A Deep Learning-Based Framework Using CNN+LSTM for Karate Kata Classification and Correctness Evaluation

Nur Abdulrahman
Department of Informatics Engineering, Faculty of Engineering, Hasanuddin University, Indonesia
abdulrahmann22d@student.unhas.ac.id

Zahir Zainuddin
Department of Informatics Engineering, Faculty of Engineering, Hasanuddin University, Indonesia
zahir@unhas.ac.id (corresponding author)

Ingrid Nurtanio
Department of Informatics Engineering, Faculty of Engineering, Hasanuddin University, Indonesia
ingrid.nurtanio@gmail.com

Received: 15 May 2025 | Revised: 5 August 2025, 12 October 2025, and 25 October 2025 | Accepted: 29 October 2025

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.12147*

ABSTRACT

This paper presents a deep learning-based framework for automated classification and correctness evaluation of karate kata sequences using human pose data. The method utilizes MediaPipe BlazePose to extract 3D body keypoints from video frames, which are subsequently transformed into 132-dimensional vectors and temporally normalized into fixed-length sequences. Two neural architectures are evaluated: a baseline Convolutional Neural Network (CNN) and a hybrid CNN combined with Long Short-Term Memory (CNN+LSTM). Both models perform dual-output predictions: multi-class kata identification and binary correctness classification. Experimental results demonstrate that the CNN+LSTM model performs better in classification accuracy and movement assessment, with up to 95.74% accuracy and F1-scores exceeding 95% for several kata classes on unseen data. The findings highlight the importance of temporal modeling for structured movement analysis and establish a foundation for intelligent martial arts evaluation systems.

Keywords-human pose estimation; karate kata; CNN; LSTM; multitask learning

I. INTRODUCTION

Human movement analysis is a crucial area in computer vision, having important applications in sports performance assessment, rehabilitation, surveillance, and human-computer interaction [1]. A fundamental challenge in this field is Human Pose Estimation (HPE), which entails estimating the spatial configuration of key locations on the human body using visual input. The emergence of deep learning methods, especially Convolutional Neural Networks (CNNs), has resulted in significant improvements in the accuracy of both 2D and 3D posture estimation, surpassing previous handmade and statistical approaches [2-6].

Despite notable progress, most pose estimation systems focus on generic and repetitive movements such as walking, running, or sitting [7-9]. In contrast, organized and domain-specific sequences, exemplified by traditional martial arts, have garnered comparatively less study. Karate is one of the most widely practiced martial arts worldwide [10], characterized by organized movement sequences, called kata, that require exact execution and sequencing. This highlights its significance for computational modeling and assessment.

Current pose estimation pipelines operate primarily on static images or isolated frames, which restricts the ability to capture temporal dynamics across sequential movements [11]. This limitation poses a significant barrier when addressing tasks that require a contextual understanding of transitions between poses, such as kata routines [2]. Additionally, most action recognition models emphasize the identification of action categories, without assessing the correctness or fidelity of execution. This capability is essential in expert-level training and feedback applications [12].

BlazePose was adopted as the foundational pose estimator to support high-fidelity keypoint extraction under real-time constraints. Designed by Google Research, BlazePose is a lightweight and mobile-optimized model capable of predicting 33 body landmarks with high spatial accuracy at over 30

frames per second on standard smartphones [13, 14]. The model architecture incorporates a two-stage approach: a Single Shot Detector (SSD) for Region Of Interest (ROI) detection, followed by a regression network that estimates keypoint coordinates via an encoder-decoder architecture. The output from the pose estimation process consists of:

$$P = \{(x_i, y_i, z_i, c_i)\}_{i=1}^{33} \quad (1)$$

where x_i, y_i, z_i present normalized spatial coordinates, and c_i indicates the confidence score for each keypoint. The overall loss function integrates Smooth L1 loss for regression accuracy and Binary Cross-Entropy for $visibility$ confidence:

$$\mathcal{L}_{pose} = \sum_{i=1}^{33}[SmoothL1((x_i, y_i, z_i) - (x_i^{gt}, y_i^{gt}, z_i^{gt})) + \lambda \cdot BCE\ (c_i, c_i^{gt})] \quad (2)$$

