

Project Acronym: STAR
Grant Agreement number: 956573 (H2020-ICT-2020-1 – Research and Innovation Action)
Project Full Title: Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines
Project Coordinator: Netcompany-Intrasoft



Funded by the Horizon 2020
Framework Programme of the
European Union

DELIVERABLE

D5.4: Digital Twins for Security and Safety – Final version

Dissemination level	PU -Public
Type of Document	Demonstrator
Contractual date of delivery	31/03/2023
Deliverable Leader	SUPSI
Status - version, date	Final - V1.0, 24/07/2023
WP / Task responsible	WP5
Keywords:	Human digital models, human digital twin, Digital twin

Executive Summary

This document provides an overview of the activities and the results achieved within M16-M27 of Task 5.2 - Digital Twins for Worker Safety. The goal of T5.2 is to develop solutions for human-machine collaboration, considering, understanding, and anticipating the humans in the production system.

The main activities carried out during T5.2 early stage are 1) data collection, analysis, and pre-processing for fatigue exertion estimation; 2) AI models development for fatigue exertion estimation 3) development of experiments to support the Philips pilot and add the mental stress estimation as part of the Fatigue Monitoring System functions; 4) Prototype implementation.

SUPSI led the task, providing its expertise to design and organise the data collection sessions, as well as the design and implementation of different machine learning models to predict the fatigue exertion and mental stress of workers. RuG supported the data analysis activity by checking and identifying issues with the data, also bringing valuable insights to the data analysis phase with dedicated analyses on time series data.

This document is an accompanying report of D5.4 (demonstrator).

Deliverable Leaders:	SUPSI
Contributors:	RuG
Reviewers:	Sungho Suh (DFKI), Cosmin-Septimiu Nechifor (SIE)
Approved by:	Charalampos Ipektsidis (INTRA)

Document History			
Version	Date	Contributor(s)	Description
V0.1	03/04/2023	SUPSI	Table of contents
V0.2	31/04/2023	SUPSI	First draft (section 1-2)
V0.3	19/05/2023	SUPSI	Second draft (section 3)
V0.4	30/06/2023	SUPSI	Section 2 completed, document refactoring, added section 4
V0.5	14/07/2023	SUPSI	Section 3 finalised; comments from reviewers implemented
V0.6	18/07/2023	RuG	Add sections 2.8 and 3.5
V0.7	18/07/2023	SUPSI	Document finalised
V1.0	24/07/2023	INTRA	QA and creation of the final submitted version

Table of Contents

- EXECUTIVE SUMMARY 2**
- TABLE OF CONTENTS..... 4**
- TABLE OF FIGURES..... 6**
- LIST OF TABLES..... 8**
- DEFINITIONS, ACRONYMS AND ABBREVIATIONS 9**
- 1 INTRODUCTION..... 10**
 - 1.1 PROBLEM DESCRIPTION AND MOTIVATION11
- 2 AI MODELS FOR FATIGUE EXERTION PREDICTION 12**
 - 2.1 DATA COLLECTION12
 - 2.2 DYNAMIC DATA PRE-PROCESSING13
 - 2.2.1 *Data cleaning*.....13
 - 2.2.2 *Features extraction*.....15
 - 2.3 LABELS OF PERCEIVED PHYSICAL FATIGUE EXERTION16
 - 2.4 STATIC DATA FROM THE QUESTIONNAIRE.....17
 - 2.5 DATA ANALYSIS AND FEATURE SELECTION19
 - 2.6 MODEL 1: RANDOM FOREST.....24
 - 2.6.1 *Training using all features*.....24
 - 2.6.2 *Training excluding accelerometer data*.....26
 - 2.6.3 *Training using only heart rate, skin temperature and galvanic skin response*.....28
 - 2.6.4 *Training after dimensionality reduction*30
 - 2.7 MODEL 2: FEED FORWARD NEURAL NETWORK.....31
 - 2.7.1 *Results with the first version of the network*.....33
 - 2.7.2 *Results with the second version of the network*35
 - 2.8 FURTHER FATIGUE DETECTION AND PREDICTION SCENARIOS36
 - 2.8.1 *Clustering and Classification*37
 - 2.8.2 *Time Series processing*38
 - 2.9 SUMMARY39
- 3 AI MODELS FOR MENTAL STRESS PREDICTION 41**
 - 3.1 DATA COLLECTION41
 - 3.2 DYNAMIC DATA PRE-PROCESSING42
 - 3.2.1 *Data cleaning*.....42
 - 3.2.2 *Features extraction*.....43
 - 3.3 DATA FROM COGNITIVE DEMANDING TESTS43
 - 3.4 STATIC DATA FROM THE QUESTIONNAIRE.....44
 - 3.5 DATA QUALITY ASSESSMENT45
 - 3.6 DATA ANALYSIS AND FEATURE SELECTION45
 - 3.7 MODEL 1: RANDOM FOREST.....49
 - 3.7.1 *Training excluding accelerometer, gyroscope, and magnetometer*.....50
 - 3.7.2 *Training by grouping by user*.....51
 - 3.8 MODEL 2: FEED FORWARD NEURAL NETWORK.....53
 - 3.8.1 *Results with the first version of the network*.....54
 - 3.8.2 *Results with the second version of the network*55
 - 3.9 SUMMARY57
- 4 FAMS MODULE IMPROVEMENTS 58**
- 5 CONCLUSIONS..... 59**

REFERENCES 61

Table of Figures

FIGURE 1: FAMS - HDT CORE INFRASTRUCTURE INTEGRATION.....10

FIGURE 2: FILTERING SIGNAL FROM THE ACCELEROMETER14

FIGURE 3: HEART RATE DISTRIBUTION USING RAW COUNTING15

FIGURE 4: HEART RATE DISTRIBUTION USING DENSITY15

FIGURE 5: GALVANIC SKIN RESPONSE DISTRIBUTION USING RAW COUNTING15

FIGURE 6: GALVANIC SKIN RESPONSE DISTRIBUTION USING DENSITY15

FIGURE 7: SKIN TEMPERATURE DISTRIBUTION USING RAW COUNTING15

FIGURE 8: SKIN TEMPERATURE DISTRIBUTION USING DENSITY15

FIGURE 9: CORRELATION MATRIX SHOWING THE CORRELATION BETWEEN NUMERICAL FEATURES20

FIGURE 10: FEATURE CORRELATION WITH THE TARGET VARIABLE21

FIGURE 11: FEATURE CORRELATION WITH THE TARGET VARIABLE IN ABSOLUTE VALUE22

FIGURE 12: OBSERVATIONS AVAILABLE FOR A GIVEN FATIGUE EXERTION LABEL, GROUPED BY WORKER.....23

FIGURE 13: OBSERVATIONS AVAILABLE FOR A GIVEN FATIGUE EXERTION LABEL, GROUPED BY COMPANY23

FIGURE 14: OBSERVATIONS AVAILABLE FOR A GIVEN FATIGUE EXERTION LABEL, GROUPED BY SESSION.....24

FIGURE 15: NUMBER OF OBSERVATIONS AVAILABLE IN TRAINING AND TEST DATASETS, FOR EACH EXERTION LABEL ...25

FIGURE 16: CLASSIFICATION REPORT OF THE OBTAINED MODEL25

FIGURE 17: CONFUSION MATRIX OF THE OBTAINED MODEL26

FIGURE 18: FEATURE IMPORTANCE ACCORDING TO THE OBTAINED MODEL26

FIGURE 19: NUMBER OF OBSERVATIONS AVAILABLE IN TRAINING AND TEST DATASETS, FOR EACH EXERTION LABEL ...27

FIGURE 20: CLASSIFICATION REPORT OF THE OBTAINED MODEL27

FIGURE 21: CONFUSION MATRIX OF THE OBTAINED MODEL27

FIGURE 22: FEATURE IMPORTANCE ACCORDING TO THE OBTAINED MODEL28

FIGURE 23: NUMBER OF OBSERVATIONS AVAILABLE IN TRAINING AND TEST DATASETS, FOR EACH EXERTION LABEL ...29

FIGURE 24: CLASSIFICATION REPORT OF THE OBTAINED MODEL29

FIGURE 25: CONFUSION MATRIX OF THE OBTAINED MODEL29

FIGURE 26: FEATURE IMPORTANCE ACCORDING TO THE OBTAINED MODEL29

FIGURE 27: VARIANCE EXPLAINED AFTER FEATURE REDUCTION30

FIGURE 28: NUMBER OF OBSERVATIONS AVAILABLE IN TRAINING AND TEST DATASETS, FOR EACH EXERTION LABEL ...30

FIGURE 29: CLASSIFICATION REPORT OF THE OBTAINED MODEL31

FIGURE 30: CONFUSION MATRIX OF THE OBTAINED MODEL31

FIGURE 31: THE ARCHITECTURE OF THE FEED FORWARD NEURAL NETWORK FOR PHYSICAL EXERTION PREDICTION. .32

FIGURE 32: NUMBER OF OBSERVATIONS AVAILABLE IN TRAINING, VALIDATION, AND TEST DATASETS, FOR EACH EXERTION LABEL33

FIGURE 33: TRAINING LOSS VS VALIDATION LOSS34

FIGURE 34: CONFUSION MATRIX OF THE BEST MODEL SELECTED THROUGH LOSS34

FIGURE 35: TRAINING ACCURACY VS VALIDATION ACCURACY34

FIGURE 36: CONFUSION MATRIX OF THE BEST MODEL SELECTED THROUGH ACCURACY35

FIGURE 37: TRAINING LOSS VS VALIDATION LOSS35

FIGURE 38: CONFUSION MATRIX OF THE BEST MODEL SELECTED THROUGH LOSS36

FIGURE 39: TRAINING ACCURACY VS VALIDATION ACCURACY36

FIGURE 40: CONFUSION MATRIX OF THE BEST MODEL SELECTED THROUGH ACCURACY36

FIGURE 41: THE KNIME WORKFLOW FOR FATIGUE CLASSIFICATION.38

FIGURE 42: FATIGUE ESTIMATION AND PREDICTION EXAMPLE WORKFLOW39

FIGURE 43: AN EXAMPLE FROM THE STROOP TEST. THE RIGHT ANSWER IS GREEN (KEY DOWN).42

FIGURE 44: A SHORT SEQUENCE FROM THE 3-BACK TEST. RIGHT ANSWERS ARE MARKED BY A CHECKMARK.....42

FIGURE 45: CORRELATION MATRIX SHOWING THE CORRELATION BETWEEN NUMERICAL FEATURES46

FIGURE 46: FEATURE CORRELATION WITH THE TARGET VARIABLE47

FIGURE 47: FEATURE CORRELATION WITH THE TARGET VARIABLE (ABSOLUTE VALUE)48

FIGURE 48: OBSERVATIONS AVAILABLE FOR A GIVEN MENTAL STRESS LABEL, GROUPED BY USER49

FIGURE 49: OBSERVATIONS AVAILABLE FOR A GIVEN MENTAL STRESS LABEL, GROUPED BY SESSION49

FIGURE 50: NUMBER OF OBSERVATIONS AVAILABLE IN TRAINING AND TEST DATASETS, FOR EACH CLASS.....50

FIGURE 51: CLASSIFICATION REPORT OF THE OBTAINED MODEL.....	50
FIGURE 52: CONFUSION MATRIX OF THE OBTAINED MODEL	50
FIGURE 53: FEATURE IMPORTANCE ACCORDING TO THE OBTAINED MODEL	51
FIGURE 54: COMPARISON BETWEEN THE PROPOSED ML MODEL AND THE BASELINE MODEL.....	51
FIGURE 55: NUMBER OF OBSERVATIONS AVAILABLE IN TRAINING AND TEST DATASETS, FOR EACH CLASS.....	52
FIGURE 56: CLASSIFICATION REPORT OF THE OBTAINED MODEL.....	52
FIGURE 57: CONFUSION MATRIX OF THE OBTAINED MODEL	52
FIGURE 58: FEATURE IMPORTANCE ACCORDING TO THE OBTAINED MODEL	53
FIGURE 59: COMPARISON BETWEEN THE PROPOSED ML MODEL AND THE BASELINE MODEL.....	53
FIGURE 60: NUMBER OF OBSERVATIONS AVAILABLE IN TRAINING AND TEST DATASETS, FOR EACH CLASS.....	53
FIGURE 61: TRAINING LOSS VS VALIDATION LOSS	54
FIGURE 62: TRAINING ACCURACY VS VALIDATION ACCURACY	54
FIGURE 63: CONFUSION MATRIX OF THE BEST MODEL SELECTED THROUGH LOSS AND ACCURACY	55
FIGURE 64: COMPARISON BETWEEN THE PROPOSED NN (v1 – BEST LOSS/ACCURACY) AND THE BASELINE MODEL.	55
FIGURE 65: TRAINING LOSS VS VALIDATION LOSS	56
FIGURE 66: TRAINING ACCURACY VS VALIDATION ACCURACY	56
FIGURE 67: CONFUSION MATRIX OF THE BEST MODEL SELECTED THROUGH LOSS.....	56
FIGURE 68: COMPARISON BETWEEN THE PROPOSED NN (v2 – BEST LOSS) AND THE BASELINE MODEL.	57

