applied sciences

MDPI

Article

Shoplifting Detection Using Hybrid Neural Network CNN-BiLSMT and Development of Benchmark Dataset

Iqra Muneer [1], Mubbashar Saddique [1,2,*], Zulfiqar Habib [2] and Heba G. Mohamed [3,*]

[1] Department of Computer Science & Engineering, University of Engineering & Technology Lahore, Narowal Campus, Lahore 39161, Pakistan; iqramuneer@uet.edu.pk
[2] Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Islamabad 45550, Pakistan; drzhabib@cuilahore.edu.pk
[3] Department of Electrical Engineering, College of Engineering, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
* Correspondence: dr.mubbashar@uet.edu.pk (M.S.); hegmohamed@pnu.edu.sa (H.G.M.)

Abstract: Shoplifting poses a significant challenge for shop owners as well as other stakeholders, including law enforcement agencies. In recent years, the task of shoplifting detection has gained the interest of researchers due to video surveillance generating vast quantities of data that cannot be processed in real-time by human staff. In previous studies, different datasets and methods have been developed for the task of shoplifting detection. However, there is a lack of a large benchmark dataset containing different behaviors of shoplifting and standard methods for the task of shoplifting detection. To overcome this limitation, in this study, a large benchmark dataset has been developed, having 900 instances with 450 cases of shoplifting and 450 of non-shoplifting with manual annotation based on five different ways of shoplifting. Moreover, a method for the detection of shoplifting is proposed for evaluating the developed dataset. The dataset is also evaluated with methods as baseline methods, including 2D CNN and 3D CNN. Our proposed method, which is a combination of Inception V3 and BILSTM, outperforms all baseline methods with 81 % accuracy. The developed dataset will be publicly available to foster in various areas related to human activity recognition. These areas encompass the development of systems for detecting behaviors such as robbery, identifying human movements, enhancing safety measures, and detecting instances of theft.

Keywords: bidirectional long short-term memory; convolutional neural network; human activity recognition; shoplifting detection

Citation: Muneer, I.; Saddique, M.; Habib, Z.; Mohamed, H.G. Shoplifting Detection Using Hybrid Neural Network CNN-BiLSMT and Development of Benchmark Dataset. *Appl. Sci.* **2023**, *13*, 8341. https://doi.org/10.3390/app13148341

Academic Editor: Christos Bouras

Received: 25 May 2023
Revised: 20 June 2023
Accepted: 25 June 2023
Published: 19 July 2023

1. Introduction

Each year goods worth billions of dollars are shoplifted globally [1]. Shoplifting (also known as retail theft) is a criminal act that reduces the profitability of businesses. It involves stealing merchandise from a retail establishment without paying for it while avoiding detection [1,2]. Shoplifting is a crime that receives significant attention among the various criminal activities that occur in stores. It also involves stealing merchandise from a retail establishment by hiding the items in one's clothing, pockets, or bag and leaving the store without paying for them. Shoplifting poses a significant challenge for store owners and other entities such as law enforcement, the government, and the justice system. A study conducted by the National Association for Shoplifting Prevention indicates that one out of every 11 individuals engages in shoplifting. Additionally, it has been reported that these thieves are apprehended only once in every 48 instances of theft [1].

There are several ways to shoplift, but the majority of them involve hiding products and exiting the business without paying. The act of "hiding," which is followed by a particular series of behaviors, is what ties these strategies together. This behavior is viewed as aberrant from the standpoint of the regular shopper. Reviewing surveillance footage may make it simple to spot instances of stealing, but as the number of customers rises, it

becomes more difficult to keep track of everyone [3]. This leads to a persistent issue, which is made worse by the fact that there are more shoplifters every year. Certain countries have implemented specific laws to address shoplifting and penalize offenders accordingly. In India, shoplifting is considered an offense against property and is dealt with under Chapter XVII of the Indian Penal Code (IPC) [4]. Section 379 of the IPC stipulates that those found guilty of theft can face imprisonment for up to three years, a fine, or both. Retailers have started using closed-circuit television (CCTV) surveillance systems to monitor customers in stores [5,6], but the real-time analysis of footage is a challenging task for humans to undertake continuously. As a result, an advanced monitoring system that analyzes human behavior and identifies shoplifting incidents may be the most effective solution. In an effort to improve public safety, more and more video surveillance cameras are being deployed in public spaces, including streets, banks, shopping centers, and retail stores, every year. Security professionals are unable to handle the enormous volumes of data produced by these camera networks in real-time. The abundance of recording devices has made monitoring a more difficult task [7].

Prior research has primarily concentrated on identifying shoplifting by analyzing human behavior using methods such as 2D CNN [8], 3D CNN [9–12], RNN [3], and LSTM [13]. Previous studies on shoplifting detection have some limitations, such as training on the small size of the dataset and limited cases of shoplifting. For example, one dataset [3] contained only 155 shoplifting videos, while another had a total of 175 video clips, with 88 clips showing normal behavior and 87 clips depicting instances of shoplifting. Another dataset, UCF-Crime [3], included a subset called UCF crime with 28 videos recorded from retail store surveillance cameras. Moreover, none of these datasets include cases of shoplifting where individuals are asked to commit theft, which would reflect how an average person might attempt to steal. As a result, a larger and higher-quality dataset is needed to develop an effective shoplifting detection system.

1.1. Contribution of This Study

The main contribution of this study is as follows: (1) a large dataset of 900 videos has been developed, (2) baseline methods 2D and 3D CNN have been implemented to evaluate the dataset, (3) proposed shoplifting detection technique using hybrid neural network CNN-BiLSTM (4) make a comparison of deep learning approaches 2D and 3D CNN with the proposed technique.

Additionally, our proposed technique of shoplifting detection may be utilized to instantly analyze video from security cameras and find instances of theft. This may be used in a variety of places, including retail establishments, marketplaces, and malls. The efficiency of current security measures to deter theft can be increased by integrating the proposed solution with them. It can also be used to generate alerts and notifications to security personnel when a potential shoplifting incident is detected, allowing them to intervene and prevent the theft from occurring. Moreover, it can help store managers, and owners identify high-risk areas and times where shoplifting incidents are more likely to occur, enabling them to allocate resources more effectively to prevent theft.

