Artificial Intelligence In Medicine 173 (2026) 103336

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Seamless monitoring of stress levels leveraging a foundational model for time sequences

Davide Gabrielli [a], Bardh Prenkaj [a,b,1], Paola Velardi [a,*,1]

[a] Sapienza University of Rome, Via Salaria 113, Rome, 00198, Italy
[b] Technical University of Munich, Richard-Wagner-Str. 1, Munich, 80333, Germany

ARTICLE INFO

Keywords:
Anomaly detection
Seamless monitoring
Foundation models
Transformers
Neurodegenerative diseases
Stress detection
Explainability

ABSTRACT

Background: Accurate and continuous monitoring of physiological stress is crucial, especially for patients with neurodegenerative diseases. Traditional monitoring methods, such as Electrocardiogram (ECG), are often invasive and limited in duration, while data from lightweight wearable devices, though more practical for seamless monitoring, typically suffers from significant quality degradation compared to clinical-grade measurements.
Motivation: The challenge lies in developing a robust, long-term, and patient-friendly stress monitoring system that overcomes the limitations of conventional approaches and the accuracy compromises of current wearables. Such a system must also provide actionable, interpretable insights for clinicians and adapt to individual patient variability.
Method: This manuscript introduces a methodology for seamless stress level monitoring by leveraging UniTS, a foundational model for time series. Our approach redefines stress detection as an anomaly detection problem, establishing a personalized baseline for each patient's physiological behavior. Furthermore, to enhance clinical utility and trust, the system integrates a Large Language Model (LLM) to generate human-readable explanations for detected anomalies.
Results: The proposed UniTS-based methodology demonstrates superior performance, outperforming 12 top-performing methods on three benchmark datasets. Crucially, it achieves performance comparable to that obtained from more invasive, clinical-grade devices (like ECG) even when utilizing data from lightweight wearable devices, thereby enabling truly seamless monitoring. Furthermore, the system has been successfully tested in a real-world environment, in the context of a project to monitor elderly patients with cognitive disorders in their homes.
Novelty: This work presents an advancement in physiological stress monitoring by offering a personalized, explainable, and continuously adaptive system. We extend and fine-tune UniTS to support contextual anomaly detection and LLM-driven explainability, addressing critical gaps in current healthcare monitoring, fostering enhanced clinician control, improved system predictability, and facilitating long-term, real-world applicability for patients with neurodegenerative conditions.

1. Introduction

High stress levels can worsen the symptoms of neurodegenerative diseases such as Alzheimer's and Parkinson's [1]. In these patients, stress is generated by mood changes such as depression and apathy, confinement at home, and other factors such as social isolation, uncertainty about the evolution of the disease, and the economic, health, personal, and family situation [2,3]. Analyzing stress levels can provide insight into disease progression and help healthcare providers adjust treatment plans accordingly. Managing stress can also improve patients' quality of life and help to inform new therapeutic strategies.

AI-based methods for stress detection are motivated by the need for early, timely, and personalized stress management solutions. These methods leverage cutting-edge machine learning algorithms to provide objective, real-time monitoring and analysis, significantly contributing to mental health and well-being (see, among others, [4]). Among the numerous techniques presented in the literature, those based on data

* Corresponding author.
E-mail address: velardi@di.uniroma1.it (P. Velardi).
[1] Equal supervision contribution.

https://doi.org/10.1016/j.artmed.2025.103336
Received 15 November 2024; Received in revised form 4 December 2025; Accepted 13 December 2025
Available online 24 December 2025

extracted from sensor devices are particularly promising due to several advantages. One of the key benefits of using sensor-based methods is their ability to provide real-time data, which is crucial for timely interventions and stress management. Secondly, they protect privacy more than solutions based on behavioral signals such as speech, gestures, and facial expressions [5], a particularly relevant aspect for elderly and frail people (see [6], among others). Third, they may support continuous monitoring, enabling more convenient home care solutions for patients and caregivers. In the literature, various sensor types are used to detect stress by leveraging different physiological signals. As surveyed in [7], commonly used devices include ECG and EDA sensors, EEG, skin temperature, respiratory and pressure sensors, as well as activity sensors such as accelerometers, gyroscopes, and RFID-based technologies. However, many of these devices – particularly clinical-grade solutions such as ECGs – are not well-suited for long-term or real-world monitoring due to their invasiveness, the need for professional assistance to wear or maintain them, and high operational costs. Even consumer wearables such as smartwatches, beacons, or fitness trackers, while more practical, still face limitations due to their generally lower precision and reliability compared to clinical-grade equipment, which can compromise the accuracy of physiological measurements in health monitoring scenarios [8].

To address these challenges, we define *seamless monitoring* as the real-time collection of physiological signals that: (i) operates without time constraints, meaning it can function indefinitely without interruptions caused by battery depletion, storage limitations, or fixed monitoring schedules, thereby enabling extended physiological surveillance[2]; (ii) ensures uninterrupted data collection that integrates naturally into the patient's daily life, even during routine activities, with data flowing continuously and reliably without user intervention; and (iii) does not require explicit actions by patients or healthcare professionals beyond occasional basic maintenance, emphasizing its autonomous and passive nature, where data is automatically collected without manual setup, calibration, or ongoing management.

In essence, *seamless monitoring* goes beyond simple continuous data capture by ensuring effortless integration into everyday life, thereby maximizing patient adherence and real-world applicability.

The main objective of the manuscript is to enable the seamless monitoring of stress levels in patients, particularly those with neurodegenerative diseases, by introducing and evaluating a novel methodology that uses a foundational time series model (UniTS) [9]. This primary objective is supported by several key sub-objectives:

- **Advancing Stress Detection with UniTS:** Develop a methodology for stress detection that builds on the UniTS model, demonstrating its superior performance compared to 12 top-performing methods in three benchmark data sets. The model is further tested in the real-world setting provided by the @HOME project.
- **Enabling Seamless Monitoring:** To achieve seamless monitoring by showing that the proposed model, when utilizing data from lightweight wearable devices, delivers performance comparable to that obtained from more invasive clinical-grade devices such as ECG.
- **Justifying Anomaly Detection:** To present stress detection as an anomaly detection problem rather than a classification task, arguing its benefits for model adaptation to individual patients, improved clinician control, and improved predictability and trust of the system.
- **Enhancing Explainability:** To equip the system with explainability features, specifically by integrating a Large Language Model (LLM) to generate human-readable explanations of detected anomalies, thus aligning with medical professionals' needs and diagnostic workflows.
- **Ensuring Reproducibility:** To make the code developed publicly available,[3] fostering reproducibility and further research in the field.

2. Related work

Our work relates to AI-based emotion monitoring from sensor data and anomaly detection from time series.

2.1. Emotion monitoring and stress detection

Table 1 summarizes the key details regarding research on emotion recognition through physiological signals, providing a qualitative comparison with the solution proposed in this paper.

Signals Used lists the physiological signals employed. Examples are electrocardiogram (ECG), heart rate (HR), galvanic skin response (GSR), respiratory tract (resp.) and electromyogram (EMG), among others. The choice of signals, as well as the related device, is crucial as it directly affects the accuracy and reliability of the detection system. Different signals provide varied information about an individual's physiological state, impacting the performance of the employed models. When Heart Rate Variability (HRV) is derived from less invasive devices like wristbands and smartwatches, the quality of measurements can significantly degrade compared to those obtained from Electrocardiogram (ECG):

1. *Accuracy and limitations of wearable devices vs. ECG:* Although some wearable devices can show a high correlation with ECG-derived HRV at rest, their accuracy frequently degrades during physical activity due to factors such as motion artifacts. Studies have shown that HRV measurements acquired from portable devices can differ from those obtained from ECG, although the absolute error may be small, it is often heterogeneous [24]. PPG-based wearables, common in smartwatches and wristbands, are more susceptible to noise, motion artifacts, and variations in peripheral circulation compared to direct electrical recordings by ECG. This can lead to significant differences in HRV parameters, especially during activities other than rest [25–27].
2. *Challenges in Signal Acquisition and Processing:* The accurate determination of inter-beat intervals (e.g., RR intervals from ECG or pulse-to-pulse intervals from PPG) is paramount for reliable HRV analysis. Wearable devices face challenges in the proper detection and filtering of signals, particularly in the presence of various environmental factors and patient movements, which can lead to inaccuracies [28,29].
3. *Performance under Ideal Conditions:* It is important to note that under highly controlled and ideal resting conditions, some studies have shown a high degree of correlation between PPG and ECG for HRV measures. However, these findings often do not extend to real-world, dynamic environments where patient movement and other factors compromise signal quality [30].

Application specifies whether the study focuses on stress (S) or emotion (E) detection, whereas the second considers nuanced mental conditions.

Device Type specifies the type of device used to collect the physiological signals, such as wearables, smartwatches, wristbands, GSR devices, or pulse oximetry. The device type influences the monitoring system's accuracy, as we previously discussed, but also practicality, comfort, and user acceptance. Only some devices allow for seamless monitoring in real-world settings, which is especially valuable for frail people.

Model Type lists the machine learning or statistical models used for classification tasks. Notice how, except [10,11], all works rely on

[2] Minimal and infrequent actions such as charging a smartwatch are acceptable, as long as they do not disrupt data collection for extended periods.

[3] https://github.com/davegabe/Wearable-Stress-Monitor

Table 1
Organization of the literature on ML methods for mood monitoring from physiological data.

	Signals used	Application	Device type	Model type	Task	Year
[10]	ECG	S	ECG, Wearable	Conv. Gaussian Mixture VAE	AD	2023
[11]	Temp., Humidity (Sweat), Steps	S	Wristband	Stacked Ens. with GB	C	2023
[12]	HR, GSR, Acc	S	Smartwatch	MLP	C	2019
[13]	Resp., GSR, HR, EMG	S	*	SVM, kNN	C	2015
[14]	BVP, HR, ST, GSR, Resp.	S	Smartwatch	RF	C	2017
[15]	Steps, Acc	S	Wristband	SVM, kNN	C	2013
[16]	GSR (+ Speech)	S	GSR Device	SVM	C	2013
[17]	EEG, ECG	S	Wearable	SVM	C	2019
[18]	HR	E	Wristband	SVM, kNN, RF	C	2021
[19]	HRV (from ECG)	E	ECG, Wearable	SVM	C	2016
[20]	SpO2, Pulse Rate	E	Pulse Oximetry	SVM	C	2018
[21]	ECM, ECG, GSR, Resp.	E	*	pLDA	C	2008
[22]	HR, BP	E	Smartwatch	Fuzzy Petri Net [23]	C	2021
[us]	HRV, HR	S + E	Wristband, ECG	Transformer	AD	2024

simple off-the-shelf machine learning models (SVM, kNN, and pLDA[4]), which might not be suitable to detect emotions or stress levels that are not in the training data distribution, and therefore have limited capacity for generalization but also for personalization. In [11], an ensemble mechanism over Gradient Boosts (GB) is applied to classify stress levels by leveraging the expertise of multiple models. In this case, a linear regression meta-model is applied to the output of the GB base models to predict the outcome. In [10], a clustering algorithm based on a Gaussian mixture convolutional variational autoencoder is employed to identify normal emotional patterns and detect anomalies. To the best of our knowledge, we are the first to use a state-of-the-art foundational model for time-series to tackle stress and abnormal emotion detection.

