Computer Methods and Programs in Biomedicine 279 (2026) 109287

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Towards lightweight stress monitoring on biometric data for IoMT environments☆

Carlos Montoya-Peña [b], Juan Francisco Gaitán-Guerrero [a], Macarena Espinilla [a], José L. López [a,*]

[a] Center for Advanced Studies in Information and Communication Technologies, University of Jaén, Jaén Higher Polytechnic School, Jaén, 23071, Spain
[b] Faculty of Engineering, Don Bosco University, School of Electronics, Soyapango, 1874, El Salvador

ARTICLE INFO

Keywords:
Stress monitoring
Biometric data
HRV
Respiratory signals
Healthcare
Machine learning
Internet of medical things

ABSTRACT

Background and objective: Stress is a physiological response mechanism that enables humans to react to perceived threats through a fight-or-flight response. While beneficial in acute situations, prolonged exposure to stress can lead to significant physical and mental health issues, making early and reliable detection essential. Although many existing approaches achieve high accuracy by relying on numerous physiological signals and features, such solutions are often unsuitable for Internet of Medical Things (IoMT) applications that increasingly rely on edge computing paradigms. In these scenarios, stress detection models must operate directly on resource-constrained devices with limited computational and energy budgets. Therefore, this work proposes a lightweight and efficient methodological framework for stress detection, specifically designed for edge-based IoMT deployment.

Methods: Eight supervised Machine Learning (ML) algorithms were evaluated: Random Forest (RF), LightGBM, CatBoost, XGBoost, Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and a Multilayer Perceptron (MLP). All models were trained using Heart Rate Variability (HRV) and respiratory features extracted from the WESAD dataset. The proposed framework combines population-level training with subject-specific adaptation and evaluates model performance under progressive dimensionality reduction using subsets of 15, 10, 8, 6, and 4 features.

Results: The proposed two-stage framework demonstrates that subject-specific adaptation significantly improves stress detection performance. XGBoost achieved the highest balanced accuracy (95.1% ± 4.7%) using 10 features, outperforming the configuration with all 15 variables. Crucially, the study identifies a reduced set of 6 features as the optimal deployment configuration; despite its further reduced feature set, it showed no statistically significant performance loss compared to the 10-feature model (95% CI: −0.0078, 0.0068) and maintained a 99.6% probability of outperforming the best models from all other architectures evaluated.

Conclusions: The results show that accurate and personalized stress detection is feasible using reduced feature sets, enabling efficient, interpretable, and real-time deployment of ML models in wearable and IoMT-based monitoring systems.

1. Introduction

Chronic stress represents a growing global concern that affects individuals of all ages, sex, and social backgrounds. It often stems from external pressures in occupational, academic and social settings, often exacerbated by economic instability and vulnerability to natural disasters [1].

Although workplaces have the potential to foster mental well-being and reduce workplace stress, they may also contribute to mental health risks, particularly for humanitarian, healthcare and emergency workers who face heavy workloads and frequent exposure to traumatic events [2].

Chronic stress can affect and modify human biology and behavior, from the molecular level to neural circuits, leading to conditions that

☆ This article is part of a Special issue entitled: 'Secure, smart, and high-performance healthcare systems' published in Computer Methods and Programs in Biomedicine.

* Corresponding author.
E-mail addresses: carlos.montoya@udb.edu.sv (C. Montoya-Peña), jgaitan@ujaen.es (J.F. Gaitán-Guerrero), mestevez@ujaen.es (M. Espinilla), llopez@ujaen.es (J.L. López).

https://doi.org/10.1016/j.cmpb.2026.109287
Received 2 October 2025; Received in revised form 10 February 2026; Accepted 14 February 2026
Available online 16 February 2026

increase the risk of developing mental disorders such as depression and anxiety [3].

Depressive and anxiety disorders are among the most prevalent mental health conditions, affecting 13.5% of women and 12.5% of men worldwide, and are responsible for 12 billion lost workdays each year. Alarmingly, around 8% of children and 14% of adolescents also live with a mental disorder. In the United States, research shows that half of all adult mental health disorders emerge by age 14, and 75% by age 24 [2].

These findings underscore the urgency of early detection and targeted mental health interventions, particularly during adolescence and early adulthood.

There are different testing protocols and methods to identify stress, which employ subjective or perception-based approaches, clinical laboratory tests, and noninvasive procedures to detect physiological parameters [4,5]

These stress states (or no stress) are typically induced, modulated, and assessed using protocols such as the Socially Evaluated Cold Pressor Test and the Socially Evaluated Stroop Test [6], height simulation scenarios through Virtual Reality [7], real-world driving tasks [8], and exposure to video games [9]. In addition, they can be observed under real-life conditions, particularly in patients with severe diseases [10].

Subjective methods for assessing stress consist of surveys that ask about the perceived level of difficulty or comfort in a given situation [11,12]. Recently, artificial intelligence-driven digital mental health tools have emerged as complementary alternatives to traditional subjective assessments, enabling stress and emotional state inference through natural language interaction, conversational analysis, and emotion-aware dialogue systems [13]. Clinical laboratory tests measure cortisol levels in saliva samples at specific time points. These samples can be collected between 20 min after the onset of the stressful situation, as this corresponds to the time required for the reaction to a stressor to be perceived [14].

However, for stress detection in real-world settings, it is essential to rely on non-invasive techniques that can be applied continuously and unobtrusively throughout daily activities [15] such as work, home life, or leisure. In this context, alternative noninvasive methods [16] assess stress states through physiological parameters such as Heart Rate (HR) [17], Respiration Rate [17], Electrodermal Activity, and HRV [18,19].

Currently, low-cost IoMT devices, particularly wearable technologies [20–23], enable continuous and real-time monitoring of the patient by collecting physiological signals on resource-constrained platforms, which inherently impose strict limitations on computation, memory, and energy consumption [17]. Within this framework, the nature of such infrastructures demands the use of lightweight models [24,25] deployed at the edge, capable of efficiently interpreting physiological signals, while maintaining low latency and reduced computational cost.

In order to support the development and evaluation of stress detection models, publicly available benchmark datasets such as WESAD [26] and SWELL-KW [27] have been widely adopted. In the work of Ramteke et al. [25], the WESAD dataset was analyzed using a hybrid Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) model, achieving an accuracy of 92.7%. Building upon this broader trend towards deep and hybrid architectures, Hole et al. [28] proposed a transfer learning-based bio-inspired ensemble model (TLBEMSE) for the preemptive detection of stress and emotional disorders using electroencephalography signals, reporting strong performance on the DEAP and INTERFACE datasets. Although effective, such deep and ensemble-based learning approaches typically involve increased architectural and computational complexity, which may limit their applicability in resource-constrained or real-world settings. More broadly, the existing literature predominantly presents complex models that demand large-scale datasets and high computational resources [22,23,29–31], and whose applicability in low-cost or resource-constrained environments has rarely been assessed.

In this context, the present work aims to address these challenges by proposing a novel two-stage subject-adaptive framework specifically designed for stress detection in resource-constrained IoMT environments. Unlike traditional comparative studies, this research contributes a systematic feature-reduction methodology that identifies the precise saturation point where predictive performance meets computational efficiency. By leveraging subject-specific adaptation and rigorous statistical validation – including Bootstrap and Nemenyi analysis – we demonstrate that it is possible to achieve statistical parity between complex architectures and highly optimized, low-dimensional models.

Other studies have employed ML algorithms such as RF to detect stress based on questionnaire assessments, as well as facial emotion recognition using SVM [32]. Additionally, models like XGBoost, LR, Decision Trees, RF, KNN, and SVM have been used to analyze HR, Respiration Rate, and skin conductance signals [15,17,20,24,33], often requiring high-cost procedures or the deployment of multiple sensors.

Numerous parameters can be extracted from HRV; for example, the SWELL-KW dataset includes up to 29 distinct features derived from this signal [34]. While this richness of information can be beneficial for modeling complex physiological responses, it also introduces challenges related to computational complexity, sensor availability, and real-time processing. Consequently, when developing systems for real-time and cost-efficient stress monitoring, it becomes essential to identify the most informative variables, as this strategy reduces the computational burden of signal processing and facilitates the design and deployment of wearable devices by minimizing the number of required sensors.

In addition, in healthcare applications, physiological signals, particularly electrocardiography (ECG) recordings [35], often exhibit pronounced inter-patient variability due to factors such as age, gender, body mass index, and genetic predispositions, which can significantly alter cardiac electrical activity across individuals. As a result, stress detection models trained solely on population-level data may fail to capture patient-specific electrophysiological patterns. To overcome these inherent inter-subject variability limitations, fine-tuning has emerged as a key transfer learning strategy for personalizing ML models in healthcare [29,30,36,37]. In the context of ECG signal processing [35], a pretrained general model is precisely calibrated to the specific biometric and physiological signal characteristics of an individual patient using a smaller, specialized dataset. This adaptation effectively balances population-level knowledge with patient-specific sensitivity, leading to improved prediction accuracy and personalization.

Consequently, this work presents a novel two-stage methodological framework for the development of optimized, subject-adaptive and lightweight models tailored to IoMT edge environments. By identifying the statistical saturation point of the feature space, this framework enables stress detection based on a minimalist set of features extracted from physiological signals. The main objective is to design efficient Artificial Intelligence models that can be deployed on IoMT devices with limited computational capacity and minimal sensing requirements, ensuring their applicability in real-world monitoring scenarios. This research presents a distinctive contribution by prioritizing computational minimalism and efficiency compared with existing solutions, which often rely on high-dimensional feature sets or computationally expensive pipelines. A key contribution is the demonstration of statistical parity between highly optimized low-dimensional models and their more complex counterparts, showing that maximum predictive performance can be maintained while significantly reducing input dimensionality.