List of Tables

TABLE 1: LIST OF FEATURES DERIVED FROM DYNAMIC DATA.....	16
TABLE 2: THE BORG CR10 SCALE USED FOR EXERTION LABELLING	16
TABLE 3: LIST OF FEATURES AVAILABLE AS STATIC DATA.....	17
TABLE 4: QUESTIONS FROM THE QUESTIONNAIRE USED TO COLLECT STATIC DATA	18
TABLE 5: FATIGUE AND ACTIVITY CLUSTERING.....	37
TABLE 6: FATIGUE CLASSIFICATION	38
TABLE 7: SINGLE ATTRIBUTE FATIGUE CLASSIFICATION	38
TABLE 8: PERFORMANCE OF MODELS IN VARIOUS CONFIGURATIONS	40
TABLE 9: FEATURES COLLECTED DURING THE STROOP TEST	43
TABLE 10: FEATURES COLLECTED DURING THE 3-BACK TEST	44
TABLE 11: FEATURES COLLECTED WITH THE NASA TLX QUESTIONNAIRE	44
TABLE 12: LIST OF FEATURES AVAILABLE AS STATIC DATA	45
TABLE 13: PERFORMANCE OF MODELS IN VARIOUS CONFIGURATIONS.....	57

Definitions, Acronyms and Abbreviations

Acronym/ Abbreviation	Title
AI	Artificial Intelligence
CPS	Cyber-Physical Systems
HDT	Human Digital Twin
MSE	Mean Squared Error
PCA	Principal Component Analysis
WP	Work Package

1 Introduction

This document describes the work carried out in Task 5.2, which targets STAR’s **Objective 4 - Human-centred simulations and digital twins for safe AI systems in manufacturing**. The goal of the task is to facilitate human-machine collaboration, considering, understanding, and anticipating the humans in the production system. The task develops an Artificial Intelligence (AI) module capable to estimate the perceived fatigue exertion of the workers, by exploiting quasi-static data about the workers (e.g., age, height, weight), as well as dynamic data collected by wearable devices (e.g., heart rate).

The document sums up the activities carried out in the last 12 months of the task and details the processes and methodologies utilised to develop machine learning and deep learning models. To better support the project’s use cases, new models have been developed to improve the prediction of the physical fatigue experienced by workers in manufacturing, as well as to detect their mental stress. The new models were constructed still using dynamic data collected from wearable devices, as well as static data from questionnaires. New features have been considered, e.g., the kind of task performed by the workers. Also, a new application has been developed to study the mental stress of workers. Thanks to this new application, new labelled datasets have been collected, which are crucial for training the new models.

The new models are available to the Fatigue Monitoring System (FaMS), the module devised within the STAR project for mental stress and physical exertion prediction. The FaMS has been further improved since its first release, to better integrate it with the Human Digital Twin (HDT) Core Infrastructure developed as part of T5.1, as well as to make it easier to configure (e.g., switch between available models). A new user manual has been published to help adopters in using FaMS.

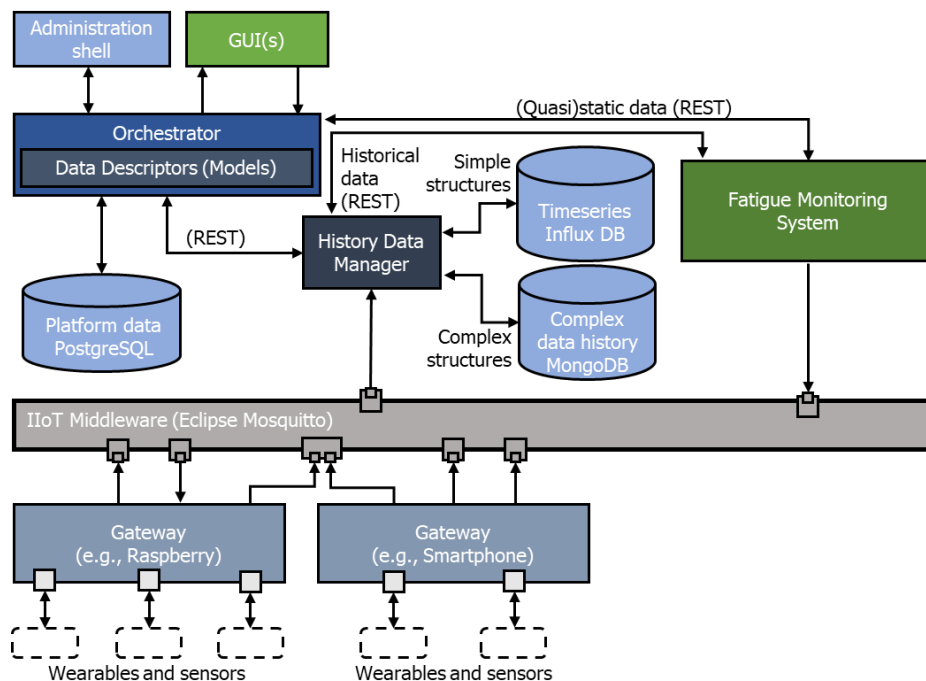


Figure 1: FaMS - HDT Core Infrastructure integration

1.1 Problem description and motivation

In the context of STAR, the Philips pilot focuses on the data quality inspection, which is manually performed by operators labelling images of products. For each image, the operator can choose the right label to describe the quality of the product. This activity is crucial to ensure the delivery of good products in the factory. Normally, visual quality inspection systems are trained based on extensive datasets and can be easily optimised due to the mass-production of products. However, when moving to lower-volume production, these extensive datasets are often not available. Therefore, flexible visual quality inspection systems that can be trained based on small, incomplete datasets are needed. The goal of the use-case is to implement a system that will make setting up automated quality inspections easier and faster by applying Active Learning. The resulting system can support the automated quality control for a new product easily (meaning with little to no data) where the production personnel are able to check the products (OK/NOK) and transfer that knowledge to the quality inspection algorithm by means of Active Learning. By relating this input created by the operators to HDT tools like fatigue monitoring, it is possible to ensure correct inputs and decrease the amount of wrong labelling, while also ensuring the physical and mental well-being of the operators doing the labelling.

Starting from the previous model for fatigue detection, a new fatigue detection model has been developed and trained on a new dataset, featuring data collected from 5 different companies (from different sectors, namely injection moulding industry, aerospace, wood industry and food industry). The new model generalises better than the previous one, thanks to the higher diversity available in the new dataset, and thus it can be easily brought to different contexts.

However, the data labelling activity is very subject to operators' judgement, which can be negatively impacted by their mental condition. A stressed operator may indeed assign a wrong label, or even don't recognise some defects from the image. To this end, as part of T5.2., the FaMS module has been extended with a new ML model devoted to the early detection of mental stress in operators. Thanks to this new model, a confidence score can be assigned to each operator evaluations, following diverse logics; for instance, a low confidence score can be assigned to those evaluations made by very stressed operators.

Developing such a new model required to collect first new datasets, possibly while monitoring operators both in relaxed and stressed conditions. Thus, SUPSI designed an experiment where operators are asked to relax for a certain amount of time (e.g., 3 minutes), then to complete some mental demanding tasks (with short breaks in between), and finally to rest again. Monitoring the operators during the activity/rest periods, the model can estimate when they are going to start feeling stressed; the new evidence is made available to the quality labelling application, which can define a proper strategy for dealing with the next labels provided by the operator (e.g., to discard them, or to double-check the same image by proposing it to a different operator).

2 AI Models for Fatigue Exertion Prediction

This section provides details about the process put in place to develop two machine learning models capable of predicting the instantaneous perceived physical fatigue exertion of workers.

The first version of FaMS embedded a multiclass classifier, capable of distinguishing between 10 levels of fatigue exertion (representing the range 1-10). However, the limited availability of fatigue exertion data hindered the training phase, with the model underperforming on certain classes.

For the new version, Random Forest classification algorithm and a feed-forward neural network have been considered.¹ The models were trained using all available features, as well as selected features only. Additionally, a data partitioning method specific for time series was exploited to create training, validation, and testing datasets.

Three main evaluation metrics were used to assess the model results: accuracy, mean squared error (MSE), and the confusion matrix. However, when dealing with a classification task, accuracy alone is not the most informative metric. Accuracy measures the percentage of correctly classified instances out of the total number of instances, but it does not provide insights into the model's performance for individual classes. Class imbalance or the difficulty of predicting certain classes can skew the accuracy results and provide an incomplete picture of the model's performance. To address this limitation, the Macro F1 score is exploited (i.e., the unweighted mean of the F1 scores calculated per class), together with the confusion matrix, a comprehensive view of the model's performance enabling the easy identification of classes harder to predict accurately and the downstream targeted improvements. Additionally, MSE, commonly used for regression models, is applied in this classification task to evaluate model performance. MSE measures the average of the squares of the errors between predicted and actual values, assigning greater weight to larger errors. This sensitivity to significant errors allows for a more accurate assessment of the model's performance. In the context of predicting fatigue exertion, the aim is to penalise the model more for predicting a significantly different level of fatigue than what is perceived by the worker, compared to a prediction that is only slightly off.

The subsequent sections discuss all the relevant steps needed for training the models, namely data collection, data pre-processing and feature selection, model training and validation.

2.1 Data collection

To train the physical exertion prediction model, the available dataset (i.e., the one already used to train the first FaMS version) has been extended so that to contain data from 5 data collection sessions, taking place in five companies. To augment the diversity within the dataset, data have been collected from different sessions involving operators dealing with different type of tasks:

- Company 1 (injection moulding industry): small parts assembly.
- Company 2 (aerospace sector): aircraft internal parts assembly.
- Company 3 (wood industry): cutting, assembly, and material handling.

¹ A feed forward neural network has been preferred to convolutional neural networks (CNNs), which are more indicated for computer vision problem thanks to their capability of taking into account the spatial structure of data to capture spatially local input patterns.

- Company 4 (food industry): manual ham greasing.
- Company 5 (injection moulding industry): screwdriving and material handling.

During each session, data were collected using the same three distinct sources exploited for the initial dataset, i.e., sensor data, quasi-static data from questionnaires, and labels indicating the worker’s perceived exertion level (using a 10-value scale derived from the Borg CR10 scale). The number of involved workers per company ranges from 1 to 7.

For sensor data, different combinations of wearable sensors have been used to collect the relevant metrics (i.e., skin temperature, galvanic skin response, heart rate, and 3-axis acceleration): Empatica E4 (wristband), Huawei Smartwatch 2, and Polar H10 (chest band). The metrics have been collected with the following (average) sampling frequencies:

- Heart rate: 1 Hz
- Galvanic Skin Response: 4 Hz
- Skin temperature: 4 Hz
- Accelerometer: 32.25 Hz

2.2 Dynamic data pre-processing

2.2.1 Data cleaning

The initial phase of this project involved preparing dynamic data obtained from sensors contained in wearable devices worn by the operators during the experiment. These data consist of six metrics that change dynamically in the short term: skin temperature, galvanic skin response, heart rate, and accelerations on the three axes (x, y, and z).

