Pose-Based Suspicious Activity Detection Using YOLOv12n and MediaPipe in Smart Surveillance Systems

Bowo Nugroho[1,*], Bima Prihasto[1], Aninditya Anggari Nuryono[1], Ahmad Rusdianto Andarina Syakbani[1], Nawfal Alhadi[1], Rionando Soeksin Putra[1], Ilham Al Basith[1]
[1]Informatika, Institut Teknologi Kalimantan;
bima @lecturer.itk.ac.id
aninditya.nuryono @lecturer.itk.ac.id
11211005@student.itk.ac.id
11201050@student.itk.ac.id
11221063@student.itk.ac.id
11221077@student.itk.ac.id
[*]Correspondence: bowo.nugroho@lecturer.itk.ac.id;

Abstract – Detecting suspicious activity in video surveillance is critical for public safety. Early identification of pre-incident behaviors, such as "looking around" nervously, can prevent crimes before they occur. However, traditional surveillance relies on continuous human monitoring, which is prone to error and resource-intensive. Existing automated methods struggle with real-time stability and produce flickering predictions. We present a two-stage pose-based detection system: YOLOv12n detects humans in each frame, and MediaPipe BlazePose extracts pose features. A Support Vector Classifier (SVC) and Neural Network (NN) then classify 15-frame temporal segments as either "still" or "looking around." On our custom dataset, SVC achieves 97.35% accuracy while NN achieves 97.00%. Temporal smoothing eliminates prediction flickering, enabling stable real-time deployment. SVC provides better precision-recall balance; NN offers greater adaptability to complex pose variations.

Keywords – Suspicious activity detection, Human pose estimation, YOLOv12n, MediaPipe BlazePose, Support Vector Classifier

I. INTRODUCTION

The exponential growth of Closed Circuit Television (CCTV) infrastructure has transformed security paradigms in public and private spaces. However, conventional surveillance systems that rely on human operators face fundamental challenges related to cognitive limitations. One well-documented phenomenon is inattentional blindness, the inability of operators to notice unexpected stimuli due to prolonged engagement in intensive visual monitoring tasks. Studies [1] reveal that operator attention declines after short monitoring sessions, causing failures in detecting critical activities. This challenge is compounded by operator fatigue, which degrades situational awareness and increases overall cognitive load [2], [3], [4].

In security contexts, early detection of suspicious activity holds strategic value, particularly for identifying pre-crime behaviors. One important indicator is "looking around" behavior, active environmental scanning by potential offenders. According to ecological criminology theories such as Routine Activity Theory and Rational Choice Theory, this scanning or environmental hypervigilance represents both an anxiety response and a risk-evaluation strategy that offenders use to assess situational readiness before committing a crime [5], [6], [7]. Real-time identification of such movement patterns is an urgent requirement for developing intelligent surveillance systems.

Unfortunately, automatic detection of subtle, rapid suspicious behaviors like "looking around" remains technically challenging. Current solutions typically employ appearance-based approaches, such as activity recognition from raw RGB image features. These methods struggle in real-world environments with varying lighting conditions, backgrounds, and visual attributes of subjects [8], [9]. As an alternative, pose-based methods, which represent the human body skeleton through keypoints, offer greater invariance to external conditions and individual differences [10], [11].

II. LITERATURE REVIEW

Recent literature demonstrates that pose-based systems, such as skeleton-based action recognition, achieve excellent performance in detecting human behavior within complex environments. Architectures like Graph Convolutional Networks (GCN) efficiently capture spatial relationships between body joints [11], while multi-scale attention mechanisms improve system accuracy for detecting specific movements in real-time [9]. For embedded surveillance systems operating on resource-constrained devices, lightweight approaches such as MediaPipe, MobileNet, and recent YOLO variants like YOLOv12n are preferred for their balance between accuracy and computational efficiency [12], [13].

