Computers, Materials & Continua
DOI: 10.32604/cmc.2023.032499
Article

Tech Science Press

Using BlazePose on Spatial Temporal Graph Convolutional Networks for Action Recognition

Motasem S. Alsawadi[1,*], El-Sayed M. El-kenawy[2,3] and Miguel Rio[1]

[1]Electronic and Electrical Engineering Department, University College London, London, WC1E 7JE, England
[2]Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology, Mansoura, 35111, Egypt
[3]Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura, 35712, Egypt
*Corresponding Author: Motasem S. Alsawadi. Emails: motasem.alsawadi.18@ucl.ac.uk, malswadi@kacst.edu.sa
Received: 20 May 2022; Accepted: 21 June 2022

Abstract: The ever-growing available visual data (i.e., uploaded videos and pictures by internet users) has attracted the research community's attention in the computer vision field. Therefore, finding efficient solutions to extract knowledge from these sources is imperative. Recently, the BlazePose system has been released for skeleton extraction from images oriented to mobile devices. With this skeleton graph representation in place, a Spatial-Temporal Graph Convolutional Network can be implemented to predict the action. We hypothesize that just by changing the skeleton input data for a different set of joints that offers more information about the action of interest, it is possible to increase the performance of the Spatial-Temporal Graph Convolutional Network for HAR tasks. Hence, in this study, we present the first implementation of the BlazePose skeleton topology upon this architecture for action recognition. Moreover, we propose the Enhanced-BlazePose topology that can achieve better results than its predecessor. Additionally, we propose different skeleton detection thresholds that can improve the accuracy performance even further. We reached a top-1 accuracy performance of 40.1% on the Kinetics dataset. For the NTU-RGB+D dataset, we achieved 87.59% and 92.1% accuracy for Cross-Subject and Cross-View evaluation criteria, respectively.

Keywords: Action recognition; BlazePose; graph neural network; OpenPose; skeleton; spatial temporal graph convolution network

1 Introduction

According to the Cisco Annual Internet Report, the amount of Machine-to-Machine (M2M) connections is expected to be the fastest-growing category in internet traffic by 2023 [1]. These connection types are related to smart homes and video surveillance applications. As a result, the amount of data related to human action recognition (HAR) will increase accordingly. Additionally, the ever-growing available visual data (i.e., uploaded videos and pictures by internet users) has also

attracted the research community's attention in the computer vision field. Therefore, finding efficient solutions to extract knowledge from these sources is imperative. Not only to help the population make their lives easier but to preserve their security. The most relevant solutions for HAR include the use of sensors attached to the human body while acting and video cameras located in the area of interest where the action is being performed. Although there have been successful applications of sensor-based HAR systems recently [2], we believe that a video-based solution can offer a more robust solution for this task.

HAR from visual data can be defined as a two-fold problem: the first part aims to extract the most relevant features from the video (also known as action representation), and the second part takes those features as input to a classification algorithm to recognize the performed action (action classification) [3]. For action representation, the solutions that have achieved the most success are based upon optical flows [4], point clouds [5], convolutional neural networks (CNN) [6,7] and landmark detection of the main joints of the human body (i.e., skeleton-data) [8]. On the other hand, for action classification, previous attempts vary from random forests [9], to recurrent neural networks (RNN) [10,11] and more recently, graph neural networks (GNN).

The skeleton-based approach for action representation has achieved a remarkable success given its lightweight and robustness [12]. It is easier for a classifier to compute a small set of cartesian coordinates than a much larger set of pixels of an image. Additionally, this representation is background-free. Therefore, the vast amount of noise of the action background is omitted using this approach. This kind of data can be represented as a graph $G = (V, E)$, where V is the set of joints and E is the set of edges that connects each of the skeleton joints (i.e., the bones). Hence, the RGB image sequence that constitutes a video can be represented as a sequence of skeleton frames as it is shown in Fig. 1. With this graph representation on place, a GNN classifier can be applied for action recognition.

Figure 1: Video representation as a skeleton-frame sequence. We extracted the skeleton from each frame of the videos. In this figure, we present a sample taken from the 'dunking basketball' class of Kinetics dataset

A GNN aims to model the relationships between entities [13]. For the purpose of action recognition using the skeleton data, it is intuitively to use a GNN to model the relationship in different joints while performing a movement. To reduce the computation needed, the graph convolutional neural networks (GCN) have been introduced [14]. These architectures generalize what CNN accomplish

in Euclidean spaces: to map a higher dimension input vector into a lower dimensional vector in an embedding space with minor loss of information. This capability has shown a great success in recent years [15]. For these reasons, we strongly believe that a skeleton-based approach using GCN is the more efficient way to achieve an accurate action recognition system.

Nowadays, there are multiple tools to extract the skeleton representation from videos [16–18]. Due to its compatibility with a wide range of commercial video cameras, the OpenPose system has become a reference for this aim. Recently, the Google Research team has released the BlazePose architecture for skeleton extraction from images oriented to mobile devices. However, currently there is a need to explore the capabilities of this library for action recognition tasks.

The first action recognition system using GCN using skeleton data is the Spatial-Temporal Graph Convolutional Network (ST-GCN). This model framework can learn both the spatial and the temporal relations between the set of nodes (i.e., the skeleton joints) during the performance of the action. To the knowledge of the authors, there has not been any previous attempts to use the BlazePose topology as an input for the ST-GCN architecture for action recognition. Most of the previous research have used the OpenPose topology in their studies [19–23] but very few works have been done using the BlazePose alternative [24,25]. Thus, in this study we present the first implementation of the BlazePose skeleton topology using this model for action recognition. To provide a valid comparison with the baseline study in [8], we test our work upon the Kinetics [26] and the NTU-RGB+D [27] benchmarks. To summarize, the contributions of this work are presented in the following list:

- We present the first results of the ST-GCN model using the BlazePose skeleton topology for action recognition.
- We provide a comparison study between the OpenPose and the BlazePose systems for skeleton extraction from videos.
- We show that just by changing the input skeleton topology for an alternative with a larger set of joints, the ST-GCN model can reach better performance with no major increase in the computational load.
- By selecting different skeleton detection thresholds, we demonstrate a positive improvement in the output of the ST-GCN model trained upon an unrestricted environment dataset, such as the Kinetics benchmark.
- Furthermore, we propose an improvement in the BlazePose topology (i.e., the *enhanced-blazepose* topology) that can achieve better results than its predecessor.
- We released the BlazePose skeleton data from Kinetics and NTU-RGB+D datasets to contribute the research community in this field (https://github.com/malswadi/blazepose-skeleton-kinetics-ntu).

