Anomaly Detection in Human Task Performance with Video-Based Contextual Reasoning

Project report in partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering (Artificial Intelligence and Machine Learning)

Submitted By

Tuhin Mondal	Enrollment No. 12021002028165
Ishita Karmakar	Enrollment No. 12021002028026
Arik Das	Enrollment No. 12021002028012
Swarnanko Saha	Enrollment No. 12021002028149
Aishi Paul	Enrollment No. 12021002028017

Under the guidance of

Prof. Sramana Mukherjee

Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning)

UNIVERSITY OF ENGINEERING AND MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

UNIVERSITY
OF ENGINEERING AND MANAGEMENT
(Established by Act XXV of 2014 of Govt. of West Bengal & recognized by UGC, Ministry of HRD, Govt. of India)
University Area, Plot No. III-B/5, Main Arterial Road, New Town, Action Area – III, Kolkata - 700160, WB, India
Admission Office: 'ASHRAM', GN-34/2, Salt Lake Electronics Complex, Kolkata - 700091, WB, India

Ph.(Office) : 913323572969
: 913323577649
Admissions : 913323572059
Fax : 913323578302
E-mail : vc@uem.edu.in
Website : www.uem.edu.in

CERTIFICATE

This is to certify that the project titled **Anomaly Detection in Human Task Performance with Video-Based Contextual Reasoning** submitted by Tuhin Mondal (University Registration No. 304202100900876 of 2021 - 2022), Swarnanko Saha (University Registration No. 304202100900861 of 2021 - 2022), Ishita Karmakar (University Registration No. 304202100900744 of 2021 - 2022), Aishi Paul (University Registration No. 304202100900735 of 2021 - 2022), and Arik Das (University Registration No. 304202100900730 of 2021 - 2022) students of the University of Engineering and Management, Kolkata, in partial fulfillment of requirement for the degree of Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning), is a Bonafide work carried out by them under the supervision and guidance of Prof. Sramana Mukherjee during 8th Semester of academic session of 2024 - 2025. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

_____ _____

Signature of Guide
<Name & Designation>

Signature of Head of the Department
Department of CSE (AI & ML)

ACKNOWLEDGEMENT

We would like to express our profound gratitude to Prof. Sramana Mukherjee, Project Mentor, Prof. (Dr.) Sudipta Sahana, Head of the Department of CSE (AI & ML), and Prof. (Dr.) Sajal Dasgupta, Vice Chancellor, University of Engineering and Management, Kolkata for their contributions to the completion of my project titled **Anomaly Detection in Human Task Performance with Video-Based Contextual Reasoning** We would like to express special thanks to our project mentor for his/her time and the efforts he/she provided throughout the semester. Your useful advice and suggestions were helpful to us during the project's completion. In this aspect, we are eternally grateful to you. We want to acknowledge that this project was completed entirely by ourselves and not by someone else. Last, but not least, we would like to extend our warm regards to our family members and peers who have kept supporting us and always had faith in our work.

Tuhin Mondal

Ishita Karmakar

Aishi Paul

Arik Das

Swarnanko Saha

TABLE OF CONTENTS

ABSTRACT

Understanding and analyzing human task performance is critical in domains such as healthcare, manufacturing, and surveillance. This study presents a computer vision-based framework for anomaly detection using YOLO, a state-of-the art object detection model, to track and interpret task-related behaviors from video inputs. The system first establishes a baseline of normal activity by learning object usage, movement patterns, and interaction sequences from initial videos of a person performing a specific task. New video inputs are then evaluated to identify anomalies such as unfamiliar object interactions, unusual pauses, and deviation in motion patterns. YOLO enables precise detection and localization of objects and interactions, forming the foundation for reliable anomaly analysis. The framework further enhances explainability by providing textual reasoning for each detected anomaly, offering actionable insights. The proposed system combines robust real-time object detection with behavior understanding to deliver a scalable and interpretable solution for task performance monitoring. Index Terms—Anomaly detection, deep learning, object detection, video surveillance, YOLO.

INTRODUCTION

Monitoring deviations in human task execution is essential in many real-world applications, including healthcare, industrial automation, and personal assistance systems. Conventional anomaly detection techniques often depend on static thresholds or handcrafted features, limiting their ability to adapt to dynamic environments and diverse human behaviors. With the advancements in deep learning and object detection, particularly through models like YOLOv5, there is a new opportunity to develop systems capable of real-time, accurate.

and context-aware performance monitoring. In this work, we utilize YOLO to develop a deep learning-based anomaly detection framework that analyzes video footage of individuals performing routine tasks. YOLO is employed to detect and track objects involved in the task, allowing the model to learn typical behavioral patterns, such as the timing of actions, object interactions, and spatial arrangements. Once a behavioral baseline is established from training videos, new inputs are analyzed to detect anomalies—ranging from interacting with unexpected objects to exhibiting irregular movement or timing. Each anomaly is accompanied by a human-readable textual explanation, enhancing the system's interpretability. The key contributions of this research include: Leveraging YOLO for object detection and behavioral modeling in task execution. A novel framework for detecting deviations from learned performance baselines using temporal and spatial features. Generating interpretable textual explanations for anomalies to support decision-making in real-time monitoring. This approach blends the precision of modern object detection with the need for personalized and explainable behavioral analytics, enabling a new class of intelligent monitoring systems.

