REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

STRESS CLASSIFICATION USING PHYSIOLOGICAL SIGNALS: A COMPREHENSIVE REVIEW OF METHODS AND APPROACHES COMBINED WITH A NOVEL CNN-BASED ECG EXPERIMENT

CLASSIFICAÇÃO DO ESTRESSE USANDO SINAIS FISIOLÓGICOS: UMA REVISÃO ABRANGENTE DE MÉTODOS E ABORDAGENS COMBINADOS COM UM NOVO EXPERIMENTO DE ECG BASEADO EM CNN

CLASIFICACIÓN DEL ESTRÉS UTILIZANDO SEÑALES FISIOLÓGICAS: UNA REVISIÓN EXHAUSTIVA DE MÉTODOS Y ENFOQUES COMBINADA CON UN NUEVO EXPERIMENTO DE ECG BASADO EN CNN

Clarissa Rodrigues[1], Sandro Rigo[1], Kauã Mark[2]

e747417
https://doi.org/10.47820/recima21.v7i4.7417

PUBLISHED: 04/2026

ABSTRACT

Accurate stress detection through physiological signals has strong potential to improve healthcare outcomes, reduce costs, and enable early intervention in stress-related disorders. This work provides a comprehensive review of recent advances in stress classification using physiological data, highlighting key methods, challenges, and emerging trends in the field. Special emphasis is given to the limitations posed by small datasets, the importance of personalized models, and the difficulties of real-time application in uncontrolled environments. In parallel, we propose and evaluate a novel convolutional neural network (CNN) architecture designed to classify electrocardiogram (ECG) signals into four distinct categories. The model demonstrates robust learning and reasonable generalization under data-constrained conditions, achieving 72.81% accuracy on an independent test set. The findings reinforce the efficacy of deep learning in stress classification and underscore the need for personalized, real-time, and multimodal approaches in future research.

KEYWORDS: Stress. Classification. Physiological signals. Deep learning. Convolutional neural networks. Electrocardiogram (ECG).

RESUMO

A detecção precisa do estresse por meio de sinais fisiológicos apresenta grande potencial para melhorar os resultados em saúde, reduzir custos e possibilitar a intervenção precoce em distúrbios relacionados ao estresse. Este estudo apresenta uma revisão abrangente dos avanços recentes na classificação do estresse com base em dados fisiológicos, destacando os principais métodos, desafios e tendências emergentes na área. Ênfase especial é dada às limitações impostas por conjuntos de dados reduzidos, à importância de modelos personalizados e às dificuldades de aplicação em tempo real em ambientes não controlados. Paralelamente, propomos e avaliamos uma nova arquitetura de rede neural convolucional (CNN) projetada para classificar sinais de eletrocardiograma (ECG) em quatro categorias distintas. O modelo demonstrou aprendizado robusto e generalização moderada em condições de restrição de dados, alcançando 72,81% de acurácia em um conjunto de teste independente. Os achados reforçam a eficácia do aprendizado profundo na classificação do estresse e ressaltam a necessidade de abordagens personalizadas, em tempo real e multimodais em pesquisas futuras.

PALAVRAS-CHAVE: Estresse. Classificação. Sinais fisiológicos. Aprendizado profundo. Redes neurais convolucionais. Eletrocardiograma (ECG).

[1] University of Vale do Rio dos Sinos - UNISINOS.
[2] Pontifical Catholic University of Rio Grande do Sul - PUCRS.

RESUMEN
La detección precisa del estrés mediante señales fisiológicas muestra un gran potencial para mejorar los resultados de la atención médica, reducir costos y permitir la intervención temprana en trastornos relacionados con el estrés. Este estudio presenta una revisión exhaustiva de los avances recientes en la clasificación del estrés utilizando datos fisiológicos, destacando los métodos clave, los desafíos y las tendencias emergentes en el campo. Se hace especial hincapié en las limitaciones que plantean los conjuntos de datos pequeños, la importancia de los modelos personalizados y las dificultades de la aplicación en tiempo real en entornos no controlados. Paralelamente, proponemos y evaluamos una nueva arquitectura de red neuronal convolucional (CNN) diseñada para clasificar las señales del electrocardiograma (ECG) en cuatro categorías distintas. El modelo muestra un aprendizaje robusto y una generalización razonable en condiciones de datos limitados, alcanzando una precisión del 72,81% en un conjunto de pruebas independiente. Los hallazgos refuerzan la eficacia del aprendizaje profundo en la clasificación del estrés y subrayan la necesidad de enfoques personalizados, en tiempo real y multimodales en futuras investigaciones.

PALABRAS CLAVE: Estrés. Clasificación. Señales fisiológicas. Aprendizaje profundo. Redes neuronales convolucionales. Electrocardiograma (ECG).

1. INTRODUCTION

Physiological signal classification has become a central research topic in biomedical signal processing, with applications spanning healthcare monitoring, affective computing, and stress detection systems. Among the available physiological modalities, electrocardiogram (ECG) signals are particularly relevant due to their non-invasive nature and their ability to capture cardiovascular dynamics associated with autonomic nervous system responses. In recent years, deep learning approaches have consistently outperformed traditional machine learning methods in time-series analysis tasks by automatically learning hierarchical feature representations directly from raw or minimally preprocessed signals (Ansari *et al.,* 2023).

