electronics

MDPI

Article

Cyber-Physical System Security Based on Human Activity Recognition through IoT Cloud Computing

Sandesh Achar [1,*,†], Nuruzzaman Faruqui [2,*,†], Md Whaiduzzaman [3,*,†], Albara Awajan [4] and Moutaz Alazab [4]

[1] Walmart Global Tech, Sunnyvale, CA 94086, USA
[2] Department of Software Engineering, Daffodil International University, Daffodil Smart City, Dhaka 1216, Bangladesh
[3] School of Information Systems, Faculty of Science, Queensland University of Technology, Brisbane 4000, Australia
[4] Intelligent Systems Department, Al-Balqa Applied University, Al-Salt 19117, Jordan
* Correspondence: sandeshachar26@gmail.com (S.A.) faruqui.swe@diu.edu.bd (N.F.); md.whaiduzzaman@qut.edu.au (M.W.)
† These authors contributed equally to this work.

Abstract: Cyber-physical security is vital for protecting key computing infrastructure against cyber attacks. Individuals, corporations, and society can all suffer considerable digital asset losses due to cyber attacks, including data loss, theft, financial loss, reputation harm, company interruption, infrastructure damage, ransomware attacks, and espionage. A cyber-physical attack harms both digital and physical assets. Cyber-physical system security is more challenging than software-level cyber security because it requires physical inspection and monitoring. This paper proposes an innovative and effective algorithm to strengthen cyber-physical security (CPS) with minimal human intervention. It is an approach based on human activity recognition (HAR), where GoogleNet–BiLSTM network hybridization has been used to recognize suspicious activities in the cyber-physical infrastructure perimeter. The proposed HAR-CPS algorithm classifies suspicious activities from real-time video surveillance with an average accuracy of 73.15%. It incorporates machine vision at the IoT edge (Mez) technology to make the system latency tolerant. Dual-layer security has been ensured by operating the proposed algorithm and the GoogleNet–BiLSTM hybrid network from a cloud server, which ensures the security of the proposed security system. The innovative optimization scheme makes it possible to strengthen cyber-physical security at only USD 4.29 ± 0.29 per month.

Keywords: cyber-physical security; human activity recognition; GoogleNet; BiLSTM; deep learning; algorithm

check for updates

Citation: Achar, S.; Faruqui, N.; Whaiduzzaman, M.; Awajan, A.; Alazab, M. Cyber-Physical System Security Based on Human Activity Recognition through IoT Cloud Computing. *Electronics* **2023**, *12*, 1892. https://doi.org/10.3390/electronics12081892

Academic Editors: Nikos Petrellis, Nikolaos Voros and Christos P. Antonopoulos

Received: 9 March 2023
Revised: 2 April 2023
Accepted: 13 April 2023
Published: 17 April 2023

1. Introduction

The field of cyber security that deals with the security of physical computing devices is called cyber-physical security. A wide range of devices, for example, desktops, laptops, servers, network switches, routers, the Internet of Things (IoT), etc., fall under the category of cyber-physical systems. As a matter of fact, every physical system associated with computing is a subset of cyber-physical systems [1]. Cyber-physical security is critical because attacks on these systems can have serious consequences, including hardware damage, service interruption, malware injection through physical ports, and data disclosure. Cybersecurity is incomplete without cyber-physical security. Organizations take various measures to protect both digital and physical assets. However, guarding physical assets 24/7 is much more challenging than digital assets [2]. The Human Activity Recognition-based Cyber-Physical Security (HAR-CPS) algorithm presented in this paper is an innovative and effective solution to beat this challenge.

One common way to secure cyber-physical infrastructure is to isolate it in a confined room and restrict access [3]. However, this is only possible for server computers that allow remote access through computer networks and is impossible to do for desktops and laptops of the office desks. Organizations hire security guards and keep the entrances locked during non-office hours. Many organizations have Closed-Circuit Television Cameras (CCTV) and monitor everything from the control room [4]. Whether secured by guards or monitored from a control room through CCTV, it requires human involvement and their undivided attention. It is beyond human capability to monitor the security status with a maximum attention level because the average attention span of adults is 20 min [5]. This is a significant vulnerability in cyber-physical security. Applying Artificial Intelligence (AI)-driven solutions is a potential way to overcome this vulnerability [6]. A literature review has shown the effective application of AI, including in healthcare [7], robotics [8], microbiology [9], image segmentation [10], and road construction [11]. HAR is a subbranch of AI that has been applied in the proposed methodology to strengthen cyber-physical system security.

The proposed HAR-CPS algorithm uses a combination of GoogleNet [12] and BiLSTM [13] networks. The BiLSTM network learns from the features extracted by GoogleNet and later automatically recognizes the activities it is trained to classify. BiLSTM networks are well known for their excellent capabilities in classifying time-dependent variables [14]. However, they are limited by their feature extraction capabilities. On the other hand, GoogleNet is an excellent CNN for extracting features [15]. However, its computational complexities impose a challenge in time-dependent classification. Combining GoogleNet and BiLSTM networks together to recognize activities from real-time video streams compensates for the weaknesses of each system and makes the classifier more effective. Depending on the level of suspicious activities, the proposed HAR-CPS generates an alarm to alert the responsible authorities. This paper also focuses on the security of the proposed security system. That is why the entire system is deployed in the cloud so that the intruders fail to attack the proposed security system physically. A USB camera connected to an IoT device to transmit the video to the cloud is the only cyber-physical component of the proposed system. IoT cloud computing combines Internet of Things (IoT) devices and cloud computing services to process, analyze, and store data from IoT devices in a more scalable, flexible, and efficient way [16]. That is why it has been used in this research project. The core contributions of the proposed system are:

- Development and training of a GoogleNet–BiLSTM hybrid network to classify designated human activities from video with an average accuracy of 73.15%.
- Creative design of the cyber-physical security system using IoT and cloud computing to ensure the cyber-physical security of the proposed security system.
- Formulation of the novel HAR-CPS algorithm to use the GoogleNet–BiLSTM hybrid network to ensure security.
- Application of Machine Vision at the Edge (Mez) to minimize the cloud resources for cost minimization.

The rest of the paper has been organized into five sections. The second section contains a literature review. The methodology has been presented in the third section of this paper. The methodology is further divided into two more subsections: Dataset and Network Architecture. The fourth section of this paper demonstrates the experimental results and performance evaluation. Finally, the paper is concluded in the fifth section.

2. Literature Review

According to the A. Ray et al., human activity recognition (HAR) is a vibrant research field [17]. HAR is a field in computer vision and machine learning that focuses on recognizing and classifying different human activities [18]. The recent advancements in this research domain demonstrate the outstanding performances of convolutional neural network (CNN)-based approaches [19]. The commercial application of HAR technology is visible in different sectors, including the healthcare sector, fitness tracking, smart homes, smart surveillance and security, and sports analysis [20]. The proposed methodology of this paper is an application of HAR in cyber-physical security. The application of HAR technology in security is not new. L. P. O. Paula et al. developed a front door security system using a human activity recognition-based approach [21]. It strengthens the security at the front door by alerting respected authorities if violent activities are detected. The concepts of the proposed paper align with this paper. However, the HAR-CPS algorithm explores the potential of applying HAR in cyber-physical security. "Cyber-physical security", abbreviated as CPS, describes safeguarding systems comprising physical and computational resources [22].

