Enhancing Security Through Real-Time Classification of Normal and Abnormal Human Activities: A YOLOv7-SVM Hybrid Approach

Ashwin Shenoy M, *Member, IAENG,* Thillaiarasu N, Ashwin Shenoy

***Abstract*—Enhancing security is currently a paramount concern for society, as traditional surveillance methods necessitate constant vigilance and monitoring of cameras, which can be inadequate. To address this issue, developing an automated security system capable of real-time detection of abnormal human activities and taking appropriate actions is imperative. This paper introduces a novel approach for classifying human actions in controlled environments by combining a support vector machine (SVM) with the deep learning model You Only Look Once (YOLOv7). The YOLOv7 model calculates the boundaries of detected targets, which are then input into the SVM to enhance classification accuracy. The results demonstrate superior classification performance compared to alternative models. In practical terms, the proposed method achieves a testing accuracy 94.24% in classifying human activities based on real-world data. This approach offers promise for preemptively identifying abnormal actions before they occur, paving the way for further advancements in security methods.**

***Index Terms*—Activity Detection, SVM, YOLOv7, Controlled Environment, Normal Activity, Abnormal Activity**

I. Introduction

TODAY, video surveillance is frequently used to secure cities and individuals. The security system's outputs provide a wealth of information. The video data contains this information. Recognizing human behaviors is a difficult and crucial issue in computer vision. Among these, the identification of anomalous activity is crucial for intelligent monitoring [1]. The purpose of activity recognition is to identify the behaviors of several agents by utilizing sequences of observations on those agents' actions and the conditions of their surroundings [2]. The expressions action and activity are commonly confused. A sequence of successive motions made by a single item is commonly referred to as an action. A walking stride is an example of an atomic movement that the limbs may convey to represent this. In terms of activity, it refers to a series of acts. Dancing, for example, involves the sequential repetition of several activities, such as waving hands, walking, leaping, and so on. Actions and activities can be ranked differently [1] [3]. The demand for public safety emphasizes the significance of video data analysis and processing. Human activity is described as something that humans either do themselves, cause to happen, or are thought to be more sophisticated, and includes coordinated actions among several people [4].Numerous studies on intelligent surveillance have been conducted because of the increased need for security and safety. It can watch individuals in huge waiting rooms, retail malls, surveillance settings, healthcare systems (hospitals, eldercare, and home nursing), vehicles on campus, in and out of cities, bridges, tunnels, and more. Numerous of these apps aim to automatically identify high-level activities, which include several fundamental (atomic) human acts. Human activities are further classified into two categories: normal activities and abnormal activities. Human deviation from normal activity and causing harm to the surroundings or to himself or herself is categorized as an abnormal activity [6]. The proposed research proposes a novel system model that combines the traditional method of an SVM with the deep learning model of YOLOv7 to detect both normal and abnormal human activities in this study. The temple dataset [14], which consists of four normal actions (walking, sitting, and praying hands) and four abnormal activities (fighting, pickpocketing, forward fall, and chain snatching) in a temple context, is used to train the SVM and YOLOv7 models. The SVM uses the YOLO model's boundary boxes to enhance classification performance.

Manuscript received October 23, 2023; revised June 22, 2024.

Ashwin Shenoy M is a Research Scholar in School of Computing and Information Technology, REVA University, Bangalore, Karnataka, India (corresponding author to provide phone:+91 7760285429; e-mail: ashwinshenoy14@ gmail.com)

Thillaiarasu. N is an Associate Professor in School of Computing and Information Technology REVA University, Bangalore, Karnataka, India (email: thillai888@gmail.com)

Ashwin Shenoy is an Assistant professor in the Department of Computer Science and Engineering, NMAM Institute of Technology affiliated to Nitte (deemed to be University), Mangalore, India (email: ashwin.shenoy@nitte.edu.in)

II. Background

A well-known object detection approach in computer vision is the YOLO method [24]. YOLO, a real-time object recognition system, processes a photo in a single forward pass using a neural network. YOLO, as opposed to conventional object identification methods, combines bounding box regression and object recognition in a single phase [7]. As a result, it can process up to 60 frames per second, which makes it quick and effective. When a picture is divided into pixels, YOLO forecasts the bounding boxes for each cell. YOLO forecasts the class probability (the chance that a certain item will be in each bounding box) and the confidence score (the chance that an object will be in each box). Moreover, YOLO [8] could predict the cell's enclosing box coordinates.

