Journal of Ambient Intelligence and Humanized Computing (2023) 14:11011–11021
https://doi.org/10.1007/s12652-022-04365-z

ORIGINAL RESEARCH

Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate

Mariano Albaladejo-González[1] · José A. Ruipérez-Valiente[1] · Félix Gómez Mármol[1]

Received: 4 March 2022 / Accepted: 28 July 2022 / Published online: 27 August 2022

Abstract

In the twentyfirst-century society, several soft skills are fundamental, such as stress management, which is considered one of the key ones due to its strong relationship with health and well-being. However, this skill is hard to measure and master without external support. This paper tackles stress detection through artificial intelligence (AI) models and heart rate, analyzing in WESAD and SWELL-KW datasets five supervised models and five unsupervised anomaly detection models that had not been tested before for stress detection. Also, we analyzed the transfer learning capabilities of the AI models since it is an open issue in the stress detection field. The models with the highest performance on test data were the anomaly detection Local Outlier Factor (LOF) with F1-scores of 88.89% in WESAD and 77.17% in SWELL-KW, and the supervised Multilayer Perceptron (MLP) with F1-scores of 99.03% in WESAD and 82.75% in SWELL-KW. However, when evaluating the transfer learning capabilities of these AI models, MLP performed much worse on the other dataset, decreasing the F1-score to 28.41% in SWELL-KW and 57.28% in WESAD. In contrast, LOF reported better transfer learning performance achieving F1-scores of 70.66% in SWELL-KW and 85.00% in WESAD. Finally, we found that training AI models with both datasets (i.e., with data from different contexts) improved the average performance of the models and their generalization; with this setup, LOF achieved F1-scores of 87.92% and 85.51% in WESAD, and 78.03% and 82.16% in SWELL-KW; whereas MLP obtained 78.36% and 81.33% in WESAD, and 79.37% and 80.68% in SWELL-KW. Therefore, we suggest as a promising direction the use of anomaly detection models or multi-contextual training to improve the transfer learning capabilities in this field, which is a novelty in the literature. We believe that these AI models combined with the use of non-invasive wearables can enable a new generation of stress management mobile applications.

Keywords Stress · Biometrics · Artificial intelligence · Machine learning · Transfer learning · Affective computing

1 Introduction

The abilities needed by the modern workforce are changing, soft skills such as stress management, communication, leadership, and critical thinking are considered essential for professional developments (Vasanthakumari 2019). In this paper, we focus on stress management, one of the key soft skills due to its relationship with health and well-being (Greene et al 2016). Besides, stress impairs working memory and cognitive flexibility (Shields et al 2016) affecting the students' and workers' performance. However, stress is hard to measure, and subjective reports through validated questionnaires need direct feedback from the user by indicating their stress levels over time, which is not convenient due to self-biases and the invested time in sustained use.

One growing solution to this problem is affective computing, which aims to develop machine systems that can recognize emotions, including stress. A common approach for automatic stress detection is affective computing with biometrics data (Mohammadi et al 2022; Motogna et al 2021) because some biometric data are linked to stress, mainly

✉ Mariano Albaladejo-González
mariano.albaladejog@um.es

José A. Ruipérez-Valiente
jruiperez@um.es

Félix Gómez Mármol
felixgm@um.es

[1] Department of Information and Communications Engineering, University of Murcia, Calle Campus Universitario, Murcia 30100, Murcia, Spain

heart and heart rate variability (Szakonyi et al 2021). Biometrics data can be collected via sensors; however, some are more invasive than others, and thus not all are viable for use in real applications. Another challenge is that individual differences increase the difficulty of using biometric data (Hu et al 2019). Due to these individual differences, AI models should be evaluated with a subset of subjects not utilized in training (Wu et al 2021). In addition, another challenge is being able to develop models that can be used in different environments, situations, and stressors; thus, a transfer learning evaluation is highly recommended for real applications.

These artificial intelligence (AI) models could measure stress in different environments such as workplaces (Khowaja et al 2020), driving (Kerautret et al 2022), education (Celdrán et al 2020), or emergencies (Pluntke et al 2019). In these environments, AI models could support the improvement of self-regulation of stress, and high stress can be reported to human resources staff, managers, or teachers to make the necessary changes in the environment and workload. Besides, the lessons learned could be applied to other soft skills where AI models can help and other affective computing applications.

In this work, we have assessed the viability of using biometrics to perform stress detection through AI models. We implemented five supervised learning models, the common approach in the state-of-the-art (Mohammadi et al 2022; Szakonyi et al 2021), and five unsupervised anomaly detection models that have not been employed before for stress detection. We also explored transfer learning in stress detection to see if the AI models applied in one environment could be effectively re-used in a different one. Transfer learning has been evaluated in two ways: training the models on one dataset and evaluating them on another dataset and training the models on two datasets, evaluating them on test data from both datasets. Wu et al (2021) are among the few authors who have also analyzed stress detection in multiple contexts, but they applied a different approach to this paper. Therefore, this study goes beyond the state-of-the-art by implementing new stress detection AI models and evaluating their transfer learning capabilities which are essential for real applications.

We establish the following research questions (RQs) for this research:

RQ1 What are the best supervised models, unsupervised anomaly detection models, and configurations for stress detection through heart rate and heart rate variability?

RQ2 What is the transfer learning performance in the context of stress detection? We split this RQ into two sub-RQs.

- RQ2.1 What is the performance of a model configured in one context when applied in another?
- RQ2.2 What is the performance of a model configured with data from multiple contexts?

We selected the above RQs due to their significance in the field of affective computing and stress detection. RQ1 explores the configuration and evaluation of unsupervised anomaly detection models to detect stress for the first time in the state-of-the-art. RQ2.1 consider possible limitations of transfer learning in stress detection. Poor transfer learning in supervised learning models implies that most AI models proposed in the state-of-the-art for stress detection cannot be used in other contexts; thus, they should not be applied in real applications where the context is not the same as the training context. Moreover, this problem could be extended to other applications in the field of affective computing. Finally, RQ2.2 tests the potential of configuring AI models in multiple datasets to improve their transfer learning capabilities which is uncommon in affective computing.

