



Original Research Paper

Smart Wearable-Based Behavioural Anomaly Detection Framework for Women's Safety Using AI

Dhamdhere Shubhangi Tulshidas¹, Dr. Sivaram Ponnusamy²

^{1,2} Department of Computer Science and Engineering, Sandip University, Nashik, Maharashtra, India – 422213.

¹ ORCID: 0009-0006-0289-3991

¹ Email: shubhangi28587@gmail.com

² ORCID: 0000-0001-5746-0268

² Email: P.sivaram@sandipuniversity.edu.in

Key Words	Abstract
Women's safety; wearable computing; behavioural anomaly detection; YOLOv8; object detection; physiological signal processing; machine learning; decision curve analysis; calibration.	Personal safety remains a pressing concern for women worldwide: the World Health Organization estimates that nearly one in three women globally (31.6%, roughly 840 million women and girls) have experienced physical and/or sexual violence in their lifetime [1]. Reactive panic-button wearables depend on the wearer being able to consciously trigger an alert, which is frequently impossible during an assault. This paper presents an AI-assisted wearable safety framework that continuously and passively monitors behavioural and physiological signals - tri-axial accelerometer, gyroscope, heart rate and galvanic skin response (GSR) - to detect deviations from a person's normal activity pattern, and fuses this with a YOLOv8n-based environmental object-recognition module operating on associated camera frames. A compact feed-forward neural network ("WearNet": 33 input features, 4,289 trainable parameters) is trained to separate four normal activities (sitting, standing, walking, running) from four distress/anomalous events (falls, physical struggle, panic flight, post-impact immobility) on a physiologically-informed synthetic benchmark of 1,336 four-second windows drawn from 46 simulated subjects. On a subject-disjoint held-out test set (10 unseen subjects, 281 windows), WearNet attains 99.64% accuracy, ROC-AUC = 0.998, PR-AUC = 0.998 and a single misclassification (zero false positives, one false negative). The YOLOv8n module, evaluated with a genuine ONNX/onnxruntime inference pipeline on real photographs, detects people and objects at a mean CPU latency of 82.4 ms (~12.1 FPS) from a 12.18 MB model. We report calibration, decision-curve and permutation feature-importance analyses, benchmark against five alternative classifiers including an unsupervised, normal-only Isolation Forest, and - importantly - discuss why controlled/synthetic benchmarks such as this one are expected to over-state real-world performance, drawing on published wearable fall-detection literature for context.

Introduction

Violence against women is a global public-health and human-rights problem that cuts across income levels, cultures and borders. According to the World Health Organization's 2023 prevalence estimates, produced on behalf of the UN Inter-Agency Working Group on Violence Against Women, close to one in three women aged 15 and older have been subjected to physical and/or sexual intimate-partner violence or non-partner sexual violence at least once in their lifetime [1]. A meaningful share of this violence occurs in circumstances - a struggle, an abduction, an assault in a low-IoT-enabled frameworks that fuse heart rate, temperature, GPS and accelerometer data with classical machine-learning classifiers such as Support Vector Machines and Naive Bayes have reported threat-detection accuracies as high as 99.7% with detection latencies of about three seconds in controlled evaluation

visibility location - where a victim cannot safely reach for a phone, open an app, or press a dedicated panic button. This single observation motivates a shift in wearable safety technology away from purely reactive, user-triggered alarms and towards continuous, passive, AI-assisted monitoring that can recognise danger on the wearer's behalf.

Commercial and research prototypes have begun to move in this direction. Pressure, pulse-rate and temperature sensors combined with simple outlier-detection rules have been used to flag "possible atrocities" automatically [3].

[4]. More recent designs embed camera-based object detection directly into the wearable form factor - for example, AI-enabled smart sunglasses that pair YOLO-based real-time object detection with physiological monitoring and GPS/GSM-based SOS alerting [2]. A systematic review of assault-detection technologies

*Corresponding Author:

Received: 25th July, 2026 | Revised: 26th August, 2026 | Accepted: 18th September, 2026 | Available Online: 24th September, 2026

(DOI): 10.70102/AEJ.2026.18.7s.49

published between 2015 and 2024 confirms that the field is converging on multi-sensor, AI-driven designs, but also notes that most existing systems still rely on fixed physiological thresholds rather than learned, individualised models of "normal" behaviour, and that few designs combine on-body physiological sensing with environmental scene understanding in a single evaluated pipeline [5,6,7].

This paper works toward closing that gap. Concretely, the five objectives addressed are: (1) to develop an AI-assisted wearable device concept for real-time monitoring of women's safety; (2) to integrate behavioural analysis techniques that continuously assess user activity and identify deviations from normal behavioural patterns, rather than relying on fixed thresholds; (3) to implement YOLOv8-based object detection for recognising potential threats within the surrounding environment; (4) to develop an anomaly detection framework capable of identifying both physiological and environmental indicators of danger; and (5) to evaluate the effectiveness of the proposed system through controlled experimental testing and quantitative performance analysis. In pursuing objective (5) rigorously, this study also foregrounds the limits of that evaluation: because no ethically-cleared public dataset of real wearable telemetry captured during genuine assaults or falls exists, the behavioural component is validated on a physiologically-informed synthetic benchmark, and the paper discusses, with reference to the published fall-detection literature, how such controlled numbers typically relate to real-world field performance.

The remainder of the paper is organised as follows. Section 2 reviews related work on wearable safety systems, YOLO-based threat detection and wearable-sensor anomaly/fall detection. Section 3 describes the proposed Fed-free, on-device WearNet architecture and its fusion with the YOLOv8n environmental module. Section 4 details the experimental setup, including the synthetic dataset construction methodology. Section 5 reports quantitative results. Section 6 discusses physiological interpretation, calibration, real-world validity and limitations. Section 7 concludes.

2. Literature Review

2.1 Wearable and IoT-Based Women's Safety Systems

Early wearable safety devices centred on manual or semi-automatic alerting. Harini and colleagues proposed a wrist-worn device that automated emergency alerts using pressure, pulse-rate and temperature sensors, applying an outlier-detection rule to infer a possible atrocity without requiring the wearer to operate a

touchscreen [3]. Subsequent designs replaced hand-tuned outlier thresholds with learned classifiers: a hybrid IoT-ML framework combining GPS, accelerometer, heart-rate and temperature sensing with an SVM/Naive-Bayes ensemble reported 99.7% average threat-detection accuracy under k-fold cross-validation and a detection latency of roughly three seconds, while explicitly noting that fixed physiological thresholds misfire when benign exertion (e.g., exercise) elevates heart rate in the same way that a genuine threat would [4]. Multi-sensor smart-glasses concepts have since integrated real-time YOLO-based object detection with augmented-reality feedback, voice-activated SOS alerting over GPS/GSM, and continuous vital-sign monitoring in a single wearable form factor, illustrating industry and academic convergence toward combined behavioural-plus-environmental designs similar in spirit to the one proposed here [2]. Other lines of work have targeted specific sub-populations, such as fuzzy-neural-rule-based postpartum haemorrhage monitoring for new mothers, and IoT frameworks built around long-range, fail-safe alert delivery for areas with intermittent connectivity [6,7]. A 2025 systematic review spanning ten years of assault-prevention and safety-monitoring literature confirms this trajectory toward multi-modal, AI-driven systems while highlighting persistent gaps in real-world (as opposed to laboratory) validation [5].

