Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.058193

ARTICLE

Improving Badminton Action Recognition Using Spatio-Temporal Analysis and a Weighted Ensemble Learning Model

Farida Asriani[1,2], Azhari Azhari[1,*] and Wahyono Wahyono[1]

[1]Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, 55281, Indonesia

[2]Electrical Engineering Department, Universitas Jenderal Soedirman, Purbalingga, 53371, Indonesia

*Corresponding Author: Azhari Azhari. Email: arisn@ugm.ac.id

Received: 06 September 2024 Accepted: 18 October 2024 Published: 18 November 2024

ABSTRACT

Incredible progress has been made in human action recognition (HAR), significantly impacting computer vision applications in sports analytics. However, identifying dynamic and complex movements in sports like badminton remains challenging due to the need for precise recognition accuracy and better management of complex motion patterns. Deep learning techniques like convolutional neural networks (CNNs), long short-term memory (LSTM), and graph convolutional networks (GCNs) improve recognition in large datasets, while the traditional machine learning methods like SVM (support vector machines), RF (random forest), and LR (logistic regression), combined with handcrafted features and ensemble approaches, perform well but struggle with the complexity of fast-paced sports like badminton. We proposed an ensemble learning model combining support vector machines (SVM), logistic regression (LR), random forest (RF), and adaptive boosting (AdaBoost) for badminton action recognition. The data in this study consist of video recordings of badminton stroke techniques, which have been extracted into spatiotemporal data. The three-dimensional distance between each skeleton point and the right hip represents the spatial features. The temporal features are the results of Fast Dynamic Time Warping (FDTW) calculations applied to 15 frames of each video sequence. The weighted ensemble model employs soft voting classifiers from SVM, LR, RF, and AdaBoost to enhance the accuracy of badminton action recognition. The E2 ensemble model, which combines SVM, LR, and AdaBoost, achieves the highest accuracy of 95.38%.

KEYWORDS

Weighted ensemble learning; badminton action; soft voting classifier; joint skeleton; fast dynamic time warping; spatiotemporal

1 Introduction

Incredible progress has been made in detecting human action recognition (HAR), which significantly impacts computer vision applications in sports analytics. Despite these developments, it is still challenging to identify dynamic and complex movements like those in badminton because precise recognition accuracy and improved management of complex motion patterns are required [1–3].

Spatiotemporal dynamics in video and skeleton data generated with deep learning techniques like convolutional neural networks (CNNs), long short-term memory (LSTM), and graph convolutional networks (GCNs) are highly effective [1,3,4]. These techniques improve recognition performance in large datasets by automatically learning hierarchical features. However, they can be resource-intensive and might not translate well to smaller, more specialized datasets, like badminton action recognition datasets.

On the other hand, when paired with handcrafted feature extraction, conventional machine learning techniques such as support vector machines (SVM), random forests (RF), and logistic regression (LR) show good performance [5–7]. Ensemble approaches improve accuracy even more by utilizing the advantages of each model [8,9]. Even with their efficacy, these models cannot fully represent the intricacy of human movement, especially in high-speed sports like badminton.

Feature extraction methods based on deep learning and handcrafted techniques have been used in recent studies. While handcrafted methods like motion trajectories and histograms of oriented gradients (HOG) are practical and straightforward, they frequently lack the sophistication required to capture the subtleties of intricate actions. Deep learning techniques utilizing 3D skeleton data provide more comprehensive spatiotemporal representations and have been extensively demonstrated in previous research [1–3]. Tasnim et al. [10] employed MobileNetV2, DenseNet121, and ResNet18 models in conjunction with transfer learning and fusion techniques, resulting in high accuracy on benchmark dataset.

Action recognition could be significantly improved by integrating 3D skeleton data, which records joint movements' temporal and spatial progression [11,12]. While skeleton-based methods have been used for human action recognition, deep learning implementations customized to badminton are still in their infancy and could be improved [13]. Furthermore, although 3D skeleton data is robust against common problems such as lighting and camera angles, it is still unclear how to extract meaningful features from this data to create scalable and resilient models.

Combining complementary color and texture features using the Choquet fuzzy integral to model complex, multi-modal distributions is one of the critical multi-view learning strategies, particularly in complex environments. Atanassov's intuitive 3D fuzzy Histon roughness index and other methods for encoding spatial information capture intricate spatial relationships, while wavelet transform image scaling improves the differentiation of spatial dynamics. Better feature extraction is achieved through content-aware feature weighting, which determines significance based on context. Additionally, the multi-view Bag of Words (BoW) model enhances scene categorization by integrating spatial and semantic information from multiple perspectives. Lastly, combining color and spatial features solves insufficient shape-based recognition [14,15].

This paper proposes a weighted ensemble learning technique and a spatiotemporal analysis of 3D skeletal data to address these issues. We suggest applying Fast Dynamic Time Warping (FDTW) for the temporal feature and using the right hip's 3D Euclidean distance as the spatial feature. The contributions of this paper are:

1. A new 3D spatiotemporal feature extraction method, in which the temporal feature is extracted using FDTW, and the spatial feature is the separation between the right hip and other skeleton coordinates.
2. A weighted ensemble learning model for badminton action recognition that combines LR, SVM, and AdaBoost.
3. A performance comparison indicates that the proposed weighted ensemble model surpasses deep learning methods, such as CNN and LSTM, on small datasets like badminton action

detection, delivering competitive accuracy with reduced computational expense for real-time applications.

2 Related Works

2.1 Three-Dimensional (3D) Human Skeleton Detection

The real-time posture estimation system, based on a deep neural network, has been researched extensively. For example, the system used convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to determine a person's posture from a single RGB camera. The system exhibited a high accuracy in estimating an individual human's three-dimensional stance and utilizing a publicly accessible dataset [8,9,16–19]. Li et al. [20] proposed a posture estimation system using a similar methodology based on an RGB-D camera. The convolutional neural network estimated the human joint positions using a depth map and the pose. The system could calculate the pose accurately and in real-time on a publicly available dataset. A novel posture estimation system was created, employing convolutional neural networks and recurrent neural network techniques. The technology was designed to precisely forecast the human body's position in real time with a sliding window mechanism. The suggested method successfully executed real-time posture estimation and achieved remarkable accuracy on a publicly available dataset [8,16,20–22]. The study uses media pipes and convolutional neural networks to measure the human body's position in three dimensions. This multi-step method accurately assesses human posture in publicly available datasets [16,20,22–24]. GCN can be used to effectively model human joints and improve performance in pose estimation and action recognition tasks. Using GCN with spatiotemporal information enhances accuracy in complex scenes [25]. Action recognition improvement is also achieved by combining behavioral dependencies and contextual cues. This method allows the model to distinguish between similar poses [26].

