for person recognition using: (a) 1D-CNN, (b) LeNet, (c) AlexNet, and (d) AlexNet-LSTM, with a 70:30 train-test data split

The confusion matrix results in Fig. 10 show the prediction results from the diagonal aspect of the matrix using AlexNet-LSTM on a 70:30 train-test split. The proposed architecture model is able to show more correct predicted results.

3.2. Classification Evaluation for Activity Recognition Using Deep Learning

In the performance evaluation of human activity recognition, the 1D-CNN, LeNet, AlexNet, and AlexNet-LSTM models were trained using motion features derived from five distinct activities: boxing, waving, clapping, running, and jogging. The evaluation results include accuracy, precision, recall, and F1-score for each model, as presented in Table 4.

Table 3. Person recognition classification report using AlexNet-LSTM on a 70:30 train-test split

Person Label Classes	Accuracy	Precision	Recall	F1-Score
Id individual 1	0.85	1.00	1.00	1.00
Id individual 2		1.00	1.00	1.00
Id individual 3		1.00	1.00	1.00
Id individual 4		0.62	1.00	0.76
Id individual 5		1.00	0.41	0.58
Id individual 6		1.00	1.00	1.00
Id individual 7		1.00	1.00	1.00
Id individual 8		1.00	0.16	0.28
Id individual 9		0.54	1.00	0.71
Id individual 10		1.00	1.00	1.00
Average	0.85	0.92	0.86	0.83

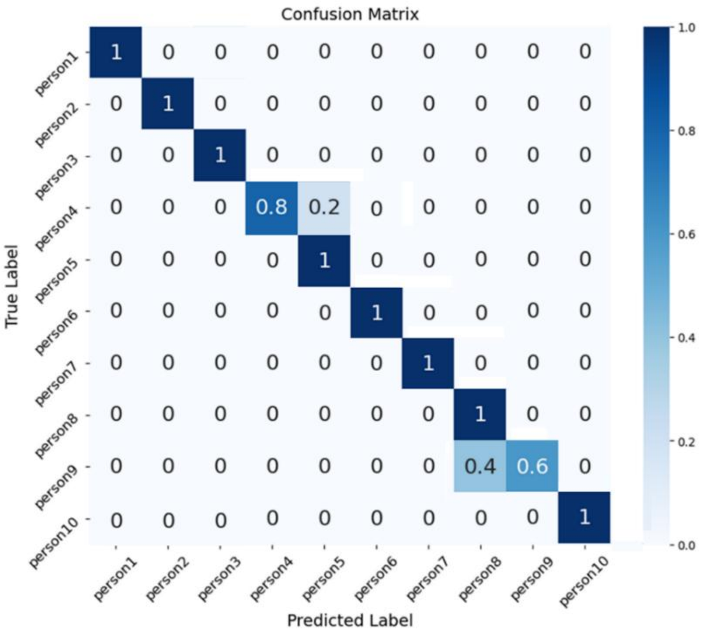


Fig. 10. Confusion matrix for person recognition using AlexNet-LSTM on a 70:30 train-test split

