Journal of Intelligent Information Systems (2024) 62:77–90
https://doi.org/10.1007/s10844-023-00806-z

RESEARCH

Semi-supervised and ensemble learning to predict work-related stress

Fátima Rodrigues[1,2] · Hugo Correia[1]

Received: 20 April 2023 / Revised: 18 July 2023 / Accepted: 19 July 2023 /
Published online: 5 August 2023

Abstract
Stress is a common feeling in people's day-to-day life, especially at work, being the cause of several health problems and absenteeism. Despite the difficulty in identifying it properly, several studies have established a correlation between stress and perceivable human features. The problem of detecting stress has attracted significant attention in the last decade. It has been mainly addressed through the analysis of physiological signals in the execution of specific tasks in controlled environments. Taking advantage of technological advances that allow to collect stress-related data in a non-invasive way, the goal of this work is to provide an alternative approach to detect stress in the workplace without requiring specific controlled conditions. To this end, a video-based plethysmography application that analyses the person's face and retrieves several physiological signals in a non-invasive way was used. Moreover, in an initial phase, additional information that complements and labels the physiological data was obtained through a brief questionnaire answered by the participants. The data collection pilot took place over a period of two months, having involved 28 volunteers. Several stress detection models were developed; the best trained model achieved an accuracy of 86.8% and a F1 score of 87% on a binary stress/non-stress prediction.

Keywords Stress · Semi-supervised learning · Ensemble learning · Classification

1 Introduction

In recent decades, there has been an increasing focus on the quality of work as an integral part of the quality of life. Indeed, work fulfils several functions in a person's life (e.g., economic, social, psychological), and determines their quality of life. Workplace stress not only has a detrimental impact on people but also adds expenses to the business due to high

✉ Fátima Rodrigues
mfc@isep.ipp.pt

Hugo Correia
1170569@isep.ipp.pt

[1] ISEP, Polytechnic Institute of Porto, Rua Dr. António Bernardino de Almeida, Porto, 4249-015 Porto, Portugal

[2] ISRC, Interdisciplinary Studies Research Center, Porto, Portugal

staff turnover, absenteeism, decreased productivity, higher risk of accidents, and a lack of innovation (Hanula-Bobbitt & Bočkutė, 2022). In a Portuguese study (Ordem dos Psicólogos Portugueses, 2023) workers were estimated to miss work on 7.4 days a year due to stress and psychological health problems. In 2022, absenteeism and reduced work efficiency cost Portuguese companies €1.8 billion and €3.5 billion, respectively, resulting in a total indirect cost of €5.3 billion per year.

According to the World Health Organization, stress is the reaction observed when people face demands and pressures that are greater than their capacity to handle them and test their coping mechanisms (World Health Organization, 2023). These demands are linked to high perceived stress, a lack of social support from colleagues and supervisors and heavy workloads. Long-term stress may be damaging since it has negative physiological effects that can lead to both mental and physical health issues (Seiler et al., 2020).

In light of such potential harmful effects, detecting levels of stress at an early stage is crucial. As a result, much progress has been done in the last years towards the development of various devices to collect real-time physiological data. Nevertheless, a reliable real-time stress detection system that is unobtrusive and completely transparent for the user has not been created yet.

There is a study (Carolan et al., 2017) that confirms that stress-related illnesses may be prevented and treated in the workplace, and those who do so are more likely to be more productive. Hence, non-intrusive stress-sensing tools that continuously monitor stress levels, with a minimal impact on workers' daily lives, could be used to regulate or relieve stress by providing personalized recommendations or even by just alerting the users of their current level of stress. These applications might result in more efficient and cost-effective interventions in stressful work situations, resulting in more productive work environments where employees can better manage their workload.

Although stress is not simple to properly identify, there are several studies that have established a correlation between stress and perceivable human features (Alberdi et al., 2016; Panicker & Gayathri, 2019). People can sometimes identify their stress levels through the recognition of external body manifestations. However, this is not always possible, in which case underlying body reactions have to be analysed for stress to be detected. Hence, using several physiological signals can be beneficial in many ways, because each signal can compensate for the lack of information in the others. Given that the amount and type of information extracted from each signal are different, and since stress is related to various factors and has diverse symptoms, a multimodal solution is best suited for an efficient stress detection.

This paper describes a Machine Learning (ML) approach to build a stress detection model based on two physiological signals and facial expressions collected from 28 volunteers, over a two-month period. Our main contribution is the validation of the ML approach using a multimodal dataset collected in a non-invasive and non-intrusive way — that is, while people carry out their daily activities in the workplace, without changing their routines in any way. We expect our solution to contribute to the mitigation of work-related stress in employees, the improvement of their mental health and the reduction of the associated socioeconomic costs.

The remainder of this paper is organized as follows. Section 2 reviews previous works on stress detection that follow a non-intrusive approach. The methodology adopted here is explained in Section 3, including the data acquisition, the labelling process to complete the missing stress labels, the derivation of features from the heart rate variability signal, the feature selection, and the development and evaluation of models. Challenges and limitations are discussed in Section 4. The main conclusions and prospects of future work are disclosed in the final section.

