International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 6, Nov-Dec 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Quality-Aware Wrist Photoplethysmography for Subject-Independent Stress-State Monitoring: ECG-Referenced Validation and Selective Prediction

Egberibo Paul Waredibamo*, Oluwaseun Omotoso**
*(Department of Human Physiology, Bayelsa Medical University, Yenagoa, Nigeria
Email: paulwaredibamoegberibo@gmail.com)
**(Department of Public Health, Lead City University, Ibadan, Nigeria
Email: oluwaseunomotoso95@gmail.com)

--************************--

Abstract:

Wearable photoplethysmography (PPG) offers scalable monitoring of short-timescale physiological stress, but motion artifact and the non-equivalence of pulse-rate variability (PRV) and electrocardiographic heart-rate variability (HRV) limit reliability. This study evaluates a quality-aware PPG machine-learning framework using the WESAD benchmark. Wrist blood-volume pulse (64 Hz), wrist accelerometry (32 Hz), and chest ECG (700 Hz) from 15 participants were segmented into 865 60-s windows with a 30-s step for baseline-versus-stress classification. Subject-independent leave-one-subject-out (LOSO) evaluation compared raw PRV, conventionally filtered PRV, PPG+ACC, quality-aware PRV, and a full PPG+ACC+quality model. Raw PRV produced the highest mean Macro-F1 (0.8379), whereas PPG+ACC achieved the highest AUROC (0.9560) and AUPRC (0.9330). The full model achieved AUROC 0.9524, AUPRC 0.9295, and the lowest Brier score (0.1164). Selective prediction reduced error from 13.33% at 79.77% coverage to 7.48% at 61.85% coverage, while accepted-window Macro-F1 increased from 0.8346 to 0.8885. ECG-referenced analysis showed strong correlation for heart rate (r = 0.841, $p = 0.0107$) but weak correlation and poor agreement for RMSSD (r = 0.209; bias = 0.1669 s), demonstrating that PPG-derived PRV should not be treated as interchangeable with ECG-derived HRV. The results support uncertainty-aware wearable stress-state monitoring while identifying physiological validity and between-subject heterogeneity as key deployment constraints.

Keywords — ***Photoplethysmography, pulse-rate variability, occupational stress, wearable sensors, motion artifact, leave-one-subject-out, selective prediction, machine learning***

--************************--

I. INTRODUCTION

Occupational stress and fatigue are dynamic states that can fluctuate substantially within a work shift and accumulate under repeated physical, cognitive, and environmental demands. Persistent exposure can adversely affect worker performance, decision-making, and safety, creating a need for monitoring approaches capable of capturing short-term physiological changes. Conventional questionnaires and psychometric instruments remain important for assessing perceived stress, workload, fatigue, and burnout; however, they provide discrete self-reported observations and are not designed for continuous physiological monitoring. Wearable sensing provides a complementary approach by enabling repeated measurement of autonomic responses during daily activity, although its practical value depends on signal integrity, motion robustness, and appropriate interpretation of the measured physiological variables.

Heart-rate variability (HRV), conventionally derived from normal-to-normal R–R intervals in electrocardiography (ECG), is widely investigated as an indicator of autonomic regulation. Occupational stress has been associated with alterations in HRV, including reductions in parasympathetically mediated indices such as the root mean square of successive differences (RMSSD), although reported effects vary across populations, tasks, recording durations, and experimental protocols [1], [2]. Continuous ECG acquisition can nevertheless be intrusive for routine workplace monitoring. Wrist-worn devices offer a less obtrusive alternative by estimating cardiac timing from photoplethysmography (PPG), which measures peripheral blood-volume changes. The beat-to-beat variability obtained from PPG is more appropriately termed pulse-rate variability (PRV). Although PRV can approximate some ECG-derived HRV characteristics under controlled conditions, the two measures are not physiologically interchangeable. Motion artifacts, pulse-transit dynamics, sensor-skin contact, sampling frequency, and peripheral vascular responses can perturb pulse timing and degrade PRV estimation [3]. These effects are particularly important in occupational settings, where movement is intrinsic to many work activities.