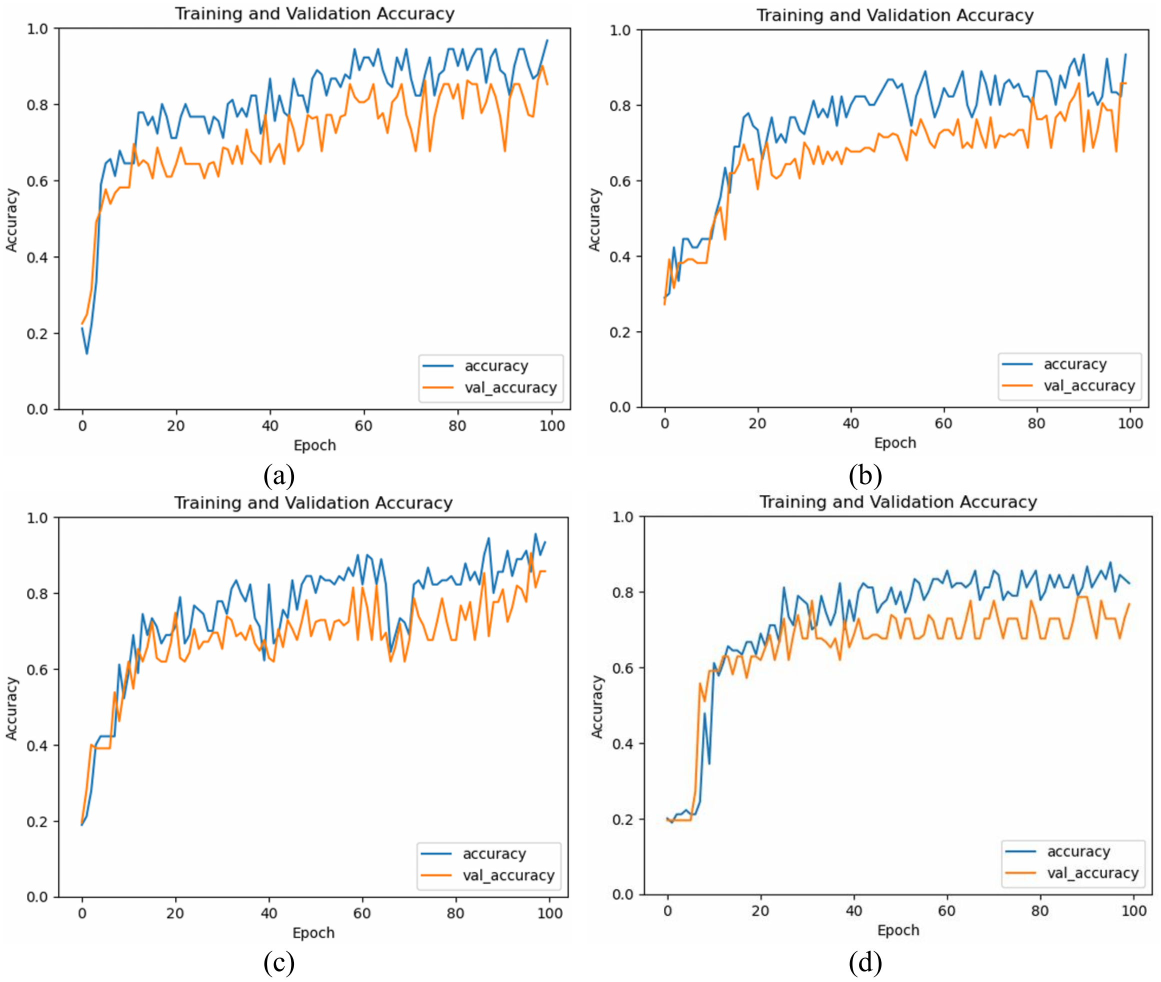
Table 4. The performance of classifier models for activity recognition

Ratio Data Train: Test	Classifier Models	Accuracy	Precision	Recall	F1-Score
50:50	1D-CNN	0.7800	0.7958	0.7307	0.7367
	LeNet	0.7600	0.8273	0.7785	0.7192
	AlexNet	0.7800	0.8969	0.7873	0.7583
	AlexNet-LSTM	0.7150	0.6838	0.6962	0.6876
60:40	1D-CNN	0.8111	0.8077	0.7981	0.7981
	LeNet	0.7278	0.6116	0.7500	0.6633
	AlexNet	0.6611	0.6305	0.6619	0.6377
	AlexNet-LSTM	0.7625	0.6328	0.7547	0.6810
70:30	1D-CNN	0.8524	0.8524	0.8377	0.8413
	LeNet	0.8571	0.8790	0.8599	0.8442
	AlexNet	**0.8571**	**0.8442**	**0.8599**	**0.8463**
	AlexNet-LSTM	0.7667	0.7551	0.7790	0.7509
80:20	1D-CNN	0.8125	0.8256	0.8004	0.7973
	LeNet	0.8167	0.8297	0.8036	0.8016
	AlexNet	0.7333	0.7823	0.7244	0.6895
	AlexNet-LSTM	0.7531	0.6895	0.7429	0.6903
90:10	1D-CNN	0.6889	0.7690	0.6786	0.6524
	LeNet	0.7333	0.7855	0.7357	0.6969
	AlexNet	0.6926	0.5722	0.6984	0.6243
	AlexNet-LSTM	0.6722	0.7017	0.6767	0.6558