MediaPipe BlazePose enables real-time 3D pose estimation with 33 keypoints, providing rich spatial and body

orientation representations that are particularly relevant for distinguishing gestures such as "still" versus "looking around." This framework operates efficiently through a two-stage inference strategy: ROI detection followed by keypoint tracking with optimal performance on mobile devices [12]. For object detection, YOLOv12n, as the first attention-based model in the YOLO family, successfully integrates CNN efficiency with the global representation power of transformers, enabling high-resolution human detection with fast inference speeds [13].

For classifying suspicious activity, commonly used models include Support Vector Classifier (SVC) and Neural Networks (NN). SVC, based on Structural Risk Minimization principles, efficiently separates classes with maximum margin in high-dimensional feature spaces and avoids overfitting on small-to-medium datasets [14], [15]. Neural Networks, particularly Multilayer Perceptron (MLP) architectures, offer high flexibility in modeling non-linear relationships and adapt to greater data complexity when paired with proper architectural design and regularization [16].

The literature also shows that temporal smoothing mechanisms improve prediction stability in real-time detection contexts by reducing flickering effects that lead to short-term classification errors. These approaches typically use probability thresholds and temporal consistency requirements as filters to validate predictions before system acceptance [17], [18].

Despite these advances, a gap remains in developing intelligent surveillance systems that are computationally efficient yet accurate in detecting "looking around" behavior in real-time. Previous studies tend to focus on explicit, long-duration human activities, while suspicious behaviors like "looking around" are characterized by short duration, small amplitude, and difficulty distinguishing from normal gestures. Therefore, an approach that combines efficiency, spatiotemporal precision, and classification capability is needed to produce a reliable detection system.

This research develops a two-stage framework integrating object detection using YOLOv12n and pose estimation using MediaPipe, processed through SVC and NN models for classifying "still" and "looking around" activities. Our main contributions are: (1) Design of 345-dimensional spatiotemporal features based on 15-frame segments; (2) Integration of additional semantic features, including face orientation and hand visibility; (3) Performance comparison of SVC and NN on a structured dataset; and (4) Application of temporal smoothing to improve detection stability in real-time scenarios. This study provides practical and theoretical contributions to intelligent surveillance, particularly in detecting pre-crime behaviors that are difficult to define explicitly in visual monitoring environments.

III. METHODOLOGY

This research presents a two-stage framework for detecting suspicious "looking around" behavior from human pose in surveillance video. First, YOLOv12n detects humans in each frame. Second, MediaPipe BlazePose extracts body pose keypoints. The resulting spatiotemporal features feed into Support Vector Classifier (SVC) and Neural Network (NN) models, which classify each segment as either "still" or "looking around" in real-time. We evaluate video data recorded in semi-naturalistic conditions at a limited scale. Figure 1 shows the overall architecture of the proposed real-time suspicious behavior detection system.

A. Object Detection Using YOLOv12n

The first stage detects humans in each video frame using YOLOv12n. This lightweight model combines CNN efficiency with transformer-based global attention, enabling fast and accurate inference on resource-constrained edge devices [13]. The detector outputs bounding boxes, which serve as input to the pose estimation stage.

B. Pose Estimation Using MediaPipe BlazePose

Given detected bounding boxes, the second stage estimates 33 body keypoints in 3D using MediaPipe BlazePose. BlazePose operates in two sub-stages: ROI estimation from the detected body region, followed by landmark regression via a lightweight CNN. This enables real-time pose extraction with sufficient precision to capture subtle posture cues, head orientation, hand position, and fine-grained movement patterns [12].

C. Spatio-Temporal Feature Extraction and Reduction

The output from BlazePose consists of 33 keypoints in spatial coordinates. The five landmarks used are: nose, left eye, right eye, left wrist, and right wrist. Each sample in the dataset is represented by a time window containing 15 consecutive video frames. Each landmark has 4 attributes: X, Y, Z coordinates, and visibility. This produces a total of 20 features per time frame, which are then converted into a 300-dimensional vector by taking the mean value per landmark for each attribute within the time window. Additional semantic features, such as head orientation and hand visibility, are added to improve classification performance.