A second challenge concerns the conventional machine-learning assumption that every acquired physiological window should yield a prediction. A wearable stress-monitoring system may encounter windows in which the PPG waveform is severely contaminated even though the classifier remains numerically confident. Treating such predictions equivalently to those obtained from high-quality measurements can undermine reliability. Accelerometry provides information about movement associated with PPG corruption, while explicit signal-quality measures can characterize whether a physiological window contains sufficient cardiac information for inference. This motivates a quality-aware formulation in which physiological features, motion context, signal quality, and predictive uncertainty are considered jointly, and the system is permitted to abstain when measurement quality or model confidence is inadequate.

Machine-learning approaches to condition monitoring similarly emphasize the importance of extracting reliable state information from imperfect sensor measurements. Deep-learning-based signal classification [4], physics-aware estimation [5], and engineered signal-selective sensing structures [6] illustrate this broader principle in non-physiological monitoring applications. While these studies do not provide evidence for physiological stress detection, they motivate the general systems perspective adopted in this work: sensing, signal conditioning, feature representation, inference, and validation should be treated as interconnected components rather than independent processing stages.

Despite increasing interest in wearable stress recognition, three related questions therefore remain important for deployment-oriented systems: whether wrist PPG-derived PRV preserves ECG-referenced variability sufficiently under realistic signal conditions; whether incorporating motion and explicit signal-quality information improves subject-independent stress discrimination; and whether uncertainty-aware abstention can improve prediction reliability by rejecting unreliable measurements. Addressing these questions requires evaluation at both the physiological-signal and machine-learning levels rather than classification accuracy alone.

Accordingly, this study develops and empirically evaluates a motion- and quality-aware PPG framework using the Wearable Stress and Affect Detection (WESAD) dataset [7], [8]. Wrist PPG, tri-axial accelerometry, and reference chest ECG are analyzed under a leave-one-subject-out protocol to prevent subject-level information leakage. The study makes four principal contributions: **(1)** ECG-referenced quantification of agreement between PPG-derived PRV and ECG-derived HRV, including heart rate, mean inter-beat interval, SDNN, and RMSSD; **(2)** a controlled ablation analysis that isolates the contributions of conventional PPG processing, accelerometry, and explicit signal-quality descriptors to subject-independent stress classification; **(3)** a selective-prediction formulation that jointly applies signal-quality and probabilistic-confidence criteria and quantifies the resulting coverage–risk trade-off; and **(4)** subject-level analysis of performance variability to expose failure modes that may be obscured by aggregate classification metrics.

WESAD is used here as a controlled laboratory benchmark for **acute stress-state discrimination**, rather than as a surrogate for occupational burnout or a clinical diagnostic instrument. The resulting analysis therefore focuses on the technical validity and reliability of the wearable inference framework while establishing the methodological basis for subsequent evaluation in free-living occupational environments.

II. MATERIALS AND METHODS

A. Dataset and Experimental Protocol

Experiments were conducted using the Wearable Stress and Affect Detection (WESAD) dataset [7], [8], which provides temporally synchronized physiological and inertial measurements acquired from 15 participants. The present study used wrist blood-volume pulse (BVP) sampled at 64 Hz, wrist tri-axial acceleration (ACC) at 32 Hz, and chest electrocardiography (ECG) at 700 Hz. ECG served exclusively as the cardiac reference for evaluating PPG-derived pulse-rate variability (PRV), whereas the experimentally assigned WESAD condition labels provided the reference for stress-state classification.

The analysis was restricted to the baseline (label 1) and stress (label 2) conditions, thereby defining a binary classification problem. Each recording was segmented into 60-s windows with a 30-s stride. To minimize contamination at transitions between experimental conditions, a window was retained only when at least 95% of its corresponding label samples belonged to a single target class. After segmentation and label-purity filtering, 865 windows remained, comprising 559 baseline and 306 stress windows.

Generalization was evaluated using leave-one-subject-out (LOSO) cross-validation. For each fold, all windows belonging to one participant were reserved for testing, while the remaining participants formed the training set. Consequently, no window from the held-out participant contributed to feature preprocessing, quality-threshold estimation, model fitting, or parameter estimation. This subject-disjoint protocol prevents window-level leakage and provides a stringent estimate of generalization to previously unseen individuals.