Therefore, the main contributions of this article can be summarized as follows:

- the proposal of a novel two-stage subject-adaptive framework that integrates population-level knowledge with personalized fine-tuning, specifically designed to overcome inter-subject variability in physiological stress responses while improving patient-level accuracy and minimizing memory footprint and training overhead;

- the identification of a statistical saturation point within the feature space through rigorous dimensionality studies, demonstrating that a lean set of reduced features maintains statistical parity with high-dimensional models;
- a comprehensive bootstrap-based evaluation of models and reduced feature sets, providing robust confidence intervals and success probabilities, demonstrating the superior performance of Gradient Boosting machines using optimized feature subsets for stress detection in IoMT environments;
- and a hardware benchmarking study quantifying the trade-offs among inference latency, memory usage, and energy consumption, providing practical deployment guidelines for edge-based stress detection.

The structure of this article is as follows. Section 2 describes the employed datasets and the preprocessing pipeline applied to the physiological signals, as well as the design and training strategy of the proposed ML models, including feature selection and dimensionality reduction. Section 3 presents the experimental setup and provides a comprehensive analysis of the obtained results, with particular emphasis on performance-efficiency trade-offs and edge deployment considerations. Finally, Section 4 summarizes the main findings of the study and outlines potential directions for future research.

2. Materials and methods

This section describes the methodological framework employed to determine the patient's stress condition from physiological signals. It first details the selected dataset and the preprocessing pipeline applied to raw ECG and respiratory data in order to extract HRV and respiratory features. Next, the design and training strategy of the proposed ML models is presented, including population-level training using Leave-One-Subject-Out (LOSO) validation and subject-specific adaptation through fine-tuning. Finally, a systematic dimensionality reduction process is conducted to evaluate model performance under progressively reduced feature sets, allowing an in-depth analysis of the trade-offs between predictive accuracy, robustness, and computational efficiency in resource-constrained IoMT environments.

2.1. Dataset

In the present work, the WESAD dataset [26] was selected because it contains raw data from ECG, Blood Volume Pulse (BVP), and chest band signals, from which HRV parameters and respiratory rate can be extracted.

The dataset comprises physiological measurements collected from 15 subjects under three affective states: neutral, amusement, and stress. Notably, subject 2 also participated in an additional rest/meditation protocol, but this scenario was not taken into consideration. An additional binary label was created to distinguish between stress and no-stress conditions, thereby framing the problem as a binary classification task to determine whether a subject is experiencing stress.

2.2. Data preprocessing

The dataset was first preprocessed by removing non-predictive variables such as timestamps, raw signal data, labels, and subject identifiers. Remaining features included statistical and spectral HRV parameters extracted from ECG and respiratory signals. Missing values were removed prior to model training.

To develop the model, R-peaks in the ECG signal were detected using the Pan-Tompkins algorithm. The ECG signals reflect the voltage oscillations or electrical activity of the heart for each subject. The synchronized data for each subject is stored in files named SX.pkl, where X denotes the subject identifier (e.g., S2.pkl, see Fig. 1). Within

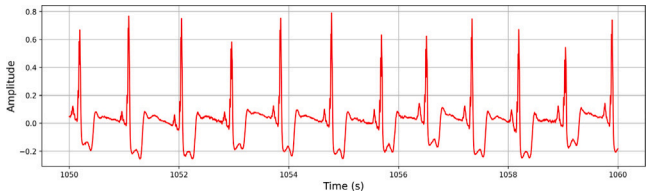


Fig. 1. Segment of ECG signal file S2.pkl (1050s to 1060s).

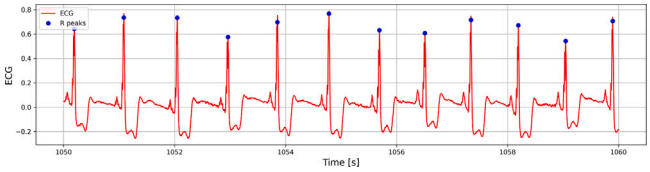


Fig. 2. Example of a 10-second ECG segment with detected R-peaks highlighted in blue.

these signals, the main inflection points were identified, with R-peaks being among the most prominent and readily distinguishable features.

A first-order Butterworth filter was applied to preserve only frequencies within the 5 Hz to 15 Hz range, where most of the useful information of the ECG signal is located. Subsequently, a derivative stage was applied to highlight abrupt changes in the signal, mainly corresponding to the R-peaks, followed by squaring the signal to ensure all values are positive. The resulting signal was then smoothed using a 150 ms moving average filter to facilitate the final identification of the R-peaks. These peaks were marked by adding a new column to the dataset, where the rows marked as a R-peak resulted in a subset of 44,207 records from the original 4,255,330 entries.

After detecting the R-peaks within the ECG signal (see Fig. 2), the instantaneous HR was calculated by measuring the inter-beat interval (ibi), defined as the time difference between two consecutive R-peaks.

$$\text{ibi}_i = t_{i+1} - t_i \tag{1}$$

where t_i and t_{i+1} represent the timestamps (in seconds) of two consecutive R-peaks.

The average HR was then derived by computing the mean distance between these R-peaks. This process enables the extraction of key HRV features essential for subsequent stress analysis.

$$\text{HR}_i = \frac{60}{\text{ibi}_i} \tag{2}$$

These calculations form the foundation for extracting HRV features. A moving window of 50 R-peaks was used to compute a set of time-domain and frequency-domain metrics, including sdnn, rmssd, sdsd, pnn20, pnn50, mean_rr, median_rr, sdrr_rel_rr, vlf, lf, hf, and the lf/hf ratio. These features are essential for evaluating the subject's stress level.

Similarly, HR can be measured applying a filter to a photoplethysmography (PPG) or BVP signal, a technique that, through the reflection of incident light on body tissue, provides information about changes in blood volume and pulse, thereby describing the mechanical activity of the heart (see Fig. 3).

For respiratory rate identification, a similar peak detection and filter procedure was applied to determine the average respiratory frequency, setting a minimum peak height above the 80th percentile and a minimum interval of 1.5 s to avoid false peaks caused by noise. The detected peaks were marked in a new column (see Fig. 4).

After extracting 15 HRV parameters, the results for all subjects were concatenated into a single dataframe and saved as one file. This dataset contains relevant records only for R-peaks; the other records without identified peaks do not include HRV parameter information and were therefore removed. The resulting dataset comprises 44,207 records.

The pipeline of the pre-processing process is shown in Fig. 5 below.

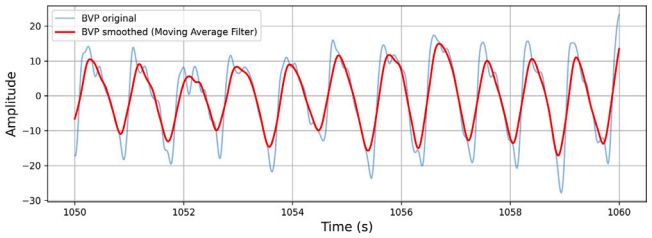


Fig. 3. Example of a 10-second BVP segment with the original signal and its smoothed version after a moving average filter.

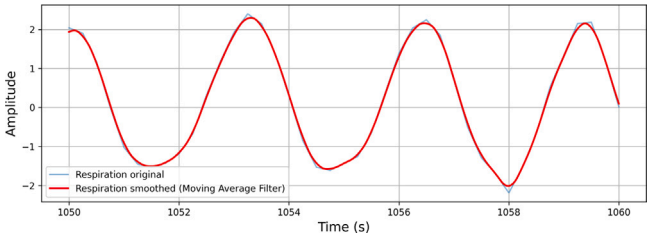


Fig. 4. Example of a 10-second respiration segment with the original signal and its smoothed version after a moving average filter.

2.3. Model architecture and training strategy

In order to evaluate the suitability of different ML approaches for stress detection in resource-constrained IoMT environments, a structured training and evaluation strategy was designed. This strategy aims to analyze both the generalization capability of population-level models and their potential for personalization at the patient level, while accounting for inter-subject physiological variability. To this end, multiple classification algorithms were implemented and assessed under a unified validation framework, allowing a fair comparison of their performance, robustness, and computational requirements.

Following this evaluation strategy, an initial analysis was conducted using LOSO validation, in which eight binary classification models were implemented and evaluated to predict stress states based on physiological variables, primarily HRV metrics and respiratory rate. The dataset was first preprocessed by removing non-predictive variables such as timestamps and raw signal data. In each LOSO iteration, all samples from one subject were held out for testing, while the remaining subjects were used for training, enabling an assessment of population-level generalization.

To identify the most suitable architectures for stress detection, a benchmark including RF, XGBoost, LightGBM, CatBoost, LR, SVM, KNN, and MLP was defined. Each algorithm was configured using a set of representative hyperparameters selected to balance predictive performance, model complexity, and computational efficiency. To ensure a robust and reproducible comparison, hyperparameters for all models were determined through a two-step selection process. First, a heuristic search was performed based on established benchmarks for resource-constrained IoMT environments to prioritize model economy. Second, these values were refined via a preliminary sensitivity analysis using a subset of the training data (independent of the final LOSO test subjects). In the next stage, all the models followed a subject-specific adaptive modeling strategy where the first 500 samples per class from each held-out subject were used for subject-specific calibration, simulating a deployment scenario in which a brief initial calibration period is available for a new user. Feature preprocessing included mean imputation and standard scaling, and reduced feature subsets ranging from 4 to 15 HRV and respiratory metrics were evaluated. This two-stage framework combines population-level model training with subject-specific fine-tuning, enabling robust, efficient, and personalized stress inference. The following two sections present the description for both validation and training approaches.

2.3.1. Population-level training with LOSO validation

A RF classifier was selected as the initial population-level model due to its strong performance during cross-validation, where it consistently demonstrated high accuracy and robustness. The model was configured with 100 decision trees and a fixed random seed to ensure reproducibility. Each tree was trained on a bootstrap sample of the training data and employed random feature selection at each split, promoting model diversity and reducing variance across the ensemble. In addition, a class-balancing strategy was applied to mitigate the inherent imbalance between stress and non-stress samples in the dataset, ensuring fair learning across both classes.