LITERATURE SURVEY

Anomaly detection in human behavior analysis has gained significant traction in recent years, particularly with the increasing accessibility of video data and advancements in deep learning techniques. Traditional approaches primarily relied on handcrafted features and statistical methods to detect. deviations in activity. For instance, Simonyan and Zisserman introduced a two-stream convolutional network to model hu man motion patterns, combining spatial and temporal features to enhance action recognition accuracy [Simonyan and Zis serman, 2014]. With the rise of deep learning, convolutional neural networks (CNNs) have revolutionized video-based hu man activity recognition. Works like that of Pirsiavash and Ramanan utilized 3D CNNs and recurrent neural networks (RNNs) to capture spatiotemporal dependencies and detect irregular behaviors in first-person video views [Pirsiavash and Ramanan, 2012]. However, these models often require extensive computational resources and large-scale datasets to generalize well across diverse task scenarios. Object detection frameworks have increasingly been applied to enhance human activity understanding. YOLO (You Only Look Once), particularly YOLOv5 developed by Ultralytics, has emerged as a lightweight yet powerful real-time object detector capable of localizing and classifying objects with high accuracy. Hou demonstrated the use of YOLOv5 for human detection and be haviour analysis in surveillance systems [Hou, 2023], showing that the integration of object detection with behavior reasoning significantly improves anomaly detection performance.

More recent works have explored explainable AI (XAI) in behavioral

analysis to provide interpretable outputs. For ex ample, Fan proposed a combination of YOLOv5 with Slow-Fast networks for real-time action detection using PyTorchVideo, enabling models to focus attention on key temporal segments for anomaly explanation [Fan, 2021]. However, few have integrated such reasoning with fast object detection like YOLO to produce real-time, interpretable insights. Our approach builds upon this body of work by leveraging YOLO not just for object detection, but as a foundation for modelling task-specific interactions and identifying behavioral anomalies. Furthermore, the inclusion of contextual textual reasoning bridges the gap between high-performance detection and human interpretability, which remains underexplored in existing literature

PROPOSED SOLUTION

The proposed system follows a comprehensive pipeline designed for real-time anomaly detection:

1) Dataset Preparation: Video footage of workplace activities was collected and annotated using Roboflow to distinguish normal and anomalous behaviors.
2) Model Selection: YOLOv5 was selected because of its superior performance and lightweight nature in object detection and real-time processing capabilities.
3) Video Processing: The frames were extracted from the video streams and analyzed by the YOLOv5 model to identify anomalies.
4) Output Generation: The detected anomalies were recorded in a CSV file with timestamps and visual annotations were overlaid on the video output.

This structured approach ensures scalability and adaptability to various workplace environments

EXPERIMENTAL SETUP AND RESULT ANALYSIS

A. Data Labeling with Roboflow :

Roboflow facilitated the annotation of video frames, catego rizing activities into “normal” (e.g., standard work tasks) and “anomalous” (e.g., accessing restricted websites). The tool’s augmentation features enhanced dataset diversity, improving model robustness.

B. Model Training :

The YOLOv5 model was fine-tuned on the annotated data set using a pre-trained base model. The training parameters included a batch size of 16, 10 epochs and an image resolution of 640×640 pixels, optimized for surveillance footage from the workplace.

C. Inference :

During inference, video frames were processed at 30 frames per second, with YOLOv5 detecting anomalies in real time. Confidence thresholds were set to 0.5 to balance precision and recall.

D. Result Storage :

Anomalies were logged in a CSV file with details such as frame number, timestamp, and anomaly type. Video output included overlaid bounding boxes and captions for visual confirmation.

E. Mathematical Foundations :

The implementation leverages several key equations under pinning YOLO's functionality

F. Bounding Box Prediction, Loss Function, and Confidence Score in YOLOv5 for Anomaly Detection :

In YOLOv5, detecting anomalies within workplace surveil lance relies on three critical aspects: bounding box pre diction, loss function optimization, and confidence score evaluation. These elements work together to ensure accurate object localization, classification, and filtering of predictions, making the system effective in real-time anomaly detection.

G. Bounding Box Prediction :

YOLO (You Only Look Once) uses a smart and efficient strategy for detecting objects in images. Instead of scanning an image multiple times like older methods, YOLO looks at the entire image once and makes all its predictions in a single pass.

Here's how it works:

• YOLO divides the input image into a grid (e.g., 7×7 or 13×13).

• Each grid cell is responsible for detecting objects whose centers fall inside it.

• For every grid cell, YOLO predicts:

– One or more bounding boxes (which define object positions),

– Confidence scores (how sure it is about object presence)

–And class probabilities (what the object is—e.g., person, car).

