Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776971

INTERNET OF THING (IoT) USING NEURAL NETWORK MODEL FOR IDENTIFYING ANTI-SOCIAL ACTIVITIES IN SURVEILLANCE MONITORING

Tuannurisan Sukitjanan[1], Sakesan Chana[2], Thaphat Chaichochok[3] and Surachet Channgam[4]

[1,2,3] Department of IoT and Information Technology, Faculty of Industrial Technology, Songkhla Rajabhat University, Thailand. E-mail: tuannurisan.su@skru.ac.th[1] , sakesan.ch@skru.ac.th[2], thaphat.ch@skru.ac.th[3]

[4] Program in Digital Business Technology, Faculty of Management Science, Chandrakasem Rajabhat University, Thailand. E-mail: surachet.c@chandra.ac.th

ABSTRACT

Sensors or electronic devices are used for remote monitoring. Real-time traffic, forest, military, commerce, and medical remote monitoring detects abnormalities. The computational complexity of computer vision-based video processing systems is substantial. This research develops a Slow-Fast Convolution Neural Network (SF-CNN) to detect and classify anomalous behaviour in surveillance videos. The suggested CNN architecture automatically learns video frames and selects the best object behavior attributes from a wide sample of films. SF-CNN learns slowly or quickly. When the frame rate is low, slow learning is enabled, and when high, rapid learning is enabled. The input video teaches both both spatial and temporal information. Actions identify humans, vehicles, and animals. All movies contain normal and aberrant activities in different circumstances. The SF-CNN architecture solves numerous limitations anomalous motions end-to-end. SF-CNN performance is tested on many benchmark datasets. The proposed method had 99.6% accuracy, higher than prior methods.

Keywords: Deep Learning, Neural Network, Video Processing, Internet of Thing

INTRODUCTION

Anit-Social activities are increasing day by day in innumerable fields. Theft, illegal, and other outlawed activities are considered anti-social activities and also must be identified immediately and protect the area as quickly as possible. It reduces the loss of data, things, and human death [1],[2]. The medical industry, forest, research centers, aerospace, vehicle stations, malls, most extensive building, etc., are some areas that need to be surveillance to avoid abnormal activities [3]. The surveillance system records the activities using CCTV cameras and continuously records them as videos. The output of the surveillance system is a video. The video is processed, and the abnormal activity is identified using the object detection and recognition method. The activity of the objects is classified as normal or abnormal. Today, there are ten (10) reasons business people need a video surveillance system. They are: Resolve Conflicts, Increase employee productivity, Reducing Theft, Better experience, Real-Time Monitoring, Enhancing safety, Digital storage, Evidence making, Access control, and Business savings [4, 5].

Though the surveillance monitoring system process is similar, the applications are different. The cameras' capacity and configuration and the surveillance systems have been changed based on the application. Some surveillance applications are given in Figure-1, which

shows surveillance in the office, road, official building, and backside of a house. Different intrusion detection systems have been proposed for security provision earlier, but it detected after the abnormal event. There is no methodology, which can stop the strange activities automatically or manually [6]. For preventing or controlling abnormal movements, location information is required with complete knowledge about the geo-region. It helps to identify the abnormal activity earlier and provides better security for the particular surveillance area. One of the best solutions to enhance the existing security is surveillance monitoring. To improve aberrant activity detection, this research designs and implements a deep learning-based convolution neural network [3, 7, 8]. Deep learning lets computers mimic human behavior. One deep learning algorithm detects street light signs, and another distinguishes a pedestrian from a cat, car, etc. It corrects grammar, spelling, repetitive words, punctuation, and more in given texts and generates a new copy with no errors. Today's fields progress more. The deep learning model helps computers classify speech, image, and text data. Deep learning achieves state-of-the-art accuracy by outperforming humans [9]. These models are trained with multi-layered neural networks and massive labeled data. Due to its accuracy, deep learning technology has many important applications. These electronic methods meet user expectations. Robotics, AI automobiles, picture caption creation, and other safety and essential applications use deep learning because it outperforms people. Deep learning [10-12] has been popular recently for the following reasons:

Deep learning requires enormous labeled data. For driverless automobiles, millions of photos and thousands of hours of footage are needed to train the computer.

Deep learning requires huge hardware and computer power. Deep learning networks take weeks to perform well. Instead of cluster or cloud computing, development teams use GPUs' huge parallel processing capacity to enhance performance and cut training time.