D. Classification Models: SVC and Neural Network

The reduced spatiotemporal features are used as input for two classification models: Support Vector Classifier (SVC) and Neural Network (NN). The SVC model uses a Radial Basis Function (RBF) kernel and is known to be superior for small to medium datasets and capable of separating two classes with maximum margin in high-dimensional feature space [15], [16]. As a comparison, a Multilayer Perceptron (MLP) architecture with two hidden layers containing 256 and 128 neurons is used. Dropout and Adam optimizer are also applied to prevent overfitting [19], [20].

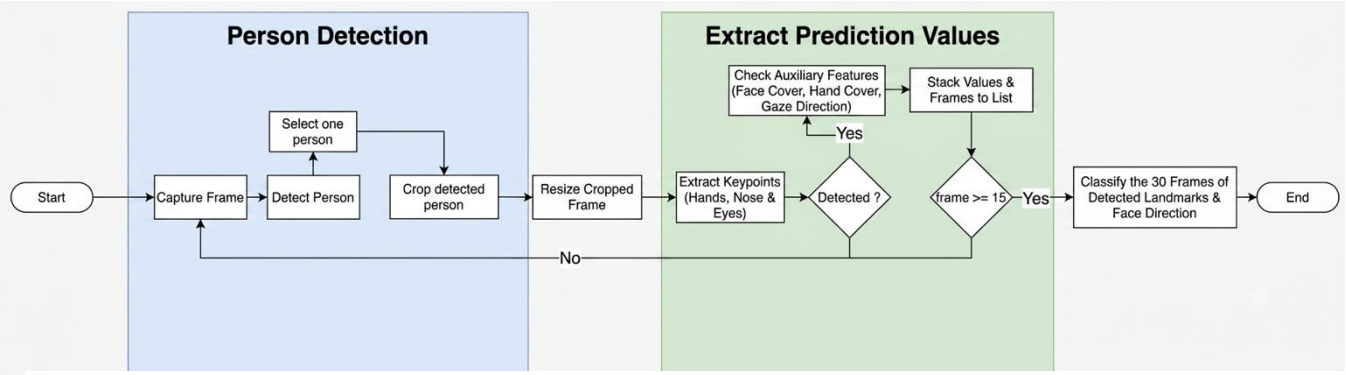

Figure 1.Architecture of a pose-based suspicious behavior detection system.

E. Temporal Smoothing for Real-Time Detection

To stabilize predictions in real-time, we apply temporal smoothing, which filters noisy frame-to-frame pose estimates. The system updates the output label only when two conditions hold: (1) prediction confidence exceeds 75%, and (2) the predicted label remains consistent for 10 consecutive frames [17], [21], [22].

F. Data Collection and Performance Evaluation

The dataset consists of 26 videos, all recorded with a single person in an indoor scenario simulating a CCTV viewing angle. Each video has a duration between 10 and 25 seconds, with room lighting kept constant and a uniform color tone, ensuring no disturbances or changes in illumination throughout the recording to minimize environmental variability. The camera is mounted at a high position with a downward tilt of approximately 30°–45°, mimicking typical CCTV installation, with viewpoint variations from the subject's right side, left side, and from behind when the subject is facing away from the camera. The videos are recorded at 30 fps and monitored continuously, while data capture is performed manually using an input button; each button press stores the most recent 15 frames, and for each frame the selected landmarks are extracted so that the 15 frames together form a single labeled record. Labeling is carried out concurrently with feature extraction, yielding a total of 400 records with a balanced 50:50 distribution between the "still" and "looking around" classes. The dataset is then randomly split at the segment level into training and testing sets with an 80:20 ratio to ensure fair evaluation and prevent data leakage between segments originating from the same video. Evaluation was performed using accuracy, precision, recall, and F1-score metrics.