The BlazePose architecture utilizes previous posture estimations with ROI tracking to provide resilience against occlusion, fluctuating lighting conditions, and significant body movement. To address the stated constraints, a methodological framework is developed that employs posture sequences obtained by BlazePose, which are then analyzed using both CNN and CNN+LSTM architectures, as seen in Figure 1. The aim is to transcend traditional activity classification to include the measurement of movement accuracy inside kata sequences, therefore fulfilling the need for specialized assessment tools in martial arts performance evaluation.

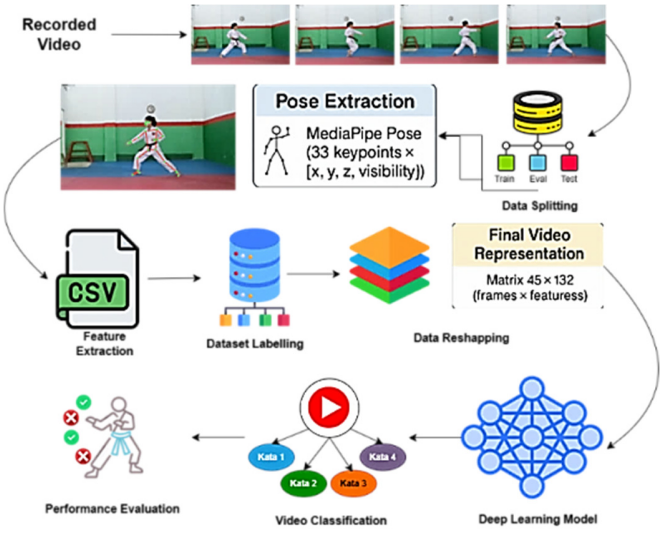

Fig. 1. Overview of the proposed classification pipeline.

HPE and action identification have become more interconnected, especially as research transitions to capturing intricate spatiotemporal dynamics from video data. CNN-based architectures have effectively classified static postures by extracting spatial information from keypoint representations [15-17]. Nevertheless, CNNs often struggle to capture the temporal correlations essential for studying motion-intensive activities, such as martial arts routines [18].

Few studies have examined the categorization of kata sequences using deep learning. In [19], transfer learning with VGG16 was used to categorize karate postures, achieving reasonable accuracy despite dataset constraints. In [12], computer vision techniques were used to categorize karate stances in both fighting and kata contexts; however, this system was deficient in temporal modeling skills. Although these efforts provided basic contributions, they focused on discrete pose categorization instead of assessing movement accuracy or sequence dynamics. To address this, hybrid models integrating CNN with LSTM networks have been suggested for action recognition, including the categorization of temporally sequenced stances. The CNN module pulls spatial information from posture data, while the LSTM captures the temporal correlations among frames. This method has been used effectively in several action recognition tasks, including the categorization of gesture and dance movements [18, 20]. Such architectures have shown promising accuracy levels while maintaining computational efficiency compared to full 3D-CNN models [21].

The progression of real-time pose estimators, such as BlazePose, has facilitated the creation of low-latency and mobile-compatible action detection systems. BlazePose utilizes a detector-tracker method that integrates the initial Region-Of-Interest (ROI) identification with a regression network to predict 33 anatomical keypoints. BlazePose has been shown to surpass OpenPose in some applications, such as fitness monitoring, while operating up to 75 times faster on a mobile device [13]. The BlazePose architecture employs a combination of heatmap and regression branches during training, but omits the heatmap during inference for enhanced performance. Its pivotal output architecture renders it very compatible with deep learning frameworks that depend on pose sequence analysis.

The assessment of movement quality or accuracy, beyond mere classication, is a rather neglected domain. Traditional action recognition primarily focuses on action labeling; however, newer studies have included models to evaluate movement precision, especially in rehabilitation or sports training contexts [12]. However, such systems often depend on domain-specific rules or handcrafted features, limiting generalizability.