Initially, standard cleaning procedures have been applied (e.g., duplicates removal, missing data filtering). Subsequently, to facilitate the downstream processing of the data, an additional parameter representing the concept of “session” has been included. This concept was deemed crucial due to the significant variations in worker activities and working conditions throughout their working shifts. Within a defined session (i.e., a timeframe), these activities and conditions are assumed to remain consistent. Furthermore, sessions can be considered as independent from each other, meaning they have no influence on one another. To identify sessions, data from sensors have been analysed looking for timeframe longer than 3 minutes where no new data came. These timeframes with no data represent the points where to split a session from the following one.

Given that sensors data coming from different devices will likely have different timestamps, simply joining values from different tables (e.g., heart rate and gsr) produces many missing values. To this end, missing values have been filled with linear interpolation. In case of very noisy signals (e.g., accelerometer), a further filtering function has been applied to smoothing the signal (e.g., savgol filter). An example of filtered signal is available in Figure 2.

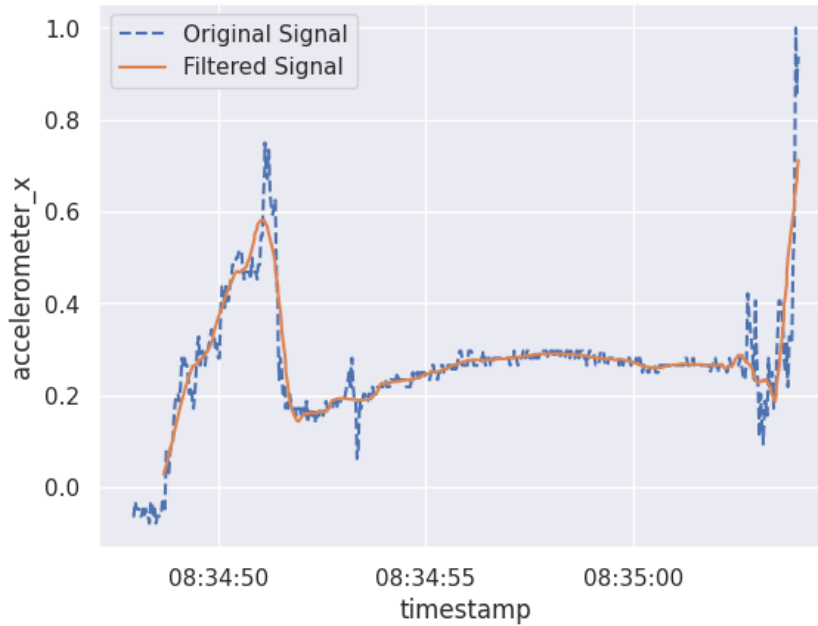


Figure 2: Filtering signal from the accelerometer

Figures from Figure 3 to Figure 8 plot the distributions of three parameters, namely the heart rate, skin temperature, and galvanic skin response. Given that the distributions among the various companies are not the same, outliers have been removed from the dataset, based on the 25th and 75th percentiles, leading to a removal of about 13.33% observations.



Figure 3: Heart rate distribution using raw counting

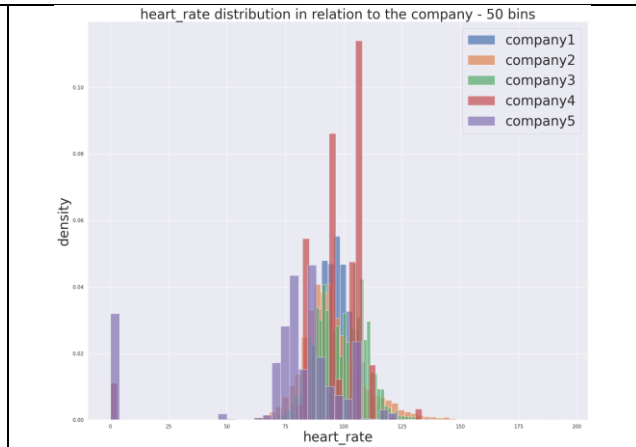


Figure 4: Heart rate distribution using density

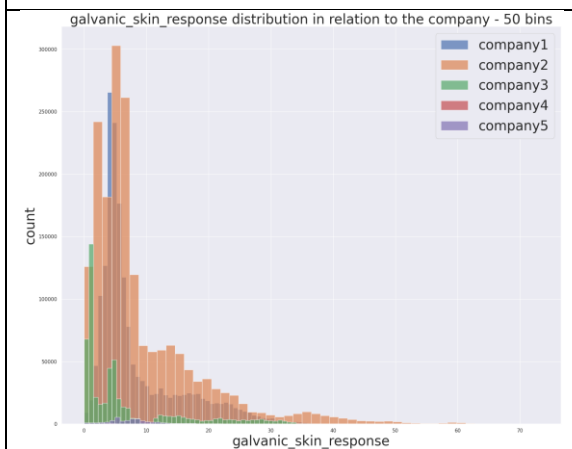


Figure 5: Galvanic skin response distribution using raw counting

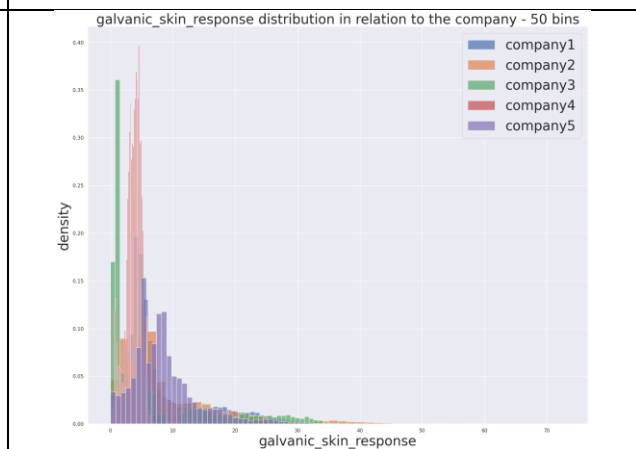


Figure 6: Galvanic skin response distribution using density

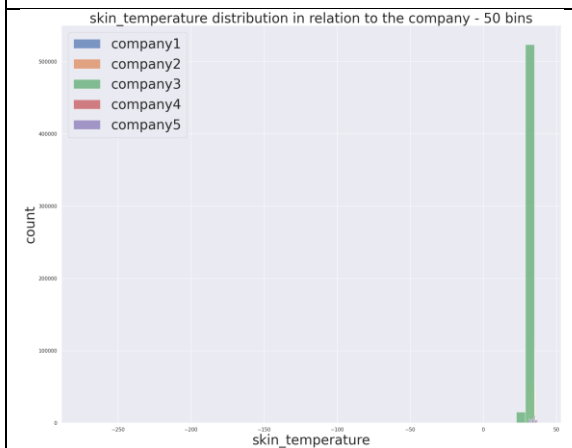


Figure 7: Skin temperature distribution using raw counting

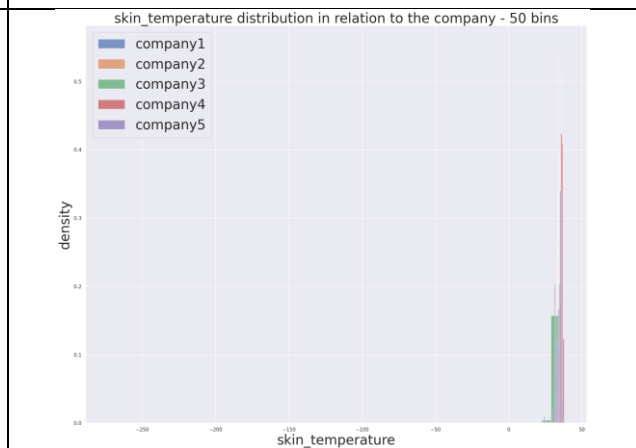


Figure 8: Skin temperature distribution using density

2.2.2 Features extraction

To augment the features available in the dataset, new ones have been automatically derived. The mean deviation was the first feature to be extracted from heart rate, skin temperature, and galvanic skin response. In addition, for these three metrics, it was decided to perform 60-

second windowing with overlapping, where in each window the maximum, minimum, average, and standard deviation values were obtained. Total acceleration was then calculated from the three components (x, y, z). Taking these four metrics into consideration, a 5-second windowing with overlapping has then been applied, where in each window the mean, standard deviation and integral were derived.

Table 1 presents the features derived from the dynamic data.

Table 1: List of features derived from dynamic data

Feature	Unit	Derived features
Accelerometer	g	$\mu_{acc_i}, i \in \{x, y, z\}$ mean value $\sigma_{acc_i}, i \in \{x, y, z\}$ standard deviation $\int_{acc_i}, i \in \{x, y, z\}$ integral $\sum \int_{acc_i}, i \in \{x, y, z\}$ sum of integrals over all axes
Galvanic skin response	Microsiemens (μS)	min_{gsr} minimum value max_{gsr} maximum value μ_{gsr} mean σ_{gsr} standard deviation MD_{gsr} mean deviation
Heart rate	bpm	min_{hr} minimum value max_{hr} maximum value μ_{hr} mean σ_{hr} standard deviation MD_{hr} mean deviation
Skin temperature	$^{\circ}C$	min_{st} minimum value max_{st} maximum value μ_{st} mean σ_{st} standard deviation MD_{st} mean deviation

2.3 Labels of perceived physical fatigue exertion

The second step was to merge the dynamic data with the fatigue exertion label data. The fatigue exertion label data were collected at a lower frequency than the sensors (during the experiment, the operator was asked every 3-5-10 minutes to provide the perceived level of exertion, based on the Borg CR10 scale, reported in Table 2); for this reason, the labels were have been matched to the sensor data with the closest timestamp (ensuring that the closest timestamp was no more than 3 minutes away). Also, missing values have been replaced with linear interpolation, assuming the fatigue exertion has a linear trend. This operation significantly increased the number of available labels for fatigue exertion, thereby preserving sensor data.

Table 2 presents the features derived from the physical fatigue exertion data.

Table 2: The Borg CR10 scale used for exertion labelling

Value	Level of Exertion	Description
1.0	Very Weak	
2.0	Weak	Light exercise
3.0	Moderate	

4.0		
5.0	Strong Heavy	Exercise
6.0		
7.0	Very Strong	
8.0		
9.0		
10.0	Extremely Strong	“Maximal” affordable exercise

2.4 Static data from the questionnaire

Static data are collected via questionnaire filled out by operators before or after conducting the experiment. These data include biographical information, health status, and work experience.

Categorical variables, e.g., “sex” and “smoker”, have been binarised, while variables with ordinal values, such as “training intensity” and “job experience”, have been converted to ordinal variables. Finally, three additional features were calculated for the static data: body mass index, waist-to-height ratio, and ponderal index.

Table 3 contains the list of parameters available as static data (either collected through the questionnaire or automatically computed).