G. System Architecture

The complete system architecture encompasses the following stages: object detection (YOLOv12n), pose estimation (BlazePose), feature extraction and reduction, classification (SVC/NN), and stabilization through temporal smoothing. In the training process, the SVC is configured with C set to 1.0, using a Radial Basis Function (RBF) kernel and gamma set to scale. This system is designed for efficiency, accuracy, and stability in detecting suspicious behavior in real-time.

Figure 1 illustrates the system workflow starting from frame acquisition, human object detection, to image cropping and resizing. The process continues with body keypoint extraction and additional features, where data from a number of consecutive frames is collected and then classified to produce the final behavior prediction.

IV. RESULTS AND DISCUSSION

The results of this research are presented based on the main stages in the developed system, covering the object detection and feature extraction process, performance evaluation of Support Vector Classifier (SVC) and Neural Network (NN) models, and system stability analysis in real-time scenarios. The results are explained systematically following the methodology subsection structure described previously, to ensure continuity and coherence between the experimental design and the obtained outputs.

A. Detection and Feature Extraction Results

The initial stage of the developed system is human object detection in video using YOLOv12n. This model was chosen due to its ability to perform detection efficiently on devices with computational limitations. Experimental results show that YOLOv12n is capable of detecting humans with high precision, including in low-light conditions or complex backgrounds. All detected human objects are then cropped and normalized to 640×640 pixel resolution using the INTER_AREA interpolation method. To keep the human body aspect ratio consistent, the letterbox padding technique is applied, which wraps the object in a white frame, ensuring body proportions remain accurate.

The feature extraction process is performed using MediaPipe BlazePose with a focus on five main landmarks: nose, right and left ear, and right and left wrists. Each landmark is extracted in the form of x, y, z coordinates and visibility value (v). Additionally, supplementary features such as face direction (facing right, left, or center), face status (looking down or upright), and hand visibility (visible or not)

Figure 3. Feature extraction

are calculated using heuristic rules. All these features are then arranged into a 345-dimensional vector, representing information from 15 consecutive video frames. This temporal representation proves effective in capturing subtle motion patterns between the two target poses: "still" and "looking around."

Figure 2 visualizes the data transformation process from visual input to structured numerical data. The flow begins with video frame sequence acquisition, where detected human objects are cropped and processed for body landmark detection. The coordinates (x, y, z) from these landmarks are then extracted and arranged from 15 consecutive frames, producing a temporal data matrix representing motion features as input for the classification model.

B. Support Vector Classifier (SVC) Performance Evaluation

The SVC model was trained using a Radial Basis Function (RBF) kernel and configured to produce prediction probabilities. Before training, all feature data were normalized using StandardScaler to balance the distribution of each feature. Evaluation of this model shows highly satisfactory results. Based on the confusion matrix, the SVC model achieves an accuracy of 97.35% on the test data, with a precision of 97.53%, a recall of 97.31%, and an F1-score of 97.35%. These results indicate excellent and balanced classification performance between precision and recall.

Further analysis of SVC performance shows that this model is capable of distinguishing two pose classes with a clear decision margin. This capability stems from the margin maximization principle that forms the foundation of SVM theory [15], [23], which enables better generalization, especially in high-dimensional domains such as those produced by the combination of spatiotemporal features. The model also shows resistance to overfitting, as evidenced by similar performance between training and testing data.

C. Neural Network Performance Evaluation

The Neural Network (NN) model used in this research consists of two hidden layers with sizes of 256 and 128 neurons, respectively. ReLU activation is applied at each hidden layer, and dropout with a ratio of 0.3 is used to prevent overfitting. The output layer uses Softmax activation to produce classification probabilities.

Training and testing results show that NN provides competitive performance. Accuracy on test data reaches 97.00%, with precision of 97.56%, recall of 96.67%, and F1-score of 97.01%. Although slightly lower than SVC, the NN model shows a stable and convergent learning curve, with consistent loss decrease trends on training and validation data. This indicates that model parameters have been optimally configured and are capable of learning from data without overfitting tendencies.