Despite these advances, stress classification using physiological signals remains a challenging problem. Real-world datasets are often limited in size, exhibit significant class imbalance, and are collected under uncontrolled conditions, which substantially hinders model generalization. Furthermore, inter-individual variability in physiological responses to stress frequently leads to performance degradation when generalized models are applied across subjects, reinforcing the need for robust architectures capable of operating under constrained data scenarios (Finseth *et al.,* 2023).

Although recent studies have proposed hybrid architectures that combine convolutional neural networks with recurrent or attention-based mechanisms to capture long-range temporal dependencies, such models typically require large annotated datasets and considerable computational resources. These requirements restrict their applicability in real-time or resource-constrained environments, such as wearable stress monitoring systems (Modi *et al.,* 2025).

Consequently, there is a strong motivation to investigate compact and computationally efficient models that balance performance, robustness, and deployability.

Convolutional neural networks (CNNs), particularly one-dimensional architectures, remain widely adopted as effective baselines for physiological time-series classification. CNNs are well suited for extracting local temporal patterns and morphological characteristics from ECG signals, offering a favorable trade-off between representational capacity and computational cost (Guhdar; Mohammed; Mstafa, 2025). When combined with appropriate preprocessing, class weighting, and regularization strategies, CNNs have demonstrated stable learning behavior even in data-limited and imbalanced scenarios.

In this context, the present work investigates a compact one-dimensional CNN architecture for multiclass temporal signal classification, with an emphasis on robustness under realistic experimental constraints. The proposed model is evaluated under conditions of limited data availability, noise, and class imbalance, with the objective of assessing whether a carefully regularized CNN can achieve reliable generalization without relying on complex hybrid architectures or extensive data augmentation.

2. STATE-OF-THE-ART RESEARCH METHODOLOGY

A non-systematic literature review was conducted using a snowball sampling strategy, starting from key studies and expanding through their reference lists. The databases consulted included ACM Digital Library, IEEE Xplore, Elsevier, Springer, and ScienceDirect. The keywords “stress,” “physiological signals,” “classification,” and “customization” were used to identify publications published from 2021 onward. A total of 27 papers were initially retrieved, of which seven were selected after relevance screening.

The review aimed to address the following research questions:

- (a) What are the physiological signals used, individually or in combination?
- (b) What are the relevant algorithms used in this context?
- (c) What are the techniques used to deal with customization challenges?

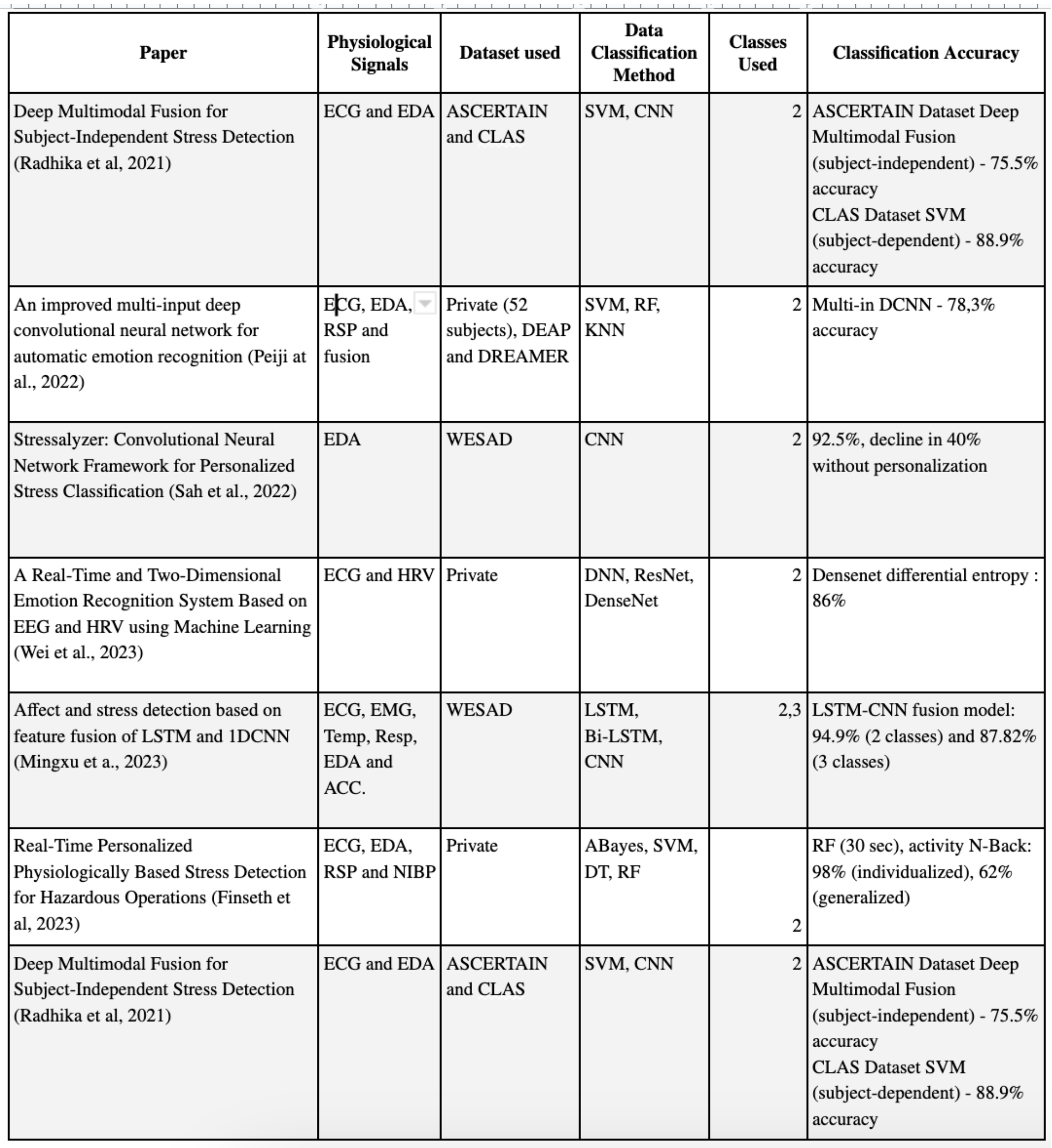
Papers considered not relevant to the research questions were removed, resulting in seven articles. To foster the analysis and support the answering of the research questions, the chosen papers were comparatively analyzed based on the following information:

