JESK J Ergon Soc Korea 2026; 45(4): 361-375
https://dx.doi.org/10.5143/JESK.2026.45.4.361
https://jesk.or.kr eISSN:2093-8462

Separating Alert Onset Rate from Alert Duration in Wearable Stress Monitoring: a Behavior-level Alert Policy Based on Signal Quality and Context Validity

Seunghee Hong[1], Hang Joo Choi[1], Byungchan Min[2]

[1]Ebenhaim, Cheonan-si, 31156
[2]Department of Industrial and Management Engineering, Hanbat National University, Daejeon, 34158

웨어러블 스트레스 모니터링에서 알림 발생률과 지속시간의 분리 평가: 신호품질·맥락타당성 기반 행동 레벨 알림 정책

홍승희[1], 최항주[1], 민병찬[2]

[1]에벤하임
[2]국립한밭대학교 산업경영공학과

Corresponding Author
Byungchan Min
Department of Industrial and Management Engineering, Hanbat National University, Daejeon, 34158
Email : bcmin@hanbat.ac.kr

Received : June 02, 2026
Revised : June 29, 2026
Accepted : July 22, 2026

Copyright@2026 by The Human Factors & Ergonomics Society of Korea. All right reserved.

Objective: This study introduces an alert-level evaluation that separates alert onset rate from alert active duration in wearable stress monitoring and uses it to assess a behavior-level alert policy based on the signal quality index (SQI) and context validity index (CVI) for reducing false alerts while preserving stress recall.

Background: Wearable physiological classifiers based on photoplethysmography, electrodermal activity, and motion signals are vulnerable to signal artifacts and context mismatch. Exercise-related physiological arousal can resemble acute stress and increase alert burden when behavioral context is ignored.

Method: We performed a secondary analysis of the publicly available Hongn dataset and reproduced three XGBoost-based classifiers for stress/rest, aerobic/sprints, and four-class context classification. The classifier outputs were post-processed using a behavior-level alert-policy framework. SQI was implemented as confidence weighting for stress evidence, whereas CVI was implemented as behavior-level context veto and rescue rules. Alert-policy parameters were specified before final evaluation as monotonic operating rules for Sensitive, Balanced, and Strict modes.

Results: Four-class accuracy closely matched the reported benchmark (0.839 vs. 0.840), whereas the stress/rest classifier remained below the reported value (0.834 vs. 0.930). In the revised alert-level analysis, the classifier-only baseline produced 5.288 false alerts/h and 8.110 exercise-context false alerts/h. Under the combined SQI+CVI alert policy, false alerts/h decreased to 1.609, 1.265, and 1.035 in the Sensitive, Balanced, and Strict modes, respectively. Exercise-context false alerts/h decreased to 0.640, 0.427, and 0.000. Recall increased from 0.325 to 0.727, 0.675, and 0.565. Ablation analysis showed that SQI-only produced the highest F1 in the Balanced mode, whereas the combined SQI+CVI alert policy provided a stronger balance between false-alert reduction, exercise-context suppression, and recall preservation.

Conclusion: Because alert onset rate and alert active duration can move ndependently,

classification accuracy alone can rank alert policies differently from alert-burden metrics; the proposed alert policy should therefore be interpreted as a balanced alert-policy design rather than a model that maximizes every metric.

Application: The separation of alert onset rate from alert active duration can serve as an operational, alert-burden-aware evaluation lens for wearable stress monitoring, and the proposed alert policy may inform future context-adaptive alerting strategies, although independent validation is required before practical application claims can be made.

Keywords: Wearable monitoring, Signal Quality Index, Context Validity Index, Behavior-level alert policy, False alert

1. Introduction

웨어러블 센서 기반 스트레스 모니터링은 연속적인 생체신호를 활용하여 사용자의 상태 변화를 추적할 수 있다는 장점이 있다. 그러나 생체신호 분류 결과를 실제 사용자 알림으로 전환할 때는 분류 정확도만으로 시스템의 유용성을 충분히 평가하기 어렵다. 반복적이거나 맥락적으로 부적절한 알림은 사용자의 주의를 분산시키고 알림 피로를 유발할 수 있으며, 임상 의사결정지원 및 생리 모니터링 연구에서도 반복 경보의 운영 부담이 지속적으로 보고되어 왔다(Ancker et al., 2017; Drew et al., 2014). 알림 부담은 단위시간당 새롭게 발생한 알림 사건의 수와 알림 상태가 활성화된 총 시간으로 구분할 수 있다. 전자는 시간당 알림 발생률, 후자는 시간당 알림 활성시간(alert minutes/h)으로 나타내며, 두 지표는 서로 독립적으로 변할 수 있다. 따라서 실제 알림 시스템은 분류 성능과 함께 두 형태의 알림 부담을 분리하여 평가할 필요가 있다.

기존 웨어러블 스트레스 연구는 주로 생체신호 분류기의 정확도와 F1-score 향상에 초점을 두었다. Hongn et al. (2025)은 유도 스트레스, 휴식, 유산소 운동 및 스프린트 조건에서 수집한 공개 생체신호 데이터셋과 XGBoost 기반 기술 검증 결과를 제시하였다. 그러나 운동 중 증가하는 심박수와 피부전도도는 급성 스트레스 반응과 생리적으로 유사하므로, 신체활동 맥락을 고려하지 않으면 운동 구간이 스트레스로 오분류될 수 있다(Hovsepian et al., 2015; Can et al., 2019; Schmidt et al., 2018). Gjoreski et al. (2017)은 손목 착용형 기기의 스트레스 인식에 맥락 정보를 결합하여 일상 환경의 거짓 양성을 줄이고 분류 성능을 개선할 수 있음을 보였다.

분류기 출력을 실제 알림 의사결정으로 변환하기 위해 임계값 조정과 시간적 후처리가 활용되어 왔다. Sarker et al. (2016)은 급변하는 스트레스 확률 시계열에서 유의한 스트레스 episode를 식별하기 위해 유의도 임계값과 시계열 패턴을 적용하였고, Can et al. (2020)은 고립된 오분류를 줄이기 위해 decision-level smoothing을 적용하였다. 신호품질 측면에서는 PPG와 EDA의 사용 가능성을 평가하여 품질이 낮은 구간을 제외하거나(Orphanidou et al., 2015; Kleckner et al., 2018), 실생활 자료에 sensor-specific artifact detection threshold를 적용하는 방법이 제안되었다(Can et al., 2019). 또한 신체활동으로 오염된 구간을 제외하거나 행동 맥락을 결합하는 activity-aware/context-aware 접근이 활용되었다(Hovsepian et al., 2015; Gjoreski et al., 2017). 이러한 방법들은 threshold 조정, temporal smoothing, signal-quality filtering 및 맥락 기반 억제가 분석 구간 또는 episode 수준의 오분류를 줄일 수 있음을 보여준다.

그러나 기존 접근의 평가는 대부분 분석 구간 또는 episode 수준의 accuracy, F1-score, false-positive rate에 머물렀으며, 분류기 출력이 사용자에게 전달되는 알림으로 변환된 뒤 새 알림이 얼마나 자주 발생하는지와 알림 상태가 얼마나 오래 유지되는지를 별도의 운영 지표로 다루지는 않았다. 한편 임상 의사결정지원 및 생리 모니터링 분야의 알림 피로 연구는 반복적이고 비관련적인 경보가 사용자 부담을 높인다는 점을 보여주었지만(Ancker et al., 2017; Drew et al., 2014), 웨어러블 스트레스 모니터링에서 운동 맥락 보정과 신호품질 제어를 결합하여 알림 발생률과 알림 활성시간을 동시에 평가한 연구는 제한적이다. 특히 동일한 분류 결과라도 신규 알림 사건 수는 감소하면서 알림 활성시간은 증가할 수 있으므로, 두 지표를 분리하지 않으면 알림 부담의 상충 관계를 파악하기 어렵다.