YOLO is compatible with a wide range of dimensions and forms. One of YOLO's primary benefits is its speed. Because YOLO can evaluate photographs in real-time, it is highly suited for robots, surveillance systems, and autonomous vehicles [8]. In addition, YOLO is more efficient than standard object detection algorithms since it only needs to process a photo once [7]. In July 2022, the YOLOv7 model became

available to the general public [11]. Overall, YOLOv7 offers a more rapid and robust network architecture, a more efficient feature integration method, enhanced performance for object identification, a more reliable loss function, and increased label assignment and model training efficiency. Researchers now have access to revolutionary methodologies, the most noteworthy of which are various convolutional networks (CNN), thanks to advancements in computer technology and the rise of deep learning. Many studies have been published that use CNN to detect activity, including YOLOv1 [12], YOLOv2 [13], YOLOv3 [14], YOLOv4 [15], YOLOv5 [16], YOLOv6 [23], YOLOv7 [20], Mask R-CNN [17, 18], and work [19], with YOLOv7 achieving the fastest and most accurate results.

YOLOv7 also provides an enhanced way for feature integration. As a result, YOLOv7 requires far less expensive computational equipment than do other deep-learning models. The YOLOv7 detector is the most advanced in the YOLO family. With the help of this network's trainable bag of goodies, real-time detectors may improve accuracy while cutting inference costs. When extend and compound scaling are used together, the target detector can successfully reduce the number of parameters and calculations, resulting in a much higher detection rate [9]. With speeds ranging from 5 to 160 frames per second, in terms of accuracy and speed, YOLOv7 outperforms traditional object detectors. It also makes fine-tuning detection models simple and offers a selection of freeware that is ready for use. The YOLOv7 configuration file makes adding new modules and creating new models a breeze. The researchers propose E-ELAN, which continuously enhances the network's learning capability without modifying the beginning gradient path [10]. It does this through the use of cardinality to expand, shuffle, and merge.

The SVM is one of its most extensively explored classifiers, and several attempts to develop reliable methods for enhancing their performance and optimizing their hyperparameters have been documented in the literature [21, 22].

This survey highlights the significance of YOLO (You Only Look Once) as a real-time object detection method in computer vision, emphasizing its speed and efficiency. It introduces YOLOv7, a newer and more robust version known for improved feature integration and object recognition performance. YOLOv7 has great accuracy and speed, making it suited for a wide range of applications, and it makes optimal use of computing resources. Additionally, the survey mentions the use of Support Vector Machines (SVMs) as a well-researched classifier to enhance performance in object detection tasks.

III. MATERIALS AND METHODS

The strategy that has been suggested places an emphasis on efficiently discriminating between normal and abnormal actions while using the least amount of processing time. In the interest of maintaining security, the system has been instructed to categorize the acts of persons and to alert them to any irregularities that may occur.

Figure 1 depicts our proposed hybrid technique for activity categorization based on YOLOv7-SVM. YOLOv7, a modern object recognition system, can recognize and track objects in real-time, and SVM is a machine-learning method that is useful for classifying activities as normal or abnormal.

A. Experimental Dataset

There are four ordinary activities (standing, sitting, strolling, and praying hands) and four abnormal activities (forward fall, fighting, chain snatching, and pickpocketing) that are included in the dataset [14]. These actions and events take place in the surroundings of the temple. Labeling has been done on more than one thousand frames in preparation for the analysis. The picture shown in Figure 2 is an example of a dataset. A demonstration of how to annotate an image using the labelImg tool is shown in Figure 3. Specifically for this investigation, Figure 3 illustrates the categorization of activities into normal and abnormal areas.

B. YOLOv7-SVM Network

With the YOLOv7-SVM model, the benefits of YOLOv7 for person identification are combined with an SVM classifier to create a model that is capable of distinguishing between normal and abnormal human actions. The purpose of this method is to identify and classify various human activities that are caught in a controlled environment.

1) Human Detection with YOLOv7: People in frames are identified with the help of YOLOv7. YOLOv7 has received accolades for its exceptional accuracy and its ability to identify objects in real time, which makes it a perfect tool for doing human recognition in a short amount of time. Each convolutional layer performs convolutions by applying several filters to the feature map that is being input. The size of the output feature map (O) is:

$$O = \left(\frac{W - F + 2P}{S}\right) + 1 \tag{1}$$

where W is the input feature map size, F is the filter/kernel size, P is the padding, and S is the stride. The overlap of two bounding boxes is measured by the IoU. It is used to evaluate the accuracy of predicted bounding boxes. The formula for calculating IoU is:

$$IoU = \frac{AreaofIntersection}{AreaofUnion} \tag{2}$$

2) Object Extraction: After detecting humans, the bounding box coordinates are taken from the image's frame. These bounding boxes define the areas of interest (ROIs) that contain the human subjects. The bounding box coordinates give the top-left corner (TLx, TLy) of the ROI:

$$TLx = x_1 \quad and \quad TLy = y_1 \tag{3}$$

ROI Width and Height: The width (Wroi) of the ROI can be calculated as:

$$W_{roi} = x_2 - x_1 \tag{4}$$

The height (Hroi) of the ROI can be calculated as

$$H_{roi} = y_2 - y_1 \tag{5}$$

Extracting the ROI: Once you know the width, height, and coordinates of the ROI's top-left corner, you may extract the ROI from the image.