Task indicates whether the model's task is classification (C) or anomaly detection (AD). Most works tackle stress and emotion monitoring as classification problems, categorizing states into predefined classes. The only exception is UAED [10], which performs an anomaly detection task. We argue that approaching stress monitoring as an anomaly detection problem is more principled and inherently more explainable than doing fine-grained emotion classification. The key reasons for this superiority include:

1. *Personalized Adaptation:* Anomaly detection establishes a unique baseline of "normal" physiological behavior for each individual patient. This personalized approach facilitates better model adaptation to individual variations, which is crucial given that physiological responses to stress can differ significantly between people. Classification, conversely, typically learns boundaries between predefined classes, which may not adequately capture individual nuances [31,32].
2. *Enhanced Clinician Control and Trust:* Anomaly detection systems align well with clinicians' diagnostic workflows. Doctors are often more interested in identifying and investigating deviations from a patient's normal physiological state than merely confirming a predefined stress category. By focusing on identifying "abnormalities", the system provides actionable insights that clinicians can visualize and investigate, enhancing their understanding and trust in the system's predictions [33,34].
3. *Intrinsic Explainability:* Anomaly detection inherently offers better explainability. By identifying deviations from a learned normal baseline, the system can provide reconstruction errors for each signal dimension, indicating which specific physiological features contributed most to a detected anomaly. This level of detail is often more challenging to achieve with classification models, which provide a categorical output without explicit reasons for the classification [35,36].
4. *Precision Medicine and Continuous Adaptation:* Anomaly detection models offer significant benefits for precision medicine by identifying unique patient profiles and continuously adapting to new data. This allows for dynamic monitoring and timely intervention based on individual physiological changes, making the system more robust and relevant for long-term health management [37,38].

[4] Extended Linear Discriminant Analysis.

2.2. Anomaly detection in time series

Following UAED [10], we cast stress monitoring as an anomaly detection problem. As previously remarked, this allows us to define a personalized baseline of "normal" physiological behavior for each patient and flag deviations as anomalies, rather than relying on fixed categorical labels.

Anomaly detection in time series has been explored using a variety of paradigms, which we summarize below.

Distance-based methods detect anomalies based on their deviation from the k-nearest neighbors in feature space. These methods typically assume that anomalies lie far from dense regions of normal data.

Density-based approaches, such as LOF and Isolation Forest, estimate the local data density and identify low-density regions as anomalous [39,40]. These methods can be sensitive to the curse of dimensionality in high-dimensional time series.

Prediction-based models detect anomalies by comparing predicted and actual values; large prediction errors are treated as anomalies. Benkabou et al. [41] proposed a Poisson-model-based predictor to capture temporal dependencies. Similarly, Shokoohi-Yekta et al. [42] introduced a discord discovery approach based on local recurrence rates in multivariate time series, which allows identifying unusual subsequences that do not recur frequently. Although effective in generic applications, these methods lack personalization and interpretability needed for clinical use.

Reconstruction-based methods train autoencoders or similar architectures to learn normal patterns. At inference time, a high reconstruction error implies a deviation from the learned baseline. Recent works such as HypAD [43] use hyperbolic embeddings to enhance the representation of rare events. Another example is the space-embedding strategy proposed by Zhang et al. [44], which projects multivariate time series into a spatial manifold to amplify differences between normal and abnormal patterns. These approaches, however, are domain-agnostic and do not address the interpretability or patient-level variability critical in medical applications.

GAN-based approaches leverage adversarial learning to generate realistic reconstructions and amplify anomaly signals. MAD-GAN [45] combines the generator and discriminator outputs for anomaly scoring. TadGAN [46] introduces cycle-consistency to improve robustness. TranAD [47] integrates transformer encoders with adversarial training and self-conditioning to capture complex temporal dependencies. USAD [48] further strengthens detection by enforcing asymmetric reconstruction loss using adversarially trained autoencoders.

Gaussian Mixture Model (GMM)-based techniques combine dimensionality reduction with probabilistic modeling. DAGMM [49] jointly optimizes an autoencoder and a GMM for unsupervised detection. UAED [10] extends this by applying a 2D convolutional variational autoencoder and modeling the latent space with GMMs, specifically adapting to physiological time series such as ECG.

Graph-based approaches such as MTAD-GAT [50] and GDN [51] model inter-sensor and temporal dependencies using attention-based message passing. These methods excel in capturing structured interactions but can be computationally expensive and opaque in interpretation.

Other deep learning models include OmniAnomaly [40], which uses stochastic latent variables in a variational framework. CAE-M [52] incorporates memory networks and MMD regularization to filter noise. MSCRED [53] models multi-resolution temporal patterns via convolutional signature matrices, while LSTM-NDT [54] applies dynamic thresholds over recurrently forecasted signals.

In this paper, we leverage **UniTS** [9], the first *foundation model* specifically designed for multivariate time series. Unlike prior work, UniTS supports multiple downstream tasks (including anomaly detection) and captures long-range temporal and inter-sensor dependencies via cross-dimensional attention. Crucially, by fine-tuning UniTS to our application domain, we redefine stress monitoring as an anomaly detection problem and integrate *LLM-based explanations* to offer interpretability tailored for clinical practitioners — capabilities not addressed by earlier approaches.

3. Method

UniTS [9] is the first foundation model[5] for multivariate time series, designed to perform in a unified way various tasks such as forecasting, imputation, classification, and anomaly detection. It employs self-attention to capture relationships across both the sequence and variable dimensions. This is particularly beneficial in our multi-sensor domain since, if one variable has an anomaly, the model can assess how this anomaly correlates with the patterns observed in other variables, thereby improving detection accuracy. Furthermore, UniTS is designed to cohesively integrate signals of different types and lengths. To achieve this, it employs multi-scale processing techniques to handle time-series data of different lengths, which involves processing the data at various temporal resolutions.

3.1. Problem definition and preliminaries

Anomaly detection aims to identify observations in a multivariate time series that significantly deviate from learned patterns of normal behavior [43,48,55]. Formally, consider a series $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ where each observation $\mathbf{x}_t \in \mathbb{R}^n$. A model f, trained on historical data $\mathbf{X}_{train} = \{\mathbf{x}_1, \dots, \mathbf{x}_t\}, t < T$, learns normal patterns in the data. For each time point t, we compute an anomaly score:

$$S_t = g(\mathbf{x}_t, f(\mathbf{X}_{train})),$$

and mark $\mathbf{x}_t$ as anomalous if $S_t > \tau$, producing the anomaly set $\mathbf{A} = \{\mathbf{x}_t \mid S_t > \tau, t = 1, \dots, T\}$. Following [48], dependencies over time are modeled using sliding windows:

$$W_t = \{\mathbf{x}_{t-K+1}, \dots, \mathbf{x}_t\},$$

transforming the series $\mathbf{X}$ into a sequence $\mathbf{W} = \{W_1, \dots, W_T\}$. Given a new window $\widehat{W}_t$, $t > T$, the task is to assign a label $y_t \in \{0, 1\}$ indicating whether the window contains an anomaly ($y_t = 1$) or not ($y_t = 0$). For simplicity, we denote training windows by W and unseen windows by $\widehat{W}$.

[5] Foundation models are pre-trained on broad data at scale and can be adapted to a wide range of downstream tasks.

3.2. Framework overview

UniTS is a versatile multi-task model built on a unified network architecture. In this study, we focus on its application to anomaly detection using multivariate time series data, specifically heart rate (HR) and heart rate variability (HRV) signals (see Fig. 1). At a high level, UniTS takes raw time series data as input, processes it through a sequence of transformer-based blocks, and reconstructs the signals to identify unexpected deviations, which are marked as anomalies. We divide this section into two parts: i.e., (1) the input preprocessing/tokenization (Section 3.2.1), and (2) the modified transformer layers to process the input tokens (Section 3.2.2).

3.2.1. Input tokenization

In UniTS, raw time series data is first converted into structured units called tokens, representing either time steps or variables. This unified token format allows the model to capture dependencies across both time and variables without task-specific preprocessing, enabling broad applicability to forecasting, classification, imputation, and anomaly detection. UniTS defines three token types: sample, prompt, and task tokens,[6] each serving a specific role in time series analysis and supporting multi-task learning.

Sample tokens. The time series[7] $\mathbf{X} \in \mathbb{R}^{T\times n}$. We split $\mathbf{X}$ into patches along the time dimension via a non-overlapping patch size of k. Then, a linear layer projects each patch into an embedding vector of length d, obtaining sample tokens $z_{\mathbf{X}} \in \mathbb{R}^{\frac{T}{k}\times n\times d}$. $z_{\mathbf{X}}$ are added with learnable positional embeddings.

Prompt tokens. They are defined as learnable embeddings $z_p \in \mathbb{R}^{p\times n\times d}$ where p is the number of tokens. These tokens incorporate the task UniTS needs to perform. Notice that, in Fig. 1, we do not illustrate the prompt tokens proposed in the original paper since we only exploit the *forecasting prompt* in our anomaly detection scenario. Nevertheless, for completeness, we invite the reader to visualize that these prompt tokens are prepended to the input series tokens as in Eq. (1).

$$z_{\text{Anomaly}} = [z_p, z_{\mathbf{X}}]_{\mathcal{T}} \in \mathbb{R}^{(p+\frac{T}{k})\times n\times d}, \tag{1}$$

where $[*, *]_{\mathcal{T}}$ is the concatenation along the time axis.

Differently from the original proposal of [9], we do not use the task tokens. Here, we want to compare the reconstructed tokens with those coming from the ground truth (GT) and pinpoint the anomalies along the series. In detail, instead of injecting task tokens into the input, we want to reconstruct the input tokens.

3.2.2. Modified transformer

The model takes the input z_{Anomaly} and processes it through N stacked blocks of a modified transformer. This design allows us to handle complex, multi-domain data that may vary in both dynamics and the number of variables.

As shown in Fig. 1, each block includes three main components: Time Multi-Head Self-Attention (MHSA) to capture temporal dependencies, Variable MHSA to model relationships across variables, and a Dynamic Feed-Forward Network (FFN) for adaptive feature transformation. Lastly, Gating modules control the flow of information between these components. In the following, we describe each component in detail.

[6] Task tokens are not used in our scenario; see [9] for details.

[7] For simplicity, we assume that the entire time series has already been split into multiple windows. Therefore, at inference, one reconstructs a single window at a time.

