

A Multimodal Physiological Dataset for Nurse Stress Detection

Luis Fernando Pareja Bernal

luis.parejabernal@bfh.ch

Bern University of Applied Sciences

Jannic Stefan Bieri

Bern University of Applied Sciences

Souhir Ben Souissi

Bern University of Applied Sciences

Christoph Golz

Bern University of Applied Sciences

data-descriptor

Keywords: Nurse Stress Detection , Machine Learning, Work-Related Stress, Physiological Signals, Multimodal Data, System Usability Scale, Questionnaire

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10826027/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Multimodal Physiological Dataset for Nurse Stress Detection

Luis Fernando Pareja Bernal^{1,*}, Jannic Stefan Bieri¹, Souhir Ben Souissi¹, and Christoph Golz²

¹Applied Machine Intelligence, Bern University of Applied Sciences, Switzerland

²School of Health Professions, Bern University of Applied Sciences, Switzerland

*luis.parejabernal@bfh.ch

ABSTRACT

Occupational stress among nurses affects well-being, workforce retention, and patient care, yet stress is rarely assessed using integrated multimodal data. This study evaluated the feasibility of combining self-reported assessments, smartwatch-derived physiological signals, and work schedule data for stress detection in nursing practice. Twenty four nurses participated in a three month field study in Switzerland. Daily questionnaire responses were combined using an expert-informed scoring equation to derive shift-level stress labels. After preprocessing, the machine learning dataset comprised 16,813 observations from 18 nurses. A CatBoost classifier achieved a weighted F1 score of 0.734 using heart rate, body temperature, shift information, temporal features, and nurse ID, outperforming models based solely on physiological data (F1 = 0.568). Work schedule analyses identified associations between recovery opportunities and work–privacy conflict, while smartwatch usability was rated as good (SUS = 75.15).

The findings demonstrate the feasibility of integrating subjective, physiological, and routinely collected organisational data in real world nursing practice. This multimodal approach provides a basis for context sensitive stress detection while highlighting the need for larger samples and person-independent model validation.

Keywords: Nurse Stress Detection , Machine Learning, Work-Related Stress, Physiological Signals, Multimodal Data, System Usability Scale, Questionnaire.

Introduction

Occupational stress among healthcare professionals is a persistent challenge for employees, healthcare organisations, and the sustainability of healthcare systems. Healthcare professionals are exposed to a combination of quantitative, emotional, physical, and organisational demands. Adverse psychosocial working conditions have been associated with stress reactions, impaired health and well-being, burnout, reduced job satisfaction, and intentions to leave the organisation or profession^{1,2}. These concerns are particularly relevant in nursing. Nurses frequently work under conditions characterised by high workload, staffing constraints, emotional demands, interruptions, and responsibility for patient safety. Shift work, overtime, irregular working hours, and limited recovery opportunities may create additional demands by disrupting sleep and the compatibility of work with private and family life. Unfavourable working conditions may consequently affect not only nurses' health and work-related outcomes but also the quality and safety of patient care³⁻⁵.

Work scheduling represents an important component of these organisational working conditions. Its potential relevance extends beyond the number of early, late, or night shifts and includes shift sequences, overtime, consecutive working days, recovery periods, schedule predictability, and changes between planned and effective duties. Nurses' influence over their working hours, the consideration of individual preferences, and the perceived fairness and transparency of scheduling processes may further shape how schedules are experienced and how well work can be reconciled with private life⁵⁻⁷.

Work-related stress is not a single, directly observable condition but a dynamic process arising from the interaction between work-related demands and the individual's appraisal of these demands in relation to available coping resources⁸. Within occupational health research, it is therefore important to distinguish between potential work-related stressors, short-term psychological, behavioural, and physiological responses, and possible longer-term health- and work-related consequences⁹.

Physiological activation forms part of an adaptive response to changing demands; however, repeated or prolonged activation may contribute to cumulative physiological strain and adverse health outcomes¹⁰.

Because these components represent different aspects of the stress process, occupational stress cannot be adequately captured by a single indicator or data source. Self-reported assessments are essential for capturing the individual's appraisal of working conditions and the subjective experience of burden. Assessments conducted close to the respective work situation can capture fluctuations between and within work shifts and reduce reliance on retrospective recall. However, repeated assessments require active participation and may be affected by non-response, declining compliance, and missing observations when implemented intensively in everyday work¹¹.

Wearable sensors can complement self-reported assessments by repeatedly or continuously recording physiological signals during routine work^{12,13}. Measurements such as heart rate and body temperature can provide time-resolved information about physiological changes occurring during work shifts. However, these signals should not be interpreted as direct or specific measures of psychological stress. Physiological responses may vary both between and within individuals and can be influenced by physical activity, circadian and thermoregulatory processes, environmental conditions, health status, shift timing, and individual characteristics¹⁴. Physiological measurements therefore need to be interpreted in relation to subjective assessments and the context in which they were recorded.

Routinely available work schedule data may provide an additional source of temporal and organisational context. Planned and effective schedules can be used to derive indicators describing shift types and frequencies, working sequences, schedule changes, consecutive working days, and potential recovery opportunities⁵. Because work schedules are generated for operational purposes, their secondary use may complement questionnaire and sensor data without requiring additional repeated data entry from participating nurses.

Nevertheless, scheduling indicators cannot automatically be interpreted as stressors. A discrepancy between a planned and an effective schedule may represent an unwanted disruption, but it may also result from a voluntary shift exchange, an individually preferred adjustment, or organisational flexibility. Similarly, the impact of a particular shift sequence may depend on the nurse's personal circumstances, preferences, perceived control, and opportunities for recovery^{6,7}. Linking scheduling indicators with self-reported assessments is therefore necessary to examine whether and under which circumstances these organisational characteristics are associated with perceived work-related burden.