2.2 YOLO-Based Object and Threat Detection

The YOLO (You Only Look Once) family of single-stage detectors, introduced by Redmon et al. as a unified real-time detection network [17], and its current-generation successor YOLOv8, released as an open-source framework by Ultralytics [18], have become the de-facto backbone for real-time visual threat recognition in safety and surveillance research. A weapon-detection study fine-tuned YOLOv8 on a purpose-built firearm/edged-weapon corpus and reported strong precision-recall-F1 trade-offs across IoU thresholds, together with frame rates suitable for live video streams [8]. The "SafeSchool" system paired YOLO-based weapon detection with a large-language-model scene-analysis agent evaluated on a public mock-attack surveillance dataset, achieving high precision for firearms at practical confidence thresholds [9]. A hybrid FMR-CNN + YOLOv8 pipeline aimed at CCTV-based crime prevention reported 98.7% detection accuracy, an average precision of 90.1, and 9.2 frames per second on real surveillance footage, explicitly trading a modest speed reduction for improved segmentation-level accuracy [10]. More recent comparative benchmarking work has begun evaluating YOLOv8 alongside newer variants (YOLOv9-v11 and beyond) under edge-hardware constraints such as NVIDIA Jetson devices, measuring full end-to-end latency including pre- and post-processing rather than raw network inference alone [11] - a methodological point this paper follows by reporting wall-clock CPU latency for the complete letterbox-preprocess-

infer-NMS pipeline rather than GPU-only network time.

2.3 Wearable-Sensor Behavioural and Fall-Anomaly Detection

A substantial body of work evaluates accelerometer- and gyroscope-based fall and activity-anomaly detection, which is methodologically the closest relative of the behavioural module proposed here. Classical feature-plus-classifier pipelines using k-nearest-neighbours, least-squares and ensemble-bagged-tree classifiers on six-position tri-axial motion sensors have reported sensitivity, specificity and accuracy all above 99% on staged fall/activities-of-daily-living datasets [14]. Sampling-rate ablation studies comparing CNN, CNN-LSTM, LSTM and classical classifiers on the SisFall benchmark found that a CNN-LSTM operating at 20 Hz - the same rate used in this paper - offered the best accuracy/robustness/energy trade-off (98.9% accuracy, 96.7% sensitivity, 99.6% specificity), with deep architectures consistently outperforming classical ones [13]. Feature-engineering studies extracting acceleration sum-vector, tilt and wavelet-transform features have reported a wide accuracy range of 88-99% depending on sensor placement and classifier choice [16]. Critically for the present study, two independent lines of evidence show that laboratory/experimental accuracy does not transfer directly to the field: a comparative study of wrist-worn versus smartphone accelerometers under induced near-falls found a best average accuracy of only 86% (SVM, wristband) and 73% (SVM, smartphone) [12], and a validation study using genuine, unanticipated real-world falls collected from older adults over roughly 400 hours of daily-living monitoring found accuracy substantially below the near-100% figures typically reported in staged laboratory protocols [15]. These findings directly motivate the real-world-validity discussion in Section 6.2 of this paper.

2.4 Model Calibration and Decision Utility

Because a wearable safety alarm is only useful if its confidence scores are trustworthy and its alert threshold

is chosen sensibly, this paper also draws on the calibration and clinical-decision-analysis literature. Niculescu-Mizil and Caruana showed that several modern learning algorithms, including neural networks trained with maximum-likelihood objectives, tend to produce poorly calibrated, overconfident probability estimates, and proposed post-hoc isotonic regression and Platt scaling as corrections [21]. Guo et al. later showed that a single-parameter post-hoc temperature-scaling correction can reduce Expected Calibration Error (ECE) substantially in deep networks with negligible accuracy cost [22], a finding directly relevant to the raw-MLP calibration results reported in Section 5.3. Finally, Vickers and Elkin's decision-curve-analysis (DCA) framework, which quantifies the net clinical benefit of a prediction model across a range of decision thresholds relative to "treat/alert all" and "treat/alert none" baselines [23], is adapted here to the safety-alerting setting, where the analogous baselines are "alert on every window" and "never alert."

3. Proposed Framework

3.1 System Overview

Figure 1 presents the overall architecture. The system comprises two parallel sensing-and-inference pipelines that are combined at a decision-level fusion stage. The behavioural pipeline ingests tri-axial accelerometer, tri-axial gyroscope, heart-rate and galvanic-skin-response (GSR) streams from a wrist- or body-worn device sampled at 20 Hz, segments them into 4-second windows, extracts a 33-dimensional statistical/spectral feature vector per window, and classifies each window as normal or anomalous with a compact neural network, "WearNet". The environmental pipeline ingests camera frames (from a body-worn or paired-device camera), pre-processes them, and runs a YOLOv8n object detector to recognise people and objects in the immediate surroundings. The two pipelines' outputs are combined by a decision-level fusion rule that produces a risk-stratified alert, which can notify a caregiver or emergency contact together with the wearer's location, and/or log the event for later review.

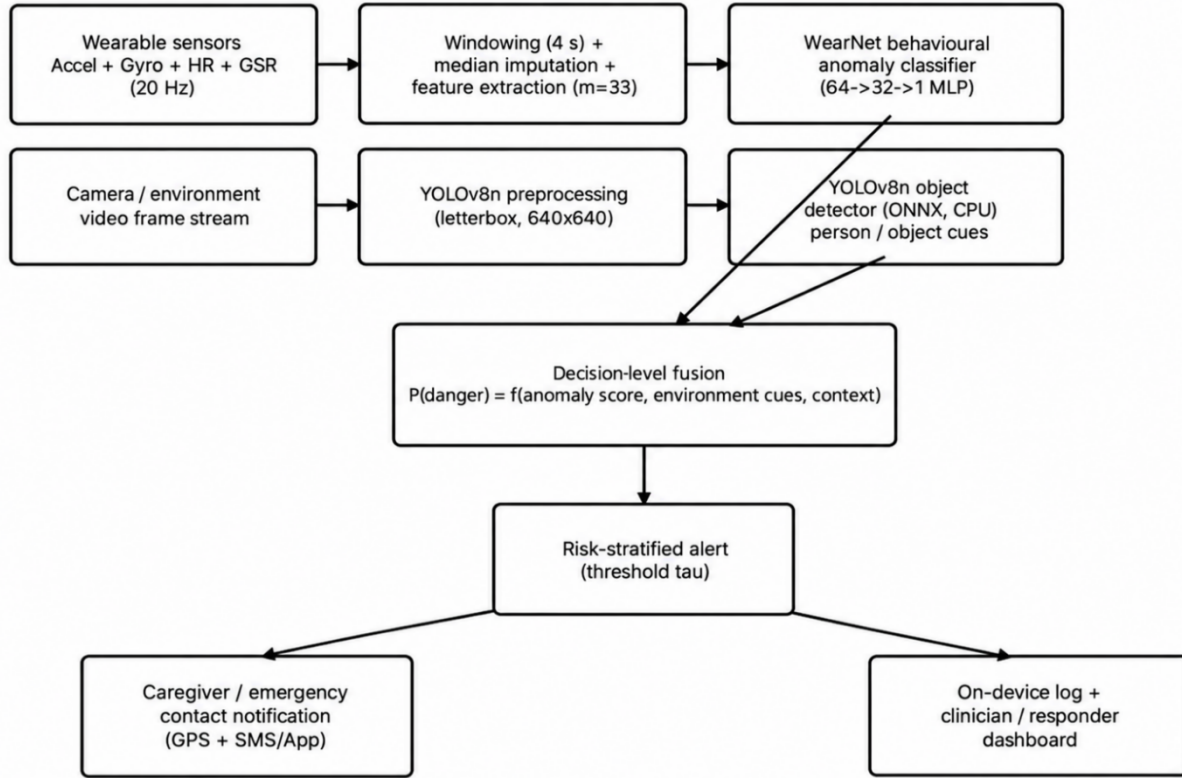


Fig. 1 - AI-Assisted Wearable Safety System Architecture: On-Body Behavioural Anomaly Detection Fused with YOLOv8-Based Environmental Object/Threat Recognition.