Office [1], Road [2], Building [3], House [4]

Figure-1. Various Applications of Surveillance System

Deep learning models are also called deep neural networks since they use neural network design. Deep neural networks can have 150 hidden layers, while regular neural networks can only have 2-3. "Deep" indicates a neural network's hidden layers. Deep learning models are trained using a huge number of labeled neural networks that can learn features directly from data. CNNs are popular deep neural networks. CNN uses 2D convolution layers and input data to convolute learned features, making it ideal for image processing.

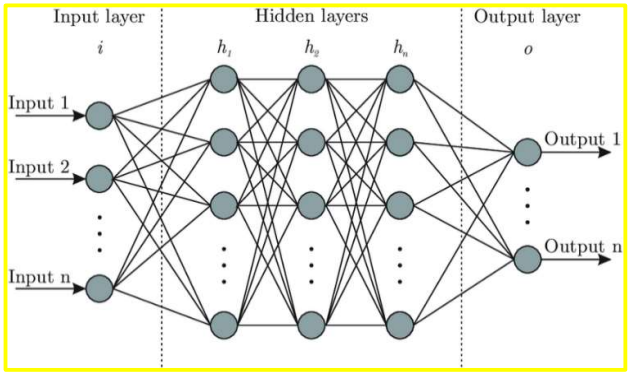


Figure-2. Structure of a Neural Networks

CNN classifies photos without manually identifying attributes. CNN directly extracts visual features. The network is trained using images, and the required features are learned simultaneously without pre-training. Automated feature extraction improves object classification in computer vision. CNNs need many hidden layers to recognize picture features. Hidden layers complicate visual features. The first and second hidden layers recognise edges and complex forms to distinguish an object.

Machine Learning Versus Deep Learning

Deep learning, which uses human-like artificial intelligence, is more efficient than machine learning. Machine learning algorithms parse data manually. It learns from data to create a model that categorizes the thing. Machine learning models use data to improve.

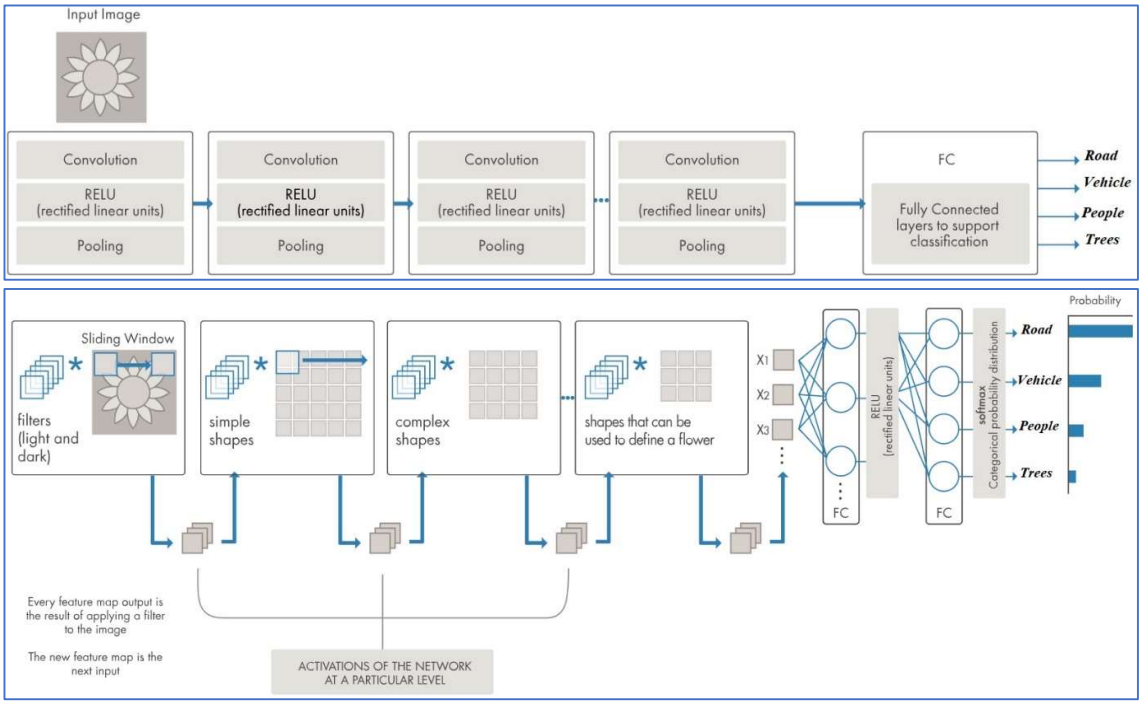


Figure-3. A Sample Layer Architecture of CNN - Training and Testing process

Inspired by the brain's Neural System, the deep learning model uses an Artificial Neural Network (ANN). ANN can examine data continually and make accurate decisions like a human brain. Data size improves deep learning. Machine learning offers many models and approaches to sort applications. Training a deep learning model requires millions of photos and hours of videos. A GPU processes data quickly. If neither is available, machine learning algorithms are better than deep learning.

