PROCEEDINGS OF
INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS AND NEW APPLICATIONS
https://proceedings.icisna.org/
3nd International Conference on Intelligent Systems and New Applications (ICISNA'25), Antalya, December 12-14, 2025.

Computer Vision-Based Behavior Analysis for Workplace Efficiency: Bakery Environment Application

Cengiz Samet Tepe[1], Ilkay Cinar[2]

[1] *Department of Computer Engineering, Selcuk University, 42250 Selcuklu, Konya, Türkiye*
248273001029@ogr.selcuk.edu.tr, ORCID: 0009-0004-6650-4794

[2] *Department of Computer Engineering, Selcuk University, 42250 Selcuklu, Konya, Türkiye*
ilkay.cinar@selcuk.edu.tr, ORCID: 0000-0003-0611-3316

***Abstract*— Nowadays, increasing efficiency in the service sector by objectively monitoring and tracking business workflows has become a critical requirement for the sustainability of businesses. Traditional monitoring methods are insufficient for providing sustainable performance analysis due to their time-consuming nature and reliance on subjective human judgment. The aim of this study is to develop a system that automatically detects and classifies employee behavior using computer vision and deep learning techniques. As part of the study, data was collected using a camera placed in a real bakery environment. Five basic classes were labeled on the created data set: cleaning, product interaction, computer use, phone use, and money interaction. The current YOLOv11 (You Only Look Once) architecture, which offers high speed and accuracy for object detection and classification, was used. According to the experimental results obtained from training the model, the system demonstrated high performance, achieving 0.9552 Precision, 0.9324 Recall, 0.9437 F1-Score, and 0.9644 mAP@50 values. These results demonstrate that the proposed system can detect employee behaviors in real-time with a high accuracy rate, allowing it to be used as an effective tool in workplace productivity enhancement and performance evaluation processes.**

***Keywords*— Behavior Analysis, YOLOv11, Computer Vision, Deep Learning, Object Detection, Service Sector.**

I. INTRODUCTION

In today's service sector, the sustainability of businesses and the strength of their competition with each other depend on the efficiency of their business processes and the performance of their employees. The behaviors exhibited by employees of businesses in this sector are decisive across a wide range, from service quality to customer satisfaction, and from occupational safety to operational efficiency. How employees allocate their time throughout the day between tasks, how long they spend on these tasks, and how often they perform them are important pieces of information for the business in terms of both process improvement and workforce planning. Especially in environments with high human interaction, such as bakeries, cafes, and restaurants, the analysis of employee behavior plays a critical role in both operational efficiency and service quality. These analyses currently rely on traditional methods such as direct monitoring by managers or manual reporting. However, manual monitoring methods cannot provide reliable and scalable solutions due to their time-consuming nature, the impossibility of continuous tracking, and the subjective judgments inherent in the human factor [1].

With the evolution of technology, Computer Vision and Deep Learning-based systems have begun to offer powerful alternatives for the automatic analysis of human behavior. Real-time object detection architectures such as Convolutional Neural Networks (CNN) and YOLO (You Only Look Once) are particularly capable of making highly accurate inferences from images. A review of the literature reveals that these technologies are widely used in the fields of occupational safety and industrial productivity.

Deep learning-based object detection algorithms are widely used, particularly in industrial settings, to ensure occupational health and safety [2], detect risky behavior [3, 4] and monitor the use of Personal Protective Equipment (PPE) [5, 6]. Furthermore, it is emphasized that they offer effective solutions in operational issues such as tracking the presence of workers in specific work areas to measure time spent, optimize workflows, and analyze worker productivity with objective data [1, 6, 7]. In these applications, different versions of the YOLO (You Only Look Once) architecture have been frequently preferred by researchers due to its real-time detection capability, high accuracy rate, and success in industrial failure detection.