Recent advances in Transformer-based pose models and Graph Convolutional Networks (GCNs) have created new opportunities for action recognition from skeletal data. Despite their performance improvements, these designs are computationally intensive and less appropriate for real-time feedback systems often used in sports environments [22].

Temporal consistency modeling is a crucial element, especially in video-based posture estimation. In [2], a multi-scale Temporal Consistency Exploration (TCE) module was introduced, which augmented video-based posture estimation without the need for optical flow, thus enhancing resilience against occlusions and motion blur. These temporal refinement approaches highlight the need for models that locate joints coherently throughout time.

Although several studies have investigated action classification using pose-based representations, only a few have concurrently integrated temporal modeling with movement

accuracy assessment, especially in organized activity domains such as karate kata. Most previous studies focused on broad action labeling, without tools to evaluate whether actions conform to technical norms or are conducted accurately. Furthermore, little research has focused on systems that can perform multi-class categorization across several kata types within a cohesive framework. The proposed solution integrates BlazePose, a lightweight real-time pose estimation model, with a CNN+LSTM architecture to effectively capture spatial configurations and temporal relationships. This enables the identification of various kata classes and the measurement of execution accuracy, providing a comprehensive solution for automated assessment in martial arts training environments.

II. METHODOLOGY

A. Data and Pose Extraction

A dataset was created specifically for this study to ensure high relevance and quality. It features recordings from seven experienced karate practitioners at Dojo Inkado Minasa Upa, Makassar. All sessions were conducted in a controlled indoor environment to minimize background noise and motion artifacts. A standard DSLR (Digital Single-Lens Reflex) camera was used to capture the performances, ensuring consistent high-quality video data for analysis. The final curated dataset consists of 100 video clips covering four distinct kata classes (*Kata 1*, *Kata 2*, *Kata 3*, *Kata 4*). Crucially, each performance was professionally annotated as "correct" or "incorrect" by a certified karate instructor. This labeling was based on established martial arts guidelines for posture, timing, and technical precision, providing a reliable ground truth for the model's correctness evaluation task.

For the machine learning pipeline, the dataset was methodologically divided into training, evaluation, and testing subsets using a fixed 3:1:3 ratio. The training set was used to optimize the model's weights, while the evaluation set guided hyperparameter tuning and early stopping protocols to prevent overfitting. The test set was held out entirely during training, serving as an unbiased sample to assess the final model's generalization performance on completely unseen data.

Frame-by-frame pose estimation was performed using MediaPipe's BlazePose technology, which offers real-time 3D human posture tracking. The model forecasted 33 anatomical keypoints for every frame; each keypoint was linked to four characteristics: x, y, and z coordinates in a normalized image space, and a $visibility$ score indicating the detection confidence. These produced a 132-dimensional posture description per frame via concatenation:

$$F_t = [x_1, y_1, z_1, v_1, \dots, x_{33}, y_{33}, z_{33}, v_{33}] \in \mathbb{R}^{132} \quad (3)$$

All pose descriptors were exported to a structured tabular format (.CSV) and subsequently aligned with their corresponding temporal indices for sequence-level modeling. This preprocessing stage resulted in a set of temporally ordered pose sequences suitable for downstream spatial-temporal representation and classification.

B. Temporal Normalization and Video Representation

Individual execution speed and artistic diversity cause karate kata performances to naturally vary in length. A temporal normalization technique was used, in which every video sequence was resampled to a set length of 45 frames, guaranteeing a consistent temporal input across all samples. Linear interpolation was used to preserve the sequential order and relative dynamics of the original motion while obtaining consistent sampling throughout time. This normalization produced consistent model input size and lower variation from non-structural temporal aberrations.