Research conducted by Sarp B. et al. used a Raspberry Pi-based security system similar to the proposed methodology [23]. However, there was no artificial intelligence applied in their approach. It was a video and audio transmission system that allows users to see outdoor activities and maintain verbal communication. The proposed HAR-CPS algorithm is much more advanced. It uses a sophisticated GoogleNet–BiLSTM network to automatically classify the activities and notify the authorities if there are any threats to cyber-physical security. The security system developed by Aldawira R. C. et al. has an innovative application of IoT, a motion sensor, and a touch sensor [24]. Despite the scope of applying HAR technology, most of the research has focused on video surveillance and simple sensor-based approaches [25–27]. Compared to these papers, the proposed HAR-CSP algorithm is more advanced and effective than most of the state-of-the-art applications of HAR in securing cyber-physical systems.

Kong M. et al. developed a real-time video surveillance system that addresses network latency challenges for real-time video communication [28]. Similar challenges have been faced in the edge-computing-enabled video segmentation research conducted by Wan S. [29]. Transmitting video in real time requires a high bandwidth and is sensitive to time delays. A significant amount of time delay caused by latency interrupts the frame sequence [30]. Moreover, video processing requires a large amount of cloud resources, which increases the expenditure. According to M. Darwich, cost minimization for video processing provided through cloud services is essential [31]. Real-time video transmission through latency-sensitive networks and video processing in the cloud are two challenges the proposed methodology face as well. A. George et al. developed an effective communication technology for real-time video transmission through a latency-sensitive network while maintaining acceptable quality using machine vision at the IoT edge (Mez) [32]. The proposed methodology uses Mez technology to manage the latency sensitivity and cloud resource usage for video processing.

Video analysis and its applications in intelligent surveillance, autonomous vehicles, video analysis, video retrieval, and entertainment rely heavily on computer-vision-based human activity recognition [33]. This paper's analysis agrees with both observation and technique of the proposed methodology. While designing a cyber-physical system security algorithm, it is best to focus on combining computer vision and machine learning. A temporary pose-based human action recognition system was created by Mazzia V. et al. [34]. In a test with 227,000 parameters, it obtained 90.86 percent accuracy, and while the paper's precision is impressive, the high computational cost renders it unsuitable for developing a cheap security system. A DCNN-based architecture using depth vision guided by Wen Q. et al. obtained a promising 93.89 percent accuracy [35]. To train robots on video datasets, this strategy overcomes the difficulty of collecting and classifying large amounts of data. The Microsoft Kinect camera is required for it, which is not cost-effective. Compared to these approaches, the proposed HAR-CPS algorithm is computationally simple and less expensive, yet is a high-performing solution to cyber-physical system security [36].

3. Methodology

A GoogleNet–BiLSTM hybrid network is employed as the classifier in the proposed HAR-CPS algorithm. A video dataset is necessary for this type of hybrid network. In this section, we explain the HAR-CPS algorithm, along with the video dataset selection criteria, dataset processing, network design, the HAR-CPS method's operating principle, and mathematical interpretations. Figure 1 provides a visual summary of the proposed approach.

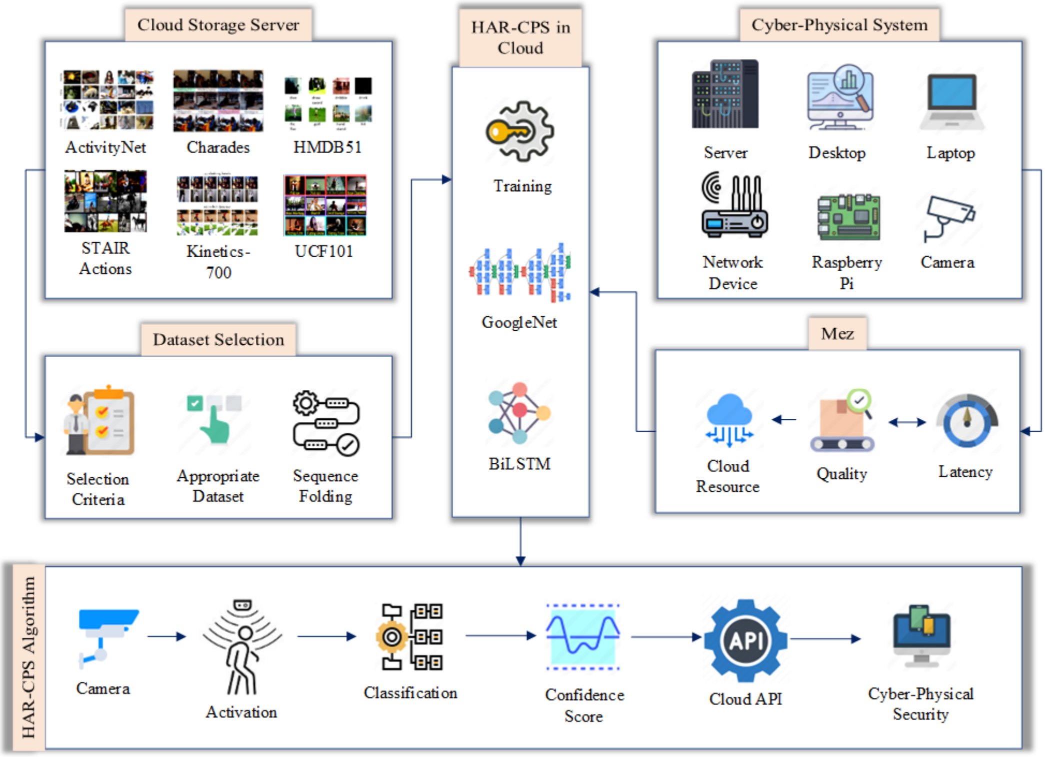

Figure 1. The overview of the proposed methodology.

3.1. Dataset Selection

The proposed HAR-CPS algorithm is an approach based on human activity recognition approach. There are multiple human activity recognition (HAR) datasets. This research has studied and analyzed the most widely used HAR datasets. These datasets are listed in Table 1 [33]. Each dataset is rich enough to train a CNN to recognize human activities. However, the purpose of this research is to recognize activities that are considered threats to the security of cyber-physical systems.

Table 1. Human activity recognition (HAR) dataset descriptions.

Dataset	Categories	Videos	Description
ActivityNet [37]	200	21,313	Activities conducted on a daily, social, and domestic basis, including games and workouts.
Charades [38]	157	66,493	Routine chores performed within the house, such as refilling glasses, folding towels, etc.
HMDB51 [39]	51	5100	Movement of the body and face, as well as contact with objects, are all included.
Kinetics-700 [40]	700	530,336	Interactions involving a single person as well as those involving many people.
STAIR Actions [41]	100	109,478	Frequent indoor activities in the house, workplace, bathroom, and kitchen, including item handling, etc.
UCF101 [42]	101	13,320	Interactions between humans and other objects, movements of the body that do not include other objects, and the utilization of various instruments.

Usually, large-scale cyber-physical systems are kept in confined rooms with limited access. Trained security personnel check the credentials of anyone who wants to access the cyber-physical systems. The proposed HAR-CPS algorithm aims to keep the physical computing infrastructure safe and monitor security breaches as a real human security guard would. Anyone accessing the cyber-physical system without proper authorization and keys to unlock the doors will apply physical force to open the door. The attacker may punch the door to break it. Someone may try to break the door by kicking or hitting it. Pushing the door is another physical force someone may use to break it. Instead of physical force, intruders may carry weapons to gain access to cyber-physical systems. We have selected five activities listed in Table 2 from this observation. These five activities are our core dataset selection criteria.