TABLE I
EXPERIMENTAL CONFIGURATION

Parameter	Setting
Participants	15
Analysis windows	865 (559 baseline; 306 stress
Wrist BVP	64 Hz
Wrist ACC	32 Hz
Chest ECG	700 Hz
Window / step	60 s / 30 s
Validation	LOSO

B. Cardiac Signal Processing and Variability Extraction

Wrist BVP and reference ECG were processed independently to preserve the distinction between pulse-derived and electrically derived cardiac timing. For the conventional PPG pipeline, BVP was detrended and band-pass filtered within the cardiac frequency range before pulse-peak localization. Detected pulse intervals were subjected to physiological plausibility screening according to

$$0.35 \le I_i \le 1.50\ \mathrm{s} \tag{1}$$

where I_i denotes the ith pulse-to-pulse interval. Residual interval artifacts were suppressed using a robust median-absolute-deviation criterion. The ECG channel was independently band-pass filtered before R-peak detection, followed by equivalent interval plausibility and outlier screening.

For an accepted sequence of N inter-beat intervals $\{I_i\}$ from i = 1 to N, mean interval, heart rate, SDNN, RMSSD, pNN50, and coefficient of variation were extracted. RMSSD and SDNN were computed as

$$RMSSD = \sqrt{\left[\frac{1}{N-1}\Sigma(\mathrm{i}=1)^{\mathrm{N}-1}(\mathrm{I}(i+1)-I_i)^2\right]} \tag{2}$$

And

$$\mathrm{SDNN} = \sqrt{\left[\frac{1}{\mathrm{N}-1}\Sigma_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{I_i}-\bar{\mathrm{I}})^2\right]} \tag{3}$$

where $\bar{\mathrm{I}}$ denotes the mean inter-beat interval. Features calculated from wrist BVP are referred to throughout as **PRV**, whereas corresponding ECG-derived quantities are designated **HRV**. This terminology avoids assuming physiological equivalence between pulse-to-pulse and R–R interval variability and permits their agreement to be evaluated explicitly.

C. Motion and Signal-Quality Characterization

Wrist accelerometry was incorporated to characterize movement likely to degrade optical pulse measurements. For the tri-axial acceleration vector $\mathrm{a(t)} = [\mathrm{a_{x(t)}}, \mathrm{a_{y(t)}}, \mathrm{a_{z(t)}}]^{\mathrm{T}}$ the instantaneous acceleration magnitude was calculated as

$$\mathrm{A(t)} = \sqrt{[\mathrm{a_x^2(t)} + \mathrm{a_y^2(t)} + \mathrm{a_z^2(t)}]} \tag{4}$$

Five motion descriptors were extracted from each analysis window: mean acceleration magnitude, standard deviation, root-mean-square magnitude, mean absolute jerk, and the 95th percentile of acceleration magnitude.

A window-level signal-quality index q ∈ [0,1] was subsequently constructed from four complementary indicators: cardiac-band spectral concentration, pulse-interval regularity, physiologically plausible beat yield, and an accelerometer-derived motion penalty. Higher values therefore represent windows with stronger cardiac periodicity, more regular pulse timing, greater plausible-beat recovery, and lower motion contamination.

To prevent information from the held-out participant from influencing rejection decisions, the quality threshold was recomputed independently within each LOSO fold using training data only. Specifically,

$$\tau_q = Q_{0.20(\{q_i:\, i \in D_{train}\})} \quad (5)$$

where $Q_{0.20}$ denotes the 20th percentile of the training-window quality-score distribution and D_{train} denotes the training set for the corresponding LOSO fold. A test window satisfied the quality criterion when $q_i \geq \tau_q$.

D. Machine-Learning Configurations and Selective Prediction

Five feature configurations were evaluated to isolate the contributions of signal preprocessing, motion information, and explicit signal-quality characterization: **Raw PRV, Conventional PRV, PPG+ACC, Quality-Aware PRV, and Full**. Raw PRV used variability descriptors extracted without the conventional filtered-PPG pathway. Conventional PRV used variability descriptors obtained from the processed PPG signal. PPG+ACC augmented conventional PRV with the five acceleration-derived motion descriptors. Quality-Aware PRV combined conventional PRV with signal-quality descriptors, whereas the Full configuration jointly incorporated PRV, acceleration, and quality information.