The rest of the paper is organized as follows. In Sect. 2 we provide a background of AI models and biometric data to measure and develop key capabilities. In Sect. 3 we present the followed methodology, and in Sect. 4 we show the results obtained. In Sect. 5 we discuss the outcomes, implication, and limitations. Finally, we present the research conclusions and future work in Sect. 6.

2 Related works

Due to the importance of soft skills in the 21st-century society, there is an interest in measuring and developing soft skills using AI, for example, to measure teamwork skills (Chopade et al 2019), presentation skills (Ochoa and Dominguez 2020), or stress management (Lin et al 2020). Therefore, AI could help manage and improve these capabilities if it is implemented as part of applications with that purpose. For example, to enhance stress management for cybersecurity professionals (Albaladejo-González et al 2021).

Depending on the user's recorded data, there are two principal approaches to measure these capabilities. One approach focuses on non-biometric data, especially user telemetry (Sikander and Anwar 2019; Sağbaş et al 2020), and computer vision (Ramos-Giraldo et al 2020). The other approach consists of measuring biometric data such as an electrocardiogram (ECG), electromyogram (EMG), electroencephalogram (EEG), electrodermal skin activity (EDA), skin temperature, and blood pressure (Arya et al 2021). In this paper, we focus on stress detection through biometric data, which belongs to the field of affective computing.

For stress detection, the use of heart rate and heart rate variability is widespread because they are linked to stress (Szakonyi et al 2021). Both can be extracted from a photoplethysmography (PPG) sensor, which is available in most smartwatches and wristbands. However, other biometric data cannot be measured without expensive and invasive research-oriented devices. For this reason, this article focuses only on heart rate measurement.

Nevertheless, stress does not necessarily manifest strictly as an increase in the heart rate, but it does show up in features extracted through its processing in time windows (Khowaja et al 2020). These features summarize the heart rate and the heart rate variability in time windows, and the typical approach is to introduce these features in supervised AI models to classify them as stress or non-stress. Additionally, heart rate features can be combined with features from other physiological signals before being analyzed by the supervised AI model (Pourmohammadi and Maleki 2020; Khowaja et al 2020; Szakonyi et al 2021). Supervised AI models require training with labeled data from all categories; therefore, they need stress and non-stress windows during the training. In addition to these models, we tested unsupervised anomaly detection AI models, which classify the inputs based on whether they belong to the distribution of previous observations. We introduced baseline windows (windows recorded in conditions without stressors) into these models, and they classified new inputs as belonging to the baseline or not. We considered stress windows or not if they belonged to the baseline or not, respectively. There are some applications of unsupervised models in the affective computing field, such as Carbonell et al (2021). However, we have not found similar applications for stress detection using biometric signals.

Transfer learning is essential for real affective computing applications, including stress detection via biometric data. Usually, researchers focus on the performance on a subset of subjects not utilized in training (Wu et al 2021), that are reserved for testing the quality of the model within the same context. For example, Mozafari et al (2021) employed a leave-one-subject-out (LOSO) cross-validation. In contrast to normal cross-validation, in LOSO cross-validation one subject never has time windows in training and validation sets simultaneously. One approach to improve the model's performance on new subjects is including a normalization to eliminate the individual differences between subjects (Pourmohammadi and Maleki 2020; Zontone et al 2019). The paper at hand assesses the application of a feature normalization to the range [0, 1] based on the subject baseline, which is a subset of data from the analyzed subject recorded without stressors. In contrast to other authors, we obtained the results with and without the normalization step for each AI model to analyze its effect and select the best option. Another approach is the application of transfer learning techniques to reduce these individual differences as Mozafari et al (2021) proposed; however, this approach is less common in the literature.

Few authors have gone a step further and trained AI models with data from different datasets to improve the model's generalization and performance. An example is Wu et al (2021), who enhanced performance on a small dataset using data from a different source. In our study, we also have configured models with data from multiple datasets (RQ2.2). Besides, we have assessed the transfer learning capabilities by evaluating the AI models on a different dataset where it was trained (RQ2.1). Therefore, this paper contains a more realistic evaluation of the AI models' performance for real applications. In addition, this is the first study, as far as we know, that utilizes unsupervised anomaly detection models for stress detection.

3 Methodology

This section presents the methodology followed to answer RQs. We divided this section into the following two subsections: Sect. 3.1 describes the dataset search and Sect. 3.2 contains the pre-processing, feature extraction, and the different AI models configurations and evaluations.

3.1 Public datasets suitable for stress detection

3.1.1 Search criteria

This search aimed to find public datasets suitable to evaluate different models of stress detection. Table 1 contains the terms searched in each source, and the obtained datasets were filtered based on the following criteria:

- The dataset was obtained from a case study that collected the ECG or PPG measurements under stress and non-stress conditions to obtain the heart rate and heart rate variability.
- The dataset reported the stress and non-stress measurements of the different subjects separately to evaluate the AI models' performance in different subjects.
- The dataset contained data from at least 15 subjects to reserve at least five subjects for the test subset (one-third).
- The dataset included stress questionnaires to verify that the subjects were really stressed and non-stressed in each phase.
- The dataset was published from 2010 onwards. This criterion excluded data from obsolete measurement technologies.

Table 1 Explored terms in the search of public datasets

Source	Terms
Referenced articles	"database", "repository", and "dataset"
Google	"stress dataset", "stress data", "stress database", "stress repository", "stress detection", "stress recognition", "emotion detection", "emotion data", "emotion dataset", "emotion database", "emotion repository", "emotion detection", and "emotion recognition"
Kaggle, UCI Machine Learning Repository, Amazon Datasets, and Google's dataset search engine	"stress" and "emotion"

3.1.2 Identified datasets

This search found the following two datasets: Multimodal Dataset for Wearable Stress and Affect Detection (WESAD) (Schmidt et al 2018) and Smart Reasoning Systems for Well-being at Work and at Home-Knowledge Work (SWELL-KW) (Koldijk et al 2014).

