

Validation design changes wearable stress-state classification performance in hospital nurses

Fuat Kahraman

Iğdır Üniversitesi

Gözde Özsezer

gozdeozsezer@hotmail.com

Çanakkale Onsekiz Mart University

Gülengül Mermer

Iğdır Üniversitesi

Article

Keywords: wearable sensors, stress classification, machine learning, model validation, electrodermal activity, explainable artificial intelligence

Posted Date: September 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10535611/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Wearable-sensor classifiers can appear highly accurate when multiple windows from the same person contribute to both model development and evaluation. We examined how validation design affected classification of three wearable-derived stress states in a public dataset collected from 15 female hospital nurses. Electrodermal activity, heart rate, skin temperature and triaxial acceleration were represented by 24 features from 60-second windows. Retained internal window-level estimates were available for six algorithms, while leave-one-identifier-out evaluation was available for XGBoost. Internal accuracy ranged from 0.810 to 0.990. XGBoost accuracy decreased from 0.948 internally to a mean of 0.661 when an identifier was held out, with approximate identifier-level values of 0.32–0.90. Electrodermal-activity features had the largest SHAP contributions; removing this modality reduced accuracy by 0.0606. The source cohort comprised 15 nurses, whereas the merged redistribution contained 18 identifiers, and the labels originated from algorithm-proposed events subsequently reviewed by nurses. The findings show that internal window-level performance does not establish transportability to an unseen nurse. Independent outcomes, reproducible participant-level evaluation, calibration and external validation are required before operational use.

Introduction

Wearable sensors can capture physiological and behavioural signals continuously in naturalistic care environments. Electrodermal activity, cardiac signals, peripheral temperature and movement have been studied as low-burden indicators of arousal and occupational strain in nurses and other populations [1–7]. Their value depends on whether the target labels are valid, model performance transports to new users and settings, and outputs can support an appropriate decision without creating unacceptable surveillance or workload.

Dense sensor streams create a distinctive evaluation problem. Thousands of adjacent windows can be generated from one person, and these windows share physiology, device placement, environmental exposures and event context. If windows from the same person contribute to model fitting and evaluation, the resulting estimate may show recognition of later segments from known users rather than performance for a previously unseen user. Health-prediction studies similarly show that model performance can deteriorate under transportability-focused evaluation [8, 9].

Previous analyses of the public nurse-stress dataset have emphasised model discrimination and explainability [1, 5]. A decision-support interpretation additionally requires transparent label provenance, participant-independent validation, calibration, uncertainty and task–technology fit [10, 11]. Without these elements, high internal accuracy can be mistaken for deployment readiness.

We therefore examined how apparent classification performance changed between internal time-aware window evaluation and a leave-one-identifier-out design. Secondary objectives were to compare retained internal results across six algorithm families, examine feature dependence using SHAP and sensor-

group ablation, and assess whether exploratory temporal maps supported a consistent nominal day–night pattern. We expected identifier-held-out performance to be lower and more heterogeneous.

Results

Dataset and analytic provenance

The source cohort comprised 15 nurses, whereas 18 identifier values appeared in the merged redistribution. Because the mapping and any exclusions were not retained, held-out results are reported as identifier-level estimates rather than assumed to represent 18 independent nurses. The available materials did not support a participant flow diagram or an exact post-windowing sample size. Table 1 summarises the data provenance and its interpretive consequences.

Table 1
Data source and analysis provenance.

Element	Description	Interpretive consequence
Source cohort	15 female hospital nurses, approximately 30–55 years old, observed during the COVID-19 pandemic.	Small, single-site occupational cohort with limited demographic and institutional transportability.
Recording	Empatica E4 electrodermal activity, heart rate, skin temperature, and triaxial acceleration.	Naturalistic, time-resolved signals; physiological arousal is not specific to psychological stress.
Analysis table	Public merged redistribution with approximately 11.5 million one-second rows and nine fields.	The redistribution contains 18 identifier values although the source report describes 15 nurses.
Outcome	Three wearable-derived stress-state labels linked to candidate events reviewed after shifts.	The target is not an independent clinical diagnosis and may partly incorporate predictor information.
Predictors	EDA, HR, skin temperature, and X/Y/Z acceleration; identifiers and time used for grouping and ordering.	Conclusions are specific to these modalities, preprocessing choices, and the retained outcome.
Note. EDA, electrodermal activity; HR, heart rate. The identifier discrepancy should be reconciled against the source data before a definitive reproducible reanalysis.		