이러한 연구 공백을 바탕으로 다음 두 가지 목적을 설정하였다. 첫째, 분류 정확도와 별도로 시간당 알림 발생률과 시간당 알림 활성시간을 분리하여 평가하는 운영 수준의 관점을 제시한다. 둘째, Hongn et al. (2025)의 XGBoost 기반 기준 분류기를 재현한 뒤, 신호품질 지수(Signal Quality Index, SQI)와 맥락타당성 지수(Context Validity Index, CVI)를 행동 레벨별 알림 정책으로 결합하여 거짓 알림과

운동 맥락 오탐의 감소 효과를 평가한다. 여기서 알림 정책(alert policy)은 제도적 의미의 정책이 아니라, 분류기 출력에 SQI 가중과 CVI veto/rescue를 적용하여 최종 알림 상태를 결정하는 알고리즘적 후처리 규칙의 집합을 뜻한다. 이러한 접근의 기여는 새로운 스트레스 분류기의 최고 성능 제시가 아니라, 알림 부담을 분리·정량화하는 평가 관점과 그 위에서 작동하는 행동 레벨 알림 정책을 제안하는 데 있다.

2. Method

2.1 Dataset

본 연구는 PhysioNet에 공개된 Hongn et al. (2025)의 Scientific Data 데이터셋을 사용하였다. 이 데이터셋은 손목 착용형 센서에서 수집한 blood volume pulse (BVP)/PPG, EDA, accelerometer (ACC), 심박수, 피부온도 신호를 포함하며, stress, rest, aerobic, sprints 조건으로 구성된다.

초기 분석 구간 후보는 496개였으나, split-record grouping, 선택된 분석 구간 구성, 세 분류기 출력 정렬 기준을 적용한 뒤 최종 행동 레벨 알림 정책 분석에는 482개 분석 구간을 사용하였다. 과제별 표본 수는 stress/rest 256개, aerobic/sprints 226개, four-class 482개였다.

2.2 Benchmark classifier reproduction

기준 분류기는 Hongn et al. (2025)의 기술 검증 절차를 따라 XGBoost (Chen and Guestrin, 2016) 중심으로 재현하였다. 재현 대상 과제는 stress/rest, aerobic/sprints, four-class classification이었다. 원 논문에서 분석 구간 선택, feature subset, resampling 방식, XGBoost 세부 하이퍼파라미터, split seed가 완전히 공개되지 않았으므로, 본 연구에서는 해당 요소를 후보군으로 두고 조합 탐색하였다.

모든 후보 설정은 동일한 80/20 train-test split 및 10-fold cross-validation 구조에서 평가하였고, resampling은 training fold 내부에서만 적용하였다. Subject-level leakage를 방지하기 위해 동일 피험자의 자료가 학습과 검증에 동시에 포함되지 않도록 분할 단위를 관리하였다(Kaufman et al., 2012; Varoquaux, 2018).

각 과제의 후보 설정은 교차 검증 성능과 재현 절차의 일관성을 기준으로 비교한 뒤, 선택된 설정을 held-out test set에 적용하였다. 재현 정확도, F1-score 및 Hongn et al. (2025)이 보고한 성능값과의 차이는 Results 3.1에 제시하였다.

알림 정책 분석을 위해 세 분류기의 확률 출력을 동일 subject/session의 분석 구간 단위로 정렬하였다. Four-class 출력은 stress, rest,

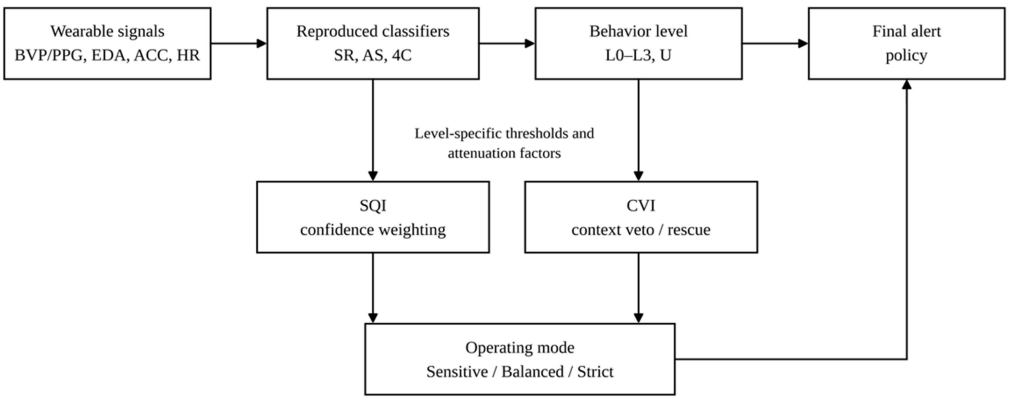


Figure 1. Behavior-level SQI/CVI alert-policy framework

aerobic 및 sprints 맥락 확률을 제공하고, stress/rest 출력과 aerobic/sprints 출력은 각각 통합 스트레스 근거와 운동 근거의 구성요소로 사용하였다. 각 출력의 재현 성능과 해석상 제한은 Results 3.1과 Discussion에서 기술하였다.

2.3 Behavior-level SQI/CVI alert-policy framework

본 연구의 SQI, CVI 및 행동 레벨 알림 정책 함수는 단일 선행연구에서 그대로 차용한 표준 공식이 아니라, 분류기 출력 이후 최종 알림 상태를 결정하기 위해 정의한 알고리즘적 후처리 함수이다. SQI 구성요소는 PPG 파형의 타당성·박동 규칙성과 움직임 아티팩트 평가(Allen, 2007; Tamura et al., 2014; Orphanidou et al., 2015), EDA 아티팩트 평가(Kleckner et al., 2018; Taylor et al., 2015), 실생활 스트레스 모니터링의 sensor-specific artifact handling (Can et al., 2019)에 근거하였다. CVI와 행동 레벨 규칙은 신체활동에 의한 생리적 혼동 및 맥락 기반 스트레스 분류 문헌(Hovsepian et al., 2015; Gjoreski et al., 2017)과 행동강도 분류 기준(Ainsworth et al., 2011; Haskell et al., 2007; Pate et al., 1995; Tremblay et al., 2017)을 참고하였다. 다만 실제 MET 값을 측정하지 않았으므로, 행동 레벨은 직접 측정된 대사량이 아니라 분류기 확률, 움직임 근거, 심박 근거 및 맥락 불확실성을 이용해 추정한 알림 정책 범주로 정의하였다. 전체 분석 절차와 행동 레벨별 SQI/CVI 알림 정책의 구조는 Figure 1에 제시하였다.

2.3.1 SQI calculation

SQI는 각 분석 구간의 생체신호가 스트레스 알림 판단에 사용할 만큼 신뢰 가능한지를 나타내는 구간 수준 신호 신뢰도 함수로 정의하였다. SQI는 PPG 진폭 타당성, BVP peak 기반 박동 규칙성, ACC magnitude 기반 움직임 억제, EDA 안정성의 네 구성요소로 계산하였다.

$$\mathrm{SQI}_i = \frac{Q_i^{\mathrm{PPGamp}} + Q_i^{\mathrm{beat}} + Q_i^{\mathrm{motion}} + Q_i^{\mathrm{EDA}}}{4} \tag{1}$$

PPG 진폭 타당성은 BVP 신호의 95th percentile과 5th percentile 차이를 training fold의 percentile 기준으로 정규화하였다. 박동 규칙성은 BVP peak 간격의 변동계수에 반비례하도록 정의하였다. 움직임 억제 점수는 ACC magnitude 변동량에 반비례하도록 정의하였다. EDA 안정성은 인접 EDA 변화량과 EDA 기울기를 결합한 아티팩트 지표에 반비례하도록 정의하였다. 모든 percentile-based normalization 기준은 정보누출을 방지하기 위해 training fold에서만 산출하고 validation 또는 test fold에는 동일 기준을 적용하였다.