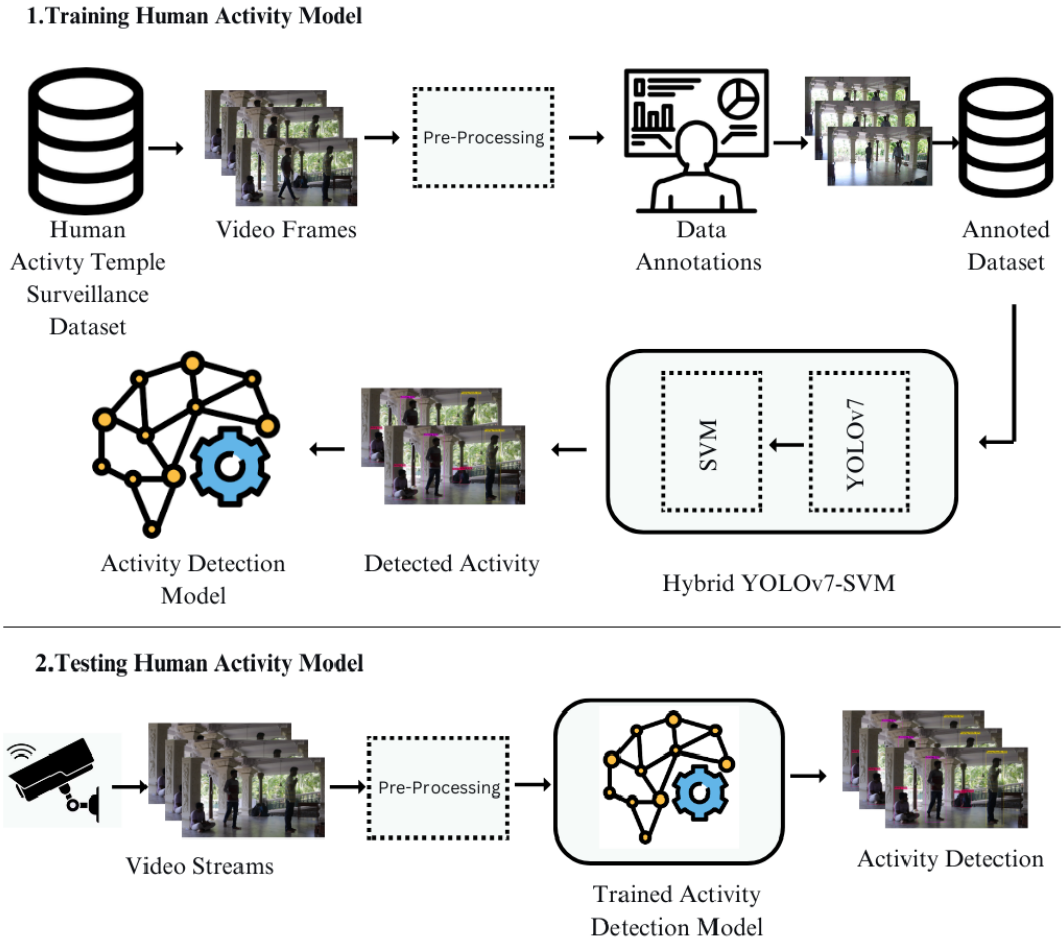

Fig. 1. A demonstration of the HAR's proposed framework. There are two stages to the framework. The hybrid YOLOv7-SVM model is first trained to extract video frames using a customized human surveillance dataset. The HAR model examines the video stream in the testing phase that follows to detect and identify various actions.

Fig. 2. Sample Dataset Images

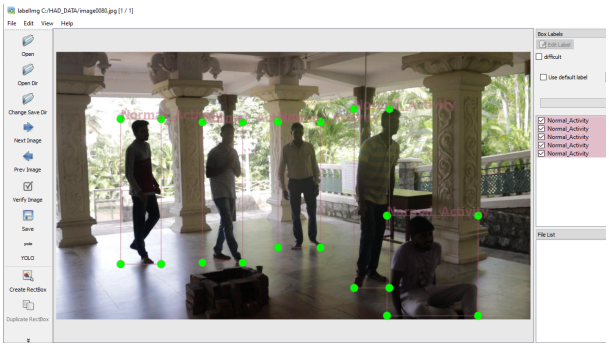

Fig. 3. Sample image annotation using labelImg tool

Feature Extraction To extract relevant information from human ROIs This stage tries to transform human-populated zones into fixed-length feature vectors that reflect important aspects of human activity.