Within each LOSO fold, missing feature values were median-imputed and predictors were standardized using parameters estimated exclusively from the training participants. Classification was performed using a histogram-based gradient-boosting classifier. For feature vector z_i , the classifier produced the posterior probability vector

$$\hat{p}_i = f_{\theta(z_i)} = [\hat{p}_{i(y=0)}, \hat{p}_{i(y=1)}] \quad (6)$$

where y = 0 represents baseline and y = 1 represents stress. The nominal class prediction was obtained from the class having the maximum posterior probability.

Prediction confidence was defined as

$$c_i = \max_y \hat{p}_{i(y)} \quad (7)$$

The Full framework was additionally evaluated as a selective classifier. A prediction was retained only when both signal quality and probabilistic confidence satisfied their respective acceptance criteria:

$$g_{i(\tau_q,\tau_c)} = \mathbb{I}(q_i \geq \tau_q)\mathbb{I}(c_i \geq \tau_c) \quad (8)$$

where $\mathbb{I}(\cdot)$ is the indicator function, τ_q is the training-derived signal-quality threshold, and τ_c is the confidence threshold. The confidence threshold was varied from 0.50 to 0.95 to characterize the trade-off between prediction availability and reliability. This formulation permits the system to abstain when either the physiological input or model confidence is insufficient rather than forcing a potentially unreliable binary decision.

E. Performance Evaluation and Statistical Analysis

Performance was evaluated at three levels: physiological agreement, stress-state classification, and selective prediction. Agreement between PPG-derived PRV and ECG-derived HRV was quantified separately for heart rate, mean inter-beat interval, SDNN, and RMSSD using mean absolute error (MAE), mean bias, Pearson correlation coefficient, and Bland–Altman 95% limits of agreement. The limits of agreement were calculated as

$$LoA = \bar{d} \pm 1.96 s_d \quad (9)$$

where $\bar{d}$ is the mean paired PRV–HRV difference and s_d is the standard deviation of the paired differences.

Stress-state classification was evaluated using balanced accuracy, Macro-F1, area under the receiver-operating-characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and Brier score. Balanced accuracy and Macro-F1 were emphasized because the final analysis corpus was class-imbalanced, containing 559 baseline and 306 stress windows.

For selective prediction, **coverage** was defined as the proportion of test windows for which a prediction was retained:

$$C(\tau_c) = \left(\frac{1}{N}\right)\sum_{i=1}^{N} g_i \quad (10)$$

where N is the total number of evaluated test windows. **Selective risk** was defined as the classification error conditional on prediction acceptance:

$$R(\tau_c) = \frac{[\sum(i=1)^N g_i \mathbb{I}(\hat{y}_i \neq y_i)]}{[\sum(i=1)^N g_i]} \quad (11)$$

where $\hat{y}_i$ and y_i denote the predicted and reference class labels, respectively. The resulting coverage–risk relationship therefore quantifies the reliability gained by withholding progressively lower-confidence predictions.

Model comparisons were conducted using paired LOSO subject-level metrics and two-sided Wilcoxon signed-rank tests, thereby treating the participant rather than individual windows as the unit of statistical comparison. Given the cohort size of 15 participants, statistical tests were interpreted as evidence for specific paired performance differences rather than as proof of universal model superiority. Between-subject dispersion was retained explicitly to characterize inter-individual variability in wearable stress-state inference.

III. RESULTS

A. ECG-Referenced PRV Validity

Agreement depended strongly on the physiological quantity. Heart rate showed Pearson r=0.841 with MAE 5.50 beats/min and bias -2.37 beats/min. Mean inter-beat interval showed r=0.829 and MAE 0.0476 s. Variability measures were substantially less concordant: SDNN yielded r=0.388 and MAE 0.0982 s, whereas RMSSD yielded r=0.209 and MAE 0.1669 s. For RMSSD, the PRV-HRV bias was +0.1669 s with 95% limits of agreement from -0.0208 to 0.3545 s. Thus, pulse-rate tracking was considerably more reliable than short-window variability equivalence.