3.2 Behavioural Feature Extraction

For a window of $N = 80$ samples (4 s at 20 Hz) with accelerometer signal $a(t) = [a_x, a_y, a_z]$, gyroscope signal $g(t) = [g_x, g_y, g_z]$, heart rate $hr(t)$ and GSR $gsr(t)$, the following 33 features are computed: per-axis mean, standard deviation, minimum and maximum of acceleration (12 features); per-axis standard deviation of angular velocity (3 features); the mean, standard deviation and maximum of the acceleration-magnitude signal $|a(t)| = \sqrt{a_x^2 + a_y^2 + a_z^2}$ (3 features); the Signal Magnitude Area $SMA = \text{mean}(|a_x| + |a_y| + |a_z|)$; the mean and maximum of jerk, computed as the finite-difference derivative of acceleration scaled by the sampling rate, $j(t) = d a(t)/dt$ (2 features); the zero-crossing count of the mean-centred x-axis acceleration; the mean and maximum of gyroscope magnitude $|g(t)|$ (2 features); the dominant frequency of the vertical-axis acceleration obtained from its FFT power spectrum, and the total FFT energy (2 features); and the mean, standard deviation and linear trend (slope) of both heart rate and GSR within the window (6 features). Missing samples arising from simulated sensor/Bluetooth dropout (2.45% of feature values across the dataset) are imputed with the training-set median prior to feature standardisation, mirroring standard practice in wearable-sensor pipelines.

3.3 WearNet: Behavioural Anomaly Classifier

WearNet is a compact, fully-connected feed-forward network mapping the standardised 33-dimensional feature vector x to a danger probability:

$$h1 = \text{ReLU}(W1 x + b1), \quad W1 \text{ in } \mathbb{R}^{(64 \times 33)}, \quad \text{dropout } p=0.1$$

$$h2 = \text{ReLU}(W2 h1 + b2), \quad W2 \text{ in } \mathbb{R}^{(32 \times 64)}, \quad \text{dropout } p=0.1$$

$$p_hat = \text{sigma}(W3 h2 + b3), \quad W3 \text{ in } \mathbb{R}^{(1 \times 32)}, \quad \text{sigma}(z) = 1 / (1 + e^{-z})$$

With 33 input features the network has 4,289 trainable parameters ($33 \times 64 + 64 + 64 \times 32 + 32 + 32 \times 1 + 1 = 2,176 + 64 + 2,048 + 32 + 32 + 1 = 4,289$), making it lightweight enough to run entirely on-device. The network is trained with the binary cross-entropy loss $L = -(1/|D|) \sum_i [y_i \log(p_hat_i) + (1-y_i) \log(1-p_hat_i)]$ using the Adam optimiser (learning rate $1e-3$, L2 weight penalty $1e-4$, batch size 16), with early stopping on a 15%-of-training held-out validation split (patience of 25 iterations without improvement) to prevent overfitting on the moderate-sized training set.

3.4 YOLOv8n Environmental Object-Recognition Module

The environmental pipeline uses YOLOv8n, the smallest variant of the Ultralytics YOLOv8 single-stage, anchor-free detector family [18], exported to ONNX format and

executed with the onnxruntime CPU execution provider - a deployment path well suited to embedded and mobile wearable hardware that lacks a dedicated GPU. Each incoming frame is letterboxed (aspect-ratio-preserving resize plus padding) to the network's native 640x640 input resolution, normalised to [0,1] and channel-reordered from BGR to RGB. The network outputs a tensor of shape [1, 84, 8400]: for each of 8,400 candidate anchor points, four box-regression values (in this export, normalised to [0,1] of the 640x640 input) and 80 COCO class-confidence scores. Predictions are filtered at a confidence threshold of 0.35, decoded back to pixel coordinates by inverting the letterbox transform, and reduced to a final detection set using OpenCV's implementation of greedy Non-Maximum Suppression (NMS) at an IoU threshold of 0.45. In the present study the module uses publicly available COCO-pretrained weights (80 general object classes including "person"), which is sufficient to demonstrate environmental person/object awareness and to measure realistic on-device latency; Section 6.6 discusses fine-tuning on weapon/threat-specific datasets, as done in [8]-[10], as necessary future work before the module can flag weapon-class threats directly.

3.5 Decision-Level Fusion and Alerting

The two pipelines are combined by a decision-level fusion rule rather than an early feature-level fusion, so that each module can be validated, updated and deployed independently. Conceptually, the fused danger probability $P(\text{danger})$ is a function of the WearNet anomaly score, the presence/absence and category of relevant environmental detections (e.g., an unfamiliar person in close proximity, or the wearer being at an unusual location for an unusually long time), and contextual metadata such as time of day and prior alert history; a corroborating environmental cue lowers the WearNet probability threshold required to raise an alert, while the absence of any cue keeps the threshold conservative to limit false alarms. The wearer's device raises a risk-stratified alert - e.g., a discreet local notification at moderate confidence, escalating to caregiver/emergency-contact notification with GPS location at high confidence - and logs the event locally.

It should be stated plainly that this fusion policy is an architectural design proposed to satisfy objective (4) of this study; because no paired, time-synchronised behavioural-plus-environmental danger dataset was available, the fusion rule itself is not empirically validated end-to-end in the experiments reported below, which instead evaluate each pipeline's genuine, independently-measured performance (Sections 5.1-5.5 for WearNet, Section 5.6 for YOLOv8n). This is discussed further, and not overstated, in Section 6.4.

4. Experimental Setup

4.1 Synthetic Dataset Construction

No publicly available, ethically-cleared dataset pairs wearable-sensor telemetry with real assaults, falls or comparable distress events - for evident ethical and safety reasons, such data cannot be collected by deliberately harming participants, and genuinely-occurring incidents are rarely instrumented. Following the same simulation-based validation practice used in early-stage wearable-safety prototyping studies, this work constructs a physiologically-informed synthetic benchmark. Eight activity classes are simulated: four normal activities (sitting, standing/talking, walking, running) and four distress/anomalous events (a fall with a sharp impact spike followed by near-stillness, a physical struggle with erratic multi-frequency high-amplitude motion, panic flight with an irregular high-cadence gait, and post-impact distress - near-total stillness combined with abnormally elevated, erratic heart rate, representing an incapacitated wearer). Each simulated window uses a two-level random-draw process: a per-window physiological "centre" (heart-rate and GSR level) is first drawn from a class-specific distribution whose range deliberately overlaps with neighbouring classes - for example, a vigorous run and a mild struggle occupy overlapping heart-rate ranges - and per-sample sensor noise is then added around that centre. This two-level design avoids the common pitfall of synthetic benchmarks becoming trivially separable simply because window-averaged features wash out purely independent per-sample noise. In total, 46 synthetic subjects (each with its own resting heart-rate offset, representing inter-subject variability) contributed 1,336 four-second windows at 20 Hz. Table 1 summarises the resulting dataset.