- Objective: Main research goals of the papers under review;
- Physiological Signals: Types of physiological signals used and their combinations;
- Dataset Used: Types of datasets used, being real data collected from some experiment, synthetic data, and also evaluating their availability for public use
- Data Classification Methods: Algorithms used in the studies for data classification;
- Classes Used: Classes extracted using different classification methods and signals;

- Accuracy of Classification: Analysis of the effectiveness of the result of the study, using different combinations of physiological signals and methods.

The papers selected for the study are described in Table 1. Most of the papers studied employ a specific architecture, typically a convolutional neural network (CNN), with multi- or monosignal inputs. There are very few papers that use multiple signals, and even fewer that address customization models, in recent years for stress classification. This denotes the importance of this challenge. Most studies used data collected specifically for the experiment. There are publicly available datasets related to emotional oscillations, including ASCERTAIN (Subramanina *et al.,* 2016), DEAP (Koelstra *et al.,* 2012), WESAD (Schmidt et al., 2018), CLAS (Markova *et al.,* 2022), and DREAMER (Katsigiannis *et al.,* 2018). The ASCERTAIN dataset contains Big Five personality scales and emotional self-ratings from 58 users, along with their electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), and facial activity data, recorded using off-the-shelf sensors while viewing affective movie clips. DEAP collected EEG data from only 32 participants, each of whom watched 40 one-minute excerpts of music videos. Participants rated each video on arousal, valence, like/dislike, dominance, and familiarity.

Table 1. Selected Studies



Paper	Physiological Signals	Dataset used	Data Classification Method	Classes Used	Classification Accuracy
Deep Multimodal Fusion for Subject-Independent Stress Detection (Radhika et al, 2021)	ECG and EDA	ASCERTAIN and CLAS	SVM, CNN	2	ASCERTAIN Dataset Deep Multimodal Fusion (subject-independent) - 75.5% accuracy CLAS Dataset SVM (subject-dependent) - 88.9% accuracy
An improved multi-input deep convolutional neural network for automatic emotion recognition (Peiji at al., 2022)	ECG, EDA, RSP and fusion	Private (52 subjects), DEAP and DREAMER	SVM, RF, KNN	2	Multi-in DCNN - 78,3% accuracy
Stressalyzer: Convolutional Neural Network Framework for Personalized Stress Classification (Sah et al., 2022)	EDA	WESAD	CNN	2	92.5%, decline in 40% without personalization
A Real-Time and Two-Dimensional Emotion Recognition System Based on EEG and HRV using Machine Learning (Wei et al., 2023)	ECG and HRV	Private	DNN, ResNet, DenseNet	2	Densenet differential entropy : 86%
Affect and stress detection based on feature fusion of LSTM and 1DCNN (Mingxu et a., 2023)	ECG, EMG, Temp, Resp, EDA and ACC.	WESAD	LSTM, Bi-LSTM, CNN	2,3	LSTM-CNN fusion model: 94.9% (2 classes) and 87.82% (3 classes)
Real-Time Personalized Physiologically Based Stress Detection for Hazardous Operations (Finseth et al, 2023)	ECG, EDA, RSP and NIBP	Private	ABayes, SVM, DT, RF	2	RF (30 sec), activity N-Back: 98% (individualized), 62% (generalized)
Deep Multimodal Fusion for Subject-Independent Stress Detection (Radhika et al, 2021)	ECG and EDA	ASCERTAIN and CLAS	SVM, CNN	2	ASCERTAIN Dataset Deep Multimodal Fusion (subject-independent) - 75.5% accuracy CLAS Dataset SVM (subject-dependent) - 88.9% accuracy

WESAD is the broadest known dataset in the area of emotional recognition, containing blood volume pulse (BVP), ECG, electrodermal activity (EDA), electromyogram (EMG), respiration, body temperature, and three-axis acceleration recorded from both a wrist- and a chest-worn device, of 15 subjects during a lab study. CLAS includes ECG, plethysmography (PPG), EDA, and accelerometer data from 62 healthy volunteers, collected while they participated in three interactive tasks and two perceptual tasks. Finally, DREAMER includes EEG and EDA data from 23 participants, along with participants' self-assessments of their affective state after each stimulus, in terms of valence, arousal, and dominance.