WESAD contains different physiological measurements from 15 subjects in different conditions. The measurements were recorded by an Empatica E4 wristband and a RespiBAN placed on the chest. The Empatica E4 measured the blood volume pulse (BVP), EDA, the skin temperature, and data collected by an accelerometer. BVP signal is the measurement of a PPG sensor. The Empatica E4 extracted from the BVP signal the inter-beat distance and heart rate, and the RespiBAN measured the ECG signal, EDA, EMG, respiration, temperature, and data collected by an accelerometer. We extracted the heart rate and heart rate variability from the BVP signal, the only signal employed from this dataset in this methodology. The authors recorded these physiological signals in four conditions. The non-stress condition, named by authors as baseline, consisted of reading magazines, and the stress condition consisted of public speaking and a mental arithmetic task. They furthermore included an amusement condition, meditation, and a rest break that our experiment did not use. The experiment was validated by collecting the following questionnaires at the end of each phase: Positive and Negative Affect Schedule, State-Trait Anxiety Inventory, Self-Assessment Manikins, and a Short Stress State Questionnaire (Schmidt et al 2018). The authors developed the study with 17 subjects, but they discarded the measurement of two subjects due to a sensor malfunction. Therefore, WESAD only contains 15 subjects.

In turn, SWELL-KW dataset was collected from an experiment executed on 25 subjects in different working conditions that consisted of writing different reports and making presentations about different topics. SWELL-KW has two stress conditions: time pressure and interruptions. In the time pressure condition, the subjects had two thirds of the time they needed in neutral conditions, and in the interruption conditions, the subjects received eight emails interrupting their tasks. The dataset contains the subjects' keyboard, applications, mouse telemetry, posture, faces, ECG, and EDA during the experiment. In this methodology, we extracted the heart rate and the heart rate variability from the ECG signal. In addition to time stress and interruption stress phases, SWELL-KW included a non-stress condition named neutral phase where the subjects were working without a deadline. Furthermore, the experiment included before each phase rest breaks and at the end of each phase, they reported the questionnaires: Rating Scale Mental Effort, NASA Task Load Index, Self-Assessment Manikin, and a Visual Analogue Scale to rate stress with values from 1 to 10 (Koldijk et al 2014). We discarded the subjects with the IDs #11 and #23 because the dataset repository commented that the measurements of subjects with the IDs #11 and #23 were not correctly recorded. In addition, we eliminated subject #7 due to poor signal quality and subject #8 because its neutral phase measurements were too short. Hence the final data collection employed included 21 users.

3.2 Evaluation of the AI models to detect stress

This subsection introduces the methodology applied in the configuration and the evaluations of supervised and unsupervised anomaly detection models (Fig. 1). We selected the following unsupervised anomaly detection models available in the well-known library *scikit-learn* (Pedregosa et al 2011): Elliptic Envelope, Isolation Forest, Local Outlier Factor (LOF), and Support Vector Machine (SVM) One Class. We also included an autoencoder from the Python toolkit *PyOD* (Zhao et al 2019) to evaluate an unsupervised neural network. The peculiarity of the anomaly detection models is that they receive reference data to classify new data as belonging or not to the reference data; therefore, we introduced data recorded in conditions without stressors as reference data. Among the supervised models, we wanted to cover a distance-based model, a neural network, a bagging model, a boosting model, and a classical linear model. Therefore, we employed the models K-Nearest Neighbor (KNN), Multi-layer Perceptron (MLP), Random Forest (RF), Adaboost, and Linear Discriminant Analysis (LDA). Again, the implementations utilized for these models were from the *scikit-learn* library. The performance metric employed in all

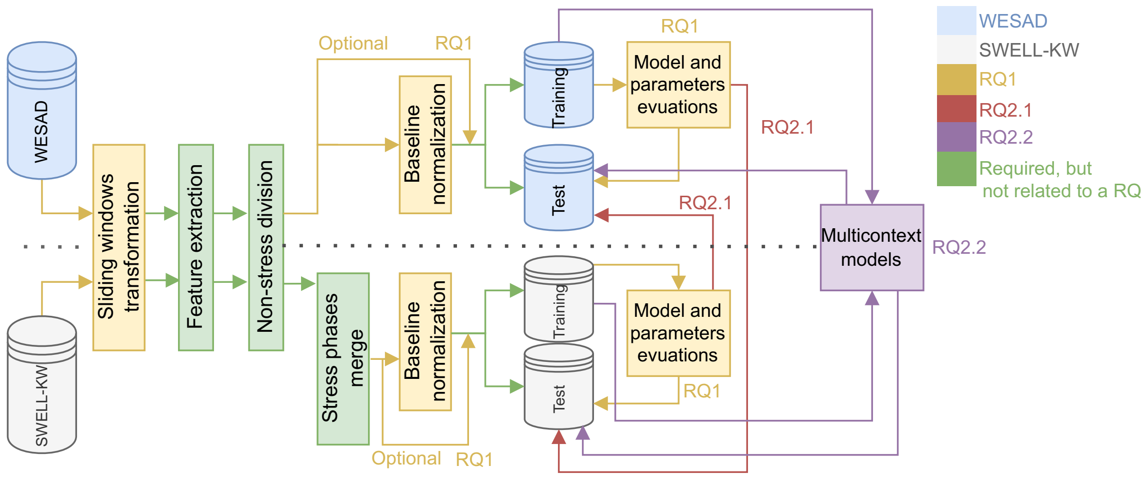

Fig. 1 Methodology applied in the AI models configurations and evaluations

the configurations and evaluations was the macro averaged version of the F1-score (Opitz and Burst 2019), giving equal weight to each class (stress and non-stress) in order to solve the imbalanced data problem.

3.2.1 Data pre-processing and feature extraction

The datasets required a previous transformation and processing to prepare them for the evaluation focused on our RQs. The first step of the transformation was to segment the ECG (SWELL-KW) and PPG (WESAD) recordings of each phase into sliding windows to create more windows from the signals. Sliding windows have a signal portion in common with the previous window and a previously unused signal portion (lan Chen et al 2017; Schmidt et al 2018; Zontone et al 2019). After the feature extraction (next step), we converted each window into one row in the dataset. Therefore, if we did not apply this process, we would only obtain one row for each phase and subject, and also, the AI models would predict the stress with too low a frequency. The initial size of the windows was 5 min, following the recommendation of the library used in the feature extraction. The overlapping between windows was 4 min and 40 s; thus, the shift was of 20 s. These values were the initial ones for the model and normalization evaluation, but later on, we also evaluated the impact of the sliding window and shift sizes for RQ1.