The remainder of the paper is structured as follows: in **Section 2** we present previous attempts to enhance the performance of the ST-GCN model. Additionally, we present previous utilization of the BlazePose system upon the ST-GCN architecture with different use-cases. For this purpose, in **Section 3** we provide a brief comparison between the previous utilized system (i.e., OpenPose) and BlazePose for skeleton extraction from videos. In **Section 4** we present the Enhanced-BlazePose topology. We provide the details for its construction and the reasons that motivate us to propose this solution. In **Section 5** we provide an explanation of the details of the chosen framework (i.e., the ST-GCN) for action recognition used in this study. These include the different layers utilized by the model and overview of its construction. Consequently, the experimental settings to achieve the results presented in this study are described in **Section 6**. In **Section 7**, the results obtained are presented and discussed in

depth. Also, we included the possible paths for future works we have found according with our results in this section. Finally, **Section 8** presents the conclusion of our study.

2 Related Work

Recently, the ST-GCN model usage has increased considerably. From stroke type recognition in tennis [28], nurse activity recognition in hospitals [29], stock price prediction [30], fall detection [31] and action recognition systems [32]. Unfortunately, this architecture presents some disadvantages. The main drawbacks include its incapacity to learn the relationship between joints that are far away from each other and it does not consider either the hierarchical structure from GCNNs or the bone information. Consequently, novel improvements to the ST-GCN have been introduced. For example, authors in [33] proposed the two-stream adaptive graph convolutional network (2s-AGCN). In their work, they utilize one of the streams to model the inter-joint relationships and the other to model the inter-bone relationships in a data-driven manner.

Recently, authors in presented the SV-GCN network. This proposal consists in a two-stream architecture to combine the RGB and the skeleton data from videos. In their work, authors in [34] presented the Spatial Temporal Graph Deconvolutional Network (ST-GDN). This model aims to alleviate the noise propagated across the node messages of the skeleton graph by using a deconvolution layer as a filter. In our previous work in, we presented a novel set of partitioning strategies that can capture more accurately the relationship between the joints of the skeleton that outperform the accuracy achieved of the baseline model in [8]. Additionally, multiple work has been done to use attention modules in the model to improve the performance [35,36].

Most of the work aforementioned utilize either the skeleton data provided by NTU-RGB and the OpenPose system, respectively. However, there is only few works done with BlazePose as input for ST-GCN. The use cases found in literature vary from yoga pose recognition [37], posture detection, posture correction [38], teach movements to robots [39], and emotion perception [40]. To the knowledge of the authors, there is no previous usage of the skeleton topology provided by the BlazePose system on the ST-GCN model for action recognition tasks.

3 OpenPose *vs.* BlazePose

The BlazePose system provides more joint information than its predecessors and allows for tracking more accurately. Our intuition is that the increase in the number of joints in the skeleton provided by the BlazePose system with respect to the other skeleton topologies (i.e., OpenPose) will provide additional information that will help to improve the performance in the ST-GCN model for this aim.

The first difference in both systems is the method used to achieve the pose estimation from images. OpenPose uses a bottom-up approach, while BlazePose uses a top-down alternative. The first method localizes the body parts in the image, then maps them into their corresponding person; the second localizes a region of interest where the person is located and then estimates the body's main joints.

To achieve the pose estimation, OpenPose implements Part Affinity Fields (PAFs) scores to estimate the confidence of linkage between each detected body part with a given person in the image [41]. With this solution, the system can see multiple persons in an image by computing each PAFs simultaneously. Thus, the runtime cost is reduced considerably. On the other hand, BlazePose focuses on on-device human pose estimation applications. For that reason, their strategy is oriented to provide a lightweight solution. To achieve that aim, they first detect a region of the body that remains mostly

rigid across all the video frames (the head of the person) using a fast on-device face detector. That information allows it to estimate other body reference landmarks (i.e., the hips and shoulders) of the pose faster.

OpenPose provides different pose formats: the BODY_25 and the COCO [42]. Among these, the COCO format is used most extensively. As the name suggests, the BODY_25 topology provides 25 key points while the COCO provides 18 (shown in Fig. 2a).

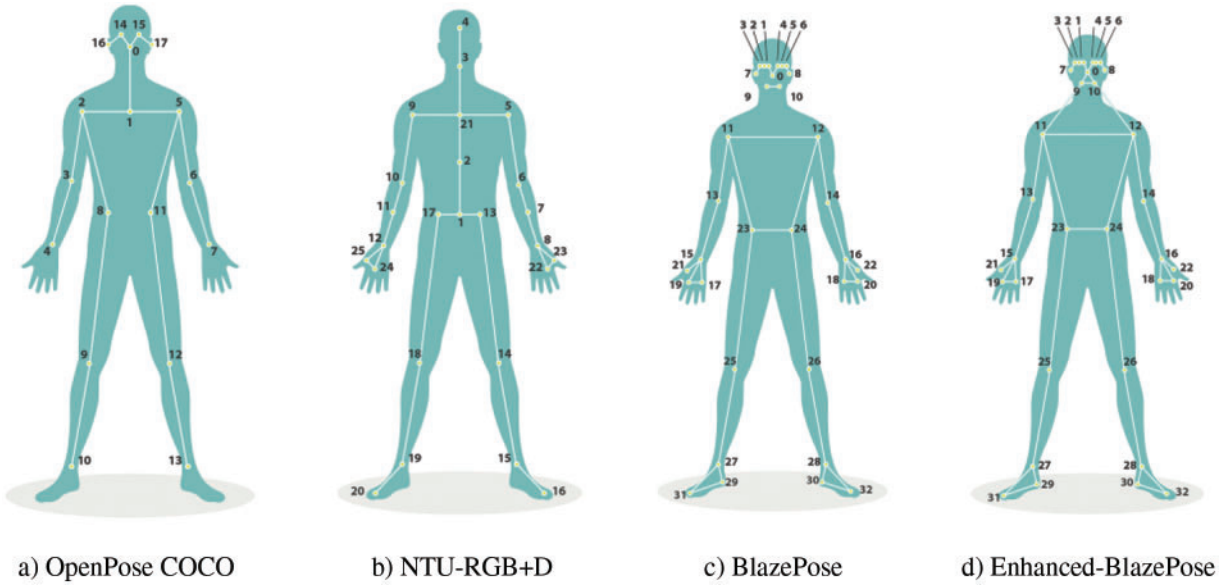


Figure 2: Skeleton topology comparison

The output skeleton provided by the BlazePose system is a superset of COCO [42], the BlazeFace [43] and a BlazePalm consisting of 33 keypoint set. As it is shown in Fig. 2c, this topology can provide a more accurate representation of the hand and the feet movements given the additional keypoints in these parts.