Think of each grid cell as a watch tower scanning a small region. If the center of an object lies within a cell's region, that cell detects and describes it.

This approach allows YOLO to:

• Detect multiple objects across locations,

• Operateinreal-time with highspeed,

• Maintain accuracy with a single neural network pass.

Why it's explainable:

• Each cell has a clear detection responsibility.

• The entire image context is still visible to the model.

• Bounding boxes make outputs visually interpretable.

Bounding Box Prediction :

The bounding box is predicted using the following equations::

$$b_x = \sigma(t_x) + c_x \quad (1)$$
$$b_y = \sigma(t_y) + c_y \quad (2)$$
$$b_w = p_w * e^{(t_w)} \quad (3)$$
$$b_h = p_h * e^{(t_h)} \quad (4)$$

Where:
- bx, by are the predicted center coordinates of the bounding box.

- bw, bh are the predicted width and height of the bounding box.
- σ is the sigmoid function, which ensures the outputs remain within a bounded range.
- tx, ty, tw, th are the raw outputs of the network.
- cx, cy are the offsets of the grid cell.
- pw, ph are the dimensions of the anchor box, which help the model learn object scales effectively.

This formulation enables precise localization of workplace anomalies such as unauthorized website access and excessive idle screen time.

Loss Function: Optimizing Detection Performance
The training process of YOLO relies on a composite loss function that balances three key objectives:
- Localization Loss: Ensures accurate bounding box predictions.
- Confidence Loss: Encourages the model to assign high

confidence to true objects and low confidence to background regions.

- Classification Loss: Helps the model correctly categorize detected objects (e.g., normal vs. anomalous activities).

The overall loss function is defined as:

$$L = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{obj\_ij} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] \quad (5)$$

$$+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{obj\_ij} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] \quad (6)$$

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{obj\_ij} (C_i - \hat{C}_i)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{noobj\_ij} (C_i - \hat{C}_i)^2 \quad (7)$$

$$+ \sum_{i=0}^{S^2} \mathbb{1}_{obj\_i} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \quad (8)$$

Where:
- $\mathbb{1}_{obj\_ij}$ is 1 if object exists in cell i and bounding box j, 0 otherwise.
- $\mathbb{1}_{noobj\_ij}$ is 1 if no object exists in cell i and box j.
- λ_{coord} and λ_{noobj} are weighting factors to balance localization precision and background suppression.
- C_i, $\hat{C}_i$ are the predicted and actual confidence scores.
- $p_i(c)$, $\hat{p}_i(c)$ are the predicted and ground-truth probabilities for class c.

This loss function ensures accurate detection and classification of anomalies.

Confidence Score: Filtering Detections
During inference, YOLO assigns a confidence score to each detected anomaly, calculated as:

$$Confidence = Pr(Object) \times IoU_{truth,pred} \quad (9)$$

Where:
- Pr(Object) is the probability that an object is present in the bounding box.
- IoU_truth,pred measures the Intersection over Union between the predicted and actual bounding boxes.

A confidence threshold (e.g., 0.5) is applied to filter out weak detections, ensuring only reliable anomalies are flagged. Lowering the threshold increases sensitivity but may introduce false positives, while increasing it improves precision at the cost of potentially missing some anomalies.

Interconnections and Model Improvements:

1. Bounding box prediction generates multiple candidate boxes per grid cell.

2. The loss function ensures accurate predictions by minimizing localization, classification, and confidence errors.

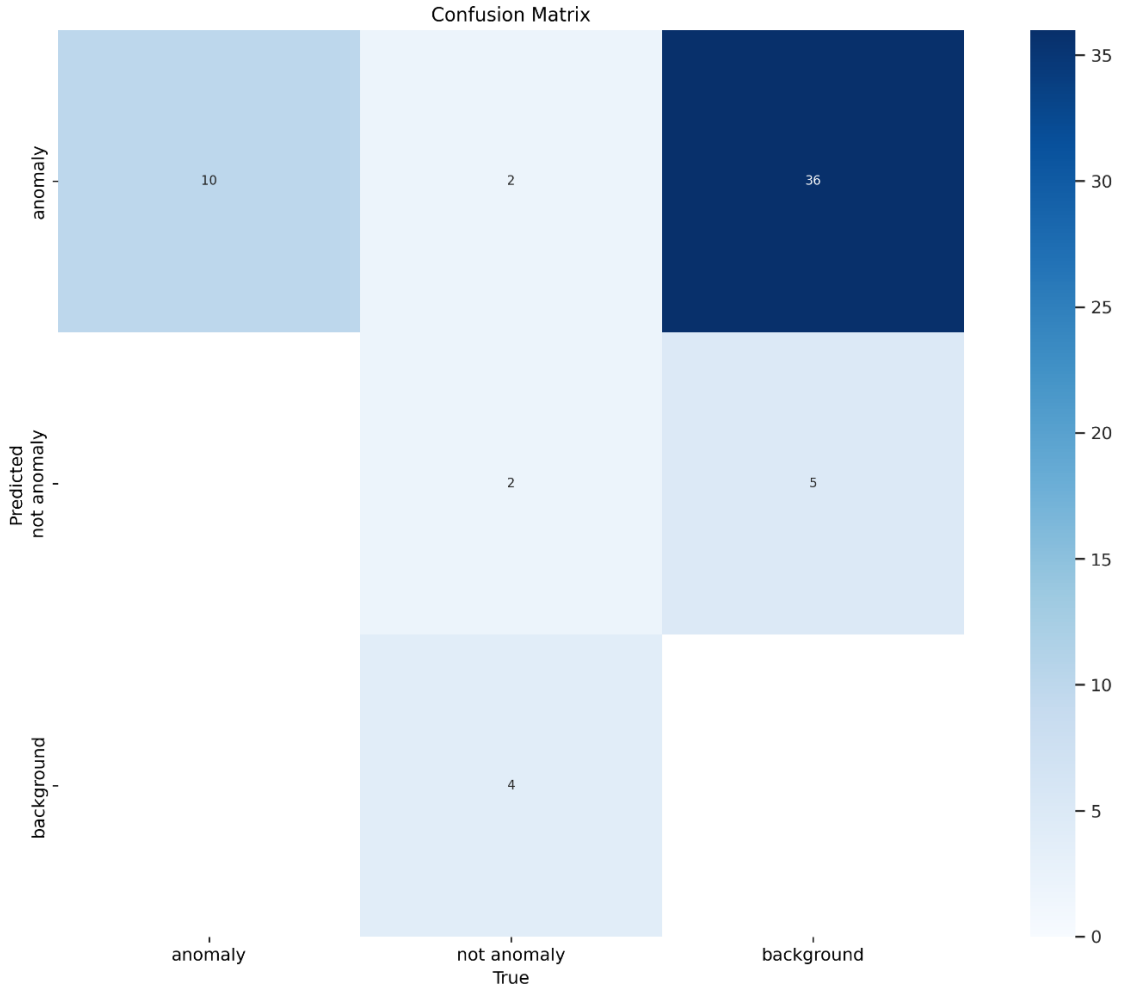
3. Confidence scores filter out unreliable predictions, ensuring high-quality detections.

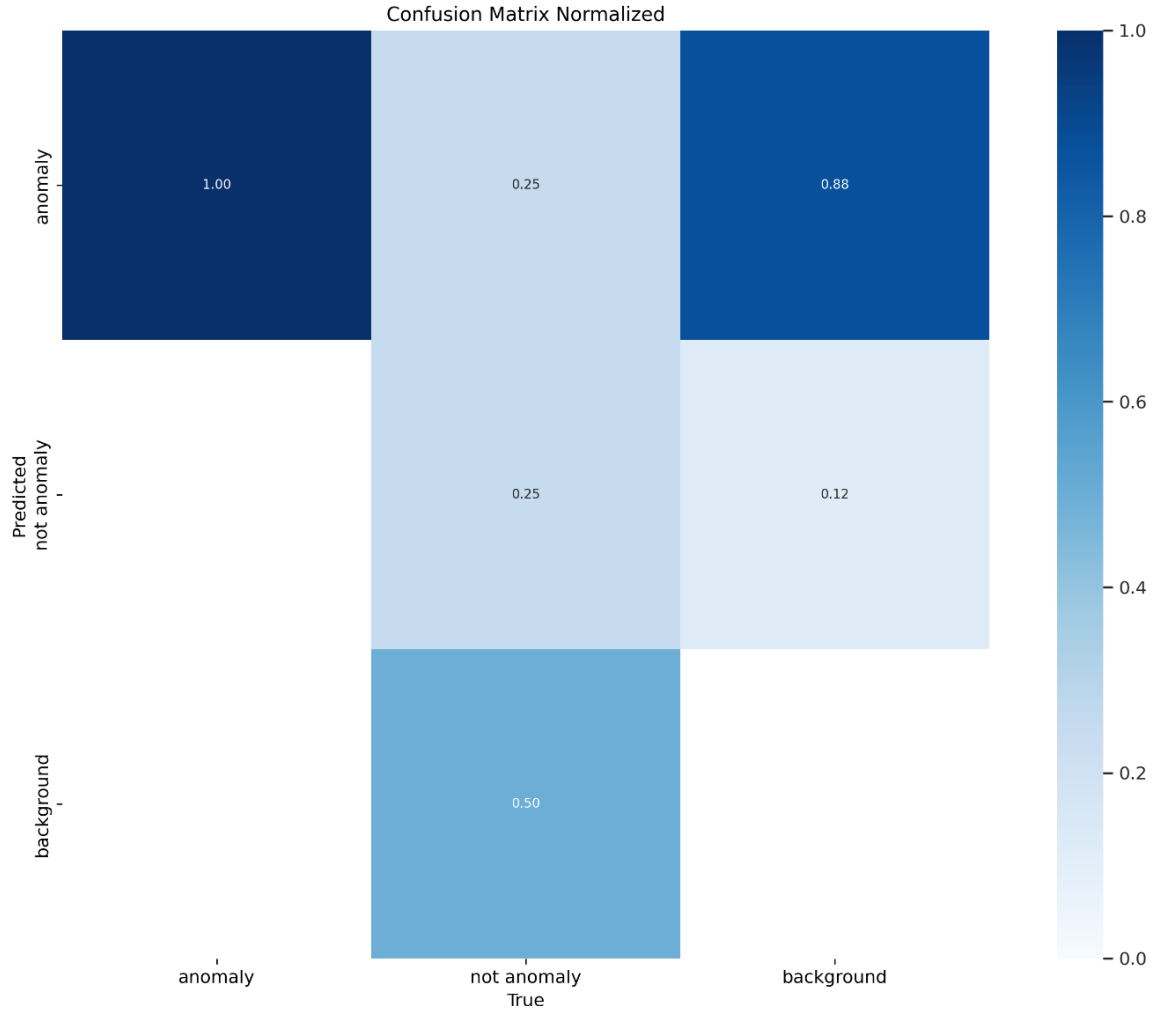
Together, these mechanisms make YOLOv5 highly effective in real-time workplace surveillance—even when trained on a limited dataset (e.g., 50 images).

