v.6, n.9, 2025

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 COMBINADA 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 José Rigo[1], Kauã Mark[1], William Frohlich[1]

e696745
https://doi.org/10.47820/recima21.v6i9.6745

RECEIVED: 08/02/2025 APPROVED: 09/02/2025 PUBLISHED: 09/17/2025

ABSTRACT

Accurate stress detection through physiological signals shows strong potential for improving healthcare outcomes, reducing costs, and enabling early intervention in stress-related disorders. This study presents 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 on 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 shows robust learning and reasonable generalization under data-constrained conditions, achieving 60.95% accuracy on an independent test set. The findings reinforce the efficacy of deep learning in stress classification and underscore the necessity 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 utilizando 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 da 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 60,95% 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] Universidade do Vale do Rio dos Sinos – UNISINOS.

RESUMEN

La detección precisa del estrés a través de señales fisiológicas presenta un gran potencial para mejorar los resultados en salud, reducir costos y posibilitar 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 principales métodos, desafíos y tendencias emergentes en el área. Se hace especial énfasis en las limitaciones impuestas por conjuntos de datos reducidos, la importancia de 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 señales de electrocardiograma (ECG) en cuatro categorías distintas. El modelo demostró un aprendizaje robusto y una generalización moderada en condiciones de restricción de datos, alcanzando un 60,95% de exactitud en un conjunto de prueba independiente. Los hallazgos refuerzan la eficacia del aprendizaje profundo en la clasificación del estrés y resaltan la necesidad de enfoques personalizados, en tiempo real y multimodales en investigaciones futuras.

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

1. INTRODUCTION

Early detection of stress is critical for reducing healthcare costs and preventing diseases associated with chronic exposure to stress. Automatic recognition of predictive factors and detection of physiological conditions (Wu *et al.,* 2021) can lead to significant benefits across domains such as education, health, traffic safety, and workplace productivity.

Although stress pattern detection is still a relatively new research topic, promising results have been reported in controlled environments (Nath *et al.,* 2023). However, several challenges persist. Individuals react differently to stress situations, and generalized models often underperform compared to personalized approaches due to inter-individual variability (Finseth *et al.,* 2023). Personalization, which accounts for user-specific information, typically yields better performance than user-independent models (Gashi *et al.,* 2023).

Preprocessing is another crucial factor, as removing noise and irrelevant data can significantly improve algorithm efficiency. Methods such as wavelet transforms and passband filters have been used, but the challenge remains to remove undesirable data without introducing distortion or bias (Chatterjee *et al.,* 2022). Wearable devices have been increasingly adopted for stress monitoring. These systems can provide real-time feedback, while customization can tailor interventions to individual needs. This study reviews the state-of-the-art in stress classification, identifies challenges, and presents an experimental CNN-based ECG classification model designed for data-constrained, noisy, and imbalanced conditions that reflect real-world applications.

2. STATE-OF-THE-ART RESEARCH METHODOLOGY

A non-systematic review was conducted using snowball sampling, starting from key studies and expanding via references. Databases searched included DL-ACM, IEEE, Elsevier, Springer, and Science Direct. The keywords “stress,” “physiological signals,” “classification,” and “customization” were used, focusing on publications since 2021. Twenty-seven papers were retrieved; after screening, seven were deemed relevant.

The review sought to answer the following 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?