While these datasets make data publicly available, they have the limitation of a very small number of participants and focus more on comparing different classification methods than on the data collection itself. Generally, recognizing such changes requires analyzing a large amount of data. Techniques such as time windows are used to reduce the data, but the challenge lies in maintaining a sufficiently large dataset so that no relevant information is lost.

Qualitative analysis indicated that the most commonly used algorithms for this purpose were convolutional neural networks (CNNs) and support vector machines (SVMs). The most commonly used devices for data collection are Empatica, Emotiv, and SHIMMER, with heart rate (HR) and GSR as the least intrusive signals and achieving the highest accuracy for stress detection. Collecting real-time data in uncontrolled environments, removing noise, and ensuring data persistence remain the biggest challenges in this area.

Additional algorithms used in this area include the comparison between machine learning (ML) algorithms, random forest (RF), explainable neural network (xNN), linear regression (LR), support vector machine (SVM), and long short-term memory (LSTM) using EDA and BVP signals (Nath *et al.,* 2021). Compared with the best-performing ML algorithm, LSTM achieved an accuracy of 81%.

The electrocardiogram (ECG) is among the most widely used physiological signals in machine learning (ML)- based stress detection systems. The electroencephalogram (EEG) signal is also increasingly common (Zontone *et al.,* 2022). Electrodermal activity (EDA) has also been used in many studies, as it is strongly correlated with stress detection. Other signals used include electromyography (EMG), blood volume pulse (BVP), respiratory rate, body temperature, and accelerometer data.

Supervised learning classification methods, such as support vector machines (SVMs) and convolutional neural networks (CNNs), are the most commonly used in the selected papers. This is due to their strong performance in both accuracy and computational efficiency. However, further studies employing new classification methods, such as unsupervised learning, artificial neural networks (ANNs), deep learning, and reinforcement learning, are needed to enable a comparative analysis among them.

While the challenges under discussion remain debated, related ethical concerns include privacy, security, and legislation, as well as the reliability of results, cost, and interoperability (Hern´andez *et al.,* 2021). Regarding reliability, other sources of emotional change can also activate the sympathetic nervous system (SNS) and elicit similar heart rate (HR) signals, and everyday situations can introduce bias in the results (Panicker *et al.,* 2019). Radhika *et al.,* (2021) compared the use of support vector machines (SVMs) and convolutional neural networks (CNNs) on the public

ASCERTAIN and CLAS datasets. For ASCERTAIN, 75.5 % accuracy was obtained using deep multimodal fusion (subject-independent), while CLAS dataset obtained 88.9 % accuracy (subject-dependent). The results show that convolutional layers affect deep multimodal fusion and that the generalization capability of subject-independent stress detection models is lower than that of subject-dependent models.

Another study proposed a multi-input deep convolutional neural network (DCNN) that extracts features from different input signals separately. Filters across channels are not shared, which mitigates inter-channel interference and enables automatic feature extraction. The study compared SVM, random forest (RF), and K-nearest neighbors (KNN) methods, with 78.3 % accuracy with DCNN as the best result. To obtain a model with stronger generalization ability, the individual and temporal differences of biological signals should be considered (Peldˇzi *et al.,* 2022).

Sah *et al.,* (2022) reported an impressive 92.5% accuracy with a CNN, which declined by 40% without customization. To address the customization challenge, the authors employed an online learning method to personalize the stress model for a specific user. In the online learning scenario, a general machine model M1 is retrained on data obtained from the user while the model is in use. The model is retrained until the performance of the personalized model (M2), on the user data is at an acceptable level. The leave-one-subject-out (LOSO) analysis was also used, in which data from one subject are removed from the training set and retained as the test set to evaluate the machine learning model trained on data from all other subjects. To quantify performance decline, the difference in the model's accuracy on the training and test set was calculated. The subject requires customization if the difference is greater than 5.

Densenet, DNN, and ResNet were compared, with Densenet achieving the highest accuracy. This suggests that a neural network model, such as an LSTM, can be used to analyze data over longer periods to improve classification accuracy (Wei *et al.,* 2022). Mingxu *et al.,* (2023) reported promising results comparing Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), and convolutional neural network (CNN) models, achieving 94.9% accuracy for 2 classes and 87.82% for 3 classes. This is because LSTM models pay greater attention to the temporal features of physiological signals, whereas CNN models focus more on the spatial correlations among physiological signals.

Finseth *et al.,* (2023) compared algorithms from prior studies: ABayes, support vector machine (SVM), decision tree (DT), and random forest (RF). RF for the N-Back activity achieved 98% accuracy with the individualized method and 62% with the generalized method. The highest 10-fold cross-validation performance for the VR-ISS across all windows and classifiers was 94%, achieved with an ABayes classifier and a window size of 30 seconds, suggesting that the personalized approach performed well. In addition to the excellent results obtained with CNNs, a recent study demonstrated the use of CNNs in conjunction with RNNs. This study, which mapped the use of deep learning (DL) on medical physiological data for 2 years, found that EEG and ECG are the most widely used physiological signals, with few studies in the multi-signal field (EEG with 79 studies, ECG with 47, and multi-signal with only 5) (Rim *et al.,* 2020).