Following resampling, each sequence was represented as a matrix $V \in \mathbb{R}^{45\,x\,132}$, where each row corresponds to a 132-dimensional feature vector extracted from one frame, and each column represents a spatial or *visibility* component of a keypoint. This matrix encapsulates spatial information, encoded in joint positions and motion patterns, reflected through their progression over time. Thus, the resultant representation retains enough resolution to enable temporal sequence modeling and static posture analysis. Using convolutional and recurrent layers within deep neural networks, this architecture allows later processing and helps the system to acquire hierarchical representations that record posture configurations and their temporal development. Notably, the normalizing process reduces inter-sample fluctuation in frame count, guaranteeing compatibility with batch-based training and lowering model optimization computational load.

C. Model Architecture

Two neural network architectures were implemented to classify karate kata sequences: a CNN-only model and a CNN combined with an LSTM (CNN+LSTM) network. Both models were designed to perform multi-task learning, simultaneously predicting the class of kata (multi-class classification) and assessing movement correctness (binary classification). Each input sample consisted of a temporally normalized matrix of shape 45×132, representing a sequence of 45 frames, each with 132 features corresponding to the 33 key points of the body (x, y, z, $visibility$) extracted via BlazePose. This matrix was processed as a temporal signal across time steps, with feature extraction performed via 1D convolutions.

The CNN-only model employed a sequence of 1D convolutional layers followed by pooling and regularization. The transformation applied by each convolutional layer is defined as:

$$Z_1 = ReLU\,(X * W_1 + b_1),\ Z_1' = MaxPool\,(Z_1)$$

$$Z_2 = ReLU\,(X * W_2 + b_2),\ Z_2' = MaxPool\,(Z_2) \quad (4)$$

where $*$ denotes 1D convolution, W_i and b_i are learnable weights and biases, and ReLU is the rectified linear activation function. After the second pooling layer, the resulting tensor was flattened into a vector $h_{CNN} \in \mathbb{R}^n$, serving as the input to two parallel classification heads.

The CNN+LSTM architecture followed an identical convolutional structure, but instead of flattening the output, the final feature maps were processed by a unidirectional LSTM

layer. The LSTM models temporal dependencies across the frame sequence using the recurrence relation:

$$h_t, c_t = LSTM\ (z_t, h_{t-1}, c_{t-1}) \quad (5)$$

where z_t denotes the input at time step t, and h_t and c_t are the hidden and cell states, respectively. The final hidden state h_T is used as the sequence-level embedding passed to the classification layers.

Both architectures shared the same output structure, consisting of two fully connected layers. A softmax layer for *kata* classification:

$$\hat{y}_{kata} = softmax\ (W_k * h + b_k) \quad (6)$$

and a sigmoid layer for movement correctness classification:

$$\hat{y}_{correct} = \sigma\ (W_c * h + b_c) \quad (7)$$

where W_k and W_c are weight matrices and $\sigma(\cdot)$ is the sigmoid activation function.

The model was trained using a composite loss function, combining categorical cross-entropy loss for multiclass classification:

$$\mathcal{L}_{kata} = -\sum_{i=1}^{c} y_i \log(\hat{y}_i) \quad (8)$$

and binary cross-entropy loss for correctness classification:

$$\mathcal{L}_{\text{correct}} = [\, y\ \log \hat{y}) + (1 - y) \log(1 - \hat{y})] \quad (9)$$

This dual-task formulation allowed shared learning of representations while supporting distinct objectives, improving generalization through multi-task inductive bias.

D. Training Procedure

For multi-output classification, both architectures were linked to two parallel dense layers with softmax activations: (i) kata type (multi-class) and (ii) movement correctness (binary). Thanks to this dual-head output, the models could concurrently deduce execution quality and action identity.

Model training was carried out using a supervised learning system with multi-output goals. The input training data consisted of resampled posture sequences along with two related labels for each instance: a category label denoting the kata class and a binary label indicating the accuracy of the movement. Default settings were used to train using the Adam optimizer. The model was constructed using binary cross-entropy for the correctness assessment output and categorical cross-entropy as the loss function for the kata classification output. The total loss function was the unweighted sum of both goals, therefore supporting simultaneous optimization of action type prediction and quality evaluation.