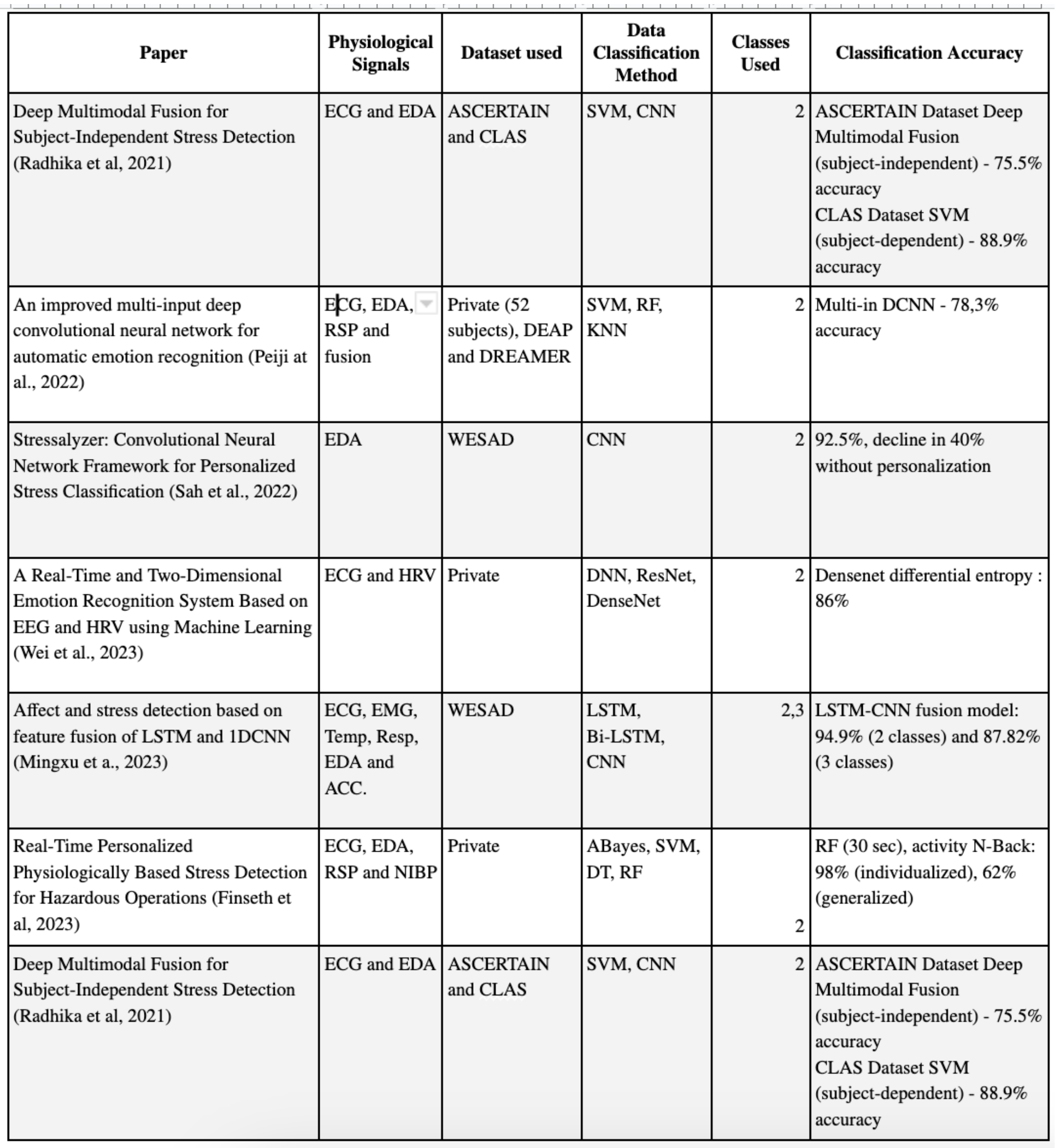
Papers considered not relevant to the research questions were removed, resulting in seven articles. To foster the analysis and support the answering of the research questions, the chosen papers were comparatively analyzed based on the following information:

- **Objective**: Main research goals of 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 its 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 studied papers are designed to use a specific structure, usually a convolutional neural network (CNN), with multi or mono-signals. There are very few papers using multi-signals and even fewer for 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 public datasets related to emotional oscillations such as 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). ASCERTAIN dataset contains big-five personality scales and emotional self-ratings of 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 only collected EEG data of 32 participants as each watched 40 one-minute long excerpts of music videos.

Participants rated each video in terms of the levels of 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 has ECG, plethysmography (PPG), EDA, and accelerometer data of 62 healthy volunteers, which were recorded while involved in three interactive tasks and two perceptive tasks. Finally, DREAMER has EEG and EDA data from 23 participants, along with the participants' self-assessment of their affective state after each stimulus, in terms of valence, arousal, and dominance.