Integrating questionnaire responses, physiological measurements, and routinely collected work schedule data may address some of the limitations inherent in each individual data source. Self-reports provide information about appraisal and perceived burden, wearable measurements capture physiological activation, and work schedules describe potential organisational and temporal stressors. Combining these complementary sources enables triangulation across different components of the stress process and may provide a more context-sensitive representation of occupational stress in real-world nursing practice¹⁵.

Machine learning and deep learning approaches have increasingly been applied to stress detection using physiological, behavioural, and multimodal data^{16,17}. Although these approaches have demonstrated promising predictive performance, the available evidence is frequently based on controlled experimental settings, retrospective analyses, small samples, or datasets providing limited information about actual working conditions. The implementation and evaluation of automated stress-detection approaches in routine healthcare environments therefore remain limited^{18,19}.

High-quality data collected during everyday clinical work are needed to develop stress-detection approaches that account for both physiological responses and the organisational context in which they occur. Combining subjective assessments, wearable sensor data, and routine work schedules may support not only the development of predictive models but also a more differentiated interpretation of the conditions under which stress-related responses are observed.

The present study therefore investigated the feasibility of integrating questionnaire responses, smartwatch-derived physiological signals, and routinely collected work schedule data from nurses during their everyday work. The study comprised two complementary analytical components. First, heart rate (HR) and body temperature (BT) measurements were temporally aligned with shift information and questionnaire-derived stress labels to construct a multimodal dataset and evaluate different combinations of physiological and contextual features using a CatBoost classifier. Second, planned and effective work schedules were transformed into scheduling-related indicators and explored in relation to perceived daily work-related burden and Work-Privacy Conflict. The perceived usability of the smartwatch-based data collection approach was also assessed.

The main contributions of this study are:

- The construction of a real-world multimodal dataset combining heart rate, body temperature, shift information, temporal variables, nurse identifiers, and questionnaire-derived stress labels collected during nurses' routine work;

- The evaluation of the feasibility and perceived usability of smartwatch-based physiological data collection in everyday nursing practice;
- The derivation of scheduling-related indicators from routinely available planned and effective work schedules and the exploratory examination of their relationships with perceived daily burden and Work-Privacy Conflict; and
- The evaluation of different combinations of physiological and contextual features for machine-learning-based stress classification using a CatBoost model.

Methods

This study aims to construct a multimodal physiological dataset suitable for clinical environments gathering features like HR, body BT, shift code, nurse ID and day of shift. The construction of the dataset involves 3 broad stages: data collection, data preprocessing and annotation.

Aim and exploratory hypotheses

The overall aim of this analysis was to examine the feasibility of using routinely available work schedule data to derive scheduling-related indicators and to explore their relationship with nurses' perceived work related burden and Work-Privacy Conflict. Specifically, the analysis focused on three aspects: first, the perceived usability of the smartwatch as part of the data collection approach; second, the feasibility of transforming planned and effective work schedules into meaningful scheduling indicators; and third, the exploratory association between these scheduling indicators and self-reported indicators of work related stress.

Given the feasibility-oriented design and the small sample size, the analyses were primarily exploratory:

- **Usability of the smartwatch:** It was expected that nurses would report at least acceptable perceived usability of the smartwatch after the three-month data collection period, indicating that the technical data collection approach was feasible in everyday nursing practice.
- **Description of scheduling indicators:** It was expected that routinely available planned and effective work schedules could be transformed into interpretable scheduling-related indicators, including shift frequencies, mismatches between planned and effective schedules, single days off, periods of more than five consecutive working days, changes from scheduled free days to workdays, and shift rotation patterns.
- **Daily burden by shift type:** It was expected that perceived daily burden would vary between work-related shift categories. In particular, shift categories reflecting direct patient care and less favourable working times, such as early, late, and night shifts, were expected to be associated with higher perceived daily burden than less frequent non-standard categories.
- **Scheduling indicators and Work-Privacy Conflict:** It was expected that scheduling indicators reflecting lower predictability or reduced recovery opportunities, such as mismatches between planned and effective schedules, single days off, more than five consecutive working days, changes from free days to workdays, and rapid shift changes, would be positively associated with Work-Privacy Conflict. In contrast, a higher number of free days was expected to be negatively associated with Work-Privacy Conflict.

Data Collection

This study used nurses and their work shifts as the primary source of multimodal physiological data. Data were collected between March and April 2025 from 24 nurses working in a hospital in Switzerland. During their shifts, participants wore a Withings smartwatch²⁰, which records a range of physiological signals, including heart rate, body temperature, step count, traveled distance, and electrocardiograms (ECGs). After each shift, participants completed a questionnaire to assess their perceived stress levels during the workday.

Questionnaire

During the three-month data collection period, nurses completed a short daily questionnaire after each recorded work shift. The daily questionnaire included items on satisfaction with the current shift, perceived work-related burden, and compatibility of the shift with private life.

In this questionnaire the nurses had to answer three simple questions:

- *Zufriedenheit täglich* (*zufrid_tgl*): indicates the nurse's overall satisfaction with the current work shift.
- *Der heutige Arbeitstag war sehr belastend für mich* (*belas_tgl*): assesses the perceived level of stress experienced during the workday.
- *Die heutige Schicht/Dienst war sehr gut vereinbar mit meinem Privatleben* (*wp_tgl*): evaluates how well the shift was compatible with the nurse's private life and daily activities.

These questionnaire responses were used alongside an expert validated equation 1 to generate the stress labels associated with the physiological data.

