IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2885∼2895
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2885-2895

Pre-driving fatigue screening from short-term heart rate variability with subject-independent validation

Tia Haryanti[1], Eri Prasetyo Wibowo[1], Wahyu Kusuma Raharja[2], Rossi Septy Wahyuni[3], Ilmiyati Sari[4]
[1]Information Technology Doctoral Program, Universitas Gunadarma, Depok, Indonesia
[2]Department of Electrical Engineering, Faculty of Industrial Technology, Universitas Gunadarma, Depok, Indonesia
[3]Department of Industrial Engineering, Faculty of Industrial Technology, Universitas Gunadarma, Depok, Indonesia
[4]Department of Informatics, Faculty of Industrial Technology, Universitas Gunadarma, Depok, Indonesia

Article Info

Article history:

Received Oct 2, 2025
Revised Mar 13, 2026
Accepted Apr 20, 2026

Keywords:

30-second electrocardiogram
Leave-one-subject-out
Pre-driving fatigue screening
Probability calibration
Short-term heart rate variability
Subject-independent validation

ABSTRACT

This study evaluates fatigue screening from 30-second electrocardiogram (ECG) recordings using short-term heart rate variability (HRV) features in a pre-driving context. The dataset comprises 99 participants (one session each) with fatigue labels derived from the Karolinska sleepiness scale (KSS), where the primary label (K1) defines non-fit as KSS $\geq$ 7. A subject-independent logistic-regression model was trained under a leave-one-subject-out (LOSO) scheme. Probabilities were calibrated using Platt scaling and evaluated through threshold-free metrics (receiver operating characteristic (ROC)-area under the curve (AUC), precision-recall (PR)-AUC) as well as calibration performance using the Brier score. The model achieved ROC-AUC =0.687 (95% confidence interval: 0.591–0.776), PR-AUC =0.621, and a Brier score of 0.200. At the operating threshold $t = 0.255$, the model achieved sensitivity of 1.000 with no false negatives, while specificity remained 0.091 (95% confidence interval: 0.030–0.140). Reliability analysis indicated reasonable calibration in the operational probability range. These findings support short-term HRV derived from ECG as a screening tool that prioritizes avoiding missed non-fit cases, paired with a triage scheme (fit/review/non-fit) to manage uncertainty near the decision threshold. Future work should incorporate ECG morphology and signal quality cues and aim to improve specificity without sacrificing sensitivity.

Corresponding Author:

Eri Prasetyo Wibowo
Information Technology Doctoral Program, Universitas Gunadarma
Depok, Indonesia
Email: eri@staff.gunadarma.ac.id

1. INTRODUCTION

Road traffic crashes remain a major global public health burden, accounting for approximately 1.19 million fatalities annually according to the 2023 World Health Organization (WHO) report [1]. Fatigue and drowsiness substantially increase crash risk, with meta-analyses reporting odds ratios around 1.3 and up to 18% of fatal crashes in the United States attributed to drowsiness [2]–[6]. Recent studies indicate that wearable-derived physiological signals, particularly heart rate variability (HRV) from electrocardiogram (ECG) and photoplethysmography (PPG) sensors, are feasible for real-world fatigue screening [7]–[9]. Short sleep duration (<6–7 h) is consistently linked to increased crash risk [10], underscoring the need for rapid, low-cost,

and non-invasive pre-driving screening tools.

Physiologically, HRV reflects autonomic nervous system modulation and has been widely studied in relation to fatigue and reduced alertness [11]. In practical settings, ECG recordings are often limited to 10–30 second. Evidence from ultra-short-term HRV studies suggests that selected time-domain features, such as root mean square of successive differences (RMSSD), can approximate longer recordings under resting conditions [12]–[18]. Consequently, HRV-based screening requires rigorous RR-interval quality control, transparent reporting of signal limitations, and conservative evaluation strategies. Methodologically, prior work often overlooks strict subject-independent validation (e.g., leave-one-subject-out (LOSO)) [19] and systematic probability calibration using reliability diagrams, Brier scores, and confidence intervals [6], [20]–[23]. Recent studies further emphasize adaptive calibration strategies to enhance probabilistic reliability under dataset shift and limited-sample conditions [24]–[26].

Recent multimodal driving datasets highlight the value of integrating wearable-derived biomarkers within fatigue research [27]. Motivated by accessibility and low cost, 30-second single-lead ECG reports from smartwatches in PDF format was utilized. The pipeline converts ECG PDFs into fixed-resolution images, extracts the signal region of interest, removes textual artifacts, and calibrates time and amplitude using the printed grid, enabling ultra-short-term HRV extraction from field data. To support operational screening, explicit RR-interval quality control, probability calibration, and uncertainty reporting are incorporated. Cross-subject generalization is ensured using LOSO evaluation. This study proposes a pre-driving fatigue screening framework based on smartwatch-derived HRV, emphasizing high recall and a three-class triage scheme (fit/review/non-fit) for risk-aware decision making in safety-critical contexts.

2. RELATED WORK

Recent reviews have examined the relationship between HRV and multidimensional fatigue or driver alertness, reporting consistent trends for indices such as RMSSD alongside substantial inter-individual variability [11], [22]. Driver-centered studies have further linked HRV features from wearable ECG and PPG sensors to drowsiness and fatigue detection in controlled and real-world settings, supporting HRV as a practical marker for pre-driving screening [7]–[9]. Evidence from ultra-short-term HRV studies indicates that RMSSD and selected time-domain indices derived from 10–30 second segments can reasonably approximate standard 5-min HRV under resting conditions [12]–[18]. Comparisons between ECG and PPG highlight the scalability of wearable PPG and its susceptibility to motion artifacts and physiological biases, emphasizing the need for conservative feature selection and quality control [28]–[30]. Recent investigations further confirm the validity and reliability of ultra-short-term HRV metrics across diverse populations and recording conditions, while underscoring their sensitivity to signal quality [31].