The NN model also shows greater adaptability for more complex data variations. This aligns with previous studies showing that Multi-Layer Perceptron (MLP) architectures have high generalization capability in pose-based classification [16], [24], especially when combined with temporal feature representation.

Table 1. Comparison of Classification Methods

Model	Evaluation Matrix			
	Accuracy	Precision	Recall	F1- Score
SVC	97.35	97.53	97.31	97.35
Neural Network	97.00	97.56	96.67	97.01

Figure 4. Test video and detection results.

Table 1 shows that both models provide very high performance, but SVC is slightly superior in accuracy and balance between precision and recall. Nevertheless, NN shows the highest precision value, indicating reliability in positive classification.

Table 2. Computational Performance Evaluation

Processing Stage	Support Vector Classifier	Neural Network
YOLOv12n Latency	34.10 ms	42.82 ms
MediaPipe Pose Latency	19.16 ms	38.00 ms
Classifier Latency	0.88 ms	87.28 ms
Total Inference Time	**61.74 ms**	**177.16 ms**
System Througut	**16.20 FPS**	**5.64 FPS**

Table 2 presents a comparative analysis of the computational performance between the Support Vector Classifier (SVC) and Neural Network (NN) models. The evaluation reveals that the SVC pipeline is significantly more efficient, achieving a total inference time of 61.74 ms and a processing throughput of 16.20 FPS. The classification stage in the SVC model introduces a negligible latency of only 0.88 ms. In contrast, the NN model exhibits a substantial computational bottleneck, requiring 87.28 ms for classification alone. This results in a much higher total inference time of 177.16 ms (5.64 FPS). These findings indicate that the SVC model is vastly superior in terms of computational speed and is better suited for real-time deployment.

D. Real-Time Detection Stability Analysis

To test system stability in real-time scenarios, a temporal smoothing mechanism is used. This mechanism requires predictions to have a minimum probability of 75% and remain consistent for 10 consecutive frames before the classification decision can be updated. Without this mechanism, inference results show significant flickering, especially when the subject changes position rapidly.

After implementing the smoothing mechanism, detection results become much more stable. The system is capable of maintaining consistent classification decisions, even when the subject makes small movements or is in non-ideal lighting conditions. Tests on real-time video show that the system is capable of working with an adequate average frame rate for direct applications, without significant performance degradation.

This stability is important for real-world implementation, as it prevents momentary detection errors that could cause false alarms or misinterpretation by security systems. The combination of rich feature representation, reliable classification models, and appropriate post-processing strategies proves capable of producing a robust suspicious activity detection system suitable for deployment in automated surveillance environments.

Overall, these results demonstrate the effectiveness of the proposed two-stage framework in detecting and classifying suspicious activity based on human pose. With both models achieving accuracy above 97%, this system can serve as a foundation for the development of future smart security systems.

Figure 3 displays system testing results on real-time video as well as a visualization of the data processing flow. The main display shows activity detection with class labels and confidence levels generated at that moment. The right side of the figure specifically visualizes 15 consecutive frames (frame sequence) captured by the system as a unified temporal input. This frame sequence is processed together to extract landmark features, allowing the model to analyze movement patterns within that time span to determine activity classification.

E. Interpretation of Detection and Feature Extraction Results

The use of YOLOv12n as an object detection model proves effective in detecting human presence in video frames. YOLOv12n is a nano version of the YOLO series specifically designed for edge computing applications with computational limitations, without sacrificing detection precision [13]. The effectiveness of YOLOv12n in this research is reflected in its

ability to maintain the shape and proportions of human objects, even in varying lighting conditions.

Feature extraction using MediaPipe BlazePose successfully produces stable and representative body landmarks. The selection of five body keypoints—nose, left and right ears, and left and right wrists—represents head and hand movements that serve as the main visual indicators of the "looking around" pose. The addition of heuristic features such as face direction and hand visibility enriches the semantic pose information, in line with findings in [12] that demonstrate the importance of integrating direction and depth predictions for reliable three-dimensional pose estimation.