For the statistical analysis, the daily item on perceived work-related burden was used: *Der heutige Arbeitstag war sehr belastend für mich* (*belas_tgl*). This item assessed the extent to which nurses perceived the respective workday as stressful.

In addition to the daily questionnaire, nurses completed a comprehensive survey after the three-month data collection period. The comprehensive questionnaire included the following instruments:

- Work-Privacy Conflict from the Copenhagen Psychosocial Questionnaire²¹.
- System Usability Scale (SUS)²²

The SUS was used to assess the perceived usability of the smartwatch after the study period. Work-Privacy Conflict was selected as the most relevant known stressor.

Furthermore, nurse managers from the two participating healthcare organisations provided anonymised work schedules for the participating nurses. Both planned and effective work schedules were made available, allowing deviations between planned and actual work schedules to be examined.

Annotation

Stress labels were generated based on the nurses' questionnaire responses. Using these responses, the following Equation 1 was applied to compute the corresponding labels:

$$\text{label} = 0.5 \cdot \text{belas_tgl} + 0.2 \cdot (10 - \text{zufri_tgl}) + 0.3 \cdot (10 - \text{wp_tgl}) \quad (1)$$

The equation assigns stress labels based on the score obtained from the nurses' questionnaire. Specifically:

- If the resulting score is less than 4, the individual is considered not stressed and is assigned a label of 0.
- If the score falls between 4 and 7 (inclusive of 4 but less than 7), the individual is considered mildly stressed and is assigned a label of 1.
- If the score is greater than or equal to 7, the individual is considered very stressed and is assigned a label of 2.

This categorization transforms the continuous questionnaire score into a discrete label suitable for further analysis and modeling.

Ethics approval and consent to participate

The Cantonal Ethics Committee of Bern, Switzerland, reviewed the study and confirmed that it did not fall under the Swiss Federal Act on Research Involving Human Beings and therefore did not require formal ethical approval (Req-2024-01388). Participation was voluntary for all employees. Prior to participation, all participants received detailed study information specifying the purpose and procedures of the study, the data collected through questionnaires and wearables, and the data obtained from organizational routine data sources. The study information also stated that an anonymized dataset would be made available for data sharing purposes. The two participating healthcare organizations were likewise informed in advance about the planned sharing of an anonymized dataset. Informed consent was obtained from all participants; no consent waiver was granted or required. Participants were free to stop completing the questionnaires, discontinue wearing the wearable, or withdraw from the study at any time.

Data Cleaning & Preprocessing

HR and BT measurements were originally recorded in separate files and later merged into a single dataset. Heart rate data were structured as one measurement per row, whereas body temperature data contained multiple measurements within a single row. To harmonize the formats, redundant heart rate entries (noise) were removed, and body temperature arrays were expanded so that each measurement occupied its own row. Because the original sampling frequencies of the two signals differed, both were resampled prior to merging. The merge itself was performed using an 'as of' join, a method well suited to time aligned physiological data.

After merging, the dataset was resampled into 10 minute intervals, ensuring that consecutive rows were spaced uniformly. Shift information was then incorporated by assigning each observation to the corresponding shift window. These processing steps introduced missing values, which were subsequently removed.

Finally, the dataset was refined by excluding days on which nurses were not working and by retaining only measurements recorded during their scheduled shifts. Complete shift schedules were available for each nurse throughout the study, enabling precise filtering.

Final Dataset

A cleaned, annotated and ready to use dataset with 16,813 observations. The dataset includes physiological and contextual variables such as HR, BT, shift code, class label, day of measurement, and nurse identifier.

Figures 1 and 2 present two plots for the same nurse, corresponding to a stressed shift and a non stressed shift, respectively. The visualizations illustrate differences in physiological signals between the two conditions.

Figure 1. Physiological signals for a nurse during a stressed shift.

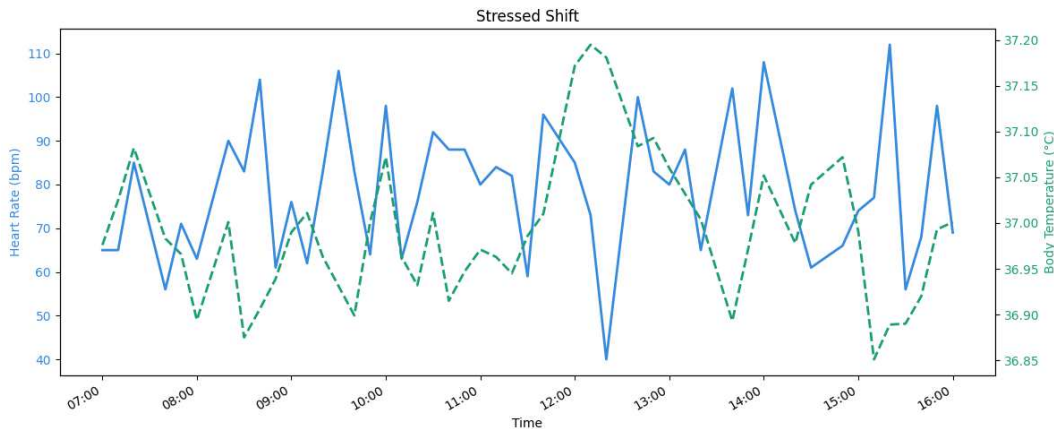


Figure 3 shows that the dataset has class imbalance, with a predominance of the no stress class ($n = 11,177$), compared to the stress classes (class 1: $n = 4,895$; class 2: $n = 740$). This issue can be addressed using several strategies, including applying class weighting during model training or reformulating the problem as a binary classification task.

Technical Validation

Statistical Analysis

The aim of the analysis was to examine the feasibility of using routinely available scheduling data to describe and explore scheduling-related indicators of work-related stress among nurses. Therefore, all analyses were exploratory in nature.