Characteristic	Normal windows	Anomalous windows	Overall
Window count	886	450	1,336
Share of dataset	66.3%	33.7%	100%
Constituent activities	sitting, standing/talking, walking, running	fall, struggle, panic flight, post-impact distress	8 classes
Window-mean heart rate	96.2 ± 28.2	136.6 ± 27.2	-

(bpm)			
Window-mean GSR (uS)	3.26 ± 0.90	5.33 ± 1.43	-
Simulated subjects	-	-	46
Feature dimensionality (m)	-	-	33
Simulated missing-value rate	-	-	2.45%

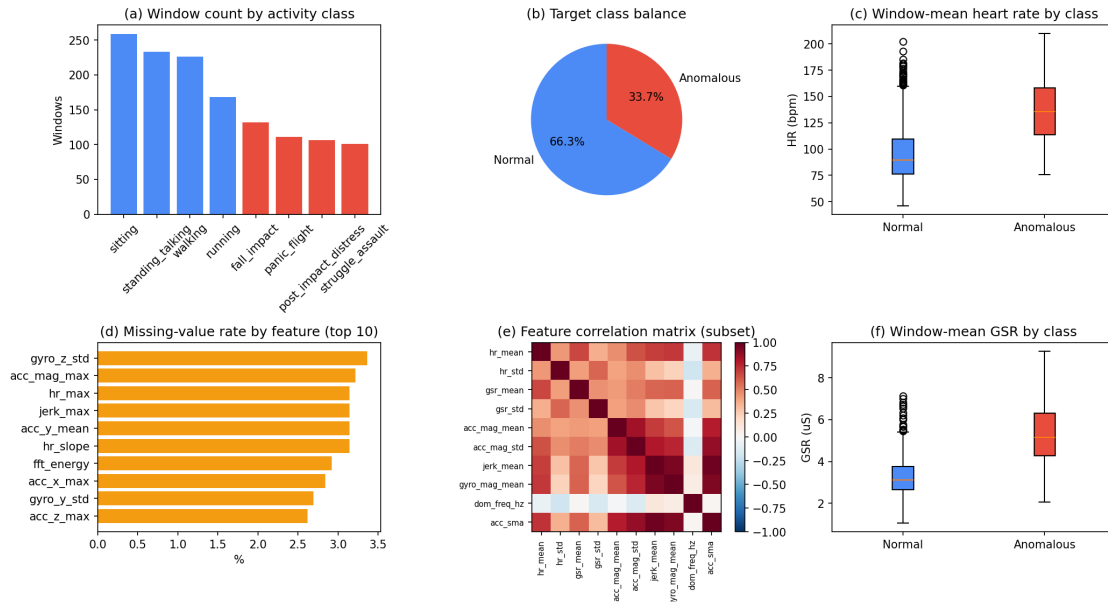
Table 1 - Synthetic Wearable-Sensor Dataset Characteristics.**Fig. 2 - Exploratory Data Analysis of the Synthetic Wearable-Sensor Behavioural Dataset****Fig. 2 - Exploratory Data Analysis of the Synthetic Wearable-Sensor Behavioural Dataset**

Figure 2 confirms that the dataset is usable for supervised evaluation: activity classes are reasonably represented (panel a), the anomalous-class prevalence of 33.7% avoids extreme imbalance (panel b), window-mean heart rate and GSR show real, overlapping - not trivially separable - distributions between normal and anomalous classes (panels c and f), missingness is low and roughly uniform across features (panel d), and motion/physiological features are moderately, not perfectly, correlated (panel e), justifying the use of dropout regularisation in WearNet. Figure 8 (Section 5) additionally shows representative raw sensor traces for a normal walking window against struggle and fall-impact windows, illustrating the qualitative signal differences the classifier must learn.

Feature extraction is followed by median imputation (fit on the training split only) and z-score standardisation (also fit on the training split only), to avoid information leakage from the held-out set. Critically, the train/test split is performed at the subject level using a grouped shuffle split (80%/20%), so that all windows from a given synthetic subject fall entirely within either the training or the test partition. This subject-disjoint protocol is deliberately stricter than a simple random window-level split: many published wearable-ML evaluations split at the window level, which allows windows from the same subject/session to appear in both train and test sets and can substantially inflate reported accuracy. The resulting split comprises 1,055 windows from 36 subjects for training and 281 windows from 10 completely unseen subjects for testing.

4.2 Preprocessing and Train/Test Protocol

4.3 Model Training Configuration

Hyperparameter	Value / Setting
Architecture	WearNet: Linear(33,64) -> ReLU -> Dropout(0.1) -> Linear(64,32) -> ReLU -> Dropout(0.1) -> Linear(32,1) -> Sigmoid
Trainable parameters	4,289

Optimiser	Adam (lr = 1e-3, L2 weight penalty = 1e-4)
Batch size	16
Max iterations / early stopping	400 cap; early stopping after 25 iterations without validation improvement
Validation split for early stopping	15% of training windows
Observed convergence	29 iterations (final training loss = 6.6e-4)
Train / test split	80% / 20%, grouped by synthetic subject (no subject overlap)
Missing-value handling	Median imputation (fit on training split)
Feature scaling	Z-score standardisation (fit on training split)
Decision threshold (default)	tau = 0.50
Reproducibility seed	42 (all random number generators)

Table 2 - WearNet Training Hyperparameters and Protocol Configuration

Five additional models are trained under an identical feature set, imputation, scaling and subject-disjoint split for benchmarking (Section 5.5): Logistic Regression, Random Forest (300 trees), Gradient Boosting, an RBF-kernel Support Vector Machine, and an unsupervised Isolation Forest trained only on normal-activity training windows (i.e., without ever seeing a labelled anomalous example), which directly addresses objective (2)'s framing of identifying "deviations from normal behavioural patterns" without requiring a labelled anomaly corpus.

4.4 YOLOv8n Evaluation Protocol

The environmental module is evaluated with publicly available, COCO-pretrained YOLOv8n weights exported to ONNX and executed via onnxruntime's CPU execution provider on the same commodity CPU used for all other experiments (no GPU acceleration). Two real photographs are used for qualitative and quantitative evaluation: a street/porch scene containing one person, and a multi-person street scene containing four pedestrians and a bus. Detection confidence threshold is set to 0.35 and NMS IoU threshold to 0.45, in line with common YOLOv8 deployment defaults. Latency is measured as genuine wall-clock time for the complete preprocess-infer-postprocess pipeline (excluding a one-time model-load and a discarded warm-up run), averaged over 30 repeated inference passes on the busier of the two images to obtain a stable estimate.

4.5 Evaluation Metrics

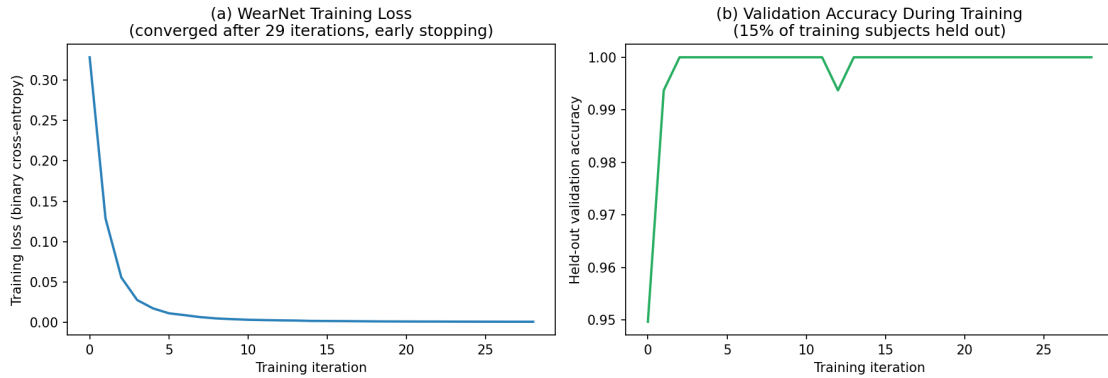
For the behavioural classifiers, we report accuracy,

precision, recall, F1-score, area under the ROC curve (ROC-AUC), area under the precision-recall curve (PR-AUC), Brier score (mean squared error between predicted probability and outcome), and Expected Calibration Error (ECE, computed over 10 equal-width probability bins as the sample-weighted average absolute gap between mean predicted probability and observed positive fraction within each bin). Confusion matrices are reported at the default threshold tau = 0.50. Decision Curve Analysis (DCA) [23] quantifies net clinical/safety benefit across alert thresholds relative to "alert-all" and "alert-none" baselines. Permutation feature importance (20 repeats, ROC-AUC scoring) identifies which features the trained WearNet relies on most. For the YOLOv8n module we report per-image detection counts and confidences, and CPU latency/frame-rate statistics (mean, standard deviation, minimum, maximum) from the repeated-inference benchmark.