While current studies prove the success of deep learning-based methods, there is a limited number of studies in the literature focusing on the service sector and, in particular,

dynamic work environments such as bakeries. The majority of existing datasets belong to industrial sites, construction sites, or laboratory environments; this makes it difficult to analyze behaviors specific to the service sector (e.g., counter cleaning, product arrangement, interaction with customers or money).

This study aims to automatically detect and classify employee behaviors using an unique dataset collected from a real bakery environment, with the goal of filling this gap in the literature. The study uses YOLOv11, one of the latest and high-performance architectures in the field of object detection. Trained on five different classes 'Cleaning', 'Product Interaction', 'Computer Use', 'Phone Use', and 'Money Interaction' the model can analyze employees' daily activities with high accuracy and in real time. The results show that the proposed system is a faster, more objective, and more reliable performance evaluation tool compared to manual monitoring methods.

II. Material And Methods

In this study, data was collected from a real work environment to identify employee behaviors specific to the service sector, and a detection mechanism was developed using YOLOv11, one of the most recent versions of YOLO. The proposed system's working method and the used methodologies are detailed under the following headings:

A. Data Collection and Preparation

The dataset used in this study was created from video footage recorded in the actual working environment of a bakery operating in the service sector. The data was collected using a fixed camera positioned to cover the bakery's main working area as much as possible, without interrupting the employees' movements, while also protecting customer privacy as much as possible. To reflect the natural behavior of employees, recordings were taken on different days, at different times of the day, and under varying lighting conditions to ensure data variability. In this way, the aim was to create a dataset based not only on controlled scenarios but also on natural workflows.

Samples were taken at specific moments in time from the collected videos, and frames were extracted at specific moments from each recording. After data cleaning, editing, and separating data believed to be of no use for training, a dataset consisting of approximately 2,500 clean data was prepared. Five classes representing the basic activities of employees in their workflow were defined. These classes were defined as 'Cleaning', 'Product Interaction', 'Computer Use', 'Phone Use', and 'Money Interaction'. This dataset was divided into two main sets: approximately 1,700 images for the training process and approximately 800 images for the validation process in order to evaluate the model's performance. The labeling process was performed manually by drawing bounding boxes around the relevant objects and actions in each image frame.

B. Data Labeling

Accurate and consistent labeling of data is critical for training deep learning models. In this study, 'labelImg', an open-source image labeling tool with a graphical user interface, was used for the labeling process of the collected images. Each image frame in the training dataset was manually reviewed, and actions belonging to the five defined behavior classes (Cleaning, Product Interaction, Computer Use, Phone Use, Money Interaction) were annotated using bounding boxes via the 'labelImg' interface. The data obtained as a result of the labeling was saved in a format directly compatible with the YOLO architecture. An example of labeling for the 'Phone Use' class is shown in Figure 1.

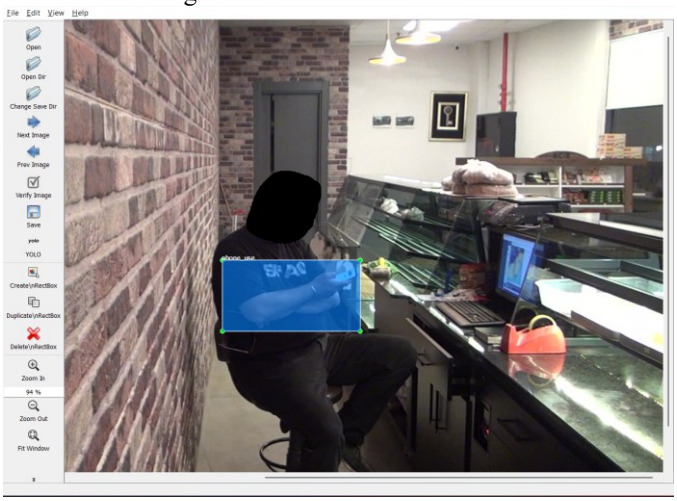

Fig. 1 Example labeling for the 'Phone Use' class

The YOLOv11 architecture, developed by Ultralytics and one of the latest versions of YOLO, has been chosen for the object detection and classification task. The YOLO (You Only Look Once) family is known for analyzing the image in a single pass, enabling both classification and localization (regression) operations at the same time, making it highly ideal for real-time applications [6, 8, 9].