1.2. Organization of This Study

The remainder of this article is structured as follows: In Section 2, previous datasets and methods used for shoplifting detection are examined. Section 3 outlines the process used to generate the dataset. The applied methods for shoplifting detection are presented in Section 4. The experimental setup is detailed in Section 5. Section 6 provides an analysis of the results. Finally, Section 7 concludes the article.

2. Related Work

This section will introduce current datasets and methods utilized for the detection of shoplifting.

Yamato et al. [14] conducted a study on categorizing instances of shoplifting using the Jubatus plug-in to extract feature values from images in order to analyze unusual customer behavior. A linear classifier and kNN classifier were used to classify surveillance video data in the proposed application and calculate the probability of shoplifting. Tsushita et al. [15] developed an algorithm for detecting violence and theft in video surveillance. Their approach involved dividing the frame into eight regions and detecting changes in the speed of the person being monitored. Nasaruddin et al. [16] proposed a method for detecting anomalies using a pre-trained C3D model to extract features and a full-link neural network for regression. They used the UCF-Crime dataset [3] to train their model on 11 classes and tested its performance on videos depicting theft, fights, and traffic incidents.

Sultani et al. [9] introduced a real-time anomaly detection approach for identifying 13 anomalous behaviors, including theft, burglary, fighting, shooting, and vandalism. They utilized a 3D CNN network for feature extraction and categorization of the samples into normal and abnormal classes. Their approach included a rating loss function and a fully connected neural network for decision-making, which was trained using the labeled data. The effectiveness of a neural network model on a dataset of real-time shoplifting videos was analyzed in [11] using a 3D CNN for feature extraction and classification. Ansari et al. [13] present an intelligent video surveillance system designed to detect shoplifting activities in retail stores. The system utilizes a convolutions neural network (CNN) to analyze the typical features of motion and appearance of a shoplifter and a deep learning module based on long-term memory (LSTM). The system identifies the shoplifting activities by analyzing the appearance and motion features in a video sequence. The performance of the proposed system is evaluated on a dataset of 177 videos, with 87 shoplifting cases, and showed promising results with accuracy = 90%.

Kirichenko et al. [3] present a novel approach for detecting shoplifting using a hybrid network that combines a 3D convolutional neural network (CNN) and gated recurrent units (GRU). The model uses the spatiotemporal features extracted from a video sequence by the 3D CNN and GRU, then uses the network to classify the actions as either normal or shoplifting. The approach is evaluated on a subset of the UCF crime dataset [8] of real-life shoplifting videos; UCF consists of 28 videos of shoplifting. The dataset was augmented artificially by dividing each video into 32 video segments. Due to the limited size of the dataset, which only contained 310 instances, it was deemed necessary to artificially increase its size. The approach taken involved horizontally mirroring each video fragment, which resulted in 620 instances. The results show improved accuracy with 93 % compared to other methods.

In previous studies, the majority of efforts have been focused on the task of shoplifting detection based on human activity recognition problems [3,13]. The task has been explored with 3D CNN [9–11], CNN + GRU [3], and CNN + LSTM [13]. In addition, the task was explored over very small datasets containing only 155 videos of shoplifting, and another containing a total of 175 video clips, with 88 clips depicting normal behavior and the remaining 87 clips showing instances of shoplifting [13], UCF-Crime dataset [3], including a subset UCF crime consisting of 28 videos recorded from surveillance cameras in retail stores.

The major limitations of the existing datasets are as follows. First, the size of these datasets is small. Second, the corpora have only limited cases of shoplifting. We need a large and highly qualitative dataset to develop an efficient shoplifting detection system. Therefore, it will not be possible to use existing datasets to accurately train and test the shoplifting detection systems. In addition, all these existing datasets lack the simulated (people are asked to commit the theft) cases of shoplifting, which reflects how a common person uses different ways to commit theft.

3. Dataset Development Process

This section presents the development process of the dataset, which includes data collection, annotation, dataset statistics and extracted features, and images from the dataset.

3.1. Data Collection

The dataset that is generated from real-life events is called a real dataset. The fact that it includes real-world events makes the data more crucial for training our model on actual events and improving performance. Using mobile devices and IP cameras, we gathered the dataset from stores where different people were performing the behavior. It was very difficult to persuade store owners to allow these activities to take place in their establishments to develop real datasets. Over the course of six months, we selected several individuals, transported them to the store, and instructed them on how to commit theft. The videos are recorded at a resolution of 352 × 640 while they were performing thievery in several places with varied backdrops and different viewpoints in order to catch every potential technique of committing the stealing. The selected store was "Saim Store".

We asked them to steal good in the following different ways.

- Picking up the items and placing them in their pocket.
- Boys setting the items into their shirts.
- Boys putting items into their jackets.
- Boys also placed the stuff inside their college bags.

3.2. Data Annotation and Statistics

In the manner described above, we made shoplifting videos by training our workforce and devoted a lot of time to this task. The data were manually annotated by an annotator in the next stage. The dataset, which included 450 videos of each class, was labeled with a '0' for non-shoplifting activity and a '1' for shoplifting activity.

After annotating the videos, we trimmed them into equal lengths, extracted all possible frames from each one, and created a csv file with a record of the number of frames in each video and the video's name. Obtaining the appropriate data in the appropriate format is a crucial step in the neural network training process. Our corpus is a real-world dataset that includes actual images that were taken from suggested videos in real life. To make the equal length of all the videos, we trimmed 3 s videos which have 90 frames. Each class has 450 videos and a total of 40,500 images per class. The data are balanced to prevent any potential bias in our model towards a particular class.