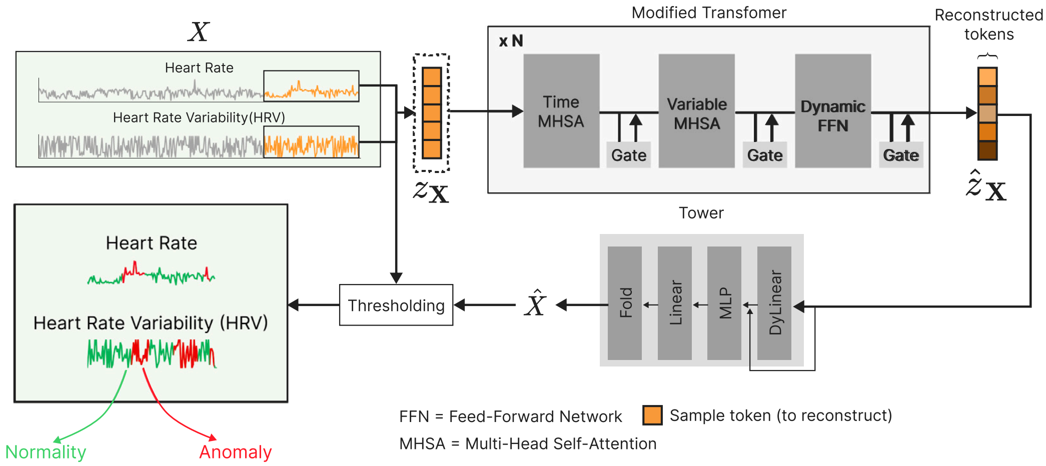

Fig. 1. Fine-tuning of the UniTS architecture for anomaly detection. The multivariate time series in input gets transformed into several tokens, which are then passed through N different blocks of the UniTS model. UniTS reconstructs the sample tokens, which are unpatched and compared against the real tokens. The comparison uses a dynamic threshold to discern between anomalies and normality. For visualization purposes, we do not illustrate the prompt tokens concatenated to the sample ones.

Time and variable MHSA. We apply a two-way self-attention mechanism that operates across both the feature and time dimensions of the data. Unlike previous methods, which typically focus on either the time dimension or the variable dimension separately, our approach considers both simultaneously. This design allows the model to handle time series samples that may vary in the number of features and the length of the time window, making it more flexible for real-world data.

Dynamic FFN. The transformer block is adapted by adding a dynamic operator to the feed-forward network (FFN) layer. This enhancement allows the FFN to model relationships between different tokens, whereas a standard FFN processes each embedding vector independently without capturing such dependencies.

Gating. To reduce interference in the latent representation space, we use gating modules after each layer. These modules dynamically adjust the scale of features in the latent space, helping to maintain stable and reliable representations throughout the network.

Tower. This module converts the processed tokens into time-point predictions, reconstructing the full time-series signal. Given the generated sample tokens $\hat{z}_{\mathbf{X}}$, the tower module rebuilds the original time-series sample as shown in Eq. (2):

$$\hat{\mathbf{X}} = \mathrm{Proj}\big(\mathrm{MLP}(\hat{z}_{\mathbf{X}} + \mathrm{DyLinear}(\hat{z}_{\mathbf{X}}))\big), \tag{2}$$

such that, the MLP is a two-layer FFN with interleaved activation functions, Proj is the unpatchify operation that converts the embeddings back into a continuous sequence, and DyLinear is a dynamic linear transformation defined as

$$\mathrm{DyLinear}(z_t, w) = \mathrm{Interp}(w) z_t, \tag{3}$$

where Interp uses bilinear interpolation to resize the weight matrix w from shape $w_{\text{in}} \times w_{\text{out}}$ to $l_{\text{in}} \times l_{\text{out}}$, enabling flexible reconstruction for different input and output lengths.

3.3. Fine-tuning UniTS for abnormal emotion detection

In UniTS, whole model *fine-tuning* and *prompt-learning* represent two distinct paradigms for adapting the foundational model to new datasets and specific tasks, each with unique advantages and operational mechanisms. Here, we want to use HR and HRV signals for abnormal emotion detection via non-invasive wearable devices. Hence, the choice between fine-tuning and prompt learning is influenced by practical considerations. By preprocessing the datasets to obtain features at a low-frequency rate (e.g., transforming ECG samples from 700 Hz to HR/HRV at 0.1 Hz), we significantly reduce the computational burden of fine-tuning UniTS. This reduction aligns with our goal to emulate real-world scenarios where vital signals are measured with smartwatches, balancing effective battery usage with reliable signal acquisition [56]. Consequently, fine-tuning becomes a more feasible and "affordable" strategy in our application, enabling the model to adapt effectively without compromising efficiency. We follow [56] to reduce the sampling rate — this is also supported in Fig. 5 of Section 4.4. Finally, we use 20 epochs and an exponential learning rate decay to fine-tune the model. In Appendix E, we provide details on the discussion between the advantages of prompt learning and fine-tuning.

3.4. Adding explainability features

Once the UniTS model is fine-tuned for anomaly detection, we leverage the output of the Tower module during inference (as illustrated in Fig. 1) as input to a Large Language Model (LLM), such as GPT-4o [57] or LLaMa-2 [58], to generate explanations for the detected anomalies. The anomaly score produced by UniTS represents the likelihood of an anomaly and is computed as the ratio between the detected anomaly energy and a predefined threshold. Scores below 1 indicate normal samples, whereas scores above 1 denote anomalies. These scores are then used to prompt the LLM, enabling it to generate interpretable explanations tailored for domain experts.

The LLM can be prompted to contextualize its explanations according to the intended audience, ensuring the generated output matches the user's level of expertise. As noted by [59], persona-based prompting allows the LLM to adjust the level of technical detail based on the user's role, such as a clinician or a system administrator. In the example prompt, the LLM is informed that it is "interacting" with a professional clinician, which guides it to generate explanations using clinical terminology and domain-specific knowledge. This capability is particularly critical in clinical settings, where precision and appropriate medical language are necessary to convey the physiological implications of detected anomalies.

Moreover, LLM-generated explanations can support multi-level decision-making by presenting both the expected physiological norms and the detected anomalies. This dual approach ensures that stakeholders are not only alerted to deviations but also informed of baseline expectations, fostering a deeper understanding of abnormal versus normal readings. The example shown in Fig. 2 illustrates this by prompting the LLM to explain why anomalies occurred while also describing the expected physiological behavior when anomaly scores

Prompt:
You are an automated system that interprets changes in HR (Heart Rate) and HRV (Heart Rate Variability) signals when emotional anomalies are detected in a patient's vital signs with respect to a resting state condition.
The data provided includes:

- **Time**: Seconds since the start of the recording
- **HR**: Heart Rate in beats per minute
- **HRV**: Heart Rate Variability (RMSSD) in milliseconds
- **Prediction**: Indicates if an anomaly was detected (1) or not (0)
- **Anomaly Energy**: Ratio of the detected anomaly energy to the threshold, indicating the intensity of the anomaly. If energy is < 1, the sample is normal; otherwise, it indicates an anomaly.

The patient is a young adult with no known medical conditions. The data was collected using a wearable ECG sensor.
Patient Data:

Time	*HR*	*HRV*	*Prediction*	*Anomaly Energy*
0	82.39	20.9	0	0.0
⋮	⋮	⋮	⋮	⋮
80	82.56	10.8	1	1.3

Provide a concise explanation structured in four sections: Analysis, Causes, Criticality, and Scores.

- Analysis: Highlight any strong, high-intensity anomalies if present, using a dotted list to describe the observed changes in HR and HRV signals. Avoid directly mentioning energy values; instead, describe the anomalies' clinical relevance and intensity. Describe any trends in HR and HRV over time, noting persistent patterns or shifts in the data that may indicate physiological or emotional responses.
- Causes: Propose likely causes of these anomalies, considering the patient's demographic, the characteristics of the detected anomalies (e.g., short, sudden, prolonged), and any environmental or measurement factors that could affect signal accuracy. Distinguish between potential true physiological changes, system errors, or data noise.
- Criticality: Assess the clinical significance of these anomalies. Consider both the impact of individual anomalies (especially high-energy ones) and the cumulative effect if anomalies are frequent or prolonged. Specify if the anomaly could require immediate follow-up, if it's non-critical but worth monitoring, or if it seems likely to be insignificant or an error.
- Scores:
  - **Significance Score** (0-10): Rate the significance of the anomaly based on how clinically relevant it is, with 0 being not relevant (e.g., noise) and 10 being highly relevant (e.g., real physiological change).
  - **Criticality Score** (0-10): Rate the potential clinical impact or urgency, with 0 being minimal concern and 10 being critical. This should consider both the intensity of the anomaly and the risk it may pose to the patient's well-being.

Ensure the explanation is brief, clinically relevant, and suitable for interpretation by a professional clinician.

Fig. 2. Prompt used to interpret HR and HRV signals, focusing on emotional anomalies, and generate a structured clinical analysis for patient data interpretation.

exceed a specified threshold. In this setup, the system interprets HR and HRV signals and produces explanations that combine reasoning for abnormal readings with descriptions of "normal" physiological patterns, all expressed in appropriate language for the clinical domain.

4. Experiments and results

4.1. Datasets

We use the following three widely adopted datasets for emotion detection:

- *DREAMER* [60] is designed for emotion recognition via low-cost devices. It consists of EEG and ECG readings recorded from 23 participants. ECG signals were captured at a sampling rate of 256 Hz using a SHIMMER wireless sensor. Movie clips were used to stimulate different emotions in participants, and at the end of each clip, participants scored their self-assessment on a 5-point arousal, valence, and dominance scale. Here, we treat states with a valence of 1 as abnormal and the rest as normal.
- *MAHNOB-HCI* [61] was collected using the Biosemi Active II system with active electrodes to capture physiological signals, including ECG and EEG, from 27 participants at a sampling rate of 256 Hz. Participants were asked to record their emotional status after each trial, consisting of multimedia content stimulation. Here, we treat the states with a valence of 1 as abnormal and the rest of the emotional states as normal. For our experiments, we selected subsets of the dataset with recordings of at least 120 s and excluded those with incomplete tracks due to recording errors.
- *WESAD* [62] has ECG and BVP signals from 15 participants at a rate of 700 Hz extracted via a RespiBAN Professional sensor worn on the chest and the Empatica E4 wristband. It contains four different emotional states. Here, we treat the stress state as abnormal and the rest as normal. The stress condition is obtained by exposing the subjects to the Tier Social Stress Test, which consists of public speaking and a mental arithmetic task.
- *@HOME* is a dataset collected as part of an ongoing pilot study aimed at detecting physiological anomalies, including stress, in real-world home environments. It is entirely acquired through consumer-grade wearables and ambient sensors, enabling continuous, unobtrusive monitoring of elderly patients with neurodegenerative disorders. The dataset includes multivariate physiological signals – heart rate (HR), heart rate variability (HRV), respiratory rate, blood pressure, blood oxygen saturation, sleep quality, and step count – sampled every minute. Emotional and

physiological anomalies are identified using UniTS, adapted for contextual anomaly detection.