2 Related work

The literature on stress detection is vast and started with studies that explore data collected from very different signals (e.g., electrodermal activity, electrocardiogram, electroencephalogram, blood pressure, skin temperature and respiration features), all of them acquired through obtrusive equipments (Alberdi et al., 2016). More recently, researchers have combined data coming from several physiological signals, having shown the advantage of detecting stress from various sources of stress (Alberdi et al., 2016; Panicker & Gayathri, 2019).

Furthermore, other works emphasise how crucial it is to monitor physiological signals since they enable users to receive quick feedback during routine tasks (Gedam & Paul, 2021). However, a drawback of these methods is the need to use wearable sensors at all times (and in specific places in the body) in order to allow for an accurate and continuous monitoring. Other approaches have tried to remove these limitations. For example, (Hilmy et al., 2021) propose to detect stress based on speech signal analysis techniques with machine learning. However, this method may result in speech misunderstanding and improper stress detection in real-life situations involving loud environments.

Sabour et al. (2021) proposes a new dataset, UBFC-Phys, collected with and without contact from participants living under social stress situations. A wristband was used to measure contact blood volume pulse (BVP) and electrodermal activity signals, while a photoplethysmography video recordings allowed to compute remote pulse signals and facial expression features. Pulse rate variability (PRV) was extracted from the BVP signal. Experimental results showed a stress state recognition accuracy of 85.48% achieved by remote PRV features.

In Park et al. (2018) the authors aimed to uncover the viability of mental stress level prediction through the Heart Rate Variability (HRV) attained from the photoplethysmography (PPG) sensors in wearable devices. The experiment involved the measurement of the subject's PPG signals for 30 seconds, three times a day, using the wearable device. At the end of the day, the participants evaluated their own mental stress using the Perceived Stress Scale (PSS). Results showed that the LF/HF feature was correlated with the subject's PSS scores, and the predicted model resulted in an average accuracy of 86.35%.

Moreover, Dalmeida and Masala (2021) presented an effective and non-invasive approach on stress detection. One of the goals was to classify stress mainly using HRV-derived metrics, obtained from wearable devices. The device considered was the Apple Watch. The beats per minute were extracted using the Apple Breathe App, which allowed to calculate the RR intervals. From the several developed models, the SVM model with the Radial Basis Function (RBF) kernel provided the best accuracy (83.33%). Lastly, a web application was built in order to test the developed models. The authors state that the application was able to predict stress conditions with a 71% prediction probability, with an increased probability of 79% in relaxed states.

Another study (Can et al., 2019) collected physiological data from 21 students in an algorithmic programming competition over the course of nine days in order to identify stress. They collected data through non-obstructive wearable devices that could be used daily. The developed system contained heart activity data from a photoplethysmogram (PPG) sensor, skin conductance from the GSR, accelerometer and temperature data. The PPG sensor was used to measure the blood flow, thus providing the RR intervals. The authors developed a three-class detection system that was capable of differentiating the stress levels in three different situations: free day, lecture and contest sessions, with 90.40% accuracy.

Another relevant study (Maxhuni et al., 2016) involving 30 employees suggests modeling stress levels using several behavioural characteristics acquired via smartphones, with the caveat that the labelled data for each individual is scarce. To surpass this limitation, the authors chose the closest individual for knowledge transfer by selecting the most comparable individual model tree. Their findings showed the stress levels of subjects for whom less data were collected may be more accurately predicted by employing data (instances or models) from similar workers. However, using this transfer learning approach for workers with different backgrounds might have a negative impact on accuracy.

In summary, the majority of works rely on sensors embedded in wearables, which can be uncomfortable or unsuitable for many daily situations. The average accuracy of the stress detection works analysed lies between 85%-90%. An objective and reliable comparison among different studies would be difficult to perform and very debatable because of the distinct data used and the different experimental conditions of each work.

As mentioned in the previous section, several physiological signals are relevant to measure various aspects of human behavior and classify different mental conditions. In this study we will combine two physiological signals — heart rate variability (HRV) and percentage of eye closure (PERCLOS) — with facial expressions and some demographic attributes of users. The HRV can be used to evaluate stress because it reflects the cardiac activity and the autonomic nervous system (ANS) response. Different stress-inducing methods were observed to stimulate a variation in the ANS response through the changes in HRV parameters (Kim et al., 2018). Although PERCLOS is mainly used as a measure of fatigue, Marquart et al. (2015) demonstrated that PERCLOS is affected by mental stress, as its value decreases in stressful situations. The facial features are used for monitoring the emotional responses due to stress (Almeida & Rodrigues, 2021).

Combining physiological signals (HRV and PERCLOS) with facial expressions qualifies as a multimodal data analysis, because each modality represents a distinct type of information or sensory input. Physiological signals capture information related to the body's physiological responses, while facial expressions provide visual clues about emotional states. The different modalities are expected to interact and exhibit correlations within the context of stress prediction. For example, stress can affect physiological responses such as changes in HRV and PERCLOS, which may be reflected in facial expressions. By considering these interactions and correlations, a multimodal analysis can capture a more comprehensive understanding of stress patterns.