The paper improves object recognition by learning video frames using two learning algorithms. They learn slowly and quickly. The paper's innovation is the deep learning model's frame rate-based video data analysis. It analyzes frames slowly and gets spatial semantics if the frame rate is low. Structures and temporal semantics are analyzed at high frame rates. This deep learning model learns spatial and temporal information for video processing and object recognition.

Limitation and Motivation

Though object identification and categorization for aberrant behaviors needs improvement, certain semi-automatic systems took more computational time, raising expense. Early research has focused on surveillance monitoring applications. The video learning method might extract a specific feature from the video, which reduces recognition accuracy. To tune results, the program changes the dataset's learning rate. Video processing models are time-consuming. Surveillance applications require a fast and accurate video recognition model. Thus, this study designs and implements a slow-fast deep CNN model to learn video frames with varied learning rates, retaining spatial and temporal characteristics to automatically improve object recognition accuracy.

Deep Learning-Based Object Detection and Recognition

The deep learning [2, 13-15] approach is ideal for video processing, object detection and recognition, pattern recognition, and speech recognition. This research suggested Deep Learning's Convolution Neural Network for object and anomaly detection. The application's wired or wireless CCTV cameras feed the system/PC's input footage. If hooked, the PC stores the video file. Routers send the video to the PC. This paper makes assumptions to clarify the method.

This paper detects abnormalities in surveillance videos using deep learning. Deep learning is inspired by neural network architectures that learn features and represent data. A neural network model has input, output, hidden, and more hidden layers. Deep learning has several massive, multilayer networks. Convolutional Neural Networks [16-18]are popular deep learning networks. (CNN). This research uses CNN architecture to detect aberrant activity. CNN automatically classifies video/image features. This study analyzes and assesses the suggested CNN architecture's performance on human, vehicle, and animal behaviors in varied backgrounds, a novel data processing method.

Proposed Approach

The proposed CNN architecture is explained in this section. The video V is divided into frames (images) F, in which various objects and their activities are normal and abnormal. Some of the specific abnormal activities are different activities than the usual activities. For improving the efficiency of video/image processing, the images are initially applied for preprocessing using a moving $\mathbf{3} \times \mathbf{3}$ average filter, which removes the noises occurring in the images. It can be represented as,

$$y_{ij} = \sum_{k=-m}^{m} \sum_{l=-m}^{m} w_{kl} x_{i+k,j+l}$$

Where the input image is represented as x_{ij}, (i, j) represents the pixels in the image, and y_{ij} represents the output image. Similarly, a linear filter with the size $\mathbf{3} \times \mathbf{3}$, is used on

$$(2m + 1)\ x\ (2m + 1)$$

having the weights w_{kl} for every k and l from $-m$ to m, equal to 1.