To avoid overfitting and stabilize training, early stopping was used with a patience of 15 epochs. Performance was tracked on a different validation set throughout training. Training stopped after seeing no change in validation loss over 15 straight epochs, and the best-performing model weights were automatically restored. Training spanned 150 epochs with a batch size of 16. Regarding training and validation accuracy and loss, model performance was monitored separately for each output branch: kata classification and correctness prediction.

III. RESULTS AND DISCUSSION

To evaluate the effectiveness of both model architectures—CNN-only and CNN+LSTM—a comparative experiment was conducted using identical training, validation, and test sets. Both models were trained using the same configuration (150 epochs, batch size = 16, early stopping with patience = 15), and evaluated on their ability to (i) classify kata types and (ii) determine the correctness of execution.

Figure 2 illustrates the training history of the CNN+LSTM model. The accuracy graph demonstrates that the model swiftly achieved excellent performance in kata classification, with validation accuracy stable between 90% and 95% after 30 epochs. In contrast, correctness classification exhibited increased variability but consistently improved, with validation accuracy of 80% in the last epochs. The associated loss curves validate this tendency, with kata classification loss diminishing swiftly and correctness loss falling more gradually.

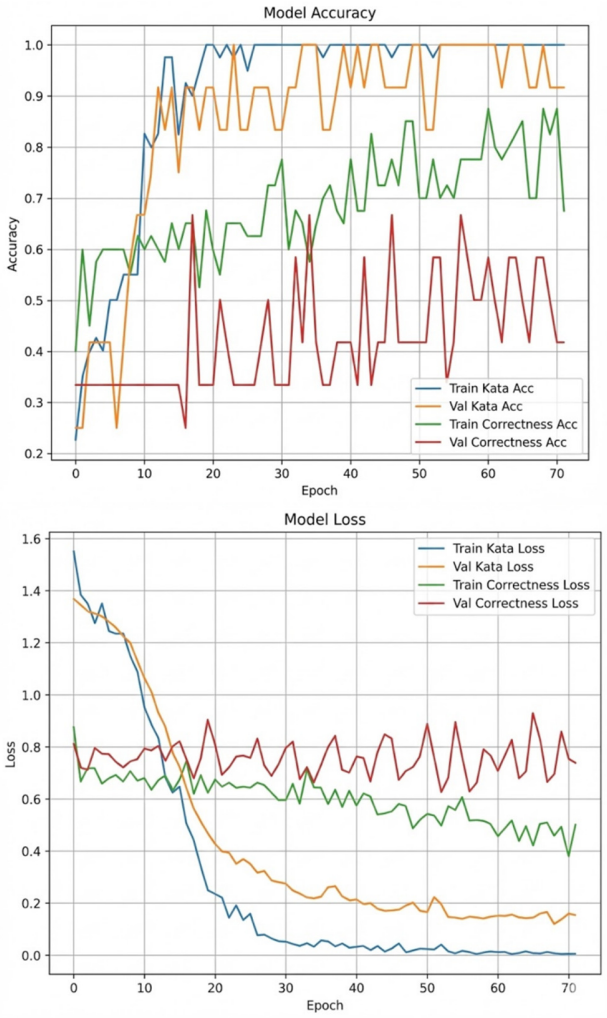

Fig. 2. Training history (accuracy and loss) for the CNN+LSTM model.

Figure 3 depicts the performance of the CNN-only model. Kata classification accuracy approached values similar to CNN+LSTM, peaking at over 100% on the training set and maintaining at 90–92% on the validation set; however, correctness classification showed less stability, with chaotic fluctuations and plateauing around 70% validation accuracy. This indicates that the CNN-only design would be less adept at capturing the sequential subtleties essential for assessing motion accuracy.