Our work differs from existing studies in several points. First, to the best of our knowledge, the combination of HRV and PERCLOS physiological signals with facial expressions and users' demographic data has not been explored in the literature before. Second, the data were collected in the office, with volunteers doing their daily activities without any restrictions imposed to them. This was possible because the physiological data and facial expressions were collected by a video-plethysmography application that runs in the background. The workers do not even notice that they are being monitored, which prevents them from biasing the data collection and permits to build a non-invasive solution. Few previous works have detected stress in a non-invasive and non-intrusive way. Finally, semi-supervised learning is used to label physiological data, since data have been classified by the collaborators, so labels are scarce. Feature selection combined with ensemble learning is implemented to find the best predict stress model.

3 Methodology

In this study, we propose a ML stress detection approach that follows the CRISP-DM methodology (Wirth & Hipp, 2000). The corresponding end-to-end pipeline will be briefly described next.

Data acquisition is the first stage, where the physiological signals are collected from the users while they are doing their regular work. In parallel, self-reported measures are collected through a questionnaire implemented in an application that periodically retrieves the user's perceived stress, among other variables, to label the physiological data.

In order to deal with the scarce labeling of data, common to many real-world applications, we apply semi-supervised learning to reduce the amount of unlabelled data. To address this issue, we use an inductive self-training method that consists of a single supervised classifier that is iteratively trained on both labelled data and data that has been pseudo-labelled in previous iterations of the algorithm (Van Engelen & Hoos, 2020).

Next, given that stress is regulated by the Autonomous Nervous System, the heart rate variability (HRV) is a reliable signal for identifying stress, because it cannot be altered by the users. HRV may be separated into time and frequency domain measures. As demonstrated in several studies (Alberdi et al., 2016; Cheng & Chen, 2022; Dalmeida and Masala, 2021; Panicker & Gayathri, 2019; Reijmerink et al., 2020; Sabour et al., 2021), time and frequency HRV characteristics are reliable stress indicators.

After data preparation, the Random Forest algorithm is used to determine the adequate number of features to be used according to the initial data collected. This algorithm measures the importance of the features by randomly sampling a set of features in the out-of-bag samples and calculating the percentage increase in the misclassification rate as compared to the out-of-bag rate with all variables intact. Then, to select the set of features that works best to predict stress, it is good practice to use more than one algorithm, so we shall use the Relief algorithm (Kira & Rendell, 1992) and an ensemble feature selection method based on the combination of several Random Forest models.

Regarding the choice of ML algorithms to use, we choose to use ensemble learning algorithms since an ensemble method combines the predictions of various machine-learning-based algorithms to provide predictions that are more accurate (Dietterich, 2000). Bagging and boosting are the most frequently used ensemble learning methods. Both methods use resampling techniques to construct different training sets for each classifier. Bagging manipulates the original training data by randomly drawing instances with replacement. Boosting also manipulates the original training data by drawing with replacement, but it uses the performance of the previous classifier(s) to calculate the probability of selecting each instance. Similarly to simple models, some ensemble methods work better than others in certain conditions (Dietterich, 2000), so we consider four ensemble machine learning classification algorithms to deal with the characteristics of the various datasets: two boosting methods — AdaBoost (AB) and Gradient Boosting (GBM) — and two bagging methods — Random Forests (RF) and Extra Trees (ET).

In order to evaluate the models' capabilities, all the experiments were conducted using the same metrics and making a data-stratified k-fold cross validation (with $k = 10$), since it provides good results with a moderate load on the processing unit.

3.1 Data acquisition

To create the dataset to develop the stress detection models two applications were used. A video-plethysmography application silently stores the participants' physiological data only using the computer video camera. This application guarantees the synchronization between the different physiological signals collected (beat per minute (bpm) and PERCLOS) and facial expressions, and stores them in files value format, without storing any images, thus complying with data protection policy. At the same time, a questionnaire application is also being executed, the main goal of which is to collect information to be used to label the physiological data in terms of stress. To this end, we automated this task by developing an application that has three questionnaires that have been clinically validated to capture the subjects perceived stress and mood states of the employees at work.

A first questionnaire, which is only presented the first time the user enters in the application, and collects demographic information and some habits of the participants (e.g., number of coffees, alcoholic beverages and cigarettes taken per day, sleep schedule, average hours of sleep per day, average perception of sleep quality in the previous week, kind of person, such as matinal, nocturnal) through 17 different questions. Data from 28 healthy employees of an insurance company were collected for a period of 8 weeks. Table 1 provides a summary of some employees' demographics collected in this first questionnaire.

As can be seen, the employees that participated in our study had heterogeneous characteristics with regard to gender, age, marital status, educational level and role in the company, which may be advantageous for the creation of the models.