4 Enhanced-BlazePose Proposal

In this study, we propose a novel skeleton topology that can improve the ST-GCN model performance even further. The motivation for the *Enhanced-BlazePose* topology is to provide an even more accurate representation of the actions by adding additional edges to the existing BlazePose topology. Our aim is to capture the relationship between the shoulders and the head during the performance of the actions.

The pseudocode for the topology definition is shown in Fig. 3. In the algorithm, the new edges are included in line 5. In practice, we considered the previous graph definition of the BlazePose topology and included the additional edges that connect the shoulder joints with those in the mouth and nose (shown in Fig. 2d).

We also noticed that the OpenPose system and NTU-RGB+D dataset (shown in Fig. 2b) have all of their joint in their skeletons connected with the rest of the graph. Contrary, the outcome of BlazePose have the joints with indexes 9 and 10 disconnected from the rest (Fig. 2c). After several

experimentation with our novel topology with all its joints connected with the rest of the graph, we show that this proposal provides positive results.

Algorithm 1 Graph

```
graph ←empty hashmap
max_distance ←1
layout ←'blaze_pose'
blaze_pose_edges ←[(0, 1), (1, 2), (2, 3), (3, 7), (0, 4), (4, 5), (5, 6), (6,
8), (9, 10), (11, 12), (11, 13), (13, 15), (15, 17), (15, 19), (15, 21), (17,
19), (12, 14), (14, 16), (16, 18), (16, 20), (16, 22), (18, 20), (11, 23), (12,
24), (23, 24), (23, 25), (24, 26), (25, 27), (26, 28), (27, 29), (28, 30), (29,
31), (30, 32), (27, 31), (28, 32)]
enhanced_blaze_pose_edges ←blaze_pose_edges+[(0, 10), (0, 9), (9, 11),
(10, 12)]
strategy ←'spatial'
no_of_nodes ←33
adjacency_matrix ← zeros(no_of_nodes, no_of_nodes)
for i, j in enhanced_blaze_pose_edges do
    adjacency_matrix[i,j] ←1
    adjacency_matrix[j,i] ←1
end for
for i = 1, 2, . . . ,no_of_nodes do
    sum ← zeros(no_of_nodes)
    for j = 1, 2, . . . ,no_of_nodes do
        sum[i] = sum[i] + adjacency_matrix[i,j]
    end for
end for
for i = 1, 2, . . . ,no_of_nodes do
    for j = 1, 2, . . . ,no_of_nodes do
        adjacency_matrix[i,j] ←adjacency_matrix[i, j]/sum[i]
    end for
end for
neighbor_set = Neighbor sets definition using max_distance and
strategy
graph ←('layout', layout)
graph ←('neighbor_set', neighbor_set)
graph ←('adjacency_matrix', adjacency_matrix)
return graph
```

Figure 3: Pseudocode for enhanced-BlazePose

5 Spatial-Temporal Graph Convolutional Network

The first action recognition system using GCN using skeleton data is the Spatial-Temporal Graph Convolutional Network (ST-GCN) this system, the authors proposed two different sets of joint connections: the *intra-skeleton* and the *inter-frame* connection sets. The first consists of the spatial connections between the joints of the skeleton. For instance, the joints of the knee of the person are connected with those in of the ankle and foot. On the other hand, the inter-frame set of connections include the connections of each joint independently across the sequence of frames of a video.

The architecture of the ST-GCN model consists of an input layer, a sequence of 9 graph convolutional modules (i.e., *ST-GCN unit*) and an output layer with Softmax regression. As shown in Fig. 4, each ST-GCN unit first extracts the relevant features of the intra-skeleton set using a GCN layer. The output of the GCN sub module has the same size as the input. Hence, the output of this sub module is treated as a learned representation of the input skeleton. Consequently, these learnt features are used as an input for a Temporal-GCN (TCN). The TCN sub-module receives the inter-frame information of these sequence of values using a fixed temporal window size. With this approach, the ST-GCN model can extract the spatial and the temporal features of a skeleton sequence to represent and classify an action.

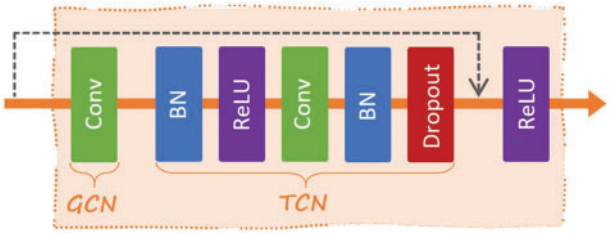


Figure 4: ST-GCN unit. The spatial features of the skeleton frames are computed in the GCN (Graph Convolutional Neural Network) stage, while the temporal features of each joint are processed in the TCN (Temporal-GCN)

The ST-GCN model also propose an additional set of layers (the *learnable edge importance weighting mask*) that aim to learn the importance of each joint during the performance of the actions. With these layers, the model can reach better performance. We considered these layers during our experiments.

6 Experimental Settings

First, we resized each of the videos into dimension of 340 $\times$ 256 pixels upon the Kinetics [26] dataset (This stage was not performed upon the NTU-RGB+D experiments). Second, we extracted the skeleton joint data from the benchmark datasets using the Python API [44] released by the BlazePose team. This tool receives a video as input and returns the horizontal (x), vertical (y) and depth (z) coordinates, in addition with the confidence score (c) of each of the 33 skeleton joints. For the Kinetics dataset, we used the x, y and c data of each joint. Conversely, we utilized the x, y and z information on the NTU-RGB+D dataset.

We did not consider the frames of the videos where the BlazePose system did not detect any skeleton. With this constraint, there were a considerable number of videos that had a small number of frames with a detected skeleton (or no skeleton detection at all in the complete video). This situation motivates us to propose different *skeleton detection thresholds* used as a criterion to select the videos for training. Strictly speaking, we generated two subsets of the BlazePose keypoints dataset: a set containing only those videos where the BlazePose system was able to detect a skeleton in 50% or more of its frames and another containing only the videos where the BlazePose system was able to detect a skeleton in 80% or more of its frames. We denominated these subsets as the *bp-50* (i.e., BlazePose with a skeleton detection threshold of 50%) and *bp-80* (i.e., BlazePose with a skeleton detection threshold of 80%) subsets. Consequently, three versions of the BlazePose keypoint data were used independently in the present study: the output of the BlazePose system with no minimum percentage of skeleton detection per video (*BlazePose* set), the bp-50 and the bp-80 sets.