First, the SUS score was calculated to assess nurses' perceived usability of the smartwatch during the study period. Second, information from the planned and effective work schedules was transformed into scheduling-related variables. These variables included: (a) number and type of shifts, (b) discrepancies between planned and effective schedules, (c) single days off, (d) more than five consecutive working days without a break, and (e) being called in on a scheduled day off.

Furthermore, shift work patterns were categorised into two broad types: (f) working the same shift type for a prolonged period

Figure 2. Physiological signals for a nurse during a non-stressed shift.

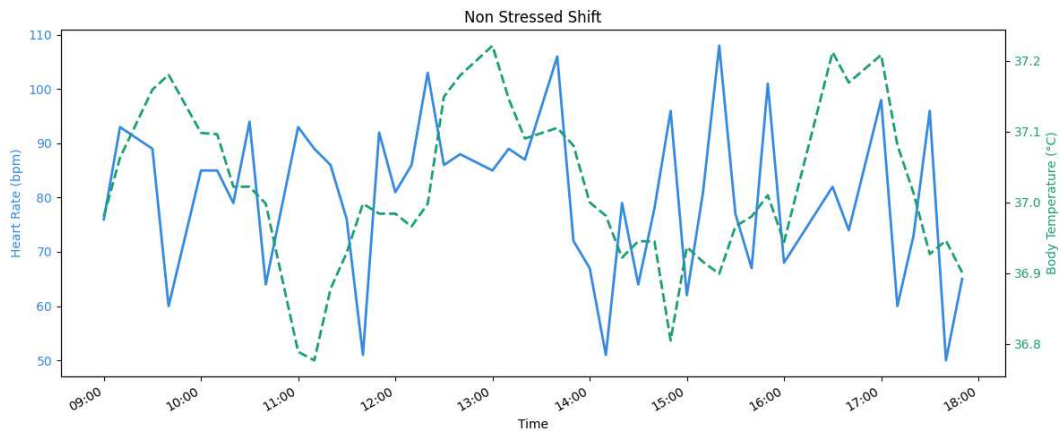
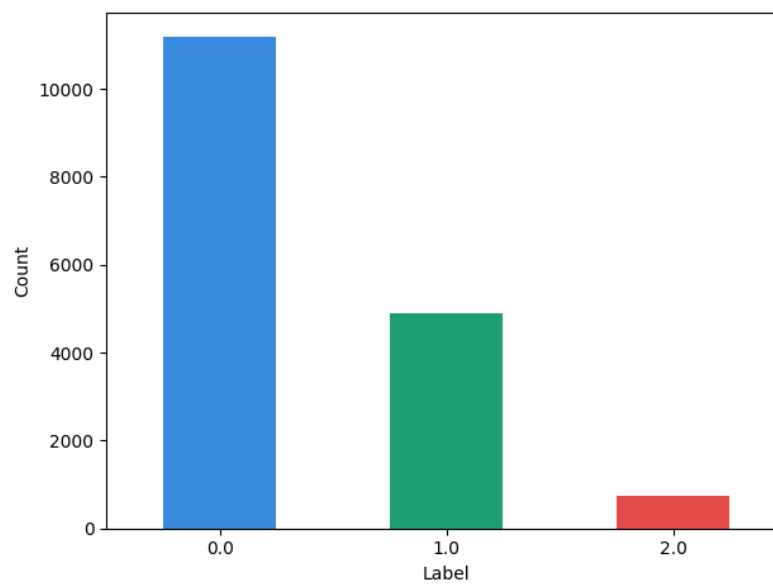


Figure 3. Dataset's label distribution



and (g) rotating shift systems. Rotating shift systems were further differentiated into (g.1) slowly rotating shift systems, defined as rotation over a period of 2 to 4 weeks, and (g.2) rapidly rotating shift systems, defined as rotation every 3 to 5 days^{5,23}.

Descriptive statistics were calculated for the scheduling-related variables. To explore whether perceived daily burden differed by shift type, daily burden ratings (*belas_tgl*) were analysed in relation to the respective shift category. Because repeated daily measurements were nested within nurses, these analyses were interpreted descriptively and, where appropriate, explored using models accounting for repeated observations within persons.

In addition, selected aggregated scheduling indicators were correlated with Work-Privacy Conflict at post-measurement. Work-Privacy Conflict was chosen because it was considered the questionnaire-based stressor most directly related to scheduling and the compatibility between work and private life. Long-term consequences such as burnout symptoms were not included in the main analysis, as the present study focused on feasibility and proximal indicators of work-related burden.

Results

The perceived usability of the smartwatch was assessed using the System Usability Scale. The mean SUS score was 75.15 (SD = 15.06), suggesting overall acceptable to good perceived usability of the smartwatch among participating nurses.

Over the three-month observation period, 24 nurses were included in the scheduling analysis. The most frequent roster category was free days ($M = 42.75$, $SD = 7.91$), followed by early shifts ($M = 19.29$, $SD = 13.33$) and late shifts ($M = 13.54$, $SD = 9.20$). Night shifts occurred less frequently on average, but showed substantial variation between nurses ($M = 7.54$, $SD = 11.94$).

In addition to the frequencies of roster categories, several scheduling-related indicators were derived from the planned and effective schedules. On average, nurses experienced 5.04 mismatches between planned and effective schedules ($SD = 4.41$) during the three-month observation period. Single days off occurred frequently, with an average of 11.46 occurrences per nurse ($SD = 3.55$). Periods of more than five consecutive working days occurred on average 1.71 times per nurse ($SD = 1.04$). Changes from a scheduled free day to a workday occurred less often, with an average of 2.21 occurrences per nurse ($SD = 3.73$). No fixed shift patterns were observed in the sample. Rapid shift changes occurred on average 3.75 times per nurse ($SD = 2.03$), whereas slow shift changes were rare ($M = 0.04$, $SD = 0.20$).