Table 2. Description of the incidents and class names.

Serial	Incident	Class
1	Trying to break the door by punching	Punch
2	Trying to kick open the door	Kick
3	Hitting on the doorknob to break it	Hit
4	Showing up in front of the door with a weapon	Weapon
4	Pushing the door to open it forcefully	Push

According to our inspection, the HMDB51 dataset contains the target categories mentioned in Table 2. This dataset has a total of 47 categories of videos. The five selected activities are a subset of these 47 categories. This is why HMDB51 is the selected dataset for this experiment. The video clips of the HMDB51 dataset are realistic and original footage. There are no animation or made-up clips. That is why these videos do not require additional filtering and feature enhancement.

3.2. The Hybrid Network Architecture

The proposed CPS algorithm combines GoogleNet and a Long Short-Term Memory (LSTM) network. GoogleNet is used to extract the features from the dataset. The LSTM network uses those features to recognize the activities in real time.

3.2.1. Sequence Folding

A BiLSTM is a recurrent neural network (RNN) that processes sequential data by collecting past and future context. Sequence folding speeds up and improves RNN training, including for BiLSTMs. The input sequence is split into smaller, fixed-length subsequences, or “folds”, in sequence folding. The BiLSTM, which comprises two independent LSTMs, a forward LSTM and a backward LSTM, processes these folds concurrently. The forward LSTM reads the subsequences from left to right and the reverse LSTM from right to left. The data are then more fully represented by concatenating the hidden states from both LSTMs at each time step [43].

Detecting suspicious activities in real time is crucial in cyber-physical security. A grayscale video stream at 30 FPS contains more than 9000 frames in a 5 min video. At the same rate, 24 h video footage contains 2.6 × 10^6 frames. The frame amount will be 3 times more if color video is streamed. Extracting features directly from the video is impractical because of this large number of frames. It introduces a very high latency. As a result, the system fails to detect suspicious activities in real time. We have used the sequence folding method defined by Equation (1) to convert the video sequence into a separate set of images.

$$\sum_{i=1}^{N} I(m_i, n_i) = \sum_{t=1}^{T} f_r((m_t, n_t), t) \tag{1}$$

where $f_r((m_t, n_t), t)$ is a time-dependent frame. This time-dependent frame is converted into time-independent individual images expressed by $I(m_i, n_i)$. These frames are sent to the cloud server. The time-independent frames minimize the latency.

3.2.2. Feature Extractor Network in Cloud

Feature extraction from image frames is computationally expensive. Resource-constrained IoT devices are not suitable for it. We used GoogleNet for feature extraction. Google Cloud has GoogleNet readily available, which is a pre-trained network. However, the entire GoogleNet has not been used. It is a 22-layer deep convolutional neural network (CNN). We used did not use the last three layers. The 19th layer is an average pooling layer. According to the GoogleNet architecture, this layer is responsible for averaging the extracted features. The research approach used in this paper utilizes GoogleNet for feature extraction. That is why the input to the BiLSTM network has been taken from the 19th layer of GoogleNet [44]. The extracted features are converted into a feature vector using Algorithm 1.

Algorithm 1 Constructing Feature Vector.

```
Input: GoogleNet, $G_N$; Frame, $F$
Output: Feature Vector, $F_s$;
Initiate: Allocate Virtual Machine, $VM$;
Start
$L_s \leftarrow VM(Size(Layers(1, G_N)))$
$L_s \leftarrow VM(Convert(L_s, F_s))$
for $i \leftarrow 1 : F$ do
    $Feature \leftarrow VM(pooling(F))$
    $F_s \leftarrow VM(Concat(Feature))$
end for
$VM(save(F_s))$
end
```

Algorithm 1 initializes the virtual machine (VM) in the cloud to extract features from the images. The number of VMs depends on the requests and the service level agreement (SLA) with the service provider. This paper initializes a single VM to construct the feature vector. Algorithm 1 takes GoogleNet and the corresponding frames as the input. Initially, it converts the frame according to the GoogleNet input layer size and stores the resized image as an L_s variable. After that, the features are extracted from video frames in a loop. In every iteration, the features are added to a feature vector F_s. When no more frames remain, the algorithm saves the feature vector. It takes 475 ms to initiate the virtual resources and an additional 711 ms to extract the features per frame. It takes 1.19 s to extract features from a one-minute video. The 1.19 s time delay is considered real time.

3.2.3. GoogleNet–BiLSTM Hybridization

The BiLSTM network is ideal for classifying sequential data, and GoogleNet is optimally designed to extract distinguishable features from images. The hybridization of these two different networks develops a system efficient in feature extraction and sequential data classification. GoogleNet–BiLSTM hybridization has been developed and studied from this observation, illustrated in Figure 2. The BiLSTM network in the experimental setup receives the video features from GoogleNet's average pooling layer. These features are passed to the BiLSTM layer. The responses from this layer are concatenated. These concatenated responses are sent to the dense layer. It follows a fully connected network architecture and a Softmax layer for classification. The classification layer has five output nodes. Each node produces a confidence score, representing the probability of being a certain class.

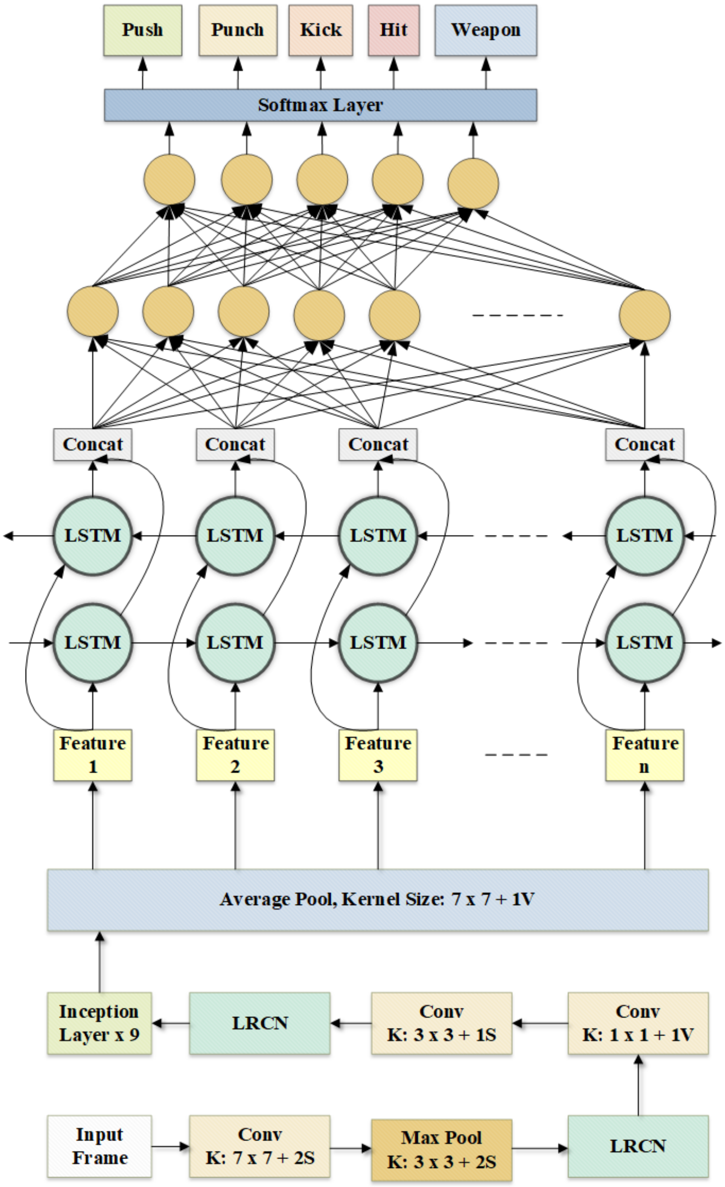

Figure 2. The GoogleNet–BiLSTM Hybridization.