Since the experiment is ongoing, we report here results from a three-month pilot involving two elderly patients with early-stage cognitive impairments. While limited in scale, this study provides rare real-world data on long-term, in-the-wild physiological monitoring — an area where public datasets are extremely scarce. Existing real-world datasets typically include very few subjects [63] or cover only short periods [64], making this pilot a valuable contribution to closing this gap.

Further details on all datasets are provided in Appendix A, and processing techniques are described in Appendix D.

4.2. Experimental setup

Compared methods. Notice that since the works in emotion monitoring do not model stress detection as an anomaly detection problem (see Table 1), we compare only with the works surveyed in Section 2.2. We compare fine-tuned UniTS with various SoTA models for multivariate time-series anomaly detection. The models compared include TranAD [47], LSTM-NDT [54], CAE-M [52], DAGMM [49], USAD [48], OmniAnomaly [40], MTAD-GAT [50], GDN [51], MAD-GAN [45], MSCRED [53], HypAD [43], and TadGAN [46]. We adopted the hyperparameters presented in their original papers for these models.

For the @HOME dataset, we adapted UniTS to perform contextual anomaly detection. This approach incorporates auxiliary physiological and environmental signals as context when predicting anomalies, reducing false positives caused by predictable physiological changes such as physical activity.[8] The detailed methodology and implementation of this framework are provided in Appendix F. Unlike controlled datasets such as DREAMER, MAHNOB-HCI, and WESAD – which rely on artificial emotion induction – the @HOME dataset captures anomalies as they naturally occur, offering a realistic benchmark for field applications. Given that exhaustive labeling is infeasible in real-world deployments, evaluation is conducted separately through retrospective clinical review of UniTS-detected anomalies and their explanations, allowing the quantification of true and false positives.

Training UniTS. For fine-tuning the UniTS model, we utilized a window size of 5, an embedding dimension of 128, and a learning rate of 5×10^{-4}. The datasets were split into 80% training and 20% testing sets. All models were trained for 20 epochs, and the F1 score was calculated as the average across a 5-fold cross-validation. We employed the Adam optimizer with weight decay set to 5×10^{-6}. The batch size was set to 64, and gradient accumulation with 32 steps was used to simulate a larger effective batch size of 2048, which helped stabilize training. The base learning rate was scaled linearly according to this effective batch size following the formula $\mathrm{LR}_{\text{effective}} = \mathrm{LR}_{\text{base}} \times \frac{\text{effective batch size}}{32}$.

Fair comparisons and threshold settings. We systematically preprocessed our datasets to ensure consistency and comparability across different sources. From all datasets, we extracted HR and HRV every 10 s using a sliding window approach based on the preceding 60 s of signal data from ECG and, when available, Blood Volume Pulse (BVP), which measures the volumetric change in blood circulation through the microvascular bed of tissue. BVP signals can be extracted non-invasively using photoplethysmography (PPG), a technique commonly integrated into smartwatches and other wearable devices. This capability is instrumental in stress detection and management, making wearable devices valuable in clinical and everyday settings. Lastly, we use the 3rd and 97th percentiles of the reconstruction error distribution as thresholds for anomaly labeling — i.e., values below the 3rd or above the 97th percentile are labeled as anomalous, and all others as normal. This non-parametric strategy adapts to skewed, subject-specific error distributions and provides a robust way to isolate extreme deviations without relying on parametric assumptions. Empirically, this percentile-based method yielded stable performance across datasets while minimizing false positives from minor physiological variability.

[8] Note that contextual information is not available nor necessary in controlled experiments such as for DREAMER, MAHNOB-HCI and WESAD.

4.3. Quantitative results

UniTS surpasses SoTA anomaly detection systems and shows remarkable stability and robustness across datasets. As shown in Table 2, UniTS outperforms all the compared systems, achieving a considerable average improvement of 22% on the second-best — i.e., 16.29% on DREAMER, 27.43% on MAHNOB-HCI, 4.07% on WESAD (ECG), and 3.53% on WESAD (BVP). We first conduct a Friedman Test on the top methods (UniTS, DAGMM, GDN, USAD, TadGAN, MAD-GAN) to show that UniTS has, statistically and significantly, the best performance across the board ($F = 20.911$, $p = .0008 < .05$). Then, we perform a Dunn post-hoc test with Holm correction where the control anomaly detector is UniTS. The test suggests that UniTS is statistically and significantly different (better) across the board on average (see Table 3). We also note that the competitive advantage of UniTS is lower in the WESAD dataset. Several systems achieve more or less similar – and in some cases slightly better – performances. We argue that WESAD is the only dataset that explicitly includes stress among the classified emotions. In contrast, the other two datasets use dimensional models of emotions (arousal and valence), resulting in a nuanced and less "clear-cut" classification. In this simpler context, many systems perform much better. BVP can be noisy and contain many artifacts, which hardens extracting meaningful features from it [65]. Hence, in WESAD (BVP), the normal and abnormal data distribution is noised and does not respect that originally coming from the ECG – i.e., compare HR and HRV distributions in Fig. 4 and notice the difference in scale in HRV coming from ECG and BVP – which we argue leads SoTA methods to underperform. Furthermore, UniTS, through its incorporated attention mechanism over the variates and the time axes, mitigates this phenomenon by giving less weight to noisy samples, maintaining high WESAD (BVP) performances.[9] Overall, UniTS exhibits remarkably stable performance and higher robustness across all datasets and systems (std of only .049), resulting from being a universal model pre-trained on many diverse signals and tasks.

For completeness purposes, in Fig. 3, we provide the reader with confidence intervals on the performances of UniTS and five top-performing SoTA methods according to the Dunn post-hoc test analysis — see Table 3. A more extensive evaluation, including additional performance metrics, model trade-offs, and a breakdown of computational complexity, is provided in Appendix G and Appendix B.

UniTs achieves performances from lightweight devices comparable to SoTA systems using more invasive and potentially accurate sensors. The only dataset collecting signals from ECG devices and wristbands (BVP) is WESAD. As shown in Table 2, UniTS delivers consistently high F1 scores with both ECG and BVP and maintains similar results across both types, while some other systems experience varied performance with BVP signals. This is a valuable result since, as discussed in Section 1, lightweight devices, such as wristbands and smartwatches, are less expensive and invasive, enabling seamless monitoring. Using non-invasive, wearable technology without sacrificing performance can expand stress-related research and applications. Future studies can integrate these devices in various settings to gather continuous stress data, leading to personalized stress management programs, early intervention systems for mental health, and better well-being monitoring.

[9] We verified that UniTS has an FPR of 0.031 in both versions of WESAD. Meanwhile, it reports an FNR of 0.172 in WESAD (ECG) and 0.159 in WESAD (BVP), suggesting that it is more capable of detecting the true anomalies (i.e., it has better recall).

Table 2
Comparison of different HR and HRV methods extracted from ECG and BVP. Bold-faced values illustrate the best; underlined ones are the second best.

	DREAMER	MAHNOB-HCI	WESAD (ECG)	WESAD (BVP)	AVG F1
LSTM-NDT [54]	0.235	0.373	0.694	0.760	0.515 ± 0.218
MSCRED [53]	0.434	0.618	0.594	0.536	0.545 ± 0.071
HypAD [43]	0.513	0.547	0.669	0.479	0.552 ± 0.072
TranAD [47]	0.515	0.622	0.474	0.717	0.582 ± 0.095
OmniAnomaly [40]	0.543	0.513	0.716	0.583	0.589 ± 0.078
MTAD-GAT [50]	0.369	0.677	0.757	0.676	0.620 ± 0.149
CAE-M [52]	0.634	0.513	<u>0.762</u>	0.590	0.625 ± 0.090
MAD-GAN [45]	0.390	<u>0.678</u>	0.735	0.750	0.638 ± 0.146
TadGAN [46]	0.528	0.657	0.705	0.680	0.643 ± 0.068
USAD [48]	0.590	0.641	0.686	0.759	0.669 ± 0.062
GDN [51]	<u>0.712</u>	0.501	0.697	<u>0.778</u>	0.672 ± 0.103
DAGMM [49]	0.619	0.596	0.744	0.733	<u>0.673 ± 0.066</u>
UniTS [9]	**0.828**	**0.864**	**0.793**	**0.800**	**0.821 ± 0.049**

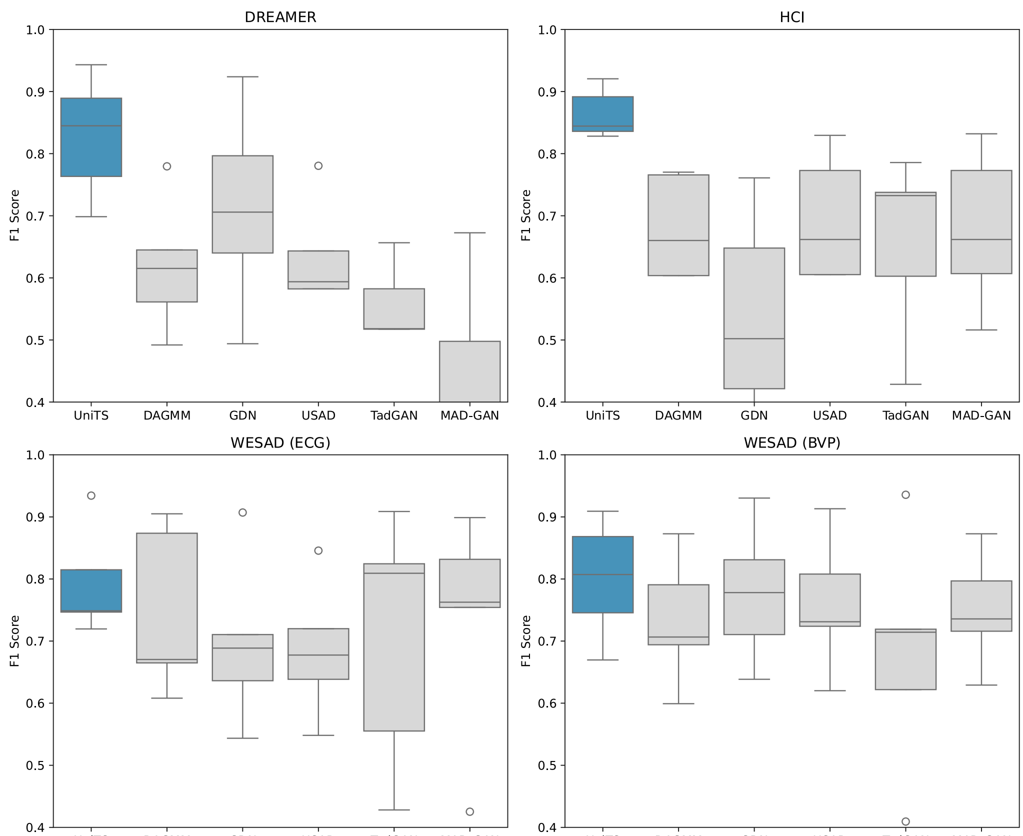

Fig. 3. Confidence intervals for the performances of UniTS and the best SoTA methods as per the Dunn post-hoc test. We highlight the method that performs best on average according to the values reported in Table 2.