The second model was based on the XGBoost algorithm, an ensemble method that builds additive decision trees and iteratively optimizes a loss function using gradient descent. The XGBoost pipeline is configured with a set of hyperparameters optimized for efficient learning and generalization across subjects. For each feature set, an XGBClassifier was trained using 100 trees, a maximum depth of 6 and a learning rate of 0.1. Using eval_metric='logloss' ensures probabilistic outputs suitable for threshold calibration per subject.

For the architecture based on the KNN algorithm, prior to training, all features were standardized using StandardScaler within a preprocessing pipeline, which is critical for KNN due to its reliance on Euclidean distances. Rather than using a fixed parameter, the number of neighbors was dynamically optimized via Grid Search among $k \in \{3, 5, 7, 9\}$, selecting the value that maximized the balanced accuracy for each specific feature set.

This approach allowed for the comparison of classifier performance as a function of feature dimensionality, contributing to an informed assessment of variable importance in the context of binary classification.

A MLP neural network was designed and trained using TensorFlow/Keras. The sequential architecture consisted of an input layer matching the number of input features, followed by four fully connected hidden layers with 128, 64, 32 and 16 neurons respectively, all using ReLU activation functions. The output layer was a single neuron with a sigmoid activation function, suitable for binary classification tasks. The model was trained using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy as the loss function. Prior to training, the input features were standardized using a standard scaler. For each LOSO fold, the model was re-initialized and trained for 20 epochs to prevent data leakage and ensure independent evaluations across subjects.

In contrast to the previous models, the CatBoost implementation leverages a set of hyperparameters optimized for stable, efficient training and robust subject-specific adaptation. The model uses 100 boosting iterations, a tree depth of 6, and a learning rate of 0.1, balancing model complexity and generalization across subjects. Logloss is selected as the objective function for binary classification, while class imbalance is handled via dynamically computed class weights for each training fold.

On the other hand, LR was implemented using a structured pipeline that integrates data imputation, feature scaling, and model training to ensure reproducible preprocessing and stable learning. The model was configured with L2 regularization to mitigate overfitting and optimized using the SAGA solver, which was selected for its efficiency in high-dimensional feature spaces and its support for warm-start initialization. This configuration ensured stable convergence over a maximum of 1000 iterations during the population-level training, while providing the necessary framework for subject-specific adaptation in subsequent stages. Additionally, a class-balancing strategy was applied to compensate for unequal distributions between stress and non-stress samples.

Similarly, the LightGBM model was configured to balance learning capacity, generalization ability, and computational efficiency. The architecture was trained using 100 boosting iterations, with a learning rate of 0.1 to control the contribution of each tree and promote stable convergence. Tree depth was limited to 6 levels to reduce overfitting,

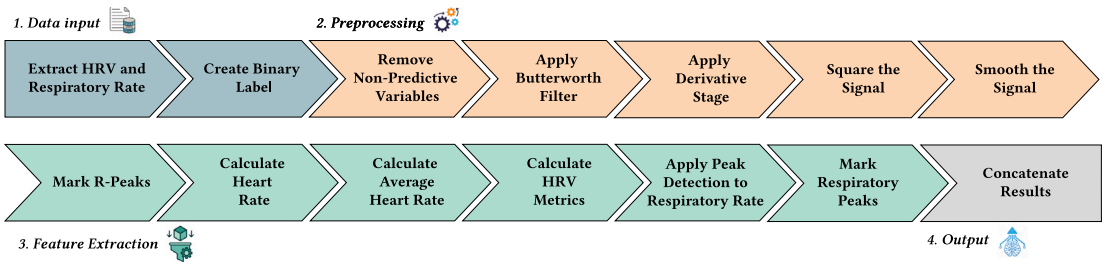


Fig. 5. Workflow of the preprocessing pipeline.

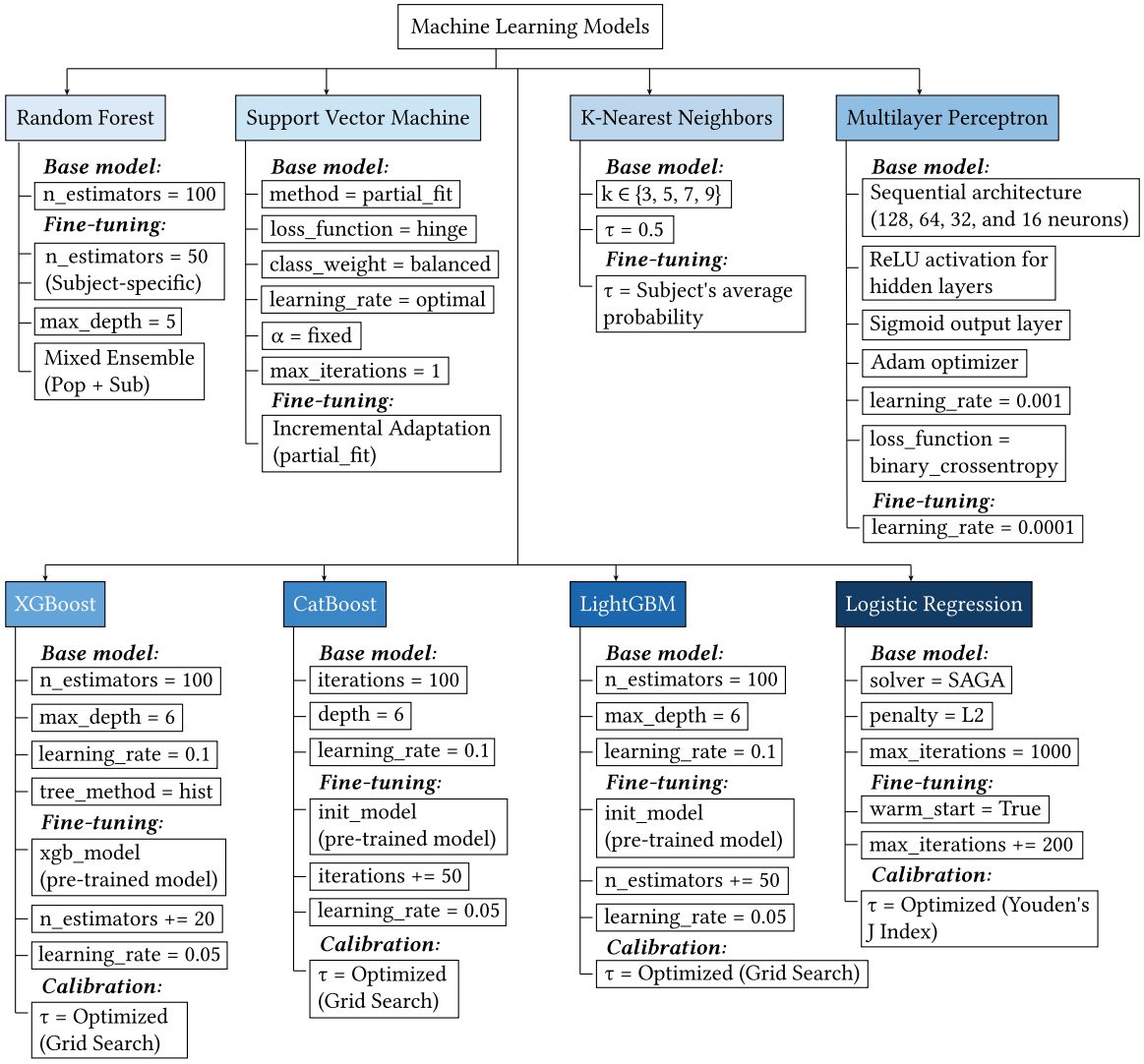


Fig. 6. Summary of the ML models evaluated in this study (RF, SVM, KNN, MLP, XGBoost, CatBoost, LightGBM, and LR), showing the population level base configuration and the corresponding subject specific fine tuning and calibration modifications applied within the proposed adaptive framework. The figure highlights the main hyperparameters, adaptation strategies, and decision threshold optimization procedures used for each model.

while the maximum number of leaves per tree was set to 31, allowing sufficient model flexibility without excessive complexity. In addition, class imbalance was addressed through dynamically adjusted class weights, ensuring proportional consideration of stress and non-stress samples across training folds.

Finally, the SVM model was implemented within a unified preprocessing pipeline that included mean imputation and feature standardization to ensure stable training behavior. The architecture utilized a Stochastic Gradient Descent (SGD) optimizer with a hinge loss function, effectively implementing a linear SVM. This approach was selected to favor computational efficiency and to facilitate the interpretation of the decision boundary while enabling incremental learning across folds. Class imbalance was handled through a balanced weighting strategy, with the effective regularization adjusting the penalty applied to misclassifications according to class frequencies.

The implementation of these eight classification models provided a comprehensive comparison of diverse ML approaches for stress detection. By leveraging both traditional algorithms and neural architectures, the study was able to explore the trade-offs between model complexity, interpretability, and predictive performance. Standardized

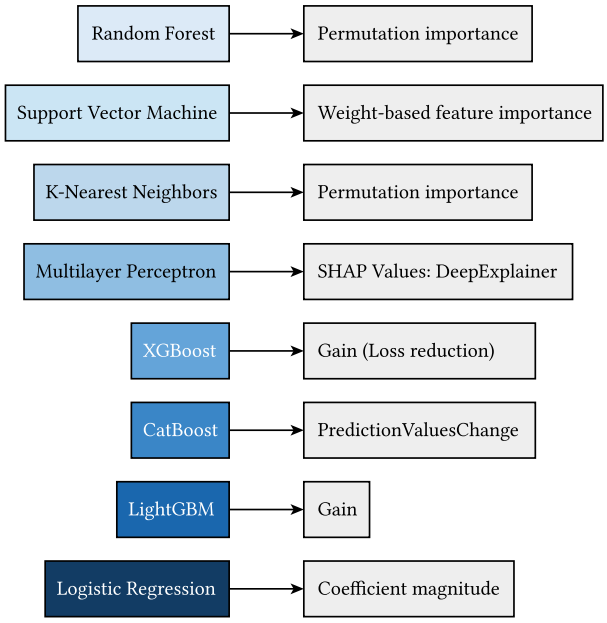


Fig. 7. Dimensionality reduction methods applied to each ML model.

preprocessing and consistent data partitioning ensured fair comparisons across models. Overall, the inclusion of explainability techniques and systematic feature reduction across experiments contributed to a robust evaluation framework, paving the way for reliable deployment of stress classification systems based on physiological signals.