Table 3: List of features available as static data

#	Field	Unit	Acquisition method	Notes
1	date	ms	Read from system	Acquisition date
2	worker		Manually assigned	Hashed UUID
3	weight	kg	Measured	
4	height	cm	Measured	
5	waistcircumference	cm	Measured	
6	gripstrength_left_avg	kg	Measured	
7	gripstrength_left_std	kg	Measured	
8	gripstrength_right_avg	kg	Measured	
9	gripstrength_right_std	kg	Measured	
10	sex		Questionnaire	male / female
11	smoker		Questionnaire	yes / no
12	weekly_trainings		Questionnaire	
13	training_intensity		Questionnaire	
14	job_experience		Questionnaire	
15	heavy_monotonous_work		Questionnaire	
16	decrease_pain		Questionnaire	
17	anxiety_past_week		Questionnaire	
18	depressed_past_week		Questionnaire	
19	job_satisfaction		Questionnaire	
20	physical_activity_pain		Questionnaire	
21	stop_until_pain_decreases		Questionnaire	
22	no_work_with_pain		Questionnaire	
23	light_work_hour		Questionnaire	

24	walk_hour		Questionnaire	
25	household		Questionnaire	
26	weekly_shopping		Questionnaire	
27	night_sleep		Questionnaire	
28	age		Computed	age = current year – year of birth (removed)
29	working_age		Computed	working age = current year – hiring year (removed)
30	max_hr	bpm	Computed	max _{hr} = 220 – #28
31	ompq		Computed	ompq = #15 + #17 + #18 + #20 + #21 + #22 + 10 * 7 – (#16 + #19 + #23 + #24 + #25 + #26 + #27)
32	body_mass_index	kg/m2	Computed	bmi = $\frac{\#3 \text{ [kg]}}{(\#4 / 100)^2 \text{ [m}^2\text{]}}$
33	Waist-to-height ratio		Computed	wsr = $\frac{\#5 \text{ [cm]}}{\#4 \text{ [cm]}}$
34	Ponderal index	kg/m3	Computed	pi = $\frac{\#3 \text{ [kg]}}{(\#4 / 100)^3 \text{ [m}^3\text{]}}$
32	company		Manually assigned	Custom ID of the worker's company

Features from #17 to #29 have been collected by asking the questions available in Table 4.

Table 4: Questions from the questionnaire used to collect static data

Field	Question	Options
heavy_monotonous_work	Is your work heavy or monotonous?	0 = Not at all 10 = Extremely
decrease_pain	Based on all things you do to cope, or deal with your pain, on an average day, how much are you able to decrease it?	0 = Can't decrease it at all 10 = Can decrease it completely
anxiety_past_week	How tense or anxious have you felt in the past week?	0 = Absolutely calm and relaxed 10 = As tense and anxious as I've ever felt
depressed_past_week	How much have you been bothered by feeling depressed in the past week?	0 = Not at all 10 = Extremely
job_satisfaction	If you take into consideration your work routines, management, salary, promotion possibilities and work mates, how satisfied are you with your job?	0 = Not satisfied at all 10 = Completely satisfied
physical_activity_pain	Physical activity makes my pain worse.	0 = Completely disagree 10 = Completely agree
stop_until_pain_decreases	An increase in pain is an indication that I should stop what I'm doing until the pain decreases.	0 = Completely disagree 10 = Completely agree
no_work_with_pain	I should not do my normal work with my present pain.	0 = Completely disagree 10 = Completely agree

light_work_hour	I can do light work for an hour.	0 = Can't do it because of pain problem 10 = Can do it without pain being a problem
walk_hour	I can walk for an hour.	0 = Can't do it because of pain problem 10 = Can do it without pain being a problem
household	I can do ordinary household chores.	0 = Can't do it because of pain problem 10 = Can do it without pain being a problem
weekly_shopping	I can do the weekly shopping.	0 = Can't do it because of pain problem 10 = Can do it without pain being a problem
night_sleep	I can sleep at night.	0 = Can't do it because of pain problem 10 = Can do it without pain being a problem

2.5 Data analysis and feature selection

At this stage, our goal was to identify the most relevant features for predicting physical fatigue exertion. Additionally, we wanted to analyse the distribution of the fatigue exertion data to determine the best strategy for partitioning the dataset for the training, validation, and testing phases of the intended model. Initially, we attempted to identify highly correlated features and considered removing them.

In Figure 9, it is possible to see how the various features related to skin temperature, galvanic skin response, and heart rate data are highly correlated with each other. Consequently, for each of these metrics, it was decided to keep the mean, standard deviation, and mean deviation, and to remove the minimum and maximum values instead. In addition, "max_hr" was also removed as it was 100% correlated with "age" (indeed, the maximum heart rate is derived directly from the age, with the formula $220 - age$). From the correlation matrix, it is possible to see that other features are also highly correlated with each other, but it was decided not to discard them.

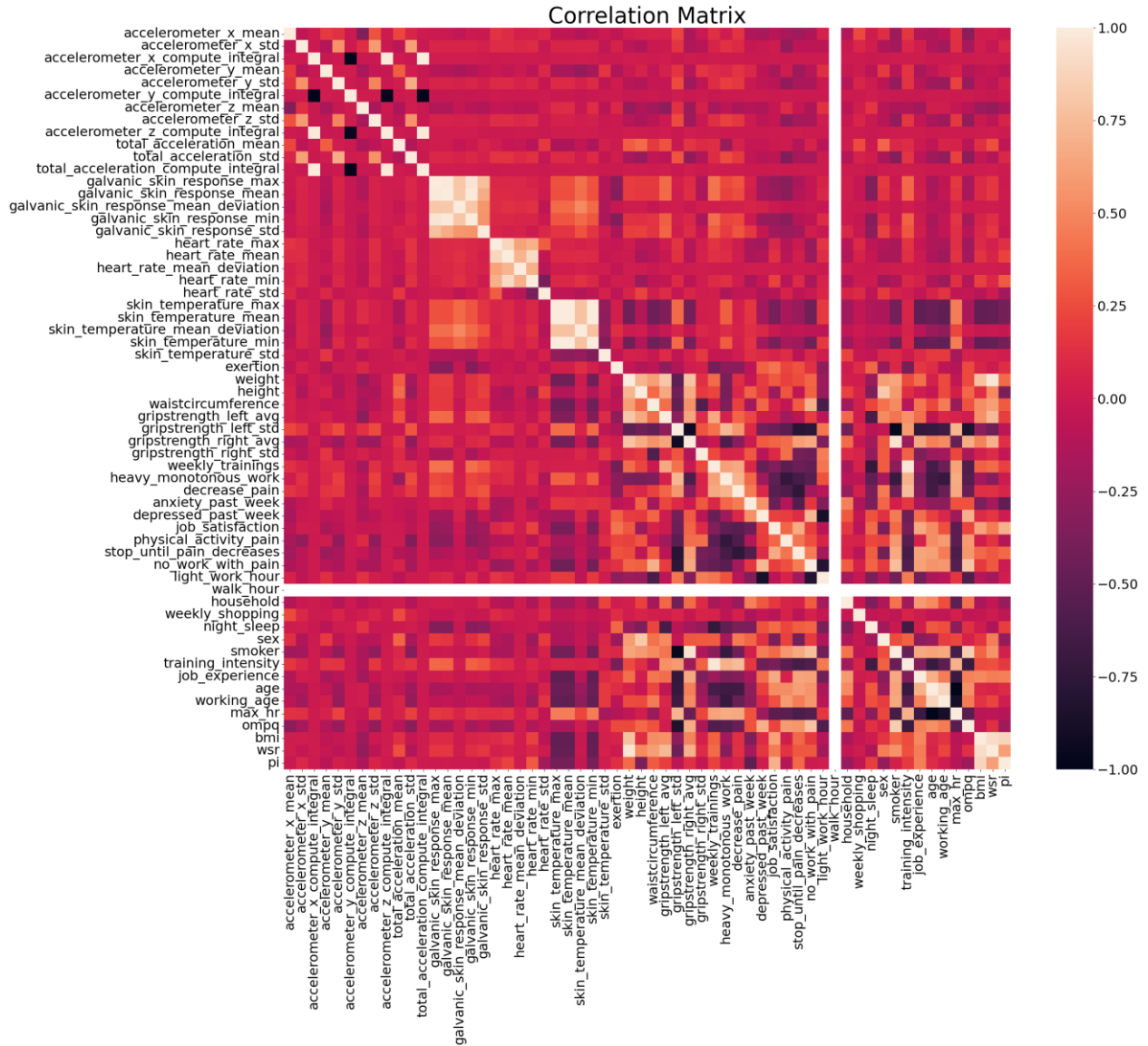


Figure 9: Correlation matrix showing the correlation between numerical features

The correlations between the features and the target variable "exertion" were then examined. From Figure 10 and Figure 11, it can be observed that the features that are most correlated are mainly related to static data. On the other hand, with regards to sensor data, the features that are most correlated with fatigue exertion are those related to the galvanic skin response.