Table 3
P-values produced by the Dunn post-hoc test with the Holm correction for the F1 metric, where the control model is UniTS against the top-performing SoTA methods. We use an α of 0.05 to reject the null hypothesis. The corrected α_{corr} is calculated as $\alpha_{org} \big/ (n - \text{rank} + 1)$, where the rank is the order of p-values from the lowest to greatest (e.g., TadGAN has rank 1, GDN has rank 5).

	p-value vs. UniTS	Corrected α_{corr}	Reject $\mathcal{H}_0$
TadGAN	2.903×10^{-3}	10.00×10^{-3}	✓
MAD-GAN	1.081×10^{-2}	1.250×10^{-2}	✓
USAD	1.277×10^{-2}	1.670×10^{-2}	✓
DAGMM	1.560×10^{-2}	2.500×10^{-2}	✓
GDN	2.198×10^{-2}	5.000×10^{-2}	✓

Lower sampling rates on lightweight devices make performance plummet by ∼54%. Fig. 5 shows the importance of the sampling rate on all datasets on ECG and BVP versions. Here, we illustrate the trend of average F1 scores of UniTS when the sampling interval increases from 10 to 60 s. The vital signals (i.e., HR and HRV) are sampled every k seconds to build the multivariate time series taken in input, with varying k. As expected, the longer the sampling interval, the less responsive the model becomes, resulting in a degradation of results. This phenomenon is even more visible when we extract the HR and HRV from BVP, having a drop of −53.89% passing from a 10 to a 60-second sampling rate. Importantly, these results align with the constraints of many consumer-grade wearables, which often compute HR/HRV at roughly one sample per minute, a setting also used in the @HOME deployment. While this lower resolution reduces accuracy,

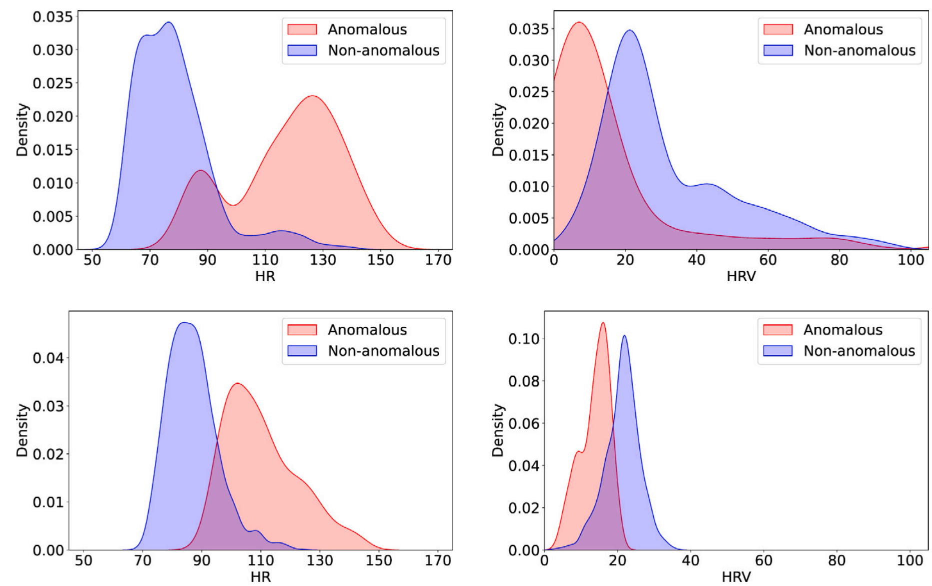

Fig. 4. Density plots for HR and HRV on the test set for WESAD for ECG (up) and BVP (down) versions. We also show the means of the distributions with dashed lines.

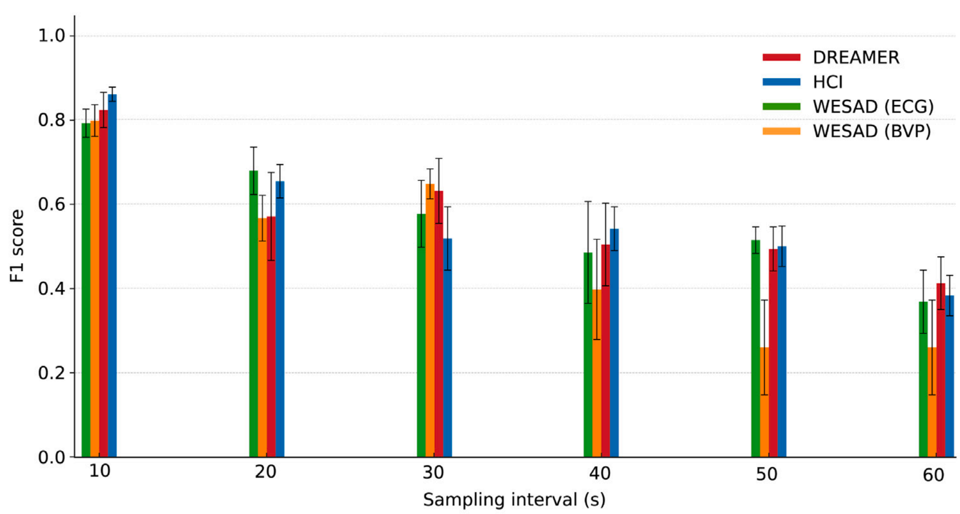

Fig. 5. Performance degradation of F1 score across different sampling intervals using UniTS on all the datasets. Notice how the overall decreasing performance trend starts to have diminishing effects, especially in WESAD (BVP).

UniTS still captures meaningful temporal patterns, demonstrating that the model remains *usable* even under strict resource constraints. At the same time, modern high-frequency wearables, such as chest-strap ECGs (125–130 Hz) or research wristbands (64 Hz), enable HR/HRV extraction on the order of 10-second windows or faster. With such richer signals, UniTS achieves substantially higher performance, benefitting from the improved temporal precision. Overall, these findings highlight a practical trade-off: resource-constrained devices necessitate low sampling rates, yielding lower but still functional performance, whereas high-precision applications benefit greatly from more frequent sampling. Thus, the choice of sampling interval should be tailored to the specific use case, balancing device limitations against the level of accuracy required.

4.4. Qualitative results

Casting the problem as one of anomaly detection enhances explainability. Anomaly detection begins by establishing a model of "normal" behavior based on the monitored signals. Next, it detects deviations from this normality. This helps clinicians understand the baseline deviations identified and align better with medical professionals' needs and practical constraints. First, clinicians are usually more interested in identifying and investigating abnormalities than confirming the norm. Anomaly detection systems align with doctors' diagnostic workflow by focusing on detecting deviations from the norm. Secondly, these models provide significant benefits for precision medicine and personalization due to their ability to identify unique patient profiles and adapt to new data continuously [66]. Furthermore, in the case of multiple monitored signals, the model should highlight which ones are more indicative of stress [67]. In detail, although anomaly detectors are not interpretable per se,[10] UniTS reconstructs the multivariate time series to provide reconstruction errors for each series' dimension. This inherent

[10] Without loss of generality, anomaly detectors produce reconstruction errors for each data point, combining all feature dimensions. These errors do not

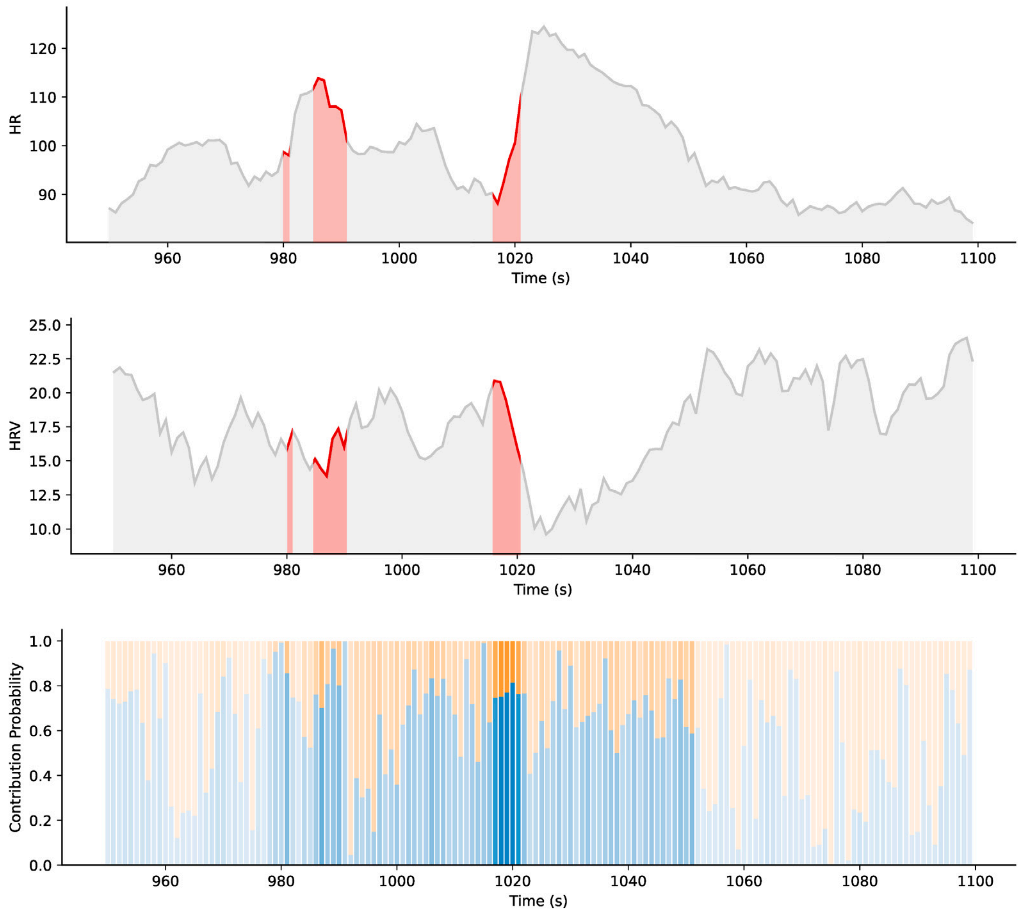

Fig. 6. HR (up) and HRV (middle) contribution in WESAD (BVP) during inference time. Red-shaded areas depict the reported anomalies. The lower bar plot illustrates the contribution probabilities based on the reconstruction error of each signal. The more transparent the contribution, the lower the chances of the signals being considered anomalous; the more opaque the contributions are, the more the chances of the signals being anomalous. Note that the transparency can be seen as visual uncertainty about the data point at a particular time step.

characteristic allows us to interpret whether anomalies occurred, which is paramount, and which features contributed to their detection.