Internal model performance

Internal window-level accuracy ranged from 0.810 for the multilayer perceptron to 0.990 for the long short-term memory and one-dimensional convolutional models. The combination of near-perfect accuracy with only moderate AUROC and unavailable multiclass averaging conventions limits ranking across algorithms. The estimates are therefore descriptive rather than evidence that the neural architectures were superior (Table 2).

Table 2
Internal window-level model performance.

Model	Accuracy	Precision	Recall	F1 score	AUROC
Random forest	0.950	0.950	0.950	0.950	0.842
XGBoost	0.948	0.948	0.948	0.948	0.850
LightGBM	0.931	0.930	0.931	0.930	0.846
Multilayer perceptron	0.810	0.810	0.810	0.810	0.821
Long short-term memory	0.990	0.990	0.990	0.990	0.857
1D convolutional neural network	0.990	0.990	0.993	0.993	0.864
Note. AUROC, area under the receiver operating characteristic curve. Averaging conventions, confidence intervals and fold-specific values were not retained.					

Identifier-held-out performance

XGBoost achieved internal accuracy of 0.948 but mean leave-one-subject-out accuracy of 0.661, an absolute difference of 0.287. Approximate identifier-specific accuracy ranged from 0.32 to 0.90. Because comparable held-out results were unavailable for the other five algorithms, the analysis does not identify the best model for a new nurse; it shows that validation level materially changed the apparent performance of the one model evaluated at both levels (Table 3 and Fig. 2).

Table 3
Validation-level performance contrast.

Validation level	Unit held out	Model(s)	Accuracy	Interpretation
Internal time-aware window validation	Later windows in the merged sequence	Six models	0.810–0.990	Secondary benchmark; windows from the same nurse may occur in development and evaluation partitions.
Leave-one-subject-out validation	All windows for one identifier	XGBoost	Mean 0.661; approximate range 0.32–0.90	Primary estimate for a previously unseen identifier; substantial between-identifier heterogeneity.
Note. LOSO, leave-one-subject-out. The identifier-held-out range is approximate because exact fold values were recovered from a retained figure rather than a machine-readable results table.				

Model explanation and sensor-group ablation

Electrodermal-activity features occupied the leading positions in the retained SHAP summary, followed by skin temperature and selected movement features; heart-rate contributions were smaller (Fig. 3). Removing electrodermal activity reduced accuracy by 0.0606, and removing skin temperature reduced it by 0.0455. Removing heart rate or any single accelerometer axis reduced accuracy by less than 0.009 (Table 4 and Fig. 4). These findings indicate which modalities the fitted pipeline relied on most, not which physiological signals caused stress.

Table 4
Sensor-group ablation results.

Feature group removed	Accuracy after removal	Accuracy reduction	Interpretation
Electrodermal activity	0.8111	0.0606	Largest reduction
Skin temperature	0.8262	0.0455	Second-largest reduction
Acceleration X	0.8630	0.0087	Small reduction
Acceleration Z	0.8647	0.0070	Small reduction
Heart rate	0.8653	0.0064	Small reduction
Acceleration Y	0.8659	0.0058	Small reduction
Note. The supplied baseline accuracy for the ablation pipeline was approximately 0.8717. Accuracy reduction is the baseline minus accuracy after removal.			

Exploratory temporal patterns

The identifier-by-nominal-period map showed heterogeneous patterns and missing identifier-period combinations. Without verified rosters and within-person inference, the retained outputs did not support claims that stress was consistently higher during night shifts or between 12:00 and 18:00. The map is retained in the Supplementary Information as an exploratory visualisation only.

Discussion

Validation design changed the apparent performance of wearable stress-state classification. Internal accuracy of up to 0.990 contrasted with mean identifier-held-out XGBoost accuracy of 0.661, and held-out results varied markedly between identifiers. Electrodermal activity and skin temperature were the modalities on which the retained XGBoost pipeline depended most. These findings support the feasibility of learning the supplied labels from sensor windows but do not establish reliable performance for a new nurse, measurement of a clinical stress construct or benefit from using model outputs in practice.

Dense wearable data provide many observations but few independent people. A chronological split may reduce direct future-to-past leakage while still allowing a model to exploit stable person-specific signatures. Holding out an entire identifier is closer to the intended use of applying a generic model to a

previously unseen person. The observed decline is consistent with broader evidence that prediction performance often deteriorates across populations and settings [8, 9]. The present held-out evaluation nevertheless remained within one source and institutional context and is not external validation.