In biosignal-based machine learning (e.g., electroencephalography (EEG), ECG, and brain-computer interfaces (BCI)), LOSO validation is widely recommended to prevent identity leakage and to obtain conservative estimates of cross-subject generalization, with prior studies reporting substantial discrepancies between within-subject and LOSO-based evaluation [18], [19], [28]. In safety-critical applications, predictive accuracy alone is insufficient, and well-calibrated probabilities are essential. Accordingly, recent studies recommend routine reporting of reliability diagrams, Brier scores, and confidence intervals, particularly for clinical screening tasks [6], [17], [20]–[23], [29], [32], [33]. Advances in adaptive and temperature-based calibration further emphasize robustness under distributional shift and limited sample sizes [24]–[26], [33]. Best practice for physiological screening models therefore includes probability calibration, reliability analysis, and uncertainty estimation via bootstrap-based confidence intervals. Reliable ultra-short-term HRV analysis additionally depends on rigorous R-peak detection and RR-interval quality control using physiological range constraints and valid RR-ratio thresholds [14]–[18].

In the sensor domain, prior studies contrast ECG and PPG for short-term HRV estimation, noting that although large-scale PPG data support population-level monitoring, they are susceptible to motion artifacts and state-dependent biases [10], [28]. For 10–30 second segments, the literature recommends emphasizing stable time-domain features and frequency ratios (e.g., RMSSD, standard deviation of normal-to-normal intervals (SDNN), and low frequency / high frequency ratio (LF/HF)) and favoring simple, well-calibrated models (e.g., logistic regression) before adopting more complex approaches [12]–[17], [22]. In screening applications, high-recall thresholding is commonly prioritized; accordingly, best practices include probability calibration,

reliability analysis, Brier score reporting, and bootstrap-based confidence intervals to ensure transparent risk assessment [6], [20]–[22]. This framework aligns with recommendations for subject-independent LOSO evaluation to prevent identity leakage and obtain conservative cross-subject performance estimates [19].

3. METHOD

This section describes the data acquisition, signal processing, feature extraction, model training, and evaluation procedures used in this study. The overall pipeline is designed to enable subject-independent fatigue screening from short-duration smartwatch ECG recordings.

3.1. Study design and cohort

This observational cross-sectional study evaluated whether HRV features derived from ECG signals can classify subjective fatigue into fit/review/non-fit, consistent with prior findings linking autonomic HRV patterns to fatigue-related states [11], [28]. The dataset comprises anonymized subjects, each contributing a single ECG recording session with a concurrent Karolinska sleepiness scale (KSS) score. All data were anonymized prior to analysis and handled in accordance with institutional ethics and privacy requirements.

3.2. Data acquisition and preprocessing

Thirty-second ECG signals from a consumer-grade smartwatch were exported as PDFs and converted into numeric waveforms (resampled to 250 Hz). R-peak detection was performed using the Pan–Tompkins algorithm with band-pass filtering (5–15 Hz), moving-window integration (150 ms), and adaptive thresholdings [13]. RR-intervals were filtered to retain physiological values (300–2,000 ms) and denoised using a five-neighbor median filter. Signal quality was assessed using RR_VALID_RATIO, with recordings meeting $\geq 75\%$ retained [12], [23], [28]. All preprocessing steps were implemented in Python using conservative default parameters suitable for 30 s recordings. Quality control further included inspection for detection jitter and verification of the absence of double-counted peaks. An example of R-peak detection and corresponding RR and beats per minute (BPM) distributions is shown in Figure 1. Figure 1(a) shows the raw ECG, Figure 1(b) shows the band-pass filtered signal (3–25 Hz), and Figure 1(c) shows the integrated signal with thresholding, applied to 30-second recordings.