After that, we employed the *Neurokit 2* (Makowski et al 2021) library to transform the windows into 52 features to be used by the AI models (feature extraction). More information about these features is available in the *Neurokit 2* documentation (Makowski et al 2021). However, the features extracted from the spectral power density in the very low (VLF) and ultra-low frequency (ULF) domains could not be calculated because they required windows too long for our datasets. Therefore, we finally extracted 50 heart rate variability features from the time, nonlinear, and frequency domains. There are more libraries for the heart rate and heart rate variability feature extraction, such as *HeartPy* and *hrv-analysis*, but we selected *Neurokit 2* because it generates the most elaborated set of features.

Then, we split each subject's windows measured in the non-stress phase. The first 50% of these windows were considered the baseline used in the normalization and the training of unsupervised anomaly detection models; the rest were considered non-stress evaluation data. Furthermore, in SWELL-KW, we had instances from two stress conditions; thus, we merged time pressure and interruptions phases into stress evaluation data.

The next step was to apply a [0, 1] normalization based on each subject's baseline. This step was optional, the goal was to assess the effect of this normalization on the model's performance for RQ1.

Finally, the datasets were split into training and test as a regular machine learning evaluation. In affective computing, a more realistic methodology is performed by dividing the data based on the subject. Therefore, the test users are different from the training ones, allowing us to evaluate the models with new subjects. The subjects of both datasets were randomly split into two-thirds for training and one-third for test, this division is shown in Table 2.

3.2.2 RQ1. Best performing AI models and configurations

The next steps were focused on identifying the best AI models and configurations for stress detection through heart rate and heart rate variability.

Model and normalization evaluation First, grid searches utilized the training subjects for model hyperparameter optimization. These grid searches employed a LOSO cross-validation with the F1-score metric of the

Table 2 Training and test dataset split

Dataset	Subjects IDs
WESAD training	#1, #4,#6, #10, #14, #15,#16, #17, #18, #19, #21, #22,#24, and #25
WESAD test	#2, #3, #5, #9, #12, #13, and #20
SWELL-KW training	#2, #4, #5, #7, #9, #10, #11, #13, #15, and #17
SWELL-KW test	#3, #6, #8, #14, and #16

Table 3 Results of the model and normalization evaluation (RQ1)

Model	Normalized dataset	F1-score WESAD test	F1-score SWELL-KW test
Autoencoder	Yes	55.44%	68.76%
Autoencoder	No	63.16%	57.06%
Elliptic Envelope	Yes	34.11%	50.71%
Elliptic Envelope	No	31.10%	49.03%
Isolation Forest	Yes	44.13%	63.48%
Isolation Forest	No	44.60%	63.48%
LOF	Yes	**78.77%**	**68.85%**
LOF	No	57.54%	65.97%
SVM One Class	Yes	32.99%	59.33%
SVM One Class	No	31.10%	55.84%
AdaBoost	Yes	81.67%	75.77%
AdaBoost	No	83.25%	48.10%
KNN	Yes	72.00%	66.57%
KNN	No	64.33%	44.99%
LDA	Yes	74.57%	54.94%
LDA	No	**94.64%**	50.59%
MLP	Yes	**86.56%**	**77.67%**
MLP	No	56.15%	42.08%
RF	Yes	78.64%	**80.72%**
RF	No	82.98%	41.61%

average of the LOSO results as the optimization objective for the supervised AI models. After that, the supervised models with the best configurations were trained with the training data and evaluated with the test data. In contrast, the unsupervised anomaly detection models were only trained with each subject's baseline for the test evaluation. The metric calculated for this evaluation was the F1-score averaged across the subjects. Besides, the evaluation was repeated ten times with different random seeds, and we calculated the average of the runs to reduce the randomness.

Window and shift evaluation After that, the top-performing unsupervised anomaly detection and supervised models were selected to continue with the window and shift sizes evaluation. We evaluated the window sizes of 120, 210, 300, and 390 s as well as the 10, 20, and 30 s shift sizes. Very large or very small window sizes reduce the performance of the AI models, and the shift between windows should not be too large for these window sizes either. After repeating the evaluation ten times, we reported the average F1-score of each window and shift from each model. We included a grid search for the model hyperparameter optimization for each window and shift size.

Dimensionality reduction evaluation Next, using the best window and shift sizes for each model and dataset, we evaluated dimensionality reduction with principal component analysis (PCA) and Tree-based feature selection (Tree FS) to test a feature extraction and feature selection method. Again we used the implementation of the library *scikit-learn*. We selected the number of principal components that at least explained 95% of the variance, which is a common criterion. In Tree FS, the threshold value was the mean of the feature importances that is the default value on *scikit-learn*. We also repeated the grid searches to optimize the models' hyperparameters for each dimensionality reduction method.

3.2.3 RQ2. Transfer learning

RQ2.1 Evaluation in another dataset Besides, we computed the performance of the best models and configurations obtained in RQ1 with the test subjects of the other dataset (which were not used as part of the training) to evaluate the transfer learning capabilities in a different context.

RQ2.2. Evaluation of the AI models configured in multiple contexts Finally, we also trained the best-performing models with the training data of both datasets simultaneously and tested those models with the test subjects of both datasets. We trained the models with the normalization, windows and shift size, and the dimensionality reduction selected in WESAD and SWELL-KW throughout the previous evaluations. We also included grid searches with the same methodology as before.