YOLOv11 has a more advanced structure in terms of speed and feature extraction compared to its previous versions. The model uses C3k2 (Cross Stage Partial with kernel size 2) blocks and the C2PSA (Convolutional block with Parallel Spatial Attention) module, which contribute to faster processing and reduced computational load to increase computational efficiency. These developments allow the model to focus more effectively on areas within the image, potentially increasing detection accuracy for objects of different sizes and locations [8]. It enables an increase in detection accuracy, particularly for smaller or poorly visible objects. This allows the model to detect small objects and subtle behavioral details with high accuracy even in environments with complex backgrounds, such as bakeries. In this study, the YOLOv11s version, a lightweight and fast variation of the model, was used to optimize the balance between speed and accuracy.

C. Experimental Setup

The labeled dataset has been divided into approximately 70% training and 30% validation sets to evaluate the system's

performance. Additionally, to validate the system under real-world conditions, a completely external video recorded on a different day, not included in the training or validation datasets, was also used. The model employed in this study was trained on an NVIDIA GeForce RTX 4060 Laptop GPU; during the training process, the hyperparameters were set to 100 epochs and a batch size of 16.

D. Performance Metrics

The performance of the proposed system has been evaluated using metrics commonly used in the classification and object detection literature. In this context, the Confusion Matrix, one of the most fundamental structures, and statistical metrics based on it, namely Precision, Recall, F1-Score, mAP@50, and mAP@50-95, have been used to measure the system's success.

1) Confusion Matrix: It is a matrix that compares the model's predictions with actual labels [10, 11]. The confusion matrix is shown in Figure 2. This matrix consists of four basic components:

- TP - True Positive
- TN - True Negative
- FP - False Positive
- FN - False Negative

Actual	Predicted	
	Positive	***Negative***
Positive	TP	FN
Negative	FP	TN

Fig. 2 Confusion Matrix representation

2) Precision: This metric shows how much of the model's positive predictions are actually correct [6]. The formula is given below:

$$Precision = \frac{TP}{TP+FP}$$

3) *Recall:* It indicates how many of the truly positive examples are correctly identified by the model [6]. The formula is given below:

$$Recall = \frac{TP}{TP+FN}$$

4) F1-Score: It is defined as the harmonic mean of precision and recall values and provides a single value summarizing the balance between the two metrics [6]. The formula is given below:

$$F1\ Score = 2x(\frac{Precision \times Recall}{Precision+Recall})$$

5) Intersection Over Union (IoU): IoU measures the overlap ratio between the predicted bounding box and the actual bounding box, yielding a value between 0 and 1. Measurements greater than 0.5 can be interpreted as 'correct detection' [10]. The formula is given below:

$$IoU = 2x(\frac{Prediction\ Box \cap Actual\ Box}{Prediction\ Box \cup Actual\ Box})$$

6) Mean Average Precision (mAP): It is the most common metric used to evaluate the overall performance of a model in object detection problems. It is the arithmetic mean of the Average Precision values calculated for each class [12].

- mAP@50 bases its accuracy assessment on a metric called IoU (Intersection over Union). It represents the average accuracy when the IoU ratio between the predicted bounding box and the actual bounding box is at the 0.50 threshold value [9, 12]. The formula is given below:

$$mAP@50 = \frac{1}{N}\sum_{i=1}^{n} AP_i^{IoU=0.5}$$

- mAP@50-95 is the average precision value calculated across different IoU threshold values (from 0.5 to 0.95, with increments of 0.05). It helps measure the model's performance more accurately [10, 12]. The formula is given below:

$$mAP@50\text{-}95 = \frac{1}{10N}\sum_{j=0}^{9}\sum_{i=1}^{N} AP_i^{IoU=0.5+0.05j}$$

III. Experimental Results

A. Model Performance

The performance results of the trained YOLOv11s model are presented in Table I. According to the results obtained, the model achieved an average mAP@50 value of 0.964 for all classes, demonstrating high detection success.