The labels also require careful interpretation. Physiological arousal may reflect physical activity, ambient conditions, hydration, illness, medication, device contact or emotional strain [2–4]. Candidate events were proposed using related wearable information before participant review, creating a possible incorporation pathway that may make sensor-based labels easier to reproduce than an independently measured psychological outcome. SHAP and ablation clarify model dependence but do not resolve construct validity or establish physiological causation.

A wearable classifier could eventually support voluntary self-monitoring, research feedback or aggregated organisational learning. It should not be used as an autonomous alarm, staffing rule, fitness-for-duty assessment or performance-management tool on the basis of these results. Decision support requires calibrated probabilities, uncertainty, actionable responses, workflow fit, user understanding and evidence that actions improve outcomes [10]. Occupational deployment would also require meaningful worker choice, role-based access, auditable secondary use and safeguards against punitive interpretation.

The study is limited by the small all-female cohort from one hospital during the COVID-19 pandemic; a non-independent outcome; the unresolved difference between 15 source nurses and 18 redistributed identifiers; unavailable exact window counts, overlap, missingness, preprocessing objects, hyperparameters and seeds; identifier-held-out results for XGBoost only; absent confidence intervals, class-specific metrics, confusion matrices and calibration; unverified work rosters; and no independent site, prospective period, impact evaluation or usability study. These limitations preclude claims of clinical effectiveness or deployment readiness.

A definitive evaluation should preregister the intended use and primary metric, collect independent time-stamped outcomes, publish a participant-window flow diagram, fit all preprocessing and tuning inside each training fold, evaluate every candidate model under identifier- and site-held-out designs, and report uncertainty, macro and class-specific metrics, calibration and decision consequences. Executable code, software versions and model specifications should be archived. In conclusion, the retained internal estimates substantially overstated performance relevant to an unseen nurse; transportability and reproducibility should therefore take precedence over headline window-level accuracy.

Methods

Study design and reporting framework

This was a retrospective secondary computational benchmarking study of an openly available wearable-sensor dataset. The analytic objective was model-performance evaluation rather than causal inference or estimation of a clinical treatment effect. Reporting was structured with TRIPOD + AI for prediction-

model evaluation and STROBE for the observational source data [11, 12]. Because the outcome is a derived stress-state label rather than a diagnosis or prognosis, TRIPOD + AI items were applied pragmatically and non-applicable items were identified in the checklist.

Data source, setting, and participants

The analysis used a public merged redistribution of the multimodal dataset reported by Hosseini et al. [1, 13]. The source study recorded more than 1250 hours of data from 15 female nurses, approximately 30–55 years old, working in a hospital during the COVID-19 pandemic. Collection occurred in two periods: 15 April to 6 August 2020 and 8 October to 11 December 2020. Empatica E4 devices captured electrodermal activity (EDA), heart rate (HR), peripheral skin temperature, and triaxial acceleration under naturalistic working conditions. The merged redistribution contained approximately 11.5 million one-second rows and nine core fields. The source report described 15 participants, whereas the redistributed table contained 18 identifier values; no verified mapping between these representations was available in the retained materials.

Outcome definition and label provenance

The target was a three-level label representing low, medium, or high wearable-derived stress state. In the source workflow, candidate events were first proposed using an algorithm operating on wearable modalities and nurses subsequently reviewed up to six of the longest events after their shifts. No additional missed events were reported in the retained source description [1]. The label therefore combines algorithmic event proposal and participant review; it is not an independently ascertained clinical diagnosis. Because candidate generation used related sensor information, partial label–predictor incorporation is possible and internal discrimination may reflect replication of the labeling process as well as physiologic pattern recognition.

Predictors and preprocessing

Predictor modalities were EDA, HR, skin temperature, and acceleration on the X, Y, and Z axes. Participant identifiers and timestamps were used only for grouping and temporal ordering. Signals were represented at one-second resolution and segmented into 60-second windows. For each modality, retained analysis records described statistical summaries including mean, standard deviation, minimum, and maximum, producing 24 window-level features. The prior analysis described forward filling and time-series interpolation for missing values, z-score standardization, and class-imbalance handling using SMOTE or class weights [14]. Exact post-windowing counts, missingness by variable, overlap between windows, preprocessing parameters, and fold-specific transformation objects were not preserved. Consequently, these steps are reported as supplied and cannot be independently reconstructed from the archived manuscript outputs alone.