For example, Fig. 6 shows how to visually support doctors in identifying which signals contributed the most to detecting anomalies. Here, we depict the contribution probability based on the reconstruction error of each signal – i.e., HR (up) and HRV (middle) – in determining anomalies. The more opaque the shade of each bar in the plot, the more likely an anomaly will occur. Notice how the opacity of the probabilities increases when there is an abrupt trend change in one (or both) of the signals, which aligns with the red areas depicting the detection of an anomaly. One can think of the transparency of each bar as a visual uncertainty of the model. Therefore, the more opaque the bars are, the more certain the model is that there is an anomaly in that particular time frame. Meanwhile, when the bars are transparent, they may indicate two things: either the model is uncertain, or the data points therein are normal. To this end, we integrate the contribution probability with detecting anomalies in the raw signals. Additionally, we invite the reader to notice the different transparency levels within the time frame where an anomaly is detected. This phenomenon is expected as not all anomalies (gradual vs. abrupt) are detected with the same confidence. Interestingly, badly reconstructing HR in this scenario pushes the model to be more confident about an anomaly occurring – see the middle portion of the third row in Fig. 6. Contrarily, even if UniTS cannot effectively reconstruct HRV, it does not lead to detecting anomalies – e.g., notice the more transparent tail-ends of the probabilities plot. By exploiting these types of explicit and inherently interpretable visualizations – i.e., the contribution probability is estimated via the reconstruction error – doctors can interpret which signal guides the prediction of an anomaly during inference and intervene accordingly. Finally, this type of analysis can be used to debug if the monitoring system functions correctly. For instance, if the contribution probability bars are opaque, yet the system does not report an anomaly, there might be something wrong with the current monitoring setup. In addition to a graphical representation of the anomaly (showing the dimensions that contribute most to the detected anomaly), we provide clinicians with a rich textual explanation, as detailed below.

Zero-shot integration with LLMs further enhances explainability. In real-world scenarios, anomalies detected by the model require further investigation by a medical professional to ensure their correctness and assess whether they could result in any significant health risks for the patient. The anomaly detection model offers useful insights, but these findings require more context and interpretation. Using an LLM, we can provide detailed explanations that help clarify the anomaly and its potential impact, as illustrated in Figure H.9 and Figure H.10 of Appendix H.

Here, we evaluated the utility and accuracy of explanations generated by GPT-4o using qualitative assessments from a medical professional. The explanations, derived from the outputs of the anomaly detection model, were structured into four sections: Analysis, Causes,

necessarily indicate which features/signals contribute to the anomalousness of the data point itself.

Table 4
Qualitative evaluation by a medical professional of GPT-4o's explanations, assessing the usefulness of each analysis, causes, criticality, and the coherence and specificity of the overall explanation. Ratings are on a 0–5 scale. Each dataset contains 4 different explanations. Here, we show averages for each dataset and the cumulative average ($\mu \pm \sigma$) over all explanation evaluations.

	Utility of Analysis	Utility of Causes	Utility of Criticality	Explanation Coherence	Explanation Specificity
DREAMER	5.00	4.75	4.50	4.25	1.25
MAHNOB-HCI	5.00	4.75	4.50	4.50	0.50
WESAD (ECG)	4.75	4.25	4.50	4.50	0.25
WESAD (BVP)	4.75	4.75	5.00	4.75	0.75
$\mu \pm \sigma$	4.875 ± 0.331	4.625 ± 0.484	4.625 ± 0.599	4.500 ± 0.500	0.688 ± 1.261

Criticality, and Scores, according to the scheme illustrated in Fig. 2. We generated an evaluation set of 16 explanations, which were reviewed by a highly experienced doctor,[11] who thoroughly examined the clinical data and images before considering the textual explanations provided by the system. Based on her evaluation, we assessed the explanations' relevance, coherence, and specificity, with the results summarized in Table 4.

The results indicate that GPT-4o provided highly useful explanations, with average utility scores of 4.875 for analysis, 4.625 for causes, and 4.625 for criticality. While the overall coherence of the explanations was strong, with an average score of 4.5, the low specificity score of 0.688 can be attributed to two main factors. First, as we already remarked, the datasets were collected in a controlled setting, which provided limited contextual information about the patient's activities or other physiological signals. This restricted the ability to contextualize the anomalies with a broader understanding of the patient's behavior or lifestyle. Second, the lack of detailed clinical history for the patient further limited the depth of the explanations. Without access to relevant background information, the model could only offer general insights based on the detected anomalies rather than providing a more personalized interpretation tied to the patient's specific condition or medical history. This limitation – resulting from the use of controlled datasets in the current experiment – is effectively overcome in the @HOME study, presented later in this section.

The medical professional was also asked to rate the significance of the anomaly to assess its relevance, as well as its criticality to evaluate the potential risk it poses to the patient.[12] These ratings were then compared with those provided by GPT-4o to determine whether numerical metrics could be generated to offer an immediately comprehensible score. Such a score would help clinicians rapidly identify anomalies that are of particular concern and may require urgent attention or further investigation. It should be noted that while GPT-4o generated ratings on a 0–10 scale (see Fig. 2), the medical professional used a more compact 0–5 scale. To use a finer-grained scale to reduce the effect of small rating errors, we requested GPT-4o to produce results on a 0–10 scale. In this way, we obtain step errors leading to a smaller fractional error when mapped to the 0–5 scale originally used by the doctor. Afterward, the GPT-4o ratings were rescaled by dividing by 2, aligning them to the 0–5 range for direct comparison with the doctor's ratings.

The results are presented in Table 5 – for error distributions, refer to Fig. 7 – which shows that GPT-4o generally assigned higher ratings across all metrics and datasets. However, the average error rates were 0.58 for significance and 0.43 for criticality, indicating that the model's assessments were comparable to the expert's. These findings suggest that by using an LLM, we can generate useful explanations that support and accelerate the review process of anomalies, providing valuable assistance to medical professionals in assessing and prioritizing potential issues.

[11] With more than 40 years of experience.

[12] We acknowledge that, in this experiment, clinical evaluation is limited by the absence of relevant factors such as overall health condition, which is instead available in the @HOME dataset.

UniTS proves effective in real-world continuous monitoring: insights from the @home pilot study. The @HOME dataset complements traditional benchmarks by introducing real-world complexity and variability into the evaluation of stress and physiological anomaly detection. All patients involved in the experiment were fully informed about its purpose and were continuously monitored by a geriatrician and their medical staff. The patients' conditions were recorded on a weekly basis through structured interviews and standardized questionnaires. As previously mentioned, several signals are monitored for each patient. For stress monitoring, we focus on heart rate (HR) and heart rate variability (HRV); in addition, stress signals can also be monitored through hyper/hypotension and sleep quality. For hyper/hypotension detection, we consider systolic and diastolic blood pressure, and for assessing sleep quality, we use metrics such as total sleep duration, deep sleep ratio, sleep onset, and wake-up time.

Evaluation in this setting is conducted retrospectively, focusing exclusively on anomalies detected by UniTS. Subsequently, these alerts are reviewed by a senior geriatrician with extensive clinical experience and a full contextual knowledge of the patients' conditions, allowing quantitative assessment in terms of confirmed true positives (TP) and false positives (FP), but not false negatives due to the absence of a complete annotation. The evaluation is based on the graphical visualization of the anomalies (e.g., Fig. 6), the explanations provided by the system, and the clinical examination of the patient.

During the observation period, UniTS identified 32 anomaly events, a number consistent with the rarity of clinically significant deviations in stable patients. Of these, 30 anomalies (93.75%) were confirmed as true positives by the expert, while only 2 (6.25%) were attributed to sensor artifacts. As shown in Table 5, all detected anomalies were considered clinically meaningful, with expert-assigned significance scores consistently ≥ 3.0. However, the corresponding criticality scores were more moderate, ranging from 2.0 to 2.4, indicating that no high-severity events occurred during the study period. Moreover, as detailed in Table 6, the specificity of the explanations improved substantially compared to the controlled benchmarks, confirming that the availability of contextual data enables the generation of highly specific and valid clinical evaluations. As previously mentioned, the patients were regularly monitored by the clinical staff, and no relevant events were observed that were not detected by our system. However, we acknowledge that this cannot be formally considered as evidence of the absence of false negatives.

Overall, our findings provide empirical evidence that UniTS not only maintains performance in controlled settings but also scales to uncontrolled, in-the-wild environments. This highlights its effectiveness for continuous, long-term monitoring where traditional classification methods are impractical due to the need for extensive labeled data or frequent user involvement.

5. Discussion and conclusion

Strengths of the proposed approach. This study introduces a number of key strengths that advance both the theoretical foundations and practical implementation of stress monitoring from physiological signals.

Table 5
Comparison of anomaly significance and criticality ratings between a medical professional and GPT-4o, showing the errors in GPT estimations for each dataset. The ratings are on a 0–5 scale. For controlled datasets (4 explanations each), and @HOME (one per anomaly). Here, we show averages for each dataset and the cumulative average ($\mu \pm \sigma$) over all explanation evaluations.

	Significance			Criticality		
	Human evaluation	GPT evaluation	Error	Human evaluation	GPT evaluation	Error
DREAMER	1.00	2.25	1.25	0.25	1.25	1.00
MAHNOB-HCI	1.00	2.13	1.13	0.50	1.00	0.50
WESAD (ECG)	1.50	2.25	0.75	0.50	1.50	1.00
WESAD (BVP)	1.50	2.13	0.63	0.75	1.00	0.25
@HOME Hyper/Hypotension	3.00	3.20	0.20	2.30	2.35	0.05
@HOME Stress	3.21	3.34	0.13	2.37	2.58	0.21
@HOME Sleep Quality	3.00	3.00	0.00	2.00	2.00	0.00
$\mu \pm \sigma$	–	–	0.58 ± 0.46	–	–	0.43 ± 0.39

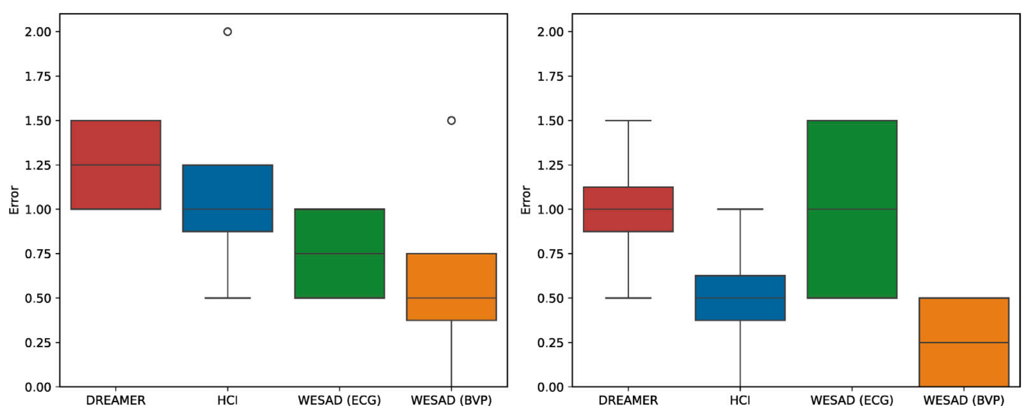

Fig. 7. Distribution of errors in GPT-4o's anomaly significance (left) and criticality ratings (right) compared to a medical professional's assessments. Ratings are on a scale from 0 to 5.