TABLE II
ECG-HRV AND PPG-PRV AGREEMENT

Metric	MAE	Bias	r
Heart rate	5.501 bpm	-2.367 bpm	0.841
Mean IBI	0.0476 s	0.0135 s	0.829
SDNN	0.0982 s	0.0981s	0.388
RMSSD	0.1669 s	0.1669 s	0.209

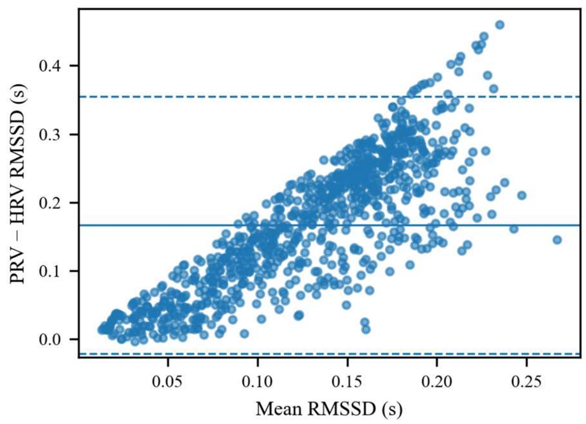

Fig. 1. Bland-Altman agreement between PPG-derived PRV and ECG-derived HRV for RMSSD.

B. Subject-Independent Stress Classification

Across LOSO folds, Raw PRV produced the largest mean thresholded Macro-F1 (0.8379) and balanced accuracy (0.8544). PPG+ACC produced the strongest discrimination, with AUROC 0.9560 and AUPRC 0.9330. The Full configuration achieved Macro-F1 0.8303, balanced accuracy 0.8449, AUROC 0.9524, AUPRC 0.9295, and the lowest mean Brier score (0.1164). Therefore, adding motion and quality information improved ranking and probability quality more clearly than fixed-threshold F1.

TABLE III
SUBJECT-INDEPENDENT LOSO PERFORMANCE

Model	BAcc	F1	AUROC	AUPRC	Brier
Raw PRV	0.8544	0.8379	0.9379	0.8990	0.1231
Conventional	0.8532	0.8362	0.9283	0.8984	0.1208
PPG+ACC	0.8490	0.8341	0.9560	0.9330	0.1165
Quality-Aware	0.8404	0.8238	0.9357	0.8962	0.1232
Full	0.8449	0.8303	0.9524	0.9295	0.1164

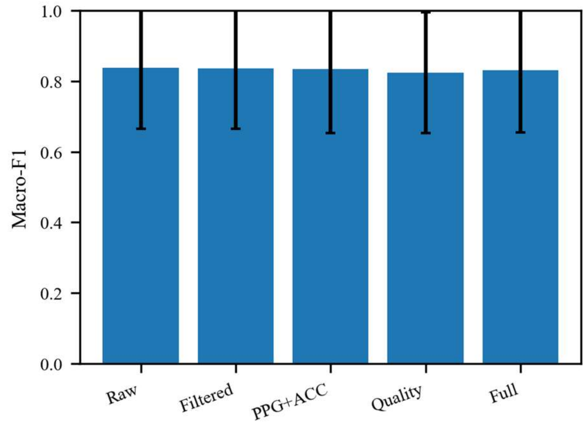

Fig. 2. Mean subject-independent LOSO Macro-F1 across model configurations; error bars denote between-subject standard deviation.

C. Paired Model Comparisons

The Full model did not significantly change Macro-F1 relative to Raw PRV (mean paired difference -0.0076, p=0.599), Conventional PRV (-0.0059, p=0.790), PPG+ACC (-0.0038, p=0.285), or Quality-Aware PRV (+0.0064, p=0.701). In contrast, Full-model AUROC exceeded Conventional PRV by 0.0241 (p=0.0107) and Quality-Aware PRV by 0.0167 (p=0.0355). AUPRC also exceeded Conventional PRV by 0.0311 (p=0.0413) and Quality-Aware PRV by 0.0333 (p=0.0355). Full and PPG+ACC were not significantly different in AUROC or AUPRC. These results indicate that accelerometry accounts for most of the observed discrimination gain, whereas explicit quality features primarily support rejection and reliability management.