3) Feature Representation and Classification: The SVM classifier uses the human ROIs' retrieved characteristics as input samples. Each feature vector corresponds to a specific human activity that needs to be classified.

$$FeatureVector = [F_1, F_2, \ldots, F_N] \quad (6)$$

where Fi represents the activation value of the ith neuron

in a specific layer.

Represent the characteristics of each ROI as a feature vector: For the i-th ROI, the feature vector can be represented as:

$\mathrm{X}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$ $\mathrm{X}_i = [x_i1, x_i2, ..., x_im]$,

where xij represents the j-th characteristic of the i-th ROI. Create the input matrix for the SVM classifier:

Combine the feature vectors of all ROIs into a matrix, X, where each row represents a feature vector:

$$\mathbf{X} = x_{11}x_{12}\cdots x_{1m}\ x_{21}x_{22}\cdots x_{2m} \vdots \ddots \vdots\ x_{n1}x_{n2}\cdots x_{nm} \tag{7}$$

This code will generate the matrix X with proper formatting and an equation number aligned to the right.

$$X = x_{11}x_{12}\ldots x_{1m}x_{21}x_{22}\ldots x_{2m} \vdots \ddots \vdots x_{n1}x_{n2}\ldots x_{nm} \tag{8}$$

Create the corresponding target or label vector. Create a target vector, y, where each element represents the class label of the corresponding ROI. Train the SVM classifier using the feature matrix and target vector. Once you have the feature matrix X and the target vector y, you can train the SVM classifier using these input samples.

YOLOv7 confidence score as YOLOconf and the SVM classification score as SVMclass.

$$Fused\_score = w \cdot YOLO\_conf + (1 - w) \cdot SVM\_class \tag{9}$$

where w is a weight factor that can be adjusted to control the influence of each algorithm's output on the final fused score.

The specific values and calculation of w will depend on the application and the performance characteristics of each algorithm.

Algorithm 1 Training Phase

1: **procedure** TRAINYOLOV7ANDSVM(labelled dataset)
2: Use a labeled dataset to train YOLOv7 to recognize objects and extract bounding boxes.
3: Using YOLOv7, extract features from the bounding boxes.
4: Based on the ground truth labels of the bounding boxes, assign labels to the features.
5: Using the labeled features and labels, train an SVM classifier.
6: **end procedure**

We implemented the YOLOv7-SVM architecture for testing throughout our experimental phase. A learning rate of 0.0013, a batch size of 64, 300 epochs, a C regularization set to 1.0, a kernel type raial basic function, and a degree of 3 were the unique hyperparameters that were set for this architecture.

IV. RESULTS AND DISCUSSION

A. Implementation details

To evaluate the performance of deep learning models and machine learning classifiers on the dataset, we ran a number of experiments. A dual NVIDIA Tesla P100 GPU with 3584 cores and a maximum throughput of 18.7 TeraFLOPS was used for training and testing our work. We worked with the Darknet DL-framework library.

Algorithm 2 Human Detection and Classification

1: **procedure** CLASSIFICATION(TestImage)
2: Use YOLOv7 to detect objects and produce bounding boxes
3: Extract features from the bounding boxes using YOLOv7
4: Classify the extracted features using the trained SVM classifier
5: Generate the final binary classification results using SVM predictions
6: **end procedure**

B. Evaluation Metrics

The predicted and ground truth object detections must be compared to calculate accuracy, precision, and recall for the YOLOv7 and SVM models used to classify human activity.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

Out of all the items that the model identified as positive (true positives + false positives), the accuracy quantifies the percentage of true positives (objects that were correctly recognized). It is calculable as:

$$precision = \frac{TP}{TP + FP} \tag{11}$$

Out of all the items in the ground truth dataset, the recall calculates the percentage of true positives. It is calculable as:

$$recall = \frac{TP}{TP + FN} \tag{12}$$

Average Precision (AP) is a scalar representation of the area under the PR curve, with larger values indicating better classifier performance.

$$AP = \int_0^1 Precision(Recall)\, dRecall \tag{13}$$

The model normally identifies two types of targets, and each class may draw a PR curve, resulting in an AP value. At the given threshold point, mAP is the average of all APs across all classes.

$$mAP = \frac{1}{class\_number} \sum_{i=1}^{class\_number} AP_i \tag{14}$$

C. State-of-the-Art comparison

In our comparison of state-of-the-art models, we assessed the performance of our YOLOv7-SVM hybrid model against several leading models: SVM, YOLOv4, YOLOv5, and YOLOv7. SVM, a traditional machine learning algorithm, achieved training and testing accuracies of 73.90% and 73.60%, respectively. YOLOv4, a sophisticated object detection model, significantly improved these figures to 90.94% and 90.70%. YOLOv5, with further enhancements in detection capabilities, achieved training and testing accuracies of

91.10% and 90.93%. YOLOv7, the latest iteration in the YOLO series, achieved even higher accuracies of 93.90% for training and 93.60% for testing. Our proposed YOLOv7-SVM hybrid model, which combines the strengths of deep learning and traditional machine learning, outperformed these models with training and testing accuracies of 94.56% and 94.24%. This comparison underscores the advancements in object detection and classification, highlighting the superior performance of our hybrid approach over both standalone machine learning and deep learning models shown in Table II and Table III.