Only one study (Li *et al.,* 2022) compared deep learning (DL) with different machine learning (ML) algorithms using the exact same dataset. The study found that DL outperformed ML, achieving 99.55% accuracy compared with 76.50% for LDA and 98.38% compared to 75.21 % for AdaBoost. When the accelerometer (ACC) data were removed, performance consistency was maintained, with 97.48% accuracy for DL, 80.34% for AdaBoost, and 93.64% for random forest (RF), compared with 76.17% for RF. However, there is not yet a sufficiently large sample size to consistently determine which of these methods is more efficient, due to parameterization and architectural complexity.

There are a few studies that compare ML and DL, and even fewer that compare DL algorithms. Prerna *et al.,* (2021) used ECG, TEMP, RESP, EMG, and EDA with ML (KNN, LDA, RD, AdaBoost, and SVM), achieving their best result of 65.73 % with RF. Other works have shown an accuracy of 92 % with LSTM (DL) compared to 96 % of SVM (Vargas-Lopez *et al.,* 2021), 95 % with ANN and 93 % with SVM (Bobade *et al.,* 2020), 88 % with the proposed DL model and 75 % with RF (Kumar *et al.,* 2021), and DL with CNN and LSTM with better results than ML (Huang *et al.,* 2022). Even fewer studies compare DL algorithms. Among them, Artificial Neural Network (ANN), SVM, Stacking Classifier, and Radial Basis Function Neural Networks (RBF) were compared, yielding the best result of 99.92 % accuracy with Stacking Classifier and the worst result of 84.46 % with RBF (Vishal Dham *et al.,* 2021). CNN-based stress detection using different signal-processing techniques to generate inputs for this architecture (Fourier Transform, cube root, and CQT) has been shown to achieve 96.6% accuracy (Gil-Martin *et al.,* 2022). Fatma (2022) analyzed CNN, LSTM, and RNN, reporting an accuracy of 93%, which was compared with the Autoregressive (AR-HMM) model previously implemented by the same author.

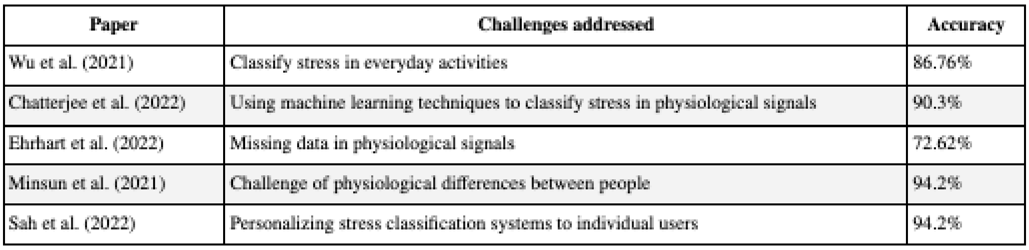
The present work identified a research gap in the use of different signals and innovation in their pre-processing stage, seeking to identify more promising combinations for the detection of stress patterns. The use of biological markers such as cortisol, recognized as being important for stress identification, was not observed in the studies. The use of this indicator is recommended to validate the data annotation step, thereby improving the accuracy of the dataset used in the experiments.

In the next section, specific challenges in this area are addressed.

Stress classification using physiological signals is a challenging task. Several challenges must be addressed to develop accurate and reliable stress detection systems. These challenges that will be focused on here are (a) the use of physiological signals in everyday activities, (b) the lack of public datasets, (c) the time series nature of physiological signals, and (d) the non-independent and identically distributed (i.i.d.) nature of physiological signals, among others.

Stress classification using physiological signals has previously been mentioned as known challenges specific to this context, such as use in everyday activities, few public datasets, an insufficient amount of existing data, and the time series nature of physiological signals that violate the machine learning assumption that the data are independently and identically distributed, thereby leading to inaccurate and biased results. (Poorya *et al,* 2022; Cederick *et al.,* 2021; Sadhana *et al.,* 2020; Pau Climent-Perez *et al.,* 2022; Finseth *et al.,* 2023), among others. Next are cited new experiments highlighting techniques with promising results when addressing these problems and the general pattern classification task, as described in Table 2:

Table 2. Main techniques in the selected papers



Paper	Challenges addressed	Accuracy
Wu et al. (2021)	Classify stress in everyday activities	86.76%
Chatterjee et al. (2022)	Using machine learning techniques to classify stress in physiological signals	90.3%
Ehrhart et al. (2022)	Missing data in physiological signals	72.62%
Minsun et al. (2021)	Challenge of physiological differences between people	94.2%
Sah et al. (2022)	Personalizing stress classification systems to individual users	94.2%

Wu *et al.,* (2021) demonstrated the application of Transfer Learning, which is the reuse of a pre-trained model to solve similar tasks, through three modules: feature extraction, domain discrimination, and a stress detector. In this study, a pre-trained VGG16 model was used, comprising a two-level BLSTM network for feature extraction with 64 LSTM units. Deep features and manually extracted features are combined with 50 and 20 units in the two layers used earlier in the domain definition module. For stress detection, the units are 30 and 10. The model is activated using the Rectified Linear Unit (ReLU), which returns 0 for negative inputs and the input value for positive inputs. During training, a batch size of 8 is used along with the Adam optimizer to minimize loss, updating the model weights based on the test data instead of the traditional stochastic gradient descent method, with its learning rate at 0.0001 and the hyperparameters selected and cross-validated 5 times.