Third, we fixed the length of the sequence of skeletons to contain only 300 frames. Therefore, if a sequence had less than 300 frames, we repeated the initial frames until reaching the desired length. Conversely, if the sequence had more than 300 frames, we deleted the exceeded frames randomly. Therefore, the spatio-temporal information of the skeleton of each video sample can be represented as a tensor with shape (33, 3, 300).

Finally, we train the ST-GCN model for 80 epochs using the spatial configuration partitioning for joint label mapping. We used the stochastic gradient descent (SGD) with learning rate decay as an optimization algorithm. The initial learning rate value was set to 0.1 with a decay factor of 0.1 every 10^{th} epoch, starting from the epoch number 20. We set a base weight decay value of 0.0001 to avoid overfitting. The batch size was varied from 16 to 256 to achieve the results we present in this study.

All the experiments were performed using the PyTorch [45] framework version 1.2 for deep learning modelling. To train the 256 batch experiments, we utilized 4 NVIDIA Tesla V100 32 GB GPUs in parallel; otherwise, we used 1 NVIDIA Tesla V100 32GB GPU. also, many artificial intelligence applications can be used in real-world problem [46–50].

7 Results and Discussion

In this section we present the performance achieved in terms of accuracy. To validate our hypothesis, we compare our results with those obtained with ST-GCN using the OpenPose topology, and also with different models that proposed modifications in the ST-GCN architecture: the 2s-AGCN, the AM-STGCN and the ST-GDN.

7.1 Kinetics

The Kinetics dataset consists of 306,245 videos corresponding to 400 classes. However, the BlazePose system presented difficulties to detect the skeletons of certain classes. For instance, the 'bartending' and 'cutting watermelon' classes had 0% of skeleton detection performance. That is, that the BlazePose system was not able to provide the skeleton information of any of the samples corresponding to those actions. To illustrate this situation, consider the examples shown in Figs. 5 and 6. It can be noticed that the face of the actor performing those actions is not shown. According to the pose estimation method of the BlazePose system explained in Section 3, the detection of the face is strictly required by this tool to predict accurately the human poses. For this reason, the BlazePose system was not able detect an accurate skeleton from any of the videos corresponding to these actions. Therefore, these classes were not considered in our experiments.

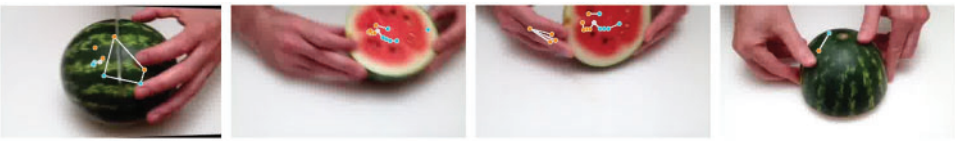


Figure 5: A sample of the BlazePose system output from the Kinetics 'cutting watermelon' class. We extracted 4 randomly selected frames from a video sample and presented them in order of appearance from left to right. As it can be seen, no face was shown about the person performing the action. Therefore, the BlazePose fails to recognize the skeleton joints

Figure 6: A sample of the BlazePose system output from the Kinetics 'bartending' class. We extracted 4 randomly selected frames from a video sample and presented them in order of appearance from left to right. Similarly to 'cutting watermelon' class, no face was shown about the person performing the action. Therefore, the BlazePose also fails to recognize the skeleton joints in this class

On the other hand, the actions 'tasting food' and 'stretching arm' were the classes with the highest skeleton detection. We show two examples of these classes in Figs. 7 and 8. This performance was expected, given that the face is strictly needed to be shown in video while performing these actions. The same situation appears with the other classes with the top performance in the skeleton extraction process.

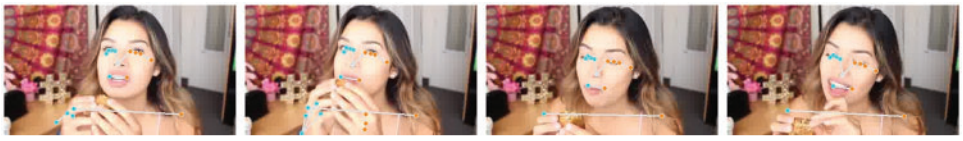


Figure 7: A sample of the BlazePose system output from the Kinetics 'tasting food' class. We extracted 4 randomly selected frames from a video sample and presented them in order of appearance from left to right. This class top class with the highest skeleton detection rate using the BlazePose system. The face of the actors are shown in most of the videos of this class

Figure 8: A sample of the BlazePose system output from the Kinetics 'stretching arm' class. We extracted 4 randomly selected frames from a video sample and presented them in order of appearance from left to right. Similarly to the 'tasting food' class, this class was also one of the top 10 classes with the highest skeleton detection rate using the BlazePose system

The comparison of the top-1 performance of the ST-GCN model obtained from using the OpenPose, BlazePose and Enhanced-BlazePose (E-BlazePose) topologies is shown in Fig. 9. The results in the plot clearly demonstrate that the BlazePose-based topologies outperform the OpenPose alternative. The information provided by the additional joints of the BlazePose topology was able to improve the accuracy performance by almost 3%. Moreover, the new edge information proposed in the e-blazepose topology added almost a 2% accuracy improvement to the action recognition.

The performance obtained with the BlazePose and the Enhanced-BlazePose topologies were improved by using the subsets with higher skeleton detection thresholds. The model trained with the BlazePose topology improved its performance by almost 13% and 14% using the bp-50 and bp-80 subsets, respectively. Furthermore, the model trained upon the Enhanced-BlazePose topology improved its performance by 11% and 14% using the bp-50 and bp-80 subsets, respectively. By using only those videos with skeleton detection in more than 80% of its frames, we improve for almost 17%

the accuracy performance with respect to the results previously achieved with the OpenPose topology. The videos of this dataset were gathered from the internet. Meaning that, there is no constraint that ensures that the body of person performing the action appears correctly inside the frames of the videos. Furthermore, the videos do not have a fixed resolution either. Although there were many frames with no skeleton detection, the different skeleton detection thresholds ensure that the input data provide valuable information and impact dramatically the performance of the output model. These improvements in the results are clearly shown in Fig. 10 and Tab. 1.