For clarity, a summary of the selected hyperparameters for each of the implemented ML models is provided in Fig. 6.

2.3.2. Subject-specific fine-tuning

To address the inter-subject physiological variability inherent to the WESAD dataset, a fine-tuning-based personalization strategy was implemented within the LOSO cross-validation framework. Following population-level training, each model was adapted through a balanced calibration process using an equal number of stress and non-stress samples from the target subject. This strategy enables equitable parameter adjustment, either by updating model weights in linear classifiers or by incrementally refining estimators in boosting-based models, thereby mitigating bias towards the majority class. Through this subject-specific adaptation, the models are better aligned with individual physiological patterns while preserving the general knowledge acquired during population-level training.

The results of four ML architectures – three Gradient Boosting variants (XGBoost, CatBoost and LightGBM), and a linear LR baseline – are presented here, having been assessed using a standardized incremental learning framework. The tree-based models shared a base configuration of 100 estimators, depth 6, and a 0.1 learning rate. Subject-specific personalization was implemented via fine-tuning (20 additional estimators for XGBoost; 50 for CatBoost/LightGBM) at a reduced learning rate of 0.05. For LR, adaptation utilized a warm-start strategy with L2 regularization ($C = 1.0$) and a SAGA solver, refining population-level coefficients over 200 iterations on subject-specific data. Complexity was controlled through structural penalties (L1/L2) and specific growth strategies (symmetric for CatBoost and leaf-wise for LightGBM) to mitigate overfitting during adaptation. Crucially, across all architectures, decision thresholds were dynamically optimized using Youden's Index (or exhaustive search) derived from subject-specific ROC analysis to maximize Balanced Accuracy. This dual adaptation of model parameters and decision boundaries enables refined residual fitting to individual patterns while preserving the robustness of population-level knowledge.

Complementing the aforementioned boosting and logistic frameworks, four additional architectures, RF, MLP, KNN, and a Stochastic Gradient Descent SVM (SVM-SGD), were integrated into this adaptive pipeline. For the RF model, a mixed ensemble strategy was employed, combining the population-level $\text{Model}_{\text{pop}}$ (100 trees, depth 10) with a subject-specific $\text{Model}_{\text{sub}}$ (50 trees, depth 5), where final inference resulted from the averaged probabilistic output of both estimators. The MLP architecture followed a sequential fine-tuning approach; the weights initialized during the LOSO phase were refined through 15 additional epochs using a significantly reduced learning rate (0.001) to prevent catastrophic forgetting of population-level features. In contrast, the KNN model utilized an instance-based adaptation where the neighborhood parameter $k \in \{3, 5, 7, 9\}$ was dynamically optimized via Grid Search for each subject, and the decision boundary was recalibrated using a localized threshold τ derived from the subject's specific feature distribution. Finally, the SVM-SGD implementation leveraged incremental learning through a partial-fit mechanism, updating the linear hyperplane coefficients using the subject-specific calibration data with an optimal learning rate schedule. Across these models, the same protocol of 500 calibration samples per class was strictly maintained, ensuring that the transition from generalized population knowledge to personalized inference was consistent across the entire benchmark of eight classification approaches.

For clarity, the specific modifications applied during the fine-tuning stage with respect to the population-level models are summarized in Fig. 6.

2.4. Dimensionality reduction

To assess the impact of feature dimensionality on classification performance, the models were each evaluated using five progressively smaller subsets of input variables: 15, 10, 8, 6, and 4 features. These subsets were derived from feature importance rankings obtained through preliminary evaluations with each respective model. This systematic reduction enabled a consistent comparison of model robustness and generalization under varying levels of input information.

The following describes the methodology used for all the models.

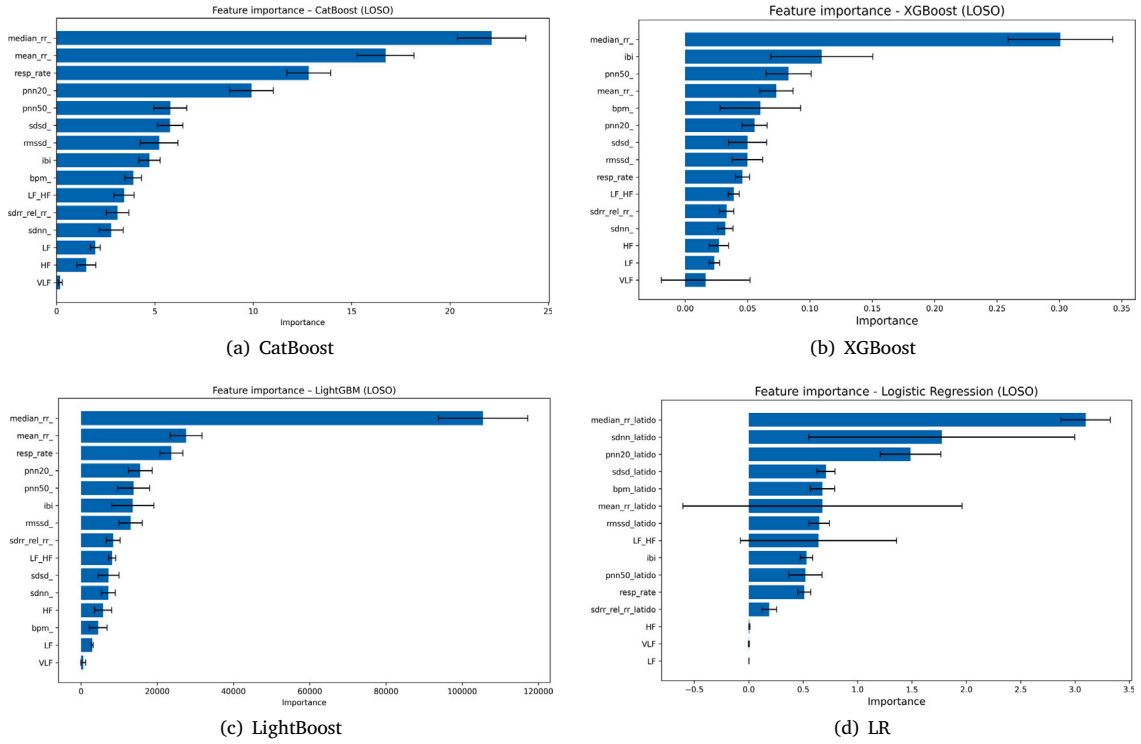
Feature importance values are calculated fold-wise within a LOSO cross-validation scheme, capturing the relevance of each variable for predictions on each left-out subject. These values are subsequently aggregated by computing the mean and standard deviation across all folds, yielding a final ranking based on their average impact on model performance. This approach identifies the most influential physiological variables, which are visualized through horizontal bar plots, with error bars reflecting the variability across subjects (see Fig. 8).

For the XGBoost model, variable importance was obtained from the feature_importances_ attribute of the trained model.

In CatBoost, the feature importance ranking is obtained using get_feature_importance method, which estimates the contribution of each predictor to the reduction of the model's loss during training.

For LightGBM model, the feature importance ranking is obtained using the feature_importance method of the trained booster, with importance_type='gain', which measures the total contribution of each feature to the reduction of the loss function across all trees.

Variable importance for the LR model is derived from the model's learned coefficients. For each fold of the LOSO cross-validation, the model is trained on standardized training data using a pipeline that combines feature scaling and L2-regularized LR with balanced class weights. The coefficients corresponding to each feature are extracted for each fold, and the absolute values are taken to reflect the magnitude of influence regardless of direction. Fold-wise coefficients are then aggregated by computing the mean and standard deviation across all folds, resulting in a final ranking of features according to their average contribution to the model's decision boundary. Features with larger absolute coefficients are considered more influential.



(a) CatBoost (b) XGBoost

(c) LightBoost (d) LR

Fig. 8. Model-based feature importance for stress classification. Subfigure (a) corresponds to CatBoost, (b) to XGBoost, (c) to LightBoost, and (d) to LR, showing which metrics emerge as the most relevant features across models.

The hierarchy of variables for the RF model was determined through permutation importance, measuring performance degradation upon feature randomization to identify critical biomarkers. For the MLP, SHAP (SHapley Additive exPlanations) values were employed via the DeepExplainer integrator, enabling the decomposition of each physiological signal's individual contribution to the neural network's output. Given that the KNN model lacks internal relevance parameters, a permutation importance framework under Nested LOSO cross-validation was applied, assessing the impact of each variable on the cohesion of stress clusters within the feature space. Finally, for the linear SVM, importance was derived directly from the absolute magnitude of the hyperplane coefficients ($|w_i|$), where higher weights indicate a superior influence on the definition of the decision boundary. To ensure the robustness of these findings, the results from each architecture were normalized and averaged across all subjects within the LOSO scheme, yielding an aggregated importance metric that mitigates individual bias and highlights biomarkers with the greatest population stability.