Candidate models

Three tree-based algorithms—random forest, XGBoost, and LightGBM—and three neural architectures—a multilayer perceptron, long short-term memory network, and one-dimensional convolutional neural

network—were evaluated. Retained outputs included aggregate internal accuracy, precision, recall, F1 score, and multiclass AUROC values. Exact hyperparameters, tuning search spaces, random seeds, software versions, early-stopping rules, and trained model objects were not available. Accordingly, the comparison is descriptive and does not establish algorithmic superiority.

Validation hierarchy and estimands

Two validation levels addressed different questions. First, a time-aware window-level split estimated performance on later windows in the merged sequence. This design reduced direct future-to-past leakage but could still place windows from the same nurse in development and evaluation partitions; it was therefore treated as a secondary internal benchmark. Second, leave-one-subject-out (LOSO) validation kept all windows for one identifier outside model fitting and repeated this process across identifiers. The retained LOSO analysis was available for XGBoost only and was designated the primary estimate of identifier-level transportability. Because no independent hospital or later prospective cohort was evaluated, LOSO is an internal-external evaluation within one dataset, not external validation.

Performance measures

Retained measures were accuracy, precision, recall, F1 score, and AUROC. The averaging convention for multiclass precision, recall, F1 score, and AUROC was not preserved. Confidence intervals, fold-level confusion matrices, class-specific sensitivity and specificity, calibration-in-the-large, calibration slope, Brier score, and decision-curve analysis were unavailable. No clinical or occupational decision threshold was defined. We therefore report point estimates without inferential comparisons and avoid claims of clinical utility.

Explainability and feature ablation

SHAP values were retained for the XGBoost model to describe how features contributed to model output [15]. Sensor-group ablation compared a baseline accuracy of approximately 0.8717 with accuracy after removing EDA, HR, skin temperature, or one acceleration axis. SHAP and ablation results describe fitted model behavior; they do not identify causal determinants of psychological stress or establish that a feature would improve downstream decisions.

Temporal mapping

Mean encoded labels were summarized by identifier and nominal day/night period in an exploratory heat map. Individual roster data were not available, and nominal day and night periods could not be linked to verified shifts. No repeated-measures model or prespecified temporal hypothesis test was available. Temporal maps were therefore interpreted descriptively and were not used to infer a night-shift or time-of-day effect.

Ethics and consent

The source data collection was approved by the University of Louisiana at Lafayette Institutional Review Board (FA19-50 INFOR), and source participants provided informed consent [1]. The present study used

only publicly available, deidentified data, involved no new recruitment or participant contact, and therefore did not require additional ethics committee review under the authors' applicable institutional policy.

Use of generative artificial intelligence

ChatGPT was used for English-language editing and structural revision of the manuscript, and Grammarly and QuillBot were used for language polishing. The authors reviewed and verified the resulting text and remain accountable for all content. Generative artificial intelligence was not used to generate data, perform statistical analyses, select observations, or create numerical results.

Data availability

The source dataset is publicly available in the Dryad Digital Repository at <https://doi.org/10.5061/dryad.5hqbzkh6f> [13]. The merged redistribution used for the retained secondary analysis is available at <https://www.kaggle.com/datasets/priyankraval/nurse-stress-prediction-wearable-sensors>. Retained model-output figures and result tables supporting the reported estimates are included in this article and its Supplementary Information. The mapping between the 15 source participants and 18 redistributed identifiers was not available and should not be inferred.

Code availability

Code associated with the source data collection and original stress-detection system is available at <https://github.com/CPHSLab/Stress-Detection-in-Nurses> [1]. The separate secondary analytical pipeline that generated the retained benchmarking outputs was not archived in the supplied materials. The reported numerical estimates are therefore traceable to retained tables and figures, but the complete secondary workflow is not computationally reproducible from the manuscript package alone.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information accompanies this paper. Correspondence and requests for materials should be addressed to Gözde Özsezer.

Funding

The authors received no specific funding for this secondary study.

Author Contribution

GÖ: conceptualization, data curation, formal analysis, methodology, visualization, writing—original draft, and project administration. FK: conceptualization, methodology, validation, and writing—review and editing. GM: conceptualization, supervision, validation, and writing—review and editing. All authors read and approved the final manuscript and agree to be accountable for the work.

Acknowledgement

The authors acknowledge the nurses who participated in the source study and the investigators who made the deidentified dataset publicly available.

Data Availability

The source dataset is publicly available in the Dryad Digital Repository at <https://doi.org/10.5061/dryad.5hqbzkh6f>. The analysis used a merged redistribution available at <https://www.kaggle.com/datasets/priyankraval/nurse-stress-prediction-wearable-sensors>. The archived executable code and a verified mapping between source participants and redistributed identifiers were not available in the materials used to prepare this report. Retained model-output figures and result tables supporting the reported estimates are included in the manuscript and Additional file 1.