3.2.4. Training the Hybrid Network

The BiLSTM network was trained with the features extracted from GoogleNet. The dataset was split into training, testing, and validation datasets with a ratio of 70:15:15. The training dataset was used to train the network. The validation dataset was used to validate the learning progress during the training. The testing dataset was kept separate and untouched during the training period. It was used to test the performance of the trained hybrid system during experimental analysis. Instead of using the entire dataset simultaneously, we used batch normalization with a mini-batch of size 16. During every iteration, the video clips were internally shuffled within the mini-batch.

Learning algorithms play a vital role in the collective performance of machine learning models. In this experiment, three widely used learning algorithms for deep neural networks have been studied. They are the Adaptive Gradient algorithm (AdaGrad) [45], the Root Mean Squared Propagation (RMSProp) [46], and the Adaptive Moment Estimation (ADAM) [47]. These learning algorithms are expressed in Equations (2)–(4), respectively.

$$\omega_i^{(t+1)} = \omega_i^t - \frac{\eta}{\sqrt{\sum_{\tau=1}^{t} g_{\tau,i}^2}} g_{t,i} \tag{2}$$

$$\omega_i^{(t+1)} = \omega_i^t - \frac{\eta}{\sqrt{(v_t) + \epsilon}} \Delta_t \tag{3}$$

$$\omega_i^{(t+1)} = \omega_i^t - m_t \left(\frac{\alpha}{\sqrt{v_t} + \epsilon} \right) \tag{4}$$

where $\omega_i^{(t+1)}$ and ω_i^t refer to the updated value of the i_{th} weight at time step $t + 1$ and t, respectively. η in Equations (2) and (3) is the learning rate. In Equation (2), $g_{t,i}$ is the loss function. Both Equations (3) and (4) contain ϵ, which adds a small constant to prevent division by zero. v_t in these equations is the exponentially decaying average of the squared gradients at time step t. The loss function in Equation (3) is measured by Δ_t. m_t and α in Equation (4) represent the first moment and learning rate, respectively. The learning algorithms adjust the weights of the hidden nodes of deep neural networks. The more efficient this process is, the better the performance of the trained network becomes. We experimented with all three of the aforementioned algorithms and analyzed the performance using a validation loss curve illustrated in Figure 3.

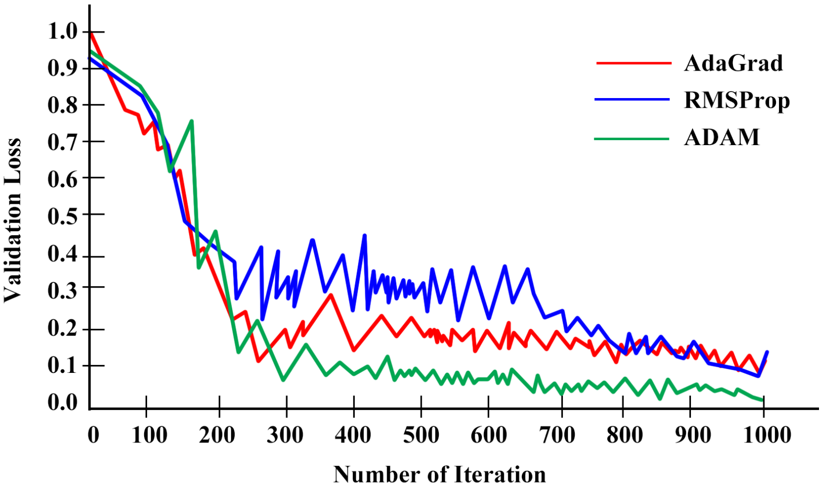

Figure 3. The optimization algorithm selection.

The validation loss curve shows that the AdaGrad learning algorithm reduces the validation loss to 250 iterations. However, there is lots of variation between 250 and 700 iterations. After that, the validation loss reduces again. Compared to this, the performance of RMSProps is much better than AdaGrad. However, the characteristics of the validation loss curve are almost similar. According to the experimental analysis in Figure 3, ADAM is the best-performing learning algorithm. That is why ADAM has been used as the learning algorithm in this research. The proposed network has been trained with 1000 iterations and 568 epochs. The learning progress is illustrated in Figure 4.

It takes 342 min and 19 s to complete the training. It was observed that the accuracy of the validation data increases sharply, and the validation loss falls sharply until the 200th iteration. After that, the slope is negligible, and the learning curve maintains smooth progress. It ends with a 72.48% validation accuracy. The initial learning rate is 0.001 and the final learning rate is 0.0001. A dynamic learning rate was used in this experiment which adjusts itself depending on the accuracy and loss.

3.2.5. HAR-CPS Algorithm

The proposed innovative HAR-CPS algorithm, presented as Algorithm 2, uses the trained GoogleNet–BiLSTM hybrid network to classify the target categories. It runs in a virtual machine provisioned through a pay-as-you-go payment method. It is more efficient to reduce the computational resources to minimize the cost. The proposed algorithm has been designed to minimize the cost. Human activity recognition is the most computationally expensive process. The algorithm calls the GoogleNet–BiLSTM hybrid network only when necessary. For the rest of the time, it performs simple linear 2D subtraction. As a result, the cost is minimized.