While these datasets have the benefit of making data publicly available, they have the limitation of a very reduced number of participants, focusing more on comparing different classification methods than on the data collection itself. Generally, for the recognition of such changes, a large amount of data must be analyzed. Techniques such as time windows are used to reduce this data, but the challenge lies in leaving the dataset at a sufficient size so that no relevant information is lost.

Qualitative analysis showed that the most commonly used algorithms for this purpose were convolutional neural network (CNN) and support vector machine (SVM). The most commonly used devices for data collection are Empatica, Emotiv, and SHIMMER, with heart rate (HR) and GSR being the least intrusive signals and with the best accuracy for stress detection. Collecting data in real-time in an uncontrolled environment, 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). When compared, it can be seen that LSTM achieved an accuracy of 81 % versus the best-performing ML algorithm among them.

The electrocardiogram (ECG) is one of the most used physiological signals in stress detection systems based on machine learning (ML) algorithms. 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 that have been used include electromyography (EMG), blood volume pulse (BVP), respiration, 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 significantly good performance in both accuracy and computational power. However, studies suggest further studies with new classification methods such as unsupervised learning, artificial neural networks (ANNs), deep learning, and reinforcement learning, so that a comparative analysis between them can be developed.

While the challenges under discussion are still being debated, there are those related to ethical concerns, such as privacy, security, and legislation, as well as the perspective of reliability of results, cost, and interoperability (Hern´andez *et al.,* 2021). Regarding reliability, other sources of emotional changes can also activate the sympathetic nervous system (SNS) and generate similar heart rate (HR) signals, as well as everyday situations can generate 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 have an impact on deep multimodal fusion, and that the generalization capability of subject-independent stress detection models is lower compared to subject-dependent models.

Another study proposed a multi-input deep convolutional neural network (DCNN) that can extract the features of different input signals separately. The filters in different channels are not shared, which solves the problem of interference between channels and achieves 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) presented an impressive 92.5 % accuracy with CNN, declining by 40 % without customization. To deal with the customization challenge, the authors 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 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, where data from one subject is removed from the training set and kept 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 needs customization if the difference is larger than = 5.

Densenet, DNN, and ResNet were compared, with Densenet achieving the highest accuracy. This suggests that a neural network model similar to LSTM can be used to analyze data over a longer period of time to improve classification accuracy (Wei et al., 2022). Mingxu et al. presented promising results comparing Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), and convolutional neural network (CNN) models, with 94.9 % accuracy for 2 classes and 87.82 % for 3 classes. This is because LSTM models pay more attention to the temporal features of physiological signal variables, while CNN models focus more on the correlation of spatial features between physiological signal variables.

Finseth *et al.,* (2023) compared different algorithms from previous papers: ABayes, support vector machine (SVM), decision tree (DT), and random forest (RF). RF for activity N-Back had 98 %

accuracy for the individualized method, and 62 % for the generalized method. The highest 10-fold cross-validation performance for the VR-ISS across all windows and classifiers was 94 % using an ABayes classifier with a window size of 30 seconds, suggesting that the personalized approach performed well. In addition to the excellent results obtained with CNN, a recent study demonstrated its use in conjunction with RNN. 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 had better accuracy than ML, with 99.55 % accuracy compared to 76.50% for LDA and 98.38 % compared to 75.21 % for AdaBoost. When the accelerometer (ACC) data was removed, the consistency of performance was maintained, with 97.48 % accuracy for DL compared to 80.34 % for AdaBoost and 93.64 % for random forest (RF) compared to 76.17 % for RF. However, there is not yet a significant sample size to consistently answer which of these methods is more efficient, due to the parameterization and complexity across architectures.