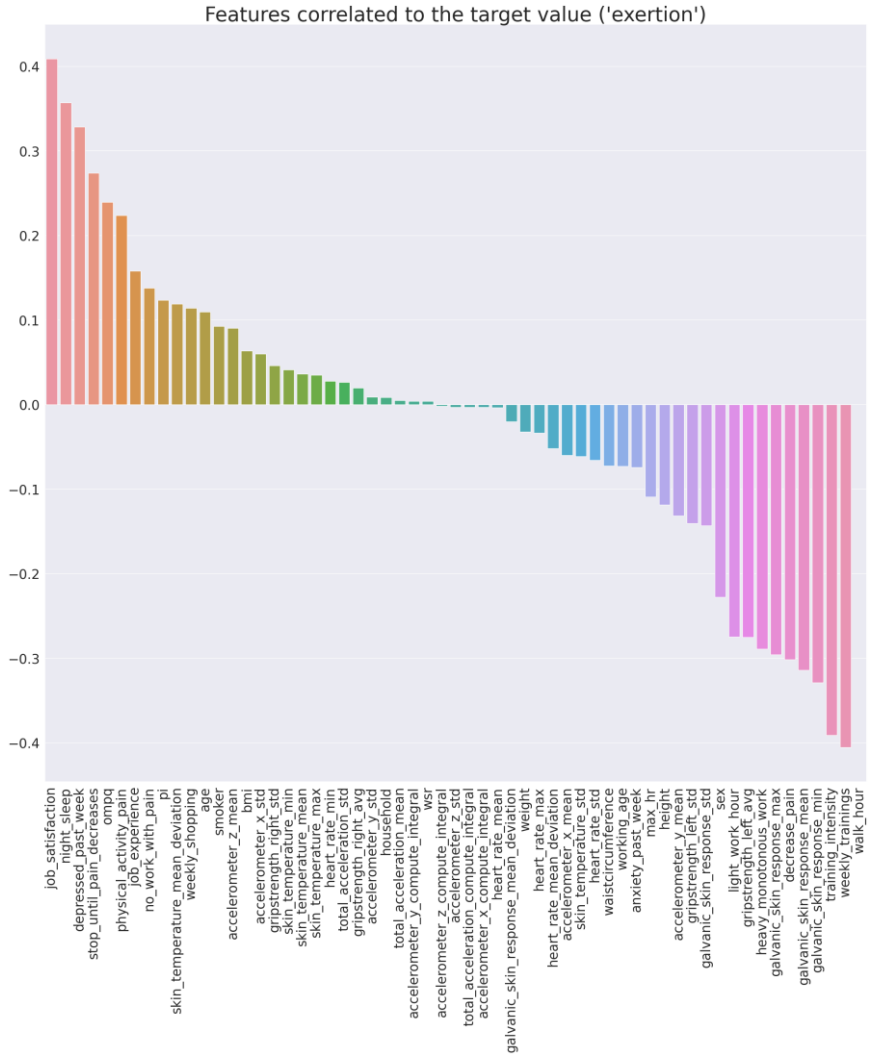


Figure 10: Feature correlation with the target variable

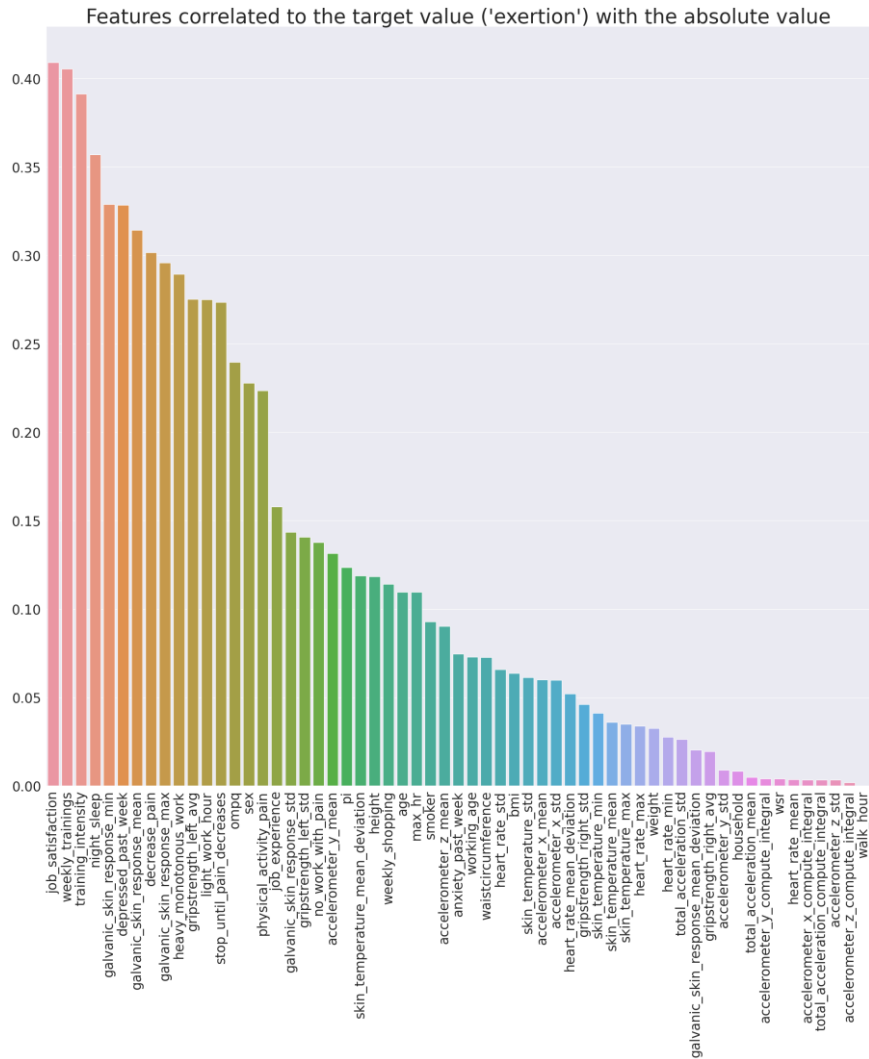


Figure 11: Feature correlation with the target variable in absolute value

Then, features with a variance below 0.05 have been removed (i.e., “walk_hour”, and “weekly_shopping”). For some features (e.g., “skin_temperature_std”, “total_acceleration_mean”, “wsr”), even if have a variance below the threshold, it has been decided to keep them, because from previous trials it seems that with certain configurations, models tend to use them to some extent.

Next, the distribution of fatigue exertion labels was analysed by worker, company, and session. Plots have been generated to visualise the number of observations available in the dataset, for each fatigue exertion label. Figures from Figure 12 to Figure 14 show the observations labelled with a certain fatigue label, grouped by worker (Figure 12), company (Figure 13),² and session (Figure 14). It is noticeable that while the range of fatigue exertion values is from 1 to 10, the collected data only encompasses values ranging from 1 to 8, mainly concentrated in the range from 3 to 7.

It can also be seen that the range of fatigue exertion is not evenly distributed among workers, companies, and sessions. Indeed, upon examining the lengths of the bars in the stacked plots, it becomes evident that certain workers, companies, or sessions exclusively occupy specific

² Company 3 was excluded from the dataset due to some misalignments between the dynamic and static data. The dataset will be fixed in future, so that to include also data from company 3 in the training phase.

ranges of fatigue exertion. Therefore, it was necessary to find an effective technique for dividing the data into training, validation, and testing sets. The following sections report on the various approaches tested.

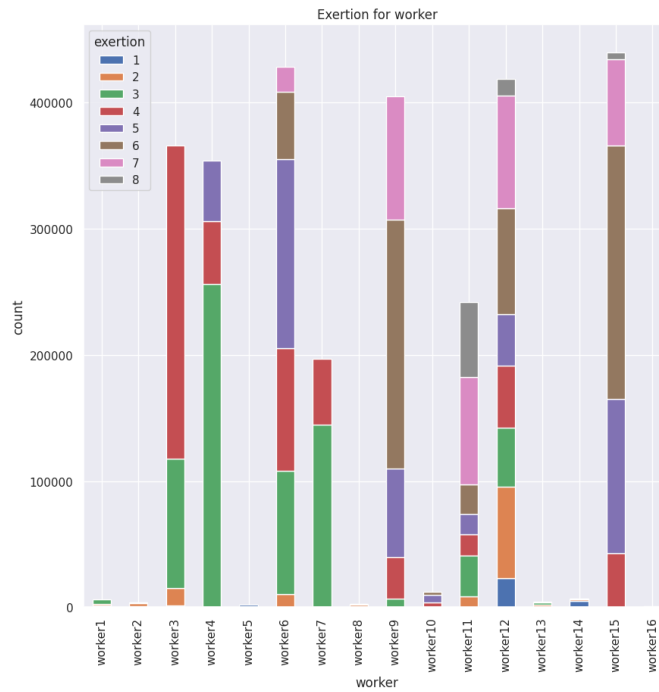


Figure 12: Observations available for a given fatigue exertion label, grouped by worker

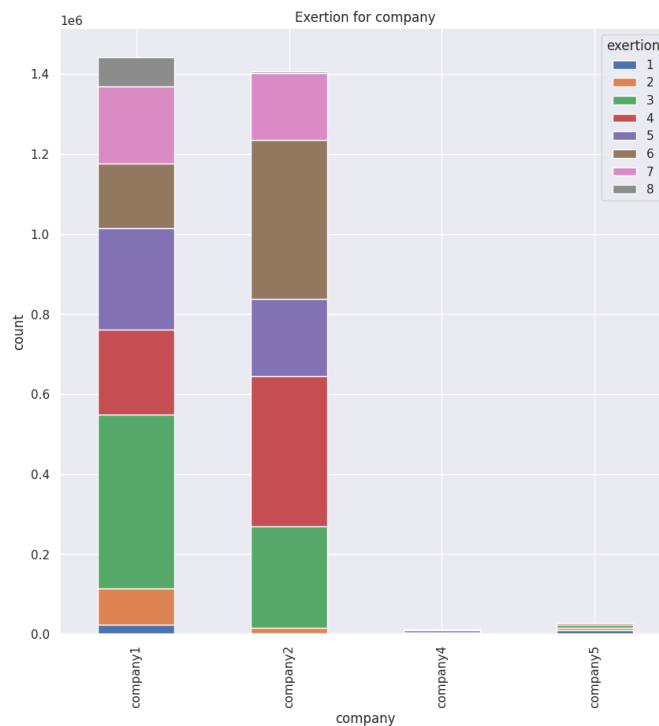


Figure 13: Observations available for a given fatigue exertion label, grouped by company

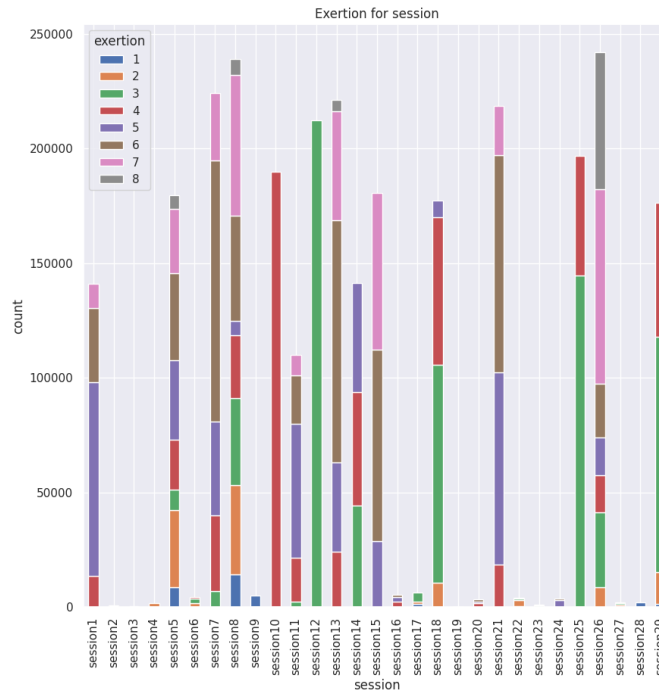


Figure 14: Observations available for a given fatigue exertion label, grouped by session

2.6 Model 1: Random Forest

The random and stratified data split is usually adopted when dividing the available dataset into training, validation, and test sets. However, this data splitting strategy is not considered optimal when dealing with time series data, as it can lead to inaccurate or misleading outcomes. Ignoring the temporal structure of the data in this partitioning approach may result in the model being trained and tested on data that does not accurately reflect the actual temporal trends. Additionally, there is a risk of including information from the training set in the test set due to the random partitioning. Thus, in this section, a more suitable approach for processing time series data is presented, which relies on a custom function developed for splitting the dataset into training and testing sets. The custom function is based on the following criterion: each identified session (see Figure 14) is further divided into four equal segments (i.e., segment containing the same number of observations); for each segment, 80% is used as training data, while the remaining 20% will serve as test set. This approach allows for less information to be shared between the training and test sets, while also ensuring that all fatigue exertion classes within the range of 1 to 8 are present in both sets.

2.6.1 Training using all features

Upon analysing the distribution of fatigue exertion in Figure 12, Figure 13, and Figure 14, it became apparent that certain classes were significantly more prevalent than others. When facing imbalanced classes, classifiers tend to exhibit bias toward the dominant class. This bias can result in accurate predictions for the dominant class while neglecting the minority classes. To address this issue, a Random Forest classifier was trained using the “balanced” mode. This approach automatically assigns weights to the classes during training, considering the frequency of each class in the dataset. Consequently, minority classes are assigned greater weights than dominant classes. This weighting scheme enables the model to prioritise the training of minority classes and counteracts the impact of class imbalance.

First, an attempt was made to utilise all available features to assess the model's performance in predicting fatigue exertion. This approach also allowed for the determination of feature importance rankings for fatigue exertion prediction according to the model. Consequently, this ranking aids in the selection of features based on their significance as determined by the model.

Figure 15 illustrates the number of observations available in training and test sets, for each exertion label (labels 0, 9, and 10 are missing because not represented in the original dataset; label 1 is also available in the test set, but the bar is not visible due to the plot visualisation).

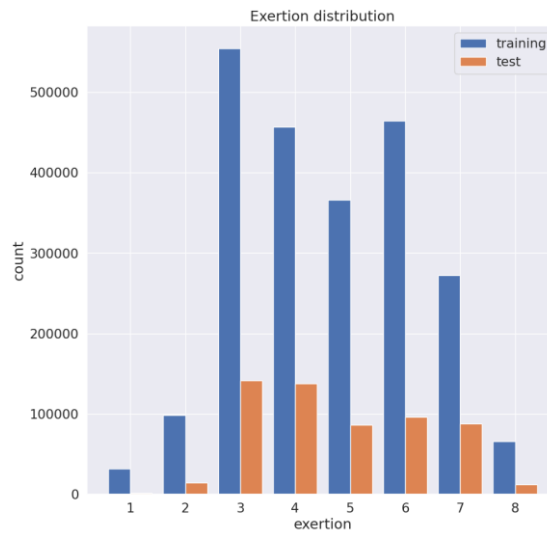


Figure 15: Number of observations available in training and test datasets, for each exertion label

Figure 16 and Figure 17 reported the results obtained on the test set with the best trained model. The accuracy achieved is approximately 65%, and the MSE is around 0.86 (considering a fatigue exertion range from 1 to 8). When analysing the confusion matrix, the model appears to struggle in correctly predicting class 2, with some difficulties also making right predictions in the range from 4 to 7. However, the results can still be considered satisfactory as most of the errors appears very close to the diagonal (which means that the model predicts a class that is very close to the right one).

	precision	recall	f1-score	support
1	0.97	0.87	0.92	1918
2	0.42	0.31	0.36	14387
3	0.89	0.84	0.86	141404
4	0.80	0.55	0.65	137670
5	0.62	0.56	0.59	86021
6	0.40	0.65	0.50	96561
7	0.54	0.56	0.55	87803
8	0.94	1.00	0.97	12102
accuracy			0.65	577866
macro avg	0.70	0.67	0.68	577866
weighted avg	0.68	0.65	0.65	577866

Figure 16: Classification report of the obtained model

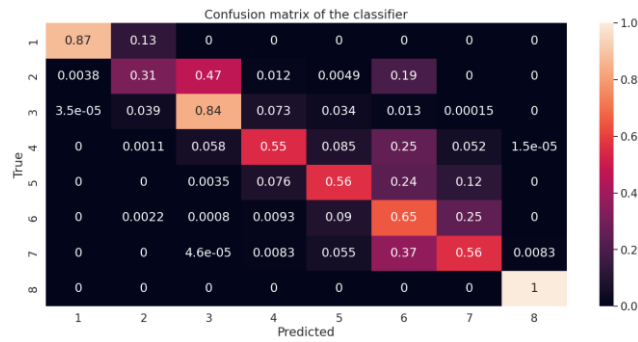


Figure 17: Confusion matrix of the obtained model

In Figure 18, it is interesting to note that the most important features are the dynamic ones, namely skin temperature, galvanic skin response, and heart rate. Among the dynamic data features, those related to acceleration appear to be less important for the model (thus, a training trial without using these features is presented in section 2.6.2). Additionally, importance is given to static features such as grip strength, height, ompq and weight.