References

1. Hosseini, S. et al. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Sci. Data*. **9**, 255. <https://doi.org/10.1038/s41597-022-01361-y> (2022).
2. Bolpagni, M., Pardini, S., Dianti, M. & Gabrielli, S. Personalized stress detection using biosignals from wearables: a scoping review. *Sensors* **24**, 3221. <https://doi.org/10.3390/s24103221> (2024).
3. Taskasaplidis, G. V., Fotiadis, D. I. & Exarchos, T. P. A review of wearable sensors and machine learning methods for stress detection. *IEEE Access*. **12**, 38219–38246. <https://doi.org/10.1109/ACCESS.2024.3373010> (2024).
4. Klimek, A., Mannheim, I., Schouten, G., Wouters, E. J. M. & Peeters, M. W. H. Wearables measuring electrodermal activity to assess perceived stress in care: a scoping review. *Acta Neuropsychiatr.* **37**, e19. <https://doi.org/10.1017/neu.2023.19> (2023).
5. Xiang, J. Z. et al. A multi-modal deep learning approach for stress detection using physiological signals: integrating time and frequency domain features. *Front. Physiol.* **16**, 1584299. <https://doi.org/10.3389/fphys.2025.1584299> (2025).
6. Feng, T. et al. A multimodal analysis of physical activity, sleep, and work shift in nurses with wearable sensor data. *Sci. Rep.* **11**, 8693. <https://doi.org/10.1038/s41598-021-87029-w> (2021).

7. Chouhan, S. A. Supervised information gain-based feature selection for multimodal physiological signals in stress prediction. *Sci. Rep.* **16**, 12169. <https://doi.org/10.1038/s41598-026-41734-6> (2026).
8. Rockenschaub, P. et al. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. *BMC Med. Inf. Decis. Mak.* **25**, 5. <https://doi.org/10.1186/s12911-024-02830-7> (2025).
9. Reps, J. M. et al. Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability. *BMC Med. Inf. Decis. Mak.* **22**, 142. <https://doi.org/10.1186/s12911-022-01879-6> (2022).
10. Parsons, C. S., Zuiderwijk-van Eijk, A. M. G., Orchard, N. A. O., Oosterhoff, J. H. F. & de Reuver, M. Task-Technology Fit of artificial intelligence-based clinical decision support systems: a review of qualitative studies. *BMC Med. Inf. Decis. Mak.* **25**, 397. <https://doi.org/10.1186/s12911-025-03237-8> (2025).
11. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378. <https://doi.org/10.1136/bmj-2023-078378> (2024).
12. von Elm, E. et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* **370**, 1453–1457. [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X) (2007).
13. Hosseini, S. et al. Multimodal sensor dataset for stress detection of nurses in a hospital. Dryad (2021). <https://doi.org/10.5061/dryad.5hqbzkh6f>
14. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357. <https://doi.org/10.1613/jair.953> (2002).
15. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30, 4765–4774 Curran Associates, (2017).