The second questionnaire collects information relevant to the working environment and job demands of employees during working days. It consists of a bi-daily questionnaire that is triggered 5 minutes before lunch time and before the end of business, both of which are defined by the user. Here questions are asked about the level of stress felt by the user during the morning/afternoon work period (period stress label), as well as the emotional state and perception of his/her productivity and the level of difficulty of tasks performed. The questionnaire also includes questions concerning the level of alertness, drowsiness and tiredness in a total of 10 different questions.

The last questionnaire consists of a simple pop-up that appears in the user's screen four times a day, twice in the morning and twice in the afternoon. This questionnaire asks the user about his/her stress level at the moment through a slider bar that ranges between 1 and 10, with the respective labels (e.g., "Not stressed", "Moderately stressed, "Extremely stressed"). This questionnaire pops up on the user's screen randomly, between the user's entry time and the scheduled lunch/leave time. The minimum difference between the generation of two pop-ups was set to 30 minutes and it was also imposed that these pop-ups may appear up to 5 minutes before the bi-diary form.

In summary, four stress labels are collected per day for each participant during the morning/afternoon work period, in which the user answers the question: *"How do you rate your stress at this moment?"* (pop-up label) and two more stress labels in which the user answers the question *"How would you rate your stress during this morning/afternoon?"* (period label). An analysis made to the values of stress pop-up labels and stress period labels (see Fig. 1) shows that, on average, all users define stress period labels slightly higher than pop-up stress labels. A Spearmans test between the two stress labels response shows a correlation coefficient of 0.760.

Table 1 Some attributes of the participants collected through the first questionnaire

Gender		Marital St.		Age		Weight (kg)		Education		Role	
Male	9	Single	9	25 - 40	9	52 - 63	15	< 9 years	1	Executive	1
Female	19	Married	16	41 - 50	11	64 - 75	6	9-12 years	4	Manager	5
		Divorced	3	>50	8	>75	7	High School	23	Collaborator	22

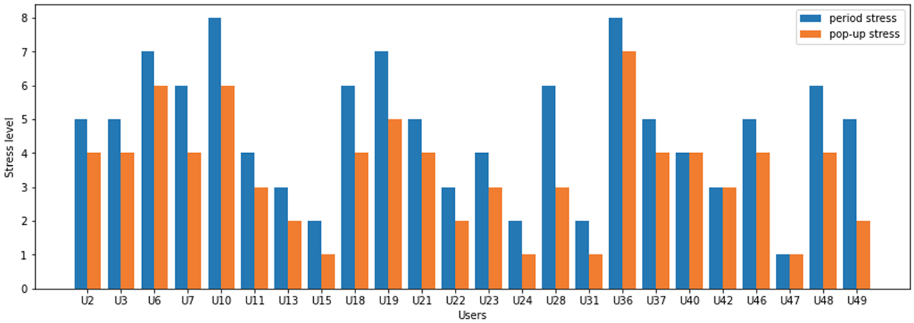

Fig. 1 Stress period vs. stress pop-up by user

In Table 2 we present the overall stress responses for the whole period of 8 weeks. Only the users with a minimum of 10 stress period labels — which is equivalent to 5 workdays — were included, thus leaving a remaining set of 24 users. The total number of stress labels, period and pop-up labels, was 4291. In order to simplify the measurements of the work-related stress, we have classified the stress-level into three classes: [1..3] as low stress, [4..6] as moderate stress and [7..10] as high stress. Results show that, during the entire monitoring period, 18 subjects perceived high stress at some point.

A total of six stress labels collected on each working day for each worker is undoubtedly insufficient to catalog all the physiological data collected. In order to deal with the scarcity of the labelled data, we apply semi-supervised learning. The data collected in the two questionnaires (bi-daily and pop-up) will only be used to develop stress prediction models to fill the unlabelled physiological data.

The objective of this study is to predict stress levels using various behavioural characteristics, especially given the limited availability of user labelled data. The ultimate goal is to stop relying on subjective, self-reported data for stress assessment and instead utilize solely objectively detected physiological data to enable continuous stress measurement.

3.2 Data labelling with semi-supervised learning

The efficiency of supervised learning is highly dependent on the volume of labelled instances. However, having a reasonable size of labelled instances is difficult, expensive and time-consuming to obtain since the availability and commitment of users are required. This is not always easy in work environments, because in real work scenarios the incorrect placement of the camera, unscheduled commitments of the collaborator and other unexpected issues

Table 2 Overall number and percentage of stress labels

Stress level	Nr. responses (%)	Nr. users
low	478 (11,1%)	20
medium	3505 (81,7%)	24
high	308 (7,2%)	18

can occur and force the interruption of data collection. There are a mix of physiological data with and without stress labels, which require the adoption of semi-supervised learning.

Since the initial data had a large portion of unlabelled data, we convert the goal attribute to a binary label. Given the stress values distribution, and in order to have a more cautious classification that aims to minimize false negatives, we considered "stress" to apply when the value of the label is at least 4, in a scale from 1 to 10. This gives the following distribution of the goal attribute: 24.5% stress, 48.2% no stress and 27.3% not classified.