As shown in Table 4, although the AlexNet-LSTM model was developed by integrating AlexNet and LSTM layers, the standalone AlexNet model outperformed the other models in terms of overall performance. Under the 70:30 data split configuration, AlexNet achieved an accuracy of 0.8571, a precision of 0.8442, a recall of 0.8599, and an F1-score of 0.8463. This can be attributed to the greater complexity and higher number of parameters in the AlexNet-LSTM model compared to AlexNet, which increases the risk of overfitting, particularly when working with small datasets. Therefore, with limited and relatively simple data, models with fewer parameters, such as AlexNet, tend to yield better performance.

Table 5 presents the performance analysis of the AlexNet model in classifying human activities across five activity categories, with a focus on a 70:30 train-test split. Based on the results obtained, the AlexNet model achieved an average accuracy of 0.86, an average precision of 0.92, an average recall of 0.86, and an average F1-score of 0.83 in recognizing human activities. The results indicate that misclassification in terms of precision is more prominent for running and jogging activities. The primary reason is that the classifier often confuses jogging with running, treating them as similar patterns, leading to a lower precision score for running, which is frequently misclassified as jogging. Additionally, misclassification also occurred for the waving activity, which was often incorrectly identified as jogging. Misclassification errors in activity recognition, particularly between activities with similar movement patterns, such as running and jogging, have technical implications. These two activities have similar movement sequences and joint positions but differ in intensity, speed, and rhythm. Classification systems that are not sufficiently sensitive to these differences can produce erroneous predictions, and these errors have varying impacts depending on the dataset and its use. This may be attributed to the unreadable leg angle data during jogging and the similarity in arm movements between the two activities. In summary, the experimental outcomes demonstrate that the AlexNet-based architecture achieves reliable performance in classifying five types of human movement activities from drone camera video data.

Fig. 11 shows the graphical representation of training and accuracy of the four models with a 70:30 train-test split for activity recognition indicates that the accuracy during training tends to stabilize after the first 10 epochs and continues to improve, albeit with minor fluctuations. Meanwhile, the training accuracy is generally higher compared to the validation accuracy. Focusing on the AlexNet model, it consistently demonstrates higher accuracy than the other models, considering the dataset size and the ratio of the split.



(a)

(b)

(c)

(d)

Fig. 11. Comparison of model accuracy for activity recognition using: (a) 1D-CNN, (b) LeNet, (c) AlexNet, and (d) AlexNet-LSTM, with a 70:30 train-test data split

The test results also produced a confusion matrix for activity recognition using AlexNet on a 70:30 train-test split shown in Fig. 12, representing the accuracy of classifying five types of activities from motion features data.

Table 5. Person recognition classification report using AlexNet on a 70:30 train-test split

Person Label Classes	Accuracy	Precision	Recall	F1-Score
running	0.86	0.76	1.00	0.86
jogging		0.68	0.51	0.58
boxing		1.00	1.00	1.00
waving		0.79	0.79	0.79
clapping		1.00	0.41	0.58
Average	0.86	0.92	0.86	0.83