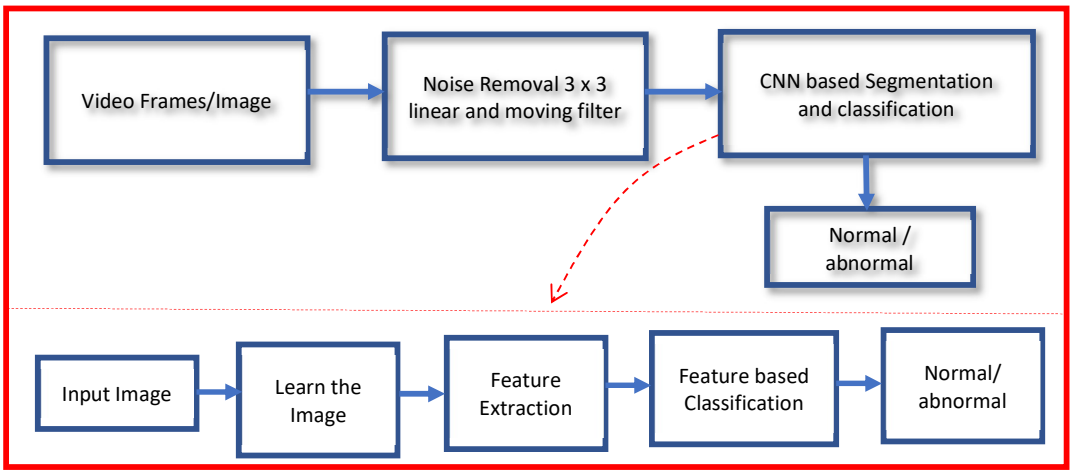
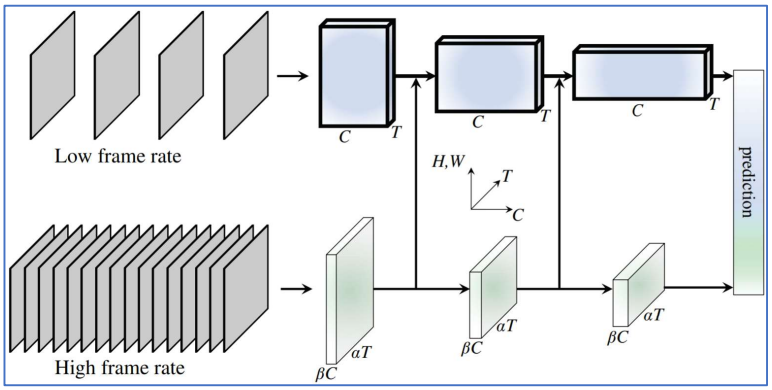


Figure-4. Deep Convolution Neural Network-Based Video Processing

Videos have more frames. One-minute videos have 100–110 frames. A program or manual process extracts video frames from input videos to create a video sequence. One action has numerous frames with the same objects and activities, wasting video processing time. Thus, selecting video frames speeds up and improves classification. 30% of video frames are labeled "normal" or "abnormal" for training to improve classification accuracy. Normal actions outnumber deviant ones. Labeling anomalous behaviors suffices for categorization, saving computing time, memory, and complexity. Processing selected and normal frames reduces time complexity. A database stores only aberrant frames and labels for training process references. Wrong/odd qualities distinguish aberrant video sequences. The remaining 70% of frames are used for testing and comparing retrieved features to taught features.

Input, output, and hidden layers form the CNN architecture. Convolutional, middle, and feature detection layers are buried. Classification layers (output layers) are fully linked and SoftMax. Resizing all frames to 32 * 32 lengthens training. Input layers deliver all input images to middle layers. The middle layer performs convolution, pooling, and rectification. Rectified

Linear Unit performs this. (ReLU). SF-CNN has three convolutional, two fully linked, and one SoftMax layer. Figure-5 depicts the suggested SF-CNN architecture. Figure-5(a) displays SF-CNN's frame rate-dependent learning mechanism. Figure-5(b) shows CNN's full capabilities.



(a).

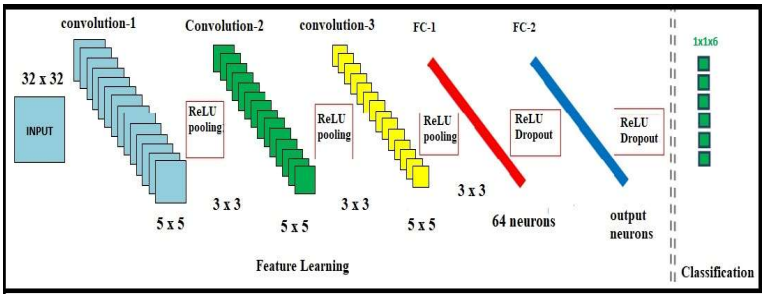


Figure-5. Proposed Deep Learning Architectures. (a). SF-CNN framework, (b). Functionality of CNN

Convolution filter activates image texture, edge, and corner information. These define frame actions. Abnormal actions are mismatched feature values. 32 First-layer convolution filters are 5-by-5-by-3. Input and color images are 5x5x3. A two-pixel pad surrounds the image's edges symmetrically. It maintains CN edges. ReLU zeros negative data to process only positive values. ReLU trains networks fastest. ReLU layer always follows pooling layer with 3 x 3 spatial pooling region and 2 pixels strides. 32 x 32 data is downsampled to 15 x 15. All three convolutions with pooling and ReLU layers are repeated to extract the most features and hidden information from the input image.