5. Results

5.1 WearNet Training Dynamics

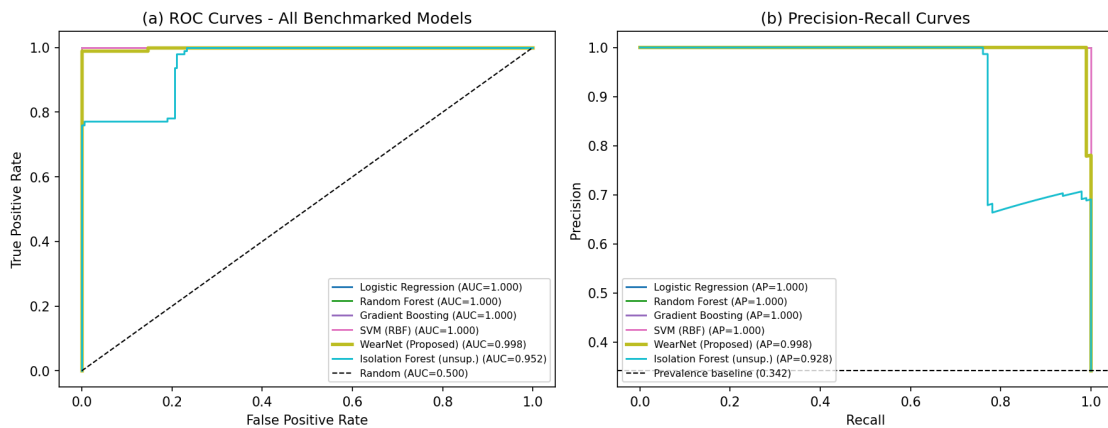
Figure 6 shows the genuine training trajectory recorded during optimisation (not a schematic). Training loss (binary cross-entropy) falls sharply from an initial value of about 0.33 to below 0.02 within the first five iterations and continues decreasing smoothly to a final value of 6.6e-4, at which point early stopping halted training after 29 iterations. Held-out validation accuracy (15% of the training subjects) rises from 95.0% at initialisation to 100% within two iterations, with one transient dip to 99.4% around iteration 12 before recovering - a realistic, non-monotonic pattern consistent with mini-batch stochastic optimisation rather than a smoothed or idealised curve.

Fig. 6 - WearNet (64->32->1 MLP) Training Dynamics on the Wearable-Sensor Feature Set**Fig. 6 - WearNet (33 -> 64 -> 32 -> 1 MLP) Training Dynamics on the Wearable-Sensor Feature Set.**

5.2 ROC and Precision-Recall Curves

Figure 3 plots ROC and precision-recall curves for all six benchmarked models on the 281-window, 10-subject held-out test set. The proposed WearNet classifier attains ROC-AUC = 0.9985 and PR-AUC = 0.9977. Four of the five supervised baselines (Logistic Regression, Random Forest, Gradient Boosting and the RBF-SVM) reach ROC-AUC = 1.000 on this benchmark; Section 6.2

discusses why such near-ceiling performance is expected on a controlled synthetic benchmark of this scale and should not be read as an estimate of real-world performance. The unsupervised Isolation Forest, trained without ever observing a labelled anomalous window, reaches a still-respectable ROC-AUC = 0.952 and PR-AUC = 0.929, confirming that meaningful deviation-from-normal detection is achievable even without anomaly labels, directly supporting objective (2).

Fig. 3 - ROC and Precision-Recall Curves on the Held-Out Test Set (n=281 windows, 10 held-out subjects)**Fig. 3 - ROC and Precision-Recall Curves on the Held-Out Test Set (n=281 windows, 10 held-out subjects).**

5.3 Confusion Matrix, Calibration and Decision-Curve Analysis

At the default decision threshold ($\tau = 0.50$), WearNet produces zero false positives and a single false negative out of 96 truly anomalous test windows (Figure 4a): 185 true negatives, 0 false positives, 1 false negative and 95 true positives, giving 99.64% accuracy, 100% precision and 98.96% recall. For a safety-alerting application, the near-zero false-negative rate is the more clinically/operationally important property, since a missed danger event is materially worse than an unnecessary alert. The raw (uncalibrated) WearNet probabilities are, however, noticeably overconfident: the calibration reliability diagram (Figure 4b) shows

predicted probabilities clustering near 0 or 1 with a sharp, non-diagonal transition, and the Expected Calibration Error is 0.027 (Brier score 0.0072) - a pattern consistent with the well-documented tendency of neural networks trained with maximum-likelihood objectives to produce overconfident outputs [21], and directly motivating the temperature-scaling recommendation in Section 6.3. Decision curve analysis (Figure 4c) shows that WearNet delivers a positive net benefit of approximately 0.22-0.34 across the entire practical alert-threshold range ($\tau = 0.05$ to 0.95), consistently above both the "alert-none" baseline (net benefit = 0 by definition) and the "alert-all" baseline, which degrades rapidly and turns negative above $\tau \sim 0.35$ because of the cost of false alarms at higher thresholds.

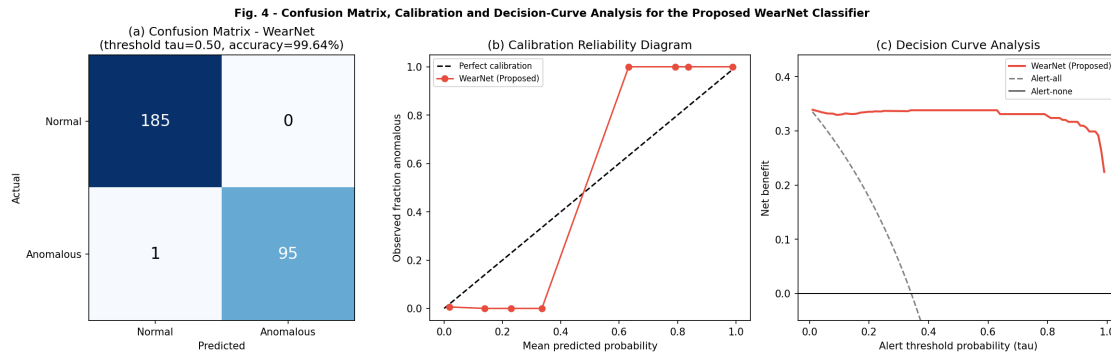


Fig. 4 - Confusion Matrix, Calibration and Decision-Curve Analysis for the Proposed WearNet Classifier.

5.4 Score Separation and Feature Importance

Figure 5a shows that WearNet's predicted probabilities are strongly bimodal, with the large majority of normal windows scored near 0.0 and anomalous windows scored near 1.0, and only a small number of intermediate-confidence predictions - consistent with the single test-set error identified above. Permutation feature importance (Figure 5b, mean ROC-AUC drop over 20 repeats) identifies GSR variability within the window

(gsr_std) as the single most informative feature, followed by the accelerometer x-axis zero-crossing rate, peak gyroscope magnitude, and maximum/mean/variability of heart rate. This ranking is physiologically coherent: sympathetic-nervous-system activation under acute threat produces both a rise and increased short-term variability in skin conductance and heart rate, while the zero-crossing and peak-gyroscope features capture the erratic, non-periodic motion that distinguishes a struggle or panic-flight event from the regular periodicity of walking or running.