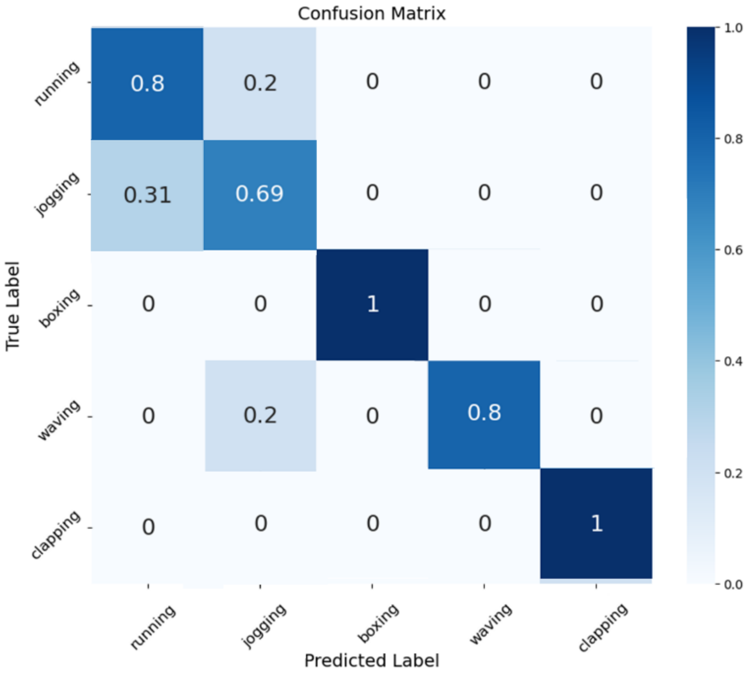


Fig. 12. Confusion matrix for activity recognition using AlexNet on a 70:30 train-test split

4. Conclusion

The findings of this study confirm the applicability and performance of deep learning approaches for recognizing person and activity using video data captured by a drone-mounted camera. The recorded video sequences involved only 10 subjects and a limited number of activities, which directly impacts the generalizability of the results. By leveraging motion features derived from joint positions and angles, the proposed models were able to classify human identity and actions with high accuracy, even under the constraints of low-resolution images. Among the evaluated architectures, the hybrid AlexNet-LSTM model outperformed the others in person recognition, achieving a classification accuracy of 0.8544, a precision of 0.9161, a recall of 0.8575, and an F1-score of 0.8332, affirming the benefits of combining spatial and temporal feature extraction. Conversely, for activity recognition, the standard AlexNet model yielded the best performance, with an accuracy of 0.8571, a precision of 0.8442, a recall of 0.8599, and an F1-score of 0.8463, suggesting that simpler models with fewer parameters may be more suitable when dealing with limited datasets. Overall, the results highlight the potential of deep learning-based frameworks in enhancing surveillance systems, particularly when applied to video data captured by drone-mounted camera. Future study could focus on expanding the data set, improving robustness against subject variability, and integrating multi-person tracking to further improve real world applicability.

Author Contribution: All authors contributed equally to the main contributor of this paper. All authors read and approved the final paper.

Funding: Ministry of Education, Culture, Research, and Technology, Republic of Indonesia

Acknowledgment: The authors acknowledge the support and funding provided by the Ministry of Higher Education, Science, and Technology (KEMDIKTISAINTEK) in part by the Center for Higher Education Funding and Assessment (PPAPT) through the Indonesia Endowment Fund for Education (LPDP) under grant of Indonesian Education Scholarship (BPI) 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] B. Kwon and T. Kim, “Toward an Online Continual Learning Architecture for Intrusion Detection of Video Surveillance,” *IEEE Access*, vol. 10, pp. 89732-89744, 2022, https://doi.org/10.1109/ACCESS.2022.3201139.

[2] M. Mohamed Zaidi *et al*., “Suspicious Human Activity Recognition From Surveillance Videos Using Deep Learning,” *IEEE Access*, vol. 12, pp. 105497-105510, 2024, https://doi.org/10.1109/ACCESS.2024.3436653.

[3] M. M. Taye, “Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions,” *Computers*, vol. 12, no. 5, p. 91, 2023, https://doi.org/10.3390/computers12050091.

[4] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, “A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning,” *Archives of Computational Methods in Engineering*, vol. 27, no. 4, pp. 1071-1092, 2020, https://doi.org/10.1016/j.jbiomech.2024.112027.

[5] H. Wang *et al*., “Markerless gait analysis through a single camera and computer vision,” *Journal of Biomechanics*, vol. 165, p. 112027, 2024, https://doi.org/10.1016/j.jbiomech.2024.112027.

[6] E. A. Tunggadewi, E. I. Agustin, and R. T. Yunardi, “A smart wearable device based on internet of things for the safety of children in online transportation,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 708-716, 2021, http://doi.org/10.11591/ijeecs.v22.i2.pp708-716.