$$A_i = P_{95}(\mathrm{BVP}_i) - P_5(\mathrm{BVP}_i), \quad Q_i^{\mathrm{PPGamp}} = \mathrm{clip}\left(\frac{A_i - A_{10}^{\mathrm{train}}}{A_{90}^{\mathrm{train}} - A_{10}^{\mathrm{train}}}, 0, 1\right) \tag{2}$$

$$\mathrm{CV}_{\mathrm{IBI},i} = \frac{\mathrm{SD}(\mathrm{IBI}_i)}{\mathrm{Mean}(\mathrm{IBI}_i)}, \quad Q_i^{\mathrm{beat}} = 1 - \mathrm{clip}\left(\frac{\mathrm{CV}_{\mathrm{IBI},i}}{\mathrm{CV}_{90}^{\mathrm{train}}}, 0, 1\right) \tag{3}$$

$$M_i^{\mathrm{ACC}} = \mathrm{SD}(\mathrm{ACCmag}_i), \quad Q_i^{\mathrm{motion}} = 1 - \mathrm{clip}\left(\frac{M_i^{\mathrm{ACC}} - M_{10}^{\mathrm{train}}}{M_{90}^{\mathrm{train}} - M_{10}^{\mathrm{train}}}, 0, 1\right) \tag{4}$$

$$A_i^{\mathrm{EDA}} = \mathrm{median}(|\Delta\mathrm{EDA}_i|) + |\mathrm{Slope}(\mathrm{EDA}_i)|, \quad Q_i^{\mathrm{EDA}} = 1 - \mathrm{clip}\left(\frac{A_i^{\mathrm{EDA}} - A_{10}^{\mathrm{EDA,train}}}{A_{90}^{\mathrm{EDA,train}} - A_{10}^{\mathrm{EDA,train}}}, 0, 1\right) \tag{5}$$

본 연구는 SQI를 모든 분석 구간에 동일하게 적용되는 일괄 차단 기준으로 사용하지 않았다. 고정 임계값 기반 SQI 차단은 거짓 알림을 줄일 수 있으나 실제 스트레스 알림까지 제거하여 재현율을 낮출 수 있다. 이에 따라 SQI는 스트레스 근거 점수를 감쇠시키는 신뢰도 가중치로 적용하였다.

$$w_{\mathrm{SQI},i} = \mathrm{clip}(0.65 + 0.35\,\mathrm{SQI}_i, 0.35, 1.00), \quad S_i^{\mathrm{adj}} = S_i \times w_{\mathrm{SQI},i} \times k_{L_i} \tag{6}$$

2.3.2 CVI calculation

CVI는 독립된 단일 점수라기보다, 운동 근거, 스트레스-운동 맥락 차이, 맥락 확신도, 구제 규칙을 결합한 맥락 신뢰도 규칙으로 정의하였다. Stress/rest 분류기의 스트레스 확률을 $P^{SR}_{stress,i}$, aerobic/sprints 분류기의 sprint 확률을 $P^{AS}_{sprint,i}$, four-class 분류기의 각 클래스 확률을 $P^{4C}_{c,i}$로 정의하였다.

$$P_i^{\text{exercise}} = P^{4C}_{\text{aerobic},i} + P^{4C}_{\text{sprints},i}, \quad \Delta_i^{\text{context}} = P^{4C}_{\text{stress},i} - P_i^{\text{exercise}} \tag{7}$$

$$H_i = -\frac{\sum_c P^{4C}_{c,i} \log(P^{4C}_{c,i})}{\log(4)}, \quad C_i = 1 - H_i \tag{8}$$

$$S_i = 0.55\, P^{\text{SR}}_{\text{stress},i} + 0.35\, P^{4C}_{\text{stress},i} + 0.10\, C_i \tag{9}$$

식 (9)의 고정 계수 0.55, 0.35 및 0.10은 각각 stress/rest 스트레스 확률, four-class 스트레스 확률 및 맥락 확신도에 부여한 주 분석 가중치이다. Table 7에서는 이 세 계수를 각각 λ_SR, λ_4C 및 λ_C로 일반화하여 표기하였으며, λ_SR+λ_4C+λ_C=1을 유지하였다. 주 분석에서는 (λ_SR, λ_4C, λ_C)=(0.55, 0.35, 0.10)으로 고정하였다.

H_i는 four-class 확률분포의 정규화 엔트로피이고, $\log(4)$는 네 클래스 분류에서 가능한 최대 엔트로피이다. C_i가 높을수록 four-class 분류기가 특정 맥락을 명확하게 판단했음을 의미한다. 로그의 밑은 numerator와 denominator에서 동일하게 사용되므로 자연로그와 base-2 로그 모두 동일한 정규화 결과를 제공한다.

$$V_i^{\text{SQI}} = \text{I}\left(\text{SQI}_i < q_{m,L_i} \text{ AND } P^{\text{SR}}_{\text{stress},i} < 0.88\right) \tag{10}$$

$$V_i^{\text{exercise}} = \text{I}\left(P_i^{\text{exercise}} > e_{m,L_i} \text{ AND } P^{4C}_{\text{stress},i} < 0.65\right) \tag{11}$$

$$V_i^{\text{margin}} = \text{I}\left(\Delta_i^{\text{context}} < \gamma_{m,L_i} \text{ AND } P^{\text{SR}}_{\text{stress},i} < 0.85\right) \tag{12}$$

구제 규칙은 보정 후 점수 S_i^{adj}가 아니라 감쇠 전 분류기 출력에 근거하였다. 이는 SQI와 행동 맥락 감쇠가 적용되기 전의 높은 스트레스 근거를 보존하기 위한 의도적 설계이다.

2.3.3 Behavior-level assignment and alert-policy parameters

행동 레벨은 L0 rest/sedentary, L1 light activity, L2 moderate activity, L3 vigorous/sprint, U uncertain context의 다섯 범주로 구성하였다 (Tremblay et al., 2017). 행동 레벨별 감쇠계수 k_L은 $L_0 = 1.00$, $L_1 = 0.92$, $L_2 = 0.72$, $L_3 = 0.48$, $U = 0.62$로 사전 고정하였다. 운영 모드별 스트레스 판단 임계값, 최소 SQI 임계값, 운동 차단 임계값, 맥락 차이 임계값 및 rescue 조건은 최종 평가 이전에 사전 고정하여 동일하게 적용하였다. 행동 레벨의 할당 기준과 감쇠계수는 Table 1에 제시하였다.

$$I_i = 0.35\, P_i^{\text{exercise}} + 0.30\, P^{\text{AS}}_{\text{sprint},i} + 0.20\, \text{ACC}_i^{\text{norm}} + 0.15\, \text{HR}_i^{\text{norm}} \tag{13}$$

알림 정책 파라미터는 최종 test set 성능을 직접 최적화하여 산출한 값이 아니라, 운영 모드별 민감도 수준을 구분하기 위해 사전에 고정한 규칙값으로 사용하였다. 파라미터는 신체활동 수준이 높을수록 움직임 아티팩트와 운동 맥락 혼동 가능성이 증가한다는 문헌적 전제와 행동강도 분류 기준(Ainsworth et al., 2011; Haskell et al., 2007; Pate et al., 1995), 그리고 민감형-균형형-엄격형의 운영 목적에 따라 단조적으로 설정하였고 최종 test set 평가 이전에 확정하였다. 단일 공개 데이터셋 기반 2차 분석이라는 한계를 고려하여, 본 연구의 결과는 확정적 성능 검증이 아니라 탐색적 알림 정책 평가에 해당한다.