Despite sharing this validation framework and standardized preprocessing, the interpretation of importance presented technical particularities according to the architecture of each model. LR was distinguished by the use of linear coefficient magnitudes following data normalization with standard scaling, allowing the absolute value of each weight (β) to reflect the predictive power of the variable. In contrast, decision tree-based models employed gain and sensitivity metrics; while XGBoost and LightGBM quantified importance through Gain – measuring the cumulative reduction of uncertainty at decision nodes – CatBoost provided an influence measure based on prediction value changes. This combination of approaches allowed for contrasting biomarker relevance from both linear contribution and non-linear interaction perspectives, using the standard deviation as a critical indicator of the stability of each signal against the biological individuality of the subjects. The combined use of model-specific and model-agnostic techniques for feature importance assessment allowed for a thorough evaluation of variable relevance across diverse classification algorithms. By systematically analyzing how input dimensionality affects performance, this approach facilitated the identification of a compact subset of key physiological features that maintain high predictive accuracy. Such insights not only enhance model interpretability but also contribute to the development of more efficient and robust stress detection systems based on physiological signals.

The dimensionality reduction techniques applied to each model are summarized in Fig. 7.

3. Experimental results

This section presents the experimental design and results of the proposed methodology. First, the experimental setup is described, including the training and testing protocol, specifically the LOSO protocol and the subject-specific fine-tuning stage, alongside the evaluation metrics, and parameter configuration for each ML model. Next, the results are reported and analyzed, highlighting the performance of the different algorithms, the relevance of the most informative features, and the comparative analysis between models. Finally, the findings are discussed in terms of their implications for stress detection in real-world IoMT scenarios.

3.1. Dataset selection and justification

As mentioned in a previous section, the WESAD dataset was used in this study due to its high quality and the richness of physiological signals relevant to stress detection. It includes raw ECG data, which was preferred because of its high sampling rate of 700 Hz, allowing precise identification of R-peaks and accurate HRV feature extraction. Additionally, it provides raw PPG signals sampled at 64 Hz a lower rate but still suitable for developing models intended for real-time applications using wearable devices. The dataset also contains data from a chest-worn respiratory band, sampled at 700 Hz, enabling the estimation of respiratory rate. Together, these signals support the extraction of

Table 1
Performance of binary classification models showing highest and lowest balanced accuracy configurations under LOSO cross-validation.

Model	n_features	Balanced accuracy (%)	Sensitivity (%)	Specificity (%)	Macro-F1 (%)
XGBoost	10	80.60 ± 11.96	75.07 ± 24.87	86.13 ± 21.06	79.72 ± 13.97
XGBoost	4	79.07 ± 13.85	69.95 ± 27.92	88.19 ± 20.88	78.25 ± 16.09
LightGBM	10	80.44 ± 12.16	77.49 ± 23.24	83.39 ± 22.66	79.08 ± 14.72
LightGBM	4	78.94 ± 15.77	78.44 ± 24.59	79.43 ± 24.22	77.23 ± 17.80
CatBoost	10	81.25 ± 12.56	78.69 ± 23.57	83.82 ± 23.97	79.82 ± 15.37
CatBoost	4	80.40 ± 13.46	78.76 ± 25.71	82.05 ± 24.38	78.48 ± 16.43
LR	15	80.33 ± 13.59	75.72 ± 28.76	84.93 ± 22.32	78.79 ± 15.81
LR	10	79.56 ± 14.22	73.93 ± 30.14	85.19 ± 21.75	78.13 ± 16.38

both HRV and respiratory parameters, which are widely recognized as indicators of autonomic nervous system activity and stress response.

The dataset comprises physiological recordings from subjects exposed to three well-defined affective states: neutral, amusement, and stress, under controlled experimental conditions. This clear labeling supports the development and evaluation of supervised classification models. Moreover, access to raw signals allows full control over preprocessing and feature extraction steps, ensuring consistency and transparency throughout the analysis pipeline.

3.2. Definition of experiments

To address the inter-subject physiological variability inherent to affective state recognition, an evaluation framework based on LOSO cross-validation was implemented using the WESAD dataset. This approach ensures that each model is evaluated on subjects whose responses are entirely unknown during the population-level training phase, thereby simulating deployment under real-world conditions. While eight models were evaluated – RF, LR, SVM, XGBoost, CatBoost, LightGBM, KNN, and MLP – the research focused on the four algorithms showing a tendency towards better performance.

The core of the proposed methodology consists of a two-phase transfer learning architecture: first, population-level training to extract general stress patterns, followed by a balanced calibration protocol using a subset of 1000 samples from the initial recordings of both classes for each test subject. This fine-tuning process enables adjustment of the decision boundary to the individual's specific baseline, employing differentiated technical adaptation mechanisms according to each algorithm's nature.

For Gradient Boosting models (XGBoost, CatBoost, and LightGBM), personalization was executed through incremental learning by adding additional estimators with a reduced learning rate of 0.05 to correct population-level residuals without degrading global knowledge. For LR, a warm-start technique was employed with the SAGA optimizer, using population-level coefficients as initial values for local optimization adapted to the new individual's distribution.

To maximize robustness, the classification threshold (τ) was determined dynamically for each subject rather than being statically fixed at 0.5. For boosting models, a threshold sweep was performed to identify the maximum Balanced Accuracy, while for LR, Youden's Index derived from the ROC curve was used. This adjustment enables the system's sensitivity to adapt to the intensity and noise characteristics of each person's signals.

Finally, to evaluate system viability on resource-constrained devices, each experiment was systematically replicated across five decreasing dimensionality configurations (15, 10, 8, 6, 4) based on a feature importance ranking, thereby analyzing performance sensitivity to progressive reduction of predictive features.

The mean and standard deviation of each metric across the fifteen folds of LOSO cross-validation were computed to analyze the stability and variability of performance.

This robust methodology enabled the comparison and selection of models with better predictive capability and generalization, ensuring the statistical validity of the results presented.

Overall, these findings highlight the effectiveness of tree-based models for this classification task, while the acceptable performance of alternative classifiers offers flexibility depending on specific precision–recall trade-offs required in practical applications.

As shown in Table 1, the four models demonstrating the highest performance consistency and statistical stability were selected for evaluation across a LOSO cross-validation scheme to assess their suitability for model adaptation. The Balanced Accuracy and Macro-F1 played a crucial role in the evaluation due to the class imbalance in the dataset. While the former ensured a fair assessment of the model across both classes, the latter was essential to balance precision and recall in the context of stress detection.

3.2.1. Edge computing performance analysis

The inference tests were conducted on a Raspberry Pi 5, featuring a 64-bit ARM Cortex processor architecture (aarch64) with four independent physical cores. The system operates at a maximum clock frequency of 2.4 GHz and a minimum of 1.5 GHz, and is equipped with 8 GB of RAM. Additionally, the processor includes a cache hierarchy comprising 2 MiB of L2 cache and 2 MiB of L3 cache. All experiments were executed on a CPU-based edge platform using Python 3.12, with an environment built on NumPy 2.0.2, scikit-learn 1.6.1, and Keras 3.10 with a TensorFlow 2.19 backend, ensuring compatibility with modern software stacks.

Inference benchmarking focused on key deployment metrics for the four selected models was conducted, including latency, memory consumption, current and model size. Latency measurements were obtained by averaging results over 1000 inference runs with a batch size of one, simulating real-time, sample-by-sample processing in an IoMT scenario. Memory usage was monitored using psutil, while model size was assessed through serialized storage measurements, allowing a comprehensive evaluation of computational efficiency on resource-constrained hardware.

To ensure high-fidelity memory metrics, the measurement script implements an isolation technique using independent subprocesses. This approach prevents memory fragmentation and ensures that the footprint of one model does not affect the next. For each model, two distinct memory values are reported: Total RAM and Net RAM. Total RAM is obtained by measuring the Resident Set Size (RSS) using the psutil library after the full model stack is loaded. Net RAM is defined as the exclusive memory overhead of the model, calculated by subtracting a baseline footprint, which comprises the Python interpreter and pre loaded inference engines, from the Total RAM.

Regarding temporal performance, a warm-up phase of 20 iterations was implemented to stabilize CPU frequency and cache states. Latency was captured with milliseconds precision using time.perf_counter, reporting average values. To ensure stable current readings (I_{max}), the experimental protocol includes a 5-second pre-execution pause to allow for manual instrument synchronization, followed by an inference loop maintained at high CPU occupancy. This procedure enables the measurement to integrate the effective current over 1000 samples, thereby mitigating idle-state fluctuations relative to the 0.54 A baseline current. By sustaining this workload, the setup ensures that the captured current draw accurately reflects the model's peak operational demand on the Raspberry Pi 5 platform.