F. Classification Model Performance and Generalization

The Support Vector Classifier (SVC) model shows the best performance in classifying the two main classes "still" and "looking around" with an accuracy and F1-Score of 97.35%. This advantage is consistent with literature showing that SVM excels in high-dimensional data situations with limited sample sizes, as well as minimal overfitting risk [14], [15]. In the context of this research, SVC is capable of finding the optimal hyperplane that separates the two classes based on the combination of spatial and temporal features generated from MediaPipe.

Conversely, the Neural Network (NN) model used, namely the Multi-Layer Perceptron (MLP) architecture, shows highly competitive performance with an accuracy of 97.00% and an F1-Score of 97.01%. This architecture adopts two hidden layers and a dropout mechanism, which aligns with best practices in neural network design to reduce overfitting and ensure stable convergence [24]. NN performance is slightly below SVC in this experiment, but shows great potential for further development, especially if the number of classes is expanded and pose variations are increased.

G. Real-Time Stability and Temporal Smoothing Evaluation

One important contribution of this research is the application of a temporal smoothing strategy to improve prediction stability in real-time implementation. The system only updates the label if two conditions are met: prediction probability above 75% and label consistency for 10 consecutive frames. This approach is effective in reducing flickering, namely, prediction changes that occur too rapidly between frames, which can confuse real-world applications such as early warning systems.

In testing on real-time video sequences, the temporal stabilization mechanism successfully maintains label continuity without losing sensitivity to pose changes. This strategy supports findings from [25], [28] that emphasize the importance of detection stability in suspicious activity-based surveillance. This mechanism also strengthens system reliability for use in real-world environments without dependence on human operators.

H. Comparison with Previous Studies

This research extends previous results in the field of Human Activity Recognition (HAR) using skeleton data. The study in [26] by Cao and Li shows that SVM is highly effective in recognizing skeleton-based activities. Additionally, the 15-frame consecutive sequence representation approach used in this research strengthens the temporal modeling approach as proposed in [17], where spatial information alone is often insufficient to distinguish activities with short duration and minimal movement expression.

Compared to texture-based methods such as Gray Level Co-occurrence Matrices (GLCM), the skeleton-based approach used in this research has advantages in terms of robustness to lighting changes and camera orientation [12]. Skeleton-based feature representation is more compact and easier to process in real-time.

I. Implications for Intelligent Surveillance Systems

The findings in this research have broad implications for the development of AI-based surveillance systems, especially in the context of resource-constrained environments such as edge devices. The proposed system can be integrated into modern CCTV architectures that adopt the Object-Based Surveillance paradigm, where physical entities are individually identified and tracked [27].

Implementation of this system also supports the natural surveillance principle from CPTED (Crime Prevention Through Environmental Design) theory, where early detection of suspicious activity can activate automatic warning systems or signal operators to focus on specific points.

Additionally, the compact model architecture and system pipeline open opportunities for deployment on mobile or embedded systems in computationally constrained environments, without sacrificing detection performance. This aligns with current computer vision trends toward lightweight and efficient models for real-time applications.

J. Limitations and Recommendations for Further Development

Although the system shows high performance, this research has several limitations that need to be noted. First, the dataset used is still limited to two main classes. This simplifies the classification task and potentially increases accuracy artificially. Testing on more complex and diverse datasets is needed to assess model robustness to real variations in human activity.

Second, the temporal approach used is still static, in the sense that it considers a fixed time window (15 frames). More adaptive approaches, such as Long Short-Term Memory (LSTM) or Transformer, can be used in the future to capture long-term temporal dynamics and more complex movement patterns.

Third, no analysis has been conducted on system robustness to extreme conditions such as low lighting, angled camera positions, or the presence of multiple individuals in a single frame. Further research can expand the scope of experiments to these conditions to improve system robustness in real-world scenarios.