Before the start of data labelling, we conducted a stratified sampling of the labelled data into two disjoint datasets, a training dataset (80%) and a validation dataset (20%). In this way, the validation dataset without any pre-processing (i.e., with only the original labels) guarantees a more rigorous and unbiased final evaluation of the developed models.

To fill the missing values, several models were developed with the training dataset, using different learning algorithms: logistic regression, two decision-tree algorithms, C5.0 and rpart, Gaussian naive Bayes and support vector machine (SVM) with the kernel rbfdot. These were chosen to represent a wide range of approaches. We performed a 10-fold cross validation in all experiments with the default parameters of the algorithms. Table 3 presents the mean (standard deviation) of accuracy and F1 measures of the three best models. The C5.0 model clearly achieved the best performance, so it was the model chosen to predict the labels for the portion of unclassified data.

From the predictions obtained with C5.0 algorithm we selected only those classified with high confidence (greater than 90%) to add to the training set. Next, the classifier was re-trained with the new classified data added to the initial dataset and the procedure was repeated. At the end of five iterations, the majority of labels were classified, with only a residual number remaining unclassified, giving a final distribution of 36.5% stress and 63.5% no stress.

3.3 Feature derivation from heart rate variability signal

Due to the several distinct user profiles that take part in the pilot, we see a significant variation in the data made accessible per user. There are many comma-separated BPM files for each user, each of which refers to a distinct time of day. This required going through a laborious data processing procedure that began with looking for empty folders and looking for missing files. Following this, it was discovered that each user had a varied amount of data, which might have led to some incoherence because some files included measurements for longer time periods than others. For shorter time periods, less precise readings will be obtained when using this information to infer HRV features. To obtain a more uniform analysis of the data, an algorithm was created to divide each csv file into 5-minute files.

Heart rate variability (HRV) is a measure of the variation in time between consecutive heart beats, whereas beats per minute (BPM) is a measure of the heart rate or the number of heart beats per minute (Kim et al., 2018). Therefore, it is not possible to obtain HRV directly from

Table 3 Performance of models to populate unlabelled data

	C5.0	SVM	rpart
accuracy	0.973 (0.0027)	0.858 (0.0178)	0.916 (0.0056)
F1	0.963 (0.0032)	0.815 (0.0197)	0.884 (0.0072)

BPM. However, HRV can be derived from the time intervals between consecutive heartbeats — also known as interbeat intervals (IBI) —, which can be calculated from BPM. The formula to convert BPM to IBI is $IBI = 60,000/BPM$ (Reijmerink et al., 2020). After obtaining IBI we perform some pre-processing operations that include filtering to remove high-frequency noise, interpolation to fill missing data points, and detrending to remove baseline trends.

Next, time-domain HRV measures, which include measures of the variation in the interval between consecutive heartbeats, were calculated directly from the IBI data. Frequency-domain HRV measures, which refer to the rate of oscillations in the heart rate signal, were obtained by applying spectral analysis to the IBI data. The following time and frequency HRV characteristics were calculated using the python package pyHRV (Gomes et al., 2019):

- HRV time domain features: Normal-to-Normal intervals (NN), heart rate average values (AVNN), standard deviations of NN intervals (SDNN), square root of the mean squared difference of successive NN intervals (RMSSD), the SDNN/RMSSD ratio, standard deviation of successive differences (SDSD), number of pairs of successive NN intervals that differ by more than 50 ms (NN50), the ratio of NN50 to the total number of NN intervals (PNN50), the same as the two last metrics but for NN intervals that differ more than 25 ms (NN25, PNN25) and 20 ms (NN20, PNN20).
- HRV frequency domain features: Power spectral density (PSD) at each of the three frequency bands, VLF (0 - 0.04 Hz), LF (0.04 - 0.15 Hz) and HF (0.15 - 0.4 Hz), peak of each frequency band (VLF peak, LF peak, HF peak), total signal power and the LF/HF ratio.

3.4 Feature selection

The training dataset has 17 attributes from the demographic questionnaire, 20 time and frequency features derived from HRV, 7 facial expressions and the PERCLOS signal. Choosing an optimal feature set is important to speed up the training process and make learnt classifiers simpler, thus being easier to interpret and to deploy.

In order to obtain the most representative features, first we determined the adequate number of features to use as input to the feature selection algorithms. This was accomplished using the RF algorithm with a backward elimination method. We calculated the cross-validation prediction error of models with sequentially reduced number of predictors (ranked by importance of the variable) via a nested cross-validation procedure. As Fig. 2 shows, 25 features

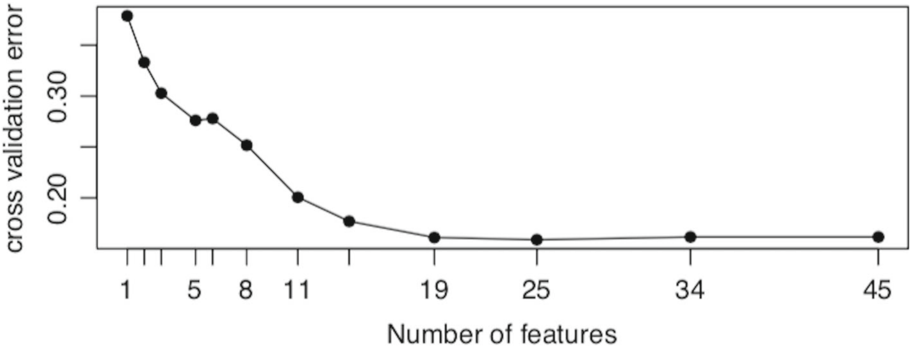

Fig. 2 Classification error rate using different numbers of features

give the minimum cross validation error (0.158), so we considered 25 as the number of features to define by each feature selection algorithm.