Table 2
XGBoost performance across different feature configurations under LOSO cross-validation with subject-specific fine-tuning.

n	BA (%)	Sens (%)	Spec (%)	M-F1 (%)
15	94.3 ± 5.4	95.1 ± 4.7	93.5 ± 8.9	92.5 ± 8.8
10	95.1 ± 4.7	95.3 ± 5.5	94.9 ± 5.7	93.7 ± 6.5
8	94.8 ± 5.0	94.7 ± 7.1	95.0 ± 6.6	93.5 ± 6.8
6	95.0 ± 4.6	95.7 ± 5.8	94.3 ± 6.1	93.2 ± 6.7
4	93.7 ± 6.5	94.2 ± 8.4	93.3 ± 8.6	91.9 ± 9.1

Table 3
CatBoost performance across different feature configurations under LOSO cross-validation with subject-specific fine-tuning.

n	BA (%)	Sens (%)	Spec (%)	M-F1 (%)
15	89.1 ± 10.8	96.7 ± 7.7	81.5 ± 17.3	81.8 ± 17.5
10	89.0 ± 10.3	95.6 ± 10.5	82.4 ± 15.5	81.6 ± 16.8
8	88.4 ± 10.3	96.0 ± 7.9	80.8 ± 18.0	81.1 ± 16.9
6	88.0 ± 9.8	95.7 ± 7.9	80.4 ± 16.8	80.4 ± 16.1
4	87.5 ± 11.4	96.0 ± 10.9	79.0 ± 16.3	78.8 ± 17.8

Table 4
LightGBM performance across different feature configurations under LOSO cross-validation with subject-specific fine-tuning, ranked by Balanced Accuracy.

n	BA (%)	Sens (%)	Spec (%)	M-F1 (%)
15	87.8 ± 11.8	94.1 ± 12.6	81.6 ± 16.2	79.3 ± 17.8
10	89.2 ± 9.8	96.0 ± 9.2	82.3 ± 17.0	81.2 ± 16.5
8	88.5 ± 9.8	98.5 ± 2.9	78.4 ± 18.8	79.7 ± 17.1
6	88.2 ± 8.5	96.8 ± 7.8	79.5 ± 15.8	78.9 ± 15.3
4	85.3 ± 11.3	98.4 ± 3.6	72.1 ± 21.2	76.3 ± 17.5

Table 5
LR performance across different feature configurations under LOSO cross-validation with subject-specific fine-tuning.

n	BA (%)	Sens (%)	Spec (%)	M-F1 (%)
15	91.7 ± 5.8	91.2 ± 9.8	92.2 ± 10.8	90.2 ± 8.6
10	90.9 ± 7.3	90.5 ± 10.6	91.3 ± 14.0	89.6 ± 10.5
8	90.9 ± 8.4	91.0 ± 11.0	90.8 ± 15.7	88.8 ± 12.6
6	90.7 ± 8.3	91.4 ± 11.2	90.0 ± 15.7	88.6 ± 12.5
4	90.2 ± 8.0	89.4 ± 11.0	91.0 ± 15.6	88.5 ± 12.3

Table 6
Subject-level Bootstrap Analysis: XGB_6 (Rank: 9.73) vs. Inter-family Models and Intra-family Ablation Variants.

Comparison (Ref: XGB_6)	Rank	95% CI	$P(>0)$
Inter-Family (Best)			
vs. RF_{10} (Bagging)	16.53	[0.012, 0.098]	0.997
vs. LR_{15} (Linear)	17.73	[0.014, 0.050]	1.000
vs. MLP_{10} (Neural Network)	18.93	[0.013, 0.093]	0.997
vs. SVM_8 (Kernel-based)	19.67	[0.015, 0.075]	1.000
vs. KNN_6 (Instance-based)	20.27	[0.026, 0.098]	1.000
vs. $CatB_{15}$ (Gradient Boosting)	21.63	[0.015, 0.106]	0.997
vs. $LGBM_6$ (Gradient Boosting)	24.10	[0.032, 0.098]	1.000
Intra-Family (Ablation)			
vs. XGB_8	9.27	[−0.006, 0.011]	0.618
vs. XGB_{10}	9.63	[−0.008, 0.007]	0.383
vs. XGB_4	11.77	[−0.002, 0.030]	0.954
vs. XGB_{15}	12.47	[−0.005, 0.023]	0.808

3.2.2. Feature ablation and statistical saturation analysis

To assess the impact of feature dimensionality on both predictive robustness and hardware efficiency, a comparative statistical ablation study was conducted across the five distinct feature-set sizes. This experimental setup evaluates 40 unique model configurations, comprising eight classifiers and five feature subsets, under identical subject level cross validation conditions. Initial performance differences were validated using Friedman and post-hoc Nemenyi tests, followed by pairwise Wilcoxon signed-rank comparisons with Holm correction to identify the minimum viable feature subset that maintains statistical parity with the best observed balanced accuracy. Following the principle of model economy, the primary objective is to pinpoint the statistical saturation point where further feature reduction leads to significant accuracy degradation while offering diminishing returns in hardware savings. To ensure the reliability of these findings beyond simple point estimates, a subject-level bootstrap resampling protocol with 10,000 iterations was established, utilizing Balanced Accuracy as the lead metric. This statistical framework allows for the calculation of Mean Deltas and 95% Confidence Intervals (CI), alongside the success probability $P(\text{diff} > 0)$, to quantify the performance gap between configurations. Finally, the experimental protocol integrates a real-time benchmarking phase on a Raspberry Pi 5 platform to measure the operational cost of each ablation level, specifically quantifying inference latency, memory footprint, and current draw, thereby providing a multi-objective validation of the framework's suitability for edge-based IoMT deployment.

3.3. Results

This section presents the findings obtained through LOSO cross-validation, evaluating both the baseline generalization capability and the impact of subject-specific refinement. Initially, the performance of all the models was assessed under a standard LOSO framework using reduced feature subsets (the best and worst performance of XGBoost, LightGBM, CatBoost, and LR models is shown in Table 1. Subsequently, a subject-specific fine-tuning stage was implemented, as summarized in Tables 2, 3, 4, 5, and 7. The data reveal a substantial improvement across all classification metrics upon integrating personalized adjustment, highlighting a notable reduction in variability (standard deviation) and a general increase in balanced accuracy, positioning XGBoost as the most robust model with performance exceeding 95% in optimal configurations.

The experimental results obtained on the Raspberry Pi platform using the two-stage testing protocol provide a comprehensive comparison of the evaluated models in terms of predictive performance and computational efficiency.

XGBoost emerges as the most accurate architecture for this classification task, achieving a peak balanced accuracy of 95.09% and a macro F1-score of 93.68% when using the top 10 features (Table 2). This performance is largely driven by a feature priority ranking dominated by the median_rr interval and the ibi (Fig. 8). However, this high predictive capability comes at the cost of increased computational demand, as XGBoost exhibits a relatively high inference latency compared to other evaluated models, such as LightGBM, reaching 0.945 ms for XGB_6, and 1.112 ms for XGB_{10} (Table 7).

In contrast, LR stands out for its remarkably small storage footprint ranging from 2.04 KB to 2.70 KB. Its variable ranking emphasizes coefficients such as the median_rr, sdnn, and pnn20 (Fig. 8), demonstrating that competitive performance can be achieved with a reduced set of linear features. Notably, this model attains a balanced accuracy above 90.00% while maintaining a I_{max} of approximately 1.030–1.100 A. On the other hand, CatBoost offers a highly balanced profile optimized for real-time inference, maintaining a consistent model size of approximately 178 KB. Its feature importance analysis highlights the median_rr, mean_rr, and resp_rate (Fig. 8) as the most influential predictors, enabling the model to sustain balanced accuracy values between 87.49% and 89.12% while keeping inference latency low, ranging from 0.329 ms to 0.638 ms. All these results are summarized in Table 7.

Finally, LightGBM demonstrates stable net memory consumption, averaging 2.34 MB. (Table 7). Although its feature importance ranking is similar to the CatBoost model when using four features, its Balanced Accuracy is significantly lower, at 85.27% (Table 3). The observed inference latency, ranging from 0.260 ms to 0.279 ms (Table 7), indicates that its single-sample processing speed on ARM-based architectures is the most optimized among all evaluated models in this implementation.

Table 7
Real-time inference performance bench marking on Raspberry Pi 5: Size, average latency and memory, and max current.

Model	Size (KB)	Lat. (ms)	Mem_{tot} (MB)	Mem_{net} (MB)	I_{max} (A)
$CatBoost_4$	177.08	0.330	183.36	2.58	1.320
$CatBoost_6$	177.43	0.388	183.36	2.58	1.320
$CatBoost_8$	177.84	0.444	183.34	2.58	1.340
$CatBoost_{10}$	178.05	0.492	183.41	2.53	1.370
$CatBoost_{15}$	178.45	0.639	183.34	2.58	1.380
$LightGBM_4$	480.28	0.260	183.19	2.33	1.230
$LightGBM_6$	479.38	0.264	183.20	2.33	1.170
$LightGBM_8$	463.21	0.267	183.19	2.31	1.260
$LightGBM_{10}$	462.37	0.272	183.11	2.34	1.160
$LightGBM_{15}$	459.44	0.279	183.17	2.31	1.130
LR_4	2.04	0.943	189.01	8.20	1.060
LR_6	2.18	0.954	188.33	8.08	1.100
LR_8	2.30	0.959	189.05	8.17	1.050
LR_{10}	2.43	0.960	189.03	8.17	1.030
LR_{15}	2.70	0.988	188.97	8.20	1.030
XGB_4	442.18	0.843	186.00	5.55	1.100
XGB_6	424.74	0.945	185.80	5.02	1.100
XGB_8	471.37	1.027	186.42	5.55	1.200
XGB_{10}	463.26	1.112	186.28	5.52	1.130
XGB_{15}	480.88	1.341	186.28	5.52	1.170

Regarding the statistical comparison and hardware-aware optimization analysis of the 40 models, comprising eight classifiers and five feature-set sizes, a Friedman test confirmed highly significant differences across configurations ($\chi^2 = 99.16, p < 0.001$). Post-hoc Nemenyi comparisons identified the top-tier performance cluster within the XGBoost family, with XGB_8, XGB_{10}, and XGB_6 achieving the best mean rankings (9.27, 9.63, and 9.73, respectively), as shown in Table 6.

An integral bootstrap analysis (10,000 resamples) was conducted using XGB_6 (Rank: 9.73) as the reference method to evaluate its suitability for deployment. In inter-family comparisons, XGB_6 demonstrated a decisive advantage over the best-performing models of every other architecture, maintaining a probability of outperforming them of $P \geq 0.997$ in all cases. Specifically, compared to RF_{10}, the analysis showed a 95% CI of [0.012, 0.098] with $P = 0.997$, while the comparison against LR_{15} yielded a 95% CI of [0.014, 0.050] with $P = 1.000$.

In intra-family ablation studies, XGB_6 demonstrated statistical parity with its larger counterparts. The comparison against XGB_{10} yielded a 95% CI of $[-0.008, 0.007]$ ($P = 0.383$), while the comparison against the full-feature XGB_{15} resulted in a 95% CI of $[-0.005, 0.023]$ ($P = 0.808$). In both cases, the confidence intervals included zero, confirming that feature reduction to six variables did not lead to a significant performance loss compared to more complex configurations.