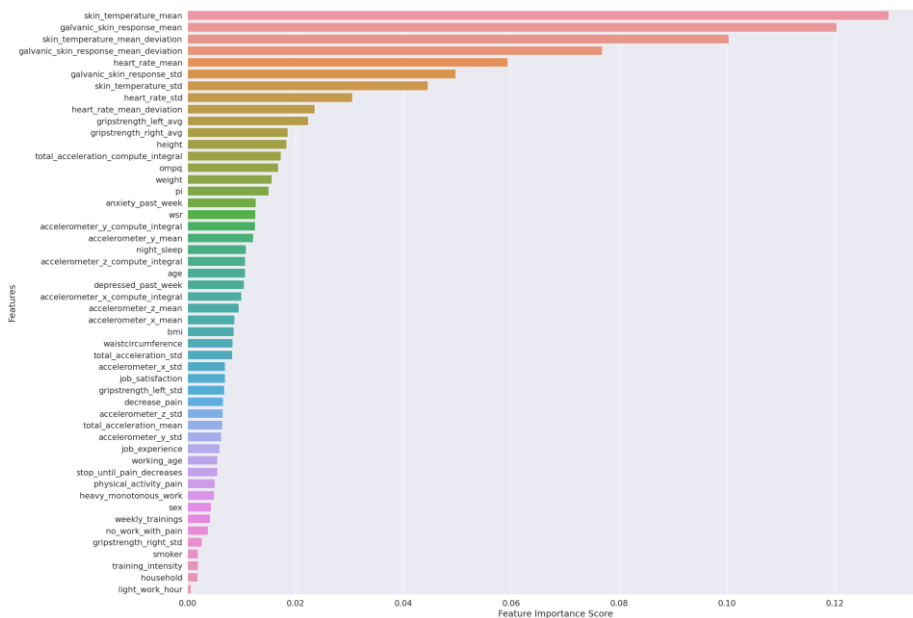


Figure 18: Feature importance according to the obtained model

2.6.2 Training excluding accelerometer data

In section 2.6.1, it was observed that the features related to acceleration are considered by the model the least important among the dynamic data features. To assess the impact on model performance, a new model was trained excluding the accelerometer-related features.

Figure 19 displays the fatigue exertion distributions between the training and test sets. Again, it should be noted that there are instances with a perceived fatigue exertion value of 1 in the test set.

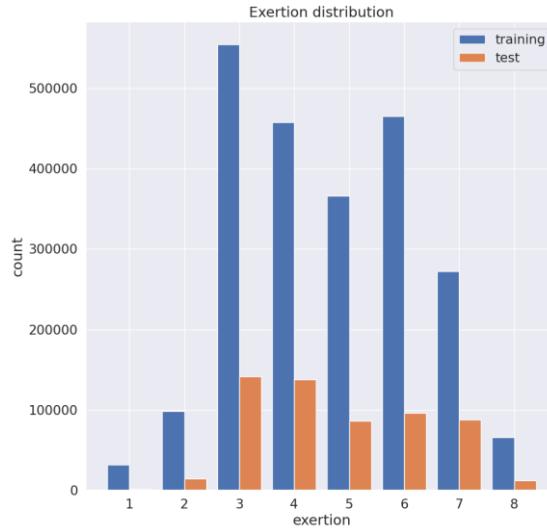


Figure 19: Number of observations available in training and test datasets, for each exertion label

The results presented in Figure 20 and Figure 21 demonstrate that excluding the acceleration features yields similar results compared to the ones obtained by training the models with all the available features. The accuracy obtained is approximately 64%, slightly lower than the 65% of the previous model. The confusion matrix also exhibits a similar pattern, indicating the model's struggle with class 2 and some difficulty in the range from 4 to 7. In the fatigue exertion range of 1 to 8, the MSE is approximately 0.8274, slightly better than the previous value of 0.8615. These results confirm that the acceleration features may not be crucial for predicting perceived fatigue exertion.

	precision	recall	f1-score	support
1	0.93	0.88	0.90	1914
2	0.40	0.29	0.33	14387
3	0.88	0.82	0.85	141392
4	0.73	0.55	0.63	137670
5	0.65	0.59	0.62	86021
6	0.43	0.65	0.51	96561
7	0.52	0.53	0.53	87803
8	0.91	1.00	0.95	12102
accuracy			0.64	577850
macro avg	0.68	0.66	0.67	577850
weighted avg	0.67	0.64	0.65	577850

Figure 20: Classification report of the obtained model

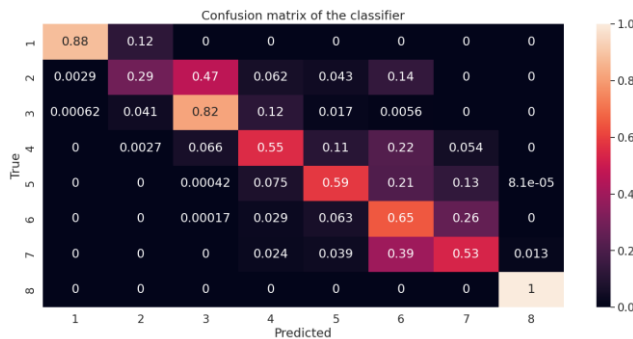


Figure 21: Confusion matrix of the obtained model

From the plot shown in Figure 22, skin temperature, galvanic skin response, and heart rate are still the most important features. In this case, the model gives importance to static features such as the age, grip strength, weight, and height.³

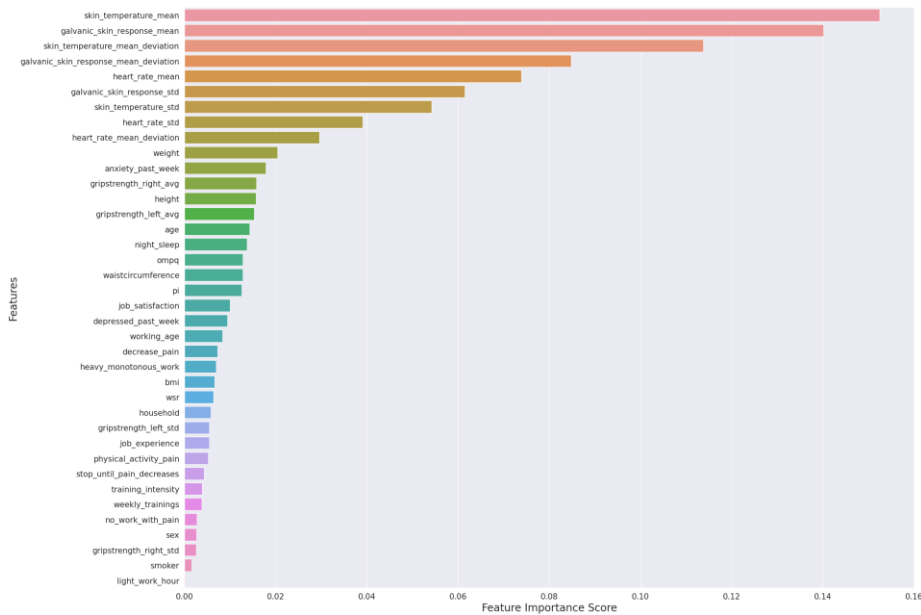


Figure 22: Feature importance according to the obtained model

2.6.3 Training using only heart rate, skin temperature and galvanic skin response

Sometimes it may not be possible to collect or update workers' static data, or the workers themselves may refuse to provide the requested information or provide false information. To overcome these scenarios, a third attempt has been made to train Random Forest classifier using dynamic data only, namely skin temperature, galvanic skin response, and heart rate. Also in this case, the model has been trained using the "balanced" mode. Figure 23 shows the fatigue exertion distributions in the training and test sets.

As can be seen from Figure 24 and Figure 25, the performance is somewhat lower than in the previous 2 experiments, with a slightly higher MSE (0.8316). However, the performance drop can be accepted to avoid situations as the ones described above; also, the model considers a very limited number of features, thus resulting in a lighter model. However, the model still struggles to distinguish between fatigue exertion level 2 and the range from 4 to 7.

From the feature importance scores plotted in Figure 26, the model continues to consider the skin temperature as the most important feature, followed by galvanic skin response, and heart rate.

³ As per importance, we refer to the Gini importance, i.e., the mean decrease of impurity. The score is computed by observing how a feature is used within the decision tree to make decision on how to divide the data set into two separate sets, with similar responses within.

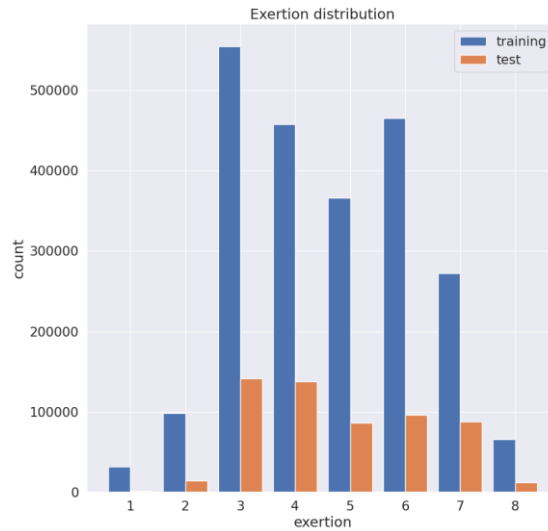


Figure 23: Number of observations available in training and test datasets, for each exertion label

	precision	recall	f1-score	support
1	0.48	0.90	0.63	1914
2	0.39	0.30	0.34	14387
3	0.83	0.81	0.82	141392
4	0.72	0.54	0.62	137670
5	0.52	0.60	0.56	86021
6	0.46	0.61	0.53	96561
7	0.56	0.58	0.57	87803
8	0.73	0.48	0.58	12102
accuracy			0.63	577850
macro avg	0.59	0.60	0.58	577850
weighted avg	0.64	0.63	0.63	577850

Figure 24: Classification report of the obtained model

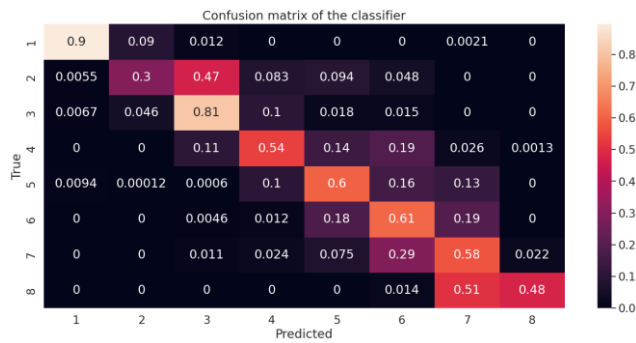


Figure 25: Confusion matrix of the obtained model

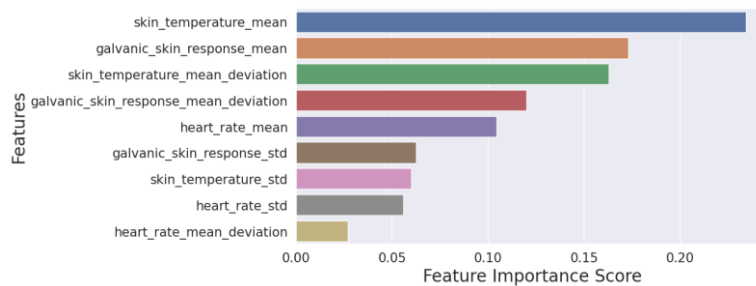


Figure 26: Feature importance according to the obtained model

2.6.4 Training after dimensionality reduction

An attempt was made to reduce the dimensionality of the features by applying the Principal Component Analysis (PCA). The data were first standardised.⁴ Then, as shown in Figure 27, it was determined which number of components could explain a percentage of the variance. With the help of the plot, it was decided to reduce the number of features to 25 components (i.e., the minimum number of features that is still capable of representing the full variance).