D. Selective Prediction and Coverage-Risk Trade-off

Quality gating alone retained approximately 79.7% of windows. When the Full model used both the quality gate and increasing probability-confidence thresholds, reliability improved monotonically as coverage decreased. At confidence threshold 0.50, 690 windows were accepted (79.77% coverage), selective risk was 13.33%, accepted-window Macro-F1 was 0.8346, and balanced accuracy was 0.8548. At threshold 0.95, 535 windows were accepted (61.85% coverage), selective risk fell to 7.48%, Macro-F1 increased to 0.8885, and balanced accuracy increased to 0.9171. Thus, abstention reduced error by 5.85 percentage points while rejecting an additional 17.93% of all windows.

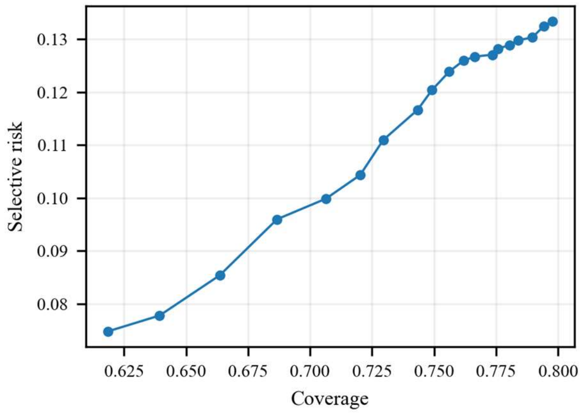

Fig. 3. Coverage-risk relationship for the full quality- and confidence-aware framework.

E. Between-Subject Heterogeneity

Performance varied substantially across held-out participants. Full-model Macro-F1 ranged from 0.325 for S10 to 0.982 for S14, with corresponding subject-to-subject variation reflected in the large standard deviations of Table III. This heterogeneity is operationally important: a strong cohort-average AUROC does not guarantee reliable performance for every new wearer. Subject-level calibration, signal-quality monitoring, and prospective validation are therefore necessary before occupational deployment.

IV. DISCUSSION

The results demonstrate that reliable PPG-based stress-state monitoring depends not only on classification performance but also on the validity of the physiological features, the availability of motion information, and the ability to identify uncertain predictions. Three findings are particularly important.

First, wrist PPG provided useful information for subject-independent discrimination between baseline and laboratory stress states, but the agreement between PPG-derived PRV and ECG-derived HRV was strongly feature dependent. Heart rate and mean inter-beat interval showed comparatively strong ECG-referenced correlations ($r = 0.841$ and $r = 0.829$, respectively), whereas agreement deteriorated substantially for SDNN ($r = 0.388$) and RMSSD ($r = 0.209$). RMSSD additionally exhibited a positive mean bias of 0.167 s and wide 95% limits of agreement, indicating substantial disagreement in short-term beat-to-beat variability. Thus, preservation of average cardiac timing does not imply accurate reconstruction of variability dynamics. This distinction is important because RMSSD and related indices are frequently interpreted as markers of parasympathetic modulation in stress research [1], [2]. The present findings therefore reinforce the need to distinguish PRV from ECG-derived HRV and to validate individual variability measures before assigning them equivalent physiological interpretations.

Second, increasing feature-set complexity did not produce a corresponding improvement in fixed-threshold classification performance. Raw PRV achieved the highest mean Macro-F1 (0.8379) and balanced accuracy (0.8544), although the differences among the evaluated configurations were small relative to the observed between-subject variability. In contrast, the PPG+ACC configuration achieved the highest AUROC (0.9560) and AUPRC (0.9330), while the Full model achieved comparable discrimination (AUROC = 0.9524; AUPRC = 0.9295) and the lowest Brier score (0.1164). These results distinguish threshold-dependent classification from probabilistic discrimination: Macro-F1 evaluates decisions at a particular operating threshold, whereas AUROC and AUPRC characterize the model's ability to rank stress and baseline observations across thresholds.

The paired subject-level analysis further clarifies this distinction. No statistically significant Macro-F1 difference was observed between the Full configuration and any of the four alternatives ($p > 0.05$). However, the Full model produced significantly higher AUROC than Conventional PRV ($p = 0.0107$) and Quality-Aware PRV ($p = 0.0355$), together with significantly higher AUPRC than these configurations ($p = 0.0413$ and $p = 0.0355$, respectively). No corresponding discrimination advantage was observed over

PPG+ACC. Taken together, these findings suggest that accelerometry contributes useful contextual information beyond PRV alone, whereas adding explicit quality information does not necessarily improve unconditional classification accuracy. Its more consequential role emerges when signal quality is incorporated into the decision process itself.