The Relief algorithm estimates the relevance of features according to how well their values distinguish between the instances of the same and other classes that are near each other. Given randomly selected instances, the algorithm searches for the k-nearest-neighbours from the same class and k-nearest-neighbours from each of the other possible classes. Based on which class the neighbours belong to, the algorithm updates the feature quality information by increasing its value if the feature separates instances with different classes well and by decreasing its value in the opposite scenario. The process of random instance selection is repeated several times, the number being pre-chosen by the user. We ran the Relief algorithm with $k = 5$ over 1000 iterations and selected the 25 top features: PNN25, Total.Power, RMSSD, happy, fear, angry, disgust, sad, perclos, avgSleepingHours, typeOfSchedule, weight, alcohol, AVNN, SDNN, PNN50, LF.Relative, sleepingHours, sleepQuality, wakeupHours, HF.Peak, LF.Peak, bestDisposition, ULF.Absolute, Lf.Normalized.

The ensemble feature selection method built is based on the combination of several Random Forest feature selection models (Ens-RF). Ensembles outperform the individual models that form the ensemble, because they are based on some form of diversity among the individual models. The ensembles herein developed used different samples of observations to obtain each model. This approach works better if the data from which we obtain the different models are highly redundant. We assumed that the necessary degree of redundancy is achieved with the k-fold method. We split the data into 10-folds to build 10 different feature selectors. For each feature the number of times it has been selected by all the feature selectors was counted and the 25 features with greater counts were chosen: PNN20, surprise, VLF.Absolute, avgSleepingHours, happy, typeOfSchedule, weight, alcohol, AVNN, PNN25, Total.Power, RMSSD, PNN50, fear, SDNN, angry, disgust, sad, perclos, firstHalfHour, HF.Absolute, Hf.Lf, Hf.Normalized, HF.Relative, LF.Absolute.

3.5 Stress modeling

This section presents the results of applying the different algorithms to the training dataset to predict the stress.

In the first experiment we develop the stress prediction models with the four ensemble algorithms using the features selected by the Relief algorithm and by the Ens-RF algorithm. However, because the accuracy of the models, with both set of features, is very similar and since the number of common features selected by both algorithms is quite high (16 features), we also ran the ensemble algorithms with the 16 features common to both: avgSleepingHours, typeOfSchedule, weight, alcohol, AVNN, SDNN, PNN50, PNN25, Total.Power, RMSSD, happy, fear, angry, disgust, sad and PERCLOS. Table 4 presents the accuracy of the models with the different sets of features.

Table 4 Accuracy of ensemble models with different sets of features

	Boosting		Bagging	
Feature selection	AB	GBM	RF	ET
Relief (25 attrib.)	0.784 (0.023)	0.812 (0.022)	0.848 (0.016)	0.833 (0.020)
Ens-RF (25 attrib)	0.774 (0.021)	0.810 (0.021)	0.848 (0.015)	0.837 (0.017)
Intersection (16 attrib)	0.778 (0.027)	0.813 (0.019)	0.853 (0.019)	0.844 (0.016))

As can be seen, the results of the boosting algorithms with the features from the intersection of both initial feature sets are quite similar to the results achieved with both initial feature sets. They are also slightly higher for the bagging algorithms and, because a smallest prediction set is preferable to deploy the model, we will consider this subset of features.

Table 5 shows the mean (standard deviation) of accuracy, F1 weighted and ROC AUC of the four ensemble algorithms with their default parameters and using the 16 intersection features.

From the results obtained it is clear that the bagging algorithms outperform the boosting algorithms. Both bagging algorithms provide better scores with default configurations.

Tran et al. (2017) show that the combination of bagging and feature selection not only helps to improve classification accuracy of bagging, but also helps to reduce the complexity of the learnt classifiers compared with the standard bagging method. Hence, we also searched for a feature subset with bagging. We ran both bagging algorithms multiple times, combining each one of the attributes with the rest of them. We used subsets comprising from two up to all the attributes. Therefore, all possible combinations among the attributes were performed. The purpose was to see how relevant the features were to the stress prediction and which minimal subset of features and method yielded the best results.

Both algorithms, RF and ET, achieved the best results with the same 16 initial features. This leads us to conclude that the initial set of 16 features selected by both feature selection algorithms is minimal.

Concerning the models performance, RF algorithm achieved a mean accuracy of 85.3% (stdev=1.9%) and the ET algorithm a mean accuracy of 84.4% (stdev=1.6%). The accuracy of the RF algorithm is therefore slightly higher.