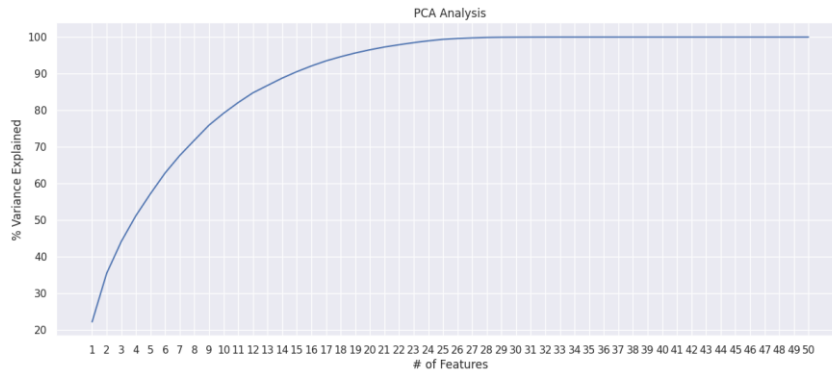


Figure 27: Variance explained after feature reduction

After that, as shown in Figure 28, the dataset was divided into training and test sets.

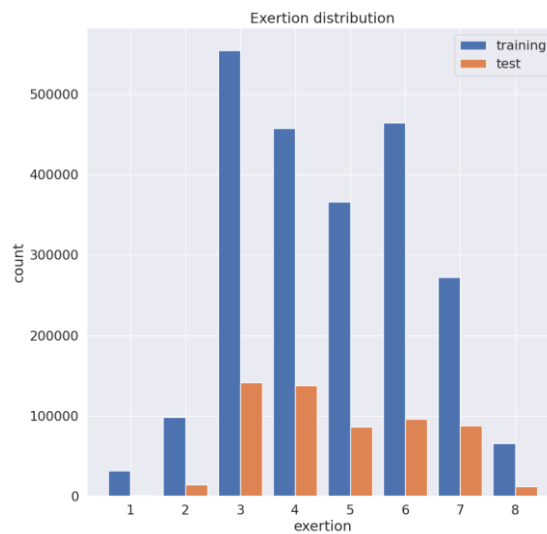


Figure 28: Number of observations available in training and test datasets, for each exertion label

As can be seen from Figure 29 and Figure 30, this approach led to a further performance drop (the MSE is about 1.0253). For example, it can be seen that the model now struggles to distinguish fatigue exertion levels 2 and 4. Consequently, reducing the number of features using PCA did not lead to an improvement in the model performance. Considering that PCA leads to a loss of interpretability of the features and, in this case, did not provide any added benefit to the model performance, it was concluded that this approach was not suitable.

⁴ As per feature standardisation, in this document we mean the process of removing the mean from the feature and scaling it to unit variance. This procedure is implemented by the StandardScaler class available in scikit-learn.

	precision	recall	f1-score	support
1	0.59	0.84	0.70	1918
2	0.32	0.26	0.28	14387
3	0.79	0.77	0.78	141404
4	0.73	0.48	0.58	137670
5	0.53	0.50	0.52	86021
6	0.39	0.67	0.50	96561
7	0.51	0.45	0.47	87803
8	0.85	0.96	0.90	12102
accuracy			0.59	577866
macro avg	0.59	0.61	0.59	577866
weighted avg	0.62	0.59	0.59	577866

Figure 29: Classification report of the obtained model

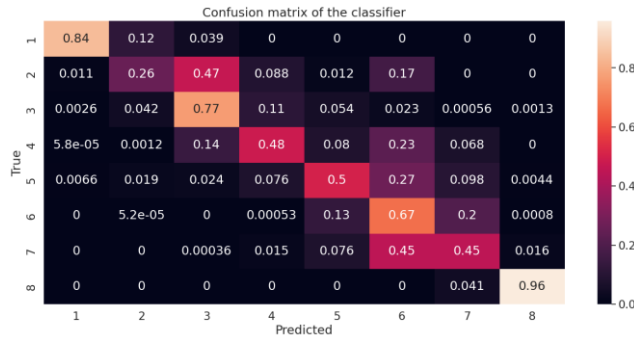


Figure 30: Confusion matrix of the obtained model

2.7 Model 2: Feed Forward Neural Network

Finally, an alternative model to Random Forest classifiers was identified in Neural Networks. Specifically, a feed forward neural network has been designed. The idea is to exploit the capability of deep neural networks to learn features in the first layers by their own, without requiring any kind of feature selection beforehand. By passing all the standardised features to the network, it can learn the relationships for the fatigue exertion prediction task. The network was trained to solve a regression task, using MSE as the loss function. In this way, during the training phase, the network also considered how much it deviated from the correct prediction, not just whether it was correct or incorrect. During the inference phase, the predicted fatigue exertion value was rounded to the nearest integer in the range of 1 to 8, simulating a classification task, despite the use of a regressor model.

The architecture of the designed network is depicted in Figure 31.

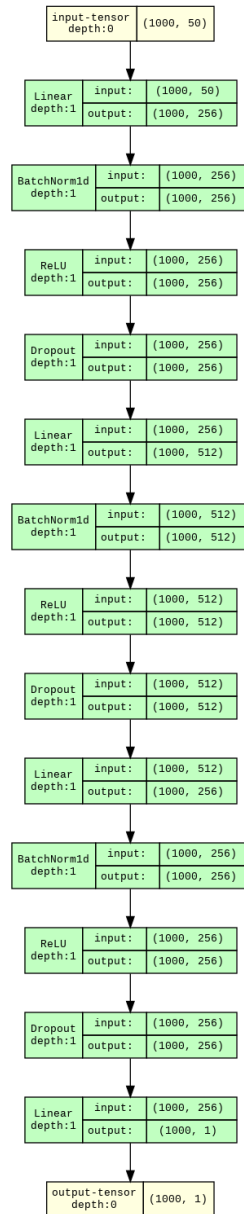


Figure 31: The architecture of the Feed Forward Neural Network for physical exertion prediction.

The dataset was divided into three subsets (i.e., training, validation, and test sets) by reusing the same custom function based on segments described in section 2.6; in this case, each segment is further divided in 3 parts: 80% for the training set, 10% for the validation, and the remaining 10% for test. In Figure 32, the fatigue exertion distribution obtained in the three data sets can be seen.

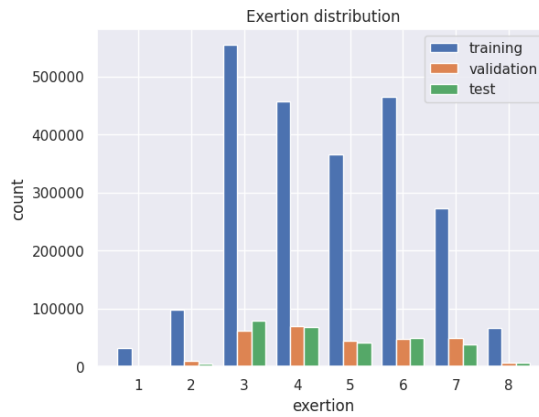


Figure 32: Number of observations available in training, validation, and test datasets, for each exertion label

Different configurations of network hyper-parameters have been tested. In the following sections, the 2 best configurations are described.

2.7.1 Results with the first version of the network

The following hyperparameters were used for the first version of the network:

- a) Optimiser: SGD
- b) Learning rate: 0.001
- c) Batch size: 1000
- d) Dropout: 0.1
- e) Epochs: 20

Below are the results obtained with this version of the network. As shown in Figure 33 and Figure 35, the model tends to overfit. In fact, the training loss continues to decrease, but the validation loss does not show much improvement and remains rather flat. If loss is taken as the benchmark metric, the best performance was achieved at epoch 16 with a value of approximately 0.65. On the other hand, if accuracy is used as the reference metric, the best performance was achieved at epoch 17 with about 55.73% of the predictions being correct. The accuracy obtained is quite low, but as explained in previous sections, it is not a very indicative metric to evaluate the model. Much more useful are the confusion matrices (see Figure 34 and Figure 36), where we can see that the model does not perform that badly overall. In fact, most of the errors are located near the diagonal of the matrix and are consequently considered less severe. As in previous tests, the model struggles to distinguish level 2 fatigue exertion, as well as levels 4, 5, and 7. Figure 33 shows the performance of the selected model using loss as the reference metric, while Figure 35 shows the performance achieved by the selected model using accuracy.

Model with best loss (MSE: 0.9060)

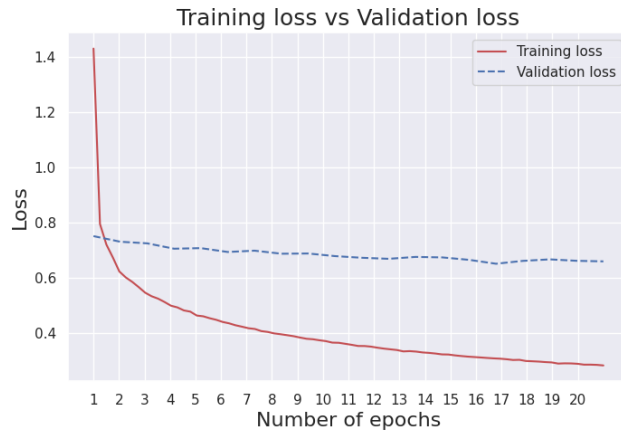


Figure 33: Training loss vs Validation loss

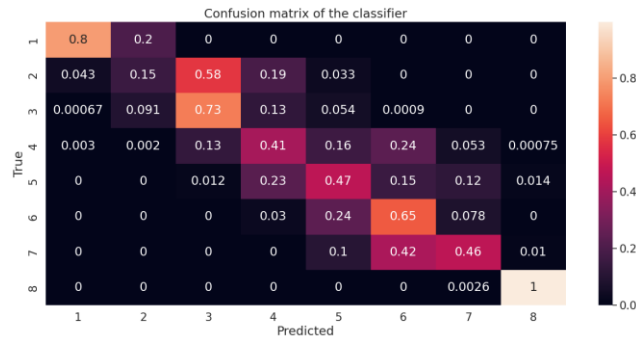


Figure 34: Confusion matrix of the best model selected through loss

Model with best accuracy (MSE: 0.9135)