4 Results

4.1 RQ1. Best performing AI models and configurations

Model and normalization evaluation Table 3 shows the results obtained in this evaluation, emphasizing in bold the higher performance on each dataset by the best unsupervised and supervised models and the selected ones. We notice in SWELL-KW that the normalization equaled or improved performance in all models. However, in WESAD, we did not observe a clear tendency. Among the unsupervised anomaly detection models, LOF obtained the best performance in both datasets (F1-score of 78.77% in WESAD and 68.85% in SWELL-KW) and was, therefore, the selected model of this type. The performances of the unsupervised anomaly detection models were considerably lower, especially in SWELL-KW. The best performing model in WESAD was LDA with the non-normalized dataset (F1-score of 94.64%), and in SWELL-KW was RF (F1-score of 80.72%) also with the normalized dataset. Nevertheless, LDA did not perform adequately in SWELL-KW (54.94% of F1-score applying the normalization and 50.59% of F1-score without it). Therefore, the MLP model with the normalized dataset was selected among the supervised models as it depicted robustness. It was the second-best model in both datasets and obtained a higher average performance (86.56%, and 77.67%) than LDA (94.64%, and 50.59%) and RF (78.64%, and 80.72%). Also, KNN, AdaBoost, and RF obtained adequate performances in both datasets.

Window and shift evaluation Fig. 2 shows the results of the window and shift evaluation. We observe that for LOF, increasing the window size requires increasing the shift size (the novelty in the window), and decreasing the window size requires reducing the shift. However, we did not detect any clear tendency in the case of the MLP model. For LOF in WESAD, the best window size was 210 s with a 20 s shift (F1-score of 88.01%), and in SWELL-KW was 300 s with a 30 s shift (F1-score of 76.63%). The best window and shift sizes for the MLP in WESAD were 390 s and 10 s (F1-score of 93.95%), and in SWELL-KW were 210 s and 30 s (F1-score of 82.75%).

Dimensionality reduction evaluation The results of the dimensionality reduction evaluation are shown in Table 4, highlighting in bold the best performance on each dataset and model. In the case of WESAD, both LOF and MLP obtained higher F1-score applying Tree FS (F1-score of 88.89% and 99.03% respectively). In SWELL, LOF improved applying Tree FS (achieving 77.17% of F1-score); however, MLP's performance decreased (reducing the

Table 4 Results of the dimensionality reduction evaluation (RQ1)

Model	Selected features	F1-score WESAD test	F1-score SWELL-KW test
LOF	ALL	88.01%	76.63%
LOF	PCA	61.26%	70.46%
LOF	Tree FS	**88.89%**	**77.17%**
MLP	ALL	93.95%	**82.75%**
MLP	PCA	51.11%	80.68%
MLP	Tree FS	**99.03%**	80.57%

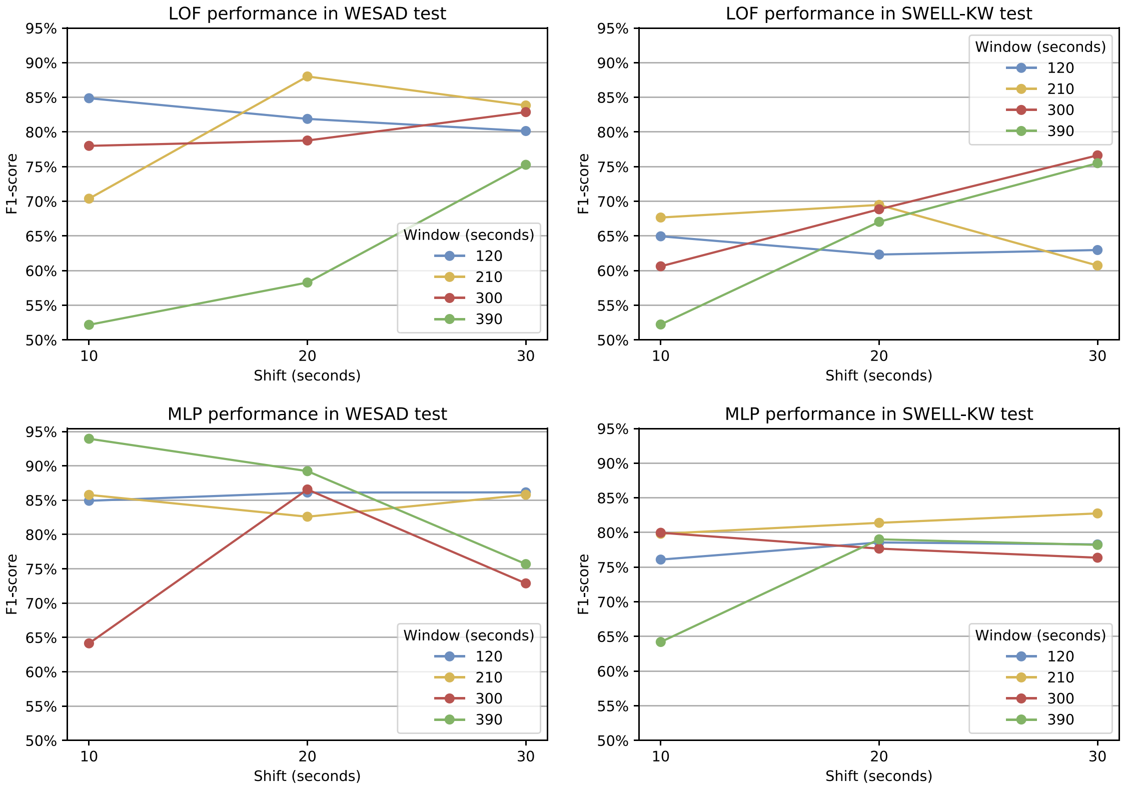

Fig. 2 Results of the window and shift evaluation (RQ1)

F1-score to 80.57%) utilizing Tree FS. In contrast, applying PCA reduced the performance in both models and datasets.

4.2 RQ2. Transfer learning

Evaluation in another dataset Table 5 shows the evaluation of the models trained with data from one dataset and evaluated on the test data from the other dataset to answer RQ2.1, emphasizing the higher performance on each test dataset and metric. The MLP decreased its performance dramatically, from 99.03% to 28.41% and from 82.75% to 57.28% in F1-score. The model trained with WESAD reported in SWELL-KW an extremely low F1-score of 28.41%, clearly showing that the model did not function correctly in a different context. In contrast, LOF reported better transfer learning capabilities between datasets because, in both experiments, LOF reported an F1-score higher than 70%. Surprisingly, the LOF model configured with SWELL-KW increased its performance from 77.17% to 85.00% of F1-score in the transfer to WESAD.