D. Ablation Experiment

The ablation experiment is a methodical breakdown of the YOLOv7-SVM hybrid model, which is intended to improve security by classifying human behaviors in real time. The SVM classifier and the YOLOv7 object identification framework are the primary elements of this study and are thought to be essential parts of the model's operation. By carefully isolating each component, researchers break down the hybrid model and assess how each component's absence affects the system as a whole. The experiment examines the effectiveness of the SVM classifier alone by surgically removing the YOLOv7 framework, providing insight into its independent capacity to distinguish between normal and pathological activity in real-time. In a similar vein, researchers explore the inherent benefits of the YOLOv7 framework alone by removing the SVM classifier, revealing its effectiveness in identifying humans.

Furthermore, the ablation experiment looks into the complex relationship between hyperparameters and model performance. Researchers investigate the various nuances that define the model's behavior by making systematic changes to hyperparameters and settings inside each component. This detailed investigation allows for a thorough grasp of how particular variables affect the model's performance in real-time activity categorization. As researchers carefully negotiate this maze of settings, they reveal secret insights into the model's inner workings, opening the way for improved hybrid architectures designed for increased security applications. Thus, using this rigorous and methodical methodology, the ablation experiment emerges as a cornerstone in the drive to develop and optimize the YOLOv7-SVM hybrid model, enabling security advances through real-time activity categorization.

A low confidence level is often used when selecting model training parameters to better extract features and detect targets. The ablation experiment was thus carried out with a confidence level of 0.5. Table II contains rigorously collected and evaluated experimental data, including model performance metrics, which provide vital insights into the effectiveness of the YOLOv7-SVM hybrid model in real-world security applications. Based on the context and information supplied, here's a potential depiction of values for the ablation experiment Table I.

These data are a hypothetical depiction of the YOLOv7-SVM hybrid model's training and testing accuracies under various experimental settings in the ablation experiment. The basic YOLOv7-SVM model has the maximum training and testing accuracies, whereas changes such as eliminating the SVM or YOLOv7 components or changing the hyperparameters result in significantly lower performance.

E. Visualization and Analysis

This research makes use of an image dataset that was acquired in a temple-controlled setting. Cross-validation is performed by dividing the training and testing datasets by 80% and 20%, respectively. As illustrated in Figure 4 and Table II, the average precision for normal activity was 94.18%, the average precision for abnormal normal activity was 92.70%, and all classes were 93.40% mAP@0.5 in the proposed model.

1) Comparative Experiment: Furthermore, the proposed hybrid of YOLOv7-SVM and several models such as YOLOv4, YOLOv5, CNN, and SVM are examined. Table II compares the proposed model's testing and training accuracy to that of the standard model. Figure 3 shows the comparison graph of the proposed model with a traditional model.

Four separate datasets were utilized to validate the proposed model. Four datasets are publicly accessible and commonly utilized in HAR research, including those from KTH [25], Kinetics [28], Weizmann [27], and IXMAS [26]. The main characteristics of these datasets are explained below.

1. KTH: The dataset comprises a collection of movies including 6 distinct activities (boxing, clapping, jogging, handshaking, running, walking), executed by over 25 individuals in four distinct settings, with three of them taking place outdoors. A frame rate of 25 frames per second (FPS) is used for encoding.

consisting of a total of 600 segments. The videos have a mean length of 5 seconds, with a total of 100 frames per video. The recordings are made at a resolution of 160 × 120 pixels, with a single color channel representing grayscale.

2. Kinetics: The dataset is called DeepMind Kinetics and it consists of videos capturing human actions. The dataset comprises 400 distinct human activity categories, each accompanied by a minimum of 400 video clips depicting the action. Each clip has a duration of around 10 seconds and is sourced from a distinct YouTube video. The acts mostly revolve around humans and include a wide array of categories, including interactions between humans and objects, such as playing instruments, as well as interactions between humans, such as shaking hands.