Table 1. Behavior-level assignment rules and attenuation factors

Behavior level	Assignment rule	k_L	Alert-policy interpretation
L0 Rest/Sedentary	$P_{\text{rest}}^{\text{4C}} \geq 0.55$, $P^{\text{exercise}} < 0.35$, and $I_i < 0.40$	1.00	Rest-dominant context; preserve stress sensitivity
L1 Light activity	Default when L0, L2, L3, and U are not assigned	0.92	Mild activity; weak context correction
L2 Moderate activity	$0.45 \leq I_i < 0.70$	0.72	Exercise-like physiology; apply moderate context veto
L3 Vigorous/Sprint	$I_i \geq 0.70$, or $P_{\text{sprints}}^{\text{4C}} \geq 0.50$, or $P_{\text{sprint}}^{\text{AS}} \geq 0.70$	0.48	Strong exercise context; apply strong context veto
U Uncertain	$H_i \geq 0.82$, or $\max(P_c^{\text{4C}}) < 0.43$, or stress-exercise conflict	0.62	Ambiguous context; require strong stress evidence

2.3.4 Final alert-decision rule

$$\text{Alert}_i^{(m)} = \text{I}\left(S_i^{\text{adj}} \geq \theta_{m,L_i}\right)\left(1 - V_i^{\text{SQI}}\right)\left(1 - V_i^{\text{exercise}}\right)\left(1 - V_i^{\text{margin}}\right) \tag{14}$$

여기서 m은 운영 모드이며, 민감형, 균형형, 엄격형으로 구성하였다. θ_{m,L_i}는 행동 레벨별 스트레스 판단 임계값, q_{m,L_i}는 최소 SQI 임계값, e_{m,L_i}는 운동 차단 임계값, γ_{m,L_i}는 스트레스-운동 맥락 차이 임계값이다. SQI/CVI 알림 정책에 사용된 핵심 수식과 각 구성요소의 해석은 Table 2에 정리하였다.

Table 2. Core equations used in the behavior-level SQI/CVI alert-policy framework

Component	Equation	Interpretation
Signal Quality Index	$\text{SQI}_i = \dfrac{Q_i^{\text{PPGamp}} + Q_i^{\text{beat}} + Q_i^{\text{motion}} + Q_i^{\text{EDA}}}{4}$	Segment-level signal reliability
SQI weight	$w_{\text{SQI},i} = \text{clip}(0.65 + 0.35\,\text{SQI}_i,\ 0.35,\ 1.00)$	Low SQI attenuates stress evidence
Context confidence	$C_i = 1 - H_i$	Certainty of four-class context classification
Stress evidence	$S_i = 0.55\,P_{\text{stress},i}^{\text{SR}} + 0.35\,P_{\text{stress},i}^{\text{4C}} + 0.10\,C_i$	Integrated stress evidence
Exercise evidence	$P_i^{\text{exercise}} = P_{\text{aerobic},i}^{\text{4C}} + P_{\text{sprints},i}^{\text{4C}}$	Probability of exercise-related context
Context margin	$\Delta_i^{\text{context}} = P_{\text{stress},i}^{\text{4C}} - P_i^{\text{exercise}}$	Stress dominance over exercise context
Intensity score	$I_i = 0.35\,P_i^{\text{exercise}} + 0.30\,P_{\text{sprint},i}^{\text{AS}} + 0.20\,\text{ACC}_i^{\text{norm}} + 0.15\,\text{HR}_i^{\text{norm}}$	Behavior-level evidence
Adjusted stress score	$S_i^{\text{adj}} = S_i \times w_{\text{SQI},i} \times k_{L_i}$	Stress score after SQI and behavior-level attenuation
Final alert	$\text{I}\left(S_i^{\text{adj}} \geq \theta\right)\left(1 - V_i^{\text{SQI}}\right)\left(1 - V_i^{\text{exercise}}\right)\left(1 - V_i^{\text{margin}}\right)$	Alert after stress, signal-quality, and context rules

2.4 Evaluation metrics and statistical analysis

주 지표는 시간당 거짓 알림 발생률(false alerts/h)로 설정하였다. 각 subject/session 내 예측 알림 상태가 0에서 1로 전환되는 시점을 알림 발생 사건(alert onset)으로 정의하였다. 시간당 거짓 알림 발생률은 non-stress 구간에서 관찰된 거짓 알림 발생 사건 수를 해당 non-stress 위험시간(hours at risk)으로 나누어 계산하였다.

운동 맥락 거짓 알림 발생률(exercise-context false alerts/h)은 non-stress이면서 운동 맥락에 해당하는 시간만을 분모로 사용하여 동일하게 계산하였다. 따라서 두 발생률은 서로 다른 time-at-risk denominator를 가지며, 직접적인 크기 비교보다는 각각 전체 non-stress 시간과 운동 맥락 non-stress 시간에서의 알림 부담으로 해석해야 한다.

시간당 거짓 알림 발생률은 신규 거짓 알림 사건의 event rate를 나타내는 반면, 시간당 알림 활성시간(alert minutes/h)은 알림 상태가 활성화된 총 시간을 나타낸다. 따라서 특정 알림 정책은 거짓 알림 발생률을 낮추면서도 실제 스트레스 알림을 더 오래 유지하여 시간당 알림 활성시간을 증가시킬 수 있다.

보조 지표는 시간당 알림 활성시간(alert minutes/h), precision, recall, F1-score, accuracy, balanced accuracy였다. 대상자 수준 대응 F1 점수($n = 36$ subjects)에 대해 Wilcoxon signed-rank test를 수행하였고, 분석 구간 수준의 예측 불일치는 exact McNemar test로 평가하였다. 다중 비교는 Holm 보정으로 제어하였다. 추가분석으로 ablation analysis, stress/rest weight sensitivity analysis, leave-one-subject-out threshold validation을 수행하였다.

$$N^{\text{onset}} = \sum_t I\,(\hat{y}_t = 1 \text{ AND } \hat{y}_{t-1} = 0), \quad \text{False alerts/h} = \frac{N^{\text{onset}}_{\text{non-stress}}}{\text{Hours}_{\text{non-stress}}} \tag{15}$$

3. Results

3.1 Benchmark reproduction

Benchmark reproduction 결과는 Table 3과 Figure 2에 제시하였다. Hongn et al. (2025)이 보고한 정확도와 비교했을 때, four-class 과제는 0.840에 대해 0.839로 거의 일치하여 accuracy gap이 -0.001이었다. Aerobic/sprints 과제는 보고값 0.910보다 높은 0.950으로 +0.040의 차이를 보였고, stress/rest 과제는 보고값 0.930보다 낮은 0.834로 -0.096의 차이를 보였다. 재현 F1-score는 stress/rest 0.864, aerobic/sprints 0.945 및 four-class 0.835였다.

Table 3. Benchmark reproduction results compared with Hongn et al. (2025)

Task	Reported accuracy	Reproduced accuracy	Accuracy gap	Reported F1	Reproduced F1	N	Selected configuration
Stress/Rest	0.930	0.834	-0.096	0.920	0.864	256	Stress_last60 + corr_filter_08 + random_over + deep_low_lr
Aerobic/Sprints	0.910	0.950	+0.040	0.910	0.945	226	Sprint_middle30 + all_available + random_over + regularized
Four-class	0.840	0.839	-0.001	0.840	0.835	482	Stress_last60 + all_available + SMOTE + stable_depth3

Four-class 분류기의 재현 성능은 Hongn et al. (2025)의 보고값과 거의 일치하였으므로 중심 맥락 분류기로 사용하였다. 반면 stress/rest 분류기는 보고된 정확도에 도달하지 못했으므로, 해당 출력은 통합 스트레스 근거 점수의 한 구성요소로 제한하였다. 차이가 발

생한 가능한 원인으로는 Hongn et al. (2025)에서 완전히 공개되지 않은 분석 구간 선택, feature preprocessing, protocol block parsing, split seed 및 내부 XGBoost 구현 차이가 있다. 따라서 이후 결과는 공개 데이터셋에 기반한 탐색적 알림 정책 평가로 해석하였다.