3.3. Images from Dataset

Some pictures from the dataset have been shown here to see the different behaviors of shoplifting. Figure 1 shows a scenario where a boy steals an item and puts it in the inner shirt pocket; the view has been taken from the front angle. Similarly, Figure 2 shows, from the side angle, a boy who is stealing the item and putting it in T-shirt's front pocket. Another behavior of shoplifting can be seen in Figure 3, in which a boy wearing a bag on his shoulder keeps the bag open intentionally. As he gets the chance, he steals the item and drops it in the fully opened bag. The most important behavior, which is very complex to identify, is shown in Figure 4. A group of two or more persons often visits the stores with the theft planning. Figure 4 shows such behavior, where a group of two people is visiting, the first person is leading, and the second one is following, and the distance between them is very narrow. In this scenario, first, they were walking normally, and upon obtaining a chance, the second one put the item in the first one's bag. Another behavior of shoplifting can be seen in Figure 5, in which a boy is wearing a closed bag on his shoulder. After covering some distance, he slightly opens some portion of his bag which is not doubtable by anyone. Upon obtaining the opportunity, he steals the item and drops it in that portion. The last Figure 6 shows a scenario where a boy steals an item and puts it in his trousers pocket, the view is taken from the back angle.

Figure 1. A boy putting the item in inner pocket.

Figure 2. A boy putting the item in T-shirt's front pocket.

Figure 3. A boy putting the item in a fully opened bag.

Figure 4. Second person is putting the item into first's bag.

Figure 5. A boy putting the item in the opened portion of closed bag.

Figure 6. A boy putting the item in trouser pocket.

3.4. Comparison with Existing Dataset

The statistics of the existing dataset and the developed dataset are presented in Table 1. The developed dataset has an equal number of instances of shoplifting and non-shoplifting, which help to better train the model, instead of a dataset that does not have an equal number of instances of shoplifting and non-shoplifting.

Table 1. Comparison of developed dataset with existing datasets of shoplifting.

Datasetsets	Shoplifting	Non-Shoplifting	Number of Videos	Dataset Length	Average Fames/s
Arroyo et al. [1]	155	755	910	2730 s	10
Ansari et al. [2]	87	88	175	1750 s	15
UCF-Crime [3]	28	1872	1900	460,800 s	30
Developed Dataset	**450**	**450**	**900**	**2700 s**	**30**

4. Methods

The section presents applied methods, including 2D, 3D deep learning (baseline), and proposed methods in detail.

4.1. Baseline Methods

CNNs have been demonstrated to have exceptional performance in computer vision tasks, particularly in recent years. One variant of CNNs, 2D and 3D CNNs, have been developed to extract spatial and temporal features from videos, respectively. Programs that utilize 3D CNNs for recognition tasks include human action recognition, 3D CNN [8–11], object recognition [17], and gesture recognition [18,19].

4.1.1. Two-Dimensional Convolutional Neural Network

A sort of neural network design known as a 2D convolutional neural network (CNN) is frequently used in image analysis tasks such as picture categorization, object recognition, and segmentation. A 2D array representing an image is commonly used as the input data for a 2D CNN. Each pixel in the array is represented by a numerical value that corresponds to the intensity or color of the associated pixel in the picture. The network comprises a number of layers, including convolution, pooling, and fully connected layers, which carry out different operations on the input data.

A 2D CNN's convolutional layers identify edges, corners, and other patterns in the input picture as features. Multiple filters, or tiny matrices of weights, are included in each convolutional layer and are applied to the input picture to produce a collection of feature maps. These feature maps draw attention to various facts of the input image and assist the network in developing its pattern and object recognition capabilities. Downsampling the feature maps and reducing the dimensionality of the input data are accomplished using pooling layers in a 2D CNN. This makes the network more effective by lowering its computational complexity. A prediction is then made using a 2D CNN's fully connected layers based on the output of the earlier levels. These layers are frequently applied in the network's final stages to divide the input picture into one or more categories. In general, a 2D CNN is an effective tool for processing image data and has been applied in a variety of

fields, including computer vision [20], imaging in the medical field [21], and autonomous cars [22].

In our case, we have utilized a 2D model to train a dataset that identifies shoplifting. Our training set consists of 81,000 images, which are separated into two distinct classes. During training, the images are inputted with a size of 600 × 600 and are passed from the input layer. In the next step, 4 Convolutional layers were used to extract low-level features for learning from images. After each convolutional layer, an activation layer was added to the network to learn more complex features and relationships between different parts of the input image with the RELU activation function. In the final step, the softmax activation function [23,24] was used for the final prediction. The 2D arrays of pixels that make up a picture are intended for analysis by 2D CNNs, which do not account for the 3D structure of the real world. They may, therefore, be unable to accurately depict the spatial connections between items in a picture. A 2D CNN's topology is normally determined before training, and the network is tuned to carry out a particular task. This implies that for the analysis of complicated or changeable data, 2D CNNs could not be as adaptable as other types of models. The quality of the picture can alter due to changes in lighting, color, and orientation, which 2D CNNs are sensitive to. As a result, they might not perform as well on pictures that are not the same as the ones they were trained on. Additionally, a 2D CNN can have a significant number of parameters, particularly for large pictures or deep networks. As a result, training and inference may be laborious and costly in terms of computing. Because of this, 2D CNNs may not be as efficient for jobs involving text or time-series data as they are for those involving image analysis [20,25].

4.1.2. Three-Dimensional Convolutional Neural Network

A sort of neural network called a 3D CNN (Convolutional Neural Network) is capable of processing three-dimensional data, such as movies or 3D photographs. An expansion of a 2D CNN, a 3D CNN convolves each filter across the input volume as opposed to a 2D picture. A 3D CNN may be utilized for a variety of tasks, including robotics [26], medical imaging, and action identification in films. A 3D CNN may identify an action by analyzing the spatio-temporal aspects of a video stream in the instance of action recognition [27]. The three-dimensional data that are sent into the network for processing make up the input layer of the network. The main components of a 3D CNN are its 3D convolutional layers. A series of learnable filters make up each 3D convolutional layer, which convolves across the input volume to produce a set of feature maps. An activation layer is often added to the network to add non-linearity after each 3D convolutional layer.

The feature maps produced by the 3D convolutional layers are down-sampled using 3D pooling layers. In order to conduct classification based on the output of the preceding layers, the fully linked layers are often added toward the end of the network. The network's output layer generates the final prediction or output. In general, a 3D CNN's structure may be modified to meet the unique needs of a particular issue, and the number and kinds of layers utilized in the network can change based on how difficult the problem is being addressed.

In our case, we have utilized a 3D model to train a dataset that identifies shoplifting. Our training set consists of 81,000 images with (frame sequence = 90) for each movie chosen for training input out of a total of 900 videos from 2 classes. During training, the images are inputted with a size of 600 × 600 and are passed from the input layer. In the next step, 8 Convolutional layers were used to extract low-level features for learning from the sequence of frames. After each 3D convolutional layer, an activation layer was added to the network image with the RELU activation function [3,28]. In the final step, the softmax activation function [23,24] was used for the final prediction.

The incorporation of an extra-temporal dimension in 3D CNNs raises the training and inference computing costs. Due to this, training big 3D CNNs on hardware with limited resources might be difficult. Since volumetric data is processed using 3D CNNs, a sizable amount of memory is needed to store the data and intermediate feature maps [29]. The

amount of input data that may be processed may be constrained as a result. While photos are plentiful and simple to get, 3D video data are still uncommon in many applications.

This can impair the performance of 3D CNNs by limiting the supply of big, high-quality training datasets. It can be difficult to evaluate the learned characteristics and comprehend how they contribute to the network's predictions because of the increasing complexity of 3D CNNs. This may reduce the 3D CNNs' transparency and interpretability, which may be a problem in some applications. In order to extract useful characteristics, 3D CNNs frequently need spatial context, which might be challenging to get in particular applications. The performance of 3D CNNs may be constrained in specific medical imaging applications, such as brain tumor segmentation, where the spatial context may not be present. Like any other deep learning model, 3D CNNs are susceptible to over-fitting, especially when the dataset is small [30].

4.2. Proposed Method

CNN has proven to be a highly effective method for image classification, outperforming other approaches. When it comes to video classification, the features of each frame in a sequence need to be extracted and inputted into a Bidirectional Long Short-Term Memory. Since CNN can identify hidden patterns in individual frames and changes in a sequence of frames, it is the best option for feature extraction. However, training a CNN from scratch requires a large dataset of images, substantial computational resources, and significant time investment for training and testing the model, making it a costly approach. Transfer learning can be used to mitigate this problem, where the last hidden layers of a pre-trained CNN can be removed, and only the image features are extracted.

The reason for selecting the Inception model is a convolutional neural network (CNN) architecture known for its ability to capture both local and global features from images effectively. It utilizes a combination of parallel convolutional filters of different sizes, enabling it to capture details at various scales. Shoplifting detection often involves analyzing video frames or images to identify suspicious activities or objects. The Inception model's robust feature extraction capabilities can be leveraged to capture relevant visual cues, patterns, and object characteristics. By using the Inception model as the visual backbone, the hybrid system can effectively process and understand the visual information from the surveillance footage. In addition, the Inception model has been used and showed exceptional performance for human activity recognition tasks [2,3,12,28].

Based on this, the method is proposed. The detailed architecture is shown in Figure 7. In the first step, the features from the frames were extracted using Inception V3 [31] with n = 7 for the 17 × 17 grid and 2 × Inception. We choose n = 7 for the 17 × 17 grid, and 2 × 2 Inception represents the module with expanded filter bank outputs. Sixteen convolution layers were used, and the final dense layer was removed. The working of inception V3 is shown in Figure 8. We just removed the last dense layer and fed the features into BILSTM [32].

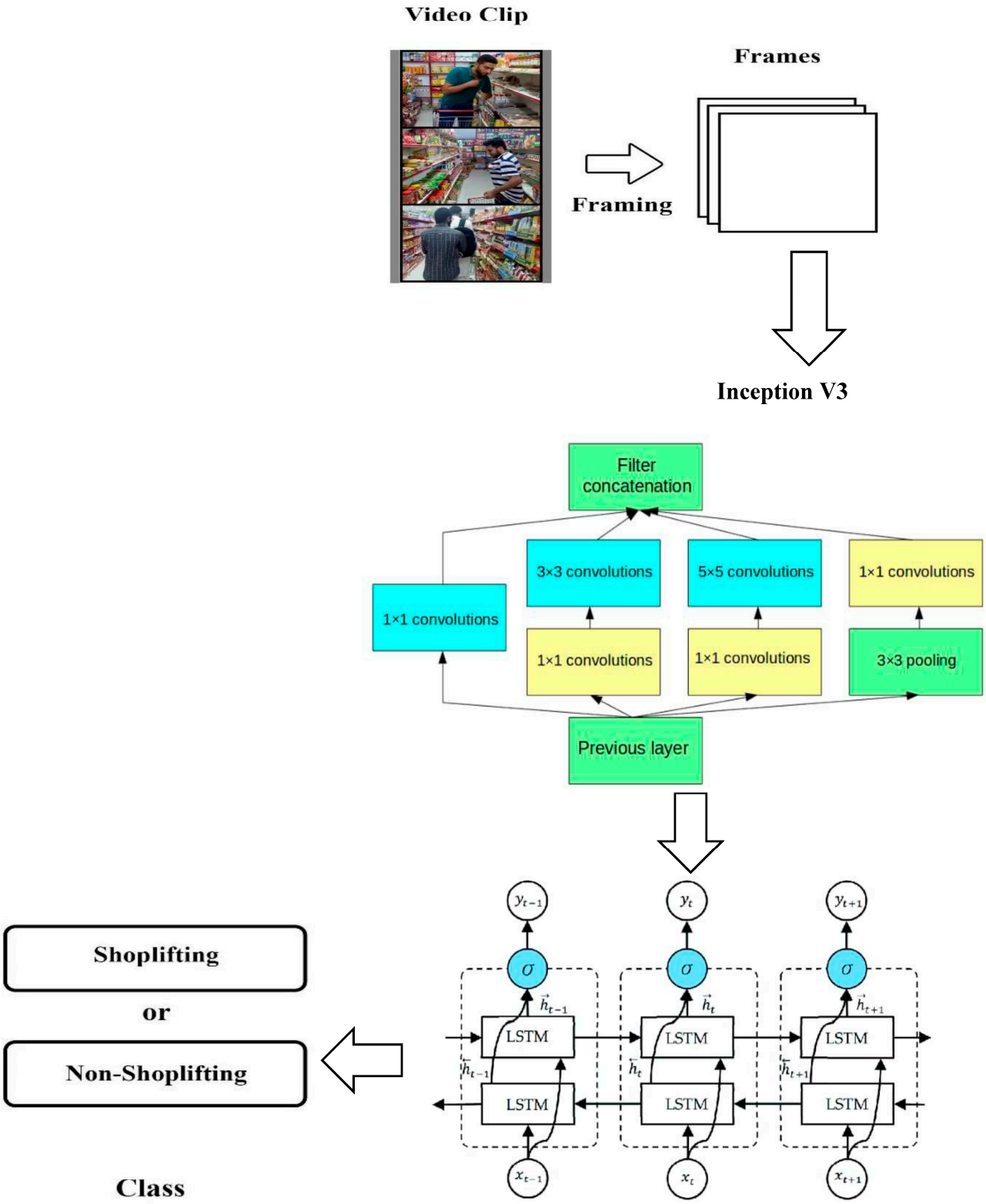

Figure 7. Proposed architecutre.

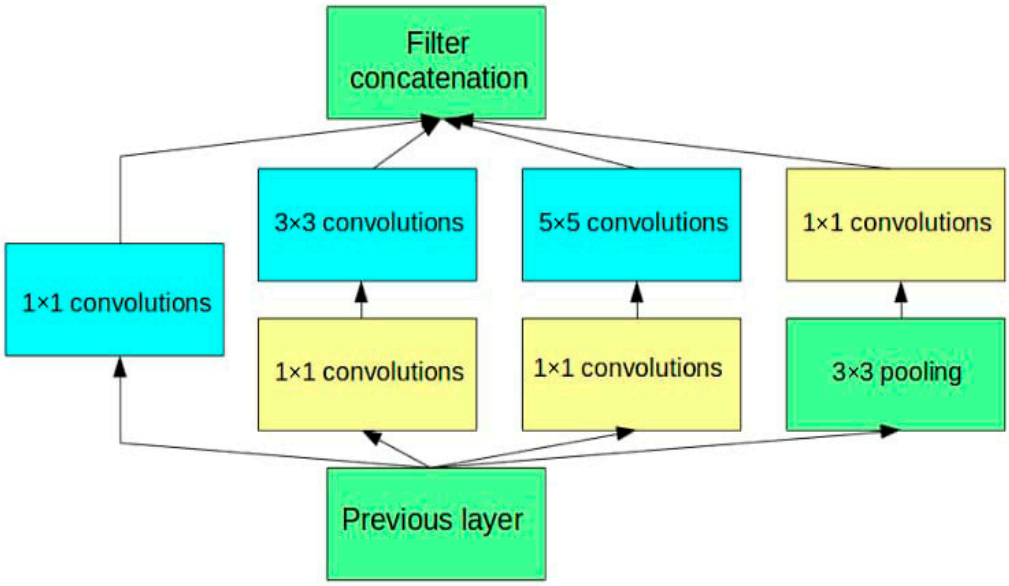

Figure 8. Architecture of Inception V3.

Bidirectional Long Short-Term Memory (BIL-STM) and Inception V3 are the two components of our proposed method, as shown in Figure 7. For the feature extraction of the frames in the proposed technique, we used parameters of a trained CNN dubbed Inception V3, which was trained on ImageNet [33] with 1.2 million training pictures, 50,000 validation images, and 100,000 testing images. Inception V3 is a deep convolutional neural network (CNN) architecture that was proposed by Google in 2015 as an improvement over the original Inception architecture [31].

In the initial Inception design, 5×5 convolution filters, which were computationally costly and produced overfitting, were utilized. These filters are replaced in Inception V3 with a mixture of 3×3 and 1×1 filters, which are more computationally effective and aid in reducing overfitting. Inception V3 starts the network with a stem module made up of pooling layers, 3×3 and 5×5 convolution filters, and other components. This aids in capturing characteristics at various scales and enhances the network's capacity to identify objects of various sizes. After each grid of inception modules, Inception V3 employs a reduction module to lower the dimensionality of the feature mappings. This is accomplished by combining 1×1 convolution filters with max pooling, which lowers the number of parameters in the network and increases its effectiveness.

After the first grid of inception modules, Inception V3 employs an auxiliary classifier to assist in regularizing the network and avoid overfitting. This classifier's predictions are mixed with the network's final predictions once it has been trained to predict the output class. The Inception V3 architecture contains 48 convolutional layers, the last of which is a fully connected layer with softmax activation followed by a layer with global average pooling. It has achieved state-of-the-art performance on several benchmark datasets, including the ImageNet large-scale visual recognition challenge (ILSVRC), where it achieved a top-5 error rate of 3.46%. BILSTM is a type of recurrent neural network (RNN) [34] that incorporates memory cells and can learn long-term dependencies in sequential data.

There are several reasons for the selection of BILSTM for the proposed model is as follows: shoplifting detection typically involves analyzing temporal sequences of actions or events, such as movements, interactions, and object trajectories, to identify suspicious behavior. Bidirectional LSTM can be applied to challenges requiring sequence-to-sequence modeling. The ability to process the input sequence both forward and backward makes this variation of the conventional LSTM capable of capturing dependencies in both past and future contexts. By incorporating BILSTM into the hybrid system, it becomes capable of modeling the temporal dynamics of shoplifting activities, detecting patterns of behavior that may be indicative of shoplifting, and distinguishing them from normal activities.

The hybrid system can benefit from the strengths of both models. The Inception model can extract high-level visual features, while the BILSTM model can effectively analyze sequential data and capture temporal dependencies. The combination of these models enables the system to have a comprehensive understanding of both visual and temporal aspects, enhancing its ability to detect shoplifting incidents accurately.

The process occurs as the input sequence in opposite directions: one that goes ahead and the other that goes backward. The final output sequence is created by concatenating the hidden state sequences that each LSTM generates. The work is shown in Figure 9.

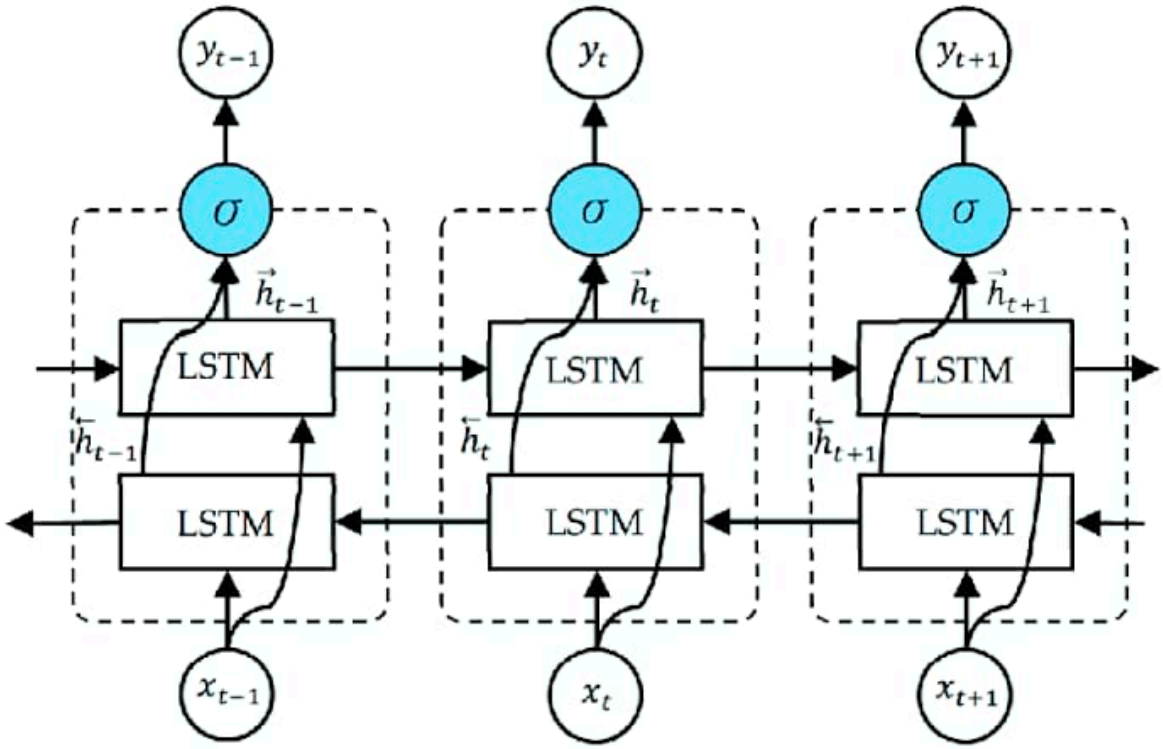

Figure 9. Architecture of Bi-LSTM.

The input sequence is processed, start to finish, by the forward LSTM and in the opposite order by the backward LSTM. The current input and the prior hidden state are used to update the hidden state at each time step in the forward LSTM, whereas the current input and the following hidden state are used to update the hidden state at each time step in the backward LSTM. By joining the forward and backward LSTM outputs at each time step, the bidirectional LSTM's output is created. As a result, a series of concatenated hidden states is produced, encoding the input sequence's past and future dependencies.

5. Experimental Setup

The developed dataset, called 'shoplift-23', has been used to evaluate the proposed and baseline methods. There are a total of 900 videos belonging to 2 classes, from which 90 frames per video are selected for training input. Each class contains 450 videos and a total of 40,500 images per class. There are a total of 81,000 frames that belong to two classes, and these frames are split into 20% of the test data.

The problem of shoplifting detection was treated as a supervised classification task. Different methods, including deep learning as baseline (2D CNN, 3D CNN) and proposed models, were applied. For experiment no 1, a 2D CNN-based method was developed, applied, and evaluated. 3D CNN-based method was developed, applied, and evaluated for experiment no 2. Finally, our proposed model was developed, applied, and evaluated as experiment no 3. All of the methods were applied using an 80:20 random split approach

of train-validation with 100 epochs and batch size 32 using Adam optimizer [35]. The data are balanced, and performance was reported using accuracy.

Accuracy is calculated by dividing the number of accurate predictions by the total number of predictions the model produced [36]. Precision is from all the classes we have predicted as positive; how many are actually positive [37]. Recall is from all the positive classes, how many we predicted correctly [38]. The F_1—Measure is defined as the harmonic mean of two other measures, Precision and Recall [39].

$$\text{Accuracy} = \text{TP} + \text{TN}/(\text{TP} + \text{FP} + \text{FN} + \text{TN}) \quad (1)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

$$F_1\text{-Measure} = 2 \times \text{Precision} \times \text{Recall}/(\text{Precision} + \text{Recall}) \quad (4)$$

6. Results and Discussion

The section presents the results of the experimentation conducted on the proposed dataset. Table 2 shows the accuracy and precision. Recall and F_1 were obtained by applying various methods for shoplifting detection. In the Table, "Methods" refers to the applied technique for the experiment used for shoplifting detection. Overall, the best result with (accuracy = 82%, precision = 88.80%, recall = 78.40 %, and F_1 = 83.01%) has been shown by our proposed method for the shoplifting detection task. It can also be observed that the proposed model is outperforming in terms of all measures with respect to accuracy, precision, recall, and F_1. Tables 3 and 4 show the confusion matrices, false positive rate, false negative rate, etc., for 2D CNN and 3D CNN, respectively. Similarly, Table 5 shows the confusion matrix for more insight details of our proposed model.

Table 2. Comparison of baseline methods and proposed method in terms of accuracy.

Methods	Convolutional Layers	Activation Units	Training Accuracy (%)	Validation Accuracy (%)	Precision (%)	Recall (%)	F_1 (%)	AUC
2D CNN	4	Relu Softmax	50.00	45.00	51.00	50.00	50.40	0.49
3D CNN	8	Relu Softmax	60.85	55.38	66.60	58.80	61.80	0.57
Proposed Method	**16**	**Relu Softmax**	**82.01**	**81.00**	**88.80**	**78.40**	**83.01**	**0.88**

Table 3. Confusion Matrix of 2D CNN.

		Predicted	
		Shoplifting	Non-Shoplifting
Actual	Shoplifting	230	220
	Non-Shoplifting	230	220

Table 4. Confusion Matrix of 3D CNN.

		Predicted	
		Shoplifting	Non-Shoplifting
Actual	Shoplifting	300	150
	Non-Shoplifting	210	240

Table 5. Confusion Matrix of Proposed Model.

Actual		Predicted	
		Shoplifting	Non-Shoplifting
	Shoplifting	400	50
	Non-Shoplifting	110	340

It shows that our proposed model based on transfer learning, which is a combination of Inception V3 + BILSTM, is more fruitful for shoplifting detection.

The reasons for the best performance are as follows:

1. Multi-scale processing to process images at different scales. This allows the network to capture features at different levels of abstraction, which helps improve its accuracy.
2. Inception V3 uses a combination of convolutional layers of different sizes and pooling layers to reduce the number of parameters in the network while still maintaining high accuracy. This efficient use of parameters allows the network to train faster, even with less data.
3. Inception V3 includes auxiliary classifiers that are used during training to encourage the network to learn more useful features, which helps improve the overall accuracy of the network.
4. Inception V3 uses various regularization techniques, such as dropout and weight decay, to prevent overfitting and improve the generalization performance of the network.
5. BiLSTMs can capture long-term dependencies in sequential data by using recurrent connections that allow information to flow through the network from one time step to the next.
6. BiLSTMs process the input sequence in both forward and backward directions, allowing the network to capture information from both past and future contexts.
7. BiLSTMs use memory cells to store information for later use. This allows the network to selectively remember or forget information based on the current input and the state of the network.
8. BiLSTMs can be trained using techniques such as gradient clipping and dropout to prevent overfitting and improve generalization performance. These techniques help to prevent the network from becoming too specialized for the training data and allow it to perform well on new, unseen data.

7. Conclusions

This paper presents a large benchmark dataset of 900 instances with 450 cases of shoplifting and 450 of non-shoplifting with manual annotation based on different ways of shoplifting. Moreover, the benchmarked dataset is evaluated using deep learning methods as baseline methods, including 2D CNN, 3D CNN, and the proposed method. A detailed comparison of deep learning methods (baseline) and the proposed method is carried out. The proposed method, which is a hybrid method based on Inception V3 and BILSTM, has outperformed all baseline methods with 81 % accuracy. In the future, we plan to investigate more deep learning and transfer learning methods for robbery and theft detection based on human activity.

Author Contributions: Conceptualization: I.M. and M.S.; methodology: I.M. and M.S.; validation: Z.H. and H.G.M.; formal analysis: I.M.; investigation, M.S.; resources, I.M. and M.S.; writing—I.M.; writing—review and editing, M.S.; visualization, I.M.; funding acquisition, H.G.M. All authors have read and agreed to the published version of the manuscript.

Funding: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023TR140), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by Ethical Committee CUI, Lahore Campus, Lahore (Approval no. EC/ZH/002/23).

Acknowledgments: Moreover, this research work would be impossible without the support of Saim Super Store Pasrur, District Sialkot Pakistan, and the authors are very thankful to the management of the store, who helped in this work.

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

References

1. Arroyo, R.; Yebes, J.J.; Bergasa, L.M.; Daza, I.G.; Almazán, J. Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert Syst. Appl.* **2015**, *42*, 7991–8005. [CrossRef]
2. Ansari, M.A.; Singh, D.K. An expert eye for identifying shoplifters in mega stores. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021*; Springer: Singapore, 2021; Volume 3, pp. 107–115.
3. Kirichenko, L.; Radivilova, T.; Sydorenko, B.; Yakovlev, S. Detection of Shoplifting on Video Using a Hybrid Network. *Computation* **2022**, *10*, 199. [CrossRef]
4. Gaur, K.D. *Textbook on the Indian Penal Code*; Universal Law Publishing: Boca Raton, FL, USA, 2009.
5. Singh, D.K. Human action recognition in video. In Proceedings of the Advanced Informatics for Computing Research: Second International Conference, ICAICR 2018, Shimla, India, 14–15 July 2018; Revised Selected Papers, Part I 2. pp. 54–66.
6. Singh, D.K.; Kushwaha, D.S. Tracking movements of humans in a real-time surveillance scene. In *Proceedings of Fifth International Conference on Soft Computing for Problem Solving*; SocProS 2015; Thapar University: Patiala, India, 2016; Volume 2, pp. 491–500.
7. Kirichenko, L.; Radivilova, T. Analyzes of the distributed system load with multifractal input data flows. In Proceedings of the 2017 14th IEEE International Conference the Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Lviv, Ukraine, 21–25 February 2017; pp. 260–264.
8. Szentannai, K.; Al-Afandi, J.; Horváth, A. Mimosanet: An unrobust neural network preventing model stealing. *arXiv* **2019**, arXiv:1907.01650.
9. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 14–19 June 2020; pp. 6479–6488.
10. Arunnehru, J.; Chamundeeswari, G.; Bharathi, S.p. Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos. *Procedia Comput. Sci.* **2018**, *133*, 471–477. [CrossRef]
11. Martínez-Mascorro, G.A.; Abreu-Pederzini, J.R.; Ortiz-Bayliss, J.C.; Garcia-Collantes, A.; Terashima-Marín, H. Criminal intention detection at early stages of shoplifting cases by using 3D convolutional neural networks. *Computation* **2021**, *9*, 24. [CrossRef]
12. Amin, J.; Anjum, M.A.; Ibrar, K.; Sharif, M.; Kadry, S.; Crespo, R.G. Detection of anomaly in surveillance videos using quantum convolutional neural networks. *Image Vis. Comput.* **2023**, *135*, 104710. [CrossRef]
13. Ansari, M.A.; Singh, D.K. ESAR, An Expert Shoplifting Activity Recognition System. *Cybern. Inf. Technol.* **2022**, *22*, 190–200. [CrossRef]
14. Yamato, Y.; Fukumoto, Y.; Kumazaki, H. Security camera movie and ERP data matching system to prevent theft. In Proceedings of the 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 8–11 January 2017; pp. 1014–1015.
15. Tsushita, H.; Zin, T.T. A study on detection of abnormal behavior by a surveillance camera image. In *Proceedings of the First International Conference on Big Data Analysis and Deep Learning*; University of Miyazaki Japan: Miyazaki, Japan, 2019; pp. 284–291.
16. Nasaruddin, N.; Muchtar, K.; Afdhal, A.; Dwiyantoro, A.P.J. Deep anomaly detection through visual attention in surveillance videos. *J. Big Data* **2020**, *7*, 87. [CrossRef]
17. Zhiqiang, W.; Jun, L. A review of object detection based on convolutional neural network. In Proceedings of the 2017 36th Chinese Control Conference (CCC), Dalian, China, 26–28 July 2017; pp. 11104–11109.
18. Chung, H.-Y.; Chung, Y.-L.; Tsai, W.-F. An efficient hand gesture recognition system based on deep CNN. In Proceedings of the 2019 IEEE International Conference on Industrial Technology (ICIT), Melbourne, VIC, Australia, 13–15 February 2019; pp. 853–858.
19. Wu, Y.; Zheng, B.; Zhao, Y. Dynamic gesture recognition based on LSTM-CNN. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 2446–2450.
20. Bhatt, D.; Patel, C.; Talsania, H.; Patel, J.; Vaghela, R.; Pandya, S.; Modi, K.; Ghayvat, H. CNN variants for computer vision: History, architecture, application, challenges and future scope. *Electronics* **2021**, *10*, 2470. [CrossRef]
21. Li, Q.; Cai, W.; Wang, X.; Zhou, Y.; Feng, D.D.; Chen, M. Medical image classification with convolutional neural network. In Proceedings of the 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), Singapore, 10–12 December 2014; pp. 844–848.
22. Gao, H.; Cheng, B.; Wang, J.; Li, K.; Zhao, J.; Li, D. Object classification using CNN-based fusion of vision and LIDAR in autonomous vehicle environment. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4224–4231. [CrossRef]
23. Sharma, S.; Sharma, S.; Athaiya, A. Activation functions in neural networks. *Towards Data Sci.* **2017**, *6*, 310–316. [CrossRef]
24. Gong, X.; Tang, B.; Zhu, R.; Liao, W.; Song, L. Data augmentation for electricity theft detection using conditional variational auto-encoder. *Energies* **2020**, *13*, 4291. [CrossRef]

25. Debella-Gilo, M.; Gjertsen, A.K. Mapping seasonal agricultural land use types using deep learning on Sentinel-2 image time series. *Remote Sens.* **2021**, *13*, 289. [CrossRef]
26. Li, L.; Ota, K.; Dong, M. Sustainable CNN for robotic: An offloading game in the 3D vision computation. *IEEE Trans. Sustain. Comput.* **2018**, *4*, 67–76. [CrossRef]
27. Ouyang, X.; Xu, S.; Zhang, C.; Zhou, P.; Yang, Y.; Liu, G.; Li, X. A 3D-CNN and LSTM based multi-task learning architecture for action recognition. *IEEE Access* **2019**, *7*, 40757–40770. [CrossRef]
28. Eckle, K.; Schmidt-Hieber, J. A comparison of deep networks with ReLU activation function and linear spline-type methods. *Neural Netw.* **2019**, *110*, 232–242. [CrossRef]
29. Niyas, S.; Vaisali, S.C.; Show, I.; Chandrika, T.; Vinayagamani, S.; Kesavadas, C.; Rajan, J. Segmentation of focal cortical dysplasia lesions from magnetic resonance images using 3D convolutional neural networks. *Biomed. Signal Process. Control.* **2021**, *70*, 102951. [CrossRef]
30. He, X.; Yang, X.; Zhang, S.; Zhao, J.; Zhang, Y.; Xing, E.; Xie, p. Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. *medrXiv* **2020**. medrXiv:2020.2004.2013.20063941.
31. Sam, S.M.; Kamardin, K.; Sjarif, N.N.A.; Mohamed, N. Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet inception-v1 and inception-v3. *Procedia Comput. Sci.* **2019**, *161*, 475–483.
32. Tao, D.; Wen, Y.; Hong, R. Multicolumn bidirectional long short-term memory for mobile devices-based human activity recognition. *IEEE Internet Things J.* **2016**, *3*, 1124–1134. [CrossRef]
33. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
34. Ansari, M.A.; Singh, D.K. Optimized Parameter Tuning in a Recurrent Learning Process for Shoplifting Activity Classification. *Cybern. Inf. Technol.* **2023**, *23*, 141–160. [CrossRef]
35. Zhang, Z. Improved adam optimizer for deep neural networks. In Proceedings of the 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), Oslo, Norway, 4–6 June 2018; pp. 1–2.
36. Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of prediction confidence and domain extrapolation of two structure–activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* **2004**, *112*, 1249–1254. [PubMed]
37. Akhtar, N.; Saddique, M.; Asghar, K.; Bajwa, U.I.; Hussain, M.; Habib, Z. Digital video tampering detection and localization: Review, representations, challenges and algorithm. *Mathematics* **2022**, *10*, 168. [CrossRef]
38. Saddique, M.; Asghar, K.; Bajwa, U.I.; Hussain, M.; Aboalsamh, H.A.; Habib, Z. Classification of authentic and tampered video using motion residual and parasitic layers. *IEEE Access* **2020**, *8*, 56782–56797. [CrossRef]
39. Asghar, K.; Sun, X.; Rosin, P.L.; Saddique, M.; Hussain, M.; Habib, Z. Edge–texture feature-based image forgery detection with cross-dataset evaluation. *Mach. Vis. Appl.* **2019**, *30*, 1243–1262. [CrossRef]