2.2 Spatiotemporal Analysis in Action Recognition

Researchers have analyzed human actions using a spatiotemporal feature analysis method based on deep learning. In action recognition, temporal sequence recordings are essential. Shahroudy et al. [27] established the NTU RGB+D dataset, which has now become one of the largest and most diversified datasets for 3D action recognition. People extensively use it to evaluate the effectiveness of spatiotemporal models. Liu et al. [28] used global context-aware attention LSTM networks on 3D frame data, which helped the model focus on important joints and time segments. Graph Convolutional Networks (GCNs) have demonstrated efficacy as a powerful instrument for spatiotemporal modeling. Spatial-Temporal Graph Convolutional Networks (ST-GCN) serve as a technique for skeleton-based action recognition. The spatial interaction between joints and the temporal dynamics along the sequence are wrapped in this network [29]. The adaptive two-stream graph convolutional network (2s-AGCN) was proposed by Shi et al. [30] to support this idea of two-stream adaptive graph convolutional networks (2s-AGCN). These networks enhance accuracy by adaptively learning the graph topology from the supplied input. Moreover, Temporal Convolutional Networks (TCNs) exhibit considerable potential in this domain. Long Short-Term Memory (LSTM) models aren't as good at sequence modeling as Temporal Convolutional Networks (TCN) models because they can't capture long-range relationships as well as TCN models [31].

Besides deep learning methods, various methodologies have contributed to spatiotemporal analysis. Action posture is encoded using a bag of 3D points from depth data, and action graph dynamics are modeled for each action. Koniusz et al. [32] portrayed that space-time action sequences can efficiently be captured in a tensor space. Such methods have already been used to recognize badminton actions

in RGB videos [33,34]. Feature extraction methods that have been applied are bounding box and histogram of oriented gradient (HOG). However, these methods produce misclassification for some classes involving the same pose during stroke [33]. The sliding window and Haar-like methods have also been applied. However, the recognition rate for different badminton players is not ideal [34]. In addition, detection of shot badminton players converts video data into sequential frames, applying sliding windows for feature extraction to detect shot timing and position with an accuracy of 95.9%. Similarly, image recognition of badminton swing motion based on a single inertial sensor transforms motion data into sequences, using sliding and action windows for accurate segmentation. This enables the Deep Residual LSTM model to recognize six swing types with over 90% accuracy, automating the extraction and recognition process for performance analysis in badminton [35,36]. Atanassov's intuitive 3D fuzzy Histon roughness index method has been discovered to detect moving objects with colored data. (IA-IA3DFHRI) [14]. The proposed method demonstrates resilience to dynamic backgrounds.

2.3 Ensemble Learning in Action Recognition

Ensemble learning has been applied to several other fields, as pointed out by prior investigations. For example, Karim et al. [37] used ensemble learning combining logistic regression, decision trees, and SVM in healthcare. Regarding handwritten recognition, Karray et al. [38] proved improved accuracy by ensemble deep learning, RF, and SVM models. The accuracy of HAR can be enhanced by integrating SVM and HMM [39]. Combining SVM, KNN, and LR with a soft voting classifier can improve accuracy to 92.78% [40]. Das et al. [41] combined ensemble models (DT, RF, SVM, and KNN) with soft voting techniques and weighted optimization using FOX optimization. Research in human action recognition has also integrated deep learning models. The suggested ensemble learning algorithm includes DNN and CNN stacked at the gated recurrent unit (GRU) to recognize human actions [42]. Kaur et al. [43] combined ResNet50 and a custom CNN in human action recognition.

2.4 Badminton Action Recognition

Badminton action recognition methods are categorized into two groups: machine learning-based classification and deep learning-based classification. Ting et al. [44] applied SVM for badminton recognition, classifying badminton strokes into ten distinct classes. Later, researchers further proved SVM's efficacy in badminton action recognition. Anik et al. [45] identified badminton stroke techniques. In this study, the accuracy of SVM reached 88%. Ghazali et al. [46] also confirmed that SVM provides superior classification results compared to decision trees, KNN, and SVM, with SVM achieving an accuracy of 83.4%. He concluded that future research should enhance recognition accuracy by expanding and refining feature extraction techniques. Another conventional machine learning method applied to BAR is AdaBoost. The recognition rates of AdaBoost are higher than those of HMM [34].

Deep learning has been applied in human action recognition. Wang et al. [47] applied the Adaptive Feature Extraction Block (AFEB) AlexNet. Rahmad et al. [4] have proven that GoogleNet has the highest classification performance compared to AlexNet, VGGNet-16, and VGGNet-19. GoogleNet provided the best accuracy for differentiating between smash and non-smash techniques [48,49]. Steels et al. [50] confirmed the accuracy of CNN in classifying nine activities using accelerometer data.

Analyzing badminton movements using skeleton key point data is a mostly unexplored area of research. Using the AlphaPose framework, Liang et al. [1] analyzed key point data from certain badminton films. Comparative investigation shows that the Long Short-Term Memory (LSTM) model surpasses the CNN model. Another model implemented is PDDRNet for 3D human pose estimation and XGBoost for classification [2]. Liu et al. [3] applied GCNN with skeletal data for badminton action recognition.

This research contributes to the spatiotemporal analysis of skeleton data and weighted ensemble learning for badminton action recognition. The use of video data processed into 3D skeleton coordinates. The data skeleton was chosen because, according to Sun et al. [51], the data skeleton has the advantage of providing 3D pose information of an object. It's simple yet informative and not sensitive to viewpoints, background, or light intensity changes. On the other hand, if the data is processed in RGB format, it will be susceptible to changes in viewpoint, light intensity, and changes in the background. The spatiotemporal feature captures the changes in the position of objects or body parts in space and time (temporal). This feature allows the model to understand the movement sequence, which is critical for recognizing actions that involve dynamic changes. In previous research, DTW was applied in human action recognition as a classifier by seeking the similarity between two sequences. In this paper, fast DTW is applied as a method for temporal feature extraction. One of the advantages of using fast DTW as a feature extraction method is its ability to reduce the number of features to a smaller amount. For example, the temporal feature in the right hip distance with dimensions 15×33, when processed with FDTW, will become a feature with dimensions 1×33.

Another contribution is the weighted ensemble of SVM, LR, RF, and AdaBoost. Machine learning classics were more efficient when compared with deep learning models. With appropriate feature extraction, machine learning models can achieve high accuracy. Ensemble learning methods that combine several classifier models have been proven to enhance accuracy.

3 The Proposed Method

Fig. 1 presents the proposed method. The first stage of video data acquisition for badminton action. The second stage is data preprocessing. The third stage is spatiotemporal data extraction. The fourth stage develops an ensemble learning model for badminton action recognition.

3.1 Data Acquisition

Data was collected using a 15MP DSLR camera with the installation as presented in Fig. 2. The video data resolution is 30 fps. An indoor badminton court served as the location for data collection. We used the existing lighting in the field without any additions. The video records a single badminton stroke technique. The execution of the badminton stroke begins from the athlete's ready position at the center of the court and returns to the center after performing the badminton stroke. The subjects included badminton athletes registered with PBSI, aged 15 to 22 years. Badminton techniques are presented in Fig. 3. The dataset consists of 1333 videos for training and 433 videos for validation.