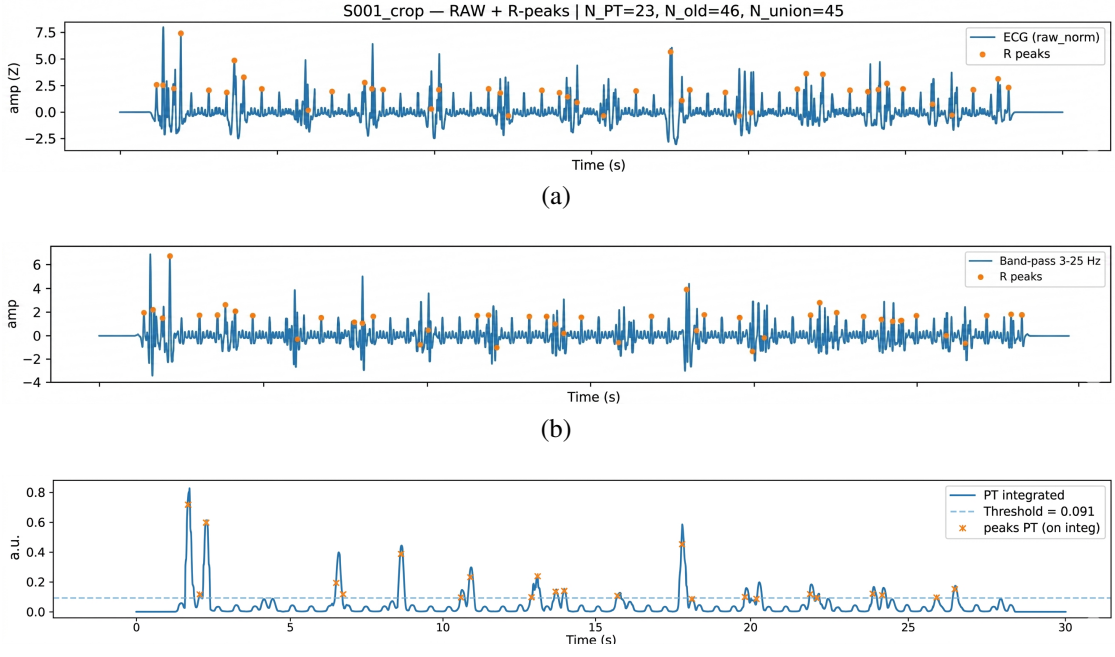


Figure 1. R-peak detection example for subject 1: (a) raw ECG, (b) band-pass filtered signal (3–25 Hz), and (c) integrated signal with thresholding, applied to 30-second recordings

3.3. Signal extraction and HRV feature set

HRV features were computed from the RR-interval series (time intervals between successive R-peaks in the ECG signal) per session.

i) Time-domain features included mean RR interval (MEAN_RR_MS), median RR interval (MED_RR_MS), standard deviation of NN intervals (SDNN_MS), root mean square of successive differences (RMSSD_MS), standard deviation of successive differences (SDSD_MS), SD1 ($SD1\_MS = RMSSD/\sqrt{2}$), SD2 ($SD2\_MS \approx \sqrt{2 \cdot SDNN^2 - 0.5 \cdot RMSSD^2}$), triangular index (TRI_IDX), pNN20, and pNN50.

ii) Frequency-domain features: the tachogram was resampled to 4 Hz, and power spectral density was estimated using Welch's method ($n_{\text{perseg}} \leq 128$). Low-frequency power (LF_PWR, 0.04–0.15 Hz), high-frequency power (HF_PWR, 0.15–0.40 Hz), total power (TOTAL_PWR, 0.04–0.40 Hz; VLF excluded), the LF/HF ratio (LF_HF), and normalized units (LF_NU and HF_NU, using a denominator of $TOTAL\_PWR - VLF$ clipped at $\geq 1 \times 10^{-6}$) were computed. $\log(LF\_HF)$ (LOG_LFHF) also computed.

iii) Non-linear and simple dynamics included sample entropy (SampEn), approximate entropy (ApEn), and linear slopes (RR_SLOPE and BPM_SLOPE) over time.

When multiple segments existed for a given (SUBJECT_ID, SESSION_ID), numeric features were aggregated using the per-session median, while labels were assigned using the maximum value (OR logic).

3.4. Ground-truth label definitions and threshold policy

The primary label (K1) defined non-fit as KSS $\geq$ 7. An additional label (K2) defined non-fit as KSS $\geq$ 8, or KSS $=$ 7 when signal quality indicators were poor (e.g., low RR valid ratio or abnormal BPM). Decision thresholds were defined under two scenarios. First, a best-F1 threshold was selected from the precision–recall curve by maximizing the F1-score. Second, a target-recall threshold (e.g., $\geq$80–90%) was selected by choosing the lowest threshold that achieves the desired recall with the highest possible precision. To manage uncertainty around the operating point, a three-class triage scheme was applied: samples were classified as non-fit if $p \geq t + \varepsilon$, fit if $p < t - \varepsilon$, and review (borderline) if $t - \varepsilon \leq p < t + \varepsilon$, with ε typically in the range of 0.002–0.005.

3.5. Modeling, validation, calibration, and evaluation

Two classifier families were employed: logistic regression and extreme gradient boosting (XGBoost). Logistic regression used *lbfgs* optimization, class balancing (class_weight=balanced), and 2,000–3,000 maximum iterations to ensure convergence on imbalanced data. XGBoost was configured as a shallow-tree ensemble (depth 3–4, 350–600 trees, learning rate 0.04–0.07, subsample 0.9–1.0, and colsample_bytree 0.8–1.0), with scale_pos_weight set near the class ratio. All features were standardized using StandardScaler or RobustScaler. Model evaluation followed a LOSO scheme, in which all sessions from one subject were held out for testing in each fold, ensuring strictly subject-independent assessment. Out-of-fould (OOF) predictions were aggregated across folds to enable unbiased performance estimation. Probability calibration was performed using platt scaling (CalibratedClassifierCV with cv="prefit"), as recommended for small physiological datasets and high-recall screening tasks [24], [26], [33]. Calibration quality was evaluated using reliability curves (10 bins) and the Brier score. Performance was assessed using threshold-free metrics (receiver operating characteristic (ROC)-area under the curve (AUC), precision-recall (PR)-AUC), and decision-level metrics (recall, precision, F1-score, specificity, negative predictive value (NPV), and confusion matrix) at the final operating point. Uncertainty was quantified using subject-level bootstrap resampling ($B = 2{,}000$) to compute 95% confidence intervals [30], [31]. Per-subject diagnostics were further analyzed to assess cross-subject performance stability.

3.6. Uncertainty management and operational procedures

Beyond reporting confidence intervals and calibration curves, operational uncertainty was addressed using a review zone ($t - \varepsilon$ to $t + \varepsilon$). Cases within this range were routed to rapid verification (e.g., signal and artifact checks, KSS reconfirmation, and feature inspection), enabling false positives to be filtered without compromising safety. The threshold t and margin ε can be retuned based on operational requirements and field trial feedback.