3. Weizmann: The Weizmann dataset has nine actors doing ten distinct sorts of motions, including running, strolling, skipping, forward leap, up-down jump, galloping, 2-hand waving, 1-hand waving, and leaning. The dataset has a total of 93 sequences recorded using a stationary camera with a resolution of 180 by 144 pixels and a frame rate of 25 frames per second. Additionally, there are 10 sequences of walking caught from various viewpoints. The acquired data undergoes background subtraction, which involves removing the backdrop and including the activities occurring in the background in the dataset.

4. IXMAS: The INRIA Xmas Motion Acquisition Sequences (IXMAS) collection comprises 13 everyday life behaviors, including checking watch, crossing arms, scratching head, sitting down, getting up, turning around, walking, waving, punching, kicking, pointing, picking, overhead throwing, and bottom-up tossing. The activities are executed three times by a group of 11 actors, resulting in a dataset containing 2154 sequences. The dataset was acquired using 5 synchronized

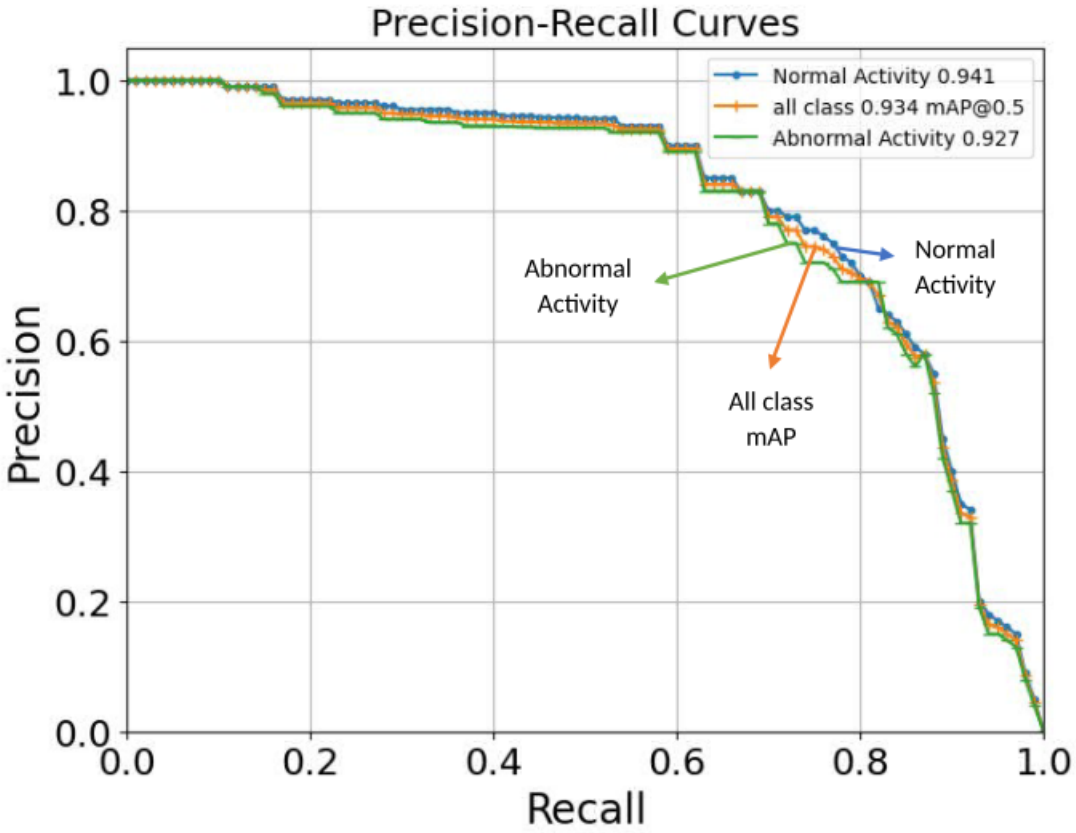

Fig. 4. The precision-recall curve.

TABLE I
ABLATION EXPERIMENT.

Sl No	Model	Training Accuracy	Testing Accuracy
1	YOLOv7-SVM (Baseline)	94.56%	94.24%
2	YOLOv7-SVM (SVM removed)	93.80%	93.50%
3	YOLOv7-SVM (YOLOv7 removed)	93.80%	93.50%
4	YOLOv7-SVM (Hyperparameters varied)	93.80%	93.50%

TABLE II
MODEL PERFORMANCE METRICS COMPARISON RESULTS DETAIL BETWEEN YOLOV4,YOLOV5, YOLOV7 AND PROPOSED MODEL

Model	Category	Precision	Recall	mAP@0.5
YOLOv4	Normal Activity	90.97	85.10	90.89
	Abnormal Activity	89.20	82.50	87.80
YOLOv5	Normal Activity	91.78	86.56	91.20
	Abnormal Activity	87.60	83.81	88.23
YOLOv7	Normal Activity	94.10	87.10	93.78
	Abnormal Activity	92.70	84.15	90.60
Proposed Model	Normal Activity	94.58	89.16	94.18
	Abnormal Activity	93.51	86.45	92.70

TABLE III
TESTING AND TRAINING ACCURACY.