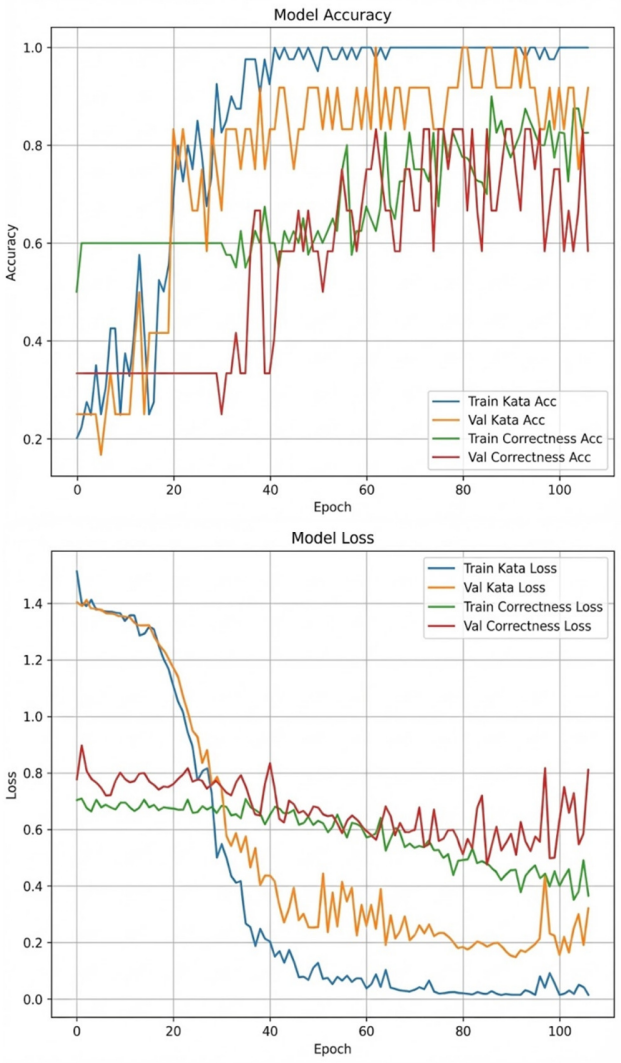

Fig. 3. Training history (accuracy and loss) for the CNN-only model

Both models included specialized label encoders for kata classification, which converted string labels (e.g., "kata1_positive", "kata2_negative") into a category format. The encoders were kept individually as .pkl files to maintain consistency during inference. The accuracy label was binary-encoded (0 for erroneous, 1 for correct) and supervised with a sigmoid activation output. The encoders facilitate the deployment of trained models in practical applications, enabling the conversion of output logits into comprehensible predictions.

The experimental findings underscore the advantages of incorporating temporal modeling through LSTM layers. The CNN+LSTM model exhibited enhanced generalization in accuracy classification, indicating that recurrent connections facilitated the capture of movement continuity and rhythm. Precise timing and synchronization between frames are crucial in karate kata, distinguishing perfect from wrong performance. In contrast, while the CNN-only model demonstrated efficiency and competitive outcomes in kata categorization, it exhibited a deficiency in robustness regarding correctness evaluation. This is probably because of its restricted receptive field and lack of temporal memory, which limits its ability to comprehend nuanced mistakes that extend across frames. In addition, the gradual reduction in loss seen in the CNN+LSTM model corroborates the assertion that sequence-level temporal characteristics are crucial for tasks requiring motion quality assessment. The enhanced stability and convergence trends in the validation measures further validate the suitability of recurrent modeling for structured movement analysis.

Both models were assessed on novel and unlabeled test video sequences to evaluate their generalization ability in realistic contexts. Table I provides a comparative analysis of model performance on the test set, emphasizing total accuracy and per-class F1 scores for each of the four kata categories.