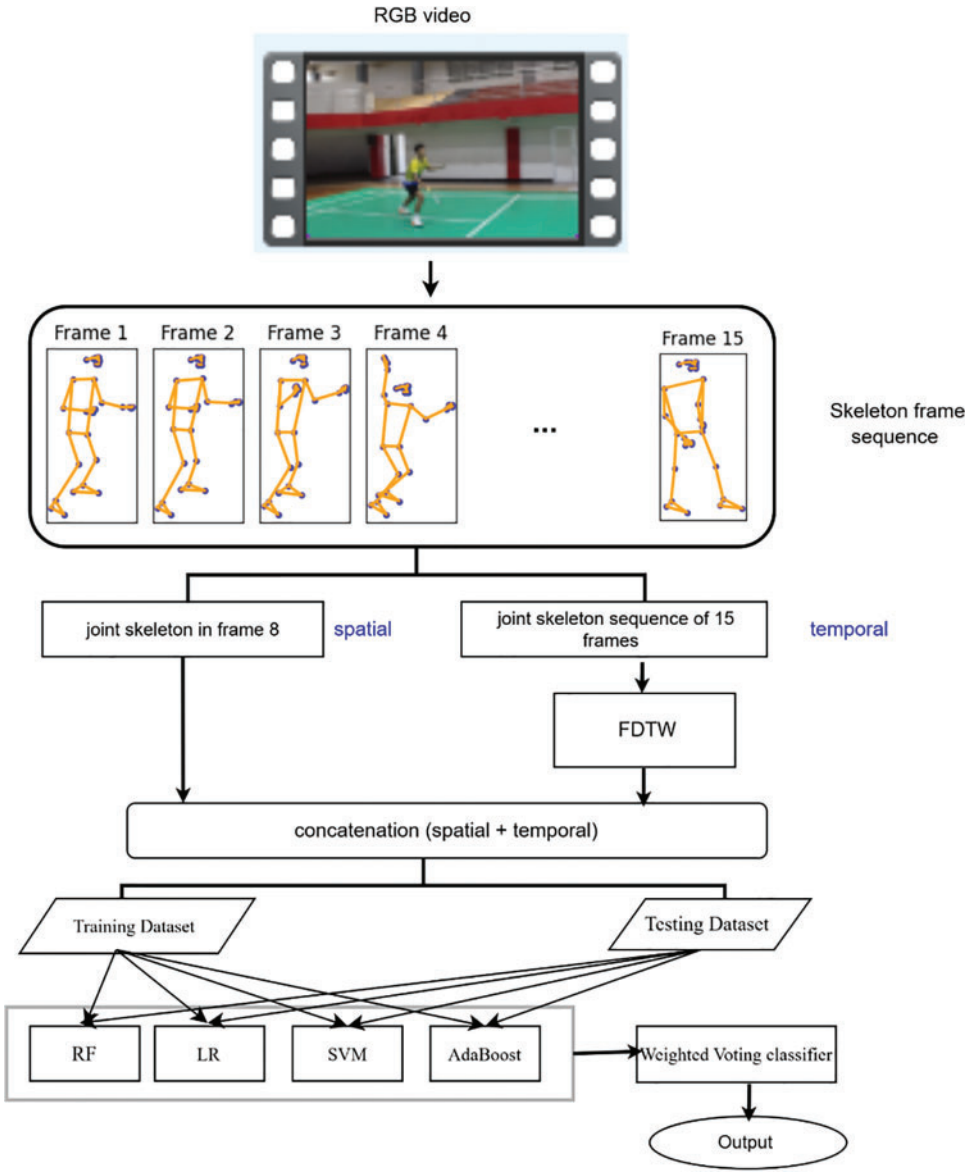

Figure 1: Proposed method

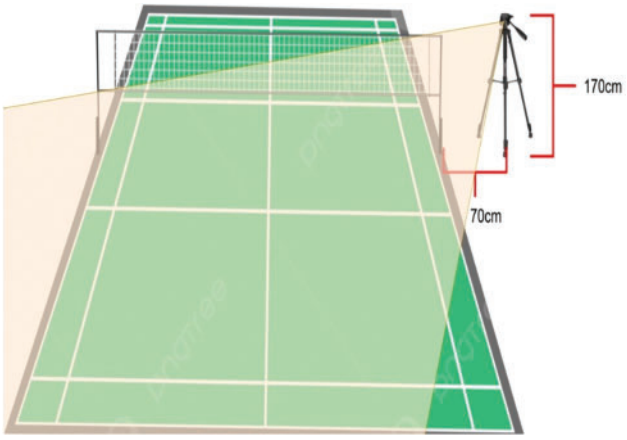

Figure 2: Camera installation

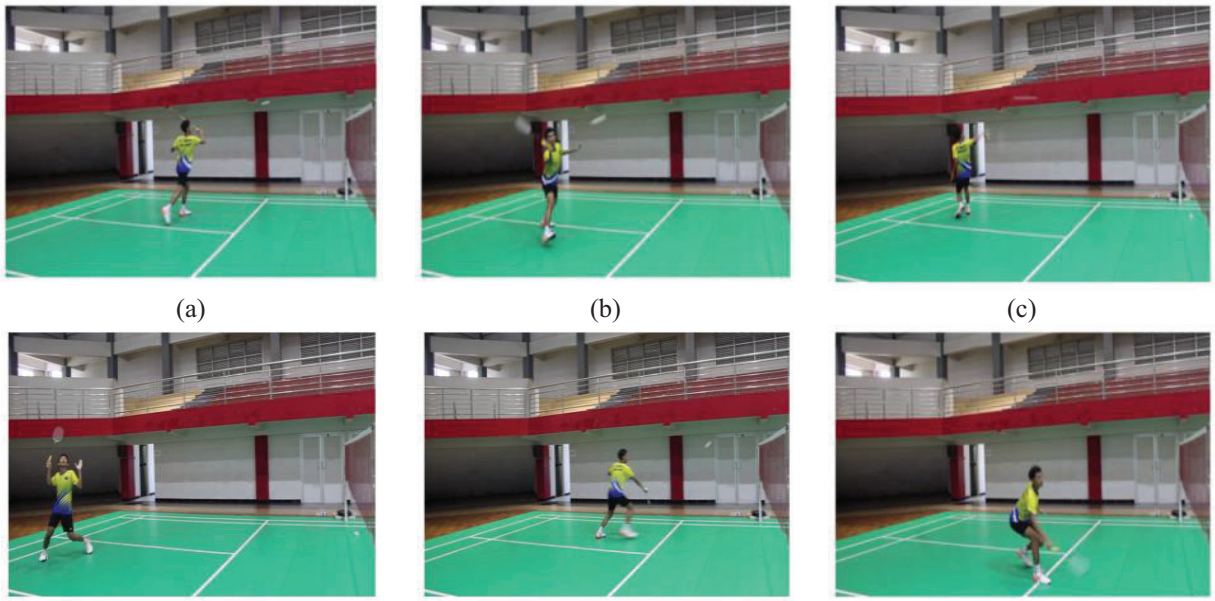

Figure 3: Badminton stroke techniques: (a) Drive backhand; (b) Drive forehand; (c) Overhead backhand; (d) Overhead forehand; (e) Underhand backhand; (f) Underhand forehand

3.2 Data Preprocessing

The data preprocessing stage segmented the T-frame video data into l-frame sequences. We perform the frame segmentation length of $l = 15$ by taking frames from 1 to T, using the frame interval as presented in Eq. (1). To extract the 3D coordinates for 33 pose landmarks, we apply media pipe to each frame. Fig. 4 presents the 33 pose landmarks. Fig. 5 presents the 15-frame pose landmarks.

$$interval = round\left(\frac{T}{l}\right) \tag{1}$$

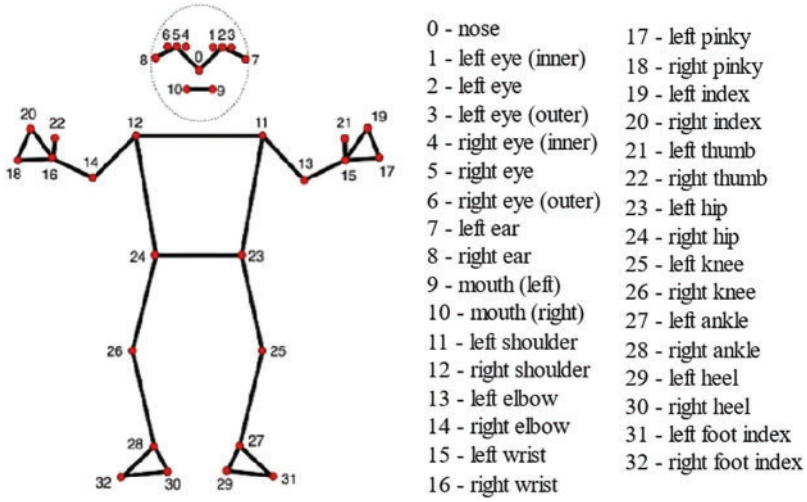

Figure 4: 33 Pose landmarks

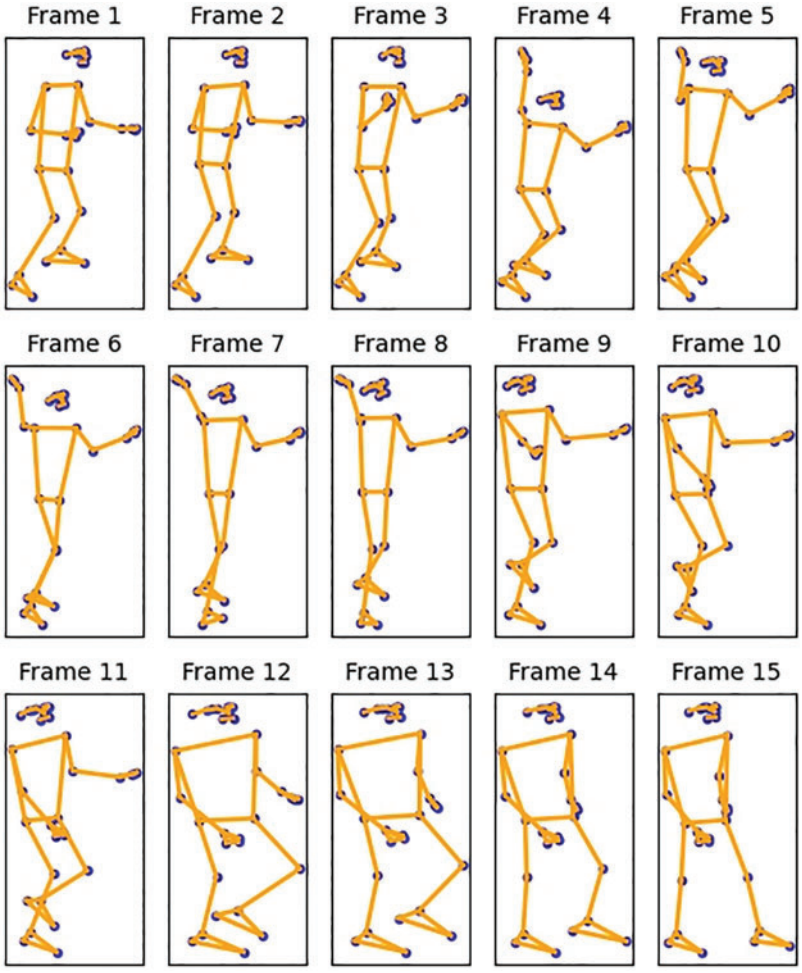

Figure 5: 15-Frame landmark pose

3.3 Feature Extraction

3.3.1 Spatial Feature Extraction

Preprocessing data produces an array with dimensions l × m × 3, where l represents the number of frames (l = 0, 1, 2, ..., 14), m represents the number of skeleton points (m = 0, 1, 2, ..., 32), and 3 represents the coordinates x, y, and z. The spatial feature is extracted by measuring the right hip distance or the three-dimensional distance between the right hip and other skeletal points. (d). spatial data is the right hip distance in frame 7. Based on the skeleton number in Fig. 4, the right hip coordinate from frame seven was written by $\left(x_{(7,24)},\ y_{(7,24)},\ xz_{(7,24)}\right)$. Eq. (2) represents the calculation of the right hip distance, while Eq. (3) represents the spatial features.

$$d\,(l,m) = \sqrt{\left(x_{(l,m)} - x_{7,24}\right)^2 + \left(y_{(l,m)} - y_{7,24}\right)^2 + \left(xz_{(l,m)} - z_{7,24}\right)^2} \tag{2}$$

$$Spatial\ feature = d(7,0), d(7,1), \ldots, d(7,32) \tag{3}$$

3.3.2 Temporal Feature Extraction

Temporal features are generated with FDTW applied to each column of right hip distance. In each data sample, d (l,m) is calculated. A sequence is defined as the order d (l,m) for each m, arranged according to the order of frames. For example, if m = 12 (right shoulder), the series is: (d (0,12), d (1,12), ..., d (14,12)). The calculation of FDTW requires a series as a reference. This paper uses data 0 (n = 0) as the reference. The temporal feature, which is the result of FDTW calculation in the form of a total distance (TD) sequence, is represented in Eq. (4). Algorithm 1 describes the FDTW algorithm.

$$Temporal\ feature = TD\,(0)\,, TD\,(1)\,, TD\,(2)\ldots., TD(32) \tag{4}$$

Algorithm 1: FDTW pseudocode

Input: X, Y, radius
Output: A minimum distance warp path between X and Y
1. // The min size of the coarsest resolution
2. Integer minTSsize = radius + 2
3. **if** (|X| ≤ minTSsize **OR** |Y| ≤ minTSsize)
4. {
5. // Base case: for a very small time series run the full DTW algorithm
6. **return** DTW (X, Y)
7. } **else** {
8. // Recursive case: Project the warp path from a coarser resolution onto the current resolution
9. // Run DTW only along the project path (and also radius cell from the project path).
10. TimeSeries shrunkX = X.reduceByHalf () //coarsening
11. TimeSeries shrunkY = Y.reduceByHalf () //coarsening
12. warpath lowResPath = FastDTW(shrunkX, shrunk, radius)
13. SearchWindow window = ExpandedResWindow (lowResPath, X, Y, radius) // Projection
14. **return** DTW (X, Y, window) //Refinement
15. }

The calculation of DTW, illustrated in Fig. 6, displays the results from a single joint skeleton of the right hip and the right hand in one video sequence, showing the calculation of DTW. The spatial data has 33 features, and the temporal data has 33 features, so the spatiotemporal data has 66 features.

This integration of spatial and temporal information provides a comprehensive representation of the athlete's movements, encompassing both the static and dynamic aspects of the badminton stroke techniques.

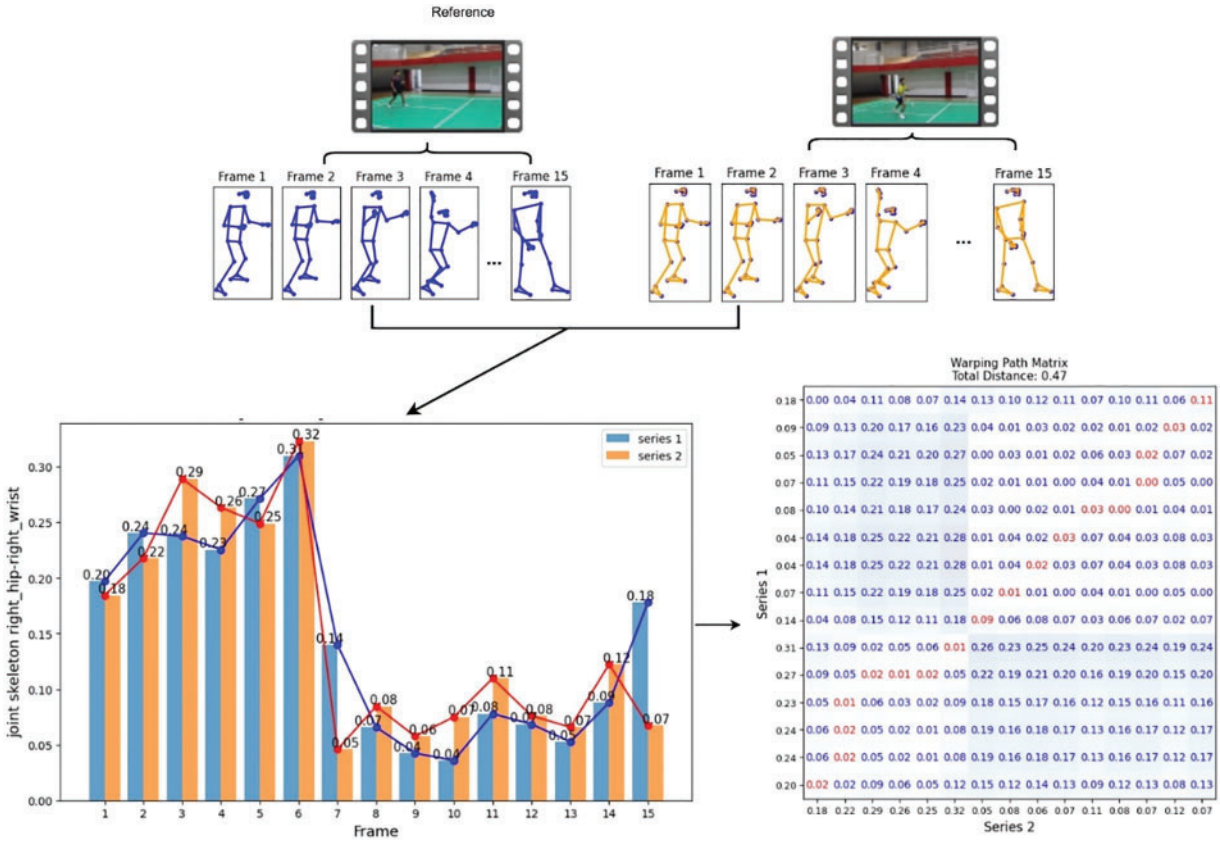

Figure 6: Calculation of DTW for skeleton points on right-hand

3.4 Machine Learning & Ensemble Model

We applied SVM, LR, RF, and AdaBoost in this research. Deep learning models (CNN and LSTM) are applied as a comparison in the machine learning model analysis. Each model was trained and evaluated individually. Five-fold cross-validation was used in validation.

The ensemble learning models in this study are arranged based on several variations of machine learning methods, as shown in Table 1, such as Ensemble-1 (E1) is configured from SVM, LR, RF, and AdaBoost, and configuration individual models for Ensemble-2 (E2), Ensemble-3 (E3), and Ensemble-4 (E4). The weighted soft voting classifier method (E1, E2, E3, and E4) can contribute to the final decision. The voting process will choose the more excellent score for more robust models.

Table 1: Ensemble model

Ensemble model	Model configuration
Ensemble-1 (E1)	SVM, LR, RF and AdaBoost
Ensemble-2 (E2)	SVM, LR and AdaBoost
Ensemble-3 (E3)	SVM, RF and AdaBoost
Ensembel-4 (E4)	SVM and AdaBoost

4 Result and Discussion

4.1 Machine Learning Models

We performed optimization by systematically testing several hyperparameter configurations for each model to determine the most effective combination that produces the maximum degree of performance. This method uses five-fold cross-validation to analyze the hyperparameters thoroughly. The training model that delivers the best performance metrics, such as accuracy, precision, recall, and F1 score, is chosen as the determinant of the best hyperparameters. The best hyperparameter model is shown in Table 2.

Table 2: Best hyperparameter

Model	Best hyperparameters
SVM	C: 10, gamma: 0.01, kernel: rbf
LR	C: 10, solver: liblinear
RF	criterion: gini, max_depth: 30, max_features: log2, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 300
AdaBoost	estimator__max_depth: 5, learning_rate: 1, n_estimators: 300

We conduct measurements of complexity, training time, and memory usage during model training. Table 3 presents the results. Eq. (5) to Eq. (10) allow for the calculation of the training time complexity (Big O). Compared to CNN and LSTM, classical machine learning (SVM, LR, RF, and AdaBoost) has lower complexity, training time, and memory usage during training.

$$Big\ O_{SVM} = O\left(n^2.d\right) until\ O\left(n^3.d\right) \tag{5}$$

$$Big\ O_{LR} = O\left(n.d\right) \tag{6}$$

$$Big\ O_{RF} = O\left(m.n.d\ \log n\right) \tag{7}$$

$$Big\ O_{AdaBoost} = O\left(m.n.d\right) \tag{8}$$

$$Big\ O_{CNN} = O\left(n.d^2.k^2.f\right) \tag{9}$$

$$Big\ O_{LSTM} = \left(n.t.h.d\right) \tag{10}$$

With n: number of samples, d: number of features, k: number of kernels, m: number of trees, t: sequence length.

Table 3: Algorithm's complexity and training cost model

Algorithm	Training time complexity	Training time (s)	Training memory (MB)
SVM	O (2.04×10^8) until O (2.98×10^{11})	0.47	0.32
LR	O (8.80×10^4)	0.36	0.10
RF	O (8.25×10^7)	4.86	0.26
AdaBoost	O (2.64×10^7)	17.04	3.12
CNN	O (2.15×10^9)	14.09	689.77
LSTM	O (4.97×10^8)	19.75	189.93

We performed the validation test using 433 distinct data points that differed from the training data. Table 4 shows that the performance of each deep learning method is greater than that of machine learning methods. For example, the accuracy of deep learning (CNN, 93.06%; LSTM, 94.06%). The precision of deep learning (CNN, 94.66%; LSTM, 96.31%), The recall of deep learning (CNN, 94.61; LSTM, 95.26), et cetera., the accuracy of deep learning (CNN, 93.06%; LSTM, 94.06%). The precision of deep learning (CNN, 94.66%; LSTM, 96.31%). The recall of deep learning (CNN, 94.61; LSTM, 95.26), etc.

Table 4: Evaluation models

Model (Individually)	Accuracy	Precision	Recall	F1 score
SVM	91.22	91.67	91.22	91.15
LR	90.53	91.14	90.53	90.48
RF	90.76	91.05	90.76	90.75
AdaBoost	92.15	91.68	91.28	91.15
CNN	93.06	94.66	95.10	94.61
LSTM	94.06	96.31	95.26	95.54

4.2 Weighted Ensemble Model

A soft voting classifier method in an ensemble model enhances accuracy and generalization in badminton action recognition. The weight value of each model SVM, LR, RF, and AdaBoost is assigned based on accuracy. We determine the weights by normalizing each model's accuracy values. Table 5 shows the comparison evaluation matrix of ensemble models. E2 achieved the highest performance with an accuracy of 95.38%, precision of 96.01%, recall of 94.89%, and F1 score of 95, 25%.

According to the validation test results, ensemble model E2, which combines SVM, LR, and AB, outperforms ensemble model E1, which combines SVM, LR, RF, and AB. The superior performance of E2 indicates that adding RF to E1 does not contribute to performance improvement and may even introduce noise that leads to a decrease in model efficiency. It also emphasizes that the selection of the right individual models is necessary to obtain an optimal ensemble model. E2 successfully leverages the strengths of SVM, LR, and AB more synergistically without the additional complexity that could weaken the overall generalization power of the model, as seen in E1. This finding shows that the

number of models in an ensemble does not always correlate with the final performance; instead, the quality and suitability of the combined models play a crucial role in determining the outcome.

Table 5: The evaluation of ensemble model validation

Ensemble model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
SVM-LR-RF- AdaBoost (E1)	94.69	95.25	94.19	94.53
SVM-LR- AdaBoost (E2)	**95.38**	**96.01**	**94.89**	**95.25**
SVM-RF- AdaBoost (E3)	93.30	93.59	92.86	93.03
SVM- AdaBoost (E4)	93.53	93.92	93.06	93.32

Refer to Fig. 7 for the evaluation results in the confusion matrix. An analysis of the confusion matrix from the top two ensemble models, E1 and E2, reveals a notable disparity in performance, especially in the categories of drive backhand and overhead backhand. Group E2 had the highest accuracy. Model E2 has superior sensitivity in detecting drive backhand patterns. Moreover, in the overhead backhand class, model E2 again demonstrates its superiority by achieving greater accuracy in differentiating this movement from other classes. That indicates a superior capability to identify patterns of overhead backhand action.

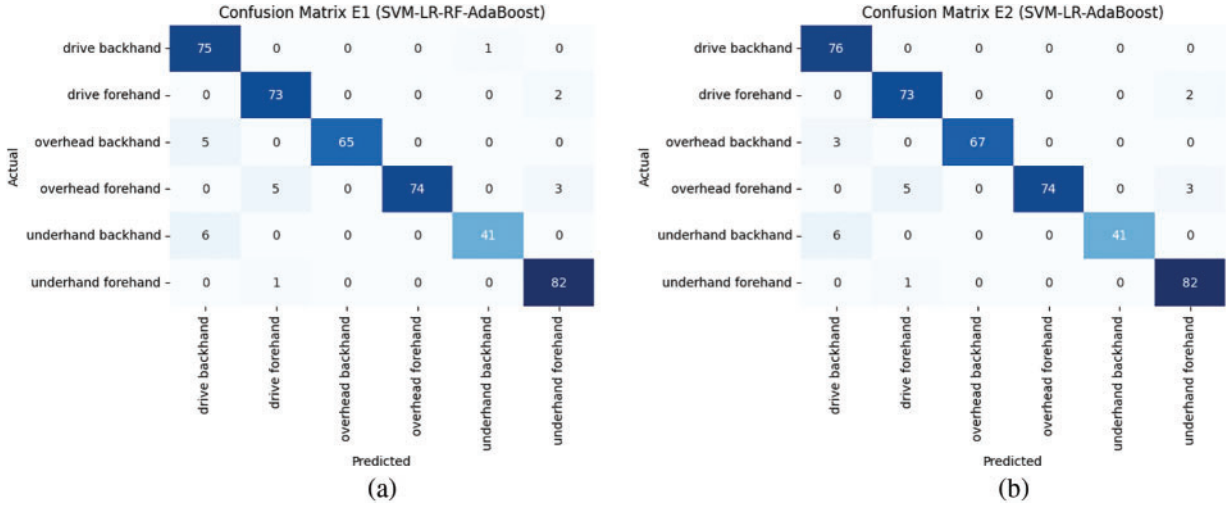

Figure 7: Confusion matrix: (a) Confusion matrix ensemble E1 (SVM-LR-RF-AdaBoost); (b) Confusion matrix ensemble E2 (SVM-LR-AdaBoost)

The classification levels for each class are presented in Fig. 8. Based on that figure, model E2 has an accuracy level of 100% for the drive backhand class, and four classes have an accuracy above 95%. In comparison, there are two classes with an accuracy below 95%, namely the overhead backhand class at 90.2% and the underhand forehand class at 87.2%. The accuracy value for the underhand backhand suggests that the model struggles to differentiate this movement from other movements with similar characteristics in this classification study.

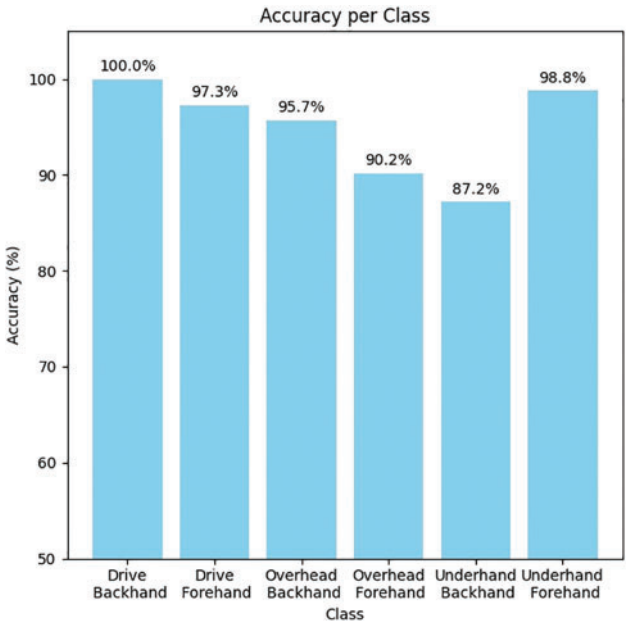

Figure 8: The accuracy of each class for E2

4.3 Analysis Result

To evaluate the results of the proposed ensemble learning model, we created a comparison table of the results of recognizing badminton stroke techniques using skeleton data as presented in Table 6. Each model in this table uses different data for its recognition process. Ensemble learning with 3D spatiotemporal skeleton features has higher accuracy compared to other models from the state of the art.

Table 6: Comparison of results and state-of-the-art

Method	Preprocessing data	Akurasi (%)
E2 (weighted ensemble SVM-LR-AdaBoost)	**Spatiotemporal 3D skeleton**	**95.38**
LSTM [1]	Skeleton data extracted by AlphaPose	80
CNN [1]	Skeleton data extracted by AlphaPose	60
PDDRNet-XGBOOST [2]	Human pose joint skeleton	93,5
GCN [3]	Human Skeleton Sequence	92
SVM [44]	RGB-D sensors	92
SVM [45]	Accelerometer and gyroscope	88.89
SVM [46]	Inertial senor	83.4
HMM [33]	Bounding box, HOG	83.33

5 Conclusion

Based on the experiments, spatiotemporal features of 3D skeleton data and ensemble models have been proven to provide high accuracy in recognizing badminton stroke techniques. The selection of machine learning models is significant in obtaining the best accuracy of the ensemble model. Based on the experiments, spatiotemporal features of 3D skeleton data and ensemble models have been proven to provide high accuracy in recognizing badminton stroke techniques. The selection of machine learning models is crucial to obtain the best accuracy of the ensemble model. E2 achieved the best performance with an accuracy rate of 95.38%.

Acknowledgement: We are truly thankful for the financial assistance from the Center for Higher Education Funding (BPPT) and the Indonesia Endowment Fund for Education (LPDP). Also, thank you to badminton athletes from Banyumas as the object of data acquisition. Thank you very much to the reviewers and editors, which helps improve the overall quality of ideas, concepts, and papers.

Funding Statement: This work is supported by the Center for Higher Education Funding (BPPT) and the Indonesia Endowment Fund for Education (LPDP), as acknowledged in decree number 02092/J5.2.3/BPI.06/9/2022.

Author Contributions: The contributions of the papers in this study are very diverse. Farida Asriani is responsible for data collection, research methodology design, experimental design, data analysis, and research reports. Azhari Azhari provides research concepts and planning, supervision, and correspondence of authors. Wahyono Wahyono plays a role in guiding and supervising research. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Liang and T. E. Nyamasvisva, "Badminton action classification based on human skeleton data extracted by AlphaPose," in *Int. Conf. Sens. Meas. Data Anal. Era Artif. Intell.*, 2023. doi: 10.1109/ICSMD60522.2023.10490491.

[2] X. -W. Zhou, L. Ruan, S. -S. Yu, J. Lai, Z. -F. Li and W. -T. Chen, "Badminton action classification based on PDDRNe," in *3rd Int. Conf. Internet, Educ. Inf. Technol.*, vol. 10, no. 17, pp. 980–987, 2023. doi: 10.2991/978-94-6463-230-9_118.

[3] J. Liu and B. Liang, "An action recognition technology for badminton players using deep learning," *Mobile Inf. Syst.*, vol. 2022, no. 1, pp. 1–10, 2022. doi: 10.1155/2022/3413584.

[4] N. A. Rahmad, N. A. J. Sufri, M. A. As'Ari, and A. Azaman, "Recognition of badminton action using convolutional neural network," *Indonesian J. Elect. Eng. Inf.*, vol. 7, no. 4, pp. 750–756, Dec. 2019. doi: 10.11591/ijeei.v7i4.968.

[5] S. Ramasinghe, K. G. M. Chathuramali, and R. Rodrigo, "Recognition of badminton strokes using dense trajectories," in *Int. Conf. Inf. Automat. Sustain.*, IEEE, Mar. 2014, pp. 1–6. doi: 10.1109/ICIAFS.2014.7069620.

[6] Z. Feng, Y. Shi, D. Zhou, and L. Mo, "Research on human activity recognition based on random forest classifier," in *Int. Conf. Control, Electron. Comput. Technol.*, IEEE, 2023, pp. 1507–1513. doi: 10.1109/ICCECT57938.2023.10140545.
[7] Z. Zaki, "Logistic regression based human activities recognition," *J. Mech. Contin. Math. Sci.*, vol. 15, no. 4, pp. 228–246, Apr. 2020. doi: 10.26782/jmcms.2020.04.00018.
[8] J. Wang *et al.*, "Deep 3D human pose estimation: A review," *Comput. Vis. Imag. Underst.*, vol. 210, no. 2, Sep. 2021, Art. no. 103225. doi: 10.1016/j.cviu.2021.103225.
[9] H. Cui and N. Dahnoun, "Real-Time short-range human posture estimation using mmwave radars and neural networks," *IEEE Sens. J.*, vol. 22, no. 1, pp. 535–543, Jan. 2022. doi: 10.1109/JSEN.2021.3127937.
[10] N. Tasnim, M. K. Islam, and J. H. Baek, "Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints," *Appl. Sci.*, vol. 11, no. 6, Mar. 2021, Art. no. 2675. doi: 10.3390/app11062675.
[11] H. Ramirez, S. A. Velastin, P. Aguayo, E. Fabregas, and G. Farias, "Human activity recognition by sequences of skeleton features," *J. Sens.*, vol. 22, no. 11, Jun. 2022, Art. no. 3991. doi: 10.3390/s22113991.
[12] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. IEEE Comput. Soci. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11988–11996. doi: 10.1109/CVPR.2019.01227.
[13] M. S. Alsawadi, E. S. M. El-Kenawy, and M. Rio, "Using blazepose on spatial temporal graph convolutional networks for action recognition," *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 19–36, 2023. doi: 10.32604/cmc.2023.032499.
[14] D. Giveki, "Robust moving object detection based on fusing Atanassov's Intuitionistic 3D Fuzzy Histon Roughness Index and texture features," *Int. J. Approx. Reasoning*, vol. 135, no. 18, pp. 1–20, Aug. 2021. doi: 10.1016/j.ijar.2021.04.007.
[15] D. Giveki, "Scale-space multi-view bag of words for scene categorization," *Multimed. Tools Appl.*, vol. 80, no. 1, pp. 1223–1245, Jan. 2021. doi: 10.1007/s11042-020-09759-9.
[16] S. Zhang, W. Chen, C. Chen, and Y. Liu, "Human deep squat detection method based on MediaPipe combined with Yolov5 network," in *Proc. 41st Chin. Control Conf.*, Hefei, China, 2022. doi: 10.23919/CCC55666.2022.9902631.
[17] J. Ma, L. Ma, W. Ruan, H. Chen, and J. Feng, "A wushu posture recognition system based on MediaPipe," in *Int. Conf. Inf. Technol. Contemp. Sports*, IEEE, 2022, pp. 10–13. doi: 10.1109/TCS56119.2022.9918744.
[18] S. H. Kim and J. Y. Chang, "Single-shot 3D multi-person shape reconstruction from a single RGB image," *Entropy*, vol. 22, no. 8, Aug. 2020, Art. no. 806. doi: 10.3390/e22080806.
[19] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, "HEMlets PoSh: Learning part-centric heatmap triplets for 3D human pose and shape estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3000–3014, Jun. 2022. doi: 10.1109/TPAMI.2021.3051173.
[20] X. Li, M. Zhang, J. Gu, and Z. Zhang, "Fitness action counting based on MediaPipe," in *Proc. Int. Congr. Imag. Signal Process., Biomed. Eng. Inf.*, IEEE, 2022. doi: 10.1109/CISP-BMEI56279.2022.9980337.
[21] L. Bai, T. Zhao, and X. Xiu, "Exploration of computer vision and image processing technology based on OpenCV," in *Proc. Int. Conf. Comput. Sci. Eng. Tech.*, IEEE, 2022, pp. 145–147. doi: 10.1109/SCSET55041.2022.00042.
[22] P. Palani, S. Panigrahi, S. A. Jammi, and A. Thondiyath, "Real-time joint angle estimation using Mediapipe framework and inertial sensors," in *Proc. Int. Conf. Bioinf. Bioeng.*, IEEE, 2022, pp. 128–133. doi: 10.1109/BIBE55377.2022.00035.
[23] S. Adhikary, A. K. Talukdar, and K. Kumar Sarma, "A vision-based system for recognition of words used in indian sign language using MediaPipe," in *Int. Conf. Imag. Inf. Process.*, IEEE, 2021, pp. 390–394. doi: 10.1109/ICIIP53038.2021.9702551.
[24] P. Padhi and M. Das, "Hand gesture recognition using DenseNet201-Mediapipe hybrid modelling," in *Int. Conf. Automat. Comput. Renew. Syst.*, 2022, pp. 995–999. doi: 10.1109/ICACRS55517.2022.10029038.
[25] W. Fu *et al.*, "Spatiotemporal correlation based self-adaptive pose estimation in complex scenes," *Digit. Commun. Netw.*, 2024. doi: 10.1016/j.dcan.2024.03.007.