Figure 9: Kinetics Training Process. As it can be noticed from one our experiment results shown in the figure, the e-blazepose topology proposed in this study outperforms the results obtained using the alternative topologies

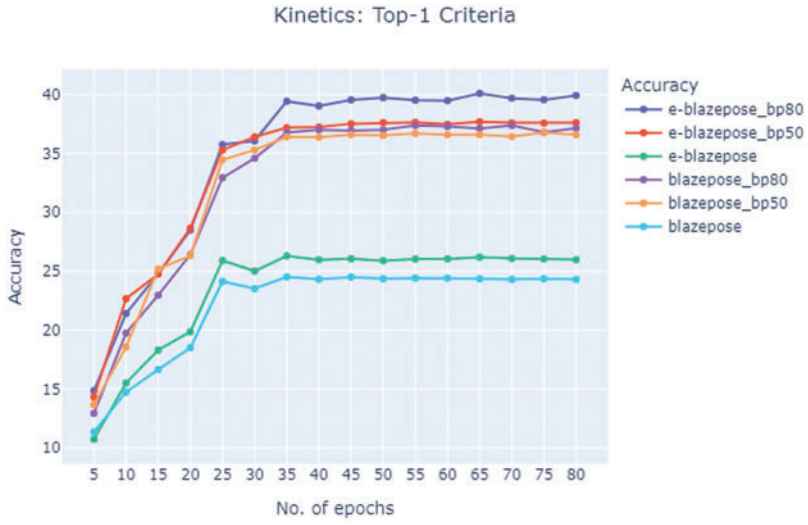


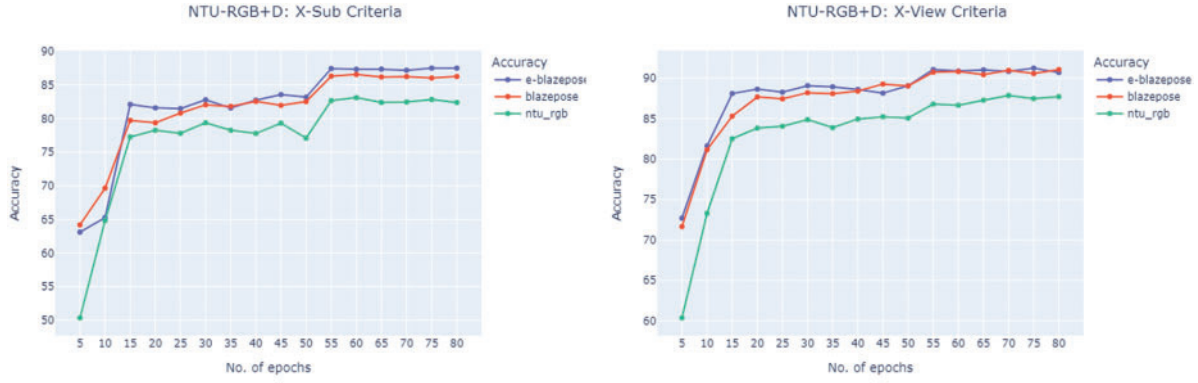
Figure 10: Kinetics performance using different skeleton detection thresholds. This criterion is able to rise the accuracy performance considerably

Table 1: Accuracy performance for ST-GCN-based models on the Kinetics dataset. In the table, 50% st and 80% st correspond to the results obtained using the bp-50 and bp-80 subsets, respectively

Method	Top-1	Top-5
ST-GCN [8]	30.7%	52.8%
AM-STGCN [20]	32.9%	55.4%
2s-AGCN [33]	36.1%	58.7%
ST-GDN [34]	37.3%	60.65%
BlazePose, 50% st	36.78%	61.69%
BlazePose, 80% st	37.38%	65.2%
E-BlazePose, 50% st	37.69%	63.65%
E-BlazePose, 80% st	40.1%	64.73%

7.2 NTU-RGB+D

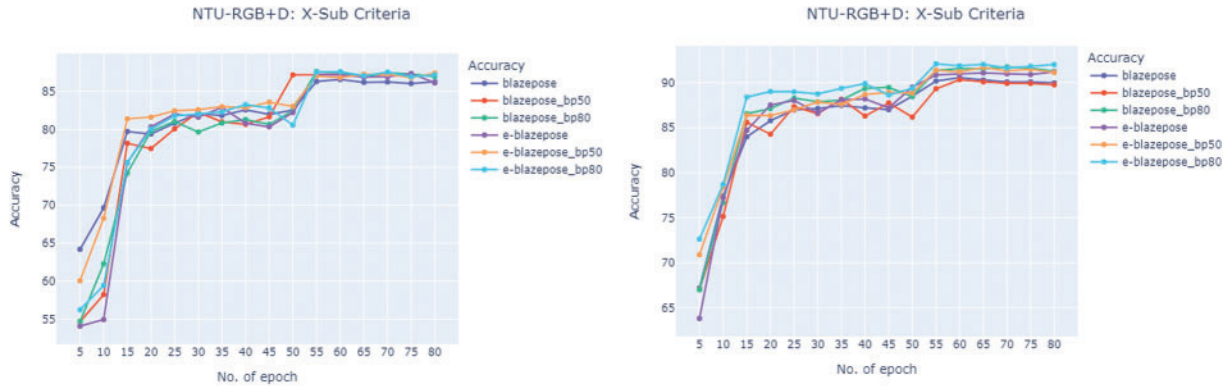
For these experiments, we utilized the Cross-Subject (X-Sub) and Cross-View (X-View) criterias proposed by the dataset authors to evaluate our results. The training process of the model using the X-Sub criteria is shown in Fig. 11a. As it can be noticed, the skeleton information provided by the BlazePose system was able to improve the performance than the NTU-RGB+D skeleton data by more than 3%. Furthermore, by using the Enhanced-BlazePose topology, we were able to improve the BlazePose topology performance by almost 1%. On the other hand, the BlazePose topology was able to improve the NTU-RGB+D data using the X-View criteria by 3.2% (results shown in Fig. 11b). However, the Enhanced-BlazePose topology reached a maximum accuracy only 0.18% higher than BlazePose for this evaluation criteria.



a) b)