Table 1.

Table 1. Frequencies of roster categories over the three-month observation period

Roster category	Mean	SD
Early shift	19.29	13.33
Free day	42.75	7.91
Illness	3.04	5.89
Office day	1.67	5.73
Half shift	1.33	5.69
Late shift	13.54	9.20
Night shift	7.54	11.94
Education day	1.96	4.29
Split shift	0.08	0.41
Mid shift	0.79	2.52
(b) Mismatch	5.04	4.41
(c) Single day off	11.46	3.55
(d) Streaks over 5	1.71	1.04
(e) Change from free to work	2.21	3.73
(f) fix shift	0	0
(g.1) slow change	0.04	0.20
(g.2) fast change	3.75	2.03

Note. Values represent the number of roster categories per nurse across the three-month observation period. $N = 24$.

Results on perceived daily burden by shift type:

Perceived daily burden was analysed across work-related shift categories using a linear mixed-effects model with random intercepts for nurses. Free days and illness days were excluded from this analysis, as the daily burden item referred to

the respective workday. The overall association between shift category and perceived daily burden did not reach statistical significance, although a descriptive tendency was observed, $F(7, 248.24) = 1.94$, $p = .064$. Estimated marginal means indicated that perceived burden was highest for early shifts (EMM = 3.54, SE = 0.27), followed by office days (EMM = 3.21, SE = 0.64), late shifts (EMM = 3.17, SE = 0.28), mid shifts (EMM = 2.88, SE = 0.72), and night shifts (EMM = 2.69, SE = 0.36). Estimates for split shifts, education days, and half shifts should be interpreted with caution because these roster categories occurred rarely, resulting in large standard errors and wide confidence intervals.

Results on correlations between aggregated scheduling indicators and Work-Privacy Conflict:

Work-Privacy Conflict at post-measurement was correlated with the aggregated scheduling indicators. Overall, the correlations were small to moderate and should be interpreted descriptively due to the feasibility-oriented study design and the small sample size. Higher Work-Privacy Conflict was moderately associated with more education days ($r = .40$), illness days ($r = .32$), and changes from a scheduled free day to a workday ($r = .28$). In contrast, a higher number of free days was moderately associated with lower Work-Privacy Conflict ($r = -.36$). The number of mismatches between planned and effective schedules showed a small negative correlation with Work-Privacy Conflict ($r = -.24$). Other scheduling indicators, including early, late, and night shifts, single days off, streaks of more than five consecutive working days, and fast or slow shift changes, showed only weak associations with Work-Privacy Conflict.

Table 2.

Table 2. Correlations between roster indicators and Work-Privacy Conflict

Roster indicator	Correlation with Work-Privacy Conflict
Early shift	.03
Free day	-.36
Illness	.32
Office day	.05
Half shift	-.15
Late shift	-.01
Night shift	-.04
Education day	.40
Mid shift	-.16
Mismatch between planned and effective schedule	-.24
Single day off	-.14
More than five consecutive working days	.10
Change from free day to workday	.28
Fast shift change	-.12
Slow shift change	-.15

Note. Values represent correlations between aggregated roster indicators over the three-month observation period and Work-Privacy Conflict at post-measurement. $N = 24$.

Stress Classification

A CatBoost classifier was trained using different feature combinations. To address class imbalance, one of the strategies applied was reformulating the original multiclass problem into a binary classification task.

CatBoost is a gradient boosting algorithm that constructs an ensemble of decision trees in a sequential manner, where each tree corrects the errors of the previous ones. Unlike traditional gradient boosting methods, CatBoost incorporates several innovations that improve performance and reduce the need for extensive hyperparameter tuning. In particular, it uses ordered boosting to reduce prediction shift and provides native handling of categorical features without extensive preprocessing, making it well suited for heterogeneous data²⁴.

Boosting-based models, such as AdaBoost, Gradient Boosting, and XGBoost, have been widely used in stress detection studies and have demonstrated strong predictive performance^{25–27}.

Table 3 shows the experimental setup.

All experiments used the same label distribution. In the training set, 66% of the samples belonged to the no stress class (label 0)

Table 3. CatBoost hyperparameter configuration used in the experiments.

Parameter	Value
Model	CatBoost
Iterations	1500
Learning rate	0.025
Depth	8
L2 leaf regularization (l2_leaf_reg)	5
Loss function	Logloss
Scale positive weight	N_0/N_1
Random seed	42
Early stopping rounds	100
Verbose	100

and 34% to the stress class (label 1). The test set had a similar distribution, with 68.6% and 31.4% of samples assigned to the no stress and stress classes, respectively. Additionally, to avoid data leakage, the dataset was split by shift, ensuring that rows belonging to the same shift did not appear in both the training and test sets.

Table 4 reports the feature combinations achieving the highest performance in terms of weighted F1 score.

Table 4. Top-performing feature combinations ranked by weighted F1-score.

Rank	Feature Combination	Weighted F1
1	No start (HR, BT, shift_code, day, ID)	0.734
2	All features (HR, BT, day, shift_code, start, ID)	0.718
3	No ID (HR, BT, day, shift_code, start)	0.670
4	No HR & ID (BT, shift_code, day)	0.656
5	Only physiological features (baseline)	0.568

Figure 4 shows the feature importance for the best performing feature set (No start (HR, BT, shift_code, day, ID)).

It is important to note that different feature combinations lead to different feature importance distributions, since the model is retrained under the same configuration and feature interactions depend on the available input set.