TABLE I. COMPARISON OF MODEL PERFORMANCE ON UNLABELED TEST VIDEOS

Metric	CNN	CNN+LSTM
Accuracy	91.49%	95.74%
F1 for Kata 1	85.71%	95.65%
F1 for Kata 2	88.89%	100%
F1 for Kata 3	95.24%	90.91%
F1 for Kata 4	96.00%	96.00%

These results show that the CNN+LSTM model consistently outperformed the CNN-only approach, particularly regarding overall accuracy and individual class discrimination (F1 scores). This further reinforces the advantage of incorporating temporal modeling through LSTM layers, allowing the system to handle nuanced variations in karate kata sequences better. The improved generalizability to unseen data suggests promising potential for real-world deployment in martial arts training and assessment contexts.

To contextualize the proposed model's performance within the existing literature, Table II presents a comparison between the proposed CNN+LSTM model and the results reported in recent similar studies. It is important to note that this is an indirect comparison, as these methods were evaluated on their own distinct datasets under different experimental conditions. The comparison serves as a general overview of the state-of-the-art performance in related tasks.

TABLE II. COMPARATIVE PERFORMANCE OF PROPOSED CNN+LSTM AND RECENT SIMILAR METHODS

Method	Accuracy (%)
Proposed CNN+LSTM	95.74
CNN+LSTM with VGG16 [19]	68.00
CNN+GRU with PoseNet [20]	94.40
3D-CNN with OpenPose [21]	74.67

IV. CONCLUSION

This study introduced a deep learning framework that uses MediaPipe-BlazePose combined with a CNN+LSTM architecture to classify karate kata sequences and assess execution accuracy. Experimental results demonstrated that the proposed CNN+LSTM model consistently outperformed the CNN-only model, particularly in capturing nuanced temporal dynamics crucial to evaluating structured martial arts movements. The primary novelty of this work lies in effectively integrating temporal modeling (via LSTM) with spatial pose estimation (via BlazePose), providing robust multi-task learning capabilities (kata identification and correctness evaluation simultaneously). This significantly improves accuracy and generalization compared to recent methods. Such an integration represents a substantial contribution towards intelligent, real-time martial arts training and assessment systems. Performance on unlabeled test data validated the resilience and generalization capabilities of the CNN+LSTM model, with F1 scores greater than 95% in many categories.

Future research directions may include:

- Incorporating attention mechanisms to further improve the interpretability and focus of temporal modeling.
- Expanding the dataset with broader kata variations, execution speeds, and athlete profiles.
- Integrating real-time feedback systems for live performance assessment.
- Exploring transformer-based architectures to enhance long-range temporal dependency modeling.

DATA AVAILABILITY STATEMENT

The dataset generated and analyzed during this current study is available from the corresponding author upon reasonable request.

REFERENCES

[1] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. T. Salo, "A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System," *Sports Medicine - Open*, vol. 4, no. 1, Dec. 2018, Art. no. 24, https://doi.org/10.1186/s40798-018-0139-y.

[2] Y. Li, K. Li, X. Wang, and R. Y. D. Xu, "Exploring temporal consistency for human pose estimation in videos," *Pattern Recognition*, vol. 103, Jul. 2020, Art. no. 107258, https://doi.org/10.1016/j.patcog.2020.107258.

[3] K. Armstrong, A. Rodrigues, A. P. Willmott, L. Zhang, and X. Ye, "Validation of Human Pose Estimation and Human Mesh Recovery for Extracting Clinically Relevant Motion Data from Videos." arXiv, 2025, https://doi.org/10.48550/ARXIV.2503.14760.

[4] R. Sun, Z. Lin, S. Leng, A. Wang, and L. Zhao, "An In-Depth Analysis of 2D and 3D Pose Estimation Techniques in Deep Learning: Methodologies and Advances," *Electronics*, vol. 14, no. 7, Mar. 2025, Art. no. 1307, https://doi.org/10.3390/electronics14071307.

[5] S. Park, J. Hwang, and N. Kwak, "3D Human Pose Estimation Using Convolutional Neural Networks with 2D Pose Information," in *Computer Vision – ECCV 2016 Workshops*, vol. 9915, G. Hua and H. Jégou, Eds. Springer International Publishing, 2016, pp. 156–169.