3.7. Implementation and reproducibility

All analyses were conducted in Python using Google Colab, primarily relying on scikit-learn and XGBoost. The implementation comprised modular scripts for ECG PDF rendering and signal extraction, RR-interval detection and quality control, HRV feature computation, subject-independent LOSO training and evaluation, probability calibration, and post-hoc analyses including bootstrap confidence intervals and per-subject diagnostics. These practices align with recent recommendations for reproducible and calibrated biomedical machine learning in small-sample settings [18], [19], [24]. All intermediate artifacts were stored in a structured directory, with fixed package versions and a global random seed (e.g., 42) to ensure reproducibility. Figure 2 illustrates the end-to-end pipeline of the proposed system. Starting from 30-second single-lead smartwatch ECG reports in PDF format, the pipeline performs signal rendering, R-peak detection, RR-interval cleaning, ultra-short-term HRV feature extraction, and session-level aggregation. A subject-independent LOSO split is then applied for model training and evaluation, followed by probability calibration. Finally, a triage-oriented decision layer assigns each session to fit, review, or non-fit categories and outputs structured prediction summaries.

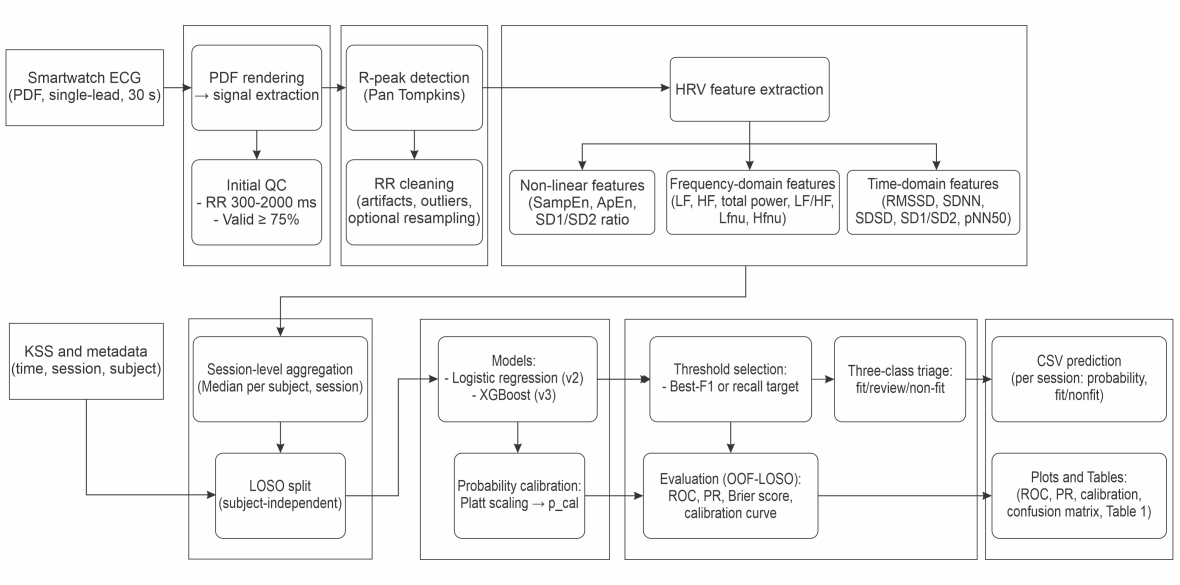


Figure 2. Flowchart of the proposed pre-driving fatigue screening method based on HRV analysis

4. RESULTS AND DISCUSSION

This section reports subject-independent experimental results and their implications for safety-oriented pre-driving fatigue screening. The evaluation is conducted under a LOSO scheme to ensure cross-subject generalization and avoid identity leakage. We present both threshold-free performance and operating-point analysis to support risk-aware decision-making.

4.1. Data characteristics

A total of 99 participants with one 30-second ECG recording per participant were analyzed. Subjective sleepiness was assessed concurrently using the KSS. Only recordings meeting predefined quality criteria were retained, including RR-intervals within the physiological range of 300–2,000 ms and a valid RR proportion of at least 75%, to ensure reliable HRV estimation. HRV features summarized substantial inter-individual variability while remaining within physiological ranges shown in Table 1.

Table 1 summarizes the cohort characteristics and HRV feature distributions. The proportion of Non-fit labels was 33.3% under the primary definition K1 (KSS $\geq$ 7) and 25.3% under the stricter definition K2 (KSS $\geq$ 8 or KSS $=$ 7 with reduced signal quality), indicating that the positive class constitutes a minority. HRV features are reported as median [IQR], including time-domain metrics (SDNN, RMSSD,

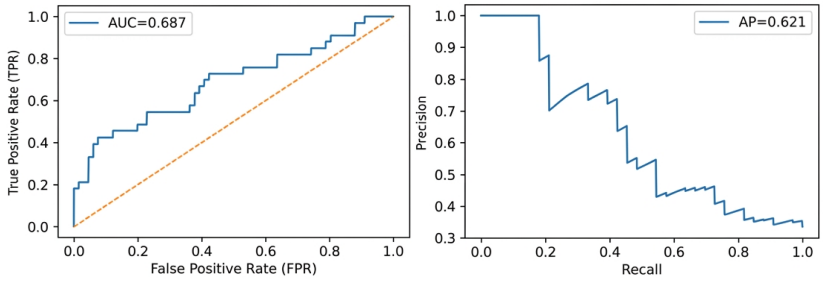
SDSD, SD1/SD2, and triangular index), frequency-domain components (LF, HF, total power, LF/HF ratio, and LFnu/HFnu), and signal quality indicators (number of RR-intervals and valid RR ratio). Overall, the HRV features fall within physiological ranges and exhibit substantial inter-individual variability, providing a representative baseline for subsequent model development and evaluation.