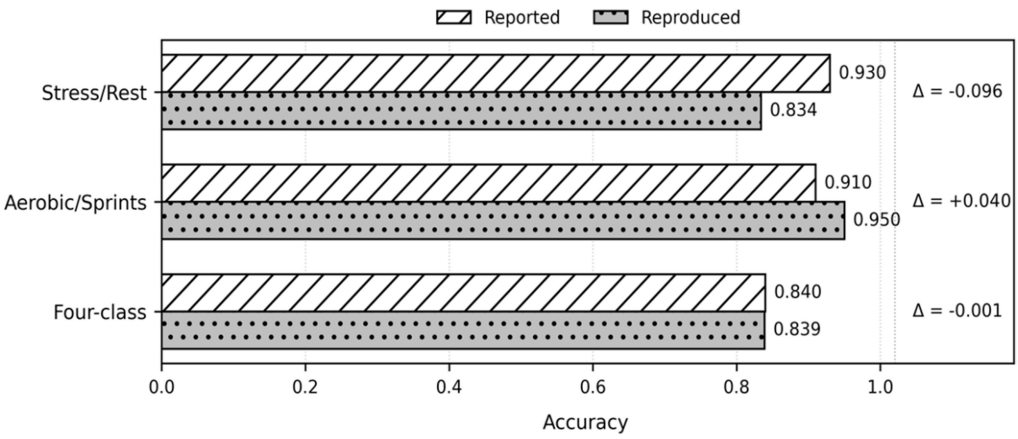


Figure 2. Benchmark reproduction accuracy compared with Hongn et al. (2025)

3.2 Main alert-level performance of the proposed alert policy

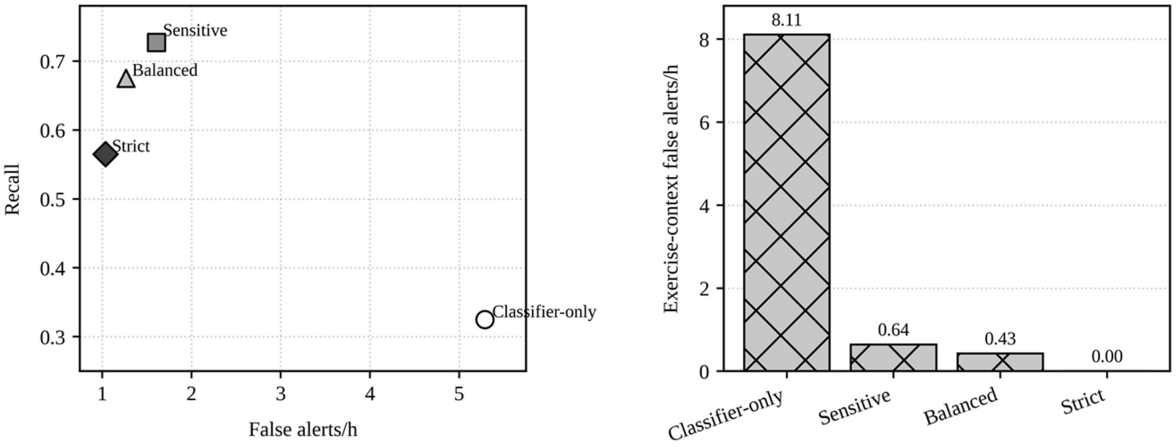
행동 레벨 SQI+CVI 알림 정책은 분류기 단독 기준선보다 시간당 거짓 알림 발생률과 운동 맥락 거짓 알림 발생률을 감소시켰다. 분류기 단독 기준선의 false alerts/h는 5.288이었고, SQI+CVI 알림 정책 적용 후 민감형, 균형형, 엄격형에서 각각 1.609, 1.265, 1.035로 감소하였다. 운동 맥락 거짓 알림 발생률은 8.110/h에서 각각 0.640/h, 0.427/h, 0.000/h로 감소하였다.

재현율은 분류기 단독 기준선의 0.325에서 민감형 0.727, 균형형 0.675, 엄격형 0.565로 증가하였다. F1-score는 0.373에서 각각 0.759, 0.740, 0.680으로 증가하였다. 이는 제안 알림 정책이 단순히 알림을 줄인 것이 아니라, 행동 레벨에 따라 스트레스 근거를 재점수화하고 운동 맥락에서 부적절한 알림을 선별적으로 억제했기 때문이다. 운영 모드별 주요 알림 수준 성능은 Table 4에 제시하였으며, 시간당 거짓 알림 발생률과 재현율의 관계 및 운동 맥락 거짓 알림 억제 결과는 Figure 3에 제시하였다.

민감형에서는 시간당 알림 활성시간(alert min/h)이 분류기 단독 기준선보다 증가하였다. 이는 오류가 아니라 event-rate 지표와 duration 지표의 차이에 따른 결과이다. 민감형은 낮은 스트레스 판단 임계값을 적용하여 더 많은 스트레스 구간을 회복하도록 설계되었다. 따라서 시간당 거짓 알림 발생률은 감소했지만, 알림 상태가 활성화된 총 시간은 증가하였다. 즉, 민감형은 거짓 알림 발생 사건을 줄이면서 재현율을 보존하는 알림 정책으로 해석해야 한다.

Table 4. Main alert-level performance of the combined SQI+CVI alert policy

Alert policy	False alerts/h	Exercise-context false alerts/h	Alert min/h	Precision	Recall	F1	Accuracy	Balanced accuracy
Classifier-only	5.288	8.110	8.833	0.439	0.325	0.373	0.651	0.565
Sensitive SQI+CVI	1.609	0.640	11.611	0.794	0.727	0.759	0.853	0.819
Balanced SQI+CVI	1.265	0.427	9.913	0.819	0.675	0.740	0.849	0.803
Strict SQI+CVI	1.035	0.000	7.976	0.853	0.565	0.680	0.830	0.760



(a) False alerts/h vs. recall (b) Exercise-context suppression

Figure 3. Alert-level performance after the behavior-level SQI/CVI alert policy

대표 비교인 Balanced SQI+CVI와 분류기 단독 기준선의 대상자 수준 F1-score 차이에 대한 Wilcoxon 부호순위 검정은 $W = 15.5$, raw $p = 3.36 \times 10^{-6}$, Holm 보정 $p = 1.34 \times 10^{-5}$로 유의하였다. 동일 비교의 exact McNemar 검정도 $b = 22$, $c = 117$, raw $p = 7.58 \times 10^{-17}$, Holm 보정 $p = 3.79 \times 10^{-16}$으로 유의하였다.

3.3 SQI × CVI interaction analysis

SQI × CVI 상호작용 분석은 분류기 단독 기준선에서 알림 후보가 된 분석 구간 중 어떤 알림이 SQI 또는 CVI에 의해 차단되었는지를 평가하였다. 세 알림 정책 모두에서 거짓 알림 차단은 주로 CVI에 의해 발생하였다. 균형형에서는 64개의 false candidate alert 중 53개가 CVI에 의해 차단되었고, 50개의 true candidate alert 중 1개만 CVI에 의해 차단되었다. SQI 단독 차단은 관찰되지 않았다.

이는 SQI가 본 프레임워크에서 독립적인 일괄 차단 규칙으로 설계되지 않았기 때문이다. 본 연구에서 SQI의 주된 역할은 보정 스트레스 점수 S_i^{adj}를 계산하는 신뢰도 가중치이다. 반면 CVI는 운동 근거와 맥락 차이를 이용하여 맥락적으로 부적절한 알림을 직접 차단하였다.

민감형과 균형형의 CVI 차단 수가 동일하게 나타난 것은 수치 오류가 아니라 interaction analysis의 정의에서 비롯된다. Table 5는 전체

Table 5. SQI × CVI interaction analysis

Alert policy	Candidate alerts	False candidates	True candidates	False blocked by SQI	False blocked by CVI	True blocked by CVI	CVI false-block rate (%)	CVI true-block rate (%)
Sensitive	114	64	50	0	53	1	82.812	2.000
Balanced	114	64	50	0	53	1	82.812	2.000
Strict	114	64	50	0	54	1	84.375	2.000

Note. Candidate alerts are alert candidates generated by the classifier-only baseline. Sensitive and Balanced produced identical CVI blocking counts because the same candidate-alert subset crossed the CVI veto boundary under both operating modes. SQI-only blocking was not observed because SQI was implemented primarily as confidence weighting through the adjusted stress score rather than as an independent exclusion rule

알림 정책 성능이 아니라 분류기 단독 기준선에서 발생한 candidate alert 중 CVI가 차단한 부분집합만을 집계한다. 민감형과 균형형은 스트레스 판단 임계값, SQI 가중, 시간당 알림 활성시간에서는 차이를 보였지만, 본 표에서 평가한 candidate subset에 대해서는 동일한 CVI veto boundary를 통과하거나 차단하였다.