This framework achieved an accuracy of 86.76% compared with other models. Chatterjee *et al.,* (2022) presented an innovative model structured with 30-second windows that apply the Fast Fourier Transform (FFT), which converts a signal from its original domain to the frequency domain (and vice versa), for each pair and then sorts the pairs in descending order by amplitude.

As a result, only the top 10 features (amplitude, frequency) of each window are used, achieving an accuracy of 90.3 % for 2 classes and 94.2 % for binary classification. Ehrhart *et al.,* (2022) use a Conditional Generative Adversarial Network (cGAN) architecture for the missing data problem, combining an LSTM with a Fully Convolutional Network (FCN) in two classes, with 16-second windows. With the help of auxiliary data generated by cGAN, this model achieved 72.62%

accuracy on the test dataset, resulting in substantial improvements in recall (+19.05%) and F1-score (+11.03%).

Classifying stress in everyday activities is challenging because it must distinguish changes in physiological data caused by stress itself from those resulting from other daily activities that elevate heart rate, for example. Minsun *et al.,* (2021) used the Empatica E4 wristband, collecting 5 physiological signals in real time: ACC, BVP, GSR, skin temperature (ST,) and heart rate (HR) that are processed by H4's internal algorithm for BVP. All the collected data is used to extract features and to train the ML algorithms. In another study, Empatica was also used to collect blood volume (BVP), electrodermal activity (EDA), and skin temperature with sample rates of 64 Hz, 4 Hz, and 4 Hz, respectively.

A major challenge in using physiological signals for detection is the rigidity of generalized models in accounting for interindividual physiological variation. Only two studies have gone deep on this. Sah et al. [11] used a 1-dimensional CNN architecture composed of two convolutional layers with filter sizes of 5. Convolution layers are followed by a global max-pooling layer and neurons. They also have dropout layers after two fully connected layers with drop-out values of 0.3. The output layer uses Softmax activation, whereas all other layers use ReLU activation. He used an online learning method to personalize the stress model to a specific user. In the online learning scenario, a general machine model M1 is retrained on data obtained from the user while using it. The model is retrained until the personalized model (M2) achieves an acceptable performance on the user data. CNN models were trained for up to 50 epochs, with a batch size of 64 and a fixed learning rate of 0.001, and leave-one-subject-out (LOSO) was used to tackle the customization challenge.

Among the algorithms, Abayes showed the best performance (Finseth *et al.,* 2023), followed by LSTM-CNN (Mingxu *et al.,* 2023). Other studies have also shown good results with CNNs and LSTMs compared with other experiments using Empathic (Akbulut, Fatma, 2022; Cosoli *et al.,* 2021; Zitouni *et al.,* 2021). LSTM has also been shown to be effective in distinguishing changes attributable to physical activity (sedentary state, treadmill running, bicycle ergometer) from those attributable to stress (Askari *et al.,* 2022).

In light of these challenges, particularly limited data availability, class imbalance, and the difficulty of generalization in multiclass settings, the following section presents an experimental CNN-based model designed to operate under these constraints. The proposed experiment aims to evaluate whether a compact and regularized convolutional architecture can achieve stable learning and reasonable performance in scenarios that more closely resemble real-world stress classification conditions.

3. EXPERIMENT: CNN MODEL USING ECG

This section presents the experimental setup and evaluation of a CNN model designed for multiclass classification of one-dimensional temporal signals. The experiment focuses on assessing the model’s robustness, generalization capability, and suitability for constrained datasets. The following subsections describe the model architecture, dataset, and preprocessing pipeline, training configuration, and quantitative results.

3.1. The dataset

The WearHealth dataset was developed as part of an interdisciplinary research project on psychophysiological stress measurement using wearable sensors, conducted between 2021 and 2025 with ethical approval and informed consent from all participants. It comprises data from 74 individuals recruited from the general population, making it larger than several widely used public stress datasets. Stress was induced using a standardized and validated protocol (TSST), enabling controlled elicitation of physiological and hormonal stress responses under realistic laboratory conditions.

The dataset integrates multimodal physiological signals, including ECG, EDA, and EMG, continuously recorded for approximately 60–70 minutes per participant, alongside heart rate and HRV measurements from a reference device. In addition, four salivary cortisol samples were collected to capture the delayed hormonal response of the HPA axis, providing a complementary and objective biomarker of stress. This combination of autonomic (fast) and endocrine (slow) responses distinguishes the dataset and enables richer analysis of stress dynamics.

Beyond physiological data, the dataset includes comprehensive psychological assessments, anthropometric measurements, and detailed protocol annotations. Data is labeled across four phases - baseline, anticipation, active stress, and recovery - allowing fine-grained temporal analysis rather than simple binary classification. This multimodal and structured design supports research on stress detection, individual variability, and model generalization, while also enabling future work on label refinement using biochemical validation.

3.2. Model architecture

The proposed model is a one-dimensional CNN model designed for multiclass classification of temporal signals. The architecture prioritizes a balance between representational capacity and computational efficiency, enabling stable learning under data-constrained and imbalanced conditions.

The network comprises three hierarchical convolutional blocks with filter sizes of 32, 64, and 128, respectively. Each convolutional layer employs a kernel size of 3 with padding to preserve temporal resolution, followed by batch normalization and ReLU activation to improve convergence stability and mitigate internal covariate shift. Max-pooling layers are applied after the first two convolutional blocks to reduce temporal dimensionality and control overfitting.