To compare the difference between classification models more rigorously, a statistical test suite was carried out on the experimental results to validate their results further and determine whether there exists a significant difference among them. We checked normality by applying the Shapiro test to the difference of cross-validation accuracy of both models. Since p-value = 0.1532 is greater than the 0.05 significance level, we can assume normality and apply the t-Student test. The p-value = 0.0048 of the t-Student test is also less than the significance level alpha = 0.05, therefore we can conclude that the difference beyween the accuracies of the models is slight but statistically significant, being the RF algorithm the most suitable algorithm to predict stress with this dataset.

In our last approach, hyper-parameter optimization was performed with GridSearch strategy on the RF algorithm to determine the best choice of parameters that would yield the highest performance. Table 6 presents the parameter grid used.

We finalize the model by training it on the entire training dataset and make predictions for the hold-out validation dataset to confirm our findings. We achieve an accuracy of 86.8% and F1 score of 87.0% on the hold-out validation dataset. A slight increase of these metrics is the result of parameters optimization of RF algorithm. With this experiment, it is possible to

Table 5 Performance of ensemble models to predict stress

	Boosting		Bagging	
	AB	GBM	RF	ET
accuracy	0.778 (0.028)	0.813 (0.019)	0.853 (0.019)	0.844 (0.016)
F1 weighted	0.775 (0.028)	0.810 (0.019)	0.852 (0.017)	0.841 (0.019)
ROC AUC	0.847 (0.014)	0.888 (0.012)	0.924 (0.010)	0.921 (0.006)

Table 6 Parameter grid for RF optimization

Parameter	Values
n_estimators	[100, 800, 1200]
max_features	['sqrt','log2']
min_samples_split	[2,5,10]
min_samples_leaf	[1,2,5]

draw the conclusion that some participant characteristics, such as weight, average sleeping hours, type of schedule practiced by the worker, alcoholic beverages taken per day, combined with some HRV features, facial expressions and PERCLOS, are good indicators of the stress level of participants.

4 Challenges and limitations

Using a multimodal dataset for monitoring behaviour patterns of individuals in their working environment has the potential to provide valuable insights concerning their stress. This research aimed to combine data from different sources, such as physiological data (from a video-plethysmography application) and demographic data (from a self-reported questionnaire). The challenges that we faced in the study arose from the integration of multiple sources of data, since data were collected in a real-life environment from heterogeneous sources. Another challenge was related to acquiring sufficient labelled data. This is often very difficult and expensive to obtain because workers are required to fill self-reported questionnaires. Finally, another limitation in our study is the large number of missing values acquired from the video-plethysmography application because workers temporarily leave their desk, e.g., due to meetings.

5 Conclusions and future work

Stress at work is common to many professions. Being aware of one's own stress level is an important step towards the prevention of diseases and increasing the productivity. This paper presented a ML approach to develop a novel classification model to help workers cope with their daily stress in the workplace. Data were collected from workers carrying out their professional activities in an uncontrolled environment. The dataset comprising measurements of heart beats per minute (BPM), percentage of eye closure (PERCLOS) and facial expressions, was entirely built from scratch, using a video-based plethysmography application. This dataset was complemented with a questionnaire answered through a desktop application, specifically developed to collect some demographic data of workers and also for the initial labelling of the physiological data collected.

Due to insufficient labelled data, semi-supervised learning was used to reduce the amount of unlabelled data. Then, we developed four different models based on ensemble learning to predict stress. The results of the experiments show that combining some HRV features, facial expressions, and demographic data along with the PERCLOS, the RF algorithm provided the best performing stress model with 86.8% accuracy, 80% recall and 87% F1 Score.

As future work, we plan to use multi-label classifiers in order to make a more assertive stress prediction, with at least three classes. We also started to develop a stress recommendation system that will integrate the stress detection model. With the help of the stress detection model, the recommendation system will provide stress-easing recommendations that adapt to the workers by considering their ratings from past recommendations, as well as their profiles. The goal is to prevent future stress episodes and, most importantly, mitigate long-term stress.

Acknowledgements The authors would like to thank the company partner, Médis - Companhia Portuguesa de Seguros de Saúde, SA, for contributing with participants for data collection. We also acknowledge the support from School of Health/Polytechnic Institute of Porto in the elaboration of questionnaires for the labelling of data.

Author Contributions Both authors contributed equally to this work.

Funding Open access funding provided by FCT|FCCN (b-on). The present work has been developed under the Project Mad@Work from ITEA3, Eureka Cluster on software innovation and COMPETE 2020 - POCI-01-0247-FEDER-046168 LISBOA-01-0247-FEDER-046222.

Availability of supporting data The data and materials used in the current study are available from the corresponding author upon reasonable request and authorization from the beneficiary entity of the project.

Declarations

Ethical Approval Not Applicable

Competing interests The authors declare that they have no competing interests.