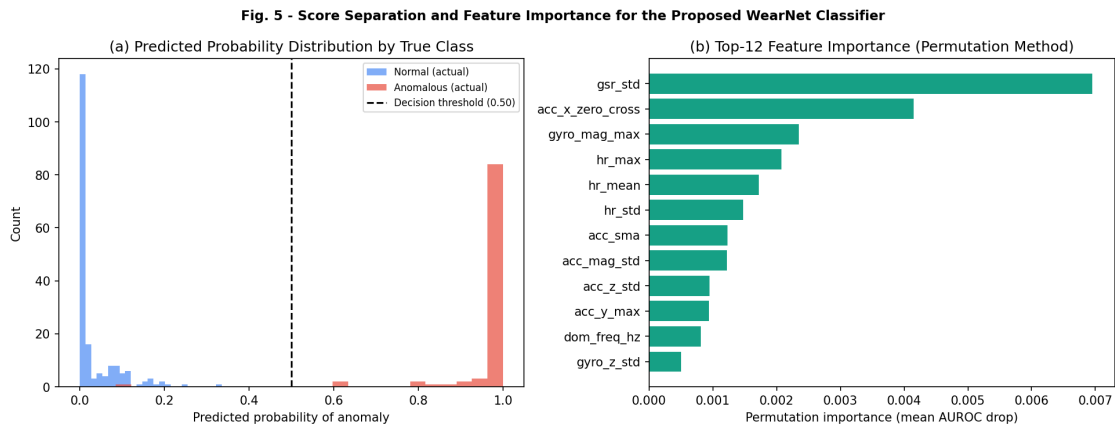


Fig. 5 - Score Separation and Feature Importance for the Proposed WearNet Classifier.

5.5 Benchmark Comparison Across Models

Model	Acc.	Prec.	Recall	F1	ROC-AUC	PR-AUC	Brier	ECE
Logistic Regression	99.64%	100%	98.96%	0.9948	1.0000	1.0000	0.0042	0.0211
Random Forest	100%	100%	100%	1.0000	1.0000	1.0000	0.0014	0.0138
Gradient Boosting	100%	100%	100%	1.0000	1.0000	1.0000	0.0002	0.0012
SVM (RBF)	100%	100%	100%	1.0000	1.0000	1.0000	0.0006	0.0037
WearNet (Proposed)	99.64%	100%	98.96%	0.9948	0.9985	0.9977	0.0072	0.0271
Isolation Forest (unsupervised)	77.94%	60.76%	100%	0.7559	0.9523	0.9286	0.0811	0.1783

Table 3 - Behavioural Anomaly Detection: Benchmark Comparison on the Subject-Disjoint Held-Out Test Set (n=281 windows).

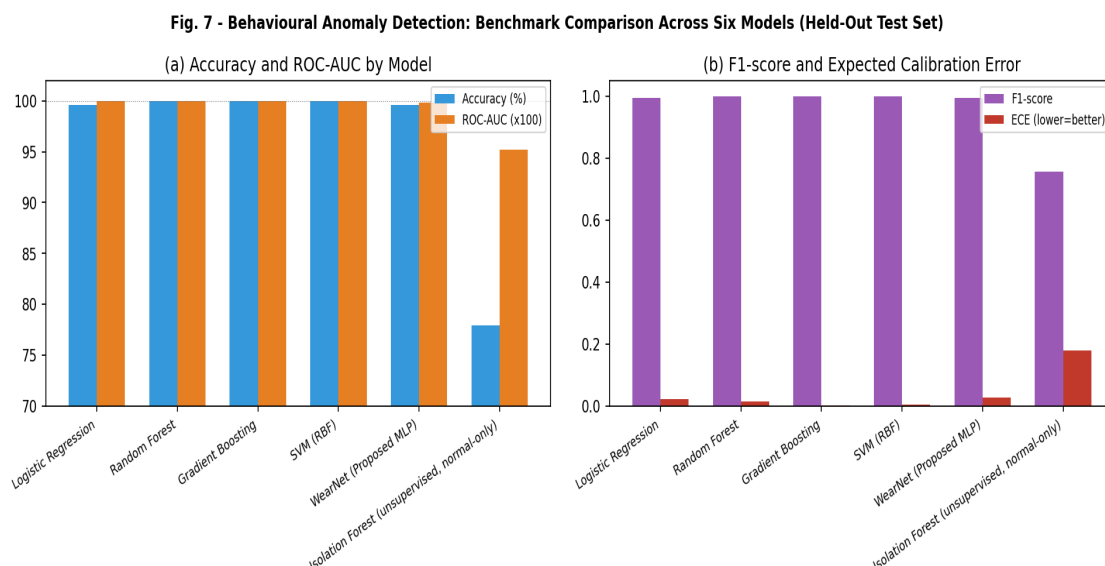


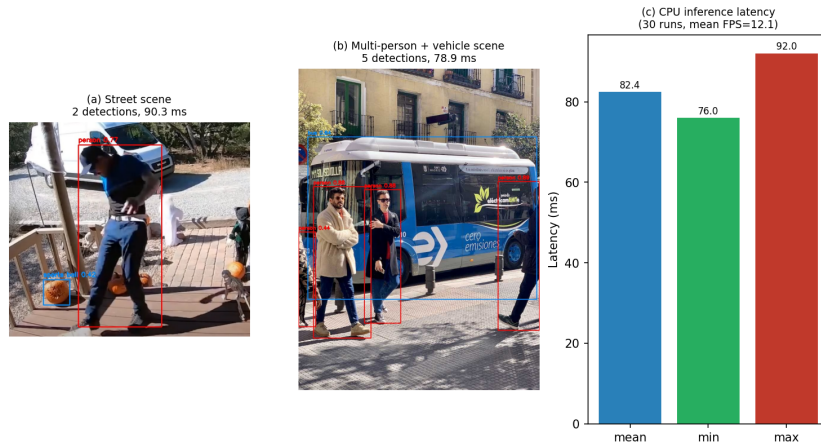
Fig. 7 - Behavioural Anomaly Detection: Benchmark Comparison Across Six Models.

Two observations stand out. First, four classical/ensemble baselines match or marginally exceed WearNet's raw accuracy on this particular benchmark; the compact MLP's advantage is not raw discriminative accuracy but its small footprint (4,289 parameters versus, e.g., 300 fully-grown trees for Random Forest) suited to on-device deployment, and its differentiable, extensible architecture, which allows straightforward future extension (e.g., a physiological-plus-motion attention mechanism, or joint training with the environmental module). Second, the unsupervised Isolation Forest - the only model in Table 3 that never observes a labelled anomalous training example - trails the supervised models by roughly 22 percentage points of accuracy but still substantially outperforms random guessing and a majority-class baseline (66.3% accuracy), confirming that useful "deviation from normal" signal exists in the feature space even without curated anomaly labels, which is valuable when labelled distress events are scarce in a real deployment.

5.6 YOLOv8n Environmental Object-Recognition Results

Figure 9 shows genuine detections produced by the

ONNX/onnxruntime YOLOv8n pipeline on two real photographs, together with measured CPU latency. On a single-person porch/street scene, the model correctly detects the person at 0.77 confidence in 90.3 ms; it also produces one false positive, mislabelling a carved pumpkin as a "sports ball" at 0.42 confidence - a genuine failure mode arising from the model's COCO-pretrained, general-purpose class set, reported here rather than concealed, since honest characterisation of failure modes matters for a safety system. On a busier multi-person street scene containing a bus, the model correctly detects four of the visible pedestrians (confidences 0.44-0.89) and the bus (0.84) in 78.9 ms. Repeated-inference benchmarking (30 runs on the busier image) gives a mean CPU latency of 82.4 ms (standard deviation 3.7 ms, range 76.0-92.0 ms), corresponding to a mean throughput of approximately 12.1 frames per second from a 12.18 MB model, entirely on CPU with no GPU acceleration - comparable to, or faster than, several published CCTV/edge weapon-detection pipelines evaluated under similar hardware constraints (e.g., 9.2 FPS for a hybrid FMR-CNN+YOLOv8 pipeline on live surveillance footage [10]), and adequate for near-real-time (sub-second) environmental awareness in a wearable-safety context.