[7] C. Zhou, D. Feng, S. Chen, N. Ban, and J. Pan, “Portable vision-based gait assessment for post-stroke rehabilitation using an attention-based lightweight CNN,” *Expert Systems with Applications*, vol. 238, p. 122074, 2024, https://doi.org/10.1016/j.eswa.2023.122074.

[8] Y. -K. Wang, C. -T. Fan, K. -Y. Cheng and P. S. Deng, “Real-time camera anomaly detection for real-world video surveillance,” *2011 International Conference on Machine Learning and Cybernetics*, pp. 1520-1525, 2011, https://doi.org/10.1109/ICMLC.2011.6017032.

[9] S. Kapoor, A. Sharma, A. Verma, and S. Singh, “Aeriform in-action: A novel dataset for human action recognition in aerial videos,” *Pattern Recognition*, vol. 140, p. 109505, 2023, https://doi.org/10.1016/j.patcog.2023.109505.

[10] R. G. Sinclair, J. Gaio, S. D. Huazano, S. A. Wiafe, and W. C. Porter, “A Balloon Mapping Approach to Forecast Increases in PM10 from the Shrinking Shoreline of the Salton Sea,” *Geographies*, vol. 4, no. 4, pp. 630-640, 2024, https://doi.org/10.3390/geographies4040034.

[11] Y. Kaya, H. İ. Şenol, A. Y. Yiğit, and M. Yakar, “Car Detection from Very High-Resolution UAV Images Using Deep Learning Algorithms,” *Photogrammetric Engineering & Remote Sensing*, vol. 89, no. 2, pp. 117-123, 2023, https://doi.org/10.14358/PERS.22-00101R2.

[12] B. Mishra, D. Garg, P. Narang, and V. Mishra, “Drone-surveillance for search and rescue in natural disaster,” *Computer Communications*, vol. 156, pp. 1-10, 2020, https://doi.org/10.1016/j.comcom.2020.03.012.

[13] A. Srivastava, T. Badal, A. Garg, A. Vidyarthi, and R. Singh, “Recognizing human violent action using drone surveillance within real-time proximity,” *Journal of Real-Time Image Processing*, vol. 18, pp. 1851-1863, 2021, https://doi.org/10.1007/s11554-021-01171-2.

[14] E. S. Rahayu, E. M. Yuniarno, I. K. E. Purnama and M. H. Purnomo, “A Combination Model of Shifting Joint Angle Changes With 3D-Deep Convolutional Neural Network to Recognize Human Activity,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 1078-1089, 2024, https://doi.org/10.1109/TNSRE.2024.3371474.

[15] B. Kwolek, A. Michalczuk, T. Krzeszowski, A. Switonski, H. Josinski, and K. Wojciechowski, “Calibrated and synchronized multi-view video and motion capture dataset for evaluation of gait recognition,” *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 32437-32465, 2019, https://doi.org/10.1007/s11042-019-07945-y.

[16] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, “Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding,” *Pattern Recognition*, vol. 48, no. 10, pp. 3102-3112, 2015, https://doi.org/10.1016/j.patcog.2014.12.016.

[17] J. H. Yoo and M. S. Nixon, “Automated markerless analysis of human gait motion for recognition and classification,” *ETRI Journal*, vol. 33, no. 2, pp. 259-266, 2011, https://doi.org/10.4218/etrij.11.1510.0068.

[18] K. M. Oikonomou, I. Kansizoglou, I. T. Papapetros and A. Gasteratos, “A Bio-Inspired Elderly Action Recognition System for Ambient Assisted Living,” *2023 18th International Workshop on Cellular Nanoscale Networks and their Applications (CNNA)*, pp. 1-6, 2023, https://doi.org/10.1109/CNNA60945.2023.10652802.

[19] M. Grimmer *et al.*, “Lower limb joint biomechanics-based identification of gait transitions in between level walking and stair ambulation,” *PLoS One*, vol. 15, no. 9, p. e0239148, 2020, https://doi.org/10.1371/journal.pone.0239148.