Fewer pooling layers can prevent data downsampling when key characteristics are removed early. Post-feature extraction CNN classifies. The network is fully linked and softMax. After ReLU, the first entirely connected layer has 64 neurons from the 32 × 32 input image. The second entirely connected layer generates categorization signals. CNN has input, middle, and final layers. Finally, SoftMax determines class distribution probability. The convolution layer weight is distributed using a random value 0.0001 (standard deviation) to reduce network loss during learning [19, 20].

SF-CNN

In this paper, deep CNN [4] is incorporated with the Slow-Fast learning method for analyzing the video segments. It comprises two parallel CNN modesl for the same input video-a Slow learning and Fast Learning. Generally, video content contains two different data: static and dynamic. The static data will not be changed or slowly changed, but the dynamic data will

continuously be changed (moving objects). According to Figure-5(a), the video frames obtained from fast streaming is input to slow frame rate learner since the slow learner can learn the output of the fast learning. The data format used in the SF learner is written as:

$$\boldsymbol{fast} = \{\boldsymbol{\alpha T}, \boldsymbol{S^2}, \boldsymbol{\beta C}\}$$

$$\boldsymbol{slow} = \{\boldsymbol{T}, \boldsymbol{S^2}, \boldsymbol{\alpha\beta C}\}$$

are used fused to create the SF learning process. The SF-CNN suggests a different methodology for transforming the data. The final one is the most efficient.

1. T-2-C (Time-2-Channel): data reshaping and transposing: $\{\boldsymbol{\alpha T}, \boldsymbol{S^2}, \boldsymbol{\beta C}\} \rightarrow \{\boldsymbol{T}, \boldsymbol{S^2}, \boldsymbol{\alpha\beta C}\}$, that is all the $\boldsymbol{\alpha}$ frames as one frame to channel.
2. TSS (Time-strided-sampling): take each $\boldsymbol{\alpha}$ frame as a sample, and it makes : $\{\boldsymbol{\alpha T}, \boldsymbol{S^2}, \boldsymbol{\beta C}\} = \{\boldsymbol{T}, \boldsymbol{S^2}, \boldsymbol{\beta C}\}$
3. TSC (Time-strided-convolution):: It performs as a three-dimensional convolution of a 5 x 12 kernel with $\boldsymbol{2\beta C}$ output channel and $\boldsymbol{\alpha}$ as stride.

Finally, global pooling operator slow and rapid learning lowers dimensionality in SF learning. The FC layer identifies output from both learners. FC classifies object behavior using SoftMax. SF-CNN is implemented via an algorithm. Any programming language is checked. SF-CNN algorithm-1.

RESULTS AND DISCUSSION

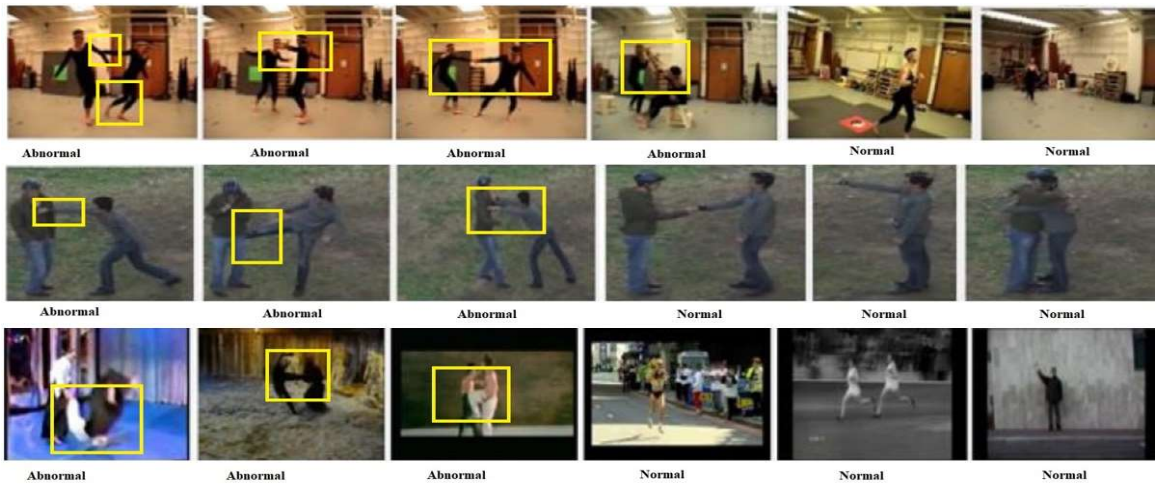
In this paper, the algorithm is implemented and verified in MATLAB software, where it contains a built-in CNN module in the Image Processing toolbox. It provides more functionalities and enables various inbuilt algorithms like regression methods, decision-making methods, and it can be chosen for performance verification. This paper uses seven datasets. Table-1 lists human, vehicle, and animal activity in the dataset. Each frame is an RGB image with varying pixel sizes. Other dataset photos are 276*236, 352*240, 360*288, etc. All photos are 32 × 32 pixels due to frame constraints. All frames have positives and negatives. 100 dataset movies were used for training and testing. 100 videos yielded 12000 frames. 5000 frames are normal, and others are problematic. The suggested framework has two experiment stages. Experiment-1 classifies normal and abnormal. Experiment-2 classifies all aberrant classes. Stochastic gradient descent with momentum trains networks. All network parameters are tweaked to acquire all output network properties. For network hyperparameter fine-tuning, the experiment uses 10–100 epochs and 0.001–0.1 learning rate. One epoch includes advancing and backward transmission of training samples with the learning rate.