Table 6
Specificity scores for explanations generated in the @HOME pilot study. Unlike the controlled datasets, the availability of contextual data allows the LLM to achieve high specificity, validating the system's real-world applicability.

Anomaly type	Explanation specificity
@HOME Hyper/Hypotension	3.90 ± 0.54
@HOME Stress	3.50 ± 0.67
@HOME Sleep Quality	3.00 ± 0.00

- *Theoretical contributions.* We propose a novel formulation of stress detection as a personalized anomaly detection problem, which represents a significant departure from traditional classification-based approaches. By establishing individualized baselines, our method supports dynamic adaptation to patient-specific physiological patterns — an essential feature for precision medicine. Furthermore, we are the first to adapt a foundational, multi-task model for time series (UniTS) to the domain of physiological monitoring. Currently deployed in a clinical study, the system continuously monitors patients in home-care environments, providing real-time alerts to healthcare professionals. This extends the theoretical landscape of time-series learning by demonstrating how pre-trained, task-agnostic temporal representations can be effectively fine-tuned for specialized, clinically relevant anomaly detection tasks. It also contributes to the growing body of work on transferability and generalization in temporal modeling.
- *Practical contributions.* Our method delivers robust, high-accuracy performance across a wide range of benchmark datasets and sensor modalities, including noisy, low-quality signals from lightweight wearable devices. This robustness enables seamless, non-invasive monitoring — a critical requirement for real-world applications involving frail or elderly patients. Importantly, we integrate Large Language Models (LLMs) to produce structured, human-readable explanations of detected anomalies. These explanations enhance clinical interpretability and trust, addressing a common barrier to the adoption of AI tools in healthcare. Our approach also tackles deployment challenges such as low sampling rates and environmental noise, making it well-suited for continuous, real-world stress monitoring.
- *Empirical validation.* As detailed in Section 4.3, UniTS consistently outperforms 12 state-of-the-art anomaly detection models across three publicly available datasets as well as a naturalistic dataset collected from elderly patients with cognitive impairments. It achieves comparable performance using both invasive sensors (e.g., ECG) and non-invasive alternatives (e.g., wristbands), demonstrating resilience to signal noise and reduced data fidelity. Additionally, we show that UniTS maintains robust performance even when signal sampling rates are decreased, highlighting its applicability in settings constrained by power consumption or hardware limitations. Results from the @HOME pilot study, a real-world long-term monitoring using consumer-grade wearables, demonstrate UniTS's robustness, with most detected anomalies clinically confirmed — highlighting its precision and practical utility in unsupervised settings.

Together, these theoretical insights and practical innovations make our approach both scientifically novel and clinically actionable.

Benefits in clinical applications. In the literature, a combination of physiological, behavioral, and psychological monitoring methods have been proposed to monitor the stress level in patients with neurodegenerative diseases, facilitated by advanced wearable devices and digital technologies [7]. Speech and video analysis, ECG, EEG, and implantable devices have shown significant promise in stress monitoring for patients with neurodegenerative diseases. These methods have been shown to provide more accurate, continuous, and detailed insights into stress levels than less invasive approaches. However, they present challenges such as higher costs, potential discomfort, privacy concerns, and the need for manual procedures. In this paper, we demonstrated that using last-generation foundational AI models can boost the performance of less invasive devices, making them comparable to those of devices with a greater impact on the daily lives of monitored patients. Monitoring the progression of patients' disease using lightweight devices has numerous advantages in clinical practice, including increased patient comfort and compliance, continuous and real-time monitoring, reduced healthcare costs, increased patient independence, personalized care, improved remote patient management, and improved health outcomes through early intervention.

Furthermore, we have shown how casting the stress detection problem as one of anomaly detection, rather than classification, improves the explainability of the model's outcomes (Section 4.4). Explainability is vital in medical applications of AI because it ensures trust and accountability, supports clinical validation, and facilitates integration into clinical workflows. In addition, anomaly detection better aligns with the clinical practice since each patient has unique physiological and behavioral baselines. Anomaly detection allows for creating personalized baselines, making identifying deviations that indicate stress easier.

Limitations and future work. We acknowledge that this study also has limitations. For example, the study compares HR and HRV extracted from a wristband and ECG device, partly due to the wider availability of benchmark datasets for these types of signals. In the future, we plan to extend our experiments to a wider variety of physiological signals that are both detectable by lightweight devices and relevant to studying mental states, such as physical activity and sleep quality. Additionally, we note that the textual explanations of the anomalies, although evaluated as high-quality w.r.t. almost all dimensions considered, have a low specificity because they do not consider the patient's past stress episodes. Currently, the lack of naturalistic (in-the-wild) datasets does not allow the detection of a mental stress state to be conditional on the patient's past history. Finally, we plan to go beyond stress monitoring to include the observation of other signals of physical and cognitive decline, such as daily activity routines and other actigraphy data, as well as contextual factors. Note that UniTS (see Section 3) *by design* allows for the integration of time series of different types and lengths. Similar to the limitations mentioned above, the obstacle to this type of experiment is rather the absence of datasets that collect, for each monitored patient, multiple signals from multiple devices in naturalistic settings.

To overcome all these limitations, we are currently collecting, in the context of the Regional project @HOME, a variety of physiological, behavioral, and environmental signals for a broader cohort of ten patients monitored in their homes. The preliminary analysis presented in this paper reflects the initial pilot phase and will be extended in future work as data collection continues across the broader patient cohort. Due to the limited scale of the pilot – two patients over a three-month period – the results, while encouraging, are not sufficient to conclusively demonstrate the effectiveness of our approach. A more comprehensive assessment will only be possible upon completion of the full study, when a larger and more diverse dataset becomes available.

We are confident that the continuation of our ongoing experiments will demonstrate more clearly the advantages of the approach presented in this paper and also constitute a valuable benchmark for other researchers.

Finally, we acknowledge that this paper does not provide a comprehensive analysis of privacy and security issues, which are particularly critical in the context of medicine and, especially, telemedicine. Regulations in this area vary across countries; however, we note that the @HOME project was conducted in Europe in full compliance with the GDPR and the draft AI Act. Patients were required to explicitly consent to the monitoring protocol, data could only be shared in anonymized form and with the approval of the oversight committee, and all anomaly reports automatically detected by the AI modules had to be explicitly validated by a physician. Furthermore, the rich explanations produced by our proposed system are intended to support clinicians in this process. A deeper examination of these privacy and security aspects, although beyond the scope of the present study, remains an important direction for future work.

CRediT authorship contribution statement

Davide Gabrielli: Validation, Software, Methodology, Formal analysis, Data curation. **Bardh Prenkaj:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Funding acquisition. **Paola Velardi:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors 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 was supported by a grant from Lazio Region, FESR Lazio 2021–2027, project @HOME (# F89J23001050007 CUP B83C2300 6240002). We sincerely thank Prof. Dr. Emilia Reda for her valuable contribution to evaluating the quality of the diagnoses provided by the system described in this article.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2025.103336.

References

[1] Knezevic Emilija, Nenic Katarina, Milanovic Vladislav, Knezevic Nebojsa Nick. The role of cortisol in chronic stress, neurodegenerative diseases, and psychological disorders. Cells 2023;12(23).

[2] Kraut Robert, Li Han, Zhu Haiyi. Mental health during the covid-19 pandemic: Impacts of disease, social isolation, and financial stressors. PLoS One 2022;17(11). e0277562.

[3] Heilman Kenneth M, Nadeau Stephen E. Emotional and neuropsychiatric disorders associated with alzheimer's disease. Neurotherapeutics 2023;19(1):99–116.

[4] Zhang Jing, Yin Hang, Zhang Jiayu, Yang Gang, Qin Jing, He Ling. Real-time mental stress detection using multimodality expressions with a deep learning framework. Front Neurosci 2022;16.

[5] Lian Hailun, Lu Cheng, Li Sunan, Zhao Yan, Tang Chuangao, Zong Yuan. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. Entropy 2023;25(10).

[6] Niemeijer Alistair R, Frederiks Brenda JM, Riphagen Ingrid I, Legemaate Johan, Eefsting Jan A, Hertogh Cees MPM. Ethical and practical concerns of surveillance technologies in residential care for people with dementia or intellectual disabilities: an overview of the literature. Int Psychogeriatr 2010;22(7):1129–42.

[7] Taskasaplidis Georgios, Fotiadis Dimitris, Bamidis Panagiotis. Review of stress detection methods using wearable sensors. IEEE Access 2024;01. 1–1.

[8] Bent Brian, Goldstein Brett A, Kibbe Warren A, Dunn John P. Digital health: Opportunities and challenges to develop precision medicine. J Am Med Assoc (JAMA) 2019;321(1):23–4.

[9] Gao Shanghua, Koker Teddy, Queen Owen, Hartvigsen Thomas, Tsiligkaridis Theodoros, Zitnik Marinka. Units: Building a unified time series model. 2024, arXiv preprint arXiv:2403.00131.

[10] Zhu Jiaqi, Deng Fang, Zhao Jiachen, Liu Daoming, Chen Jie. Uaed: Unsupervised abnormal emotion detection network based on wearable mobile device. IEEE Trans Netw Sci Eng 2023;10(6):3682–96.

[11] Al-Atawi Abdullah A, Alyahyan Saleh, Alatawi Mohammed Naif, Sadad Tariq, Manzoor Tareq, Azam Muhammad Farooq-i, Khan Zeashan Hameed. Stress monitoring using machine learning iot and wearable sensors. Sensors 2023;23(21):8875.

[12] Can Yekta Said, Chalabianloo Niaz, Ekiz Deniz, Ersoy Cem. Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. Sensors 2019;19(8):1849.

[13] Ghaderi Adnan, Frounchi Javad, Farnam Alireza. Machine learning-based signal processing using physiological signals for stress detection. In: 2015 22nd Iranian conference on biomedical engineering. ICBME, 2015, p. 93–8.

[14] Gjoreski Martin, Luštrek Mitja, Gams Matjaž, Gjoreski Hristijan. Monitoring stress with a wrist device using context. J Biomed Informatics 2017;73:159–70.

[15] Sano Akane, Picard Rosalind W. Stress recognition using wearable sensors and mobile phones. In: 2013 humaine association conference on affective computing and intelligent interaction. 2013, p. 671–6.

[16] Kurniawan Hindra, Maslov Alexandr V, Pechenizkiy Mykola. Stress detection from speech and galvanic skin response signals. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems. 2013, p. 209–14.

[17] Ahn Joong Woo, Ku Yunseo, Kim Hee Chan. A novel wearable eeg and ecg recording system for stress assessment. Sensors 2019;19(9):1991.

[18] Bulagang Aaron Frederick, Mountstephens James, Teo Jason. Multiclass emotion prediction using heart rate and virtual reality stimuli. J Big Data 2021;8:1–12.