Table 1. Cohort characteristics and HRV feature summary

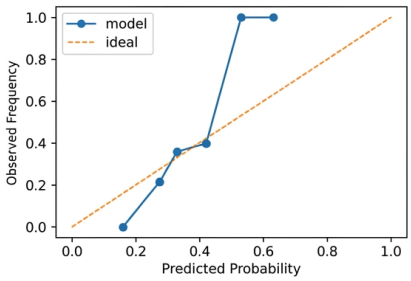
Overall cohort			
Characteristic	Value	Characteristic	Value
Participants, n	99	SDNN, ms [IQR]	91.7 [68.9–149.6]
Sessions, n	1	RMSSD, ms [IQR]	122.0 [88.2–226.3]
Sessions/participant [IQR]	1 [1–1]	SDSD, ms [IQR]	123.3 [89.0–228.8]
Non-fit (K1), %	33.3%	SD1, ms [IQR]	86.3 [62.4–160.1]
Non-fit (K2), %	25.3%	SD2, ms [IQR]	89.2 [74.7–135.3]
Heart rate, bpm [IQR]	110.7 [110.3–111.1]	Triangular index [IQR]	4.5 [3.5–8.0]
Mean RR, ms [IQR]	533.0 [518.0–545.7]	LF power, ms^2 [IQR]	179.0 [32.6–2095.3]
		HF power, ms^2 [IQR]	1056.6 [282.0–6321.3]
		Total power, ms^2 [IQR]	1305.8 [335.9–1083.7]
		LF/HF ratio [IQR]	0.2 [0.1–0.4]
		$\log(1 + \mathrm{LF}/\mathrm{HF})$ [IQR]	0.2 [0.1–0.3]
		$\mathrm{LF_{nu}}$ [IQR]	0.1 [0.1–0.2]
		$\mathrm{HF_{nu}}$ [IQR]	0.8 [0.6–0.9]
		RR count, n [IQR]	48.0 [45.5–50.0]

4.2. Threshold-free model performance under OOF-LOSO evaluation

We evaluated threshold-free performance using OOF predictions under the LOSO scheme, with logistic regression (v2) as the primary model for the K1 label Non-fit = KSS $\geq$ 7). The evaluation focuses on discrimination (ROC-AUC), minority-class precision (PR-AUC), and probability calibration (Brier score and reliability curves). The OOF-LOSO evaluation results of the primary K1 model are summarized in Figure 3. Figure 3(a) presents the ROC curve, indicating moderate discrimination performance with an AUC of 0.687. Figure 3(b) shows the precision–recall curve with an average precision of 0.621, reflecting the trade-off between recall and precision under class imbalance. Figure 3(c) illustrates the reliability diagram with a Brier score of approximately 0.200, demonstrating adequate probability calibration in the operational range.



(a) (b)



(c)

Figure 3. OOF-LOSO performance evaluation of the primary K1 model: (a) ROC curve, (b) precision–recall curve, and (c) reliability diagram with Brier score

Subject-level bootstrap analysis yielded a 95% confidence interval of 0.591–0.776 for ROC-AUC, indicating moderate cross-subject variability. The model shows stable discrimination and reasonable precision, with calibration close to the ideal diagonal and mild over-confidence at high probabilities. The stricter K2 labeling exhibits similar trends but lower stability; therefore, subsequent analyses focus on the K1 scheme for safety-oriented screening.

4.3. Operating point performance and robustness analysis

A final threshold of $t = 0.255$ was selected from the precision–recall curve to minimize false negatives and prioritize non-fit detection. At this operating point, the confusion matrix is reported in Figure 4. Decision-level metrics are reported in Table 2.

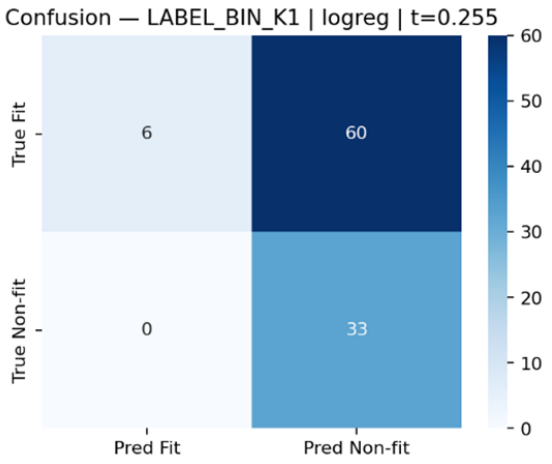


Figure 4. Confusion matrix of the primary K1 model at the selected operating threshold ($t = 0.255$)

Table 2. Final operating point for the K1 label ($t = 0.255$)

Threshold	AUC	PR_AUC	Brier	TP	FP	FN	TN	Recall	Precision	F1
0.255	0.6873	0.6210	0.2001	33	60	0	6	1.000	0.3548	0.5238

Practically, this decision prioritizes high sensitivity for detecting non-fit cases, consistent with safety-oriented screening, while acknowledging that some fit cases may be flagged as false positives and require follow-up review. Good probability calibration, reflected by the Brier score in Table 2, supports stable threshold selection by aligning predicted probabilities with observed event frequencies. In addition, a three-class triage scheme around the operating threshold ($t \pm \varepsilon$) designates borderline cases as review, routing them to manual verification without reducing vigilance for clearly non-fit cases.