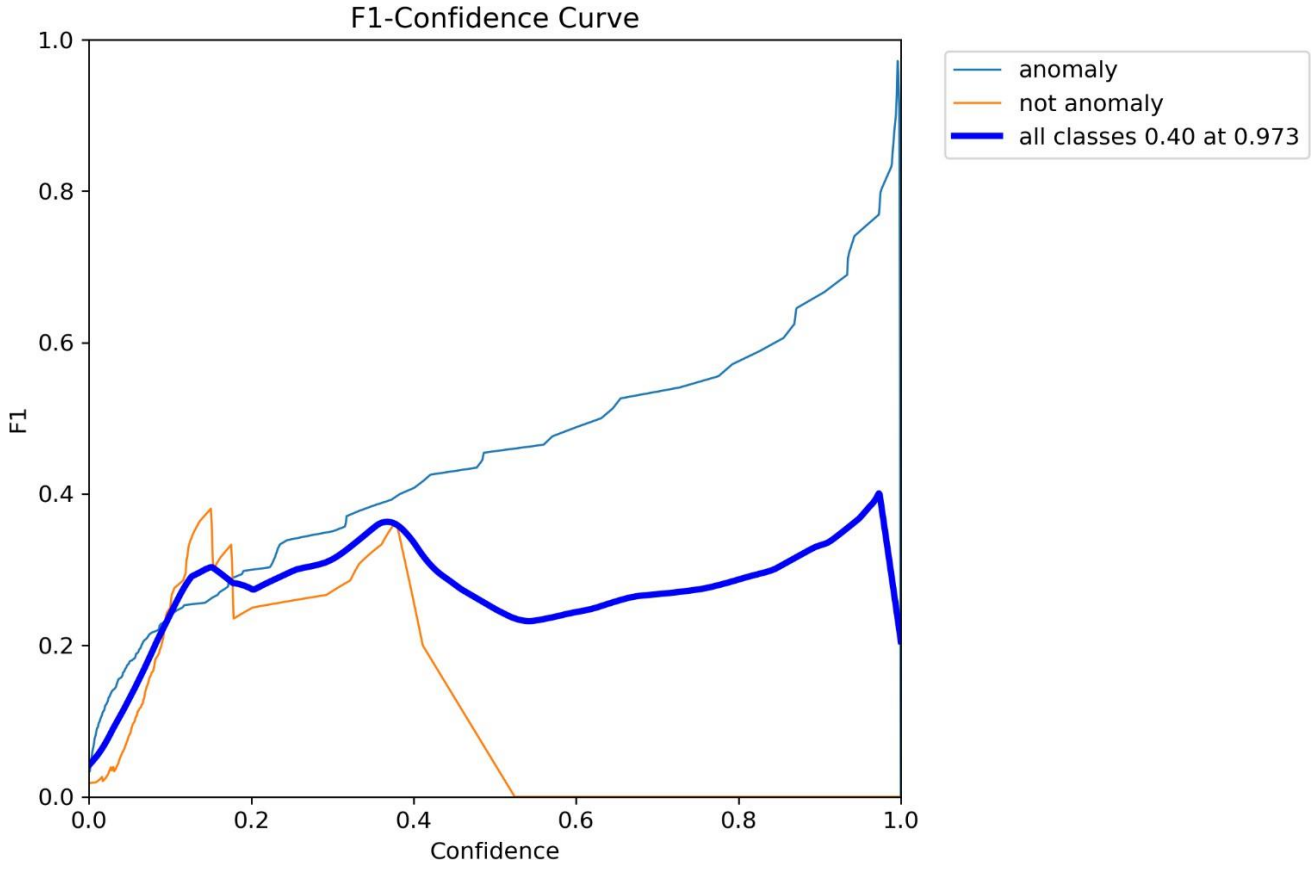
Future improvements may include:

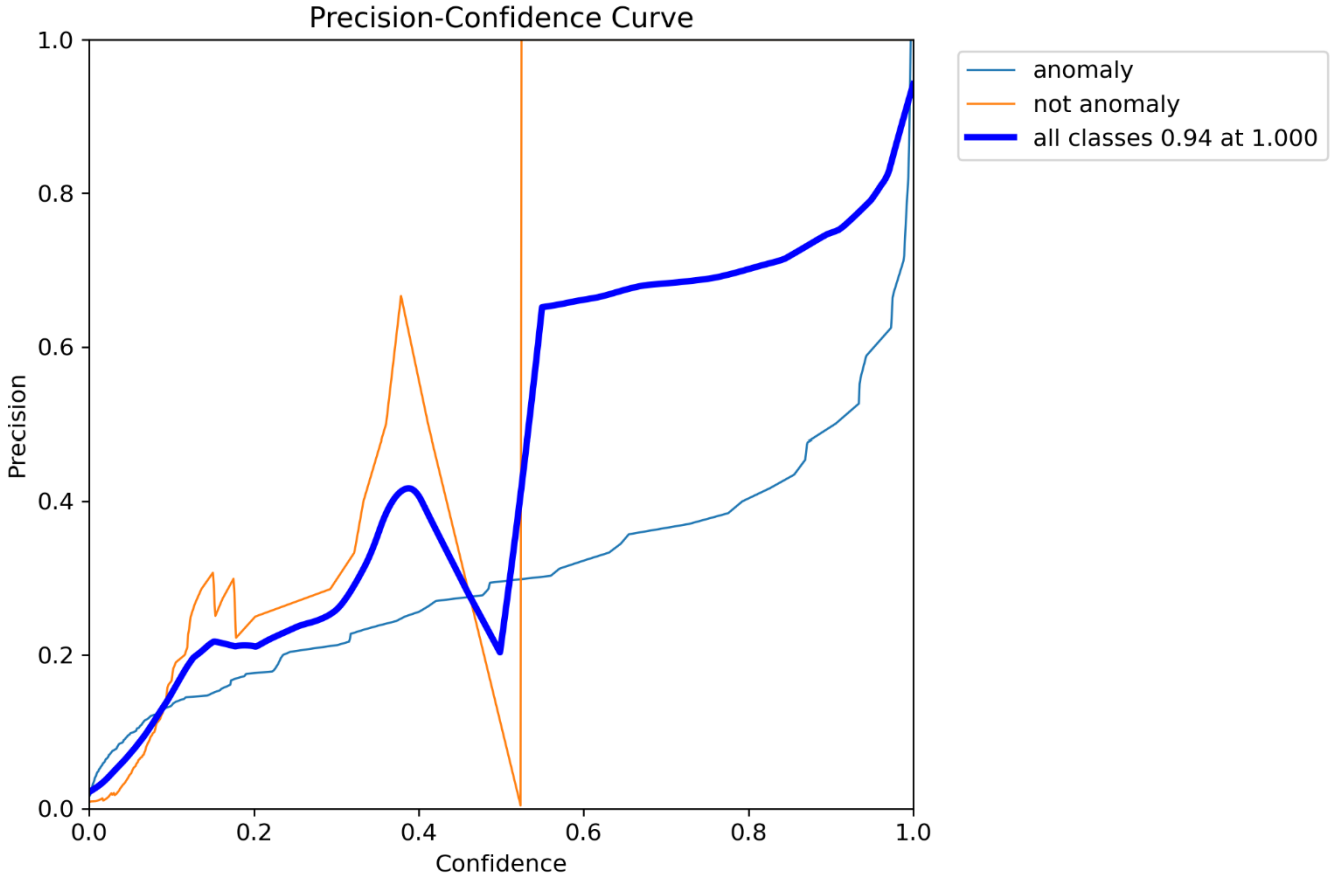
- Expanding the dataset to include a broader range of anomalies.