[20] M. A. R. Ahad, M. Ahmed, A. Das Antar, Y. Makihara, and Y. Yagi, “Action recognition using kinematics posture feature on 3D skeleton joint locations,” *Pattern Recognition Letters*, vol. 145, pp. 216-224, 2021, https://doi.org/10.1016/j.patrec.2021.02.013.

[21] M. Ota, H. Tateuchi, T. Hashiguchi, and N. Ichihashi, “Verification of validity of gait analysis systems during treadmill walking and running using human pose tracking algorithm,” *Gait Posture*, vol. 85, pp. 290-297, 2021, https://doi.org/10.1016/j.gaitpost.2021.02.006.

[22] H. Ullah and A. Munir, “Human Activity Recognition Using Cascaded Dual Attention CNN and Bi-Directional GRU Framework,” *Journal of Imaging*, vol. 9, no. 7, p. 130, 2023, https://doi.org/10.3390/jimaging9070130.

[23] J. Stenum, M. M. Hsu, A. Y. Pantelyat, and R. T. Roemmich, “Clinical gait analysis using video-based pose estimation: Multiple perspectives, clinical populations, and measuring change,” *PLOS Digital Health*, vol. 3, no. 3, p. e0000467, 2024, https://doi.org/10.1371/journal.pdig.0000467.

[24] M. Mundt, Z. Born, M. Goldacre, and J. Alderson, “Estimating Ground Reaction Forces from Two-Dimensional Pose Data: A Biomechanics-Based Comparison of AlphaPose, BlazePose, and OpenPose,” *Sensors*, vol. 23, no. 1, p. 78, 2023, https://doi.org/10.3390/s23010078.

[25] N. Nakano *et al.*, “Evaluation of 3D Markerless Motion Capture Accuracy Using OpenPose With Multiple Video Cameras,” *Frontiers in Sports and Active Living*, vol. 2, 2020, https://doi.org/10.3389/fspor.2020.00050.

[26] R. T. Yunardi, T. A. Sardjono and R. Mardiyanto, “Motion Capture System based on RGB Camera for Human Walking Recognition using Marker-based and Markerless for Kinematics of Gait,” *2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 262-267, 2023, https://doi.org/10.1109/ISCAIE57739.2023.10164935.

[27] J. W. Kim, J. Y. Choi, E. J. Ha, and J. H. Choi, “Human Pose Estimation Using MediaPipe Pose and Optimization Method Based on a Humanoid Model,” *Applied Sciences*, vol. 13, no. 4, p. 2700, 2023, https://doi.org/10.3390/app13042700.

[28] Y. Lin, X. Jiao, and L. Zhao, “Detection of 3D Human Posture Based on Improved Mediapipe,” *Journal of Computer and Communications*, vol. 11, no. 02, pp. 102-121, 2023, https://doi.org/10.4236/jcc.2023.112008.

[29] Z. Liu, Q. Liu, W. Xu, Z. Liu, Z. Zhou, and J. Chen, “Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing,” *Procedia CIRP*, vol. 83, pp. 272-278, 2019, https://doi.org/10.1016/j.procir.2019.04.080.

[30] M. Y. Heravi, Y. Jang, I. Jeong, and S. Sarkar, “Deep learning-based activity-aware 3D human motion trajectory prediction in construction,” *Expert Systems with Applications*, vol. 239, p. 122423, 2024, https://doi.org/10.1016/j.eswa.2023.122423.

[31] L. Zhang, “Applying Deep Learning-Based Human Motion Recognition System in Sports Competition,” *Frontiers in Neurorobotics*, vol. 16, 2022, https://doi.org/10.3389/fnbot.2022.860981.

[32] T. H. Lee *et al.*, “Comparative Analysis of 1D – CNN, GRU, and LSTM for Classifying Step Duration in Elderly and Adolescents Using Computer Vision,” *International Journal of Robotics and Control Systems*, vol. 5, no. 1, pp. 426-439, 2025, https://doi.org/10.31763/ijrcs.v5i1.1588.