3.4 Ablation analysis

단순 activity filter는 거짓 알림을 크게 줄였지만 재현율을 회복하지 못하였다. Balanced mode에서 activity filter only의 false alerts/h는 1.150이었으나 recall은 0.318로 분류기 단독 기준선의 0.325와 거의 유사하였다. 이는 운동 맥락에서 알림을 단순히 차단하는 방식이 알림 부담은 낮출 수 있지만 스트레스 민감도 개선에는 충분하지 않음을 보여준다. 균형형 모드에서 각 구성요소의 제거에 따른 성능 변화는 Table 6과 Figure 4에 제시하였다.

Table 6. Ablation comparison in the balanced mode

Model	False alerts/h	Exercise-context false alerts/h	Alert min/h	Precision	Recall	F1	Accuracy	Balanced accuracy
Classifier-only	5.288	8.110	8.833	0.439	0.325	0.373	0.651	0.565
Activity filter only	1.150	0.213	4.360	0.817	0.318	0.458	0.759	0.642
Strict activity filter	1.035	0.000	4.307	0.831	0.318	0.460	0.761	0.644
SQI-only	1.724	1.067	11.576	0.785	0.734	0.758	0.851	0.820
CVI-only	1.379	0.640	10.180	0.808	0.682	0.739	0.846	0.803
SQI+CVI	1.265	0.427	9.913	0.819	0.675	0.740	0.849	0.803

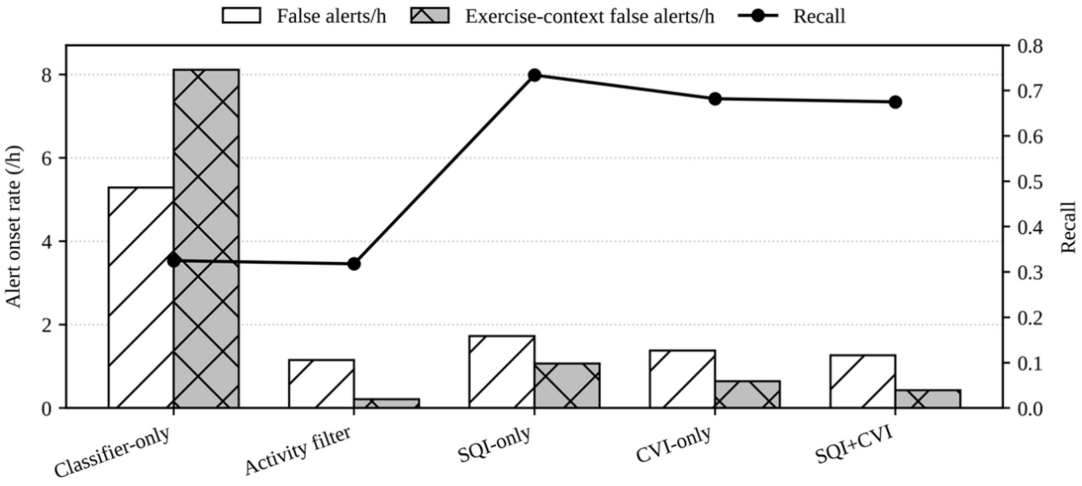


Figure 4. Ablation comparison in the Balanced mode

SQI-only는 Balanced mode에서 F1-score가 0.758로 가장 높았지만, exercise-context false alerts/h가 1.067이고 alert min/h가 11.576으로 상대적으로 컸다. 반면 SQI+CVI 결합 알림 정책은 F1-score 0.740을 유지하면서 false alerts/h를 1.265, exercise-context false alerts/h를 0.427, alert min/h를 9.913으로 낮추었다. 따라서 제안 프레임워크의 기여는 단일 F1-score 최대화가 아니라, 시간당 거짓 알림 발생률, 운동 맥락 오류, 시간당 알림 활성시간 및 재현율 사이의 운영상 균형을 달성하는 데 있다.

3.5 Stress/Rest weight sensitivity analysis

Stress/rest 분류기의 재현 성능이 Hong et al. (2025)의 보고값에 미달하였으므로, 통합 스트레스 근거 점수의 가중치가 최종 알림 정책 성능에 미치는 영향을 민감도 분석으로 평가하였다. 식 (9)의 0.55, 0.35 및 0.10을 각각 λ_SR, λ_4C 및 λ_C로 표기하였다. λ_SR은 stress/rest 스트레스 확률의 가중치, λ_4C는 four-class 스트레스 확률의 가중치, λ_C는 맥락 확신도의 가중치이며, 세 가중치의 합은 1로 고정하였다. 주 분석은 (λ_SR, λ_4C, λ_C)=(0.55, 0.35, 0.10)이었다. 이때 stress/rest 출력은 rest와 stress를 직접 구분하는 유일한 이진 스트레스 근거이고, four-class stress probability는 맥락 분류와 함께 산출된 스트레스 근거이다. 맥락 확신도는 단독 판정 근거가 아니라 확률분포의 불확실성을 보조적으로 반영하는 항이므로, Table 7의 민감도 분석에서는 λ_C=0.10을 유지한 채 λ_SR과 λ_4C를 재배분하였다. Balanced mode에서 λ_SR을 낮추거나 제거해도 false alerts/h는 1.265 수준으로 유지되었으나 recall과 F1-score는 감소하였다. 특히 four-class dominant 조건에서는 recall이 0.558, F1-score가 0.656으로 낮아졌다.

이 결과는 두 가지를 의미한다. 첫째, false alerts/h 억제는 주로 CVI veto 규칙에 의존하므로 stress/rest 가중치와 상대적으로 독립적이며, 이는 설계 의도와 일치한다. 둘째, recall과 F1-score의 감소는 stress/rest 출력이 스트레스 민감도 유지에 기여함을 보여준다. 따라서 본 연구는 stress/rest 분류기의 완전 재현을 주장하지 않지만, 해당 출력이 통합 스트레스 근거의 일부로 사용될 때 알림 정책 성능에 기여한다는 점을 민감도 분석으로 확인하였다.

Table 7. Stress/Rest weight sensitivity analysis in the balanced mode

Condition	λ_SR	λ_4C	λ_C	False alerts/h	Exercise-context false alerts/h	Alert min/h	Precision	Recall	F1
Original	0.55	0.35	0.10	1.265	0.427	9.913	0.819	0.675	0.740
Reduced SR	0.25	0.65	0.10	1.265	0.000	9.329	0.807	0.623	0.703
Four-class dominant	0.00	0.90	0.10	1.265	0.000	8.440	0.796	0.558	0.656
Equal SR/4C	0.45	0.45	0.10	1.265	0.427	9.727	0.816	0.662	0.731

Note. λ_SR, λ_4C, and λ_C are the weights assigned in Eq. (9) to the stress/rest stress probability, four-class stress probability, and context confidence, respectively; λ_SR+λ_4C+λ_C=1. The primary analysis used (0.55, 0.35, 0.10)

3.6 Subject-level alert-policy threshold validation

알림 정책 임계값의 안정성을 점검하기 위해 leave-one-subject-out threshold validation을 수행하였다. Training subject만을 이용해 threshold를 선택한 뒤 held-out subject에서 평가한 결과, Balanced mode 기준 F1-score는 평균 0.736이고 recall은 0.804이었다. 다만 시간당 알림 활성시간(alert min/h)은 15.781, exercise-context false alerts/h는 1.407로 증가하였다. 따라서 이 분석은 주 운영 성능표를 대체하기보다, 알림 정책 임계값이 held-out subject에서도 일정 수준의 recall과 F1을 유지함을 확인하는 보조 검증으로 활용하였다.