There are few studies that compare ML and DL, and even fewer that compare DL algorithms to each other. 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 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 to each other. Among them, Artificial Neural Network (ANN), SVM, Stacking Classifier, and Radial Basis Functions Neural Networks (RBF) have been compared, with 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 for stress detection with different signal processing techniques to generate inputs for this architecture (Fourier Transform, cube root, and CQT) has been shown to have 96.6 % accuracy (Gil-Martin *et al.,* 2022). Fatma (2022) analyzed CNN, LSTM, and RNN, presenting a result of 93 %, which was compared with the Autoregressive (AR-HMM) implemented previously 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 suggested to validate the data annotation step, ensuring a better accuracy in the dataset used for the experiments.

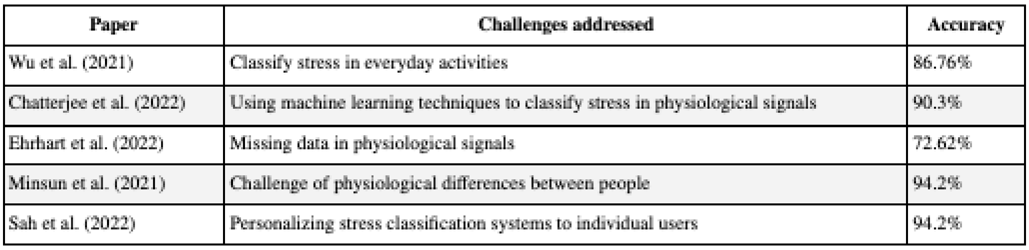
In the next section, some of the specific challenges in this area are addressed.

3. SPECIFIC CHALLENGES IN USING PHYSIOLOGICAL SIGNALS FOR STRESS CLASSIFICATION

Stress classification using physiological signals is a challenging task. There are several challenges that need to be addressed in order to develop accurate and reliable stress detection systems. These challenges that will be focused 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 mentioned known challenges specific to this context, such as use in everyday activities, few public datasets, insufficient amount of existing data and 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, composed of a two-level BLSTM network for feature extraction, with 64 LSTM neurons. 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 Units (ReLu) function, which returns 0 when given a negative value, and returns the same value for any positive value. During training a batch size of 8 is used along with 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 hyper parameters selected and cross-validated 5 times.

This framework achieved a promising result of 86.76 % accuracy when compared to other models. Chatterjee *et al.* (2022) presented an innovative model structured with 30-second windows applying Fast Fourier Transformation (FFT), which converts a signal from its original domain to the frequency representation (or vice versa) for each of the pairs, sorting descendingly by its amplitude.

As a result, only the top 10 features (amplitude, frequency) of each window are used, achieving accuracy of 90.3 % for 2 classes and 94.2 % for binary classification. Ehrhart *et al.* (2022) uses a Conditional Generative Adversarial Network (cGAN) architecture for the missing data problem, combining LSTM with Fully Convolutional Network (FCN) in two classes with 16 second windows. With the help of the auxiliary data generated by cGAN this model achieved 72.62 % accuracy on the test dataset, generating large improvements in terms of recall (+19.05 %) and F1-score I(+11.03%).

The classification of stress in everyday activities has the challenge of differentiating the change in physiological data caused by the stress itself from those that are the effect of other daily activities that cause heart rate elevation, 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 observed using physiological signals for detection is the rigidity of generalized models in accounting for physiological differences between people. Only two studies have gone deep on this. Sah *et al.* [11] used a 1-dimensional CNN architecture composed of two convolutional layers with filters and a kernel size of 5 and respectively. Convolution layers are followed by a global max-pooling layer and neurons. They also have drop-out layers after two each fully connected layer with drop-out values of 0.3. The output layer has Softmax activation, and all other layers have 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. The model is retrained until the performance of the personalized model (M2) on the user data is at an acceptable level. CNN models were trained for epochs with a batch size of and a fixed learning rate of 0.001, and leave-one-subject-out (LOSO) was used to tackle the customization challenge.

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

4. EXPERIMENT: CNN MODEL USING ECG

Cardiovascular conditions often manifest through distinctive patterns in electrocardiogram (ECG) signals, making accurate ECG classification critical for both clinical diagnostics and remote patient monitoring. Recent advances in deep learning—particularly convolutional neural networks (CNNs)—have demonstrated superior performance compared to traditional machine learning (ML) algorithms such as support vector machines (SVM) and random forest (RF) in extracting spatial and temporal dependencies from biomedical time-series data.