Figure 11: NTU-RGB+D Training Performance. (a) X-View evaluation criteria. X-Sub evaluation criteria. In this figure, we compare the performance improvements of different skeleton topologies on the ST-GCN model. The Enhanced-BlazePose (e-blazepose) can reach higher performance that the other alternatives. (b) Similarly, to the X-Sub criteria, we compare the performance improvements of different skeleton topologies on the ST-GCN model. The BlazePose-based skeleton topologies can reach higher performance

The training performance with different skeleton detection thresholds on the NTU-RGB+D dataset is shown in Fig. 12. By examining the performance both evaluation criterias (X-Sub and X-View), it can be noticed that there is no significant improvement in using the subsets with higher skeleton detection thresholds for this dataset. The reason of this is due to the fact that the NTU-RGB+D dataset consists of videos recorded in a restricted environment. Consequently, the BlazePose system was able to detect a skeleton in the vast majority of frames in the videos. Therefore, all three versions of the BlazePose keypoint datasets (i.e., the *BlazePose*, *bp-50* and *bp-80*) contain almost the same information. Hence, the results variance is mainly caused because of the random initialization and learning process of the final models; not because of major differences in the input data.



a) X-Sub b) X-View

Figure 12: NTU-RGB+D Training Performance using different skeleton detection thresholds. This criteria does not provide a significant improvement on restricted environments

We present a comparison with previous ST-GCN enhancement models in Tab. 2. The results show that the ST-GDN model provide a more significant improvement that different skeleton detection threshold for restricted environments. However, the results achieved with our alternative input topologies outperform the ST-GCN model performance in where no alterations in the architecture has been made (Figs. 9 and 11). Therefore, this motivate us to hypothesize that the performance of the ST-GDN model can be further improved if the BazePose-based topologies are used as an input.

Table 2: NTU-RGB+D accuracy results. In the table, 50% st and 80% st correspond to the results obtained using the bp-50 and bp-80 subsets, respectively

Method	X-View	X-Sub
ST-GCN [8]	81.5%	88.3%
AM-STGCN [20]	83.34%	91.4%
2s-AGCN [33]	88.8%	95.1%
ST-GDN [34]	89.7%	95.9%
BlazePose, 50% st	87.3%	90.34%
BlazePose, 80% st	87.62%	91.75%
E-BlazePose, 50% st	87.42%	91.69%
E-BlazePose, 80% st	87.59%	92.1%

Finally, we have found that the bottom-up approach with PAFs used by OpenPose makes it ideal in use-cases where multiple persons are interacting with each other in a particular action. However, this system is not as portable as the BlazePose system and it requires a considerable computation capacity to operate. Conversely, the BlazePose system is unable to track the person performing the action when there are multiple persons in the area. For instance, if a second person is located nearer to the camera than the actor, the system gives the second person a higher priority. Therefore, it changes the tracking of the skeleton joints from one actor to the other. However, the easy-to-use API provided by the BlazePose team allows it to be a practical solution for lightweight applications (such as those mobile-oriented) when a single actor is involved. Regarding the computation load required during training and inference using the ST-GCN model, there is not a significant increment in the cost between choosing the BlazePose topology over the OpenPose alternative. This is because the changes in the number of joints are only reflected in the input layer and the importance weighting layers, while the rest of the hidden layers remain the same. For instance, the trainable parameters needed to train a ST-GCN using the BlazePose skeleton ascend to 3.2 M; while the trainable parameters needed to train this model using the OpenPose topology are 3.17 M. As a result, the changes in the input skeleton only represent an increment of 0.71% in the trainable parameters.

7.3 Statistical Analysis

From the statistical perspective, the achieved results were analyzed to confirm the findings of the proposed approach. Tab. 3 presents a statistical analysis for the achieved results of bp-0 and ebp-0. On the other hand, one sample t-test is performed and the results are recorded in Tab. 4. The results presented in these tables emphasize the stability and effectiveness of the proposed approach.

Table 3: Statistical analysis of the achieved results

	bp-0	ebp-0
Number of values	36	36
Minimum	0	0
25% Percentile	0.3997	0.2458
Median	0.6159	0.5353
75% Percentile	0.8695	0.7415
Maximum	1	1
Range	1	1
95% CI of median		
Actual confidence level	97.12%	97.12%
Lower confidence limit	0.4562	0.2524
Upper confidence limit	0.7827	0.6532
Mean	0.6158	0.4922
Std. Deviation	0.2632	0.286
Std. Error of Mean	0.04387	0.04766
Lower 95% CI of mean	0.5267	0.3954
Upper 95% CI of mean	0.7049	0.589
Coefficient of variation	42.74%	58.10%
Quadratic mean	0.6683	0.5672

(Continued)

Table 3: Continued

	bp-0	ebp-0
Lower 95% CI of quad. mean	0.5839	0.4737
Upper 95% CI of quad. mean	0.7431	0.6474
Skewness	−0.3097	−0.0408
Kurtosis	−0.7527	−1.186
Sum	22.17	17.72

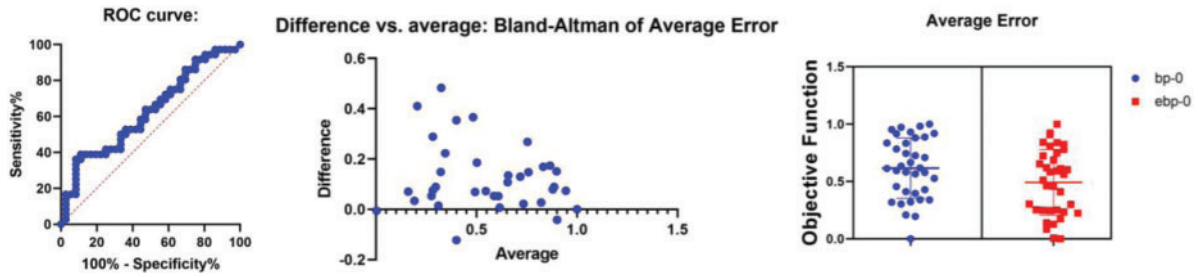
Table 4: One sample t-test for assessing the achieved results

	bp-0	ebp-0
Theoretical mean	0	0
Actual mean	0.6158	0.4922
Number of values	36	36
One sample t test		
t, df	t = 14.04, df = 35	t = 10.33, df = 35
P value (two tailed)	<0.0001	<0.0001
P value summary	****	****
Significant (alpha=0.05)?	Yes	Yes
How big is the discrepancy?		
Discrepancy	0.6158	0.4922
SD of discrepancy	0.2632	0.286
SEM of discrepancy	0.04387	0.04766
95% confidence interval	0.5267 to 0.7049	0.3954 to 0.5890
R squared (partial eta squared)	0.8492	0.7529

On the other hand, Fig. 13 depicts a visual analysis of the achieved results. In this figure, the ROC is shown in Fig. 13a to prove the accuracy of the propose approach. In addition, the difference *vs.* average mapping is shown in Fig. 13b. This figure confirms the robustness of the proposed approach. Finally, the average error *vs.* the objective function is shown in Fig. 13c. Overall, this analysis emphasizes the efficiency of the proposed approach.