TABLE I
General Performance Results Of The Model

F1-Score	Presicion	Recall	mAP@50	mAP@50-95
0.9437	0.9552	0.9324	0.9644	0.6783

When examining Table I, it can be seen that the model has a high precision value of 0.955, meaning that 95.5% of its positive predictions are correct. The recall value of 0.932 indicates that the model can detect the vast majority of actual actions without missing them.

B. Class-Based Analysis

To detail the success of detecting different behaviors in the bakery environment, the performance metrics obtained for each class are provided in Table II.

TABLE II
CLASS BASED PERFORMANCE RESULTS

Class	Presicion	Recall	mAP@50	mAP@50-95
money_interaction	0.997	1.000	0.995	0.794
computer_use	0.968	0.977	0.992	0.884
phone_use	0.971	0.968	0.984	0.676
cleaning	0.963	0.851	0.921	0.529
product_interaction	0.877	0.866	0.930	0.508

When examining class-based results, it is observed that an almost flawless detection rate was achieved, particularly in the "Money Interaction" class, with a Recall of 1.000 and a mAP@50 of 0.995. This can be explained by the fixed location of the cash register area, the proximity of the images to the camera capturing them, and the highly distinctive visual characteristics of money transactions. Similarly, "Computer Use" and "Phone Use" usage were also detected with very high accuracy due to the distinct forms of the objects.

In the "Product Interaction" and "Cleaning" classes, the mAP values were 0.93 and 0.92, respectively. The reason these classes have relatively lower scores compared to others can be explained by the fact that the angles at which these actions are performed are similar, they are relatively farther from the camera, and sometimes both actions occur in the same location.

C. Graphical Performance Analysis

When examining the confusion matrix in Figure 3, it can be seen that the model has a high accuracy rate. In particular, the model's detection success is quite high in the 'computer_use' (185) and 'phone_use' (181) classes.

The 'money_interaction' class, on the other hand, has demonstrated an almost flawless performance with 169 correct predictions.

When examining the relationship between the 'cleaning' and 'product_interaction' classes, it was observed that the system could accurately distinguish the cleaning action from other actions, but incorrectly assigned 4 data points belonging to the product_interaction class to the 'cleaning' class due to the visual and spatial similarity of hand movements.

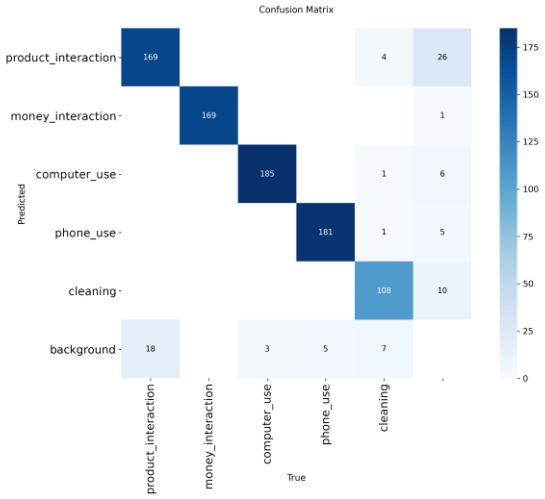

Fig. 3 Confusion Matrix

The most important point in the F1 Score-Confidence curve graph presented in Figure 4 is where the thick blue line, representing the weighted average for all classes, peaks. The model achieved maximum performance at a confidence threshold value of 0.467, obtaining an F1 score of 0.94.