However, as observed in the state-of-the-art review (Section 2), the highest-performing models in the literature (e.g., CNN on WESAD achieving 92.5%.

The present work addresses this gap by designing and evaluating a custom CNN architecture for multiclass ECG classification under data-constrained, noisy, and imbalanced conditions, which more closely resemble real-world healthcare environments.

4.1. Model Architecture

The proposed CNN model was designed for both accuracy and robustness in small-data regimes, incorporating four convolutional blocks with progressively increasing filter sizes (32, 64, 128, 256) to enable hierarchical feature abstraction. Each block is followed by batch normalization and LeakyReLU activation, which help stabilize training and prevent vanishing gradients. An adaptive average pooling layer preserves temporal context while reducing feature dimensionality, and the classification stage consists of two fully connected layers with dropout applied in both convolutional and dense components to mitigate overfitting. This architecture integrates elements proven effective in related studies, such as DCNNs for emotion recognition achieving 78.3% accuracy, while explicitly addressing common challenges including class imbalance and signal noise.

4.2. Dataset and Preprocessing

The dataset comprised ECG recordings in CSV format, of which only 10% representing approximately four patients was used for training and evaluation, a choice that mirrors realistic clinical constraints such as privacy regulations, annotation costs, and acquisition limitations. Preprocessing involved converting all entries to numeric values, imputing missing data using column-wise means, applying z-score normalization, employing stratified sampling to preserve class distribution, and using weighted random sampling during training to address class imbalance.

This methodology contrasts with the large-scale, balanced datasets commonly used in the reviewed literature, such as ASCERTAIN, DEAP, and DREAMER.

4.3. Training Configuration

The model was trained using a cross-entropy loss function with class weights to address imbalance, optimized with the Adam algorithm incorporating weight decay for regularization. A ReduceLROnPlateau learning rate scheduler, triggered by validation loss trends, was employed to enhance convergence stability, while early stopping was applied after 10 epochs without improvement to prevent overfitting. Training proceeded for 66 epochs, achieving a minimum validation loss of 0.838, with a stable and gradual reduction in loss that reflects effective feature extraction without overfitting, despite the limited and high-dimensional dataset.

4.4. Results

Evaluation on the independent test set resulted in an overall accuracy of 60.95% The model achieved its highest recall for Class 3 (0.7469) and Class 2 (0.7160), while the highest precision was observed for Class 1 (0.7949), albeit with a notably lower recall of 0.4339, reflecting a conservative prediction strategy with high specificity. The macro-average F1 score was 0.6063, and the weighted-average F1 score was 0.6043, indicating relatively balanced performance across classes. As shown in Figure 4.1, the confusion matrix reveals clear separation for Classes 2 and 3, but substantial overlap between Classes 0 and 1, likely due to morphological similarities in their ECG waveforms, a limitation also documented in prior ECG-based classification studies.

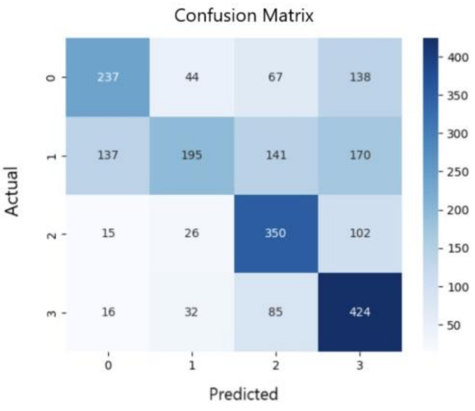


Figura 1. Confusion Matrix

5. CONCLUSION

The convolutional neural network (CNN) model developed for ECG-based classification demonstrates effective learning and reasonable generalization, particularly under data-constrained conditions. Its architecture, designed with an emphasis on feature extraction and regularization, performs well in identifying specific rhythm patterns, though it faces challenges in distinguishing between morphologically similar signals. This limitation is reflective of broader issues in the field of stress classification using physiological signals, which remains in its early stages and must overcome several key obstacles before widespread adoption is feasible. One of the most significant limitations is the limited availability of large, publicly available datasets that accurately capture real-world stress conditions. Existing datasets are often small and limited in scope, restricting the ability to train and validate robust classification models.