Sl No	Model	Training	Testing
1	SVM	73.90%	73.60%
2	YOLOv4	90.94%	90.70%
3	YOLOv5	91.10%	90.93%
4	YOLOv7	93.90%	93.60%
5	Proposed Method	94.56%	94.24%

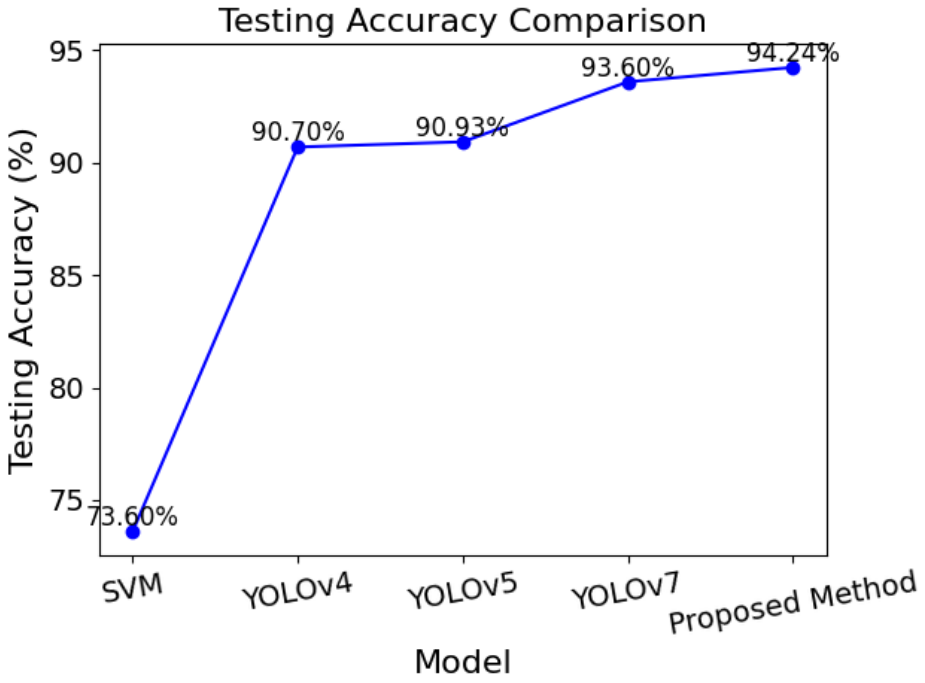

Fig. 5. Testing Accuracy Comparison graph.

and calibrated firewire cameras, including the silhouettes and visual hulls. The average accuracy comparison across different datasets is presented in Table IV.

V. CONCLUSION AND FUTURE SCOPE

The study provided a method for classifying human activity into normal and abnormal categories using YOLOv7-SVM in a controlled setting. First, the target data was extracted from the annotated temple dataset. The YOLOv7-SVM models are trained using the suggested model. The SVM gets boundary boxes from the YOLOv7 model and improves classification performance. We also ran our dataset via different deep-learning algorithms. The proposed model performed better in classification, with an accuracy of 94.24%. In the future, this research will be extended to collective abnormal actions to strengthen the accuracy of our classification and to be applied to predicting the activity before it occurs, allowing us to prevent abnormal activity.

REFERENCES

[1] A. S. M and N. Thillaiarasu, "A Survey on Different Computer Vision-Based Human Activity Recognition for Surveillance Applications," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1372-1376, doi: 10.1109/ICCMC53470.2022.9753931.

[2] B. Gottfried and H. K. Aghajan, "Behaviour Monitoring and Interpretation- BMI: Smart Environments," Vol. 3, Ios Press,2009.

[3] H. Hendriks-Jansen, Catching ourselves in the act: Situated activity, interactive emergence, evolution, and human thought. MIT Press, 1996.

[4] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 18, issue 11, pp. 1473-1488,2008.

[5] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A Review on Video-based Human Activity Recognition," Computers, Multidisciplinary Digital Publishing Institute, Vol. 2, issue 2, pp. 88-131,2013.

[6] J. K. Aggarwal and M. S. Ryoo, "Human Activity Analysis: A Review," ACM Computing Surveys (CSUR), ACM, Vol. 43, issue 3, pp. 1-47,2011.