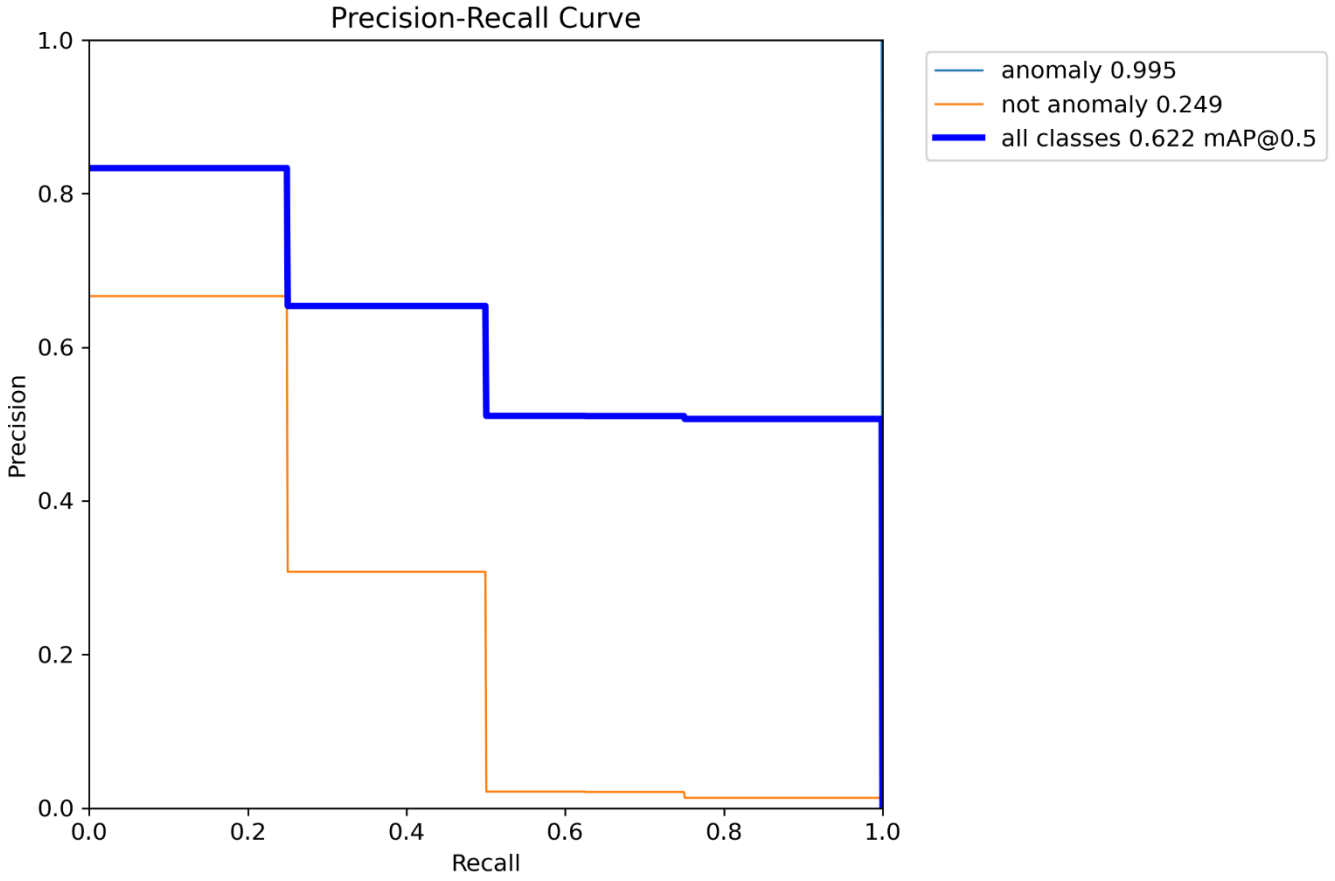
- Training with advanced models like YOLOv11n for enhanced accuracy and robustness.