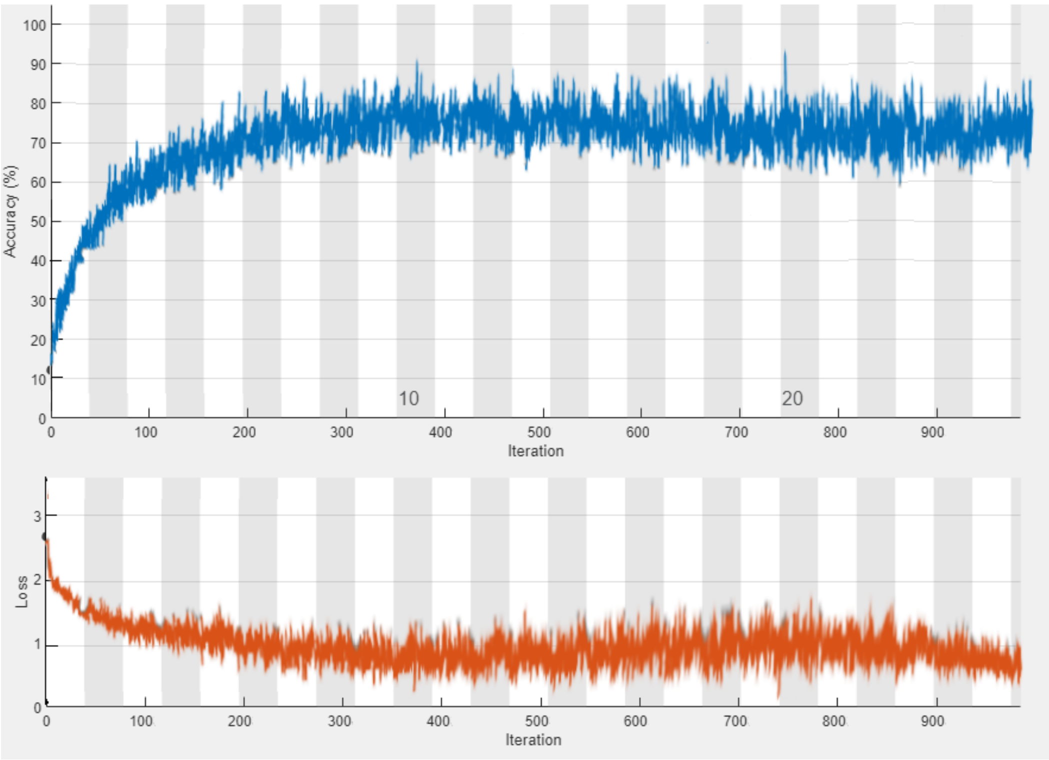

Figure 4. The learning curve with validation training accuracy and validation loss.

Algorithm 2 takes the CCTV video stream and HTTP Live Streaming (HSL) request as inputs. Initially, it initiates a variable i, reads the frames from the video stream, and stores the first frame in the $F[i]$ array. When there is a frame, the while loop is activated. In this loop, the HLS request is accepted for each frame and the frames are continuously read and stored in the $F[i]$ array. The frame difference is calculated by taking the difference between two successive frames. If there is more than a 70% difference between two frames, the proposed HAR-CPS algorithm sends the frame to the GoogleNet–BiLSTM hybrid network. This network classifies the frame and returns the predicted class with a confidence score. If the confidence score is higher than 80%, an alarm is generated according to the identified action. Otherwise, Algorithm 2 does not take any action.

Algorithm 2 applies GoogleNet–BiLSTM to recognize human activities only when two successive frames have more than 70% dissimilarity. Once two successive frames have more than 70% difference, the proposed HAR-CPS algorithm passes the frame to the GoogleNet–BiLSTM network. It predicts human activity on the video stream and returns a confidence score. If the confidence score is more than 80%, an alert is generated through a security API.

3.3. Latency and Cloud Resource Optimization Using Mez

The original Mez architecture was built to link several IoT camera nodes simultaneously. The edge server is linked to it through a wireless network [32]. In the suggested setup, only one camera is linked to a Raspberry Pi 4. Unlike the original Mez system, the proposed system communicates with the cloud server over a licensed 4G spectrum. As a result, a modified Mez architecture, as shown in Figure 5, was adopted in this experiment. This architecture includes a 4G network sensor to check network quality. It exchanges data with the Network Latency Controller.

Algorithm 2 The HAR-CPS Algorithm

```
Input: CCTV Video Stream, $v_s$; HLS Request, $H_l$
Initiate: Allocate Cloud Resource;
Output: Alert, $a$;
Start
$i \leftarrow 0$
$F[i] \leftarrow read(v_s)$
while $v_s = True$ do
    $i \leftarrow i + 1$
    Accept HLS Request
    $F[i] \leftarrow read(v_s)$
    $d \leftarrow difference(F[i-1], F[i])$
    if $d \geq 0.70$ then
        $[p, s] \leftarrow GoogleNetBiLSTM[F[i]]$
        if $s \geq 0.80$ then
            $a \leftarrow class(p)$
            $SecurityAPI(a)$
        end if
    else
        $NoAction$
    end if
end while
end
```

The Pi server in Figure 5 is the subscriber in the subscriber–publisher messaging system. It uses the Remote Procedure Call (RPC) protocol to communicate with the edge server through the Broker model. The same communication protocol is used in the Pi Camera Node (PCN), which is the publisher of the messaging system. The IoT camera node also uses the Broker model to communicate with the edge server. The edge server has a persistent storage and log management system, which stores threshold values, network quality information, and every event log. The Network Latency Controller (NLC) in Figure 5 is connected to an NTA00002B Nemo Outdoor 5G NR Drive Test sensor manufactured by Keysight Technologies, Inc. [48]. It senses the 5G network parameters, including bandwidth, throughput, latency, traffic volume, signal intensity, discontinuity, and interference. Depending on the bandwidth demand, availability, and current throughput, the NLC adjusts the knob values of the Mez to maintain a quality–latency trade-off.

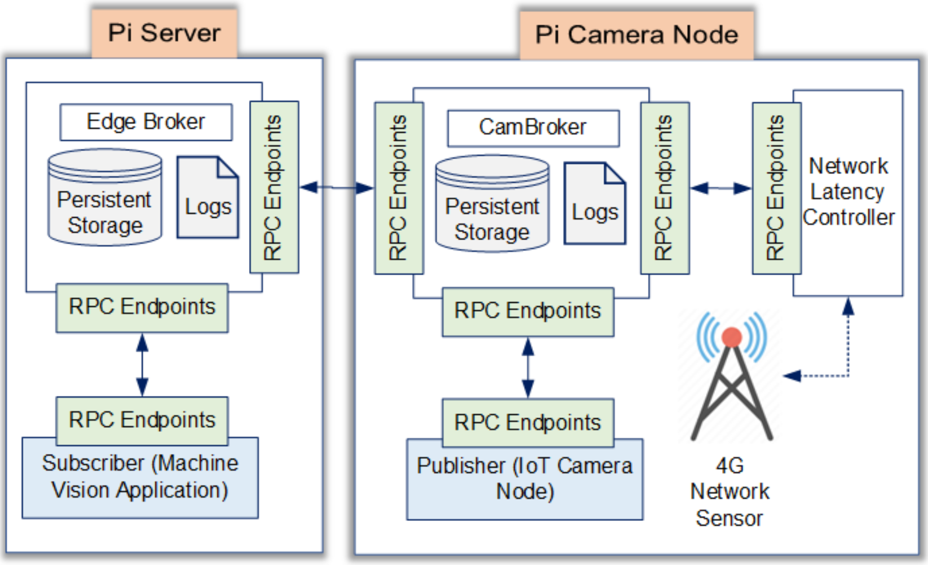

Figure 5. The Mez architecture.

3.3.1. Latency vs. Quality Trade-Off

The suggested system uses Mez technology's latency vs. quality trade-off capabilities. The frame quality may be adjusted using five different knob settings depending on the application precision requirements. Table 3 lists the possible knob settings, their functions, the influence on frame size reduction, and the application scopes.

Table 3. The knob configuration and effects.

Knob	Role	Frame Size Reduction	Scope
1	Resolution Adjustment	84%	Resolutions: 1312 × 736, 960 × 528, 640 × 352, and 480 × 256
2	Colorspace Modification	62%	Colorspaces: BGR, Grayscale, HSV, LAB, and LUV
3	Blurring	46%	Kernel size: 5 × 5, 8 × 8, 10 × 10, and 15 × 15
4	Artifact Removal	98%	Countour-based approach
5	Frame Differincing	40%	Linear frame difference-based method

3.3.2. Cloud Resource Optimization

The proposed HAR-CPS algorithm optimizes cloud resource usage using Mez [32] technology. The empirical analysis shows that keeping the first knob setting listed in Table 3 at 940 × 528 resolution reduces the frame size by 8%, lowering the cloud resource usage for video processing. The grayscale colorspace has been used, which reduces the frame size by 11%. Although Table 3 shows that blurring reduces the frame size, the proposed methodology does not use this knob. It has been observed that blurring the video downgrades the feature quality extracted by GoogleNet. However, artefact removal and frame difference knobs have been used, and they reduced the frame size by 14% and 16%, respectively. After applying Mez technology, the average frame size reduction was 49%. As a result, cloud resource usage was reduced by almost 50%.