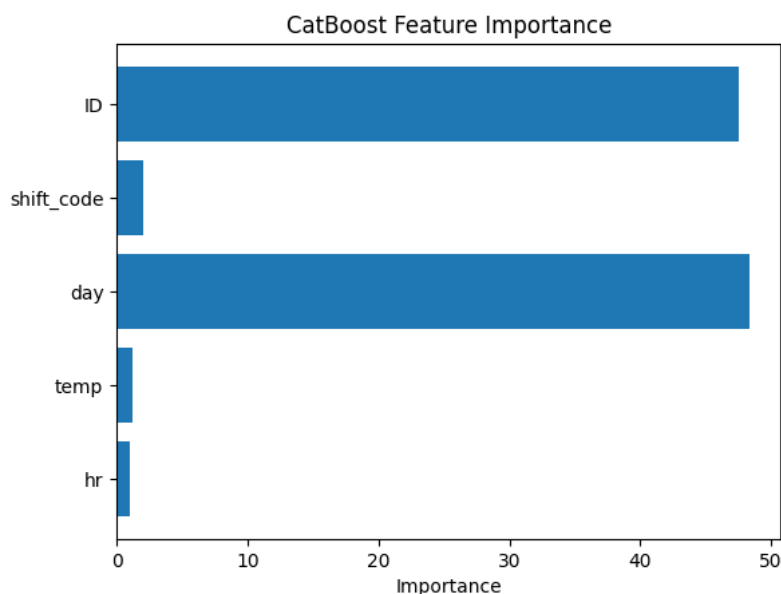
Discussion

The present study investigated whether routinely available work schedule data, combined with self-reported and physiological measures, can be used to characterise work-related burden and support the identification of stress among nurses. The findings provide initial support for the feasibility of this approach: scheduling data could be transformed into interpretable indicators describing shift patterns, schedule stability, recovery opportunities, and changes between planned and effective work, while the smartwatch-based data collection showed acceptable to good usability (mean System Usability Scale score of 75.15). Together, these findings suggest that combining routinely available organisational data with wearable and self-reported measures is technically feasible in everyday nursing practice.

Routine scheduling data provided information beyond simple descriptions of shift frequency. Indicators such as the number of free days, single days off, mismatches between planned and effective schedules, changes from scheduled free days to workdays, and shift rotation patterns captured distinct aspects of nurses' working conditions. A higher number of free days was associated with lower Work-Privacy Conflict, whereas changes from a scheduled free day to a workday were associated with higher conflict (supporting the interpretation that recovery opportunities and predictability of private time are relevant to how nurses experience the interaction between work and private life).

At the same time, scheduling characteristics could not be interpreted uniformly as indicators of burden. Despite descriptive differences between shift categories, no statistically significant overall association emerged between shift category and perceived daily burden. Early, office, and late shifts showed higher estimated burden, while night shifts showed somewhat lower estimates; however, the rare occurrence of categories such as split shifts, education days, and half shifts resulted in imprecise estimates. Shift type therefore appears to be one component of a broader combination of organisational and individual factors shaping daily experience, rather than a factor that systematically produces greater burden on its own.

Figure 4. Feature importance of set of features with highest weighted F1 (No start (HR, BT, shift_code, day, ID)) score



The results concerning schedule changes similarly underscore the importance of context. Although changes from planned free days to workdays were associated with higher Work-Privacy Conflict, mismatches between planned and effective schedules showed a small negative association, contrary to expectation. A plausible interpretation is that not every schedule change represents an unwanted disruption: some may be negotiated, voluntary, or experienced as organisational flexibility. An indicator based solely on the occurrence of a mismatch therefore cannot distinguish potentially stressful changes from potentially beneficial ones, since routine administrative data describe what happened without explaining why or how it was experienced.

The stress-classification results offer a complementary perspective. The CatBoost model achieved its highest weighted F1-score of 0.734 using heart rate, body temperature, shift code, day, and participant ID while excluding the start feature. The model using all available features achieved a slightly lower F1-score of 0.718, and a model without participant ID achieved 0.670, while a model based on physiological features alone achieved a substantially lower F1-score of 0.568. This indicates that physiological information alone was less effective for distinguishing stress classes than combinations incorporating contextual and scheduling-related information (consistent with the assumption that stress in nursing is unlikely to be determined by physiological characteristics alone).

The strong performance of the feature set containing participant ID nonetheless requires caution. Because participant ID is not itself a meaningful work-related stressor, its contribution may reflect individual-specific differences in stress levels, physiology, behavior, or measurement patterns rather than generalizable relationships between work schedules, physiology, and stress. The comparison with the model excluding ID (F1-score of 0.670) illustrates the importance of distinguishing predictive performance that reflects generalizable patterns from performance driven by participant-specific characteristics. Shift-level splitting (ensuring rows from the same shift did not occur in both training and test data) reduced the risk of direct data leakage and provided a more appropriate performance assessment than random splitting of individual observations.

Taken together, the findings suggest that the value of routinely available scheduling data lies less in identifying a single characteristic that explains stress and more in providing contextual information that can be combined with physiological and self-reported measures. The comparatively weaker performance of the physiological-only baseline supports this interpretation: scheduling information may help contextualize physiological responses, while physiological measures may capture information that schedules alone cannot provide. This combination of data sources represents a promising direction for stress-monitoring approaches in nursing, though the current results remain exploratory and do not establish causal relationships between scheduling conditions and stress or Work-Privacy Conflict.

Data Records

The Multimodal Physiological Dataset for Nurse Stress Detection is publicly available on [here](#).

The dataset is provided as a single comma-separated values (.csv) file, with its fields detailed in Table 5.