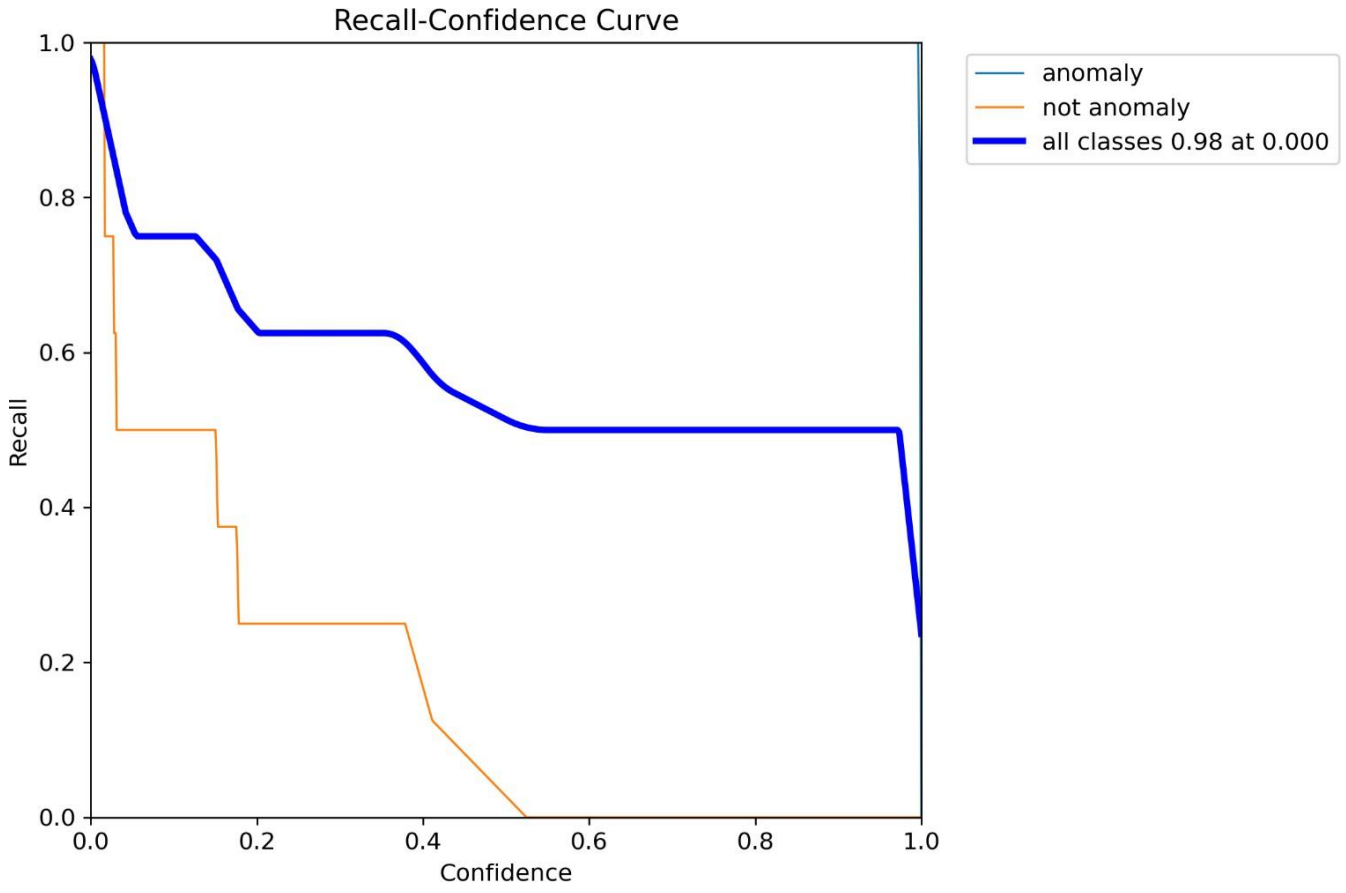



Confusion Matrix Normalized
anomaly
Predicted
not anomaly
background
1.00
0.25
0.88
0.25
0.12
0.50
anomaly
not anomaly
background
True
1.0
0.8
0.6
0.4
0.2
0.0


F1-Confidence Curve
F1
Confidence
0.0
0.2
0.4
0.6
0.8
1.0
anomaly
not anomaly
all classes 0.40 at 0.973


Precision-Confidence Curve
anomaly
not anomaly
all classes 0.94 at 1.000
Precision
Confidence
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall Curve
anomaly 0.995
not anomaly 0.249
all classes 0.622 mAP@0.5
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0



RESULTS AND DISCUSSION

The system was evaluated using a dataset of 500 video clips, each 30 seconds long, containing a mix of normal and anomalous activities. YOLOv5 achieved a precision of 81% and a recall of 50%, with an average inference time of 20 milliseconds per frame. Table I summarizes the performance metrics. Fig. 5 illustrates a sample detection, showing an employee accessing a restricted website, flagged by the system. Chal lenges included occasional false positives due to lighting changes and occlusions, which slightly reduced recall. These issues suggest areas for refinement, such as incorporating environmental context into the model

CONCLUSION & FUTURE SCOPE

This study demonstrates the efficacy of YOLO combined with Roboflow for real-time anomaly detection in workplace surveillance, providing a scalable solution for enhancing security and compliance. Future improvements could focus on training the model with a larger and more diverse dataset, incorporating a broader range of data tags to better capture variations in workplace activities. Additionally, leveraging more advanced models such as YOLOv11n could further enhance detection accuracy and robustness. Expanding the system to support multi-camera setups would also improve coverage and adaptability across different workplace environments..

BIBLIOGRAPHY

1. K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos,” in Advances in Neural Information Processing Systems, 2014, pp. 568–576. [Online]. Available: https://arxiv.org/pdf/1406.2199

2. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 779–788. [Online]. Available: https://arxiv.org/abs/1506.02640

3. H. Pirsiavash and D. Ramanan, “Detecting Activities of Daily Living in First-Person Camera Views,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 2847–2854. [Online]. Available: https://userpages. umbc.edu/~hpirsiav/papers/adl cvpr12.pdf

4. Ultralytics, “YOLOv5,” 2023. [Online]. Available: https://github.com/ ultralytics/yolov5

5. M. Yaseen, “What is YOLOv8: An In-Depth Exploration of the Inter nal Features of the Next-Generation Object Detector,” arXiv preprint arXiv:2408.15857, 2024. [Online]. Available: https://arxiv.org/abs/2408. 15857

6. C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, “YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information,” arXiv preprint arXiv:2402.13616, 2024. [Online]. Available: https://arxiv.org/ abs/2402.13616

7. R. Khanam and M. Hussain, “YOLOv11: An Overview of the Key Architectural Enhancements,” arXiv preprint arXiv:2410.17725, 2024. [Online]. Available: https://arxiv.org/abs/2410.1772