Evaluation of the AI models configured in multiple contexts Finally, Table 6 reports the performance of LOF and MLP when built using the training data from both datasets to answer RQ2.2, highlighting in bold the best performance on each test dataset and metric. We implemented two multi-dataset LOF models and two multi-dataset MLP models, one with the parameters (normalization, window and shift sizes, and dimensionality reduction) selected from WESAD and the other from SWELL-KW, in order to have a fairer comparison. The multi-dataset LOF models obtained better performance in SWELL-KW and higher average performance. In addition, the multi-dataset MLP models obtained a higher average performance than the previous single-dataset MLP models.

5 Discussion

5.1 Model configurations and performance

The experiments performed as part of RQ1 have raised several interesting findings. For example, in the case of the LOF model, we observed that increasing the window size required increasing the shift size (the novelty in the window) and decreasing the window size required reducing the shift. This phenomenon might be due to an existing proportion between the window and shift sizes where larger windows require longer shifts. However, this might be context- or model-dependent, not generalizing to other scenarios. Besides, we observed that normalization improves the performance of all the tested AI models in SWELL-KW and six of the ten models tested in WESAD. Pourmohammadi and Maleki (2020) also obtained the

Table 5 F1-score of the single-dataset models in the transfer learning evaluation (RQ2.1)

Model	Training/ configuration	Test dataset	F1-score	Accuracy	Precision	Recall
LOF	WESAD	WESAD	88.89%	89.18%	83.20%	**100.00%**
LOF	WESAD	SWELL-KW	70.66%	80.20%	88.45%	88.00%
LOF	SWELL-KW	WESAD	85.00%	85.53%	84.30%	90.64%
LOF	SWELL-KW	SWELL-KW	77.17%	87.03%	87.40%	**97.81%**
MLP	WESAD	WESAD	**99.03%**	**99.04%**	**98.29%**	99.85%
MLP	WESAD	SWELL-KW	28.41%	31.39%	32.86%	11.26%
MLP	SWELL-KW	WESAD	57.28%	63.76%	41.27%	53.63%
MLP	SWELL-KW	SWELL-KW	**82.75%**	**88.64%**	**93.01%**	92.68%

Table 6 F1-score of the multiple-dataset models in the transfer learning evaluation (RQ2.2)

Model	Parameters	Test dataset	F1-score	Accuracy	Precision	Recall
LOF	WESAD	WESAD	**87.92%**	**88.29%**	**82.84%**	**100.00%**
LOF	WESAD	SWELL-KW	78.03%	83.91%	85.89%	93.58%
LOF	SWELL-KW	WESAD	85.51%	86.14%	80.95%	**100.00%**
LOF	SWELL-KW	SWELL-KW	**82.16%**	**86.50%**	**87.68%**	**95.09%**
MLP	WESAD	WESAD	78.36%	80.06%	70.23%	93.00%
MLP	WESAD	SWELL-KW	79.37%	83.05%	82.89%	90.96%
MLP	SWELL-KW	WESAD	81.33%	82.38%	76.62%	93.04%
MLP	SWELL-KW	SWELL-KW	80.68%	83.91%	86.45%	89.36%

highest performance applying a Median Absolute Deviation normalization, and Zontone et al (2019) applied a normalization to the range [0, 1] (same that we applied) but did not compare the performance of the AI models without the normalization. Therefore, we consider that applying a normalization may improve the stress prediction through AI models in most cases; however, the conclusion is not indisputable and there may be exceptions.

Another finding was that PCA dimensionality reduction reduced the performance of LOF and MLP in both datasets. In contrast, Tree FS improved the performance of LOF in both datasets and the MLP performance in WESAD. PCA is a feature extraction algorithm that combines different features, and Tree FS is a feature selection algorithm that chooses the best features. The reduction of the performance observed in PCA suggests that PCA does not work well because the features are too different to group them in common dimensions, and Tree FS works better because some features are not very important and add noise. Sriramprakash et al (2017); Pourmohammadi and Maleki (2020) also reported higher performance applying feature selection, but they did not test PCA.

Lastly, in RQ1, LOF and MLP configured with WESAD obtained in WESAD test an F1-score of 88.89% and 99.03%, respectively. In the original publication of WESAD (Schmidt et al 2018), the authors obtained using only BVP signal an F1-score of 84.18%, our LOF and MLP overcome this performance, but the methodology followed was not the same. In contrast, LOF and MLP configured in SWELL-KW test obtained an accuracy of 87.03% and 88.64%, respectively. Sriramprakash et al (2017) achieved a 92.75% of accuracy with an SVM, but again it is not a fair comparison because the methodology applied was different. These results prove that the stress detection in new subjects from the same context is feasible, as other authors have achieved, but with less extensive evaluations than ours (Pourmohammadi and Maleki 2020; Sriramprakash et al 2017; Schmidt et al 2018).

In contrast, in RQ2.1, we have observed that transfer learning for stress detection between contexts has failed in the case of MLP, the best-supervised model. The different sensor types, the measurement environments, or the stressors might be the cause. Interestingly, the multi-dataset MLP models trained with both datasets for RQ2.2 works well in both of them. However, it might not be feasible to train a model with datasets covering all possible stress contexts. Therefore, the efforts should target the development of models that can generalize well across contexts. For example, LOF has better transfer learning capabilities between datasets (RQ2.1). Therefore, the best-unsupervised model reported better transfer learning than the best-supervised model. Thus, we suggest using unsupervised over supervised models for real applications of stress detection that are expected to generalize across contexts, in contrast to previous work, which mostly employed supervised models. However, we emphasize that conducting additional experiments and tests with the unsupervised models is essential because only one of the five models reported this high performance