When examining class-based performances, it is observed that the 'money_interaction' (orange line) class has the highest stability, hovering close to 1.0. The 'computer' and 'phone_use' classes also exhibit similarly high stability. Although the 'cleaning' and 'product_interaction' classes show an earlier downward trend compared to other classes after the 0.80 confidence threshold, they still demonstrate high performance. The fact that the curves remain horizontal for a long time across the graph and only experience a decline at very high confidence thresholds (e.g., after 0.85) indicates that the model stands behind its own predictions.

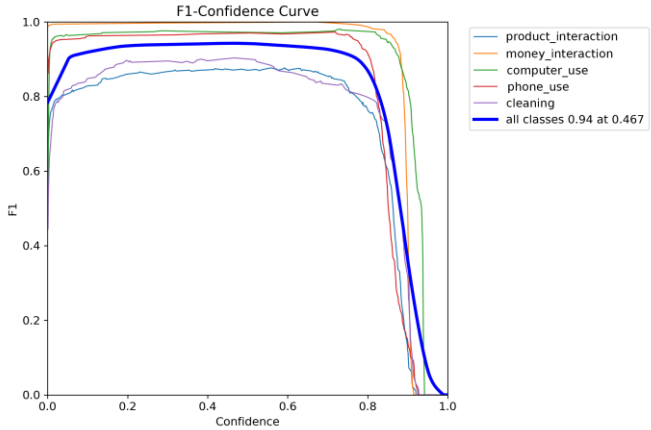

Fig. 4 F1 Score-Confidence Curve

In the Precision-Confidence curve graph shown in Figure 5, it can be observed that as the confidence threshold increases, the precision value steadily approaches 100% for all classes. This increase indicates that the rate of false positives produced in the model's high-confidence predictions is nearly zero. In particular, the fact that the "all classes" curve reaches 1.00 (perfect) certainty at a confidence threshold of 0.940 shows that the system can produce nearly error-free results when operating above this threshold value.

In the class-based analysis, the 'money_interaction' (orange) and 'computer_use' (green) classes achieve over 0.95 certainty even at very low confidence thresholds, demonstrating the model's success in identifying these objects. The relatively more complex 'product_interaction' and 'cleaning' classes, while showing some fluctuation at low confidence values, achieve a similar level of success to the other classes after the 0.60 confidence threshold.

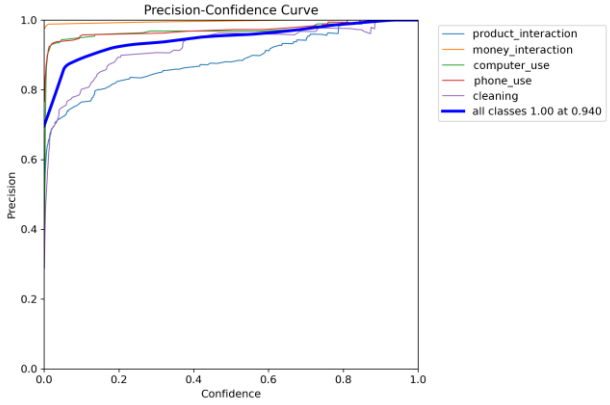

Fig. 5 Precision-Confidence Curve

Figure 6 presents the Recall-Confidence curve, which analyzes how the model's ability to detect classes (Recall) changes in response to an increasing confidence threshold. At the starting point of the graph (confidence threshold 0.000), the overall sensitivity for all classes being at the 0.97 level indicates that the model can successfully capture 97% of the classes (the False Negative rate is very low) when no filtering is applied.

The general performance curve indicated by the thick blue line follows a horizontal trajectory from 0.00 to a confidence interval of approximately 0.70, proving that the miss rate remains at a minimum level for a long time even when we increase the model's prediction confidence. In class-based differentiation, the 'money_interaction' (orange) and 'computer_use' (green) classes have the most resilient structure; they do not experience sensitivity loss until the confidence threshold approaches 0.90. In contrast, in the 'cleaning' and 'product_interaction' classes, where visual complexity is higher, the curve shows a more pronounced downward trend after the 0.60 confidence threshold.