4. Discussion

본 연구의 주요 발견은 SQI와 CVI를 결합한 행동 레벨 알림 정책이 분류기 단독 기준선보다 알림 품질을 개선했다는 점이다. 다만 추가분석 결과는 제안한 알림 정책이 모든 지표에서 최고 성능을 보이는 모델이 아님을 보여주었다. SQI-only는 Balanced mode에서 가장 높은 F1-score를 보였지만, 운동 맥락 거짓 알림 발생률과 시간당 알림 활성시간이 상대적으로 컸다. 반면 SQI+CVI 결합 알림 정책은 F1-score를 일정 수준 유지하면서 false alerts/h, exercise-context false alerts/h 및 alert min/h를 함께 낮추었다. 따라서 본 연구의 기여는 단일 성능 지표의 최대화가 아니라, 알림 발생률, 운동 맥락 오류, 알림 활성시간 및 재현율 사이의 운영상 균형을 달성하는 행동 레벨 알림 정책을 제시한 데 있다.

Stress/rest 분류기의 성능 미달은 본 연구의 중요한 한계이다. 최적화 탐색 후에도 stress/rest 정확도는 0.834로 Hongn et al. (2025)의 보고값 0.930에 도달하지 못했다. 이에 따라 본 연구는 stress/rest 분류기의 완전 재현을 주장하지 않는다. Four-class 분류기는 보고값과 거의 일치했으므로 중심 맥락 분류기로 사용하였고, stress/rest 출력은 통합 스트레스 근거의 한 구성요소로 제한하였다. 민감도 분석 결과, stress/rest 가중치를 줄이거나 제거해도 거짓 알림 억제 효과는 유지되었으나 재현율과 F1-score는 감소하였다. 이는 stress/rest 출력의 불확실성이 존재하지만, 해당 출력이 스트레스 민감도 유지에는 일정 부분 기여함을 의미한다.

본 연구에서 SQI와 CVI는 서로 다른 역할을 수행하였다. SQI는 낮은 신호품질 구간을 직접 제거하는 일괄 차단 규칙이 아니라, 보정 스트레스 점수를 계산하는 신뢰도 가중치로 설계되었다. 반면 CVI는 운동 근거와 맥락 차이를 이용해 운동으로 설명될 가능성이 높은 스트레스 알림을 직접 차단하였다. 따라서 SQI 단독 차단이 관찰되지 않았다는 결과는 SQI가 무효하다는 의미가 아니라, SQI의 기능이 직접 차단보다 점수 보정에 집중되어 있음을 의미한다.

민감형에서 시간당 알림 활성시간(alert min/h)이 증가한 점은 중요한 운영상 상충 관계이다. 시간당 거짓 알림 발생률(false alerts/h)은 신규 거짓 알림 사건의 빈도를 나타내는 발생률 지표이고, 시간당 알림 활성시간은 알림 상태가 활성화된 총 시간을 나타내는 duration 지표이다. 민감형은 재현율 보존을 목표로 하므로 더 많은 스트레스 구간을 회복하고 알림을 더 오래 유지할 수 있다. 따라서 민감형은 거짓 알림 발생률을 낮추면서도 전체 알림 활성시간을 증가시킬 수 있다.

Introduction에서 정리한 맥락 기반 스트레스 인식 연구와 같은 방향으로, 본 연구의 행동 레벨 맥락 veto는 균형형에서 운동 맥락 거짓 알림 발생률을 8.110/h에서 0.427/h로 감소시켰다. 기존 연구가 맥락 정보를 이용한 분류 성능 향상에 주로 초점을 둔 반면(Gjoreski et al., 2017; Hovsepian et al., 2015; Can et al., 2019; Schmidt et al., 2018), 본 연구는 맥락 보정을 알림 정책 단계의 veto와 rescue로 구현하고 그 효과를 알림 발생률과 활성시간의 관점에서 정량화하였다.

신호품질 측면에서 SQI는 고정 임계값 기반 이진 품질 게이팅과 달리 스트레스 근거 점수를 감쇠시키는 신뢰도 가중으로 작동하였다(Kleckner et al., 2018; Orphanidou et al., 2015). 이는 실제 스트레스 근거를 과도하게 제거하지 않으면서 CVI와 결합할 수 있도록 한 설계이다. 또한 균형형에서 시간당 거짓 알림 발생률이 5.288/h에서 1.265/h로 감소한 결과는 반복 경보에 따른 알림 피로 문제(Ancker et al., 2017; Drew et al., 2014)를 웨어러블 스트레스 모니터링의 운영 지표로 구체화한다. 즉, 본 연구는 선행연구가 제시한 신호품질과 맥락의 중요성을 알림 발생률과 시간당 알림 활성시간이라는 사용자 부담 지표로 연결하였다.

본 연구의 독창성은 새로운 스트레스 분류 알고리즘에 있지 않다. 본 연구는 분류 정확도 대신 시간당 알림 발생률과 시간당 알림 활성시간을 분리하여 평가하는 운영 수준의 관점, 그리고 신호품질과 맥락타당성을 행동 레벨 알림 정책으로 통합하는 설계에 초점을 두었다. 맥락 기반 오탐 억제와 신호품질 기반 알림 제어 자체는 선행연구에서 다루어졌으나, 본 연구는 이를 행동 레벨 운영 모드와 알림 부담 지표 위에서 결합하고 평가하였다.

본 연구에는 몇 가지 한계가 있다. 첫째, stress/rest 분류기는 Hongn et al. (2025)이 보고한 성능값을 완전히 재현하지 못하였다. 둘째, 알림 정책 파라미터는 사전 정의된 운영 규칙으로 사용되었으나, 완전한 confirmatory validation을 위해서는 nested 피험자 수준 validation에서 정책 임계값을 training fold 내부에서만 선택하는 절차가 필요하다. 셋째, 본 연구의 행동 레벨은 실제 MET 측정값이 아니라, 분류기 확률과 움직임·심박 근거를 이용해 추정한 알림 정책 범주이다. 넷째, 본 분석은 단일 공개 데이터셋의 2차 분석이므로 독립 외부 데이터셋 검증이 필요하다. 다섯째, 단순 activity filter, SQI-only, CVI-only와의 비교는 추가되었으나, 다른 context-aware 방법과의 직접 비교는 후속 연구에서 더 엄밀하게 수행되어야 한다.

한편, 본 연구의 비교 기준선은 재현된 분류기의 기본 임계값 출력이라는 점도 해석상 유의해야 한다. 제안 알림 정책에서 관찰된 재현율 개선에는 행동 레벨 재임계화의 효과가 함께 포함되며, 단순 activity filter 조건(재현율 0.318)이 맥락 차단만으로는 재현율이 개선되지 않음을 보여 알림 정책의 기여를 부분적으로 분리한다. 다만 알림 정책 고유의 효과를 더 엄격히 분리하려면 운영 임계값을 조정한 분류기와 비용민감(cost-sensitive) 분류기를 추가 기준선으로 설정하여 동일한 알림 수준 지표에서 비교하는 절차가 필요하다.

또한 본 연구의 leave-one-subject-out 검증에서는 운동 맥락 거짓 알림 발생률과 시간당 알림 활성시간이 주 성능표보다 증가하였다.

이는 보고된 운영 성능이 배치 환경의 상한이 아닐 수 있음을 시사하므로, 독립적인 외부 데이터셋과 상이한 기기·집단에서의 검증을 통해 실사용 성능을 확정하는 절차가 선행되어야 한다.

5. Conclusion

본 연구는 Hongn 공개 웨어러블 생체신호 데이터셋의 XGBoost 기반 분류기를 재현하고, 세 분류기 출력 위에 행동 레벨별 SQI/CVI 알림 정책을 적용하였다. 제안 프레임워크는 단일 임계값이 아니라 행동 레벨과 맥락에 따라 신호품질 지수와 맥락타당성 지수의 역할을 전환하는 구조를 갖는다.

추가분석 결과, SQI+CVI 결합 알림 정책은 모든 단일 지표에서 최고 성능을 보이지는 않았지만, false alerts/h, exercise-context false alerts/h, alert min/h 및 재현율 사이의 균형을 제공하였다. Balanced mode에서 SQI+CVI 알림 정책은 분류기 단독 기준선 대비 false alerts/h를 5.288에서 1.265로 낮추었고, exercise-context false alerts/h를 8.110에서 0.427로 낮추었으며, 재현율을 0.325에서 0.675로 높였다.

따라서 본 연구의 기여는 분류 정확도 대신 시간당 알림 발생률과 시간당 알림 활성시간을 분리하여 평가하는 운영 수준의 관점을 제시하고, 그 위에서 알림 품질을 개선하는 행동 레벨 알림 정책을 설계한 데 있다. 특히 단일 지표(정확도·F1) 기준의 알림 정책 순위가 알림 부담 지표 기준의 순위와 달라질 수 있음을 보였다. 다만 stress/rest 분류기 재현 성능이 보고값에 미달하였고 독립 외부 검증이 수행되지 않았으므로, 본 결과는 확정적 시스템 성능이 아니라 탐색적 알림 정책 설계 근거로 보아야 한다.

Acknowledgement

This research was supported by the Startup Growth Technology Development Program of the Ministry of SMEs and Startups, Republic of Korea (RS-2025-23523821).

References

Ainsworth, B.E., Haskell, W.L., Herrmann, S.D., Meckes, N., Bassett, D.R., Jr., Tudor-Locke, C., Greer, J.L., Vezina, J., Whitt-Glover, M.C. and Leon, A.S., 2011 Compendium of Physical Activities: a second update of codes and MET values, *Medicine and Science in Sports and Exercise*, 43(8), 1575-1581, 2011. doi:10.1249/MSS.0b013e31821ece12

Allen, J., Photoplethysmography and its application in clinical physiological measurement, *Physiological Measurement*, 28(3), R1-R39, 2007. doi:10.1088/0967-3334/28/3/R01

Ancker, J.S., Edwards, A., Nosal, S., Hauser, D., Mauer, E. and Kaushal, R., Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system, *BMC Medical Informatics and Decision Making*, 17(1), 36, 2017. doi:10.1186/s12911-017-0430-8

Can, Y.S., Chalabianloo, N., Ekiz, D. and Ersoy, C., Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study, *Sensors*, 19(8), 1849, 2019. doi:10.3390/s19081849

Can, Y.S., Chalabianloo, N., Ekiz, D., Fernandez-Alvarez, J., Riva, G. and Ersoy, C., Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches, *IEEE Access*, 8, 38146-38163, 2020. doi:10.1109/ACCESS.2020.2975351

Chen, T. and Guestrin, C., XGBoost: a scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, 785-794, 2016. doi:10.1145/2939672.2939785

Drew, B.J., Harris, P., Zègre-Hemsey, J.K., Mammone, T., Schindler, D., Salas-Boni, R., Bai, Y., Tinoco, A., Ding, Q. and Hu, X., Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients, *PLoS ONE*, 9(10), e110274, 2014. doi:10.1371/journal.pone.0110274

Gjoreski, M., Luštrek, M., Gams, M. and Gjoreski, H., Monitoring stress with a wrist device using context, *Journal of Biomedical Informatics*, 73, 159-170, 2017. doi:10.1016/j.jbi.2017.08.006

Haskell, W.L., Lee, I.M., Pate, R.R., Powell, K.E., Blair, S.N., Franklin, B.A., Macera, C.A., Heath, G.W., Thompson, P.D. and Bauman, A., Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association, *Circulation*, 116(9), 1081-1093, 2007. doi:10.1161/CIRCULATIONAHA.107.185649

Hongn, A., Bosch, F., Prado, L.E., Ferrández, J.M. and Bonomini, M.P., Wearable physiological signals under acute stress and exercise conditions, *Scientific Data*, 12, 520, 2025. doi:10.1038/s41597-025-04845-9

Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M. and Kumar, S., cStress: towards a gold standard for continuous stress assessment in the mobile environment, *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 493-504, 2015. doi:10.1145/2750858.2807526

Kaufman, S., Rosset, S., Perlich, C. and Stitelman, O., Leakage in data mining: formulation, detection, and avoidance, *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15, 1-21, 2012. doi:10.1145/2382577.2382579

Kleckner, I.R., Jones, R.M., Wilder-Smith, O., Wormwood, J.B., Akcakaya, M., Quigley, K.S., Lord, C. and Goodwin, M.S., Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data, *IEEE Transactions on Biomedical Engineering*, 65(7), 1460-1467, 2018. doi:10.1109/TBME.2017.2758643

Orphanidou, C., Bonnici, T., Charlton, P., Clifton, D., Vallance, D. and Tarassenko, L., Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring, *IEEE Journal of Biomedical and Health Informatics*, 19(3), 832-838, 2015. doi:10.1109/JBHI.2014.2338351

Pate, R.R., Pratt, M., Blair, S.N., Haskell, W.L., Macera, C.A., Bouchard, C., Buchner, D., Ettinger, W., Heath, G.W., King, A.C., Kriska, A., Leon, A.S., Marcus, B.H., Morris, J., Paffenbarger, R.S., Jr., Patrick, K., Pollock, M.L., Rippe, J.M., Sallis, J.F. and Wilmore, J.H., Physical activity and public health: a recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine, *JAMA*, 273(5), 402-407, 1995. doi:10.1001/jama.1995.03520290054029

Sarker, H., Tyburski, M., Rahman, M.M., Hovsepian, K., Sharmin, M., Epstein, D.H., Preston, K.L., Furr-Holden, C.D., Milam, A., Nahum-Shani, I., al'Absi, M. and Kumar, S., Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data, *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4489-4501, 2016. doi:10.1145/2858036.2858218

Schmidt, P., Reiss, A., Dürichen, R., Marberger, C. and Van Laerhoven, K., Introducing WESAD, a multimodal dataset for wearable stress and affect detection, *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 400-408, 2018. doi:10.1145/3242969.3242985

Tamura, T., Maeda, Y., Sekine, M. and Yoshida, M., Wearable photoplethysmographic sensors - past and present, *Electronics*, 3(2), 282-302, 2014. doi:10.3390/electronics3020282

Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A. and Picard, R.W., Automatic identification of artifacts in electrodermal activity data, *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1934-1937, 2015. doi:10.1109/EMBC.2015.7318762

Tremblay, M.S., Aubert, S., Barnes, J.D., Saunders, T.J., Carson, V., Latimer-Cheung, A.E., Chastin, S.F.M., Altenburg, T.M. and Chinapaw, M.J.M., Sedentary Behavior Research Network (SBRN) - terminology consensus project process and outcome, *International Journal of Behavioral Nutrition and Physical Activity*, 14, 75, 2017. doi:10.1186/s12966-017-0525-8

Varoquaux, G., Cross-validation failure: small sample sizes lead to large error bars, *NeuroImage*, 180, 68-77, 2018. doi:10.1016/j.neuroimage.2017.06.061

Author listings

Seunghee Hong: zeele82622@gmail.com
Highest degree: PhD, Kagawa University, Kagawa, Japan
Position title: Principal Researcher, Ebenhaim
Areas of interes: Human foctor, Biosignals, ADAS, Biosignal-based resource allocation and risk management, Digital healthcare

Hang Joo Choi: gnada8581@gmail.com
Highest degree: Completion of Master's Program at Hanbat National University, Daejeon, Korea
Position title: CEO, Ebenhaim
Areas of interest: Human foctor, Biosignals, Location Data, Digital Twin, ESG, ISO45001, ISO30414, Biosignal-based resource allocation and risk management, Digital healthcare

Byungchan Min: bcmin@hanbat.ac.kr
Highest degree: PhD, The University of Electro-Communications, Tokyo, Japan
Position title: Professor, Department of Industrial and Management Engineering, Hanbat National University
Areas of interest: Human Factors, Work Design and Measurement, Affective Engineering