Table-1. Dataset Information

Dataset	Videos	Total Frames	Frames - normal	Frames – abnormal
CMU Graphics Lab Motion [17]	11	2477	1209	1268
UT-Interaction dataset (UTI) [18]	54	5609	2706	2903
Peliculas Dataset (PEL) [19]	2	368	100	268
Hockey Fighting Dataset (HOF) [20]	12	1800	900	900
Web Dataset (WED) [21]	10	1280	640	640
UCSD-AD [22]	5	600	480	120

High-accuracy deep learning requires more inputs. Deep learning requires powerful GPUs. Intel core i7 runs MATLAB-2017 for CNN experiments. The experiment classifies photos in binary. Different datasets provide normal and abnormal photos, which are categorised. Human-related videos show walking, pointing, hugging, and shaking. Human-related videos show inappropriate kicking, shoving, and punching. The suggested CNN's classes are compared to the dataset's labeled classes to evaluate performance. Images show results.

Figure-7 shows CNN-classified normal and abnormal images from the experiment. Comparing and Figure-7 photos lets you assess the proposed CNN's performance. CNN classifies abnormalities in each frame using a binary classification model. Figure-7 shows frames with yellow bounding boxes highlighting aberrant activity. Figure-7 depicts pushing, kicking, fighting, walking in the wrong direction, crowds in the wrong locations, fighting, beating, cars on the roadside, cycles, cars, trucks, and jeeps in pedestrian roads. The experiment's learning rate can be used to discover abnormalities.



(a). Abnormality Detection in First three dataset

(b). Abnormality Detection in Four dataset

Figure-7. Abnormality Detection using proposed CNN

In both experiments, normal and abnormal, including all abnormal classes, are more accurate using the proposed CNN, which is understood from Figure-7. The performance of the proposed CNN is high and is evident by comparing both results given in Figure-6 and Figure-7. The abnormal detection accuracy is increased since the testing process is always compared with the training process. Hence, for human interacted classification, the training classes are highly accurate and are used in the testing process.