4.4. Subject-level robustness and error analysis

Model robustness to sampling variability and threshold perturbations was evaluated using subject-level bootstrap analysis under the LOSO scheme. At the final operating point ($t = 0.255$), bootstrap resampling ($B = 2{,}000$) yielded a ROC-AUC of 0.687 (95% confidence interval: 0.591–0.776), sensitivity of 0.970, specificity of 0.091, positive predictive value (PPV) of 0.348, and NPV of 0.857, indicating consistent decision-level behavior. Probability calibration was acceptable, with a Brier score of approximately 0.200, confirming preserved high recall for non-fit cases. A narrow threshold sweep ($t \pm \varepsilon$) showed that slight decreases in the threshold maintained FN $= 0$ without disproportionate FP increases, whereas slight increases introduced FN with limited FP reduction, indicating a stable operating region for safety-oriented screening. Subject-level error analysis revealed that misclassifications were dominated by sparse false positives associated with probabilities near the decision boundary, reflecting local uncertainty

rather than systematic model failure. Figure 5 illustrates the distribution of calibrated predicted probabilities (p_{cal}) for fit and non-fit classes.

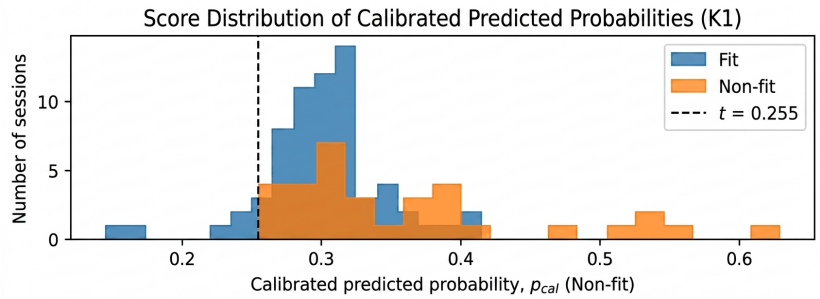


Figure 5. Score distribution of calibrated probabilities (p_{cal}) for fit and non-fit classes (K1)

To manage predictive uncertainty, we implemented a three-class triage scheme (fit/review/non-fit) using a narrow threshold band ($t \pm \varepsilon$, $\varepsilon \approx 0.003$). Borderline cases were assigned to the review category for secondary checks, while clearly non-fit cases remained prioritized. Feature-level analysis indicated substantial overlap between fit and non-fit groups, suggesting that some false positives arise from intrinsic physiological variability rather than systematic misclassification. Notably, the absence of false negatives at the final operating point confirms that the primary safety objective was achieved. The remaining false-positive burden reflects an operational trade-off that can be managed through lightweight follow-up workflows. Performance at the final threshold is summarized in Figure 4 and Table 2.

4.5. Discussion

This study demonstrates that HRV features derived from short-duration, single-lead smartwatch ECG recordings can support subject-independent fatigue screening under a safety-oriented objective. The emphasis on high recall for non-fit cases under LOSO evaluation is consistent with early screening scenarios in which minimizing missed fatigue cases is prioritized. The concentration of false positives near the decision threshold suggests local predictive uncertainty rather than systematic model failure, supporting the use of calibrated probabilities for operational decision making. The proposed three-class triage scheme (fit/review/non-fit) further manages this uncertainty by routing borderline cases to secondary verification, balancing safety and workload. The observed moderate discrimination likely reflects the limitations of ultra-short ECG segments and self-reported fatigue labels. Compared with studies relying on longer or laboratory-grade and multimodal recordings, these results highlight the potential of HRV features extracted from smartwatch ECG PDFs for early screening under realistic deployment constraints. Accordingly, the proposed system is intended as a screening aid and should not be used as a standalone diagnostic tool.

5. CONCLUSION

This study investigated the feasibility of subject-independent pre-driving fatigue screening using HRV features extracted from short-duration, single-lead smartwatch ECG recordings stored as PDF files. The results demonstrate that meaningful fatigue screening can be achieved using ultra-short ECG segments from consumer-grade devices when combined with conservative validation, probability calibration, and a safety-oriented operating strategy. By prioritizing high recall for Non-fit cases and introducing a three-class triage scheme (fit/review/non-fit), the proposed approach supports practical decision making while explicitly managing uncertainty around the decision threshold. Despite limitations related to cohort size, self-reported fatigue labels, PDF-derived ECG signals, and the absence of external validation, this work indicates that short-term HRV from wearable devices can serve as a useful decision-support tool for early fatigue screening rather than a replacement for clinical judgment. Future work will focus on expanding the cohort, incorporating additional physiological or contextual features, and validating the approach across devices, populations, and operational settings.

ACKNOWLEDGMENTS

The authors would like to thank the Doctoral Program in Information Technology, Gunadarma University, for providing laboratory facilities and technical support, including the Microelectronics and Image Processing Laboratories. The authors also sincerely thank Prof. Michel Paindavoine from Université de Bourgogne Europe for his valuable suggestions and discussions.

FUNDING INFORMATION

This research was funded by the Ministry of Education, Culture, Research, and Technology of Indonesia under the Doctoral Dissertation Research, Contact Number: 0419/C3/DT.05.00/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Tia Haryanti	✓	✓	✓		✓	✓		✓	✓		✓			
Eri Prasetyo Wibowo	✓	✓					✓			✓		✓	✓	✓
Wahyu Kusuma Raharja		✓		✓		✓				✓				
Rossi Septy Wahyuni				✓		✓				✓				
Ilmiyati Sari			✓		✓					✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The data that support the findings of this study were used with permission and are available from the corresponding author, [EPW], upon reasonable request. The data are not publicly available due to privacy and ethical restrictions related to human-subject research.

REFERENCES

[1] WHO, "Global status report on road safety 2023," *World Health Organization*, 2023. Accessed: Dec. 27, 2025, [Online]. Available: https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

[2] S. Bioulac *et al*., "Risk of motor vehicle accidents related to sleepiness at the wheel: a systematic review and meta-analysis," *Sleep*, vol. 40, no. 10, Oct. 2017, doi: 10.1093/sleep/zsx134.