3.4. Results analysis

Among the evaluated classifiers, XGBoost emerged as the most robust architecture (Table 6). Notably, XGB_{10} achieved the highest overall performance (95.1%±4.7% BA), while XGB_6 maintained a nearly identical accuracy (95.0%±4.6%). This marginal difference of only 0.1% underscores a statistical saturation point at 6 features, where further inclusion of variables offers diminishing returns (Table 7).

The statistical ablation analysis identifies XGB_6 as the optimal deployment configuration, effectively representing the saturation point of the feature space. The intra-family bootstrap analysis is particularly conclusive: the confidence interval for the comparison with XGB_{10} ($[-0.008, 0.007]$) and the balanced probability ($P = 0.383$) prove that the additional four features provide no measurable predictive gain. By leveraging this parity, the proposed configuration eliminates the requirement for additional respiratory sensors or complex derivation algorithms, adhering strictly to the principle of model economy. This justifies the transition to a leaner 6-feature model, which maintains the high accuracy of the XGBoost family while reducing computational and sensing overhead.

Furthermore, the inter-family analysis highlights the robustness of the proposed framework. XGB_6 outperformed models with higher complexity or alternative architectures, such as $CatB_{15}$ and $LGBM_6$, with success probabilities of 99.7% and 100%, respectively. This suggests that the combination of gradient boosting and subject-specific fine-tuning is far more effective for stress detection than simply increasing feature dimensionality or using alternative resource-intensive architectures.

XGB_6 strikes a superior balance between predictive reliability and operational efficiency. By achieving statistical parity with the best possible XGBoost configurations and clear superiority over all other model families.

The findings of this research carry significant implications for the design and deployment of next-generation IoMT systems. From a technological perspective, the identification of a 6-feature statistical saturation point proves that high-fidelity stress monitoring does not require high-dimensional data pipelines. This enables the development of hardware architectures where sensor duty cycles can be optimized to reduce power consumption, thereby extending the battery life of wearable devices, a critical factor for long-term ambulatory monitoring. Furthermore, the achievement of sub-millisecond inference times on edge hardware such as the Raspberry Pi 5 implies that complex subject-adaptive personalization can be performed entirely on-device. This eliminates the need for cloud-based processing, significantly enhancing data privacy and enabling real-time, zero-latency biofeedback for the user.

From a clinical and methodological standpoint, the study demonstrates that a two-stage subject-adaptive framework effectively mitigates the challenge of inter-subject physiological variability. The high success probability of the XGB_6 model ($P > 0.9967$) against more complex architectures suggests that personalized small-scale models are superior to generalized large-scale models for individualized health metrics. This shifts the paradigm for digital health interventions, suggesting that future stress-management systems should prioritize localized, subject-specific refinement over massive data aggregation. Ultimately, these implications provide a viable blueprint for transitioning from laboratory-based stress detection to proactive, privacy-preserving, and energy-efficient clinical monitoring in real-world settings.

4. Conclusion

This study has presented a novel, subject-adaptive framework for stress detection, specifically designed for the constraints of IoMT environments. By integrating a two-stage personalization strategy with Gradient Boosting architectures, we addressed the critical challenge of inter-subject variability in physiological signals while maintaining high computational efficiency.

The core of our findings lies in the identification of a statistical saturation point within the feature space. Through rigorous bootstrap analysis (10,000 iterations), we demonstrated that a minimalist configuration of only 6 physiological features (XGB_6) achieves statistical parity ($P = 0.3827$, CI including zero), as shown in Table 6, with high-dimensional models. This reduction in dimensionality does not incur a performance penalty; rather, it facilitates a leaner data-sensing pipeline that is essential for long-term wearable monitoring.

Furthermore, the robustness of the XGB_6 model was validated through a comprehensive inter-family comparison, where it consistently outperformed more complex architectures, including RF, MLP, and other boosting variants, with success probabilities exceeding 99.6% (Table 6). Real-time benchmarking on a Raspberry Pi 5 platform further confirmed the practical feasibility of our approach, demonstrating that the proposed framework operates well within the thermal and energy limits required for edge deployment.

Overall, the comparative evaluation of the proposed architectures highlights clear differences in how each model balances predictive reliability against computational cost. Tree-based ensemble methods, particularly XGBoost, demonstrate superior robustness for physiological signal interpretation by exploiting complex feature interactions, which translates into leading classification performance. LR shows that a simpler linear formulation can achieve a negligible storage footprint (<3 KB), while preserving stable predictive behavior, though its inference speed on ARM-based hardware remains slightly higher than more optimized boosting engines such as LightGBM and CatBoost, despite remaining faster than the more complex XGBoost.

CatBoost and LightGBM offer an attractive compromise for real-time physiological monitoring. CatBoost efficiently processes features within a responsive framework, maintaining a compact size with latencies between 0.33 and 0.64 ms. Notably, LightGBM proved to be the most computationally efficient architecture for single-sample inference on the Raspberry Pi 5 platform, achieving the lowest latencies (0.26–0.28 ms) among all evaluated approaches (Table 7).

The benchmarking process confirms that no single model is universally optimal for all stress monitoring scenarios; instead, model selection must be guided by the specific operational constraints of the target application. XGBoost emerges as the most accurate architecture when predictive performance is the primary objective, as its ability to exploit complex feature interactions consistently outperforms alternative approaches. However, for real-time systems where energy efficiency and low latency are critical requirements, such as wearable or battery-operated devices, specific trade-offs must be considered. While LR offers the highest energy efficiency with the lowest current draw (1.030 A), CatBoost and LightGBM represent more suitable solutions for time-sensitive applications, achieving superior temporal efficiency with sub-millisecond inference times as low as 0.26 ms (Table 7).

In summary, these results validate the proposed lightweight framework as a practical approach for stress detection in IoMT environments, demonstrating that effective deployment can be achieved by systematically balancing accuracy against computational constraints.

A key outcome of this research is the demonstrated feasibility of achieving high-quality stress classification using a substantially reduced set of physiological variables. The results show that models operating on as few as six to ten carefully selected features can maintain robust predictive performance without a proportional loss in accuracy. This dimensionality reduction is particularly relevant for practical IoMT deployment, as it lowers computational demands, reduces energy consumption, and enables stress monitoring with minimal sensing hardware. Furthermore, identifying a compact and stable subset of physiological predictors enhances model interpretability, which is essential for clinical acceptance and trust in real-world healthcare applications.

Careful selection of both the number and type of input variables is crucial to optimize binary stress classification, especially in scenarios with sensor constraints or the need for clinical interpretability. These findings suggest that simplification is possible without significant performance loss up to a certain point, benefiting both interpretability and computational efficiency.

These findings have significant implications for the design of next-generation wearable health monitors. By demonstrating that high-accuracy stress detection can be maintained with a minimal 6-feature subset, this study paves the way for less intrusive, more energy-efficient devices that do not rely on bulky respiratory sensors. This reduction in sensing requirements directly translates to lower hardware costs and extended battery life, facilitating long-term ambulatory monitoring. Nonetheless, while the results provide a robust framework for optimized stress detection, certain considerations should be noted to contextualize the findings. Expanding this research to larger and more heterogeneous groups would further consolidate the universality of this 6-feature subset across different physiological profiles. Additionally, while the Raspberry Pi 5 demonstrated excellent efficiency, future stages of this research could explore even more resource-constrained hardware, such as ultra-low-power microcontrollers, to test the absolute lower bounds of the model's operational footprint.

Future work could focus not only on classifying the presence or absence of stress, but also on identifying subcategories or severity levels within the stress class. Such stratification may help uncover correlations with clinical biomarkers, such as cortisol levels, thereby enhancing both physiological interpretability and clinical applicability. Moreover, as the proposed approach relies on short-term physiological measurements, these systems could be implemented for real-time stress monitoring in wearable or ambulatory settings, investigating the impact of long term factors, such as circadian rhythms or daily baseline shifts, on model stability.

CRediT authorship contribution statement

Carlos Montoya-Peña: Writing – original draft, Validation, Methodology, Investigation, Data curation. **Juan Francisco Gaitán-Guerrero:** Writing – original draft, Validation, Methodology, Investigation. **Macarena Espinilla:** Supervision, Resources, Project administration, Funding acquisition. **José L. López:** Writing – original draft, Supervision, Project administration, Investigation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Macarena Espinilla Estevez reports financial support was provided by Spain Ministry of Science and Innovation. Macarena Espinilla Estevez reports financial support was provided by Consejería de Universidad, Investigación e Innovación. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by grant PID2021-127275OB-I00 funded by MICIU/AEI/10.13039/501100011033, by "ERDF A way of making Europe", grant PDC2023-145863-I00 funded by MICIU/AEI/10.13039/501100011033, and by "European Union NextGenerationEU/PRTR", grant PID2024-156412OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU" and grant M.2 PDC_000756 funded by Consejerıa de Universidad, Investigación e Innovación and by ERDF Andalusia Program 2021-2027. Funding for open access charge: Universidad de Jaén/CBUA.

References

[1] International Society for Traumatic Stress Studies, Climate change and trauma: Understanding the psychological impact, 2024, https://istss.org/wp-content/uploads/2024/09/ISTSS-Briefing-Paper_Climate-Change-Final.pdf. (Accessed 15 January 2026).

[2] World Health Organization, World mental health report: Transforming mental health for all, 2022, https://www.who.int/publications/i/item/9789240049338. (Accessed 15 January 2026).

[3] R.A. Ross, S.L. Foster, D.F. Ionescu, The role of chronic stress in anxious depression, Chronic Stress. (Thousand Oaks) 1 (2017) http://dx.doi.org/10.1177/2470547016689472, 2470547016689472.