Another critical challenge lies in the need for personalized systems that can account for the considerable variability in physiological stress responses across individuals. Models trained on generalized data may not perform consistently for all users, whereas personalized models—tailored to individual physiological profiles—have shown clear promise in improving classification accuracy by adapting to unique stress signatures. Additionally, developing systems capable of real-time operation in uncontrolled environments remains a substantial hurdle. To be practical in applications such as wearable health monitors or in-vehicle stress detection, these systems must function reliably despite dynamic conditions and environmental noise, which can introduce artifacts that degrade performance.

To address these challenges, recent research has explored several novel approaches. Deep learning models, particularly those based on convolutional and recurrent architectures, have proven effective in automatically extracting complex features from physiological signals, uncovering latent patterns that are often missed by traditional methods. Personalized modeling continues to gain traction as an effective strategy, with user-specific models consistently outperforming generalized ones in stress detection. However, real-time classification in uncontrolled environments remains an open problem, demanding solutions that balance computational efficiency with robustness to noise.

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/Bi-LSTM) 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 dataset size, and leveraging transfer learning from large ECG repositories. Collectively, these refinements have the potential to bridge the performance divide between generalized and personalized models, paving the way for accurate, adaptable, and deployable stress classification systems in real-world healthcare applications.

REFERENCES

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. DOI: 10.1109/ICIRCA48905.2020.9183244.

CAN, Y. S.; CHALABIANLOO, N.; EKIZ, D.; FERNANDEZ-ALVAREZ, J.; RIVA, G.; ERSOY, C. Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches. **IEEE Access**, v. 8, p. 38146–38163, 2020.

DELMASTRO, F.; MARTINO, F. D.; DOLCIOTTI, C. Cognitive training and stress detection in MCI frail older people through wearable sensors and machine learning. **IEEE Access**, v. 8, p. 65573–65590, 2020.

DHAM, V. et al. Mental stress detection using artificial intelligence models. **Journal of Physics: Conference Series**, v. 1950, 012047, 2021.

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. DOI: 10.1145/3397482.3450732.

GEDAM, S.; PAUL, S. A review on mental stress detection using wearable sensors and machine learning techniques. **IEEE Access**, v. 9, p. 84045–84066, 2021. DOI: 10.1109/ACCESS.2021.3085502.

GJORESKI, M.; LUŠTREK, M.; GAMS, M.; GJORESKI, H. Monitoring stress with a wrist device using context. **Journal of Biomedical Informatics**, v. 73, p. 159–170, 2017. DOI: 10.1016/j.jbi.2017.08.006.

HUANG, J.; LIU, Y.; PENG, X. Recognition of driver's mental workload based on physiological signals: a comparative study. **Biomedical Signal Processing and Control**, v. 71, parte A, p. 103094, 2022. DOI: 10.1016/j.bspc.2021.103094.

KUMAR, A.; SHARMA, K.; SHARMA, A. Hierarchical deep neural network for mental stress state detection using IoT based biomarkers. **Pattern Recognition Letters**, v. 145, p. 81–87, 2021. DOI: 10.1016/j.patrec.2021.01.030.

NATH, R. K.; THAPLIYAL, H.; CABAN-HOLT, A. Machine learning based stress monitoring in older adults using wearable sensors and cortisol as stress biomarker. **Journal of Signal Processing Systems,** 2021. DOI: 10.1007/s11265-020-01611-5.

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,** v. 21, n. 3155, 2021. DOI: 10.3390/s21093155.

ZUBAIR, M.; YOON, C. Multilevel mental stress detection using ultrashort pulse rate variability series. **Biomedical Signal Processing and Control**, v. 57, 2020. Art. 101736.