Third, selective prediction provided a direct mechanism for controlling operating reliability. For the Full framework, applying the quality criterion with a confidence threshold of 0.50 retained 79.77% of windows with a selective risk of 13.33%. Increasing the confidence requirement to 0.95 reduced coverage to 61.85% but decreased selective risk to 7.48%; over the same operating range, Macro-F1 among accepted windows increased from 0.8346 to 0.8885. The improvement was therefore obtained by withholding uncertain predictions rather than by assuming that every physiological window was sufficiently reliable for inference. This behavior is particularly relevant to wearable monitoring, where motion artifacts, poor sensor contact, and transient physiological disturbances may produce measurements for which a definitive classification is inappropriate. The coverage-risk relationship makes this trade-off explicit and allows an operating threshold to be selected according to the relative consequences of unavailable predictions and erroneous alerts.

Performance nevertheless varied substantially across participants. For the Full model, subject-level Macro-F1 ranged from 0.325 to 0.982, despite considerably stronger aggregate performance. Such dispersion indicates that population-average metrics can obscure important individual failure modes and reinforces the importance of subject-disjoint validation. Inter-individual differences in physiological response, peripheral pulse morphology, movement, and sensor coupling may all contribute to this variability. Consequently, high aggregate discrimination should not be interpreted as evidence of uniformly reliable performance across users. Future systems may require individualized calibration, adaptive decision thresholds, or subject-adaptive representations to reduce this variability.

The engineering condition-monitoring studies in [4] and [5] provide methodological parallels rather than biomedical evidence. Their relevance lies in the broader requirement to transform imperfect sensor measurements into reliable state estimates while maintaining defensible signal processing and out-of-sample validation. Similarly, the signal-selective sensing principles illustrated in [6] provide an engineering analogy for treating sensing and inference as components of an integrated measurement system. The physiological conclusions of the present study, however, are derived from the WESAD experiments and ECG-referenced analysis rather than from these non-biomedical applications.

A. Occupational Interpretation and Limitations

The findings should be interpreted within the experimental scope of WESAD. The dataset provides a controlled laboratory stress paradigm and therefore does not reproduce the full range of conditions encountered during occupational activity, including prolonged shifts, task-dependent movement, thermal exposure, changing posture, variable sensor-skin coupling, sleep restriction, and cumulative physical or cognitive fatigue. Accordingly, the present results establish performance for laboratory **stress-state discrimination**, not occupational burnout detection. Burnout is a multidimensional and longitudinal construct that requires appropriate psychometric and contextual assessment; it should not be inferred directly from short PPG windows.

Several additional limitations warrant consideration. The analysis included only 15 participants and exhibited substantial between-subject performance variability. The baseline and stress classes were also imbalanced, with 559 and 306 windows, respectively. Although LOSO evaluation prevented windows from the held-out participant from entering model development, the overlapping 60-s windows are temporally correlated within each participant. Furthermore, the signal-quality index was constructed from predefined physiological, spectral, and motion indicators rather than learned or independently validated as a standardized PPG quality measure. WESAD was also collected under controlled experimental conditions and does not provide an independent occupational cohort for external validation. These factors limit the extent to which the reported performance can be generalized to unconstrained workplace monitoring.

Future validation should therefore move from laboratory benchmarking to prospective occupational evaluation. A field study should acquire synchronized wrist PPG and accelerometry during complete work shifts, with reference ECG collected during selected periods and ecological momentary assessments used to capture perceived stress and fatigue. Task type, physical activity, recovery periods, and relevant environmental conditions should also be recorded to distinguish physiological changes associated with psychological stress from those produced by physical exertion or context. Particular attention should be given to subgroup and participant-level error patterns, rejection rates, calibration stability, and the effect of sensor-quality variation under free-living conditions.

Finally, deployment should preserve the distinction between **physiological monitoring and diagnosis**. Short-window wearable outputs are more appropriately interpreted as estimates of acute stress-related physiological state or strain than as direct measures of burnout, clinical stress disorders, or worker capability. In occupational applications, selective predictions and longitudinal summaries could support voluntary worker-centered health and recovery monitoring, but the present evidence does not support their use for clinical diagnosis, individual productivity assessment, or automated employment decisions.

V. CONCLUSIONS

This study investigated the reliability of PPG-based stress-state monitoring by jointly examining physiological validity, subject-independent classification, motion information, signal quality, and selective prediction. Evaluation on 15 WESAD participants and 865 analysis windows showed that increasing model complexity did not uniformly improve classification performance. Raw PRV produced the highest mean Macro-F1 of 0.8379, whereas incorporating accelerometry improved discriminative performance, with PPG+ACC achieving the highest AUROC and AUPRC of 0.9560 and 0.9330, respectively. The Full configuration maintained comparable discrimination (AUROC = 0.9524) and achieved the lowest Brier score of 0.1164, indicating improved probabilistic reliability despite not maximizing threshold-dependent classification performance.

A key finding was the benefit of allowing the monitoring system to abstain from unreliable predictions. For the Full framework, increasing the confidence requirement reduced coverage from 79.77% to 61.85% while decreasing selective risk from 13.33% to 7.48%. This demonstrates that prediction availability and reliability can be explicitly balanced rather than forcing a stress-state decision from every physiological window. ECG-referenced analysis further showed that the validity of PPG-derived PRV is feature dependent: heart-rate and mean-interval estimates exhibited substantially stronger agreement with ECG than variability measures such as RMSSD. PRV should therefore not be treated as directly interchangeable with ECG-derived HRV, particularly for short-term variability assessment.

Overall, the results support a **quality- and uncertainty-aware approach to wearable stress-state monitoring** in which motion context, physiological signal validity, and prediction confidence are considered alongside classification performance. The findings establish technical feasibility under controlled laboratory conditions but do not constitute evidence for occupational burnout detection or clinical diagnosis. Future work should evaluate the framework prospectively in free-living occupational environments, particularly among shift workers, and investigate subject adaptation, motion-intensive activities, sensor-quality variation, and longitudinal measures of fatigue and recovery before practical workplace deployment.

REFERENCES

[1] S. Järvelin-Pasanen, S. Sinikallio, and M. P. Tarvainen, “Heart rate variability and occupational stress—systematic review,” Industrial Health, vol. 56, no. 6, pp. 500–511, 2018, doi: 10.2486/indhealth.2017-0190.

[2] L. Kane, D. Powell, K. R. Martin, C. Rees, J. Curran, and D. Ball, “Continuous heart rate variability monitoring, stress and recovery in doctors: a systematic review and meta-analysis,” Occupational Medicine, vol. 75, no. 9, pp. 630–639, 2025, doi: 10.1093/occmed/kqaf101

[3] S. Ismail, U. Akram, and I. Siddiqi, “Heart rate tracking in photoplethysmography signals affected by motion artifacts: a review,” EURASIP Journal on Advances in Signal Processing, vol. 2021, Art. no. 5, 2021, doi: 10.1186/s13634-020-00714-2.

[4] O. A. Adegboye, T. F. Komolafe, I. M. Joseph, and O. O. Kajero, “Electromagnetic Signal Classification Using Deep Learning for Pipeline Structural Health Monitoring,” International Journal of Scientific Research and Engineering

Development, vol. 7, no. 6, pp. 1317–1322, Nov.–Dec. 2024.

[5] O. A. Adegboye, T. F. Komolafe, and I. M. Joseph, “Physics-Aware ML Models for Pipeline Leak Detection and Quantification,” International Journal of Scientific Research and Engineering Development, vol. 7, no. 6, pp. 1311–1316, Nov.–Dec. 2024.

[6] O. A. Adegboye, K. M. Udofia, and A. Obot, “Inverted C-Shaped Slots Loaded Exponential Tapered Triple Band Notched Ultra Wideband (UWB) Antenna,” International Journal of Engineering Research & Technology, vol. 12, no. 4, Apr. 2023.

[7] P. Schmidt and A. Reiss, “WESAD (Wearable Stress and Affect Detection),” UCI Machine Learning Repository, 2018, doi: 10.24432/C57K5T.

[8] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing WESAD, a multimodal dataset for wearable stress and affect detection,” in Proc. 20th ACM Int. Conf. Multimodal Interaction, 2018.