[3] B. C. Tefft, "Drowsy driving in fatal crashes, United States, 2017–2021," *Research Brief*, Washington, D.C.: AAA Foundation for Traffic Safety, 2024.

[4] D. J. Gottlieb, J. M. Ellenbogen, M. T. Bianchi, and C. A. Czeisler, "Sleep deficiency and motor vehicle crash risk in the general population: a prospective cohort study," *BMC Medicine*, vol. 16, no. 1, Dec. 2018, doi: 10.1186/s12916-018-1025-7.

[5] M. Sprajcer *et al*., "How tired is too tired to drive? A systematic review assessing the use of prior sleep duration to detect driving impairment," *Nature and Science of Sleep*, vol. 15, pp. 175–206, Apr. 2023, doi: 10.2147/NSS.S392441.

[6] T. S. Filho, H. Song, M. P.-Nieto, R. S.-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Machine Learning*, vol. 112, no. 9, pp. 3211–3260, Sep. 2023, doi: 10.1007/s10994-023-06336-7.
[7] Z. AlArnaout, C. Zaki, Y. Kotb, M. AlAkkoumi, and N. Mostafa, "Exploiting heart rate variability for driver drowsiness detection using wearable sensors and machine learning," *Scientific Reports*, vol. 15, no. 1, Jul. 2025, doi: 10.1038/s41598-025-08582-2.
[8] H. Liu, Y. Zhou, C. Jiang, and C. Zhang, "Investigating heart rate variability of metro drivers before an inappropriate stop at platform with wearable devices," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2680, no. 4, pp. 192–209, Apr. 2026, doi: 10.1177/03611981251372089.
[9] G. Perrotte, J.-L. Vercher, and C. Bougard, "Postural and physiological indicators of drowsiness at the wheel compared in partially and conditionally autonomous on-road driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 116, Jan. 2026, doi: 10.1016/j.trf.2025.103444.
[10] A. Natarajan, A. Pantelopoulos, H. E.-Farinas, and P. Natarajan, "Heart rate variability with photoplethysmography in 8 million individuals: a cross-sectional study," *The Lancet Digital Health*, vol. 2, no. 12, pp. e650–e657, Dec. 2020, doi: 10.1016/S2589-7500(20)30246-6.
[11] A. G. Srinivasan *et al*., "Heart rate variability as an indicator of fatigue: A structural equation model approach," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 103, pp. 420–429, May 2024, doi: 10.1016/j.trf.2024.04.015.
[12] E. Krause *et al*., "Evaluating heart rate variability with 10 second multichannel electrocardiograms in a large population-based sample," *Frontiers in Cardiovascular Medicine*, vol. 10, May 2023, doi: 10.3389/fcvm.2023.1144191.
[13] F. Shaffer, Z. M. Meehan, and C. L. Zerr, "A critical review of ultra-short-term heart rate variability norms research," *Frontiers in Neuroscience*, vol. 14, Nov. 2020, doi: 10.3389/fnins.2020.594880.
[14] L. Wu, P. Shi, H. Yu, and Y. Liu, "An optimization study of the ultra-short period for HRV analysis at rest and post-exercise," *Journal of Electrocardiology*, vol. 63, pp. 57–63, Nov. 2020, doi: 10.1016/j.jelectrocard.2020.10.002.
[15] J. W. Kim, H. S. Seok, and H. Shin, "Is ultra-short-term heart rate variability valid in non-static conditions?," *Frontiers in Physiology*, vol. 12, Mar. 2021, doi: 10.3389/fphys.2021.596060.
[16] F. Landreani *et al*., "Assessment of ultra-short heart variability indices derived by smartphone accelerometers for stress detection," *Sensors*, vol. 19, no. 17, Aug. 2019, doi: 10.3390/s19173729.
[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, Jul. 2017, pp. 1321–1330.
[18] C.-H. Hung, W.-A. Lu, J. C. Pagaduan, C.-D. Kuo, and Y.-S. Chen, "Agreement of ultra-short-term heart rate variability measure after different repeated bouts of sprint ability tests," *Science Progress*, vol. 107, no. 3, Jul. 2024, doi: 10.1177/00368504241262150.
[19] S. Kunjan *et al*., "The necessity of leave one subject out (LOSO) cross validation for EEG disease diagnosis," in *Brain Informatics*, Cham, Switzerland: Springer, 2021, pp. 558–567, doi: 10.1007/978-3-030-86993-9_50.
[20] T. Dimitriadis, T. Gneiting, and A. I. Jordan, "Stable reliability diagrams for probabilistic classifiers," *Proceedings of the National Academy of Sciences*, vol. 118, no. 8, Feb. 2021, doi: 10.1073/pnas.2016191118.
[21] P. C. Austin, D. S. Lee, and B. Wang, "The relative data hungriness of unpenalized and penalized logistic regression and ensemble-based machine learning methods: the case of calibration," *Diagnostic and Prognostic Research*, vol. 8, no. 1, Nov. 2024, doi: 10.1186/s41512-024-00179-z.
[22] R. D. Riley *et al*., "Evaluation of clinical prediction models (part 2): how to undertake an external validation study," *BMJ*, vol. 384, Jan. 2024, doi: 10.1136/bmj-2023-074820.
[23] R. Maqsood *et al*., "Validity of ultra-short-term heart rate variability derived from femoral arterial pulse waveform in a British military cohort," *Applied Psychophysiology and Biofeedback*, vol. 49, no. 4, pp. 619–627, Dec. 2024, doi: 10.1007/s10484-024-09652-3.
[24] S. A. Balanya, J. Maroñas, and D. Ramos, "Adaptive temperature scaling for Robust calibration of deep neural networks," *Neural Computing and Applications*, vol. 36, no. 14, pp. 8073–8095, May 2024, doi: 10.1007/s00521-024-09505-4.
[25] T. Hasegawa, K. Fujino, and S. Sakai, "Analytical softmax temperature setting from feature dimensions for model- and domain-robust classification," *Neural Computing and Applications*, vol. 37, no. 33, pp. 27985–28016, Nov. 2025, doi: 10.1007/s00521-025-11488-9.
[26] W. Huang, G. Cao, J. Xia, J. Chen, H. Wang, and J. Zhang, "H-calibration: rethinking classifier recalibration with probabilistic error-bounded objective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 10, pp. 9023–9042, Oct. 2025, doi: 10.1109/TPAMI.2025.3582796.
[27] X. Tao *et al*., "A multimodal physiological dataset for driving behaviour analysis," *Scientific Data*, vol. 11, no. 1, Apr. 2024, doi: 10.1038/s41597-024-03222-2.
[28] Y. Tanoue *et al*., "The validity of ultra-short-term heart rate variability during cycling exercise," *Sensors*, vol. 23, no. 6, Mar. 2023, doi: 10.3390/s23063325.
[29] Y. Yu, S. Bates, Y. Ma, and M. I. Jordan, "Robust calibration with multi-domain temperature scaling," *36th Conference on Neural Information Processing Systems*, 2022, pp. 27510-27523, doi: 10.5555/3600270.3602265.
[30] T. M. Chieng, Y. W. Hau, Z. Omar, C. W. Lim, C.-M. Ting, and S. Mandala, "Validity and reliability of the ultra-short-term heart rate variability features in predicting ventricular tachyarrhythmia," *Biomedical Signal Processing and Control*, vol. 110, Dec. 2025, doi: 10.1016/j.bspc.2025.108173.
[31] C. Besson, A. L. Baggish, P. Monteventi, L. Schmitt, F. Stucky, and V. Gremeaux, "Assessing the clinical reliability of short-term heart rate variability: insights from controlled dual-environment and dual-position measurements," *Scientific Reports*, vol. 15, no. 1, Feb. 2025, doi: 10.1038/s41598-025-89892-3.
[32] C. Tomani, D. Cremers, and F. Buettner, "Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration," in *Computer Vision – ECCV 2022: 17th European Conference*, Cham, Switzerland: Springer, 2022, pp. 555–569, doi: 10.1007/978-3-031-19778-9_32.
[33] Z.-Q. Cheng *et al*., "Towards calibrated robust fine-tuning of vision-language models," in *Advances in Neural Information Processing Systems 37*, 2024, pp. 12677–12707, doi: 10.52202/079017-0403.