5.2 Implications

In real applications, the user baseline should be measured to apply the normalization, and if we use unsupervised anomaly detection models, to configure the AI model with these data as a reference. The baseline should be measured in non-stress conditions such as reading magazines, or alternatively, if the user wears a heart rate sensor all day, it is easier to establish their baseline heart rate levels. This process is similar to a calibration process when unsupervised models learn the baseline measurements of the subject. The process should be done at least once for each subject. Nevertheless, it might be wise to re-do this process after a while because the biometrics of the subject might change over time. Also, it could be interesting to get a large baseline data sample in different non-stress activities. This stress detector could be employed in workplaces or educational environments to alert the user of extended stress levels over time that could affect their health. In a smartwatch, the screen could display recommendations to make the user aware of the current situation nudging the user to improve its habits. Also, the stress could be reported to human resources staff, managers, or teachers to make changes to reduce the stress of their employees and students. Besides, the users could report their perceived stress levels through validated questionnaires such as State-Trait Anxiety Inventory or through direct input to the application. This stress feedback from users could be used to evaluate the model's performance or for a reinforcement learning approach. Finally, we believe that these kinds of stress management applications have a high potential to decrease the stress levels of their users; however, the ethical implications have to be considered for their use in real applications (Muller et al 2021). Therefore, more work is needed regarding user adoption, as well as key applications and contexts.

Besides, our research is focused on binary classification (stress or no stress). However, for real applications, evaluating stress as a continuous variable or an ordinal variable with several stress categories would also be interesting, enabling the prediction of more precise stress levels. Some alternatives to accomplish this objective could include the use of different predicted variables when collecting the data and training the models or using the soft output (probability) of the models instead of the hard outputs (predicted class). These approaches could allow to generate nuances of the stress levels.

5.3 Limitations

The main limitation of our work is that we only employed two datasets. In order to test the generalization of these findings, new datasets would be required. Probably due to the fact that biometrics are considered sensitive data, there are not many public datasets in this direction. Also, the number of subjects in both datasets is lower than 30, which is not statistically significant. Also, like the rest of the unsupervised anomaly detection models, LOF classifies inputs as an anomaly or non-anomaly, and we are assuming anomalous cases as stress, but there could be other cognitive states affecting the heart rate features. Finally, we want to clarify that stress detection is not a binary classification problem with well-defined categories, but a complex biopsychosocial construct. Stress detection is a fuzzy problem with a lot of overlap between the categories. Also, in a real application, the recording device is affected by external noise and could record imprecise measurements. External factors could affect the heart rate and not only the stress, such as caffeine or heart rate diseases.

Despite the aforementioned limitations, we believe our study has generated significant novel findings, given the importance of stress management in today's society. In addition, we found that transfer learning had not been investigated enough in affective computing, specifically in stress detection; therefore, our contributions in that sense are novel in the literature. Manufacturers of the popular smartwatches and wristbands that offer stress detection do not usually report how their AI models have been developed and tested and if they resolved transfer learning problems. This kind of research can lead to a new generation of stress management applications for users across multiple contexts.

6 Conclusions and future work

In our first experiments to obtain the best AI models and configuration for stress detection (RQ1), LOF achieved an F1-score of 88.89% in the WESAD dataset and 77.17% in SWELL-KW dataset, and MLP obtained 99.03% in WESAD and 82.75% in SWELL-KW. Evaluating the models on the other dataset from a different context not used in the configuration (RQ2.1), MLP decreased its performance significantly, reducing the F1-score to 28.41% in SWELL-KW and 57.28% in WESAD. In contrast, LOF showed much better transfer learning between datasets, achieving F1-scores of 70.66% in SWELL-KW and 85.00% in WESAD. In order to improve the performance of the models in both datasets, especially due to the low performance of the MLP, we configured and trained the models in both databases (RQ2.2). Multi-dataset LOF models obtained an F1-score of 87.92% and 85.51% in WESAD and 78.03% and 82.16% in SWELL-KW. Multi-dataset MLP models (RQ2.2) also achieved adequate performances, obtaining F1-scores of 78.36%, and 81.33% in WESAD, and 79.37% and 80.68% in SWELL-KW. Therefore, multi-dataset models improve transfer learning and generalization; however, it might not be viable to record stress levels across different contexts for each user purposely. Therefore, we suggest considering unsupervised over supervised models for stress detection, but we emphasize that it is essential to perform additional studies to validate these results in other datasets and other AI models, such as EigenClass (Erkan 2020) or FPFS-EC (Memis et al 2021).

These results show that the typical state-of-the-art approach of applying supervised AI models for stress detection fails to generalize to other contexts. Therefore, we do not recommend their use in real applications, and we suggest employing unsupervised anomaly detection models, especially LOF, which obtained promising results. Mainly, the paper at hand shows the importance of considering transfer learning for real applications of AI models in the affective computing field, including for stress detection.

Acknowledgements This study was partially funded by the Spanish Government Grants IJC2020-044852-I and RYC-2015-18210, co-funded by the European Social Fund, as well as by the COBRA project (10032/20/0035/00), granted by the Spanish Ministry of Defense and by the SCORPION project (21661-PDC-21), granted by the Seneca Foundation of the Region of Murcia, Spain.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Albaladejo-González M, Strukova S, Ruipérez-Valiente JA, et al (2021) Exploring the affordances of multimodal data to improve cybersecurity training with cyber range environments. In: Investigación en Ciberseguridad. Ediciones de la Universidad de Castilla-La Mancha, https://doi.org/10.18239/jornadas_2021.34.52

Arya R, Singh J, Kumar A (2021) A survey of multidisciplinary domains contributing to affective computing. Comput Sci Rev 40(100):399. https://doi.org/10.1016/j.cosrev.2021.100399

Carbonell MF, Boman M, Laukka P (2021) Comparing supervised and unsupervised approaches to multimodal emotion recognition. PeerJ Comput Sci 7:e804. https://doi.org/10.7717/peerj-cs.804

Celdrán AH, Ruipérez-Valiente JA, Clemente FJG et al (2020) A scalable architecture for the dynamic deployment of multimodal learning analytics applications in smart classrooms. Sensors 20(10):2923. https://doi.org/10.3390/s20102923

Chen LI, Zhao Y, fei Ye P, et al (2017) Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. Exp Syst Appl 85:279–291. https://doi.org/10.1016/j.eswa.2017.01.040

Chopade P, Edwards D, Khan SM et al (2019) CPSX: using AI-machine learning for mapping human-human interaction and measurement of CPS teamwork skills. In: 2019 IEEE International Symposium on Technologies for Homeland Security (HST), https://doi.org/10.1109/hst47167.2019.9032906