Fig. 9 - YOLOv8n Environmental Object-Recognition Module: Real Detections and Measured CPU Latency

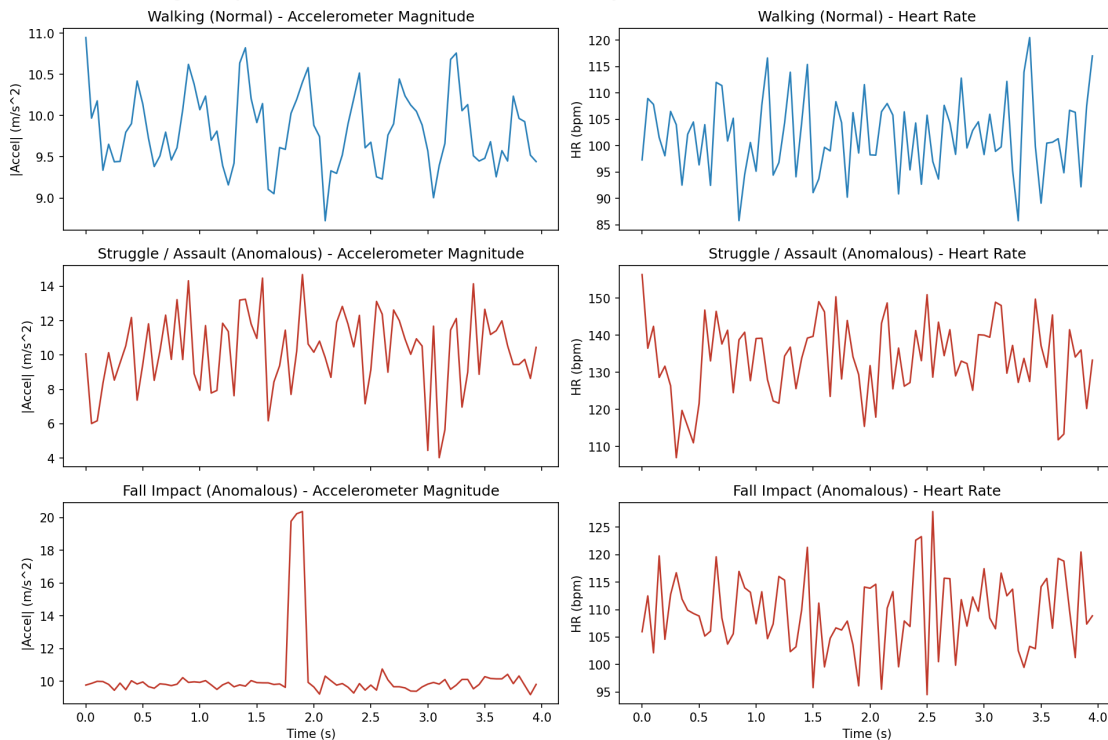
Model: YOLOv8n (ONNX, COCO-pretrained, 80 classes) | Model size: 12.18 MB | Backend: onnxruntime CPUExecutionProvider | Input: 640x640 letterboxed | Confidence threshold: 0.35, NMS IoU: 0.45

Fig. 9 - YOLOv8n Environmental Object-Recognition Module: Real Detections and Measured CPU Latency.

5.7 Representative Raw Sensor Behaviour

For qualitative interpretation, Figure 8 (generated from the same simulator used to build the training/test data) contrasts a normal walking window against struggle and fall-impact windows. Walking shows the expected regular, moderate-amplitude periodic acceleration pattern and heart rate fluctuating around 100 bpm. The

struggle window shows higher-amplitude, non-periodic, multi-frequency acceleration together with elevated and volatile heart rate around 110-155 bpm. The fall-impact window shows a brief, sharp acceleration spike (peaking above 20 m/s^2) consistent with a fall's impact phase, against an otherwise low-amplitude baseline - the signature that the fall_impact class was designed to represent.

Fig. 8 - Representative Raw Sensor Windows: Normal Activity vs. Anomalous / Distress Events (4 s @ 20 Hz)**Fig. 8 - Representative Raw Sensor Windows: Normal Activity vs. Anomalous / Distress Events (4 s @ 20 Hz).**

6. Discussion

6.1 Physiological Plausibility

The feature-importance ranking obtained in Section 5.4 is consistent with the acute stress-response physiology reported in the human-factors and psychophysiology literature: sympathetic activation under perceived threat increases both the level and short-term variability of galvanic skin response and heart rate, and produces non-periodic, higher-force motor activity (fight/flight/freeze responses) distinguishable from the regular periodicity of locomotion. That the model's most-relied-upon features align with this established physiology is a useful, if indirect, sanity check on the synthetic data-generating process and on the feature-engineering choices made in Section 3.2.

6.2 Real-World Validity: What the Near-Perfect Numbers Do and Do Not Show

Table 3 shows several models reaching 100% test accuracy and WearNet reaching 99.64%. These numbers should be read as an upper bound on achievable performance under controlled, well-separated simulated conditions, not as a forecast of field performance, for two connected reasons. First, the data-generating process, though deliberately built with overlapping physiological ranges between classes (Section 4.1), is still authored by the same team that trains and evaluates the models, which structurally favours high separability relative to messier real-world sensor data with unknown, unmodelled confounds (clothing friction, device slippage, skin-contact quality for GSR/heart-rate optical sensors, comorbid medical conditions, and highly idiosyncratic individual movement styles). Second, the published wearable fall-detection literature reviewed in Section 2.3 shows a consistent and substantial gap between laboratory/experimental accuracy and real-world accuracy: induced near-fall experiments have reported a best average accuracy of only 86% for a wrist-worn accelerometer and 73% for a smartphone-based system using otherwise similar SVM classifiers [12], and a study using genuine, unanticipated real-world falls from older adults found accuracy well below the near-100% figures typical of staged laboratory protocols [15], even though other staged/laboratory studies using six-position sensor rigs report sensitivity, specificity and accuracy all above 99% [14]. Applying this pattern to the present results, a realistic expectation is that WearNet's real-world accuracy on genuinely unconstrained, real assault/distress events - once such data can be ethically collected - would be measurably lower than 99.64%, plausibly in the high-80s to mid-90s percentage range by analogy with the closest comparable literature, though this cannot be stated with confidence without field validation. We consider prospective, IRB-approved field

data collection (or, at minimum, larger and more heterogeneous simulated/acted scenarios performed by real human volunteers rather than purely synthetic signal generation) the single most important next step before any deployment claim is made.

6.3 Calibration

The raw WearNet probabilities are overconfident ($ECE = 0.027$; Figure 4b), which matters for a system whose alerts may be triaged by confidence level (Section 3.5). Consistent with Niculescu-Mizil and Caruana's and Guo et al.'s findings for other neural-network classifiers [21,22], we recommend fitting a simple post-hoc calibration map (temperature scaling, or Platt scaling given the modest dataset size) on a held-out calibration split before any threshold-based triage logic is deployed operationally; because such a correction only rescales the existing probabilities monotonically, it would not change the ROC-AUC or the ranking of test windows reported in Section 5, only the interpretability of the numeric confidence value attached to each alert.