Figures

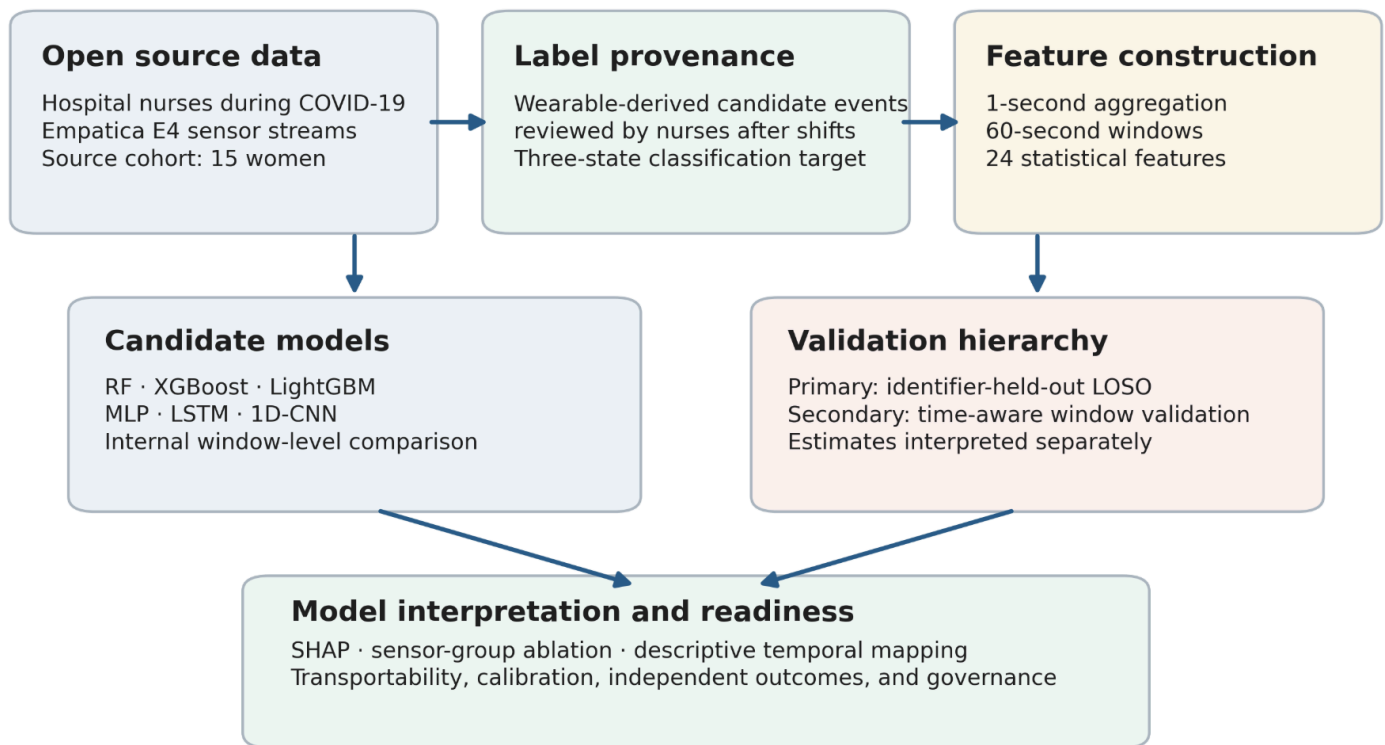


Figure 1

Study workflow and validation hierarchy. Open wearable data were represented using 60-second feature windows. Internal window-level model comparison was interpreted separately from identifier-held-out leave-one-subject-out evaluation. Model interpretation included SHAP, sensor-group ablation and descriptive temporal mapping. LOSO, leave-one-subject-out; RF, random forest; MLP, multilayer perceptron; LSTM, long short-term memory; 1D-CNN, one-dimensional convolutional neural network.