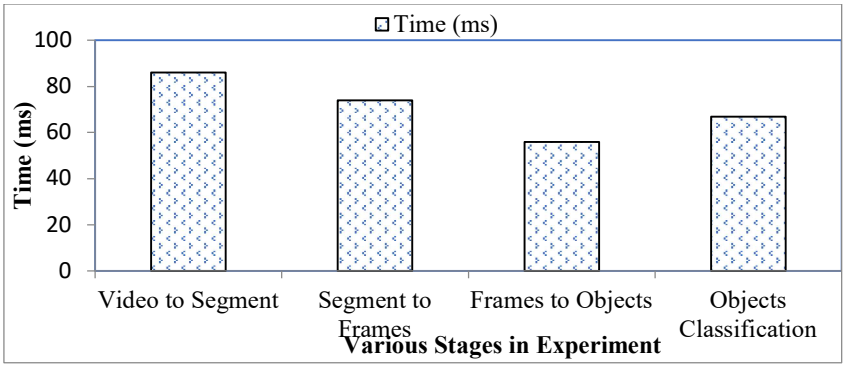


Figure-8. Time Complexity Calculation

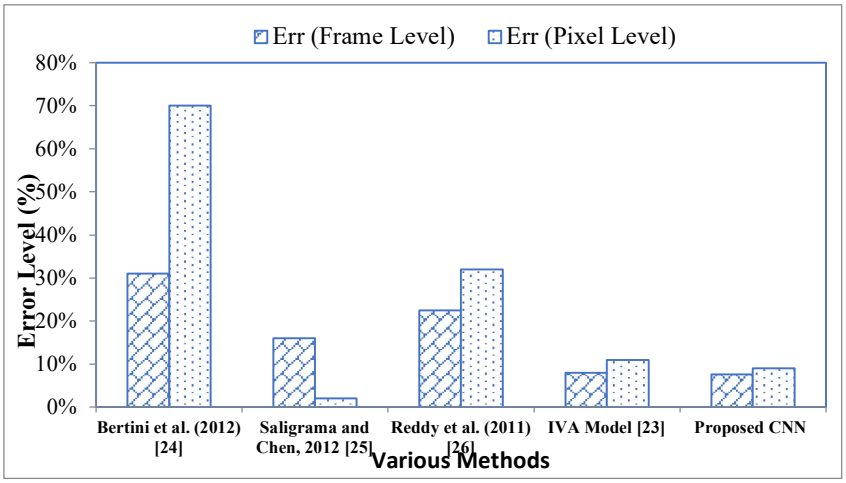


Figure-9. Dataset Based Error Analysis

Time and computational complexity can be used to evaluate the CNN technique. Figure-8 shows the calculated experiment's time complexity. Video to segment, segment to frames, object detection, and classification times are displayed. From the results, video processing takes longer than image processing. Thus, object identification and categorization are faster when video frames are transformed.

Sources, power failures, and format conversions can create video data mistakes. It ruins object detection and classification and image processing. Thus, frame- and pixel-level input frame faults must be calculated. Figure-9 illustrates the error prediction results from the proposed CNN technique and current methods as [19]All approaches calculate and compare low frame-level and high pixel-level faults. Using pixel levels, the proposed CNN approach has decreased frame error.

Eliminating error frames improves object detection and classification. The experiment tests 70% of frames and calculates categorization accuracy. Table-2 shows experiment object classification performance. Table-2 lists the investigation's total frames and accurately classified frames. 6035 of 12134 frames from all seven datasets are typical. All 6099 frames are aberrant, predefined, and verified by previous study.

Table-2. Performance Calculation

Data	Total Frames before the error	Total frames after error	Normal frames	Abnormal frames
DB	13,170	11,134	5,935	6,544
Proposed CNN	13,170	11,134	5,934	6,544

Table-2 shows that the suggested CNN architecture classifies 6034 frames as normal from 6035 and 6098 frames as abnormal from 6099. Table-2 calculates TP, TN, FP, FN, Sensitivity, Specificity, and accuracy to evaluate performance. Performance measurements are used to calculate accuracy. Figure-10 shows accuracy. Comparison results show that the suggested CNN architecture outperforms alternative methods. The proposed CNN scored

99.6%, greater than the literature survey techniques. Epochs and learning rate determine classification accuracy. CNN learning rates are 0.001, 0.01, and 0.1 for epochs 10–50. Different datasets, learning rates, and epochs are used to calculate accuracy. Table-3 shows findings. Epochs boost precision. Thus, the suggested approach uses 100 epochs to improve accuracy. The findings show that the suggested CNN has better time complexity, object detection, and classification accuracy than other techniques.

Table-3. Accuracy Based on Learning Rate

Learning Rate	**Max. No. of. Epochs**	**Dataset Accuracy (%)**					
		CMU	**UTI**	**PEL**	**HOF**	**WED**	**UCS-AD**
0.001	10	99.66	64.8	91.43	58.5	89.21	99.9
	20	99	56.55	91.43	99	99	99
	30	99	57.74	91.43	99	99	99
	40	99	70.18	91.43	99	99	99
	50	99	99.15	91.43	99	99	99
0.01	10	99	54.9	91.43	99	99	99
	20	99	99.6	91.43	99	99	99
	30	99	99.87	87.98	99	99	99
	40	99	98.84	99	99	99	99
	50	99	99.8	99	99	99	99
0.1	10	46.64	54.9	8.12	99	99	99
	20	0	0	8.13	99	99	99
	30	0	54.9	0	99	99	99
	40	0	54.9	91.37	99	99	99
	50	0	0	9.62	99	99	99

CONCLUSION

This research develops a deep learning-based surveillance system anomaly detection technique. Thus, this work creates a CNN architecture for education, information extraction, and video frame surveillance abnormality classification. This article identifies abnormalities across datasets. The goal is to create a universal anomalous identification system for human, animal, and vehicle surveillance[21, 22]. The deep learning model improves accuracy. Performance is tested by varying the learning rate and epochs. More epochs can boost accuracy.