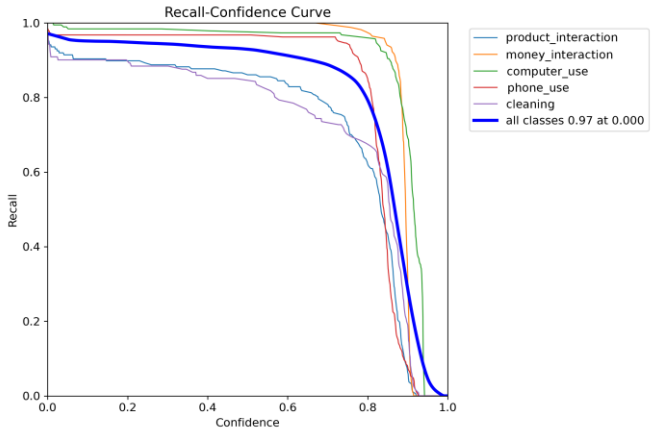

Fig. 6 Recall-Confidence Curve

D. *Visual Detection Results*

In addition to verifying the system's numerical success with evaluation metrics, testing was performed on an external video not included in the training and validation datasets to ensure it could also be validated under real-world conditions. The system's sample detection outputs are presented in Figure 7:

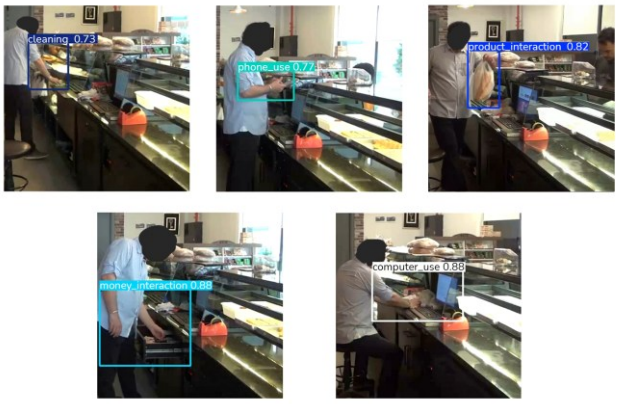

Fig. 7 Samples of detection images taken at random times from the test video

The system was able to detect target classes in this video, recorded on a different day than the day the images in the training set were obtained, as shown in Figure 7. These observational results demonstrate that the model has a successful prediction capability on new data that was not included in the training.

IV. CONCLUSION

In this study, an objective system based on deep learning has been developed for the automatic detection and analysis of employee behaviors in the service sector. To this end, frames were extracted from videos recorded during actual working hours using a single fixed camera; following preprocessing steps, five basic behavior classes (product interaction, phone

interaction, cleaning, computer interaction, money transactions) were defined and the frames were manually labeled. The proposed method was tested on a unique dataset collected from a real bakery environment by training the current YOLOv11 architecture.

Experimental results show that the developed system can accurately detect employees' daily activities (cleaning, product interaction, computer usage, etc.) with an average of 0.964 mAP@50 and 0.955 Precision value. In particular, the 99.5% success rate achieved in the “Money Interaction” class proves the system's reliability in tracking critical business processes. Although the system's overall performance is high, classification errors were occasionally observed in the “Cleaning” and “Product Interaction” classes, as can be seen in the tests on the video stream and in the confusion matrix. The main reason for this is that both classes sometimes occur in similar locations (on the countertop) and the movements of employees (reaching, wiping, etc.) are highly similar. It has been assessed that these momentary confusions, which occur especially when cleaning actions are performed in areas very close to the products, do not have a statistically significant negative impact on overall performance.

The results show that the proposed system can offer a much faster, more sustainable, and objective solution compared to traditional manual monitoring methods. This system provides business managers with a powerful decision support mechanism for increasing employee productivity, optimizing business workflows, and conducting fair performance evaluations.