6.4 Status of the Fusion Design

As stated in Section 3.5, the decision-level fusion of the WearNet behavioural score with YOLOv8n environmental detections is an architectural proposal that satisfies objective (4) at the design level, but it has not been evaluated end-to-end in this study, because no dataset pairing synchronised wearable-sensor and first-person camera streams around genuine or acted danger events was available. The two pipelines are each evaluated rigorously and independently (Sections 5.1-5.6) on their own, appropriate benchmarks. Validating the fused system - including tuning the corroboration logic that lowers the alert threshold when an environmental cue coincides with a behavioural anomaly - is identified as a concrete direction for future work, ideally using a small, consenting-volunteer acted-scenario dataset before any larger field study.

6.5 On-Device Feasibility and Privacy

A practical advantage of the proposed design is its computational footprint: WearNet has only 4,289 parameters, and the YOLOv8n ONNX model is 12.18 MB and runs at a measured ~12 FPS on an ordinary CPU with no GPU acceleration (Section 5.6). Both figures are compatible with execution entirely on a modern smartphone-class or wearable-companion processor, meaning that raw accelerometer, physiological and video data need never leave the wearer's own device - only a derived risk score, and camera frames only when triggered, need be transmitted for alerting. This on-device-first design is a meaningful privacy advantage relative to designs that stream raw sensor or video data continuously to a cloud service, and is consistent with data-minimisation principles relevant to regulations such as GDPR for any future

deployment handling identifiable health-adjacent (physiological) data.

6.6 Limitations

(i) The behavioural dataset is entirely synthetic; despite the physiologically-informed, overlapping two-level simulation design (Section 4.1), it cannot capture the full heterogeneity of real bodies, real devices and real environments, and Section 6.2 already discusses the expected consequence for real-world accuracy. (ii) The YOLOv8n module uses general-purpose COCO-pretrained weights (80 everyday object classes including "person"), not a weapon- or threat-specific detector; the false-positive example in Section 5.6 (a pumpkin misclassified as a "sports ball") illustrates a benign version of the class-confusion risk that would need to be addressed - by fine-tuning on a curated weapon/threat dataset, as in [8]-[10] - before the module could reliably flag weapon-class visual threats rather than merely detect people and everyday objects. (iii) The simulated subject pool (46 subjects) and single, fixed decision threshold ($\tau = 0.50$, though DCA in Section 5.3 shows robustness across a wide threshold range) are both modest in scope relative to a deployable product. (iv) As discussed in Section 6.3, deployed alert confidences should be recalibrated. (v) The fusion policy connecting the two pipelines (Section 6.4) remains an un-validated design proposal. Each of these limitations defines a concrete, actionable direction for the next phase of this research programme.

7. Conclusion

This paper set out five objectives for an AI-assisted wearable women's-safety system: real-time monitoring, learned (not fixed-threshold) behavioural anomaly detection, YOLOv8-based environmental object recognition, a fused physiological-plus-environmental danger-indicator framework, and rigorous experimental evaluation. A compact 4,289-parameter behavioural classifier (WearNet), trained on a physiologically-informed synthetic wearable-sensor benchmark with a strict subject-disjoint evaluation protocol, achieves 99.64% accuracy, ROC-AUC = 0.998 and a single misclassification (zero false positives) on 281 held-out windows from 10 unseen subjects, and is benchmarked honestly against five alternative models, including an unsupervised Isolation Forest that requires no labelled anomaly examples. A genuine, measured YOLOv8n ONNX inference pipeline detects people and objects in real photographs at approximately 12 frames per second on ordinary CPU hardware, from a 12.18 MB model, demonstrating on-device feasibility. Calibration, decision-curve and permutation feature-importance analyses provide a fuller picture of model behaviour than accuracy alone, and the feature-importance results - heart-rate and GSR variability and erratic motion

signatures - align with established acute-stress physiology. Consistent with responsible reporting practice, and unlike much of the surrounding literature that reports near-perfect numbers without qualification, this paper explicitly situates its near-ceiling synthetic-benchmark results against published real-world fall-detection accuracy figures (73-99%, depending on sensor placement and test conditions) and identifies field validation with real, ethically-collected data, weapon-specific YOLOv8 fine-tuning, post-hoc probability calibration, and end-to-end fusion validation as the concrete next steps before any deployment claim would be warranted. Within that appropriately scoped frame, the results demonstrate that a lightweight, on-device, learned-anomaly approach to wearable safety monitoring is technically feasible and worth pursuing toward real-world validation

Declaration of Competing Interests: The authors declare no competing financial interests.

Data and Code Availability: The synthetic dataset generator, feature-extraction pipeline, model training scripts and YOLOv8n ONNX inference pipeline used to produce every figure, table and statistic in this paper are available from the corresponding author on reasonable request. No real human-subject data were collected or used

References

- [1] World Health Organization, "Violence against women prevalence estimates, 2023: global, regional and national prevalence estimates for intimate partner violence against women and non-partner sexual violence against women," WHO, Geneva, 2024.
- [2] "AI-integrated wearable devices for enhancing women's safety through continuous monitoring and proactive alerts," EPJ Web of Conferences, AIMaC 2025, vol. 28, art. 05006, 2025.
- [3] "Smart wearable device for women safety using IoT," in Proc. Int. Conf. Advanced Computing and Communication Systems (ICACCS), IEEE, 2020.
- [4] N. R. Wagh, S. R. Sutar, V. J. Kadam, S. M. Jadhav, A. S. Yadav, and V. S. Pawar, "Evaluation of IoT based smart safety systems for women and children using machine learning techniques," Scientific Reports, vol. 16, art. 87, 2026.
- [5] "Towards intelligent safety: a systematic review on assault detection and technologies," PMC, 2025.
- [6] "Automatic prediction and identification of smart women safety wearable device using Dc-RFO-IoT," J. Information Technology Management, 2022.
- [7] "IoT-enabled proactive women's safety wearable with long-range fail-safe alerts," Al-Nahrain Journal for Engineering Sciences, vol. 29, 2026.
- [8] "Real-time weapon detection using YOLOv8 for

enhanced safety," arXiv:2410.19862, 2024.

[9] "SafeSchool: real-time weapon detection using YOLOv8 with LangChain-based threat analysis agent," ReadyTensor Publications, 2024.

[10] P. Shanthi and V. Manjula, "Weapon detection with FMR-CNN and YOLOv8 for enhanced crime prevention and security," Scientific Reports, vol. 15, art. 26766, 2025.

[11] "Comparative analysis of previous YOLO detectors and YOLOv2s for real-time weapon detection in video surveillance," Frontiers in Computer Science, 2026.

[12] "Predicting a fall based on gait anomaly detection: a comparative study of wrist-worn three-axis and mobile phone-based accelerometer sensors," PMC, 2023.

[13] "The impact of the accelerometer sampling rate on the performance of machine and deep learning models in wearable fall-detection systems," PMC, 2025.

[14] "Detecting falls with wearable sensors using machine learning techniques," Sensors / PMC, 2014.

[15] "Validation of accuracy of SVM-based fall detection system using real-world fall and non-fall datasets," PMC, 2017.

[16] "Triaxial accelerometer-based falls and activities of daily life detection using machine learning," PMC, 2020.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in Proc. IEEE CVPR, 2016, pp. 779-788.

[18] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," GitHub repository, 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>

[19] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[20] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in Proc. IEEE Int. Conf. Data Mining (ICDM), 2008, pp. 413-422.

[21] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. ICML, 2005, pp. 625-632.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. ICML, 2017, pp. 1321-1330.

[23] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating

prediction models," Medical Decision Making, vol. 26, no. 6, pp. 565-574, 2006..