[19] Guo Han-Wen, Huang Yu-Shun, Lin Chien-Hung, Chien Jen-Chien, Haraikawa Koichi, Shieh Jiann-Shing. Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine. In: 2016 IEEE 16th international conference on bioinformatics and bioengineering. BIBE, IEEE; 2016, p. 274–7.

[20] Hakim Lutfi, Wibawa Adhi Dharma, Pane Evi Septiana, Purnomo Mauridhi Hery. Emotion recognition in elderly based on spo 2 and pulse rate signals using support vector machine. In: 2018 IEEE/aCIS 17th international conference on computer and information science. ICIS, IEEE; 2018, p. 474–9.

[21] Kim Jonghwa, André Elisabeth. Emotion recognition based on physiological changes in music listening. IEEE Trans Pattern Anal Mach Intell 2008;30(12):2067–83.

[22] Lin Qiang, Li Tongtong, Shakeel P Mohamed, Samuel RDinesh Jackson. Advanced artificial intelligence in heart rate and blood pressure monitoring for stress management. J Ambient Intell Humaniz Comput 2021;12:3329–40.

[23] Cardoso Janette, Valette Robert, Dubois Didier. Fuzzy petri nets: an overview. IFAC Proc Vol 1996;29(1):4866–71.

[24] Dobbs WC, Fedewa MV, MacDonald HV, Holmes CJ, Cicone ZS, Plews DJ, Esco MR. The accuracy of acquiring heart rate variability from portable devices: A systematic review and meta-analysis. Sports Med 2019;49(3):417–35.

[25] Reali P, Tacchino G, Rocco G, Cerutti S, Bianchi AM. Heart rate variability from wearables: A comparative analysis among standard ecg, a smart shirt and a wristband. Stud Health Technol Inform 2019;261:128–33.

[26] Cernian AI, Ciobanu AB, Ciobanu V, Costache C. Comparison of hrv parameters from ecg and ppg signals. Sensors 2022;22(19):7481.

[27] Lazazzara M, Sciaraffa N, Vainich O, Conforti F. Validity of ultra-short-term hrv analysis using ppg—a preliminary study. Sensors 2022;22(20):7995.

[28] Georgiou K, Lakkas K, Papanikolaou G, Dousis E. Can wearable devices accurately measure heart rate variability? a systematic review. Folia Med 2018;60(1):8–16.

[29] Bashar Sk M, Han D. Wearable devices for remote monitoring of heart rate and heart rate variability—what we know and what is coming. Sensors 2022;22(22):8903.

[30] Charlton PH, Birrenkott T, Smith L, Marten M, Tarassenko L, Young D. A comparison of photoplethysmography and ecg recording to analyse heart rate variability in healthy subjects. J Med Eng Technol 2009;33(8):638–46.

[31] Veeravalli B, Rao BASP, Rao PN. Real-time, personalized anomaly detection in streaming data for wearable healthcare devices. In: 2018 IEEE 4th international conference on smart computing (smartComp). IEEE; 2018, p. 403–8.

[32] Valerio A, Demarchi D, O'Flynn B, Tedesco S. Development of a personalized anomaly detection model to detect motion artifacts over ppg data using catch22 features. IEEE Sensors J 2024.

[33] Alshammari N, Moharram S, Alsaiari A, Alghamdi S, Alshammari A, Alharbi S. Designing interpretable ml system to enhance trust in healthcare: A systematic review to proposed responsible clinician-ai-collaboration. 2023, arXiv preprint arXiv:2311.11055.

[34] Rizvi SA, Khan M, Khan MF, Singh PK. Hybrid deep learning for iot-based health monitoring with physiological event extraction. Healthcare 2025;13(12):1419.

[35] Li Y, Wu Y, Gui L, Jin Y. A survey on explainable anomaly detection. 2022, arXiv preprint arXiv:2210.06959.

[36] Zhang Y, Zhu S, Liu S, Ma Y, Wang Q, Zhang X, Sun X, Zhao Y, Li T, Zhu T. An explainable anomaly detection framework for monitoring depression and anxiety using consumer wearable devices. 2025, arXiv preprint arXiv:2505.03039.

[37] Khan AM, Alam N, Zeeshan K, Rahman MA, Al-Matin M. Anomaly detection framework for wearables data: A perspective review on data concepts, data analysis algorithms and prospects. IEEE Access 2022;10:125195–211.

[38] Pavithra D, Deepa T, Kiruthiga R, Sudhakar S, Muthumeenakshi P. Application of lstm networks for continuous patient monitoring and anomaly detection in wearable health devices. Indian J Sci Technol 2024;17(37):3909–21.

[39] Wang Zhiwei, Chen Zhengzhang, Ni Jingchao, Liu Hui, Chen Haifeng, Tang Jiliang. Multi-scale one-class recurrent neural networks for discrete event sequence anomaly detection. 27th ACM SIGKDD Conf Knowl Discov Data Min 2021;3726–34.

[40] Su Ya, Zhao Youjian, Niu Chenhao, Liu Rong, Sun Wei, Pei Dan. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019, p. 2828–37.

[41] Benkabou Seif-Eddine, Benabdeslem Khalid, Kraus Vivien, Bourhis Kilian, Canitia Bruno. Local anomaly detection for multivariate time series by temporal dependency based on Poisson model. IEEE Trans Neural Networks Learn Syst 2021.

[42] Shokoohi-Yekta Hassan, Chen Yanping, Keogh Eamonn. A novel computational approach for discord search with local recurrence rates in multivariate time series. Soft Comput 2019;23:12201–15.

[43] Flaborea Alessandro, Prenkaj Bardh, Munjal Bharti, Sterpa Marco Aurelio, Aragona Dario, Podo Luca, Galasso Fabio. Are we certain it's anomalous? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 2896–906.

[44] Zhang Wei, Wang Xue, Bai Shuo, Wang Liang. A space-embedding strategy for anomaly detection in multivariate time series. Neurocomputing 2022;489:92–103.

[45] Li Dan, Chen Dacheng, Shi Lei, Jin Baihong, Goh Jonathan, Ng See-Kiong. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In: ICANN. 2019.

[46] Geiger Alexander, Liu Dongyu, Alnegheimish Sarah, Cuesta-Infante Alfredo, Veeramachaneni Kalyan. Tadgan: Time series anomaly detection using generative adversarial networks. In: 2020 IEEE international conference on big data (big data). IEEE; 2020, p. 33–43.

[47] Tuli Shreshth, Casale Giuliano, Jennings Nicholas R. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. Proc VLDB Endow 2022;15(6):1201–14.

[48] Audibert Julien, Michiardi Pietro, Guyard Frédéric, Marti Sébastien, Zuluaga Maria A. Usad: Unsupervised anomaly detection on multivariate time series. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020, p. 3395–404.

[49] Zong Bo, Song Qi, Min Martin Renqiang, Cheng Wei, Lumezanu Cristian, Cho Daeki, Chen Haifeng. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International conference on learning representations. 2018.

[50] Zhao Hang, Wang Yujing, Duan Juanyong, Huang Congrui, Cao Defu, Tong Yunhai, Xu Bixiong, Bai Jing, Tong Jie, Zhang Qi. Multivariate time-series anomaly detection via graph attention network. In: 2020 IEEE international conference on data mining. ICDM, IEEE; 2020, p. 841–50.

[51] Deng Ailin, Hooi Bryan. Graph neural network-based anomaly detection in multivariate time series. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35, 2021, p. 4027–35.

[52] Zhang Yuxin, Chen Yiqiang, Wang Jindong, Pan Zhiwen. Unsupervised deep anomaly detection for multi-sensor time-series signals. IEEE Trans Knowl Data Eng 2021;35(2):2118–32.

[53] Zhang Chuxu, Song Dongjin, Chen Yuncong, Feng Xinyang, Lumezanu Cristian, Cheng Wei, Ni Jingchao, Zong Bo, Chen Haifeng, Chawla Nitesh V. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, 2019, p. 1409–16.

[54] Hundman Kyle, Constantinou Valentino, Laporte Christopher, Colwell Ian, Soderstrom Tom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018, p. 387–95.

[55] Prenkaj Bardh, Velardi Paola. Unsupervised detection of behavioural drifts with dynamic clustering and trajectory analysis. In: IEEE transactions on knowledge and data engineering. 2023.

[56] Bieber Gerald, Kirste Thomas, Gaede Michael. Low sampling rate for physical activity recognition. In: Proceedings of the 7th international conference on pErvasive technologies related to assistive environments. PETRA '14, New York, NY, USA: Association for Computing Machinery; 2014.

[57] Achiam Josh, Adler Steven, Agarwal Sandhini, Ahmad Lama, Akkaya Ilge, Aleman Florencia Leoni, Almeida Diogo, Altenschmidt Janko, Altman Sam, Anadkat Shyamal, et al. Gpt-4 technical report., 2023, arXiv preprint arXiv:2303.08774.

[58] Touvron Hugo, Lavril Thibaut, Izacard Gautier, Martinet Xavier, Lachaux Marie-Anne, Lacroix Timothée, Rozière Baptiste, Goyal Naman, Hambro Eric, Azhar Faisal, et al. Llama: Open and efficient foundation language models. 2023, arXiv preprint arXiv:2302.13971.

[59] Xie Qiujie, Feng Qiming, Zhang Tianqi, Li Qingqiu, Zhang Yuejie, Feng Rui, Gao Shang. Human simulacra: A step toward the personification of large language models. 2024, CoRR,.

[60] Katsigiannis Stamos, Ramzan Naeem. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. IEEE J Biomed Health Inform 2017;22(1):98–107.

[61] Soleymani Mohammad, Lichtenauer Jeroen, Pun Thierry, Pantic Maja. A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput 2011;3(1):42–55.

[62] Schmidt Philip, Reiss Attila, Duerichen Robert, Marberger Claus, Laerhoven Kristof Van. Introducing wesad a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM international conference on multimodal interaction. 2018, p. 400–8.
[63] Reali Federica, et al. Personalized digital phenotyping from longitudinal physiological data in real-life settings. Npj Digit Med 2023;6(1):1–12.
[64] Ciman Matteo, Wac Katarzyna. Smart health monitoring through wearable sensors: real-world evaluation and challenges. In: Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, vol. 3, ACM; 2019, p. 1–20.
[65] Luo Simon, Zhou Jianlong, Duh Henry Been-Lirn, Chen Fang. Bvp feature signal analysis for intelligent user interface. In: Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems. CHI EA '17, New York, NY, USA: Association for Computing Machinery.; 2017, p. 1861–8.
[66] Parvin Parvaneh, Chessa Stefano, Kaptein Maurits, Paternò Fabio. Personalized real-time anomaly detection and health feedback for older adults. J Ambient Intell Smart Environ 2019;11(09):453–69.
[67] Ates H Ceren, Ates Cihan, Dincer Can. Stress monitoring with wearable technology and ai. Nat Electron 2024;1–2.