7.4 Future Research and Open Challenges

Given the advantages of the human pose estimation approaches used by the OpenPose and BlazePose systems, we propose a development on a hybrid technique based upon them as a future work. This could use an object detector to predict a fixed part of the human body as a reference for fast inference of the person localization (BlazePose approach) and using PAFs for an accurate joint prediction from there (OpenPose solution). Another path could be to propose a modification in the framework such that we could find the optimal skeleton detection threshold in a data-driven manner for HAR recognition purposes. One possibility is to add an extra layer in early stages. The positive results we achieved using different thresholds on data recorded in unrestricted environments, encourage us to foresee this as future work.



a) b) c)

Figure 13: Visual plots for analyzing the achieved results

Regarding the Enhanced-BlazePose topology proposed in this study, it is unable to provide additional information were the skeleton joints from where the new set of edges are constructed (i.e., joints with indexes 0, 9, 10, 11 or 12) are not accurately predicted by the BlazePose system (either for obstruction or a weak inference). In this scenarios, no additional information is provided with respect to the baseline topology (BlazePose). This can be solved by considering the values of the joints surrounding each eye to connect the head with the shoulders. With this new proposal, the additional edges will not depend on any of the individual joints.

Additionally, we intend to evaluate the experiments in this study using different partitioning strategies. In particular, using the full distance, connection and index splits proposed in previous work.

8 Conclusion

In this research, we present the first implementation of the BlazePose skeleton topology upon the ST-GCN architecture for action recognition. To provide a valid comparison with the base model in, we have selected Kinetics and NTU-RGB+D benchmark datasets. We proposed the use of different skeleton detection thresholds that can rise the model performance when the visual data has been recorded in unrestricted environments. Finally, we have provided a comparison study between the OpenPose and the BlazePose systems for skeleton data extraction. We have shown that the supplementary information of the feet and hands of the BlazePose topology does allows this tool to provide a more accurate information about the action performed. Moreover, the Enhanced-BlazePose topology proposed in this study can achieve even higher performance.

Acknowledgement: The authors acknowledge the support of King Abdulaziz City of Science and Technology.

Funding Statement: The authors received no specific funding for this study.

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

References

[1] Cisco, "Cisco annual internet report (2018–2023)," in White Pap. Cisco public, 1–35, 2018. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf.

[2] O. Nafea, W. Abdul, G. Muhammad and M. Alsulaiman, "Sensor-based human activity recognition with spatio-temporal deep learning," *Sensors*, vol. 21, no. 6, pp. 2141, 2021.
[3] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv:1806.11230*, 2018. [Online]. Available: https://arxiv.org/abs/1806.11230.
[4] K. Kinoshita, M. Enokidani, M. Izumida and K. Murakami, "Tracking of a moving object using one-dimensional optical flow with a rotating observer," in *9th Int. Conf. on Control, Automation, Robtics and Vision*, Singapore, pp. 1–6, 2007.
[5] H. Fan, X. Yu, Y. Ding, Y. Yang and M. Kankanhalli, "PSTNET: Point spatio-temporal convolution on point cloud sequences," in *Int. Conf. on Learning Representations*, Virtual, pp. 1–24, 2020.
[6] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun *et al.,* "A closer look at spatiotemporal convolutions for action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6450–6459, 2018.
[7] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.
[8] S. Yan, Y. Xiong and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *32nd AAAI Conf. on Artificial Intelligence*, New Orleans, Louisiana, USA, pp. 7444–7452, 2018.
[9] U. M. Nunes, D. R. Faria and P. Peixoto, "A human activity recognition framework using max-min features and key poses with differential evolution random forests classifier," *Pattern Recognit. Lett.*, vol. 99, pp. 21–31, 2017.
[10] J. Liu, G. Wang, P. Hu, L. Duan and A. C. Kot, "Global context-aware attention lstm networks for 3D action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1647–1656, 2017.
[11] G. Yao, X. Liu and T. Lei, "Action recognition with 3D ConvNet-GRU architecture," in *Proc. of the 3rd Int. Conf. on Robotics, Control and Automation*, Chengdu, China, pp. 208–213, 2018.
[12] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang *et al.,* "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3595–3603, 2019.
[13] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu *et al.,* "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.
[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv Preprint arXiv1609.02907*, 2016. [Online]. Available: https://arxiv.org/abs/1609.02907.
[15] J. Xiang, X. Hu and J. Ou, "Action recognition network based on temporal spatial temporal mode," in *2020 IEEE 5th Int. Conf. on Signal and Image Processing (ICSIP)*, Nanjing, China, pp. 298–302, 2020.
[16] H. Sun and R. Grishman, "Lexicalized dependency paths based supervised learning for relation extraction," *Computer Systems Science and Engineering*, vol. 43, no. 3, pp. 861–870, 2022.
[17] Z. Cao, G. Hidalgo, T. Simon, S. Wei and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.
[18] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang *et al.,* "BlazePose: On-device real-time body pose tracking," in arXiv: 2006. 10204, 2020. [Online]. Available: https://arxiv.org/abs/2006.10204.
[19] L. Shi, Y. Zhang, J. Cheng and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 12026–12035, 2019.
[20] Y. Kong, L. Li, K. Zhang, Q. Ni and J. Han, "Attention module-based spatial-temporal graph convolutional networks for skeleton-based action recognition," *Journal of Electronics and Imaging*, vol. 28, no. 4, pp. 43032, 2019.
[21] Y. Liu, R. Ma, H. Li, C. Wang and Y. Tao, "RGB-D human action recognition of deep feature enhancement and fusion using two-stream ConvNet," *Journal of Sensors*, vol. 2021, no. 1, pp. 1–10, 2021.