Table 5. Description of the dataset fields.

Feature	Description
start	Timestamp of the observation/event
hr	Heart rate measurement
temp	Body temperature
day	Calendar day
ID	Nurse identifier
label	Stree label
shift_code	Categorical code for work shift
start_datetime	Start of shift
end_datetime	End of shift

Limitations and Future Work

Limitations

Several limitations of this study should be considered when interpreting the results. First, the size of the physiological dataset and the distribution of stress classes may limit the generalizability of the findings, particularly for minority stress categories. The relatively small number of participants also increases the risk that the models learned participant specific patterns rather than stress related characteristics. In particular, nurse identity may have influenced model performance, potentially influencing the model's predictions and leading to optimistic performance results

A further limitation emerged during data collection. The number of available stress labels did not always match the number of days worked by the participating nurses. Ideally, each workday would have been associated with a corresponding stress label obtained through the daily questionnaires. However, nurses did not consistently complete these questionnaires, resulting in workdays for which physiological measurements were available but no stress label was recorded.

Although these unlabeled instances still contained useful information, they could not be included in the supervised learning experiments. Consequently, only labeled samples were used for model training and evaluation, reducing the effective dataset size and increasing class imbalance.

Future Work

Future work should focus on expanding the dataset by recruiting a larger and more diverse group of nurses across different departments, hospitals, and work environments. Increasing the number of participants and collecting data over longer periods would improve the robustness of the developed models. Particular attention should be given to increasing the representation of minority stress classes to mitigate class imbalance and improve performance across all stress levels.

In addition, future studies should investigate strategies to reduce missing labels, such as improving questionnaire compliance through reminders or integrating alternative labeling approaches. Exploring semi supervised or self supervised learning methods may also allow the utilization of unlabeled data that are currently excluded from the analysis.

Finally, applying model interpretability techniques could provide valuable insights into which physiological signals contribute most strongly to stress detection. Improved interpretability would enhance model transparency, support trust in healthcare applications, and provide a deeper understanding of the relationship between occupational stress and physiological indicators.

Usage Note

The Multimodal Physiological Dataset for Nurse Stress Detection is designed to support research on occupational stress monitoring using physiological and work related data collected in real clinical environments. The dataset combines heart HR, BT, shift data, temporal variables, and nurse IDs with stress labels derived from questionnaire responses and expert validated scoring equation.

The dataset can be used to investigate a variety of machine learning, including binary and multiclass stress classification, personalized stress detection, feature importance analysis, and analysis of how physiological signals change during work shifts. The inclusion of shift type and temporal information enables studies on the relationship between work schedules and stress, while the availability of nurse identifiers supports research on subject-dependent and subject-independent modeling approaches.

Researchers may employ the dataset to evaluate traditional machine learning algorithms, deep learning architectures, transfer learning methods, and explainable artificial intelligence techniques. Furthermore, the dataset provides a foundation for developing passive stress monitoring systems that can be integrated into wearable devices and clinical decision support applications.

Data Availability

The complete dataset is available in the GitHub repository at [Physiological-Stress-Nurse-Dataset](#). All files are distributed under the MIT License, which allows unrestricted use, modification, and distribution provided that the original license and copyright notice are retained.

Code Availability

The code containing the experiments is available in the same GitHub repository at [Physiological-Stress-Nurse-Dataset](#). All files are distributed under the MIT License, which permits unrestricted use, including use, modification, and distribution, provided that the original license and copyright notice are included in all copies or substantial portions of the software. The code for data preprocessing and analysis of SUS, work-privacy conflict and schedules can be found in the Supplementary Information under Supplementary Code S1.

Acknowledgements

AI-based tools were used solely for linguistic assistance and text revision and did not contribute original scientific content, data analysis, or conclusions. All output was critically reviewed and edited by the authors.

1 Research funding

This research did not receive funding.

Author contributions

L.F.P.B. conceptualised and conducted the preprocessing, cleaning and labeling of the data, the execution of the experiments and prepared the corresponding sections of the manuscript. J.S.B and C.G. conceptualised and conducted the preprocessing and analysis of the survey and scheduling data. S.B.S. and C.G supervised the research, contributed to the study design, and writing, reviewed and edited the manuscript. C.G. provided domain expertise in occupational health and nursing, contributed to the design of the stress labeling equation, and supported clinical validation of the methodology. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Peter, K. A. *et al.* Factors associated with health professionals' stress reactions, job satisfaction, intention to leave and health-related outcomes in acute care, rehabilitation and psychiatric hospitals, nursing homes and home care organisations. *BMC Heal. Serv. Res.* **24**, DOI: [10.1186/s12913-024-10718-5](https://doi.org/10.1186/s12913-024-10718-5) (2024).