[33] S. Nam and S. Lee, “JT-MGCN: Joint-temporal Motion Graph Convolutional Network for Skeleton-Based Action Recognition,” *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6383-6390, 2021, https://doi.org/10.1109/ICPR48806.2021.9412533.

[34] H. Wang, B. Yu, K. Xia, J. Li, and X. Zuo, “Skeleton edge motion networks for human action recognition,” *Neurocomputing*, vol. 423, pp. 1-12, 2021, https://doi.org/10.1016/j.neucom.2020.10.037.

[35] S. S. Patil, S. S. Pardeshi, and A. D. Patange, “Health Monitoring of Milling Tool Inserts Using CNN Architectures Trained by Vibration Spectrograms,” *CMES - Computer Modeling in Engineering and Sciences*, vol. 136, no. 1, pp. 177-199, 2023, https://doi.org/10.32604/cmes.2023.025516.

[36] J. A. Gamble and J. Huang, “Convolutional Neural Network for Human Activity Recognition and Identification,” *2020 IEEE International Systems Conference (SysCon)*, pp. 1-7, 2020, https://doi.org/10.1109/SysCon47679.2020.9275924.

[37] Irfanullah, T. Hussain, A. Iqbal, B. Yang, and A. Hussain, “Real time violence detection in surveillance videos using Convolutional Neural Networks,” *Multimedia Tools and Applications*, vol. 81, no. 26, pp. 38151-38173, 2022, https://doi.org/10.1007/s11042-022-13169-4.

[38] H. Lee and D. Shin, “Beyond Information Distortion: Imaging Variable-Length Time Series Data for Classification,” *Sensors*, vol. 25, no. 3, p. 621, 2025, https://doi.org/10.3390/s25030621.

[39] J. Zhu, H. Chen and W. Ye, “Classification of Human Activities Based on Radar Signals using 1D-CNN and LSTM,” *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1-5, 2020, https://doi.org/10.1109/ISCAS45731.2020.9181233.

[40] R. T. Yunardi, T. A. Sardjono, and R. Mardiyanto, “Enhancing Surveillance Vision-Based Human Action Recognition Using Skeleton Joint Swing and Angle Feature and Modified AlexNet-LSTM,” *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 1, pp. 754-768, 2025, https://doi.org/10.22266/ijies2025.0229.53.

[41] V. B. Semwal *et al.*, “Development of the LSTM Model and Universal Polynomial Equation for All the Sub-Phases of Human Gait,” *IEEE Sensors Journal*, vol. 23, no. 14, pp. 15892-15900, 2023, https://doi.org/10.1109/JSEN.2023.3281401.

[42] M. -K. Yi, K. Han and S. O. Hwang, “Fall Detection of the Elderly Using Denoising LSTM-Based Convolutional Variant Autoencoder,” *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18556-18567, 2024, https://doi.org/10.1109/JSEN.2024.3388478.

[43] J. Engel, J. Sturm and D. Cremers, “Camera-based navigation of a low-cost quadrocopter,” *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2815-2821, 2012, https://doi.org/10.1109/IROS.2012.6385458.

[44] N. Gageik, P. Benz and S. Montenegro, “Obstacle Detection and Collision Avoidance for a UAV With Complementary Low-Cost Sensors,” *IEEE Access*, vol. 3, pp. 599-609, 2015, https://doi.org/10.1109/ACCESS.2015.2432455.

[45] R. T. Yunardi, T. A. Sardjono, and R. Mardiyanto, “Skeleton-Based Gait Recognition Using Modified Deep Convolutional Neural Networks and Long Short-Term Memory for Person Recognition,” *IEEE Access*, vol. 12, pp. 121131-121143, 2024, https://doi.org/10.1109/ACCESS.2024.3451495.

[46] A. K. Ozcanli and M. Baysal, "Islanding detection in microgrid using deep learning based on 1D CNN and CNN-LSTM networks," *Sustainable Energy, Grids and Networks*, vol. 32, p. 100839, 2022, https://doi.org/10.1016/j.segan.2022.100839.