[22] C. Yang, A. Setyoko, H. Tampubolon and K. Hua, "Pairwise adjacency matrix on spatial temporal graph convolution network for skeleton-based two-person interaction recognition," in *2020 IEEE Int. Conf. on Image Processing (ICIP)*, Virtual, pp. 2166–2170, 2020.
[23] M. S. Alsawadi and M. Rio, "Skeleton-Split framework using spatial temporal graph convolutional networks for action recognition," in *2021 4th Int. Conf. on Bio-Engineering for Smart Technologies (BioSMART)*, Paris, France, pp. 1–5, 2021.
[24] A. S. B. Pauzi, F. B. Mohd Nazri, S. Sani, A. M. Bataineh, M. N. Hisyam *et al.,* "Movement estimation using mediapipe BlazePose," in *7th Int. Visual Informatics Conf., IVIC 2021*, Kajang, Malaysia, pp. 562–571, 2021.
[25] A. Kulikajevas, R. Maskeliūnas, R. Damaševičius, J. Griškevičius and K. Daunoravičienė, "Exercise abnormality detection using BlazePose skeleton reconstruction," in *Int. Conf. on Computational Science and Its Applications*, Cagliari, Italy, pp. 90–104, 2021.
[26] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier *et al.,* "The kinetics human action video dataset," in arXiv: 1705. 06950, 2017. [Online]. Available: https://arxiv.org/abs/1705.06950.
[27] A. Shahroudy, J. Liu, T. Ng and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 1010–1019, 2016.
[28] M. Skublewska-Paszkowska, P. Powroznik and E. Lukasik, "Learning three dimensional tennis shots using graph convolutional networks," *Sensors*, vol. 20, no. 21, pp. 6094, 2020.
[29] X. Cao, W. Kudo, C. Ito, M. Shuzo and E. Maeda, "Activity recognition using ST-GCN with 3D motion data," in *Adjunct Proc. of the 2019 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing and Proc. of the 2019 ACM Int. Symp. on Wearable Computers*, London, United Kingdom, pp. 689–692, 2019.
[30] X. Hou, K. Wang, C. Zhong and Z. Wei, "ST-Trader: A spatial-temporal deep neural network for modeling stock market movement," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 5, pp. 1015–1024, 2021.
[31] Y. M. Galvão, L. Portela, J. Ferreira, P. Barros, O. A. D. A. Fagundes *et al.,* "A framework for anomaly identification applied on fall detection," *IEEE Access*, vol. 9, pp. 77264–77274, 2021.
[32] Y. Jiang, K. Song and J. Wang, "Action recognition based on fusion skeleton of two kinect sensors," in *2020 Int. Conf. on Culture-oriented Science & Technology (ICCST)*, Beijing, China, pp. 240–244, 2020.
[33] M. S. Alsawadi and M. Rio, "Skeleton split strategies for spatial temporal graph convolution networks," *Computers, Materials & Continua*, vol. 71, no. 3, pp. 4643–4658, 2022.
[34] E. -S. M. El-Kenawy, S. Mirjalili, F. Alassery, Y. Zhang, M. Eid *et al.,* "Novel meta-heuristic algorithm for feature selection, unconstrained functions and engineering problems," *IEEE Access*, vol. 10, pp. 40536–40555, 2022.
[35] N. Heidari and A. Iosifidis, "On the spatial attention in spatio-temporal graph convolutional networks for skeleton-based human action recognition," in *2021 Int. Joint Conf. on Neural Networks (IJCNN)*, Virtual, pp. 1–7, 2021.
[36] A. Abdelhamid, E.-S. M. El-kenawy, B. Alotaibi, M. Abdelkader, A. Ibrahim *et al.,* "Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm," *IEEE Access*, vol. 10, pp. 49265–49284, 2022.
[37] U. Bahukhandi and S. Gupta, "Yoga pose detection and classification using machine learning techniques," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 12, pp. 186–191, 2021.
[38] D. Neogi, N. Das and S. Deb, "FitNet: A deep neural network driven architecture for real time posture rectification," in *2021 Int. Conf. on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Bahrain, pp. 354–359, 2021.
[39] A. Abdelhamid and S. Alotaibi, "Optimized two-level ensemble model for predicting the parameters of metamaterial antenna," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 917–933, 2022.
[40] D. Sami Khafaga, A. Ali Alhussan, E. M. El-kenawy, A. E. Takieldeen, T. M. Hassan *et al.,* "Meta-heuristics for feature selection and classification in diagnostic breast cancer," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 749–765, 2022.

[41] Z. Cao, T. Simon, S. Wei and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, pp. 1302–1310, 2017.
[42] D. Sami Khafaga, A. Ali Alhussan, E. M. El-kenawy, A. Ibrahim, S. H. Abd Elkhalik *et al.,* "Improved prediction of metamaterial antenna bandwidth using adaptive optimization of LSTM," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 865–881, 2022.
[43] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran and M. Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," arXiv Preprint arXiv1907. 05047, 2019.
[44] MediaPipe, 2022. [Online]. Available: https://mediapipe.dev.
[45] A. Abdelhamid and S. R. Alotaibi, "Robust prediction of the bandwidth of metamaterial antenna using deep learning," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 2305–2321, 2022.
[46] E.-S. M. El-Kenawy, A. Ibrahim, S. Mirjalili, Y. Zhang, S. Elnazer *et al.,* "Optimized ensemble algorithm for predicting metamaterial antenna parameters," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 4989–5003, 2022.
[47] A. Ibrahim, H. Abutarboush, A. Wagdy, M. Fouad and E. -S. M. El-kenawy, "An optimized ensemble model for prediction the bandwidth of metamaterial antenna," *Computers, Materials & Continua*, vol. 71, pp. 199–213, 2021.
[48] E. -S. M. El-kenawy, A. Ibrahim, N. Bailek, B. Kada, M. Hassan *et al.,* "Sunshine duration measurements and predictions in saharan algeria region: An improved ensemble learning approach," *Theoretical and Applied Climatology*, vol. 147, no. 3–4, pp. 1015–1031, 2022.
[49] A. Salamai, A. Ageeli and E. -S. M. El-kenawy, "Forecasting E-commerce adoption based on bidirectional recurrent neural networks," *Computers, Materials and Continua*, vol. 70, no. 1, pp. 10. 32604, 2021.
[50] E. -S. M. El-kenawy, A. Ibrahim, S. Mirjalili, Y. Zhang, S. Elnazer *et al.,* "Optimized ensemble algorithm for predicting metamaterial antenna parameters," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 4989–5003, 2022.