Overall, this discussion confirms that the two-stage approach combining YOLOv12n and MediaPipe, with support from classification models and temporal stabilization strategies, offers a promising and reliable solution for intelligent surveillance systems. Going forward, adaptation to more complex environments and integration with broader decision-making systems will be important directions for further development.

V. CONCLUSION

This research has successfully developed a suspicious activity detection system based on human pose with an efficient and accurate two-stage approach. By combining object detection using YOLOv12n, pose extraction through MediaPipe BlazePose, and classification using Support Vector Classifier (SVC) and Neural Network (NN), this system is capable of achieving classification accuracy above 97% in distinguishing between "still" and "looking around" poses. The temporal smoothing mechanism applied also successfully improves prediction stability in real-time scenarios, making this system suitable for implementation on resource-constrained automated surveillance devices. The main contribution of this study lies in the optimization of a lightweight detection pipeline while maintaining high performance, as well as in the formulation of spatiotemporal feature representation capable of capturing subtle differences in movement patterns. This enriches the body of research in the field of Human Activity Recognition (HAR), particularly in the context of AI-based surveillance systems. The implications of these findings include potential implementation in public security systems, early detection of suspicious behavior, and integration with automatic warning systems. For future research, testing on more complex datasets, the use of adaptive temporal model architectures such as LSTM or Transformer, and testing in more varied real-world conditions are recommended.

REFERENSI

[1] Marois, A., Lafond, D., Williot, A., Vachon, F., & Tremblay, S. (2020). Real-time gaze-aware cognitive support system for security surveillance. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64(1), 1145-1149. https://doi.org/10.1177/1071181320641274

[2] Monteiro, T., Li, G., Skourup, C., & Zhang, H. (2020). Investigating an integrated sensor fusion system for mental fatigue assessment for demanding maritime operations. Sensors, 20(9), 2588. https://doi.org/10.3390/s20092588

[3] Peukert, M., Claus, L., & Meyer, L. (2025). Subjective and objective fatigue dynamics in air traffic control. Industrial Health, 63(5), 431-442. https://doi.org/10.2486/indhealth.2024-0206

[4] Uğurlu, Ö., Sana, F., Uğurlu, H., & Şenbursa, N. (2025). Analysis of the impact of officers of the watch's mental and physical health issues on collision and grounding accidents. Inquiry the Journal of Health Care Organization Provision and Financing, 62. https://doi.org/10.1177/00469580251371875

[5] Gomes, V. and Cardoso, M. (2024). The effect of sleepiness in situation awareness: a scoping review. Work, 78(3), 641-655. https://doi.org/10.3233/wor-230115

[6] Cardone, C., & Hayes, R. (2012). Shoplifter Perceptions of Store Environments: An Analysis of how Physical Cues in the Retail Interior Shape Shoplifter Behavior. Journal of Applied Security Research, 7(1), 22–58. https://doi.org/10.1080/19361610.2012.631178

[7] Centre for the Protection of National Infrastructure (CPNI). (2016). Hostile reconnaissance: Understanding and countering the threat. London: CPNI.

[8] Kang, J., Shin, J., Shin, J., Lee, D., & Choi, A. (2022). Robust Human Activity Recognition by Integrating Image and Accelerometer Sensor Data Using Deep Fusion Network. Sensors, 22(1), 174. https://doi.org/10.3390/s22010174

[9] Li, T., Zhang, R., & Li, Q. (2020). Multi Scale Temporal Graph Networks For Skeleton-based Action Recognition. arXiv preprint arXiv:2012.02970. https://doi.org/10.48550/arXiv.2012.02970

[10] Li, X., Onie, S., Liang, M., Larsen, M., & Sowmya, A. (2022). Towards building a visual behaviour analysis pipeline for suicide detection and prevention. Sensors, 22(12), 4488. https://doi.org/10.3390/s22124488