To further enhance generalization, dropout regularization with a rate of 0.3 is applied after the final convolutional block. Instead of flattening the feature maps, an adaptive average pooling layer aggregates temporal information into a fixed-length representation, enabling the model to handle variable input lengths while preserving salient temporal patterns.

The classification stage consists of two fully connected layers. The first dense layer maps the extracted features into a 128-dimensional latent space using ReLU activation, while the final layer outputs four neurons corresponding to the target classes. This architectural design enables effective hierarchical feature extraction while maintaining robustness against noise and class imbalance.

3.3. Dataset and preprocessing

The dataset consists of one-dimensional temporal signal recordings stored in CSV format, where each sample corresponds to a fixed-length time window associated with one of four target classes. Input features and labels were loaded from separate files to maintain modularity in data handling.

All input values were converted to numeric format, and missing entries were addressed through column-wise mean imputation to ensure numerical stability without discarding samples. Class labels were encoded as integer values to support supervised multiclass classification.

To reflect realistic data availability constraints, only 75% of the total dataset was used in the experiments. Within this subset, stratified sampling was used to preserve the class distribution across splits. The data was divided into training, validation, and test sets, with 80% allocated to training and validation and 20% reserved for independent testing.

Feature normalization was performed using z-score standardization, fitted exclusively on the training set and subsequently applied to the validation and test sets to prevent data leakage. Finally, an additional channel dimension was added to each input sample to adapt the data format to one-dimensional convolutional layers.

This preprocessing pipeline ensures numerical consistency, preserves temporal structure, and mitigates the effects of class imbalance, thereby aligning the experimental setup with real-world deployment scenarios. This methodology contrasts with the large-scale, balanced datasets commonly used in the reviewed literature, such as ASCERTAIN, DEAP, and DREAMER.

3.4. Training configuration

Model training was conducted using a supervised learning framework optimized for imbalanced multiclass classification. The loss function employed was categorical cross-entropy with class weights computed from the training data distribution, ensuring proportional contribution from minority classes during optimization.

The Adam optimizer was used with a learning rate of 0.001 and a weight decay factor of 1 × 10^{-5}” to provide regularization. Training was performed using mini-batches of size 64, balancing computational efficiency and gradient stability.

A ReduceLROnPlateau learning rate scheduler monitored validation loss and dynamically reduced the learning rate when performance stagnation was detected. Additionally, early stopping was applied with a patience of 7 epochs, terminating training when validation loss no longer improved and restoring the best-performing model parameters.

Training was run for up to 50 epochs, though convergence was typically achieved earlier due to early stopping. This configuration promoted stable convergence, reduced overfitting, and improved generalization performance under limited data conditions.

3.5. Results

The trained model was evaluated on an independent test set using accuracy, precision, recall, and F1-score as performance metrics. The proposed CNN achieved an overall classification accuracy of 72.81%, indicating strong performance for a four-class temporal signal classification task under constrained conditions.

Class-wise analysis showed that Classes 0 and 1 achieved the highest performance, with F1 Scores exceeding 0.79, indicating reliable discrimination of these signal patterns. In contrast, Classes 2 and 3 exhibited lower recall values, suggesting partial overlap in their temporal characteristics, an issue commonly reported in multiclass time-series classification.

The macro-averaged F1-score was 0.7265, while the weighted F1-score reached 0.7263, indicating balanced performance across classes despite dataset imbalance. Training and validation loss curves exhibited stable, consistent convergence without significant divergence, confirming the effectiveness of the adopted regularization and optimization strategies.

Overall, the results demonstrate that the proposed CNN architecture achieves robust generalization and competitive performance in data-limited and noisy environments, supporting its applicability to practical temporal signal classification problems.

4. CONCLUSION

This study presented a one-dimensional convolutional neural network (CNN) model for temporal signal classification under realistic constraints, including limited dataset size, class imbalance, and signal noise. By prioritizing architectural simplicity and effective regularization, the proposed approach demonstrated stable learning and consistent generalization, reinforcing the suitability of CNNs as practical baselines for physiological signal analysis.

The experimental results indicate that the model can learn discriminative temporal representations despite data constraints. Such behavior is consistent with recent findings in the literature, which emphasize that CNN-based models remain competitive for ECG and physiological time-series classification when appropriately designed and trained (Ansari *et al.,* 2023). Rather than pursuing maximal accuracy through highly complex architectures, this work emphasizes robustness, reproducibility, and computational efficiency.

Nevertheless, the results also reveal persistent challenges inherent to multiclass stress classification. Overlapping temporal characteristics between classes and inter-subject variability continue to limit separability and performance gains, particularly in generalized models. These limitations are widely acknowledged in recent studies and reflect fundamental issues associated with physiological signal variability in real-world environments (Xiang *et al.,* 2025).

Future research directions include exploring architectures with enhanced temporal modeling capabilities, such as attention mechanisms or hybrid CNN-Transformer designs, which have shown promise in capturing broader contextual information from physiological signals (Modi *et al.,* 2025). Additionally, increasing the dataset size, improving annotation strategies, and leveraging transfer learning from larger ECG repositories may further enhance model robustness while preserving generalization.