2. Gilles, I. *et al.* Work experiences of healthcare professionals in a shortage context: analysis of open-ended comments in a swiss cohort (scohpica). *BMC health services research* **25**, 520 (2025).
3. Holland, P., Tham, T. L., Sheehan, C. & Cooper, B. The impact of perceived workload on nurse satisfaction with work-life balance and intention to leave the occupation. *Appl. nursing research* **49**, 70–76, DOI: [10.1016/j.apnr.2019.06.001](https://doi.org/10.1016/j.apnr.2019.06.001) (2019).
4. Van Bogaert, P. *et al.* Predictors of burnout, work engagement and nurse reported job outcomes and quality of care: a mixed method study. *BMC nursing.-London* **16**, 5, DOI: [10.1186/S12912-016-0200-4](https://doi.org/10.1186/S12912-016-0200-4) (2017).
5. Martins, T., Musy, S. N. & Simon, M. Nurse shift patterns, staffing and their association with perceived workload: Sequence analysis of multicentre data. *Int. J. Nurs. Stud. Adv.* 100420 (2025).
6. Renggli, F. J., Gerlach, M., Bieri, J. S., Golz, C. & Sariyar, M. Integrating nurse preferences into ai-based scheduling systems: Qualitative study. *JMIR Form Res* **9**, e67747, DOI: [10.2196/67747](https://doi.org/10.2196/67747) (2025).
7. Gerlach, M., Renggli, F. J., Bieri, J. S., Sariyar, M. & Golz, C. Exploring nurse perspectives on ai-based shift scheduling for fairness, transparency, and work-life balance. *BMC nursing* **24**, 1161, DOI: [10.1186/s12912-025-03808-0](https://doi.org/10.1186/s12912-025-03808-0) (2025).
8. Bakker, A. B. & de Vries, J. D. Job demands–resources theory and self-regulation: New explanations and remedies for job burnout. *Anxiety, Stress. & Coping* **34**, 1–21, DOI: [10.1080/10615806.2020.1797695](https://doi.org/10.1080/10615806.2020.1797695) (2021).
9. Gonzalez-Mul , E., Kim, M. M. & Ryu, J. W. A meta-analytic test of multiplicative and additive models of job demands, resources, and stress. *J. Appl. Psychol.* **106**, 1391 (2021).
10. O'Connor, D. B., Thayer, J. F. & Vedhara, K. Stress and health: A review of psychobiological processes. *Annu. review psychology* **72**, 663–688 (2021).
11. Wrzus, C. & Neubauer, A. B. Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment* **30**, 825–846 (2023).
12. Frazier, S. E. & Parker, S. H. Measurement of physiological responses to acute stress in multiple occupations: A systematic review and implications for front line healthcare providers. *Transl. Behav. Medicine* **9**, 158–166, DOI: [10.1093/tbm/iby019](https://doi.org/10.1093/tbm/iby019) (2019).
13. Fauzi, M. A. & Yang, B. Continuous stress detection of hospital staff using smartwatch sensors and classifier ensemble. In *pHealth 2021*, 245–250, DOI: [10.3233/SHTI210607](https://doi.org/10.3233/SHTI210607) (IOS Press, 2021).
14. Weber, J., Angerer, P. & Apolin rio-Hagen, J. Physiological reactions to acute stressors and subjective stress during daily life: A systematic review on ecological momentary assessment (ema) studies. *PloS one* **17**, e0271996 (2022).
15. Lukan, J. *et al.* Work environment risk factors causing day-to-day stress in occupational settings: a systematic review. *BMC public health* **22**, 240 (2022).
16. Giannakakis, G. *et al.* Review on psychological stress detection using biosignals. *IEEE transactions on affective computing* **13**, 440–460 (2019).
17. Richer, R. *et al.* Machine learning-based detection of acute psychosocial stress from body posture and movements. *Sci. Reports* **14**, 8251 (2024).
18. Ikae, C. *et al.* A scoping review of natural language processing for detecting work-related stress among health professionals. *Discov. Comput.* **29**, 14, DOI: [10.1007/s10791-025-09886-7](https://doi.org/10.1007/s10791-025-09886-7) (2026).
19. Pareja Bernal, L. F., Ben Souissi, S. & Golz, C. Machine learning-based detection of workplace stress using wearable and multimodal data: a systematic literature review. *Front. Artif. Intell.* **Volume 9 - 2026**, DOI: [10.3389/frai.2026.1837195](https://doi.org/10.3389/frai.2026.1837195) (2026).
20. Withings. Scanwatch 2 (2026). Accessed: 2026-06-08.
21. Lincke, H.-J. *et al.* Copsoq iii in germany: validation of a standard instrument to measure psychosocial factors at work. *J. Occup. Medicine Toxicol.* **16**, 50 (2021).
22. Brix, T. J., Janssen, A., Storck, M. & Varghese, J. Comparison of german translations of the system usability scale–which to take? In *German Medical Data Sciences 2023–Science. Close to People. Proceedings of the 68th Annual Meeting of the German Association of Medical Informatics, Biometry, and Epidemiology eV (gmds) 2023 in Heilbronn, Germany*, 96–101 (SAGE Publications 1 Oliver’s Yard, 55 City Road, London, EC1Y 1SP, 2023).
23. Chang, Y.-S., Wu, Y.-H., Chen, H.-L. & Hsu, C.-Y. Is one day off sufficient for re-adaptation to a daytime routine after two consecutive nights of work? *Ergonomics* **61**, 162–168 (2018).
24. Dorogush, A. V., Ershov, V. & Gulin, A. Catboost: Gradient boosting with categorical features support. In *Proceedings of the NeurIPS 2017 Workshop on Learning with Limited Labeled Data* (2018).

25. Rodrigues, F. & Correia, H. Semi-supervised and ensemble learning to predict work-related stress. *J. Intell. Inf. Syst.* **62**, 77–90, DOI: [10.1007/s10844-023-00806-z](https://doi.org/10.1007/s10844-023-00806-z) (2024).
26. Hossain, T. *et al.* Quantifying work-stress in surgical intensive care unit nurses: A multimodal activity-aware physiological analysis. In *Proceedings of the 2025 ACM International Symposium on Wearable Computers (ISWC '25)*, 127–133, DOI: [10.1145/3715071.3750406](https://doi.org/10.1145/3715071.3750406) (ACM, 2025).
27. Nägelin, M. L. *The Stress in Your Data: Towards the Automated Detection of Work-Related Stress*. Doctoral thesis, ETH Zurich (2023). DOI: [10.3929/ethz-b-000658582](https://doi.org/10.3929/ethz-b-000658582). ETH Diss. No. 29790.