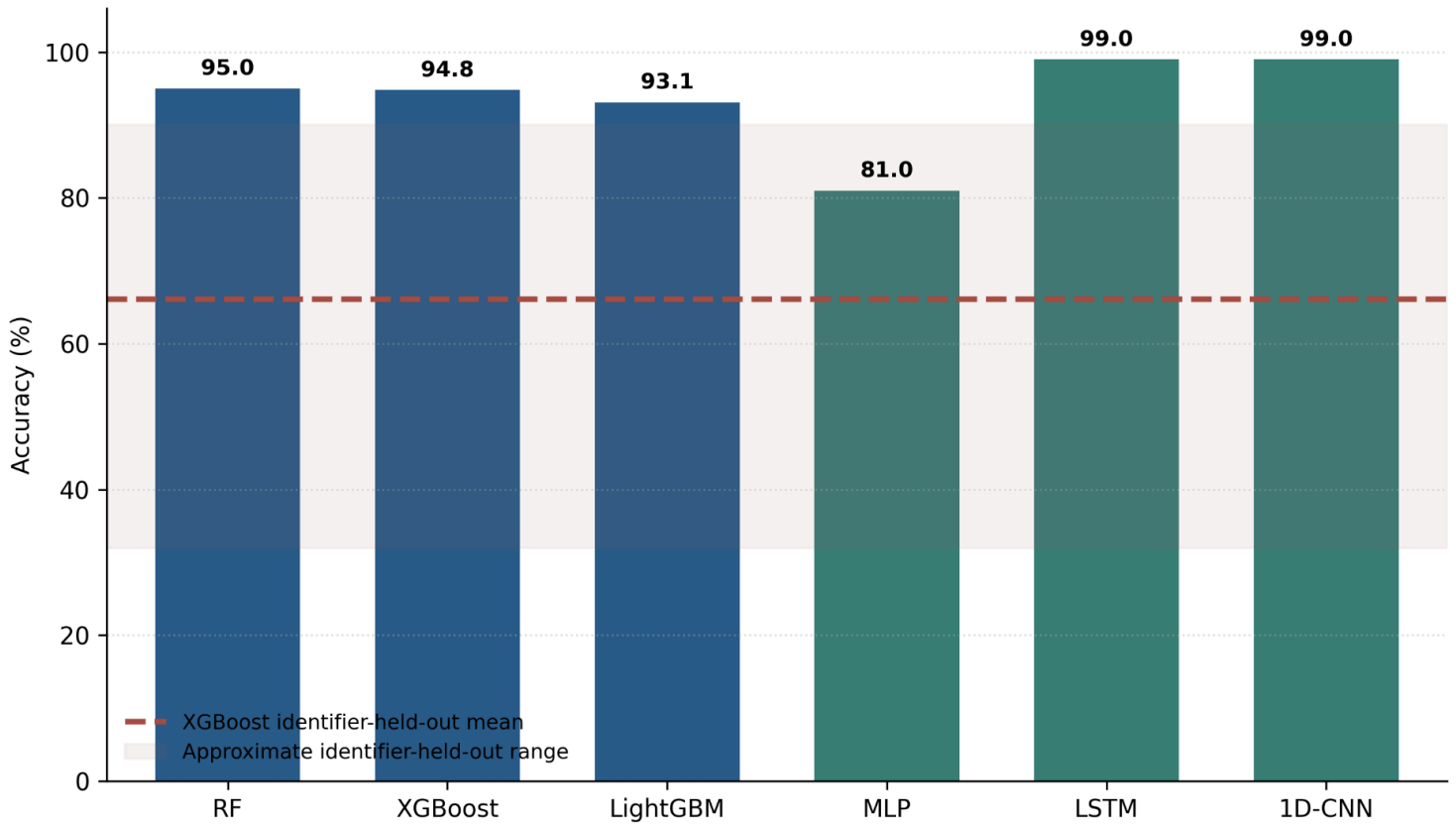


Figure 2

Internal and identifier-held-out accuracy. Bars show retained internal window-level accuracy for six candidate models. The dashed line shows mean XGBoost identifier-held-out accuracy (66.1%), and the lightly shaded area shows the approximate identifier-level range (32%–90%). The two estimates answer different validation questions.

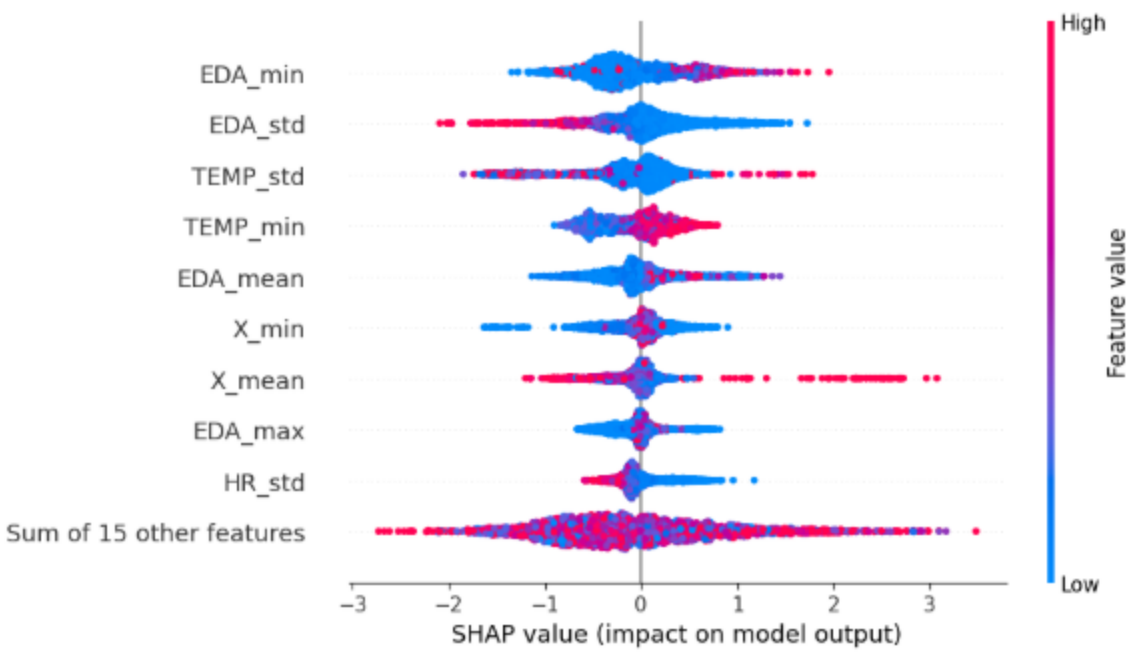


Figure 3

SHAP summary for the retained XGBoost model. Horizontal position represents contribution to model output and colour represents feature value. Electrodermal-activity features had the largest contributions. SHAP values describe fitted-model behaviour and should not be interpreted causally.

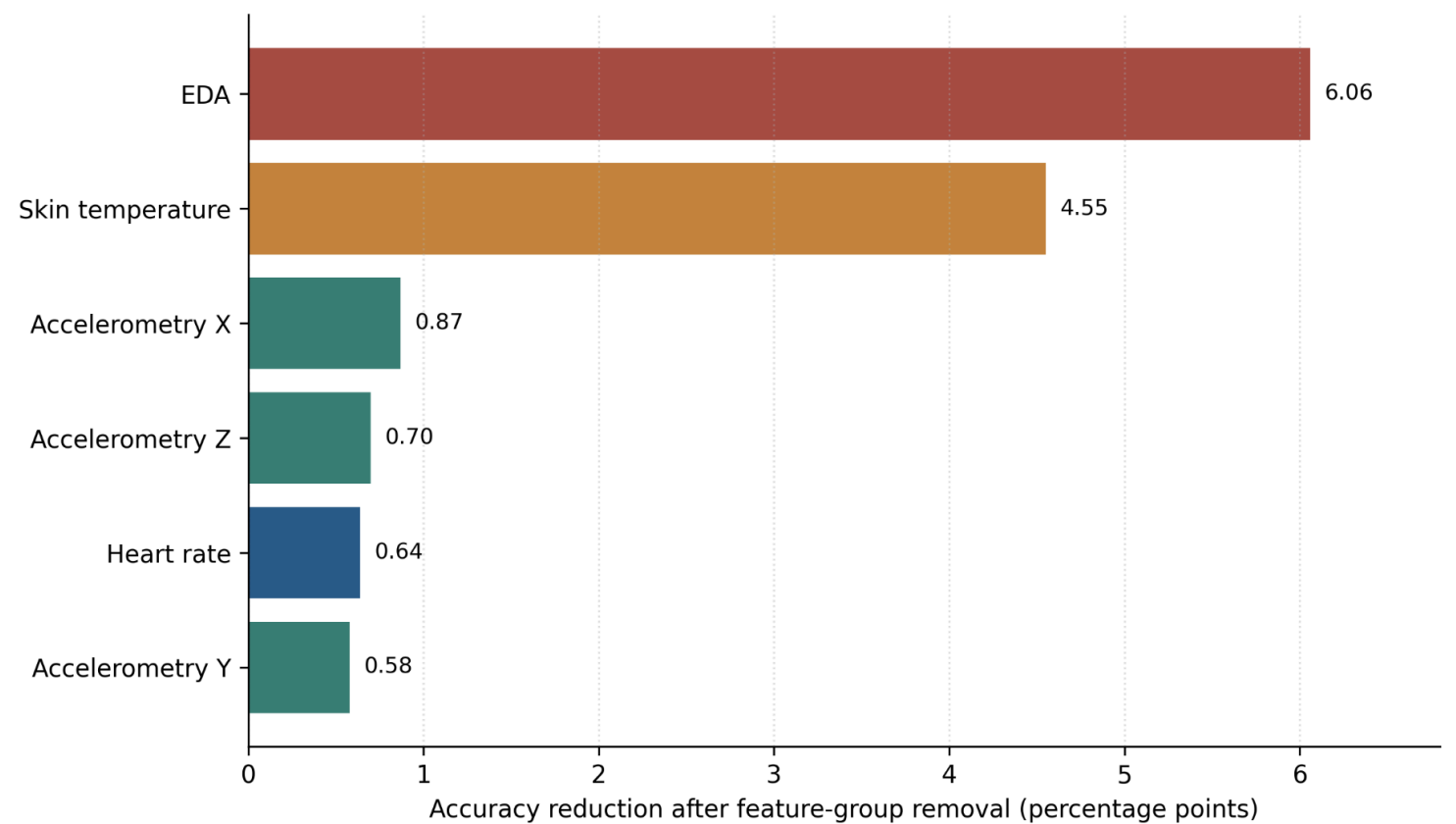


Figure 4

Sensor-group ablation. Bars show the absolute decrease in accuracy after removing each sensor group from the retained pipeline. Electrodermal activity and skin temperature produced the largest reductions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [02SupplementaryInformationScientificReports.docx](#)