[6] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 5137–5146, https://doi.org/10.1109/CVPR.2018.00539.

[7] J. Stenum, K. M. Cherry-Allen, C. O. Pyles, R. D. Reetzke, M. F. Vignos, and R. T. Roemmich, "Applications of Pose Estimation in Human Health and Performance across the Lifespan," *Sensors*, vol. 21, no. 21, Nov. 2021, Art. no. 7315, https://doi.org/10.3390/s21217315.

[8] H. L. Cornman, J. Stenum, and R. T. Roemmich, "Video-based quantification of human movement frequency using pose estimation: A pilot study," *PLOS ONE*, vol. 16, no. 12, Dec. 2021, Art. no. e0261450, https://doi.org/10.1371/journal.pone.0261450.

[9] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 538–552, Sep. 2012, https://doi.org/10.1109/JSTSP.2012.2196975.

[10] "Top 50 Most Popular Martial Arts in the World 2025," *Country Cassette*, May 21, 2024. https://countrycassette.com/most-popular-martial-arts-in-the-world/.

[11] Y. W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng, "Forecasting Human Dynamics from Static Images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3643–3651, https://doi.org/10.1109/CVPR.2017.388.

[12] J. Echeverria and O. C. Santos, "Toward Modeling Psychomotor Performance in Karate Combats Using Computer Vision Pose Estimation," *Sensors*, vol. 21, no. 24, Dec. 2021, Art. no. 8378, https://doi.org/10.3390/s21248378.

[13] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking." arXiv, 2020, https://doi.org/10.48550/ARXIV.2006.10204.

[14] S. Zhang, C. Wang, W. Dong, and B. Fan, "A Survey on Depth Ambiguity of 3D Human Pose Estimation," *Applied Sciences*, vol. 12, no. 20, Oct. 2022, Art. no. 10591, https://doi.org/10.3390/app122010591.

[15] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Pose-Based CNN Features for Action Recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3218–3226, https://doi.org/10.1109/ICCV.2015.368.

[16] M. H. D. L. Cruz, U. Solache, A. Luna-Álvarez, S. R. Zagal-Barrera, D. A. Morales López, and D. Mujica-Vargas, "CNN 1D: A Robust Model for Human Pose Estimation," *Information*, vol. 16, no. 2, Feb. 2025, Art. no. 129, https://doi.org/10.3390/info16020129.

[17] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes." arXiv, 2017, https://doi.org/10.48550/ARXIV.1711.00199.

[18] G. Ercolano and S. Rossi, "Combining CNN and LSTM for activity of daily living recognition with a 3D matrix skeleton representation," *Intelligent Service Robotics*, vol. 14, no. 2, pp. 175–185, Apr. 2021, https://doi.org/10.1007/s11370-021-00358-7.

[19] B. G. Priya and D. M. Arulselvi, "Action Classification for Karate Dataset Using Deep Learning," in *International Conference INTELINC*, Oct. 2018, pp. 212–216.

[20] N. U. R. Malik, S. A. R. Abu-Bakar, U. U. Sheikh, A. Channa, and N. Popescu, "Cascading Pose Features with CNN-LSTM for Multiview Human Action Recognition," *Signals*, vol. 4, no. 1, pp. 40–55, Jan. 2023, https://doi.org/10.3390/signals4010002.

[21] S. M. Saeed, H. Akbar, T. Nawaz, H. Elahi, and U. S. Khan, "Body-Pose-Guided Action Recognition with Convolutional Long Short-Term Memory (LSTM) in Aerial Videos," *Applied Sciences*, vol. 13, no. 16, Aug. 2023, Art. no. 9384, https://doi.org/10.3390/app13169384.

[22] G. Lan, Y. Wu, F. Hu, and Q. Hao, "Vision-Based Human Pose Estimation via Deep Learning: A Survey," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 1, pp. 253–268, Feb. 2023, https://doi.org/10.1109/THMS.2022.3219242.