References

Alberdi, A., Aztiria, A., & Basarab, A. (2016). Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review. *Journal of Biomedical Informatics, 59*, 49–75. https://doi.org/10.1016/j.jbi.2015.11.007

Almeida, J., Rodrigues, F. (2021). Facial expression recognition system for stress detection with deep learning. In ICEIS (1), 256-263. https://www.scitepress.org/Papers/2021/104742/104742.pdf

Can, Y. S., Chalabianloo, N., Ekiz, D., et al. (2019). Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors, 19*(8), 1849. https://doi.org/10.3390/s19081849

Carolan, S., Harris, P. R., Cavanagh, K. (2017). Improving employee well-being and effectiveness: systematic review and meta-analysis of web-based psychological interventions delivered in the workplace. Journal of Medical Internet Research, *19*(7), Article e271. https://doi.org/10.2196/jmir.7583

Cheng, J. C., & Chen, A. L. P. (2022). Multimodal time-aware attention networks for depression detection. *J Intell Inf Syst, 59*, 319–339. https://doi.org/10.1007/s10844-022-00704-w

Dalmeida, K. M., & Masala, G. L. (2021). HRV features as viable physiological markers for stress detection using wearable devices. *Sensors, 21*(8), 2873. https://doi.org/10.3390/s21082873

Dietterich, T. G. (2000). Ensemble methods in Machine Learning. In multiple classifier systems, First International Workshop, MCS. Cagliari, Italy, June 21–23, Proceedings 1. Springer, Berlin Heidelberg, 1–15 https://doi.org/10.1007/3-540-45014-9_1

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

Gomes, P., Margaritoff, P., Silva, H. (2019). pyHRV: Development and evaluation of an open-source python toolbox for heart rate variability (hrv), Proc. International Conference on Electrical, Electronic and Computing Engineering (icetran), 822-828.

Hanula-Bobbitt, K., Bočkutė, K. (2022). Stress management in the education sector. Master thesis, Tampere University of Applied Sciences, Finland 2022. https://urn.fi/URN:NBN:fi:amk-2022121429695

Hilmy, M. S. H., Asnawi, A. L., Jusoh, A. Z. et al. (2021). Stress classification based on speech analysis of MFCC feature via Machine Learning. In 8th International Conference on Computer and Communication Engineering (ICCCE) 339-343, IEEE. https://doi.org/10.1109/ICCCE50029.2021.9467176

Kim, H. G., Cheon, E. J., Bai, D. S. et al. (2018). Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investigation, *15*(3), 235. https://doi.org/10.30773/pi.2017.08.17

Kira, K., & Rendell, L. A. (1992). The feature selection problem: traditional methods and a new algorithm. *AAAI, 2*, 129–134. https://doi.org/10.5555/1867135.1867155

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing, 3*, 2854–2861. https://doi.org/10.1016/j.promfg.2015.07.783

Maxhuni, A., Hernandez-Leal, P., Sucar, et al. (2016). Stress modelling and prediction in presence of scarce data. *Journal of Biomedical Informatics, 63*, 344–356. https://doi.org/10.1016/j.jbi.2016.08.023

Ordem dos Psicólogos Portugueses, (2023). O Custo do stress e dos problemas de saúde psicológica no trabalho em Portugal, Contributo OPP. https://www.ordemdospsicologos.pt/pt/noticia/4466, last accessed 27 Feb 2023

Panicker, S. S., & Gayathri, P. (2019). A survey of machine learning techniques in physiology based mental stress detection systems. *Biocybernetics and Biomedical Engineering, 39*(2), 444–469. https://doi.org/10.1016/j.bbe.2019.01.004

Park, J., Kim, J., Kim, S. P. (2018). Prediction of daily mental stress levels using a wearable photoplethysmography sensor. In TENCON IEEE Region 10 Conference, 1899-1902. https://doi.org/10.1109/TENCON.2018.8650109

Reijmerink, I., van der Laan, M., & Cnossen, F. (2020). Heart rate variability as a measure of mental stress in surgery: A systematic review. *Int. Arch. Occup. Environ. Health, 25*, 1–17. https://doi.org/10.1007/s00420-020-01525-6

Sabour, R. M., Benezeth, Y., De Oliveira, P., et al. (2021). UBFC-Phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3056960

Seiler, A., Fagundes C.P., Christian L. M. (2020). The impact of everyday stressors on the immune system and health. In Stress challenges and immunity in space, 71-92. Springer, Cham. https://doi.org/10.1007/978-3-030-16996-1_6

Tran, C. T., Zhang, M., Andreae, P., et al. (2017). Bagging and feature selection for classification with incomplete data. In Applications of Evolutionary Computation: 20th European Conference, Evo Applications, Amsterdam, Springer International Publishing. https://doi.org/10.1007/978-3-319-55849-3_31

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning, 109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6

World Health Organization (WHO), (2023). Occupational health: stress at the workplace, https://www.who.int/news-room/questions-and-answers/item/ccupational-health-stress-at-the-workplace, last accessed 5 Mar 2023

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international Conference on the Practical Applications of Knowledge Discovery and Data Mining, Vol. 1, 29-39. http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf