FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

Cascaded Segmentation-Classification Framework Optimized for Sports Action Recognition Using Still Images: A Comparative Analysis of PSO and Grid Search

Audrey Huong (PhD)[1*], Ser Lee Loh (PhD)[2], Kok Beng Gan (PhD)[3], Xavier Ngu (PhD)[4]

[1]Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat Johor, Malaysia; Email: audrey@uthm.edu.my
[2]Centre for Robotics and Industrial Automation, Fakulti Teknologi dan Kejuruteraan Elektrik, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia; Email: slloh@utem.edu.my
[3]Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia; Email: kbgan@ukm.edu.my
[4]RF EMC Centre Malaysia Sdn. Bhd., Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia; Email: xavier@uthm.edu.my

***Correspondence:** *Audrey Huong (PhD)*, Email: *audrey@uthm.edu.my*

ABSTRACT- The conventional method of evaluating human actions and activities requires integrating complex hardware and interpretation systems. This article proposes using a hybrid segmentation-classification system to recognize human sports action automatically based on still frames. This study tested the proposed framework on a public dataset containing images of humans performing three athletic actions: dancing, performing martial arts, and playing net sports. This framework combined a two-stage U-Net and EfficientNet-B0, and GoogleNet, and is optimized using particle swarm optimization (PSO) and grid search for cascaded segmentation and classification problems. Results indicated that PSO improves segmentation and classification accuracies by 5 % and 60 %, respectively, compared to the conventional grid search. The PSO segmentation results showed good agreement between the predicted mask and its ground truth, with overlapping scores of 0.85-0.9. The consecutive classification experiments revealed a slight superiority in the performance of EfficientNet-B0 with an accuracy of 0.86-0.88 over GoogleNet (~0.82-0.84), which also showed a lower convergence efficiency. While no significant difference was observed in the recall scores of EfficientNet between the weighted and unweighted loss methods, GoogleNet showed a considerable improvement in the true positive rates from 0.6177 to 0.7160 using the weighted loss strategy. Despite using the weighted loss method, there were negligible improvements in the performance of grid search-optimized networks. The poor classification results suggest low model generalization ability using manual tuning. This study concluded that the PSO-optimized cascaded segmentation-classification framework could potentially leverage advancements in human movement evaluation and rehabilitation assessment for sports applications. This system can be improved by adopting larger datasets or collaborative machine learning training to enhance model convergence for practical rehabilitation applications to assess users' physical and fitness performance.

Keywords: Sports action, EfficientNet, Optimization, Segmentation, Grid search.

ARTICLE INFORMATION

Author(s): Audrey Huong (PhD), Ser Lee Loh (PhD), Kok Beng Gan (PhD), and Xavier Ngu (PhD);
Received: 03/03/2025; **Accepted:** 28/08/2025; **Published:** 10/12/2025;
E- ISSN: 2347-470X;
Paper Id: IJEER 0303-03;
Citation: 10.37391/ijeer.130414
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-13/ijeer-130414.html

Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Sports science is a multidisciplinary field that studies the function of the human body and physical activity to improve sports performance. Sports posture and action recognition analysis are important aspects and aims of physical education and training when evaluating the suitability of sports movement and activities while preventing sports injuries and trauma. Statistics by the National Safety Council (NSC) reported that approximately 3.7 million patients were treated in emergency departments in 2023 for sports and recreational-related injuries and trauma [1]. Many of these injuries can be prevented with proper training programs.

The traditional physical performance analysis relied on input and evaluations of sports and rehabilitation specialists and fitness practitioners [2]. Individualized training or exercise plans are prescribed through consistent monitoring and regular identification of risk factors to promote rehabilitation, optimize functional motor capacity, and improve sports performance outcomes. These practices are important to help athletes maximize their performance, and the activities are normally

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

carried out within a sports institution or care facility. Over the years, home-visiting services have been made available for individuals, especially the elderly with poor mobility and poor cognition, allowing them to engage in physical activity and exercise safely [2, 3]. Nonetheless, traditional manual assessment methods are subjective due to the inconsistency of human nature, making them inefficient, expensive, and error-prone [2].

Following the growing demand for professional fitness delivery and facility resources, technology-assisted sports rehabilitation and injury prevention have become popular because of their remarkable efficiency and flexibility [2, 4-5]. Recent technologies in this field include using state-of-the-art sensors and immersive tools like virtual reality (VR) to monitor human posture and rehabilitation performance [6]. Yang and Meng [7] established an all-dimensional stereoscopic display system to capture the sports movement reconstructed in VR space for computer-aided motion correction to optimize motion technical skills in athletes. Others in [8, 9] used wearable devices to gather critical information through embedded sensors, such as position, orientation, and configuration, which can be transformed into real-time virtual information for analysis. These studies take full advantage of information and communication technology (ICT) and the Internet of Things (IoT) technologies to expand patients' experience. Ambulatory wearable systems developed based on an Inertial Momentum Unit (IMU) consisting of accelerometers, gyroscopes, magnetometers, or pressure-sensitive or mechanical sensors are popular for posture estimation in physical rehabilitation assessment [4, 10-11].

Chrono-photography and computer vision (CV) technology using cameras are receiving increased attention in sports injury research to predict human movement and posture using still images or video sequences due to their cost-effectiveness and simplicity. This strategy has been used with deep learning (DL) methods. Wang [12] proposed using a deep 3D convolutional neural network (CNN) and graph theory to recognize standard and wrong demonstrative sports (*i.e.*, gymnastics) movements based on a video sequence. The prediction is based on the extraction of different key angle features, *e.g.*, angle features of wrist-elbow-shoulder and neck-shoulder-elbow. Instead of meticulous key descriptors and angle calculations, [13] proposed using an extreme learning machine approach to fuse the scores of manuals and DL feature kernels from the long-term recurrent convolutional networks in predicting sports micromovement. CNN and transformer-based architecture performance were compared in [14] for sports (ball games) recognition tasks based on video clips. The results showed exceptional performance in detecting various action scenes using the CNN-based models with a prediction accuracy of over 90 %. Whilst most of these studies predict human motion or activities based on spatiotemporal features extracted from a series of movements from a video clip, others use still or scenic photographs to recognize sports action. Mottaghi et al. [15] classified sports (free wrestling) action using silhouette-skeleton features to overcome the conventional skeleton-based features that required expensive depth or stereo cameras and tracking methods. The study mapped the silhouettes to a fixed coordinate system origin and converted the image into a skeleton through morphological operations-based thinning; the latter is transformed into a graph that reveals information about the body joints and parts for classification using a support vector machine (SVM) and k-nearest neighbor (KNN) classifiers [15]. Robust Deep Active learning (RDAL) was used in [16] to create gaze shift paths in capturing relevant representations in a dancing scene, followed by manifold learning principles as feature extractors to isolate deep features and SVM for action recognition. Alavigharahbagh et al. [17] demonstrated gradient-based analyses of image frames in silhouette extraction, whose outputs were enhanced with active contour and color-based image segmentations for segmenting the motion region. The features of key points on the human body were used for action prediction using Long-Short Term Memory (LSTM).

Despite the good recognition accuracies using DL techniques, the high requirements of hardware resources and training data robustness (*i.e.*, high intra-class variability and low inter-class variance) [13, 17-19] remain a practical barrier to the widespread adoption of this technology in this field. There are also data quality and imbalance issues, and low availability of annotated data [13-14, 19-21]; these challenges are often compounded by a lack of inter-class and intra-class variability that does not represent a larger population, limiting the number of features that can be extracted. Because of the latter issue, most features of the majority class have been learned, dominating model performance and causing a bias in the decisions in favor of the majority class [22]. Thus, the model may unnecessarily early converge with limited local features that may not represent the patterns effectively, limiting the network's learning ability [19, 22].

Many previous studies approach these challenges by either integrating DL with uniquely formulated feature extraction or meticulous mathematical analyses [16], or increasing the complexity of their model (*e.g.*, multi-dimensionality) [12, 17] to improve model learning capacities. Some less expensive and easily accessible approaches to address insufficient dataset problems include implementing innovative methods, such as simple linear iterative clustering, partitioning an image into several meaningful fragments [16], or using domain randomization to generate synthetic datasets [20]. Other efforts to prevent class imbalance and model overfitting include assigning weights to individual features for optimal feature selection [16, 19, 21] or improving the model learning by tuning the related hyperparameters to accelerate model convergence [19-20]. Traditional manual tuning, either through random best guess or grid search, is popular for improving the robustness of CNN models due to their simplicity and ease of implementation; they also work well for smaller and intermediate parameter ranges [23]. Nonetheless, automatic and adaptive optimization approaches, such as genetic algorithms or PSO [12, 19, 24], can explore the search space more effectively and provide a higher quality solution for optimizing the entire population [24]. This paper aims to further progress in human activity and action recognition by introducing a cascaded segmentation and classification framework optimized using PSO and grid search.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

A weighted cross-entropy loss is also adopted to address the class imbalance problem. Unlike the earlier works in [12-17], the proposed approach uses still images extracted from video clips as input in the detection process. It is a cost-effective alternative without expensive and complex electronics and sensor systems, making it economically suitable for non-point-of-care implementation.

2. MATERIALS AND METHODS

The following subsections describe data handling methods, the proposed optimization frameworks, and the hybrid architecture designed for sports action recognition.

2.1. Dataset and data handling

Numerous public datasets are available for research on human detection and action tracking. The dataset used in this study consists of images and their masks downloaded from Kaggle [25]. This public dataset was developed from the Martial Arts, Dancing and Sports (MADS) data source (http://visal.cs.cityu.edu.hk/research/mads/) consisting of videos of five professional sportspersons performing martial arts (tai-chi and karate), dancing (hip-hop and jazz) and sports actions (basketball, volleyball, football, rugby, tennis, and badminton) for human pose tracking estimation. The videos were recorded using multiple cameras from different angles. The videos were split into still images in this original dataset before the person in each frame was manually extracted and segmented for ground-truth pose data. The process produces 1,192 still images and their corresponding masks. These images and masks have a consistent size of 384 × 512 pixels, as shown in *figure 1*.

The masks in RGB format are converted to grayscale images and binarized using the MATLAB imbinarize function, with pixel "0" representing the background and "1" denoting the object's pixel. The images and their masks were randomly split into training, validation, and testing purposes using a random seed number of 1 and a split ratio of 70/15/15 %, corresponding to 1,830/392/393 samples, respectively. The classification process required grouping samples in these sets into different actions. The dataset contains images representing five dynamic athletic events: hip-hop and jazz dance moves, badminton, tennis, volleyball movements, as well as karate and tai-chi movements. Due to the low sample size, these activities collapsed into three: dancing, performing martial arts, and playing net sports. Thus, unlike the segmentation problem, where the samples were randomly split without considering the types of athletic activities, the classification task deals with imbalance data (*i.e.*, an inconsistent number of samples under each action), as shown in *table 1*.

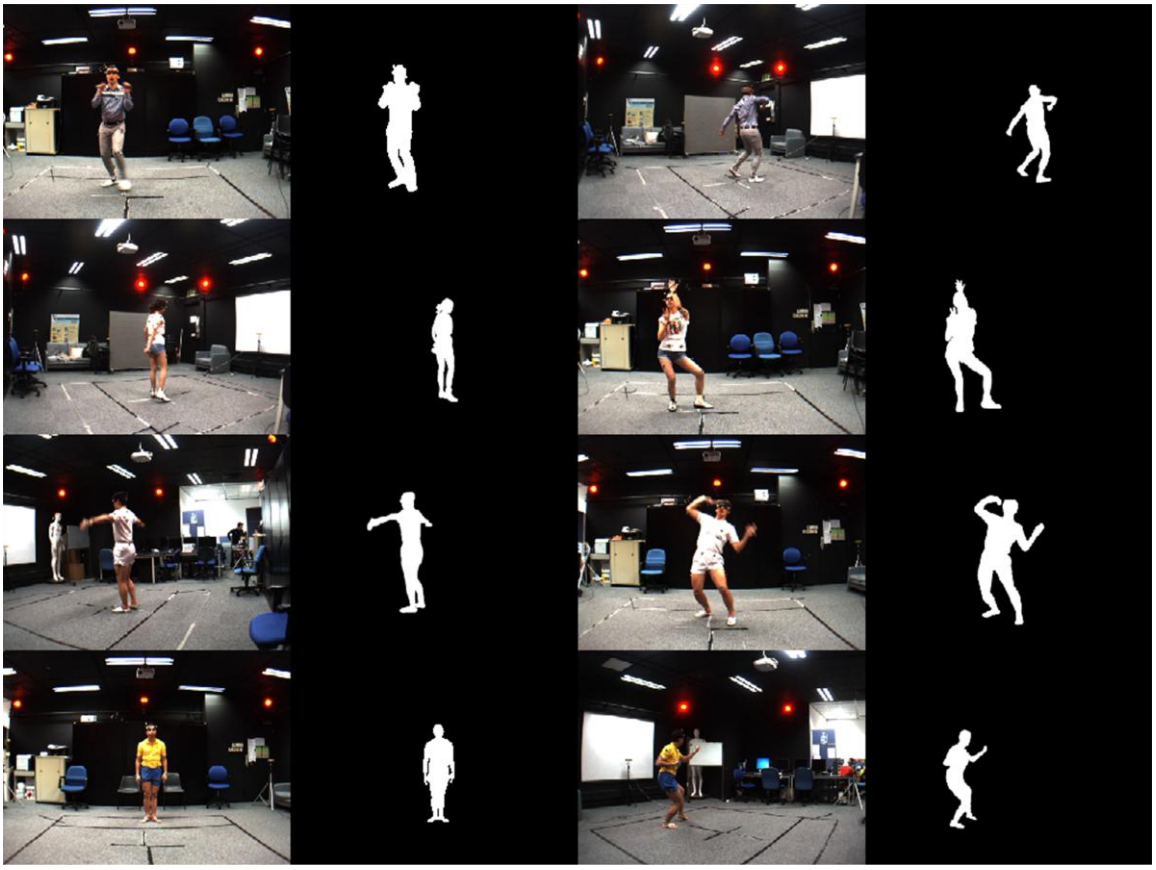


Figure 1. Still images of individuals performing different actions extracted from videos captured at different angles and their ground-truth masks

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

Table 1. The number of different athletic action images under each dataset

Dataset	Athletic action images		
	Dance	Martial arts	Net sports
Training	412	357	65
Validation	104	55	20
Testing	93	73	13

2.2. Segmentation-classification framework

The proposed cascaded framework shown in *figure 2* involves segmenting object pixels in a scene using semantic segmentation, as discussed in *section 2.2.1*. This process produces a silhouette mask for the athletic action classification task described in *section 2.2.2*. This study enhances the deep learning model's generalizability for object segmentation and action recognition by adopting automatic PSO and manual grid search methods to tune important hyperparameters in *section 2.3*. This strategy is combined with a weighted cross-entropy loss to address class imbalance.

2.2.1. Segmentation model

Semantic segmentation is the first process in this framework to automatically track humans in a scene and classify their poses. This process produces a binary mask highlighting the human pixels. This study input images of the original size of 512 × 384 into the fully convolutional U-Net model shown in *figure 3*. This U-shaped model consists of a two-stage encoder to extract the most significant features in the dataset and a two-stage decoder to recover the features and image dimension, producing a segmentation map. These two sections are connected by a bridge network consisting of two convolutions. The detected human pixels were post-processed to estimate the activity pose, as discussed in *section 2.2.3*.

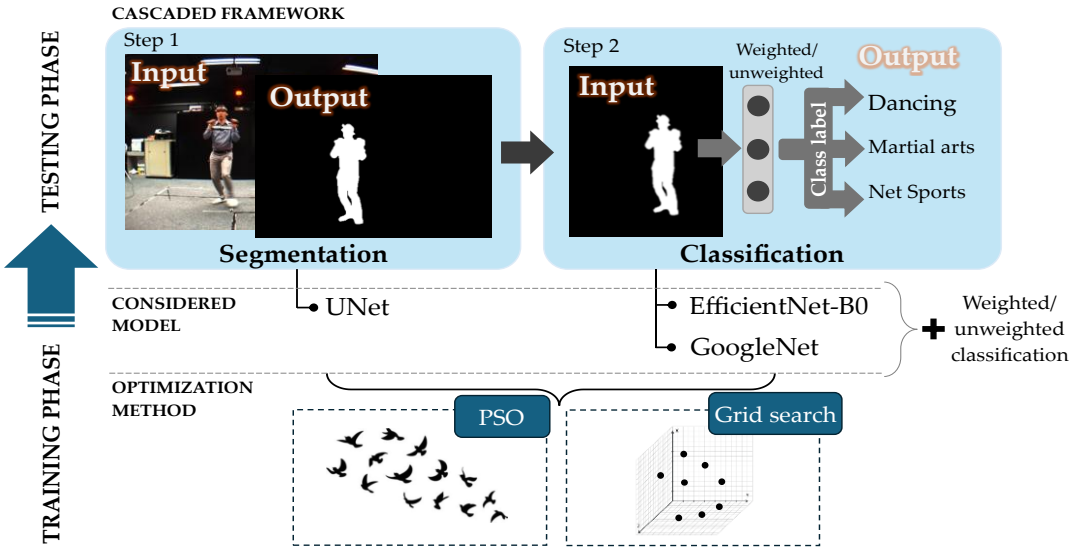


Figure 2. The proposed strategies and cascading segmentation-classification pipeline for athletic action recognition. The deep learning models are optimized using particle swarm optimization (PSO) and grid search, and added with a weighted loss function

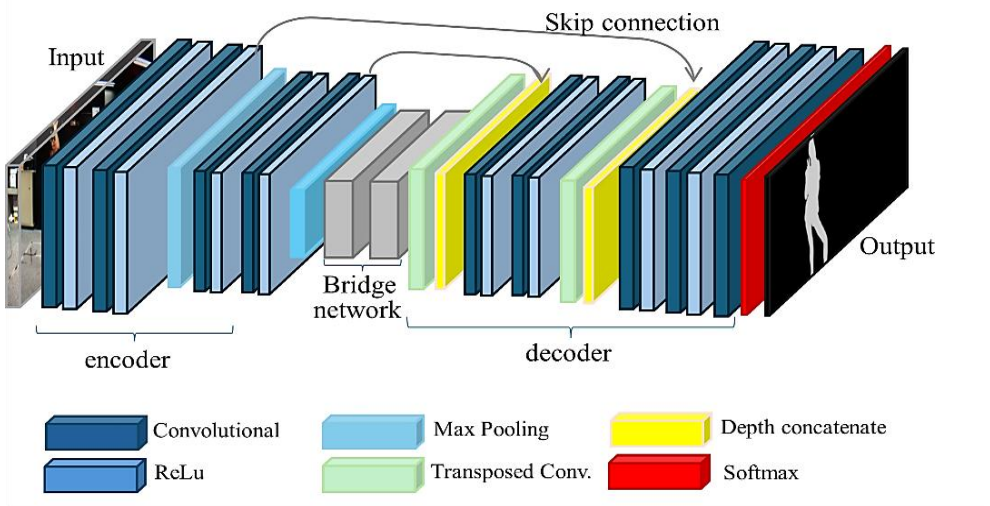


Figure 3. Two-stage encoder-decoder U-Net segmentation model architecture

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

2.2.2. Classification model

The segmented human silhouette is input into a CNN for action recognition. The networks shown in *figure 4* are *(a)* EfficientNet-B0 and *(b)* GoogleNet. EfficientNet-B0 is a state-of-the-art model that adopts an innovative scaling strategy to efficiently scale the network's width, depth, and resolution to improve model generalizability without excessive fine-tuning and complexity. *Figure 4(a)* shows the 260-layer deep network's scaling structure, consisting of 4 million total learnables. GoogleNet, shown in *figure 4 (b)*, is a 22-layer deep CNN based on the inception module. This network has 5.9 million learnables to extract various features while effectively reducing the training parameters. Since significant differences exist between the size of each class label, this study considered both weighted and unweighted loss systems as the output of these models, as shown in *figure 2*. The class weight, W, given for each class, j, is calculated based on the total number of samples ($N = 834$), bin count information in *table 1*, and number of classes ($n = 3$) in *eq. (1)* as 0.6748, 0.7787, and 4.2769 for dancing, martial arts, and net sports classes, respectively.

$$W_j = \frac{N}{n \cdot bincount(j)} \quad (1)$$

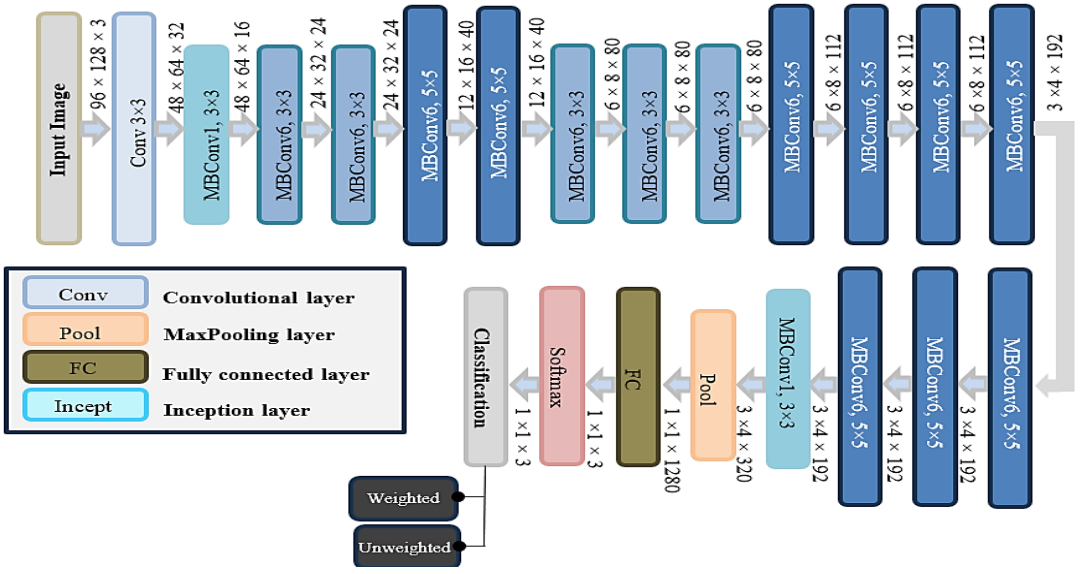
These class weights (W_j) are incorporated in the calculation of the cross-entropy loss function (L_{ce}) that learn to balance the class contributions and improve a model's predictions by measuring the discrepancy between the predicted probability of class j (P_j) and the true probability of the class, T_j [26]. The weighted cross-entropy loss formula is given in *eq. (2)*. The variable W_j is omitted from the formula when the unweighted cross-entropy loss is considered.

$$L_{ce} = -\frac{1}{n}\sum_{j=1}^{j=n} W_j T_j \log P_j \quad (2)$$

EfficientNet-B0 is a complex model with a deeper architecture and a higher parameter count than GoogleNet. Therefore, pre-experimental testing used this model to choose the hyperparameter range to account for the limitation in computational resources. The pre-experimental findings showed a significant trade-off between image size and mini-batch value. A good image size preserves the original image resolution and quality, allowing important information to be learned during training. However, it produces poor classification results using the same minibatch size defined for the segmentation problem, *i.e.*, a maximum value of 32. Nonetheless, a further increase in the minibatch value leads to early termination of the jobs due to insufficient GPU memory. Therefore, the size of the input images was reduced to ¼ of its original size (*i.e.*, 96 × 128 ×3) for classification purposes, while the minibatch value was increased to 128. The networks in *figure 4* are trained one at a time with the weighted and unweighted loss functions in *eq. (2)*. Networks with a weighted cross-entropy loss are referred to in the following text as weighted loss networks, while unweighted loss is used to describe networks with an unweighted cross-entropy loss function.

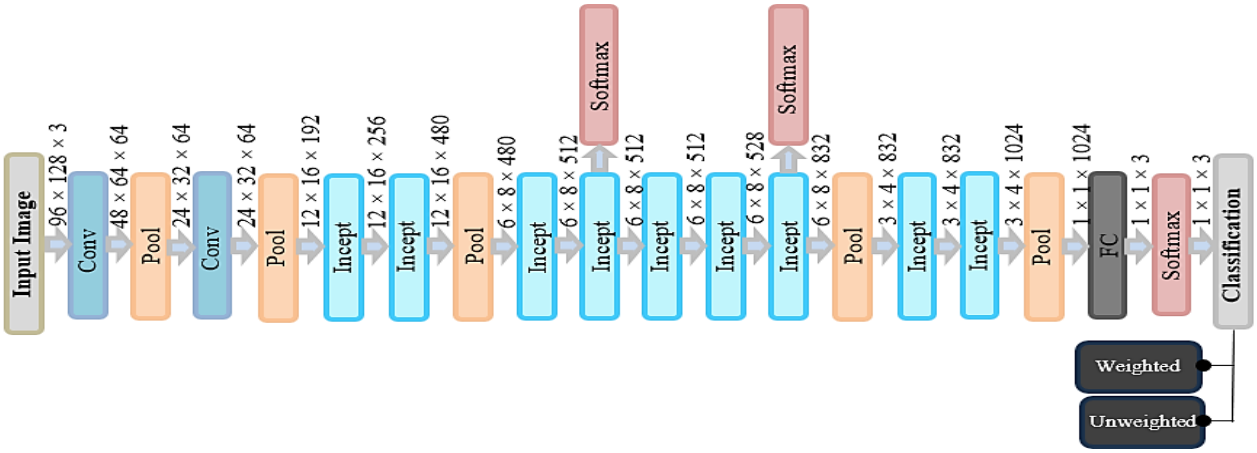
2.2.3. Post-segmentation processing and action classification

This paper classified actions based on human silhouettes to preclude the model's ability to learn individual-specific characteristics rather than human action patterns. *Figure 5* summarizes the image processing and action classification flow. The segmentation process produces a predicted binary mask for each input test image. Although the binary masks are expected to show the borders of a human silhouette, the results may contain some background pixels misclassified as objects and vice versa. Therefore, during the testing and deployment phase, the largest blob (*i.e.*, predicted object pixels) in an image is determined based on the largest area of each connecting pixel using the *regionprops* function in MATLAB. Next, a pruning operation is applied to trim the connectivity length of fewer than 15 pixels, reducing spurious branches and removing unnecessary information. After the cleaning process, the mask is converted into RGB color space, a format compatible with the input of the CNNs, using the concatenation method and resized to 96 × 128 according to the input size of the models in *figure 4* to estimate the performed action.



(a) EfficientNet-B0

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X



(a) GoogleNet

Figure 4. Network architecture of (a) EfficientNet-B0 and (b) GoogleNet with weighted and unweighted cross-entropy loss functions for action

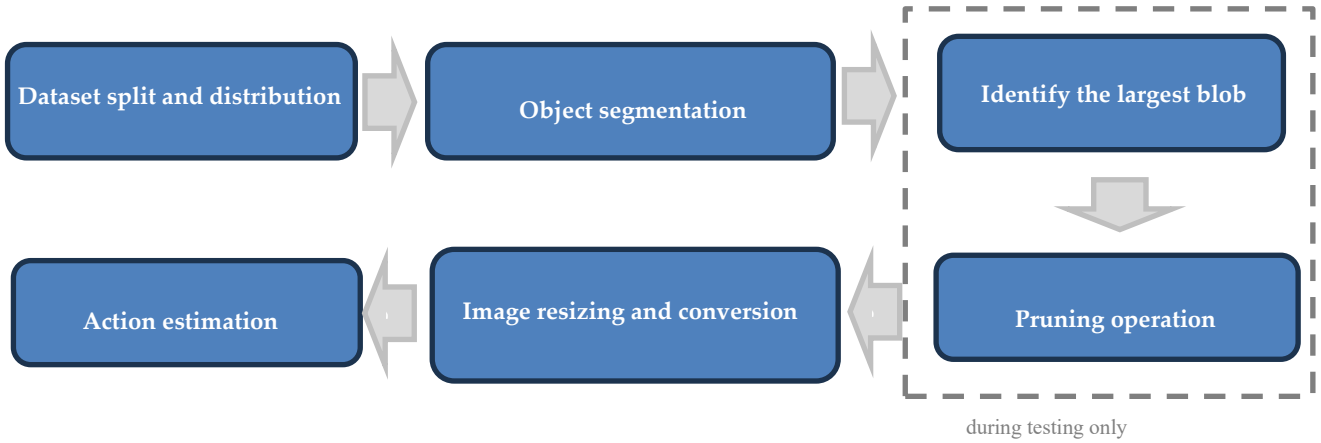


Figure 5. Image processing and analysis flow

2.3. Objective-Based Search

Model training parameters are important determinants of a model's generalization ability. The hyperparameters commonly adjusted to increase model convergence and accuracy include learner type, epoch number, mini-batch size, and initial learning rate. This work used particle swarm optimization (PSO) to search for the optimal hyperparameter combinations by iteratively minimizing the objective function defined in *eq. (3)*.

$$f(T_{acc}, V_{acc}, t) = (1 - T_{acc}) \cdot 10^2 + (1 - V_{acc}) \cdot 10^3 + {}^{t}/_{1000} \quad (3)$$

This search process optimizes the solutions by reducing errors in the training (T_{acc}) and validation accuracies (V_{acc}), and training time (t) at each iteration, where a higher penalty weight has been assigned for validation errors. This objective function is also used to identify the optimal point in the grid search by evaluating the model's performance using predefined solutions.

2.3.1 Particle swarm optimization (PSO)

In this experiment, twenty search particles (possible solution points) were randomly launched into the search space defined in *table 2* to train segmentation and classification models, respectively. The objective function of all particles is evaluated, and the best objective function, p_{best}, is chosen among the neighbors. The position and velocity of each particle i are updated based on the weighted sum of (i) its previous velocity, v_i, (ii) the difference between the best position the particle has seen and its current position, ΔP, and (iii) the difference between global best and particle current position, ΔG, to generate new velocity vector as follows [27]:

$$v_i' = W \cdot v_i + c_1 \cdot u_1 \cdot \Delta P + c_2 \cdot u_2 \cdot \Delta G \quad (4)$$

This work used the default inertia weight range, $W = [0.1, 1.1]$, whereas u_1 and u_2 are uniformly distributed random vectors ranging [0, 1]. c_1 and c_2 are weighting factors for self-adjustment and social adjustment, with the same default value of 1.49. The particle's position is updated based on the old position, x_i, and updated velocity, v_i', given by:

$$x_i' = x_i + v_i' \quad (5)$$

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

The velocity and position of each particle are allowed to be updated ten times in search of the best global solution. Early termination criteria include when cross-validation accuracy failed to increase after 100 evaluations, *i.e.*, when the model is not sufficiently adjusting to new data, or the maximum total iteration number has been reached. Different regularization strategies can be used to prevent overfitting by adding a penalty to the sum of the squares of the model's weights [28]. However, for the sake of simplicity, this paper used ridge regularization, or L2 regularization, with a value of $1e^{-4}$, which is the default regularization method for model training in MATLAB software. A summary of the PSO search algorithm is shown in *figure 6*.

2.3.2. Grid search optimization

This paper also compares PSO search efficiency against the standard manual search. For this purpose, the grid search approach that systematically evaluates each predefined solution is used to optimize the model training. Unlike the PSO, which finds a candidate solution based on the neighborhood information, the grid points (*i.e.*, potential solutions or hyperparameter combination sets) are regularly spaced on the grid, and the current solution does not affect the search process [26].

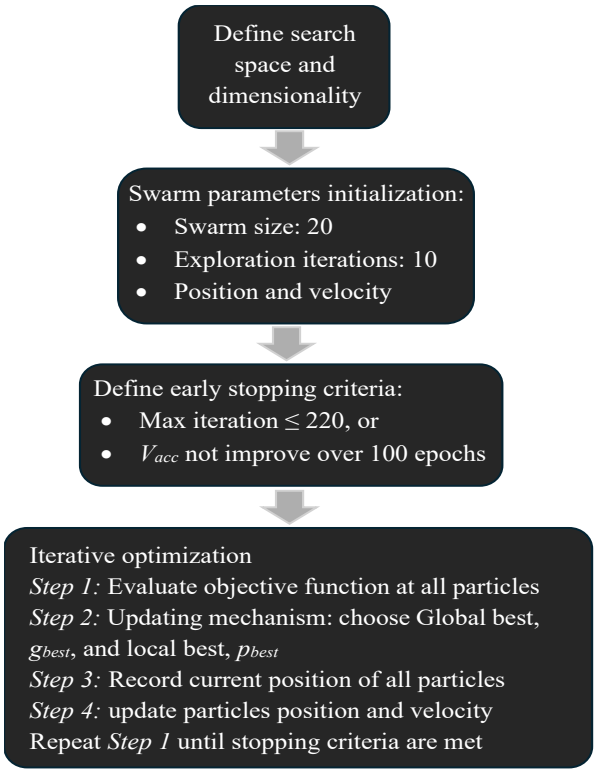


Figure 6. The process flow of the PSO method

Table 2 shows the discrete points considered under the grid search experiment. They are of the same range of hyperparameter values considered in the PSO. Each solution consists of the set of values formed from the hyperparameters in the table, $\{\alpha, \varepsilon, \beta, \psi\}$. This combination process generated 192 possible solutions, comparable to the number of search iterations in the PSO experiment (*i.e.*, 220 solutions). This research design is essential to mitigate potential biases in the analysis and reporting due to the choice of hyperparameters and the number of evaluated solutions. Once all positions are evaluated, the optimal solution is decided based on the combination that produces the lowest objective function value in *eq. (3)*. The same stopping criteria in *section 2.3.1* are used in this experiment for consistency in the comparisons.

Table 2. The search space boundary defined for segmentation and classification problems

Training hyperparameter	PSO Search Space		Grid Search
	Lower limit	Upper limit	Predefined point
Learner type, α	1:3→{*Adam*, *Sgdm*, *RMSProp*}		[1, 2, 3]
Epoch no., ε	50	200	[50, 100, 150, 200]
Mini-batch size, β	8	32^a / 128^b	[8, 16, 24, 32][a]/ [8, 32, 64, 128][b]
Initial learning rate, ψ	$1e^{-4}$	$1e^{-1}$	[$1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$]

[a] segmentation and [b] classification tasks.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

2.4. Performance Evaluation Metrics

This study used various metrics to analyze the efficiency of each network in the cascade system. The quality of segmented maps was tested using Intersection over Union (*IoU*) and Dice similarity (*DS*) in *eqs. (6)* and *(7)* by comparing the areas of ground-truth (*GT*) and predicted (*Pred*) masks for pixel "0" representing the background and "1" the object (human). Meanwhile, the classification performance is evaluated using some of the common metrics. Specifically, classification accuracy (*ACC*), precision (*PREC*), recall or sensitivity (*RECALL*), specificity (*SPEC*), and *F1*-score expressed in *eqs. (8)-(12)*.

$$IoU\ (GT, Pred) = \frac{|GT \cap Pred|}{|GT \cup Pred|} \tag{6}$$

$$DS\ (GT, Pred) = \frac{2(GT \cap Pred)}{GT + Pred} \tag{7}$$

$$ACC = \frac{(TP+TN)}{(FN+FP+TP+TN)} \tag{8}$$

$$PREC = \frac{TP}{(TP+FP)} \tag{9}$$

$$RECALL = \frac{TP}{(TP+FN)} \tag{10}$$

$$SPEC = \frac{TN}{(TN+FP)} \tag{11}$$

$$F1 = \frac{2 \cdot PREC \cdot SENS}{PREC + SENS} \tag{12}$$

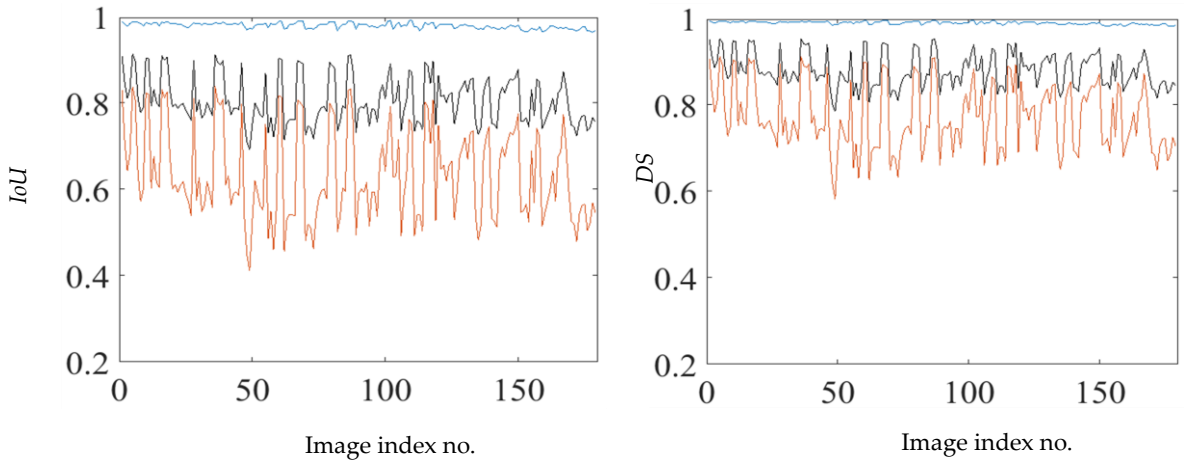
TP and *TN* are the true positive and true negative rates, respectively. *FP* defines the false positive percentage, whereas *FN* denotes a false negative.

3. RESULTS

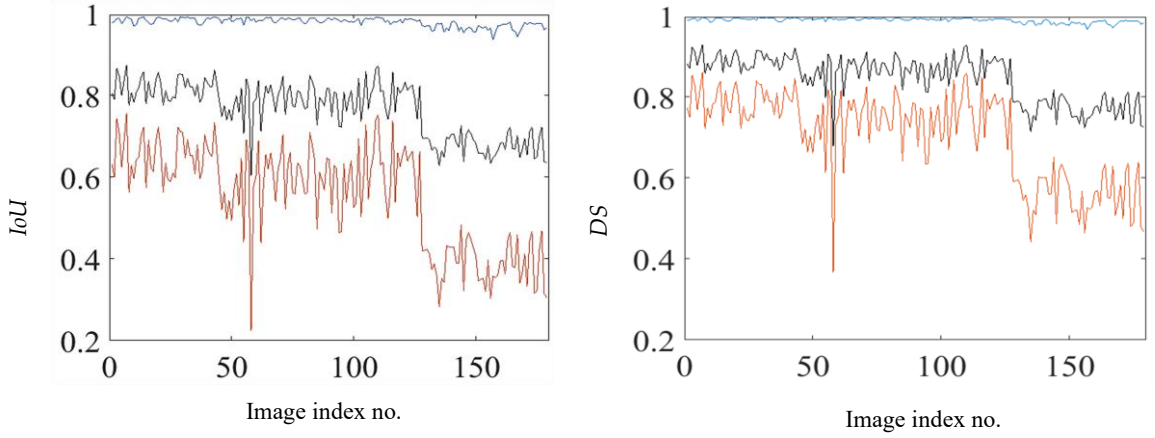
The pixel-level segmentation performance of the PSO and grid search optimized models tested on 179 testing samples is shown in *figure 7*. These models are evaluated using *IoU* and *DS* metrics in terms of their ability to distinguish objects (red lines) from background pixels (blue lines) in an image. Also shown in the figures are the mean values calculated by averaging the accuracies from the object and background pixels segmentation. Based on the means of *IoU* and *DS*, the best and worst performing results are chosen as test image index numbers 5 and 49, respectively, for the PSO optimized U-Net, whereas image index numbers 7 and 58 are identified from the grid search experiment. These segmentation results revealed an overall better performance of the PSO by 5 % compared to the manual counterpart. The segmentation output of the best and worst performing images from the PSO technique is shown in *figure 8* for reference and comparison. Also shown in the diagram are the calculated mean *IoU* and *DS* measures for the predicted masks. The confusion matrix of the classification models tested on the processed segmented masks is shown in *figures 9* and *10*, respectively, for PSO and grid search optimized networks with and without a weighted loss function. The above-mentioned metrics calculated from these confusion matrices are presented in *figure 11*. Results from the PSO experiments are plotted as cross markers (+), whereas the bar graphs show the evaluated performance scores from the grid search.

Since the classification is based on segmentation accuracy, the relationship between the object detection performance and classification accuracy is compared using IBM SPSS Statistics 23.0 with a one-way ANOVA (Analysis of Variance) test. This study tested the relationship between the misclassification probability (using results from all models) of each testing image and its average *IoU* and *DS* scores using a confidence level, σ, of 95 %. PSO and grid search results are compared using the same test. The statistical results show no association between the PSO-optimized segmentation efficiencies and the classification prediction probabilities, with a significance value, ρ, given by 0.650 and 0.925, for testing against mean *IoU* and *DS*, respectively. The same test was repeated to investigate the relationship between the grid search segmentation and classification efficiencies against the *IoU* and *DS* results. The statistical output shows a significant association with an ρ = 0.000 for both tests.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X



(a) Particle swarm optimization (PSO)



(b) Grid-search optimization

Figure 7. The intersection over union (*IoU*) and dice similarity (*DS*) scores of (a) PSO and (b) grid-search optimized system tested on 179 independent test images (indicated by the index number). The blue line represents the *IoU* and *DS* scores of the background class, while the red line represents the object segmentation results. The dark line shows the average performance of the background and object segmentation

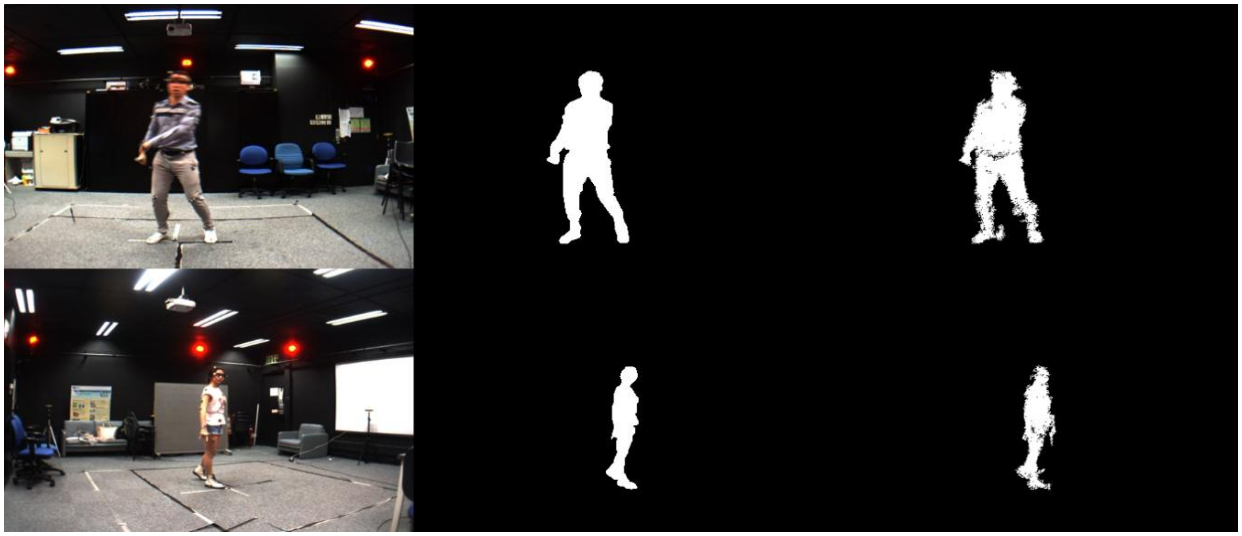


Figure 8. (Top) The best and (bottom) worst performing PSO-optimized segmentation results. Left to right: original image, ground-truth, and predicted mask. Also shown in the diagrams are the calculated intersection over union (*IoU*) and dice similarity (*DS*) scores in detecting the background and the object

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

The PSO search processes found the optimal hyperparameters combinations $\{\alpha, \varepsilon, \beta, \psi\}$ given by {*Adam*, 182, 32, 0.0001} for the segmentation model and {*Sgdm*, 99, 256, 0.078}, {*Sgdm*, 175, 105, 0.0361}, {*RMSProp*, 200, 235, 0.0001}, and {*RMSProp*, 124, 42, 0.0001} for weighted and unweighted loss EfficientNet-B0, and GoogleNet, respectively. Meanwhile, the predefined position of {*RMSProp*, 100, 128, 0.001}, {*Adam*, 150, 64, 0.0001}, {*RMSProp*, 200, 64, 0.001}, and {*Adam*, 100, 32, 0.0001} was identified as the optimal solution in the grid search for weighted and unweighted loss EfficientNet-B0, and GoogleNet, respectively. This manual tuning experiment also determined {*Adam*, 150, 32, 0.0001} as the best hyperparameters for training the U-Net segmentation model.

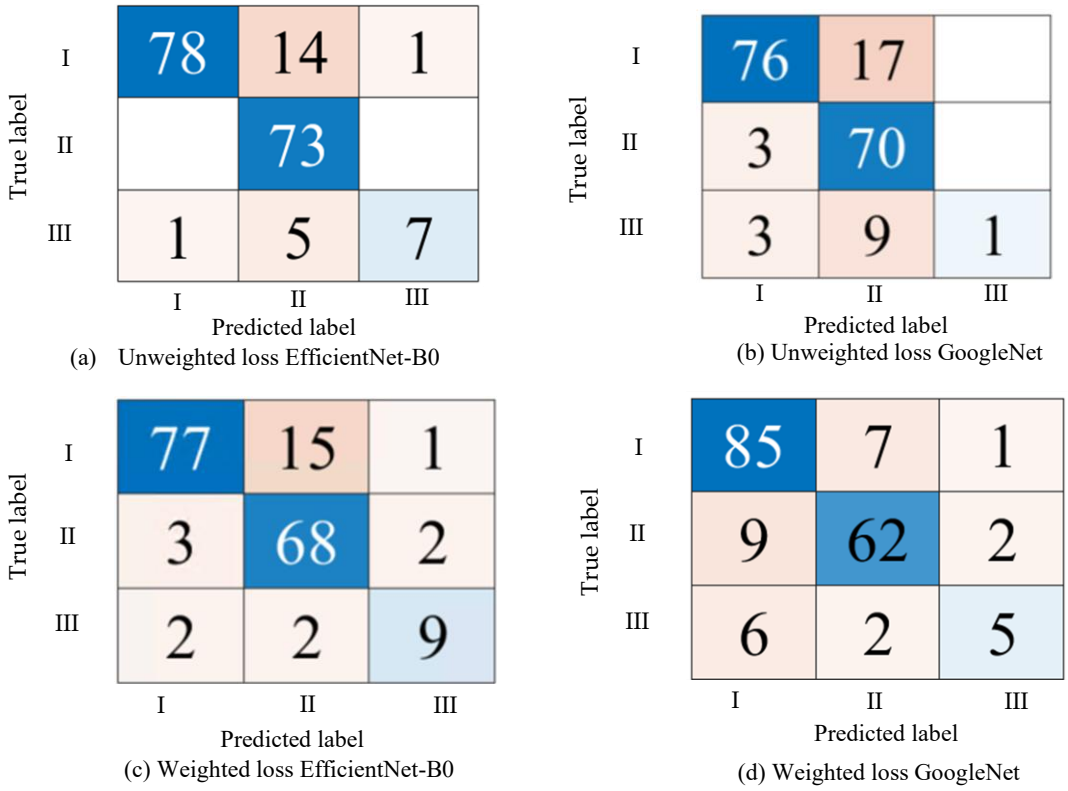


Figure 9. Particle Swarm Optimization (PSO)-optimized EfficientNet and GoogleNet classification performance. Testing confusion matrix of (a) unweighted loss EfficientNet-B0, (b) unweighted loss GoogleNet, (c) weighted loss EfficientNet-B0, and (d) weighted loss GoogleNet. Class label I: dancing, II: doing martial arts, and III: playing net sports

4. DISCUSSION

Physical activity assessment is a crucial aspect of sports rehabilitation, as it improves one's physical fitness while minimizing sports injuries. Some of the traditional methods used in this area of research are commonly associated with sophisticated and complex wearable sensors and devices, often implemented under highly controlled settings to detect actions carried out by individuals. This study proposed using a cascaded deep learning system to detect athletic actions automatically with color photographs as input. This framework, enforced by the PSO and grid-search-based optimization, was shown to produce different classification accuracies for networks with and without a weighted loss function. The segmentation results in *figure 7* showed slightly superior performance of the PSO by 5 % in the evaluated IoU and DS scores compared with the manual grid search. Even though their segmentation performances are comparable, the PSO outperforms the conventional manual search in action classification tasks. The analysis in *figure 11* unveiled a noteworthy reduction by an average of 60 % in all metrics evaluated for the grid search results in *figure 10*. The same search space, as shown in *table 2*. These experimental results showed the insufficiency of the grid search in decision-making and confirmed the efficiency of neighborhood search. Therefore, more emphasis will be on discussing the PSO results in the remainder of this paper.

The results in *figure 7* show that the PSO-optimized two-stage U-Net performed exceptionally well in detecting the background pixels using this dataset, with high *IoU* and *DS* values of 0.98-1. There is a noticeable decrease in the *IoU* and *DS* scores of object detection, ranging between 0.4 and 0.9, but this did not compromise the subsequent classification accuracy, as evidenced by the ANOVA test (ρ = 0.650 and 0.925). The results in *figure 8* (bottom) show that the object's silhouette border can be seen, even in the sample with the lowest average segmentation performance, given by *IoU* = 0.692 and *DS* = 0.785. Nonetheless, this paper does not rule out the possibility that noises within the detected object region may distort the silhouette's shape, leading to a misclassification.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

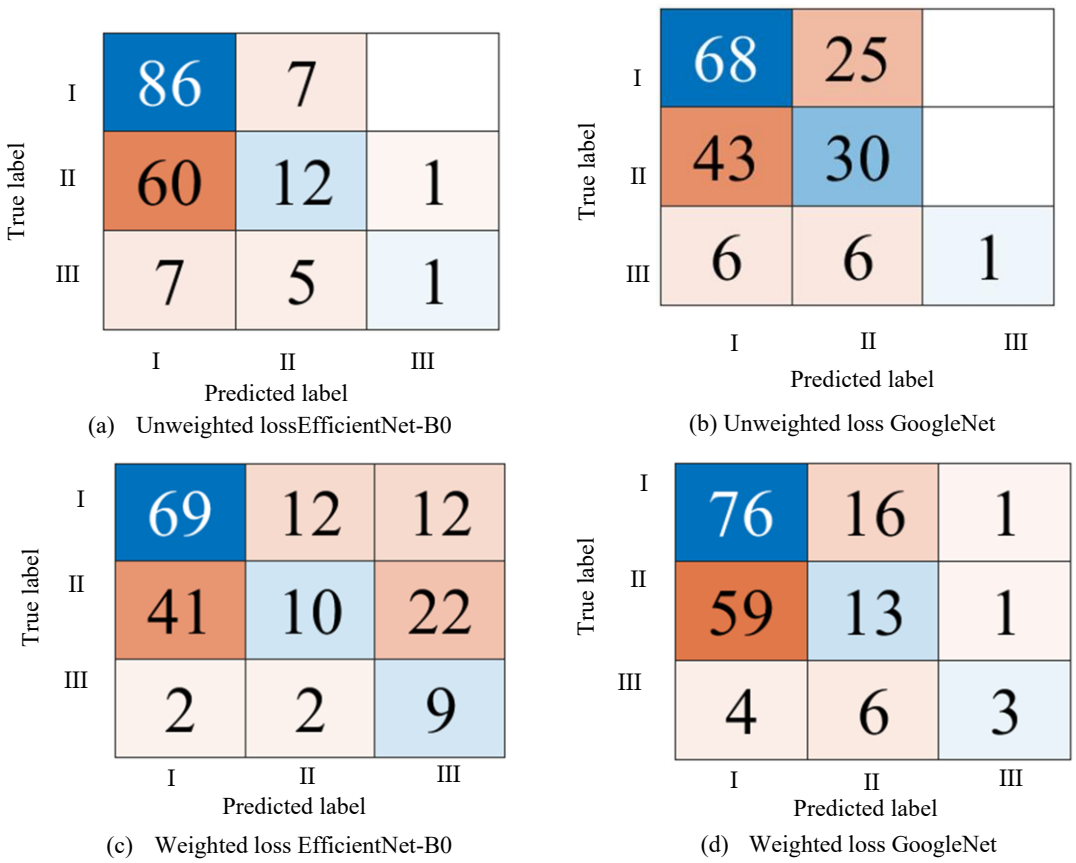


Figure 10. Grid search-optimized networks classification performance. Confusion matrix of classification results of (a) unweighted loss EfficientNet-B0, (b) unweighted loss GoogleNet, (c) weighted loss EfficientNet-B0, and (d) weighted loss GoogleNet. Class label I: dancing, II: doing martial arts, and III: playing net sports

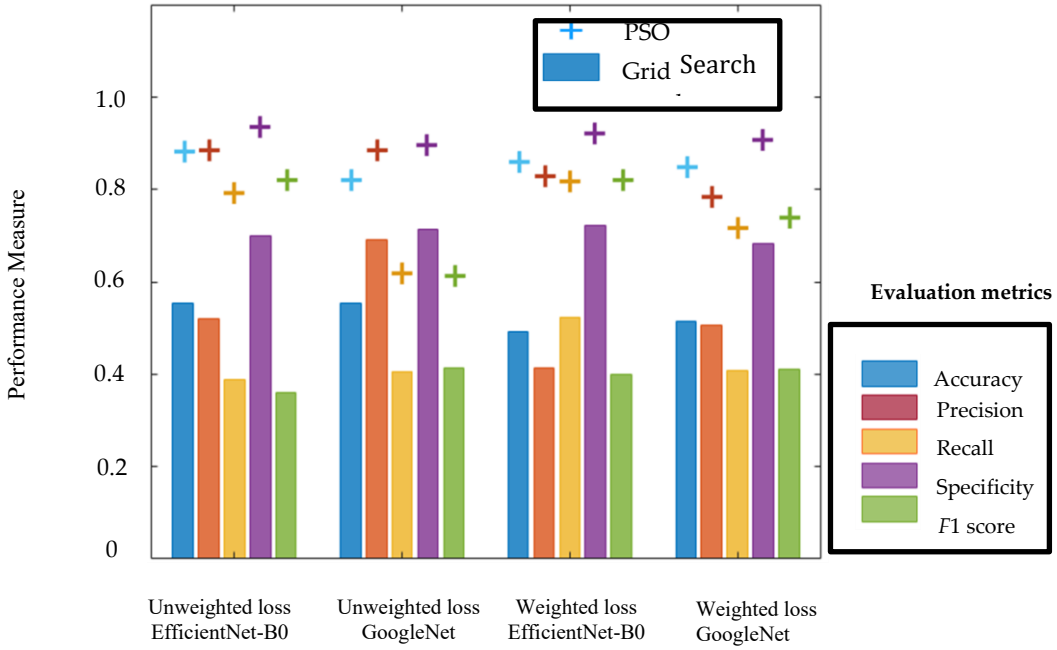


Figure 11. Classification performance measures of Unweighted loss EfficientNet-B0 and GoogleNet, and Weighted loss Efficient-B0 and GoogleNet optimized using Particle Swarm Optimization (PSO) (cross "+" marker) and Grid search (bar graph). The different colors in the plot represent the evaluated performance measures

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

The results in *figure 7* show that the PSO-optimized two-stage U-Net performed exceptionally well in detecting the background pixels using this dataset, with high *IoU* and *DS* values of 0.98-1. There is a noticeable decrease in the *IoU* and *DS* scores of object detection, ranging between 0.4 and 0.9, but this did not compromise the subsequent classification accuracy, as evidenced by the ANOVA test ($\rho = 0.650$ and 0.925). The results in *figure 8* (bottom) show that the object's silhouette border can be seen, even in the sample with the lowest average segmentation performance, given by $IoU = 0.692$ and $DS = 0.785$. Nonetheless, this paper does not rule out the possibility that noises within the detected object region may distort the silhouette's shape, leading to a misclassification.

Small data size is a factor that is well-known for decreasing model learning capacities [16, 19-21], and it is a key challenge identified in this work. This problem is further exacerbated by the imbalanced training samples in *table 1* that can lead to a model overfitting the majority class [22]. This study attempted to overcome this problem by adopting a weighted loss function to impose a higher penalty for inaccurate classifications of the minority class (*i.e.*, playing net sports). The PSO experiments showed that introducing the weighted loss strategy improves both the classification models' true positive and recall rates in *figure 11*; these changes are, however, negligible in the case of grid search, especially when comparing the results of the weighted and unweighted loss GoogleNet in *figure 10(b)* and *(d)*. The results of the PSO-optimized weighted loss networks in *figure 9(c)* and *(d)* show a drop in the number of misclassifications in the minority class (*i.e.*, class label III (playing sports)) compared to the unweighted loss confusion matrix in *figure 9(a)* and *(b)*. Thus, the recall scores were improved by at least 3 % and 15 % in *figure 11* (in yellow cross markers) for the PSO-optimized EfficientNet and GoogleNet, respectively. These effects are evident in GoogleNet, which showed noticeable classification improvements by 0.1 in most evaluated metrics. Since greater emphasis has been placed on the identification performance of minority classes during the training, the sufficiency of the model in learning features of other classes has been reduced. This leads to an increase in the misclassification rates of other actions. Another factor that can also impact the classification results is the efficiency of the optimization process implemented on different models. The PSO optimization results revealed that SGDM is the best solver for EfficientNet-B0 evaluated on this dataset, while RMSProp is determined for GoogleNet. The optimal initial learning rate of 0.0001 is found for GoogleNet with and without a weighted loss function, which may suggest the possibility of premature convergence to a local minimum. This problem, however, is not observed in EfficientNet-B0, implying that PSO works better with EfficientNet-B0. Therefore, this issue can be due to the inadequate feature extraction capability of GoogleNet, rendering the solution at the boundary points.

An investigation into the misclassified samples of the PSO experiment found that eight images failed to be classified correctly by all classification models. Most of which are from class label III (playing net games). This is further confirmed by the statistical analysis showing a strong correlation between the class labels and misclassifications with a $\rho = 0.000$. Thus, variations in the postures between class labels and the increased dataset may support the possibility of improved classification performance. The results showed that complex models like EfficientNet performed better than the residual-based GoogleNet. EfficientNet-B0 can learn better features with dynamicity in network width, depth, and resolution. This is also why this model showed less pronounced differences in the overall model's efficiency with the weight loss strategies, wherein the largest difference is less than 0.05, as shown in *figure 11* (indicated by cross "+" markers).

Most misclassified results consisted of images portraying the object in a stance position and distance away, where the human silhouette is not prominent. An upright stance is a common posture that prepares our body for the next reactions or activities when carrying out different athletic actions. The latter involved postural changes in the limbs, stride, arm swing movements, and body tilting; hence, these coordinated movements can provide important cues for the prediction. However, at certain postures, such as normal stance, the limbs are close to the trunk, depriving the networks of features for action estimation. In that regard, images with a poor angle of view, where human posture cannot be reliably determined, can also contribute to classification errors.

Although not explicitly discussed and shown in this study, a brief and simple experiment has been carried out to explore the efficiency of the proposed system when used on unpredictable and diverse real-world images from Google images. The experiment found this system performs poorly on these images, even when the best performing PSO-optimized U-Net and EfficientNet are used. Most real-world images contain rich scenes and diverse color information and have high background complexity, contrasting with the unique and homogeneous background images used in this research. This explains the insufficiency of the segmentation network in adequately recognizing humans in the images. The poor segmentation results led to low action classification accuracies. Nonetheless, it must be mentioned here that such segmentation tasks may be better accomplished by a network trained on skin detection datasets, such as that demonstrated in one of our previous publications in [29]. Since the classification is based on the detected silhouette, the improved segmentation efficiency would allow the classifiers to process the segmented mask efficiently, improving classification performances.