[11] Karim, M., Khalid, S., Aleryani, A., Khan, J., Ullah, I., & Ali, Z. (2024). Human action recognition systems: a review of the trends and state-of-the-art. Ieee Access, 12, 36372-36390. https://doi.org/10.1109/access.2024.3373199

[12] Chen, H., Jiang, Y., & Ko, H. (2022). Pose-guided graph convolutional networks for skeleton-based action recognition. Ieee Access, 10, 111725-111731. https://doi.org/10.1109/access.2022.3214812

[13] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). BlazePose: On-device Real-time Body Pose tracking. ArXiv preprint arXiv:2006.10204. https://doi.org/10.48550/arXiv.2006.10204

[14] Tian, Y., Ye, Q., & Doermann, D. (2025). YOLOv12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524. https://doi.org/10.48550/arXiv.2502.12524

[15] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128-138. https://doi.org/10.14445/22312803/IJCTT-V48P126

[16] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249-268.

[17] Gavrilescu, M., & Vizireanu, N. (2019). Feedforward neural network-based architecture for predicting emotions from speech. Data, 4(3). https://doi.org/10.3390/data4030101

[18] Ramirez, H., Velastin, S. A., Aguayo, P., Fabregas, E., & Farias, G. (2022). Human Activity Recognition by Sequences of Skeleton Features. Sensors, 22(11). https://doi.org/10.3390/s22113991

[19] Yang, S., & Berdine, G. (2024). Confusion matrix. The Southwest Respiratory and Critical Care Chronicles, 12(53), 75–79. https://doi.org/10.12746/swrccc.v12i53.1391

[20] Yeh, W., Jiang, Y., Tan, S., & Yeh, C. (2021). A new support vector machine based on convolution product. Complexity, 2021(1). https://doi.org/10.1155/2021/9932292

[21] Balakrishnan, V., Loganathamoorthy, D., Kumar Ayyasamy, R., Alhashmi, S. M., & P, C. (2025). Online suicide ideation detection (OnSIDe): a context-aware transfer learning approach using BERT-CNN. PeerJ Computer Science, 11, e3239. https://doi.org/10.7717/peerj-cs.3239

[22] Alharbi, N., Aljohani, L., Alqasir, A., Alahmadi, T., Alhasiri, R., & Aldajan, D. (2024). Improved automatic drowning detection approach with YOLOv8. Engineering, Technology & Applied Science Research, 14(6), 18070–18076. https://doi.org/10.48084/etasr.8834

[23] Alhammad, N., & Al-Dossari, H. (2021). Dynamic segmentation for physical activity recognition using a single wearable sensor. Applied Sciences, 11(6), 2633. https://doi.org/10.3390/app11062633

[24] Kulkarni, S. R., Lugosi, G., & Venkatesh, S. S. (1998). Learning pattern classification—A survey. *IEEE Transactions on Information Theory*, *44*(6), 2178–2206. https://doi.org/10.1109/18.720536

[25] Castro, W., Oblitas, J., Santa-Cruz, R., & Avila-George, H. (2017). Multilayer perceptron architecture optimization using parallel computing techniques. PLoS ONE, 12(12). https://doi.org/10.1371/journal.pone.0189369

[26] Cheng, J., Zhang, X., Chen, X., Ren, M., Huang, J., & Luo, P. (2022). Early Detection of Suspicious Behaviors for Safe Residence from Movement Trajectory Data. ISPRS International Journal of Geo-Information, 11(9). https://doi.org/10.3390/ijgi11090478

[27] Cao, Y., & Li, T. (2025). SVM action recognition model based on skeletal key point analysis with posture sensors to help sports training. *BMC Sports Science, Medicine and Rehabilitation*, *17*, Article 220. https://doi.org/10.1186/s13102-025-01260-w

[28] Gorodnichy, D. O., & Mungham, T. (2008). Automated video surveillance: challenges and solutions. ACE Surveillance (Annotated Critical Evidence) case study. *Proceedings of the NATO SET-125 Symposium on Sensor and Technology for Defence against Terrorism*, Mannheim, Germany.