BIOGRAPHIES OF AUTHORS

Tia Haryanti is a doctoral student in the Department of Information Technology, Universitas Gunadarma, Depok, Indonesia. She received the M.T. in Industrial Engineering from Universitas Gunadarma. Her research focuses on pre-driving fatigue assessment and early screening using physiological signals. In this study, she led the methodology, ECG signal processing pipeline, modeling, and manuscript preparation. She has presented parts of this work and contributes to projects supported by the PDD–BIMA grant (2025). She can be contacted at email: tiaharyanti@staff.gunadarma.ac.id.

Eri Prasetyo Wibowo is a prominent professor at Universitas Gunadarma, Indonesia, with a distinguished academic and research career. He completed his B.S. degree in Electronics and Instrumentation at Universitas Gadjah Mada in 1991, followed by an M.S. degree in Information Systems at Universitas Gunadarma in 1994. He earned his Ph.D. in Electronics Informatics from Universit´e de Bourgogne, France, in 2005. He is an active member of the IEEE and contributes to the EACOVIROE project, promoting the Erasmus mundus-funded master VIBOT program. His primary research interests include CMOS sensor de sign for face tracking and recognition, development of ADCs for video applications, VLSI design, and real-time image processing. Through his work, he has made substantial contributions to advancing electronics and informatics, particularly in sensor technology and image processing systems. He is member of IEEE. He can be contacted at email: eri@staff.gunadarma.ac.id.

Wahyu Kusuma Raharja is with the Department of Electrical Engineering, Universitas Gunadarma, Depok, Indonesia. His research spans embedded systems, physiological signal acquisition, and data management, including work on heartbeat and arrhythmia monitoring devices, internet of things (IoT) applications for telemonitoring, and biomedical signal processing. In this study, he contributed to the design of signal acquisition and quality-control procedures, data curation, and validation of ECG extraction and preprocessing steps. He also provided technical review and editing to ensure rigor in signal processing and feature extraction. He can be contacted at email: wahyukr@staff.gunadarma.ac.id.

Rossi Septy Wahyuni is with the Department of Industrial Engineering, Universitas Gunadarma, Depok, Indonesia. Her research interests include educational technology, interactive learning media, and the integration of technology in instructional design. In this study, she contributed to contextualization of the research problem, integration of human-factors considerations in the screening framework, and interpretation of usability and practical deployment aspects. She also reviewed and edited the manuscript for clarity and educational relevance. She can be contacted at email: rossysw@staff.gunadarma.ac.id.

Ilmiyati Sari received her B.S. degree in Mathematics from University of Indonesia in 2009, her M.S. in Mathematics from University of Indonesia in 2012 and her doctoral in Universitas Gunadarma in 2018. She has published extensively in the area of video processing. Her research interests include dynamic models, statistics, image processing, video processing, and machine learning. She can be contacted at email: ilmiyati@staff.gunadarma.ac.id.