[4] T. Mathobela, C. Stein, C. Vincent-Lambert, et al., The effect of assessor visibility on student stress and anxiety in emergency care simulation assessments, BMC Med. Educ. 24 (1) (2024) 1043, http://dx.doi.org/10.1186/s12909-024-06020-x.

[5] F. Sangtarash, H. Choobsaz, M. Zarrin, S. Salari, E. Mokari Manshadi, A.A. Esmaeili, S.H. Mozaffari, B. Hatef, The relationship between the depression and anxiety stress survey questionnaire, salivary cortisol and heart rate variability, J. Mod. Rehabil. 19 (1) (2024) 80–89, http://dx.doi.org/10.18502/jmr.v19i1.17512.

[6] E. Rudics, A. Buzás, A. Pálfi, Z. Szabó, Á. Nagy, E.A. Hompoth, J. Dombi, V. Bilicki, I. Szendi, A. Dér, Quantifying stress and relaxation: A new measure of heart rate variability as a reliable biomarker, Biomedicines 13 (1) (2025) http://dx.doi.org/10.3390/biomedicines13010081.

[7] L.M. Vela, H. Crandall, T. Lim, F. Zhang, A. Gibbs, A.R.J. Mitchell, A. Condon, L.M. Diamond, H. Zhang, B. Sanchez, IoMT-enabled stress monitoring in a virtual reality environment and at home, IEEE Internet Things J. 10 (12) (2023) 10649–10661, http://dx.doi.org/10.1109/JIOT.2023.3240099.

[8] J. Healey, R. Picard, Detecting stress during real-world driving tasks using physiological sensors, IEEE Trans. Intell. Transp. Syst. 6 (2) (2005) 156–166, http://dx.doi.org/10.1109/TITS.2005.848368.

[9] G.L. Wagener, A. Melzer, Self-reported and physiological stress indicators and the moderating role of the Dark Tetrad in violent and non-violent gaming, Physiol. Behav. 288 (2025) 114724, http://dx.doi.org/10.1016/j.physbeh.2024.114724.

[10] A.J. Roth, A.B. Kornblith, L. Batel-Copel, E. Peabody, H.I. Scher, J.C. Holland, Rapid screening for psychologic distress in men with prostate carcinoma, Cancer 82 (10) (1998) 1904–1908, http://dx.doi.org/10.1002/(SICI)1097-0142(19980515)82:10<1904::AID-CNCR13>3.0.CO;2-X.

[11] K. Sipos, M. Sipos, The development and validation of the hungarian form of the STAI, 2, 1978, pp. 51–61, Cross-Cultural Anxiety.

[12] A.J. Roth, A.B. Kornblith, L. Batel-Copel, E. Peabody, H.I. Scher, J.C. Holland, Rapid screening for psychologic distress in men with prostate carcinoma: a pilot study, Cancer: Interdiscip. Int. J. Am. Cancer Soc. 82 (10) (1998) 1904–1908.

[13] T. Khan, S.M. Parida, S. Swain, A. Mishra, G. Dawal, S.N. Mohanty, M.I. Khan, Revolutionizing mental health counseling with serenity: An emotion-detecting chatbot, J. Comput. Biophys. Chem. (2024) 1–13, http://dx.doi.org/10.1142/S2737416524410011.

[14] V. LeBlanc, G. Mastoras, C. Hicks, P. MacGregor, C. O'Rielly, A. Petrosoniak, W. Tavares, The stressed heart: Validity evidence supporting mobile heart rate variability applications to detect psychological stress in healthcare learners, Med. Educ. 59 (7) (2025) 729–738, http://dx.doi.org/10.1111/medu.15629.

[15] A.I. Siam, S.A. Gamel, F.M. Talaat, Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques , Neural Comput. Appl. 35 (17) (2023) 12891–12904, http://dx.doi.org/10.1007/s00521-023-08428-w.

[16] S. Pourmohammadi, A. Maleki, Stress detection using ECG and EMG signals: A comprehensive study, Comput. Methods Programs Biomed. 193 (2020) 105482, http://dx.doi.org/10.1016/j.cmpb.2020.105482.

[17] H. Hijry, S. Meesam Raza Naqvi, K. Javed, O.H. Albalawi, R. Olawoyin, C. Varnier, N. Zerhouni, Real time worker stress prediction in a smart factory assembly line, IEEE Access 12 (2024) 116238–116249, http://dx.doi.org/10.1109/ACCESS.2024.3446875.

[18] T. Amira, I. Dan, B. Atta, G. Said, B. Az-eddine, W.-w. Katarzyna, Stress level classification using heart rate variability, Adv. Sci. Technol. Eng. Syst. J. 4 (3) (2019) 38–46, http://dx.doi.org/10.25046/aj040306.

[19] A. Chatterjee, et al., Stress management with HRV following AI, semantic ontology, genetic algorithm and tree explainer, Sci. Rep. 15 (1) (2025) 5755, http://dx.doi.org/10.1038/s41598-025-87510-w.

[20] S. Gedam, S. Paul, A review on mental stress detection using wearable sensors and machine learning techniques, IEEE Access 9 (2021) 84045–84066, http://dx.doi.org/10.1109/ACCESS.2021.3085502.

[21] G. Vos, K. Trinh, Z. Sarnyai, M. Rahimi Azghadi, Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review, Int. J. Med. Inform. 173 (2023) 105026, http://dx.doi.org/10.1016/j.ijmedinf.2023.105026.

[22] M. Quadrini, A. Capuccio, D. Falcone, S. Daberdaku, A. Blanda, L. Bellanova, G. Gerard, Stress detection with encoding physiological signals and convolutional neural network, Mach. Learn. 113 (8) (2024) 5655–5683, http://dx.doi.org/10.1007/s10994-023-06509-4.

[23] M. Khandelwal, A. Sharma, Machine learning and deep learning techniques to detect mental stress using various physiological signals: A critical insight, WIREs Data Min. Knowl. Discov. 15 (3) (2025) http://dx.doi.org/10.1002/widm.70035.

[24] D. Ghose, A. Chatterjee, I.A.M. Balapuwaduge, Y. Lin, S.P. Dash, Investigating lightweight and interpretable machine learning models for efficient and explainable stress detection, Front. Digit. Health 7 (2025) http://dx.doi.org/10.3389/fdgth.2025.1523381.

[25] R.B. Ramteke, G.O. Gajbhiye, V.R. Thool, Acute mental stress level detection: ECG-scalogram based attentive convolutional network, Frankl. Open 10 (2025) 100233, http://dx.doi.org/10.1016/j.fraope.2025.100233.

[26] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K. Van Laerhoven, Introducing WESAD, a multimodal dataset for wearable stress and affect detection, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 400–408, http://dx.doi.org/10.1145/3242969.3242985.

[27] S. Koldijk, M. Sappelli, S. Verberne, M.A. Neerincx, W. Kraaij, The SWELL knowledge work dataset for stress and user modeling research, in: Proceedings of the 16th International Conference on Multimodal Interaction, ICMI '14, ACM, 2014.

[28] K.R. Hole, D. Anand, S.N. Mohanty, et al., TLBEMSE: Design of a transfer learning-based bio-inspired ensemble model for preemptive detection of stress and emotional disorders, Neural Comput. Appl. 37 (2025) 20591–20616, http://dx.doi.org/10.1007/s00521-025-11160-2.

[29] H. Yu, A. Sano, Passive sensor data based future mood, health, and stress prediction: User adaptation using deep learning, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2020, pp. 5884–5887, http://dx.doi.org/10.1109/embc44109.2020.9176242.

[30] M. Kyrou, I. Kompatsiaris, P.C. Petrantonakis, Deep learning approaches for stress detection: A survey, IEEE Trans. Affect. Comput. 16 (2) (2025) 499–517, http://dx.doi.org/10.1109/taffc.2024.3455371.

[31] S. Yang, Y. Gao, Y. Zhu, L. Zhang, Q. Xie, X. Lu, F. Wang, Z. Zhang, A deep learning approach to stress recognition through multimodal physiological signal image transformation, Sci. Rep. 15 (1) (2025) http://dx.doi.org/10.1038/s41598-025-01228-3.

[32] A. Chakraborty, J.S. Banerjee, R. Bhadra, A. Dutta, S. Ganguly, D. Das, S. Kundu, M. Mahmud, G. Saha, A framework of intelligent mental health monitoring in smart cities and societies, IETE J. Res. 70 (2) (2024) 1328–1341, http://dx.doi.org/10.1080/03772063.2023.2171918.

[33] S. Panneer Selvam, M. Subramani, F.K. Jiavana, A.S. Ramachandran, Machine learning-based human stress detection model employing physiological sensory data, Arab. J. Sci. Eng. 50 (21) (2025) 17419–17439, http://dx.doi.org/10.1007/s13369-024-09927-1.

[34] K. Nkurikiyeyezu, K. Shoji, A. Yokokubo, G. Lopez, Thermal comfort and stress recognition in office environment, in: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies, SCITEPRESS - Science and Technology Publications, 2019.

[35] C. Ding, T. Yao, C. Wu, J. Ni, Deep learning for personalized electrocardiogram diagnosis: A review, 2024, http://dx.doi.org/10.48550/ARXIV.2409.07975.

[36] A. Ahmed Al Dossary, M. Chollet, A. Vinciarelli, From speech and PPG to EDA: Stress detection based on cross-modal fine-tuning of foundation models, in: Proceedings of the 27th International Conference on Multimodal Interaction, ICMI '25, ACM, 2025, pp. 87–95, http://dx.doi.org/10.1145/3716553.3750753.

[37] S. Wang, S. Zhang, W.-L. Chen, D. Truong, T.-P. Jung, From theory to application: Fine-tuning large EEG model with real-world stress data, 2025, http://dx.doi.org/10.48550/ARXIV.2505.23042.