Experimental results show that the suggested CNN outperforms competing methods. Anomaly categorization accuracy is 99.6%. The technology is totally autonomous and ideal for any monitoring system because it classifies abnormalities separately.

REFERENCES

1. Zhu, Z., Z. Xu, and J. Liu, *Flipped classroom supported by music combined with deep learning applied in physical education.* Applied Soft Computing, 2023. **137**.
2. Yu, C., X. Bi, and Y. Fan, *Deep learning for fluid velocity field estimation: A review.* Ocean Engineering, 2023. **271**.
3. Zhao, W., et al., *A learnable sampling method for scalable graph neural networks.* Neural Netw, 2023. **162**: p. 412-424.
4. Rajeshkumar, G., et al., *Smart office automation via faster R-CNN based face recognition and internet of things.* Measurement: Sensors, 2023. **27**.
5. Jiang, H., et al., *A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation.* Comput Biol Med, 2023. **157**: p. 106726.
6. You, K., C. Zhou, and L. Ding, *Deep learning technology for construction machinery and robotics.* Automation in Construction, 2023. **150**.
7. Zhu, G., et al., *Various solutions of the (2+1)-dimensional Hirota-Satsuma-Ito equation using the bilinear neural network method.* Chinese Journal of Physics, 2023.
8. Xie, G., et al., *A life prediction method of mechanical structures based on the phase field method and neural network.* Applied Mathematical Modelling, 2023. **119**: p. 782-802.
9. Mohammad-Rahimi, H., et al., *Deep learning: A primer for dentists and dental researchers.* J Dent, 2023. **130**: p. 104430.
10. Rayadurgam, V.C. and J. Mangalagiri, *Does inclusion of GARCH variance in deep learning models improve financial contagion prediction?* Finance Research Letters, 2023.
11. Ravikumar, K.C., et al., *Challenges in internet of things towards the security using deep learning techniques.* Measurement: Sensors, 2022. **24**.
12. Mohammed, A. and R. Kora, *A comprehensive review on ensemble deep learning: Opportunities and challenges.* Journal of King Saud University - Computer and Information Sciences, 2023. **35**(2): p. 757-774.
13. Fernando, K.R.M. and C.P. Tsokos, *Deep and statistical learning in biomedical imaging: State of the art in 3D MRI brain tumor segmentation.* Information Fusion, 2023. **92**: p. 450-465.
14. khelili, M.A., et al., *Deep learning and metaheuristics application in internet of things: A literature review.* Microprocessors and Microsystems, 2023. **98**.
15. Tanveer, M., et al., *Deep learning for brain age estimation: A systematic review.* Information Fusion, 2023. **96**: p. 130-143.
16. Alshehri, D.M., *Blockchain-assisted internet of things framework in smart livestock farming.* Internet of Things, 2023. **22**.

17. Huang, R., X. Yang, and P. Ajay, *Consensus mechanism for software-defined blockchain in internet of things.* Internet of Things and Cyber-Physical Systems, 2023. **3**: p. 52-60.
18. Si, L.F., M. Li, and L. He, *Farmland monitoring and livestock management based on internet of things.* Internet of Things, 2022. **19**.
19. Hemmati, A. and A.M. Rahmani, *The Internet of Autonomous Things applications: A taxonomy, technologies, and future directions.* Internet of Things, 2022. **20**.
20. Alshammari, H.H., *The internet of things healthcare monitoring system based on MQTT protocol.* Alexandria Engineering Journal, 2023. **69**: p. 275-287.
21. Siegel, J.W., et al., *Greedy training algorithms for neural networks and applications to PDEs.* Journal of Computational Physics, 2023. **484**.
22. Seo, M. and S. Min, *Graph neural networks and implicit neural representation for near-optimal topology prediction over irregular design domains.* Engineering Applications of Artificial Intelligence, 2023. **123**.