Overall, this work presents a reproducible and scalable CNN-based framework that aligns with current deep learning trends while accounting for the practical constraints of physiological signal classification. By emphasizing generalization under limited-data conditions, the proposed approach provides a solid foundation for future investigations and real-world stress-monitoring applications.

The proposed CNN directly targets a notable research gap by focusing on low-data, noisy, and imbalanced ECG scenarios that are underrepresented in the stress classification literature. While its current performance is competitive with other generalized approaches under similar constraints, future enhancements aim to narrow the gap with state-of-the-art personalized systems. These include integrating recurrent layers (LSTM/BiLSTM) to capture long-term temporal dependencies; applying attention mechanisms to improve temporal feature weighting; combining ECG with complementary modalities, such as electrodermal activity (EDA) and electroencephalography (EEG), for multimodal fusion; expanding the dataset size; and leveraging transfer learning from large ECG repositories. Collectively, these refinements have the potential to bridge the performance gap between generalized and personalized models, enabling accurate, adaptable, and deployable stress classification systems in real-world healthcare applications.

REFERENCES

ANSARI, Y.; MOURAD, O.; QARAQE, K.; SERPEDIN, E. Deep learning for ECG arrhythmia detection and classification: an overview of progress for 2017–2023. **Frontiers in Physiology**, Lausanne, v. 14, 2023.

ASKARI, M.; SETAREHDAN, S. K.; SHOKROLLAHI, A.; ALIREZAEI, M. Stress detection from physiological signals using deep learning methods: distinguishing stress from physical activity. **Biomedical Signal Processing and Control**, Amsterdam, v. 73, 2022.

BOBADE, P.; VANI, M. Stress detection with machine learning and deep learning using multimodal physiological data. *In:* INTERNATIONAL CONFERENCE ON INVENTIVE RESEARCH IN COMPUTING APPLICATIONS (ICIRCA), 2., 2020. **Proceedings** […]. [S. l.]: IEEE, 2020. p. 51–57.

CHATTERJEE, R.; NAG, A.; DUTTA, A.; MUKHERJEE, S. Stress classification using physiological signals with frequency-domain feature selection. **Biomedical Signal Processing and Control,** Amsterdam, v. 72, 2022.

FINSETH, T.; KJELDGAARD, M.; SKOV, M. B.; HANSEN, L. K. Personalized versus generalized models for physiological stress detection: challenges and opportunities. **Sensors**, Basel, v. 23, n. 3, 2023.

GARG, P.; SANTHOSH, J.; DENGEL, A.; ISHIMARU, S. **Stress** detection by machine learning and wearable sensors. *In:* INTERNATIONAL CONFERENCE ON INTELLIGENT USER INTERFACES – COMPANION, 26., 2021, New York. **Proceedings** […]. New York: ACM, 2021. p. 43–45.

GEDAM, S.; PAUL, S. A review on mental stress detection using wearable sensors and machine learning techniques. **IEEE Access**, New York, v. 9, p. 84045–84066, 2021.

GUHDAR, M.; MOHAMMED, A. O.; MSTAFA, R. J. Advanced deep learning framework for ECG arrhythmia classification using 1D-CNN with attention mechanism. **Knowledge-Based Systems**, Amsterdam, v. 295, 2025.

HUANG, J.; LIU, Y.; PENG, X. Recognition of driver's mental workload based on physiological signals: a comparative study. **Biomedical Signal Processing and Control,** Amsterdam, v. 71, pt. A, p. 103094, 2022.

KUMAR, A.; SHARMA, K.; SHARMA, A. Hierarchical deep neural network for mental stress state detection using IoT-based biomarkers. **Pattern Recognition Letters**, Amsterdam, v. 145, p. 81–87, 2021.

MINGXU, L.; ZHANG, Y.; LIU, H.; WANG, Z. Affect and stress detection based on feature fusion of LSTM and CNN using physiological signals. **Sensors**, Basel, v. 23, 2023.

MODI, N.; KUMAR, Y.; MEHTA, K.; CHAPLOT, N. Physiological signal-based mental stress detection using hybrid deep learning models. **Discover Artificial Intelligence**, Cham, v. 5, 2025.

RIM, B.; LEE, J.; PARK, J.; KIM, H. Deep learning applications on medical physiological signals: a comprehensive review. **Sensors**, Basel, v. 20, n. 20, 2020.

SAH, S.; DAS, A.; CHOUDHURY, S. StressNet: convolutional neural network framework for personalized stress classification. **IEEE Access**, New York, v. 10, 2022.

VARGAS-LOPEZ, O.; PEREZ-RAMIREZ, C. A.; VALTIERRA-RODRIGUEZ, M.; YANEZ-BORJAS, J. J.; AMEZQUITA-SANCHEZ, J. P. An explainable machine learning approach based on statistical indexes and SVM for stress detection in automobile drivers using electromyographic signals. **Sensors**, Basel, v. 21, n. 9, 2021.

WU, Y.; ZHANG, J.; LIU, Y.; WANG, J. Transfer learning-based stress detection using physiological signals in daily activities. **IEEE Access**, New York, v. 9, 2021.

XIANG, J.-Z.; WANG, Q.-Y.; FANG, Z.-B.; ESQUIVEL, J. A.; SU, Z.-X. A multimodal deep learning approach for stress detection using physiological signals. **Frontiers in Physiology**, Lausanne, v. 16, 2025.