[7] Z. Xue, R. Xu, D. Bai, and H. Lin, "YOLO-tea: A tea disease detection model improved by YOLOv5," Forests, vol. 14, no. 2, p. 415, 2023. Available: https://doi.org/10.3390/f14020415

[8] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," Proc. Comput. Sci., vol. 199, pp. 1066–1073, 2022. Available: https://doi.org/10.1016/j.procs.2022.01.135

[9] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv preprint arXiv:2207.02696, 2022. Available:https://doi.org/10.48550/arXiv.2207.02696

[10] F. Yang, X. Zhang, and B. Liu, "Video object tracking based on YOLOv7 and DeepSORT," arXiv preprint arXiv:2207.12202, 2022. Available:https://doi.org/10.48550/arXiv.2207.12202

[11] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," arXiv, arXiv:2207.02696, 2022.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 779–788,2016.

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in Proceedings of the Cvpr2017, Honolulu, HI, USA, pp. 187–213,2016.

[14] A. Shenoy, M. and N. Thillaiarasu, "Enhancing Temple Surveillance through Human Activity Recognition: A Novel Dataset and YOLOv4-ConvLSTM Approach," 1 Jan. 2023, pp. 11217–11232.

[15] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv, arXiv:2004.10934, 2020.

[16] R. Couturier, H. N. Noura, O. Salman, and A. Sider, "A Deep Learning Object Detection Method for an Efficient Clusters Initialization," arXiv, arXiv:2104.13634, 2021.

[17] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in Proceedings of the ICCV, Venice, Italy, 22–29 October 2017.

[18] Y. Wu, A. Kirillov, F. Massa, W. Y. Lo, and R. Girshick, "Detectron2," 2019. Available online: https://github.com/facebookresearch/detectron2

[19] Q. Gao, J. Liu, Z. Ju, L. Zhang, Y. Li, and Y. Liu, "Hand Detection and Location Based on Improved SSD for Space Human-Robot Interaction."

[20] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv, arXiv:2207.02696, 2022.

[21] W. Dudzik, J. Nalepa, and M. Kawulok, "Evolving data-adaptive

TABLE IV
ACCURACY OF THE PROPOSED MODEL WITH DIFFERENT DATASETS.

Sl No	Dataset	Accuracy
1	KTH	93.90%
2	Kinetics	90.94%
3	Weizmann	93.10%
4	IXMAS	92.90%

support vector machines for binary classification," Knowledge-Based Systems, vol. 227, p. 107221, 2021.
[22] M. A. Chandra and S. S. Bedi, "Survey on SVM and their application in image classification," International Journal of Information Technology, vol. 13, pp. 1-11, 2021.
[23] Y. Li and X. Zhang, "Object Detection for UAV Images Based on Improved YOLOv6," IAENG International Journal of Computer Science, vol. 50, no. 2, pp759-768, 2023.
[24] N. Tran, H. Nguyen, H. Luong, M. Nguyen, K. Luong, and H. Tran, "Recognition of Student Behavior through Actions in the Classroom," IAENG International Journal of Computer Science, vol. 50, no. 3, pp1031-1041, 2023.
[25] KTH, "KTH - recognition of human actions dataset," 2004. Available:https://www.csc.kth.se/cvap/actions/
[26] EPFL, "IXMAS actions – new views and occlusions," 2006. Available:https://www.epfl.ch/labs/cvlab/data/data-ixmas10/
[27] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2247–2253, 2007.
[28] W. Kay et al., "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.

Ashwin Shenoy M is a dedicated individual with a strong background in the fields of computer science and engineering. As a Research Scholar at Reva University in Bangalore, India, Ashwin is actively engaged in cutting-edge research, contributing to advancing knowledge in the field. His dual role as both an educator and a researcher demonstrates his commitment to fostering the growth and development of future professionals in the fields of computer science and engineering. Ashwin Shenoy's academic and research pursuits make him an essential asset in the pursuit of excellence within his institution and the broader academic community.

Dr.N.Thillaiarasu currently working as an Associate Professor in the School of Computing and Information Technology, REVA University, Bengaluru, from January 2021. He has also served as an assistant professor at Galgotias University, Greater Noida, from July 2019 to December 2020. He worked for 7.3 years as an assistant professor in the Department of Computer Science and Engineering, SNS College of Engineering, Coimbatore. He obtained his B.E. in Computer Science and Engineering from Selvam College of Technology in 2010 and his M.E. in Software Engineering from Anna University Regional Center, Coimbatore, in 2012. He received his Ph.D. from Anna University, Chennai, in 2019. His areas of interest include cloud computing, security, IoT, and machine learning.

Ashwin Shenoy Currently serving as an assistant professor in the Department of Computer Science and Engineering at the NMAM Institute of Technology, affiliated to Nitte (deemed to be university), Mangalore, India, Ashwin Shenoy brings a wealth of knowledge and experience to the academic community.