Erkan U (2020) A precise and stable machine learning algorithm: eigenvalue classification (EigenClass). Neural Comput Appl 33(10):5381–5392. https://doi.org/10.1007/s00521-020-05343-2

Greene S, Thapliyal H, Caban-Holt A (2016) A survey of affective computing for stress detection: evaluating technologies in stress detection for better health. IEEE Consumer Electron Magzi 5(4):44–56. https://doi.org/10.1109/mce.2016.2590178

Hu X, Chen J, Wang F et al (2019) Ten challenges for EEG-based affective computing. Brain Sci Adv 5(1):1–20. https://doi.org/10.1177/2096595819896200

Kerautret L, Dabic S, Navarro J (2022) Detecting driver stress and hazard anticipation using real-time cardiac measurement: a simulator study. Brain Behav 12(2). https://doi.org/10.1002/brb3.2424

Khowaja SA, Prabono AG, Setiawan F et al (2020) Toward soft real-time stress detection using wrist-worn devices for human workspaces. Soft Comput 25(4):2793–2820. https://doi.org/10.1007/s00500-020-05338-0

Koldijk S, Sappelli M, Verberne S et al (2014) The SWELL knowledge work dataset for stress and user modeling research. In: Proceedings of the 16th International Conference on Multimodal Interaction. ACM, https://doi.org/10.1145/2663204.2663257

Lin Q, Li T, Shakeel PM et al (2020) Advanced artificial intelligence in heart rate and blood pressure monitoring for stress management. J Ambient Intell Hum Comput 12(3):3329–3340. https://doi.org/10.1007/s12652-020-02650-3

Makowski D, Pham T, Lau ZJ et al (2021) NeuroKit2: a python toolbox for neurophysiological signal processing. Behav Res Methods 53(4):1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Memis S, Enginoglu S, Erkan U (2021) Numerical data classification via distance-based similarity measures of fuzzy parameterized fuzzy soft matrices. IEEE Access 9:88,583–88,601. https://doi.org/10.1109/access.2021.3089849

Mohammadi A, Fakharzadeh M, Baraeinejad B (2022) An integrated human stress detection sensor using supervised algorithms. IEEE Sens J 22(8):8216–8223. https://doi.org/10.1109/jsen.2022.3157795

Motogna V, Lupu-Florian G, Lupu E (2021) Strategy for affective computing based on HRV and EDA. In: 2021 International Conference on e-Health and Bioengineering (EHB). IEEE, https://doi.org/10.1109/ehb52898.2021.9657654

Mozafari M, Goubran R, Green JR (2021) A fusion model for cross-subject stress level detection based on transfer learning. In, (2021) IEEE Sensors Applications Symposium (SAS). IEEE. https://doi.org/10.1109/sas51076.2021.9530085

Muller H, Mayrhofer MT, Veen EBV et al (2021) The ten commandments of ethical medical AI. Computer 54(7):119–123. https://doi.org/10.1109/mc.2021.3074263

Ochoa X, Dominguez F (2020) Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting. Br J Edu Technol 51(5):1615–1630. https://doi.org/10.1111/bjet.12987

Opitz J, Burst S (2019) Macro F1 and Macro F1. (2):1–12. https://arxiv.org/abs/arXiv:1911.03347

Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Pluntke U, Gerke S, Sridhar A, et al (2019) Evaluation and classification of physical and psychological stress in firefighters using heart rate variability. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), https://doi.org/10.1109/embc.2019.8856596

Pourmohammadi S, Maleki A (2020) Stress detection using ECG and EMG signals: a comprehensive study. Comput Methods Programs Biomed 193(105):482. https://doi.org/10.1016/j.cmpb.2020.105482

Ramos-Giraldo P, Reberg-Horton C, Locke AM et al (2020) Drought stress detection using low-cost computer vision systems and machine learning techniques. IT Professional 22(3):27–29. https://doi.org/10.1109/mitp.2020.2986103

Sağbaş EA, Korukoglu S, Balli S (2020) Stress detection via keyboard typing behaviors by using smartphone sensors and machine learning techniques. J Med Syst 44(4). https://doi.org/10.1007/s10916-020-1530-z

Schmidt P, Reiss A, Duerichen R, et al (2018) Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. ACM, https://doi.org/10.1145/3242969.3242985

Shields GS, Sazma MA, Yonelinas AP (2016) The effects of acute stress on core executive functions: a meta-analysis and comparison with cortisol. Neurosci Biobehav Rev 68:651–668. https://doi.org/10.1016/j.neubiorev.2016.06.038

Sikander G, Anwar S (2019) Driver fatigue detection systems: a review. IEEE Trans Intell Transp Syst 20(6):2339–2352. https://doi.org/10.1109/tits.2018.2868499

Sriramprakash S, Prasanna VD, Murthy OR (2017) Stress detection in working people. Procedia Comput Sci 115:359–366. https://doi.org/10.1016/j.procs.2017.09.090

Szakonyi B, Vassányi I, Schumacher E, et al (2021) Efficient methods for acute stress detection using heart rate variability data from ambient assisted living sensors. BioMed Eng OnLine 20(1). https://doi.org/10.1186/s12938-021-00911-6

Vasanthakumari S (2019) Soft skills and its application in work place. World J Adv Res Rev 3(2):066–072. https://doi.org/10.30574/wjarr.2019.3.2.0057

Wu J, Zhang Y, Zhao X (2021) Stress detection using wearable devices based on transfer learning. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, https://doi.org/10.1109/bibm52615.2021.9669904

Zhao Y, Nasrullah Z, Li Z (2019) Pyod: a python toolbox for scalable outlier detection. J Mach Learn Res 20(96):1–7. http://jmlr.org/papers/v20/19-011.html

Zontone P, Affanni A, Bernardini R, et al (2019) Stress detection through electrodermal activity (EDA) and electrocardiogram (ECG) analysis in car drivers. In: 2019 27th European Signal Processing Conference (EUSIPCO). IEEE, https://doi.org/10.23919/eusipco.2019.8902631

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.