4. Results and Performance Evaluation

The proposed cyber-physical security algorithm based on human activity recognition is a deep-learning-based approach that runs on a cloud server. The performance of the system was evaluated from two different perspectives. First, the proposed GoogleNet–BiLSTM hybrid network was evaluated. After that, the performance of the cloud system was studied.

4.1. Performance of the GoogleNet–BiLSTM Hybrid Network

The performance of the proposed GoogleNet–BiLSTM hybrid network was evaluated using state-of-the-art machine learning performance evaluation metrics. The literature review showed that machine-learning-based image classification where CNN or LSTM networks are utilized use accuracy, sensitivity, specificity, false positive rate (FPR), and false negative rate (FNR) evaluation metrics [7]. The mathematical definitions of these evaluation metrics are listed in Table 4. These values are calculated from the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which are obtained from the confusion matrix illustrated in Figure 6.

Table 4. The evaluation metrics used in this research.

Evaluation Metrics	Mathematical Expression	Role
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Classification accuracy
Sensitivity	$\frac{TP}{TP+FN}$	Correct identification of actual positive cases
Specificity	$\frac{TN}{TN+FP}$	True negative rate
False positive rate	$1 - Specificity$	Type I error
False negative rate	$1 - Sensitivity$	Type II error

The performance of the proposed GoogleNet–BiLSTM network in terms of the state-of-the-art machine learning evaluated metrics listed in Table 4 [7]. The performance of the proposed network is detailed in Table 5. The experimental result shows that the proposed system best classifies the "kick" category. The average classification accuracy is 73.15%. The average sensitivity, specificity, false positive rate, and false negative rate are 71.52%, 72.22%, 28.48%, and 27.78%, respectively.

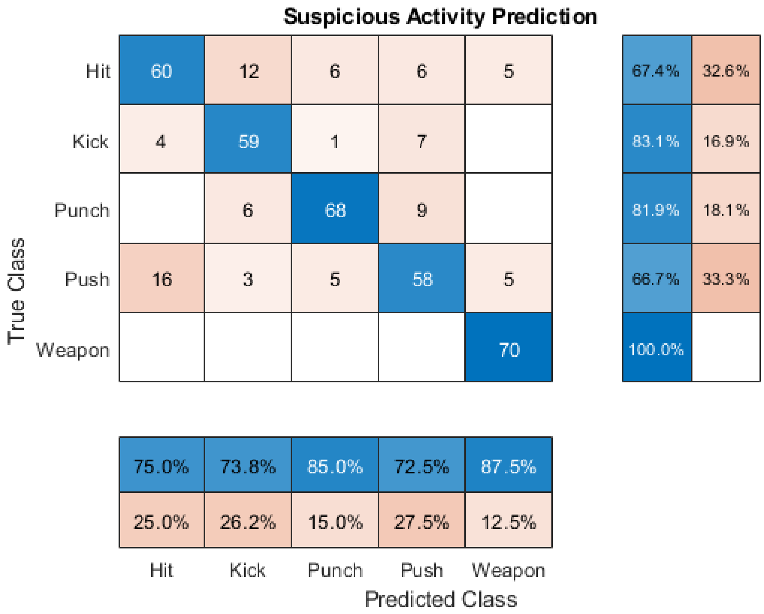

Figure 6. The confusion matrix for performance analysis.

Table 5. Classification performance of the GoogleNet–BiLSTM network.

Activity	Accuracy	Sensitivity	Specificity	FPR	FNR
Hit	73.10%	70.0%	62.2%	30.0%	37.8%
Kick	76.78%	61.3%	80.3%	38.7%	19.7%
Punch	71.47%	80.0%	75.3%	20.0%	24.7%
Push	68.63%	72.5%	66.7%	27.5%	33.3%
Weapon	75.79%	73.8%	76.60 %	26.2%	23.4%

4.1.1. Performance Comparison

The performance of the proposed system was compared with four different models. These models are BiLSTM, CNN [49], MLP [50], and LSTM [51]. The experimental dataset has different lengths of videos. We categorized them into 30 s and 60 s video clips. This experiment was conducted to understand the effect of the proposed system on video clips with different durations. The results of the experiment are detailed in Table 6, demonstrating that the proposed system outperforms other similar approaches.

Table 6. Performance comparison of the proposed system with different models and video lengths.

Model Name	Frame Sequence							
	30 s Clips				60 s Clips			
	Accuracy	Precision	Recall	F-1 Score	Accuracy	Precision	Recall	F-1 Score
BiLSTM	70.45%	68.41%	65.41%	62.40%	72.45%	69.74%	68.41%	58.41%
CNN	63.47%	65.71%	63.91%	60.84%	65.44%	69.71%	62.48%	57.94%
MLP	65.71%	62.78%	65.46%	61.75%	66.78%	65.17%	65.17%	55.17%
LSTM	67.40%	64.71%	66.34%	65.37%	68.41%	62.47%	66.34%	62.78%
Proposed Model	74.17%	72.85%	67.46%	66.74%	74.79%	73.01%	68.70%	67.41%

4.1.2. Resource Optimization Performance

The proposed GoogleNet–BiLSTM hybrid network runs in a cloud server, which handles the video stream from the proposed system [52]. Cloud resource optimization is a major contribution of the proposed methodology. A pay-as-you-go payment scheme is used to implement the HAR-CPS algorithm. This means that the expenditure increases with resource usage. The cloud resource optimization statistics over 60 min (averaged every 10 min) are listed in Table 7. The statistical data show that the optimization scheme used in this paper is most effective in primary memory usage reduction. It reduces the primary memory consumption by 64.44%. It has a positive effect on CPU usage as well. The proposed HAR-CPS system uses 0.45% less CPU after resource optimization. The average disk writing time is 0.12 MB/s after using the Mez, which is a 43.58% reduction. According to the SLA with the cloud service provider, based on the computational resource usage listed in Table 7, the predicted monthly cost of providing cyber-physical security using the proposed system is USD 4.29 only. This is predicted by the pay-as-you-go payment system with the probability of ±6.82% deviation.

Table 7. The cloud resource optimization statistics over 60 min.

Time	Without Mez			With Mez		
	CPU (%)	Memory (MB)	Disk (MB/s)	CPU (%)	B (MB)	Disk (MB/s)
10	0.2	151	0.10	0.1	37	0.13
20	0.8	155	0.20	0.5	47	0.13
30	1.1	90	0.10	0.4	57	0.13
40	1.2	78	0.30	0.1	36	0.13
50	0.7	120	0.30	0.3	50	0.07
60	0.7	140	0.30	0.6	34	0.13

5. Limitations and Future Scope

The experimental results and performance evaluation demonstrate the acceptability of the proposed HAR-CPS algorithm to strengthen the security of cyber-physical systems. Despite the impressive performance, it has several limitations, which have been discussed in this section. However, instead of considering them as limitations, these have been considered as the future scope of this research. These limitations are:

5.1. Limited Number of Actions

The proposed algorithm effectively classifies five human actions that are potential threats to cyber-physical system security. However, more actions may be considered as a security risk that this paper has not considered. The limited number of actions is a significant limitation of this research. The GoogleNet–BiLSTM hybrid network has the potential to learn to classify hundreds of different types of actions. This requires datasets with more categories. The subsequent version of the proposed HAR-CPS will be trained to categorize more human activities to ensure more rigorous cyber-physical security.

5.2. Camera–Subject Angle Sensitivity

The proposed system's accuracy is sensitive to the viewing angle between the subject and the camera. The intruders must be within a 40 to 60 degrees viewing angle. Although the camera is placed to maintain this particular viewing angle, it is still considered a weakness of the system. A geometrical image transformation algorithm is a potential solution to reduce the camera–subject angle sensitivity. Subsequent research on the proposed HAR-CPS algorithm will explore this opportunity.

5.3. Security of the HAR-CPS Device

A significant portion of the proposed HAR-CPS algorithm runs on a cloud server. As a result, it is secured from cyber-physical attacks. However, imaging and IoT devices are kept on the premises and are vulnerable to cyber-physical attacks. A creative camouflage deployment model is a potential solution to this problem, opening new research opportunities.

It is beyond the scope of any approach to ensure 100% security. There are always weaknesses in security systems. The proposed HAR-CPS system is no different. It is effective in strengthening cyber-physical security within its application domain. The limitations of the proposed system pave the path to conducting more research in this domain and to developing a better version of the HAR-CPS algorithm.

6. Conclusions

Cyber-physical security is the protection of critical infrastructure systems that are integrated with computer networks and software. Both physical and digital components are affected in the case of a cyber-physical security breach. Firewalls, intrusion detection systems, frequent vulnerability assessments, and other forms of cyber and physical security, such as access control and surveillance, must be put in place to ensure the safety of these systems. However, implementing cyber-physical system surveillance and security is more challenging than software-based cybersecurity. The human activity recognition-based cyber-physical security (HAR-CPS) algorithm rises to this challenge with flying colors. It reduces the necessity of human intervention in cyber-physical security surveillance and automatically recognizes suspicious activities with an average accuracy of 73.15%. The innovative classifier based on a GoogleNet–BiLSTM network and the algorithm are run on the cloud server, away from the cyber-physical system. As a result, the proposed system remains secured when the cyber-physical system is under attack. The effective application of Mez technology automatically adjusts the video quality to tolerate the latency sensitivity and prevents real-time video transmission interruption. It also reduces the frame size, which optimizes the cloud server expenditure. That is why the innovative HAR-CPS algorithm strengthens cyber-physical security at only USD 4.29 $\pm$ 0.29 per month.

Author Contributions: All authors contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Available on request.

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

References

1. Duo, W.; Zhou, M.; Abusorrah, A. A survey of cyber attacks on cyber physical systems: Recent advances and challenges. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 784–800. [CrossRef]
2. Zhao, Z.; Xu, Y. Performance based attack detection and security analysis for cyber-physical systems. *Int. J. Robust Nonlinear Control* **2023**, *33*, 3267–3284. [CrossRef]
3. Hammoudeh, M.; Epiphaniou, G.; Pinto, P. Cyber-Physical Systems: Security Threats and Countermeasures. *J. Sens. Actuator Netw.* **2023**, *12*, 18. [CrossRef]
4. De Pascale, D.; Sangiovanni, M.; Cascavilla, G.; Tamburri, D.A.; Van Den Heuvel, W.J. Securing Cyber-Physical Spaces with Hybrid Analytics: Vision and Reference Architecture. In Proceedings of the Computer Security: ESORICS 2022 International Workshops: CyberICPS 2022, SECPRE 2022, SPOSE 2022, CPS4CIP 2022, CDT & SECOMANE 2022, EIS 2022, and SecAssure 2022, Copenhagen, Denmark, 26–30 September 2022; Springer: Berlin/Heidelberg, Germany, 2023; pp. 398–408.
5. Jadhao, A.; Bagade, A.; Taware, G.; Bhonde, M. Effect of background color perception on attention span and short-term memory in normal students. *Natl. J. Physiol. Pharm. Pharmacol.* **2020**, *10*, 981–984. [CrossRef]
6. Del Giudice, M.; Scuotto, V.; Orlando, B.; Mustilli, M. Toward the human–centered approach. A revised model of individual acceptance of AI. *Hum. Resour. Manag. Rev.* **2023**, *33*, 100856. [CrossRef]
7. Faruqui, N.; Yousuf, M.A.; Whaiduzzaman, M.; Azad, A.; Barros, A.; Moni, M.A. LungNet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data. *Comput. Biol. Med.* **2021**, *139*, 104961. [CrossRef]
8. Chakraborty, P.; Yousuf, M.A.; Zahidur Rahman, M.; Faruqui, N. How can a robot calculate the level of visual focus of human's attention. In Proceedings of the International Joint Conference on Computational Intelligence: IJCCI 2019; Springer: Berlin/Heidelberg, Germany, 2020; pp. 329–342.
9. Trivedi, S.; Patel, N.; Faruqui, N. Bacterial Strain Classification using Convolutional Neural Network for Automatic Bacterial Disease Diagnosis. In Proceedings of the 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 19–20 January 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 325–332.
10. Trivedi, S.; Patel, N.; Faruqui, N. NDNN based U-Net: An Innovative 3D Brain Tumor Segmentation Method. In Proceedings of the 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, NY, USA, 26–29 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 538–546.
11. Arman, M.S.; Hasan, M.M.; Sadia, F.; Shakir, A.K.; Sarker, K.; Himu, F.A. Detection and classification of road damage using R-CNN and faster R-CNN: A deep learning approach. In Proceedings of the Cyber Security and Computer Science: Second EAI International Conference, ICONCS 2020, Dhaka, Bangladesh, 15–16 February 2020; Proceedings 2; Springer: Berlin/Heidelberg, Germany, 2020; pp. 730–741.
12. Ibrahim, Y.; Wang, H.; Adam, K. Analyzing the reliability of convolutional neural networks on gpus: Googlenet as a case study. In Proceedings of the 2020 International Conference on Computing and Information Technology (ICCIT-1441), Tabuk, Saudi Arabia, 9–10 September 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
13. Wei, X.; Wu, J.; Ajayi, K.; Oyen, D. Visual descriptor extraction from patent figure captions: A case study of data efficiency between BiLSTM and transformer. In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, Cologne, Germany, 20–24 June 2022; pp. 1–5.
14. Zhang, X.; Kim, T. A hybrid attention and time series network for enterprise sales forecasting under digital management and edge computing. *J. Cloud Comput.* **2023**, *12*, 1–21. [CrossRef]
15. Yang, L.; Yu, X.; Zhang, S.; Long, H.; Zhang, H.; Xu, S.; Liao, Y. GoogLeNet based on residual network and attention mechanism identification of rice leaf diseases. *Comput. Electron. Agric.* **2023**, *204*, 107543. [CrossRef]
16. Pflanzner, T.; Kertész, A. A taxonomy and survey of IoT cloud applications. *EAI Endorsed Trans. Internet Things* **2018**, *3*, Terjedelem-14. [CrossRef]
17. Ray, A.; Kolekar, M.H.; Balasubramanian, R.; Hafiane, A. Transfer Learning Enhanced Vision-based Human Activity Recognition: A Decade-long Analysis. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100142. [CrossRef]
18. Beddiar, D.R.; Nini, B.; Sabokrou, M.; Hadid, A. Vision-based human activity recognition: A survey. *Multimed. Tools Appl.* **2020**, *79*, 30509–30555. [CrossRef]
19. Park, H.; Kim, N.; Lee, G.H.; Choi, J.K. MultiCNN-FilterLSTM: Resource-efficient sensor-based human activity recognition in IoT applications. *Future Gener. Comput. Syst.* **2023**, *139*, 196–209. [CrossRef]
20. Kulsoom, F.; Narejo, S.; Mehmood, Z.; Chaudhry, H.N.; Bashir, A.K. A review of machine learning-based human activity recognition for diverse applications. *Neural Comput. Appl.* **2022**, *34*, 18289–18324. [CrossRef]
21. Paula, L.P.O.; Faruqui, N.; Mahmud, I.; Whaiduzzaman, M.; Hawkinson, E.C.; Trivedi, S. A Novel Front Door Security (FDS) Algorithm using GoogleNet-BiLSTM Hybridization. *IEEE Access* **2023**, *11*, 19122–19134. [CrossRef]
22. Kobara, K. Cyber physical security for industrial control systems and IoT. *IEICE Trans. Inf. Syst.* **2016**, *99*, 787–795. [CrossRef]
23. Sarp, B.; Karalar, T. Real time smart door system for home security. *Int. J. Sci. Res. Inf. Syst. Eng.* **2015**, *1*, 121–123.
24. Aldawira, C.R.; Putra, H.W.; Hanafiah, N.; Surjarwo, S.; Wibisurya, A. Door security system for home monitoring based on ESp32. *Procedia Comput. Sci.* **2019**, *157*, 673–682.
25. Sanjay Satam, S.; El-Ocla, H. Home Security System Using Wireless Sensors Network. *Wirel. Pers. Commun.* **2022**, *125*, 1185–1201. [CrossRef]

26. Banerjee, P.; Datta, P.; Pal, S.; Chakraborty, S.; Roy, A.; Poddar, S.; Dhali, S.; Ghosh, A. Home Security System Using RaspberryPi. In *Advanced Energy and Control Systems*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 167–176.
27. Tao, J.; Wu, H.; Deng, S.; Qi, Z. Overview of Intelligent Home Security and Early Warning System based on Internet of Things Technology. *Int. Core J. Eng.* **2022**, *8*, 727–732.
28. Kong, M.; Guo, Y.; Alkhazragi, O.; Sait, M.; Kang, C.H.; Ng, T.K.; Ooi, B.S. Real-time optical-wireless video surveillance system for high visual-fidelity underwater monitoring. *IEEE Photonics J.* **2022**, *14*, 7315609. [CrossRef]
29. Wan, S.; Ding, S.; Chen, C. Edge computing enabled video segmentation for real-time traffic monitoring in internet of vehicles. *Pattern Recognit.* **2022**, *121*, 108146. [CrossRef]
30. Ujikawa, H.; Okamoto, Y.; Sakai, Y.; Shimada, T.; Yoshida, T. Time distancing to avoid network microbursts from drones' high-definition video streams. *IEICE Commun. Express* **2023**, *12*, 126–131. [CrossRef]
31. Darwich, M.; Ismail, Y.; Darwich, T.; Bayoumi, M. Cost Minimization of Cloud Services for On-Demand Video Streaming. *SN Comput. Sci.* **2022**, *3*, 226. [CrossRef]
32. George, A.; Ravindran, A.; Mendieta, M.; Tabkhi, H. Mez: An adaptive messaging system for latency-sensitive multi-camera machine vision at the iot edge. *IEEE Access* **2021**, *9*, 21457–21473. [CrossRef]
33. Kong, Y.; Fu, Y. Human action recognition and prediction: A survey. *Int. J. Comput. Vis.* **2022**, *130*, 1366–1401. [CrossRef]
34. Mazzia, V.; Angarano, S.; Salvetti, F.; Angelini, F.; Chiaberge, M. Action Transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognit.* **2022**, *124*, 108487. [CrossRef]
35. Qi, W.; Wang, N.; Su, H.; Aliverti, A. DCNN based human activity recognition framework with depth vision guiding. *Neurocomputing* **2022**, *486*, 261–271. [CrossRef]
36. Hesse, N.; Baumgartner, S.; Gut, A.; Van Hedel, H.J. Concurrent Validity of a Custom Method for Markerless 3D Full-Body Motion Tracking of Children and Young Adults based on a Single RGB-D Camera. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 1943–1951. [CrossRef]
37. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
38. Sigurdsson, G.A.; Gupta, A.; Schmid, C.; Farhadi, A.; Alahari, K. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv* **2018**, arXiv:1804.09626 .
39. Sharma, V.; Gupta, M.; Pandey, A.K.; Mishra, D.; Kumar, A. A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets. *Appl. Artif. Intell.* **2022**, *36*, 2093705. [CrossRef]
40. Carreira, J.; Noland, E.; Hillier, C.; Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv* **2019**, arXiv:1907.06987.
41. Yoshikawa, Y.; Lin, J.; Takeuchi, A. Stair actions: A video dataset of everyday home actions. *arXiv* **2018**, arXiv:1804.04326.
42. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
43. Liu, G.; Guo, J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* **2019**, *337*, 325–338. [CrossRef]
44. Kumar, V.; Tripathi, V.; Pant, B. Exploring the strengths of neural codes for video retrieval. In *Machine Learning, Advances in Computing, Renewable Energy and Communication: Proceedings of MARC 2020*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 519–531.
45. Lydia, A.; Francis, S. Adagrad—An optimizer for stochastic gradient descent. *Int. J. Inf. Comput. Sci.* **2019**, *6*, 566–568.
46. Turitsyn, S.K.; Schafer, T.; Mezentsev, V.K. Generalized root-mean-square momentum method to describe chirped return-to-zero signal propagation in dispersion-managed fiber links. *IEEE Photonics Technol. Lett.* **1999**, *11*, 203–205. [CrossRef]
47. Newey, W.K. Adaptive estimation of regression models via moment restrictions. *J. Econom.* **1988**, *38*, 301–339. [CrossRef]
48. Berlt, P.; Altinel, B.; Bornkessel, C.; Hein, M.A. Concept for Virtual Drive Testing on the Basis of Challenging V2X and LTE Link Scenarios. In Proceedings of the 2022 16th European Conference on Antennas and Propagation (EuCAP), Madrid, Spain, 27 March–1 April 2022; IEEE: PIscataway, NJ, USA, 2022; pp. 1–5.
49. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; IEEE: PIscataway, NJ, USA, 2017; pp. 1–6.
50. Riedmiller, M.; Lernen, A. Multi layer perceptron. In *Machine Learning Lab Special Lecture*; University of Freiburg: Breisgau, Germany, 2014; pp. 7–24.
51. Bin, Y.; Yang, Y.; Shen, F.; Xu, X.; Shen, H.T. Bidirectional long-short term memory for video description. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 436–440.
52. Hossen, R.; Whaiduzzaman, M.; Uddin, M.N.; Islam, M.J.; Faruqui, N.; Barros, A.; Sookhak, M.; Mahi, M.J.N. Bdps: An efficient spark-based big data processing scheme for cloud fog-iot orchestration. *Information* **2021**, *12*, 517. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.