[47] S. Guessoum *et al*., "The Short-Term Prediction of Length of Day Using 1D Convolutional Neural Networks (1D CNN)," *Sensors*, vol. 22, no. 23, p. 9517, 2022, https://doi.org/10.3390/s22239517.

[48] U. Ileri, Y. Altun, and A. Narin, "An Efficient Approach for Automatic Fault Classification Based on Data Balance and One-Dimensional Deep Learning," *Applied Sciences*, vol. 14, no. 11, p. 4899, 2024, https://doi.org/10.3390/app14114899.

[49] A. Zaibi, A. Ladgham, and A. Sakly, "A Lightweight Model for Traffic Sign Classification Based on Enhanced LeNet-5 Network," *Journal of Sensors*, vol. 2021, no. 1, pp. 1-13, 2021, https://doi.org/10.1155%2F2021%2F8870529.

[50] S. Balasubramaniam, Y. Velmurugan, D. Jaganathan, and S. Dhanasekaran, "A Modified LeNet CNN for Breast Cancer Diagnosis in Ultrasound Images," *Diagnostics*, vol. 13, no. 17, p. 2746, 2023, https://doi.org/10.3390/diagnostics13172746.

[51] S. G. Lee, Y. Sung, Y. G. Kim, and E. Y. Cha, "Variations of AlexNet and GoogLeNet to improve Korean character recognition performance," *Journal of Information Processing Systems*, vol. 14, no. 1, pp. 205-217, 2018, https://doi.org/10.3745/JIPS.04.0061.

[52] N. N. A. A. Hamid, R. A. Razali, and Z. Ibrahim, "Comparing bags of features, conventional convolutional neural network and alexnet for fruit recognition," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 333-339, 2019, http://doi.org/10.11591/ijeecs.v14.i1.pp333-339.

[53] K. Prastika and Lina, "Application of individual activity recognition in the room using CNN Alexnet method," *IOP Conference Series: Materials Science and Engineering*, 2020, https://doi.org/10.1088/1757-899X/1007/1/012162.

[54] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021, https://doi.org/10.1016/j.ymssp.2020.107398.

[55] S. Lu, S. H. Wang, and Y. D. Zhang, "Detection of abnormal brain in MRI via improved AlexNet and ELM optimized by chaotic bat algorithm," *Neural Computing and Applications*, vol. 33, pp. 10799-10811, 2021, https://doi.org/10.1007/s00521-020-05082-4.

[56] E. Mohan *et al.*, "Thyroid Detection and Classification Using DNN Based on Hybrid Meta-Heuristic and LSTM Technique," *IEEE Access*, vol. 11, pp. 68127-68138, 2023, https://doi.org/10.1109/ACCESS.2023.3289511.

[57] N. N. Jafery, S. N. Sulaiman, M. K. Osman, N. K. A. Karim, Z. H. C. Soh, and N. A. M. Isa, "Comparative Analysis of Hybrid 1D-CNN-LSTM and VGG16-1D-LSTM for Lung Lesion Classification," *Journal of Electrical Engineering and Technology*, vol. 20, pp. 2617-2630, 2025, https://doi.org/10.1007/s42835-025-02182-w.

[58] M. R. Ahmed, S. Islam, A. K. M. M. Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," *Expert Systems with Applications*, vol. 218, p. 119633, 2023, https://doi.org/10.1016/j.eswa.2023.119633.

[59] S. E. Mansour and A. Sakhi, "Detection of Sealing Defects in Canned Sardines Using Local Binary Pattern and Perceptron Techniques for Enhanced Quality Control," *International Journal of Robotics and Control Systems*, vol. 5, no. 1, pp. 585-598, 2025, https://doi.org/10.31763/ijrcs.v5i1.1737.

[60] D. C. E. Saputra, E. I. Muryadi, I. Futri, T. A. Win, K. Sunat, and T. Ratnaningsih, "Revolutionizing Anemia Classification with Multilayer Extremely Randomized Tree Learning Machine for Unprecedented Accuracy," *International Journal of Robotics and Control Systems*, vol. 4, no. 2, pp. 758-778, 2024, https://doi.org/10.31763/ijrcs.v4i2.1379.