[26] Y. Wang, Y. Xia, and S. Liu, "BCCLR: A skeleton-based action recognition with graph convolutional network combining behavior dependence and context clues," *Comput. Mater. Contin.*, vol. 78, no. 3, pp. 4489–4507, 2024. doi: 10.32604/cmc.2024.048813.

[27] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 1010–1019. doi: 10.1109/CVPR.2016.115.

[28] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018. doi: 10.1109/TPAMI.2017.2771306.

[29] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018. doi: 10.1609/aaai.v32i1.12328.

[30] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 12026–12035.

[31] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," Mar. 2018, *arXiv:1803.01271*.

[32] P. Koniusz, L. Wang, and A. Cherian, "Tensor representations for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 648–665, Feb. 2022. doi: 10.1109/TPAMI.2021.3107160.

[33] W. T. Chu and S. Situmeang, "Badminton video analysis based on spatiotemporal and stroke features," in *Proc. Int. Conf. Multimed. Retr.*, 2017, pp. 448–451. doi: 10.1145/3078971.3079032.

[34] J. Long and S. Rong, "Application of machine learning to badminton action decomposition teaching," *Wirel. Commun. Mobile Comput.*, vol. 2022, no. 9, pp. 1–10, 2022. doi: 10.1155/2022/3707407.

[35] N. Tanaka, H. Shishido, M. Suita, T. Nishijima, Y. Kamed and I. Kitahara, "Detection of shot information using footwork trajectory and skeletal information of badminton players," *Int. Conf. Sport Sci. Res. Technol. Support*, pp. 112–119, 2023. doi: 10.5220/0012162700003587.

[36] Z. Chu and M. Li, "Image recognition of badminton swing motion based on single inertial Sensor," *J. Sens.*, vol. 2021, no. 1, 2021. doi: 10.1155/2021/3736923.

[37] A. Karim, A. Azhari, M. Shahroz, S. B. Belhaouri, and K. Mustofa, "LDSVM: Leukemia cancer classification using machine learning," *Comput. Mater. Contin.*, vol. 71, no. 2, pp. 3887–3903, 2022. doi: 10.32604/cmc.2022.021218.

[38] M. Karray, N. Triki, and M. Ksantini, "A new speed limit recognition methodology based on ensemble learning: Hardware validation," *Comput. Mater. Contin.*, vol. 80, no. 1, pp. 119–138, 2024. doi: 10.32604/cmc.2024.051562.

[39] H. Wu, W. Pan, X. Xiong, and S. Xu, "Human activity recognition based on the combined SVM&HMM," in *Int. Conf. Inf. Automat.*, Hailar, China, 2014. doi: 10.1109/ICInfA.2014.6932656.

[40] S. Jindal, M. Sachdeva, A. K. S. Kushwaha, and I. K. Gujral, "Performance evaluation of machine learning based voting classifier system for human activity recognition," *Kuwait J. Sci.*, pp. 1–12, 2022. doi: 10.48129/kjs.splml.19189.

[41] S. Das, A. Maity, R. Jana, R. Biswas, A. Biswas and P. K. Samanta, "Automated improved human activity recognition using ensemble modeling," in *Int. Conf. Recent Adv. Elect. Electron. Ubiquitous Commun. Comput. Intell.*, IEEE, 2024, pp. 1–6. doi: 10.1109/RAEEUCCI61380.2024.10547875.

[42] T. H. Tan, J. Y. Wu, S. H. Liu, and M. Gochoo, "Human activity recognition using an ensemble learning algorithm with smartphone sensor data," *Electronics*, vol. 11, no. 3, Feb. 2022. doi: 10.3390/electronics11030322.

[43] R. Kaur and D. Veer Sharma, "Human action recognition using an ensemble deep learning model for video datasets," *J. Harbin Eng. Univ.*, vol. 44, no. 7, pp. 1006–7043, 2023.

[44] H. Y. Ting, K. S. Sim, and F. S. Abas, "Automatic badminton action recognition using RGB-D sensor," *Adv. Mater. Res.*, vol. 1042, pp. 89–93, 2014. doi: 10.4028/www.scientific.net/AMR.1042.89.

[45] M. A. I. Anik, M. Hassan, H. Mahmud, and M. K. Hasan, "Activity recognition of a badminton game through accelerometer and gyroscope," in *Int. Conf. Comput. Inf. Technol.*, IEEE, Feb. 2017, pp. 213–217. doi: 10.1109/ICCITECHN.2016.7860197.

[46] N. F. Ghazali, N. Shahar, and M. A. As'ari, "Badminton strokes recognition using inertial sensor and machine learning approach," in *Int. Conf. Intell. Cybern. Technol. Appl.*, IEEE, 2022, pp. 1–5. doi: 10.1109/ICICyTA57421.2022.10037897.

[47] Z. Wang, M. Guo, and C. Zhao, "Badminton stroke recognition based on body sensor networks," *IEEE Trans. Hum. Mach. Syst.*, vol. 46, no. 5, pp. 769–775, Oct. 2016. doi: 10.1109/THMS.2016.2571265.

[48] N. A. Rahmad and M. A. As'ari, "The new Convolutional Neural Network (CNN) local feature extractor for automated badminton action recognition on vision-based data," *Int. Conf. Emerg. Comput. Technol. Sport.*, vol. 1529, no. 2, Jun. 2020. doi: 10.1088/1742-6596/1529/2/022021.

[49] N. A. Rahmad, M. A. As'Ari, K. Soeed, and I. Zulkapri, "Automated badminton smash recognition using convolutional neural network on the vision-based data," *Mater. Sci. Eng.*, vol. 884, no. 1, pp. 1–7, 2020. doi: 10.1088/1757-899X/884/1/012009.

[50] T. Steels, B. Van Herbruggen, J. Fontaine, T. De Pessemier, D. Plets and E. De Poorter, "Badminton activity recognition using accelerometer data," *J. Sens.*, vol. 20, no. 17, pp. 1–16, Sep. 2020. doi: 10.3390/s20174685.

[51] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, pp. 3200–3225, Mar. 2023. doi: 10.1109/TPAMI.2022.3183112.