This study concluded that the proposed image-based strategy has the potential to predict performed athletic actions. However, several shortcomings remain in this work, notably in the limitations of the dataset and analysis methods, which require further research and improvements before this strategy can be applied in practice. Potential efforts include using big datasets to better understand and identify critical features for better computer-aided diagnosis. The same problem has been investigated by [20], who proposed using large, combined datasets, including synthetic data, to enhance model learning. Prayitno [19] proposed a federated learning method by collaborative machine learning training to overcome the limited dataset availability, which is achieved using a central server that

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

coordinates the model's weight and parameters, while ensuring patients' information and medical data remain confidential. These approaches can be considered future work for this paper. The current work depends on the detected human silhouette in the prediction to prevent individual-specific feature recognition. Therefore, another possible area for future research that can be explored is to combine a skeletonized algorithm with a CNN to highlight body part features, including arms, legs, hips, trunk, and head, for action recognition. Such a system can also be extended to field applications to assess physical rehabilitation performances and outcomes. In addition, integrating the system with IoT could facilitate instantaneous monitoring of activities and strengthen the healthcare system.

5. CONCLUSION

This paper proposed using an image-based strategy to overcome the limitations of traditional methods that demand the use of complex and expensive sensors for activity recognition tasks. This framework consisted of a joint segmentation and classification pipeline optimized using the PSO and conventional grid search to classify three athletic events: dancing, martial arts, and playing net sports. The grid search was shown to perform relatively poorly in segmentation and classification tasks compared to its automatic counterpart. The performance of the proposed two-stage PSO-optimized U-Net model evaluated on the independent test samples showed relatively good object detection and segmentation results, with average *IoU* and *DS* scores of 0.85 and 0.9, respectively. The problem of premature convergence was observed in GoogleNet, which is overcome with a weighted loss approach, producing an observable improvement in the classification results compared to EfficientNet-B0. While the change is remarkable, especially in the evaluated recall scores, which increased from 0.6177 to 0.7160 in the PSO-optimized GoogleNet with the weighted loss method, not much difference is observed in the grid search experiment. EfficientNet adopts a compound scaling method during training, leading to more robust, consistent, and accurate feature extraction. Their learning capacity is improved with the PSO. However, the lack of dataset richness hindered their application in real-life scenarios. In conclusion, there is potential to improve the data analysis and signal processing techniques and introduce more efficient body part tracking before extending this technological solution for more cost-effective and real-time assessment of physical activities to improve athlete performance and enhance users' experience.

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

REFERENCES

[1] National Safety Council (2024). Sports and recreational injuries. Available: https://injuryfacts.nsc.org/home-and-community/safety-topics/sports-and-recreational-injuries (accessed on 3 Jan 2025).

[2] Liao, Y.; Vakanski, A.; Xian, M.; Paul, D.; Baker, R. A review of computational approaches for evaluation of rehabilitation exercises. Comput. Biol. Med. 2020, 119.

[3] Ebeling, P.R.; Cicuttini, F.; Scott, D.; Jones, G. Promoting mobility and healthy aging in men: a narrative review. Osteoporos Int. 2019, 30, 1911–1922.

[4] Pan, C.T.; Lee, M.C.; Huang, J.S.; Chang, C.C.; Hoe, Z.Y.; Li, K.M. Active assistive design and multiaxis self-tuning control of a novel lower limb rehabilitation exoskeleton. Machines 2022, 10, 318.

[5] Seçkin, A.Ç.; Ateş, B.; Seçkin, M. Review on wearable technology in sports: concepts, challenges and opportunities. Appl. Sci. 2023, 13, 10399.

[6] Sierotowicz, M.; Connan, M.; Castellini, C. Human-in-the-loop assessment of an ultralight, low-cost body posture tracking device. Sensors 2020, 20, 890.

[7] Yang, Y.; Meng, L. Physical education motion correction system based on virtual reality technology. Int. J. Emerg. Technol. Learn. 2019, 14, 105-116.

[8] Li, C.H.J.; Liang, V.; Chow, Y.T.H.; Ng, H.Y.; Li, S.P. A mixed reality-based platform towards human-cyber-physical systems with iot wearable device for occupational safety and health training. Appl. Sci. 2022, 12, 12009.

[9] Tavakoli, M.; Carriere, J.; Torabi, A. Smart wearable technologies, and autonomous intelligent systems for healthcare during the COVID-19 pandemic: An analysis of the state of the art and future vision. Adv. Intell. Syst. 2020, 2, 2000071.

[10] Qiu, S.; Liu, L.; Zhao, H.; Wang, Z.; Jiang, Y. MEMS inertial sensors-based gait analysis for rehabilitation assessment via multi-sensor fusion. Micromachine 2018, 9, 442.

[11] He, Z.; Liu, T.; Yi, J. A wearable sensing and training system: towards gait rehabilitation for elderly patients with knee osteoarthritis. in IEEE Sens. J. 2019, 19, 5936-5945.

[12] Wang, H. Recognition of wrong sports movements based on deep neural network. IIETA 2020, 34, 663-671.

[13] Zhang, L. Applying Deep Learning-Based Human Motion Recognition System in Sports Competition. Front. Neurorobot. 2022, 16, 1-12.

[14] Hussain, A.; Khan, N.; Munsif, M.; Kim, M. J.; Baik, S. W. Medium Scale benchmark for cricket excited actions understanding. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2024.

[15] Mottaghi, A.; Soryani, M.; Seifi, H. Action recognition in freestyle wrestling using silhouette-skeleton features. Eng. Sci. Technol. Int J. 2020, 23, 921-930.

[16] Xiang, L.; Gao, X. Perceptual feature integration for sports dancing action scenery detection and optimization. in IEEE Access 2024, 12, 122101-122113.

[17] Alavigharahbagh, A.; Hajihashemi, V.; Machado, J.J.M.; Tavares, J.M.R.S. Deep learning approach for human action recognition using a time saliency map based on motion features considering camera movement and shot in video image sequences. Information 2023, 14, 616.

[18] Yuan, Y.; Qin, W.; Ibragimov, B.; Zhang, G.; Han, B.; Meng, M.Q.H.; Xing, L. Densely connected neural network with unbalanced discriminant and category sensitive constraints for polyp recognition. IEEE Trans. Autom. Sci. Eng. 2020, 17, 574–583.

[19] Prayitno, C.R.; Shyu, K.T.; Putra, H.C.; Chen, Y.Y.; Tsai, K.S.; Hossain, M.T.; Jiang, W.; Shae Z.Y. A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications. Appl. Sci. 2021, 11, 11191.

[20] Liu, C.; Zhang, W.; Xu, W.; Lu, B.; Li, W.; Zhao, X. Substation Inspection safety risk identification based on synthetic data and spatiotemporal action detection. Sensors 2025, 25, 2720.

[21] Alshirbaji, A.T; Jalal, N.A.; Docherty, P.D.; Neumuth, T.; Möller, K. Robustness of convolutional neural networks for surgical tool classification in laparoscopic videos from multiple sources and of multiple types: a systematic evaluation. Electronics 2022, 11, 2849.

[22] Tsai, T.H.; Wang, C.Y. Wafer map defect classification using deep learning framework with data augmentation on imbalance datasets. J Image Video Proc. 2025, 6, 2025.

[23] Kunal, K.; Tafeer, A.; Shantanu, G.; Rajat, G. Optimizing CNN hyperparameters for satellite image segmentation: A grid search approach. Int. J. Eng. Res. Manag. 2020, 4, 140–148.

[24] Zhao, J.; Wang, J.; Han, M.; Fan, J. Adaptive particle swarm optimization of u-net parameters for marine aquaculture image segmentation. In 2025 13th International Conference on Intelligent Control and Information Processing (ICICIP), Muscat, Oman, 2025.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 13, Issue 4 | Pages 735-748 | e-ISSN: 2347-470X

[25] Zhang, W.; Liu, Z.; Zhou, L.; Leung, H.; Chan, A.B. Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3D human pose estimation. Image Vis. Comput. 2017, 61, 22-39.

[26] Yamany, W.; Moustafa N.; Turnbull, B. OQFL: An optimized quantum-based federated learning framework for defending against adversarial attacks in intelligent transportation systems. IEEE Trans. Intell. Transp. Syst. 2023, 24, 893-903.

[27] Engelbrecht, A. P. In Computational Intelligence: An Introduction, 2nd ed.; Wiley, New Jersey, U.S., 2007.

[28] Kuang, S.; Wang, L. Identification and analysis of consensus RNA motifs binding to the genome regulator CTCF. NAR Genom Bioinform. 2020, 6.

[29] Huong, A.; Ngu, X. An optimized semantic segmentation framework for human skin detection. Int. J. Integr. Eng. 2024, 16, 293-300.