In future studies, the system will be developed by attempting to prevent misclassifications occurring in real time through improvements made to the data set and calculation algorithms. Additionally, instead of using only the YOLOv11 version, training and testing will be conducted using other versions of YOLO and different variations of these versions, with plans to redesign the system using the model that demonstrates the highest performance. Furthermore, work is planned on automatically reporting not only the detection of behaviors but also their time-based analysis.

ACKNOWLEDGMENT

This study is derived from the unpublished master’s thesis of Cengiz Samet TEPE, supervised by Dr. Ilkay CINAR.

REFERENCES

[1] J. Li, X. Zhao, G. Zhou, M. Zhang, D. Li, and Y. Zhou, "Evaluating the work productivity of assembling reinforcement through the objects detected by deep learning," *Sensors,* vol. 21, no. 16, p. 5598, 2021, doi: 10.3390/s21165598.

[2] E. Guney, H. Altin, A. E. Asci, O. U. Bayilmis, and C. Bayilmis, "YOLO-based personal protective equipment monitoring system for workplace safety," *JITSI: Jurnal Ilmiah Teknologi Sistem Informasi,* vol. 5, no. 2, pp. 77-85, 2024, doi: 10.62527/jitsi.5.2.238.

[3] R. A. Maulana, "Implementation of YOLO (You Only Look Once) Algorithm for Drowsiness Detection as An Additional Safety Feature in the Operation of Crane Equipment in Real Time," *Jurnal Inotera,* vol. 10, no. 1, pp. 113-120, 2025, doi: 10.31572/inotera.Vol10.Iss1.2025.ID458.

[4] O. Önal and E. Dandıl, "Video dataset for the detection of safe and unsafe behaviours in workplaces," *Data in Brief,* vol. 56, p. 110791, 2024, doi: 10.1016/j.dib.2024.110791.

[5] K. Patel, V. Patel, V. Prajapati, D. Chauhan, A. Haji, and S. Degadwala, "Safety helmet detection using YOLO v8," in *2023 3rd international conference on pervasive computing and social networking (ICPCSN)*, 2023: IEEE, pp. 22-26, doi: 10.1109/ICPCSN58827.2023.00012.

[6] A. S. Ludwika and A. P. Rifai, "Deep learning for detection of proper utilization and adequacy of personal protective equipment in manufacturing teaching laboratories," *Safety,* vol. 10, no. 1, p. 26, 2024, doi: 10.3390/safety10010026.

[7] A. K. Das, V. Kamthane, U. Purwar, D. C. Mohanty, and B. K. Depuru, "Enhancing Workplace Efficiency and Security Through Intelligent Employee Surveillance," *International Journal of Innovative Science and Research Technology,* vol. 9, no. 3, 2024, doi: 10.38124/ijisrt/IJISRT24MAR2142.

[8] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725,* 2024, doi: 10.48550/arXiv.2410.17725.

[9] M. Hussain, "YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection," *Machines,* vol. 11, no. 7, p. 677, 2023, doi: 10.3390/machines11070677.

[10] A. S. Ozer and I. Cinar, "Real-Time and Fully Automated Robotic Stacking System with Deep Learning-Based Visual Perception," *Sensors,* vol. 25, no. 22, p. 6960, 2025, doi: 10.3390/s25226960.

[11] Z. Dolmaz and I. Cinar, "Detection of DDOS Attacks in Software-Based Systems in Cyberspace Using Machine Learning," *Journal of Technology and System Information,* vol. 2, no. 4, pp. 1-22, 2025, doi: 10.47134/jtsi.v2i4.5033.

[12] M. A. Ozuberk and I. Cinar, "YOLOv8-Based Threat Detection Model for Dangerous Objects and Violent Behaviors," in *2025 10th International Conference on Computer Science and Engineering (UBMK)*, 2025: IEEE, pp. 208-213.