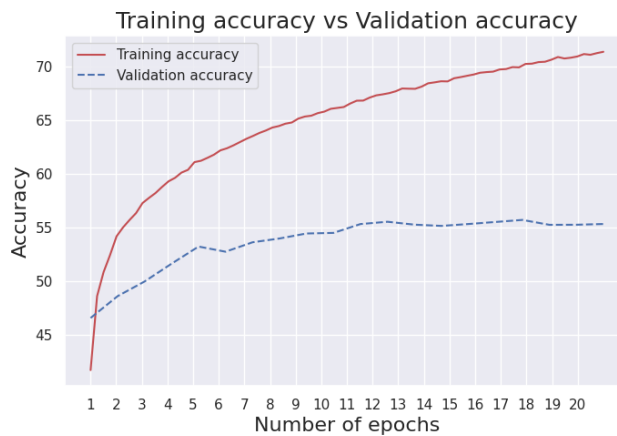


Figure 35: Training accuracy vs Validation accuracy

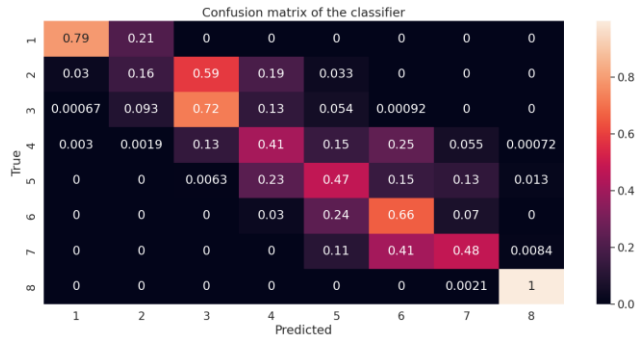


Figure 36: Confusion matrix of the best model selected through accuracy

2.7.2 Results with the second version of the network

In this case, the following hyperparameters were used to train the second version of the network:

- a) Optimiser: SGD
- b) Learning rate: 0.01
- c) Batch size: 1000
- d) Dropout: 0.2
- e) Epochs: 20

Again, as depicted in Figure 37 and in Figure 39, the model shows a tendency to overfit. The best loss value was obtained at epoch 14 with a value of about 0.6308, while the best accuracy achieved corresponds to about 57.42%, obtained at epoch 20. Looking at the confusion matrices (see Figure 38 and Figure 40), one can see the improvement that the network has achieved with the use of the new hyperparameters. Indeed, the model makes fewer errors away from the diagonal of the matrix, which is a good result. However, the model continues to struggle in distinguishing fatigue exertion levels 2, 4, 5, and 7.

Model with best loss (MSE: 0.9588)

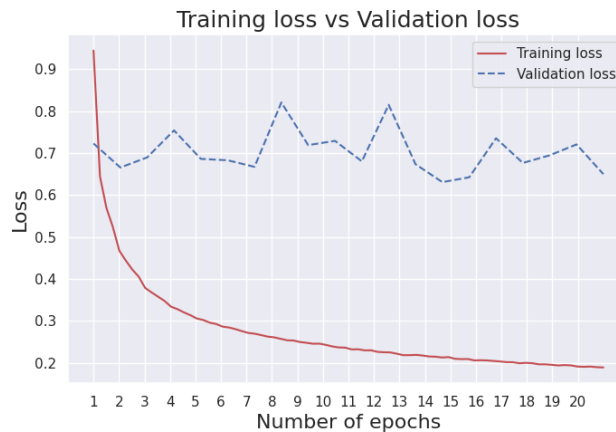


Figure 37: Training loss vs Validation loss

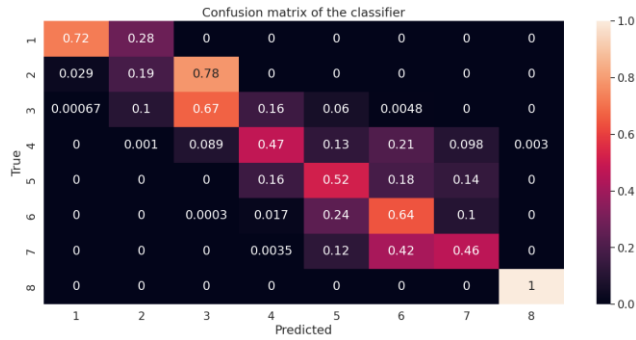


Figure 38: Confusion matrix of the best model selected through loss

Model with best accuracy (MSE: 0.9444)

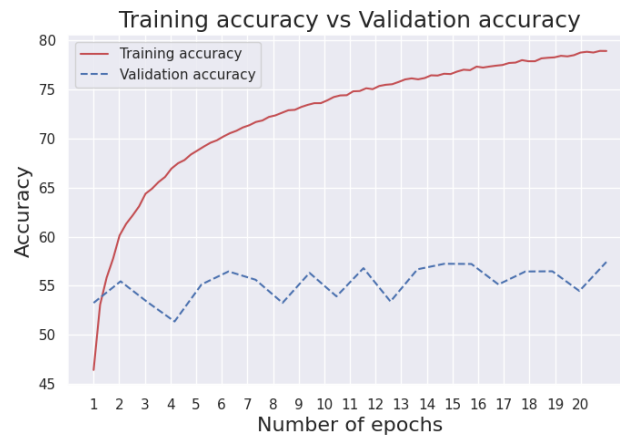


Figure 39: Training accuracy vs Validation accuracy

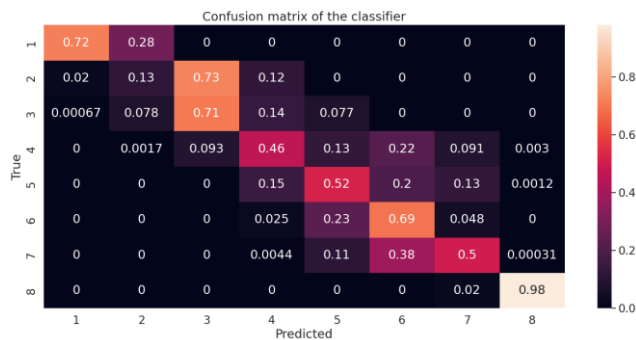


Figure 40: Confusion matrix of the best model selected through accuracy

2.8 Further fatigue detection and prediction scenarios

Beside the design of machine learning models for fatigue exertion prediction, additional data analyses were performed on the available datasets aiming to explore the following research questions:

1. To what extent can we group data from workers physical measurements and relate them to different states (e.g., work activity, fatigue/no-fatigue: useful when no information about fatigue is available)
2. To what extent can we estimate the current exertion level of a worker from past and current physical measurements? (e.g., useful when no information about fatigue is available)

3. To what extent can we estimate the current exertion level of a worker from past and current physical measurements and past exertion levels? (Useful when previous exertion levels are available, e.g., because they have been already estimated)
4. To what extent can we predict future exertion levels from past physical measurement and exertion levels?

Answering this research question is of interest when obtaining measurements not previously tagged with exertion levels (labelled data), and without having the benefit of available pre-trained models for exertion prediction or estimation. To answer the questions above, clustering over time series data was performed.

2.8.1 Clustering and Classification

The aim of the experiment was to assess whether groupings in the data could be identified. Relevant groupings could be linked to (a) grouping the data according to the type of work activity performed, without prior information about it (b) grouping the data according to fatigue levels, without prior information about it. In other words, the outcomes of the clustering exercise were reviewed in terms of clustering performance assessed via the clustering silhouette coefficient. All attributes recorded in the time series were included in the experiments. As these attributes involves measurements with different sampling rates, data synchronisation was first performed. For practical processing, low sampling rate data records (for example temperature) were upsampled to synchronise with the high sampling rate ones (e.g., acceleration). Clustering was performed with k-Means, Hierarchical, and DBSCAN methods. Specifically, the extent to which either (a) the type of activity (b) the fatigue stage, could be associated with the time series data. An example of outcomes from clustering is shown in Table 5. The results were highly similar for all types of clustering, indicating that while the 'Handling' type of activity could relatively easily be identified in the data, other aspects were confused, i.e., it was not possible for example to associate detected clusters with fatigue state; similarly, 'Walking' type of activities could not easily be associated with data groupings. Nonetheless, it could be said that on the basis of such outcome, it may make sense to include clustering as a first stage processing of non-labelled data, just so as to detect the type of work activity, and then perform fatigue detection/estimation/prediction on similarly grouped data, i.e., data from similar types of activities.

Table 5: Fatigue and Activity Clustering

Number of clusters: 2		Cluster1: Walking	Cluster 2: Handling		Cluster1: No Fatigue	Cluster 2: Fatigue
Clustering Silhouette coefficient	Average value for activity: 0.343	-0.016	0.651	Average value for fatigue: 0.015	0.02	0.01

Next, the data were used alongside their data labels regarding the fatigue state, aiming to classify them according to the fatigue state. To explore this, we transformed the multi-class classification problem into a binary classification one, by assigning a fatigue state to the highest exertion levels, and non-fatigue state to the lowest ones, with experiment repetitions for different threshold logic. Experiments were performed with the KNIME Analytics platform involving Decision Trees, Rule Learners, Logistic Regression, and k-NN classifiers, with the latter outperforming the rest for 1 and 3 nearest neighbours, as reported in Table 6. In all

cases, when performing the experiments using the activity clusters identified earlier resulted in performance improvements. However, this outcome could only be seen as evidence, which is nonetheless not statistically significant, given the somewhat limited range and volume of involved data.

Table 6: Fatigue Classification

Classifier	Overall accuracy
DT2	0.775
RL	0.662
REGR (LRL)	0.625
REGR (WEKA)	0.688
KNN	0.925

It is interesting to observe that a single measurement attribute can account for a significant part of the class separability for each experiment subject (each person), as seen in Table 7 (using the single attribute for classification).

Table 7: Single Attribute Fatigue Classification

Subject	Overall accuracy
Heart rate	0.575
Ankle data	0.675
Hip data	0.675
Wrist data	0.613
Torso data	0.725
Step data	0.613

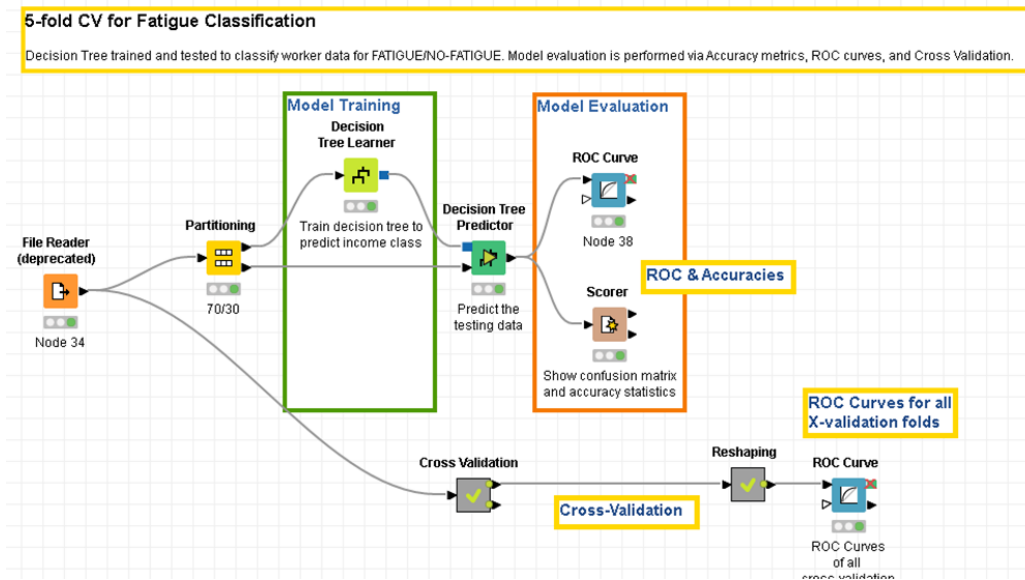


Figure 41: The KNIME workflow for fatigue classification.

2.8.2 Time Series processing

To explore the research questions 2, 3, and 4, the problem was formulated in a time series setting. In the simplest case this involved a simple auto-regression over the time series data. As before, the time series of the data sets were synchronised and upsampled to allow for joint processing. An example of a linear time series model implemented in a KNIME workflow is shown in Figure 42.

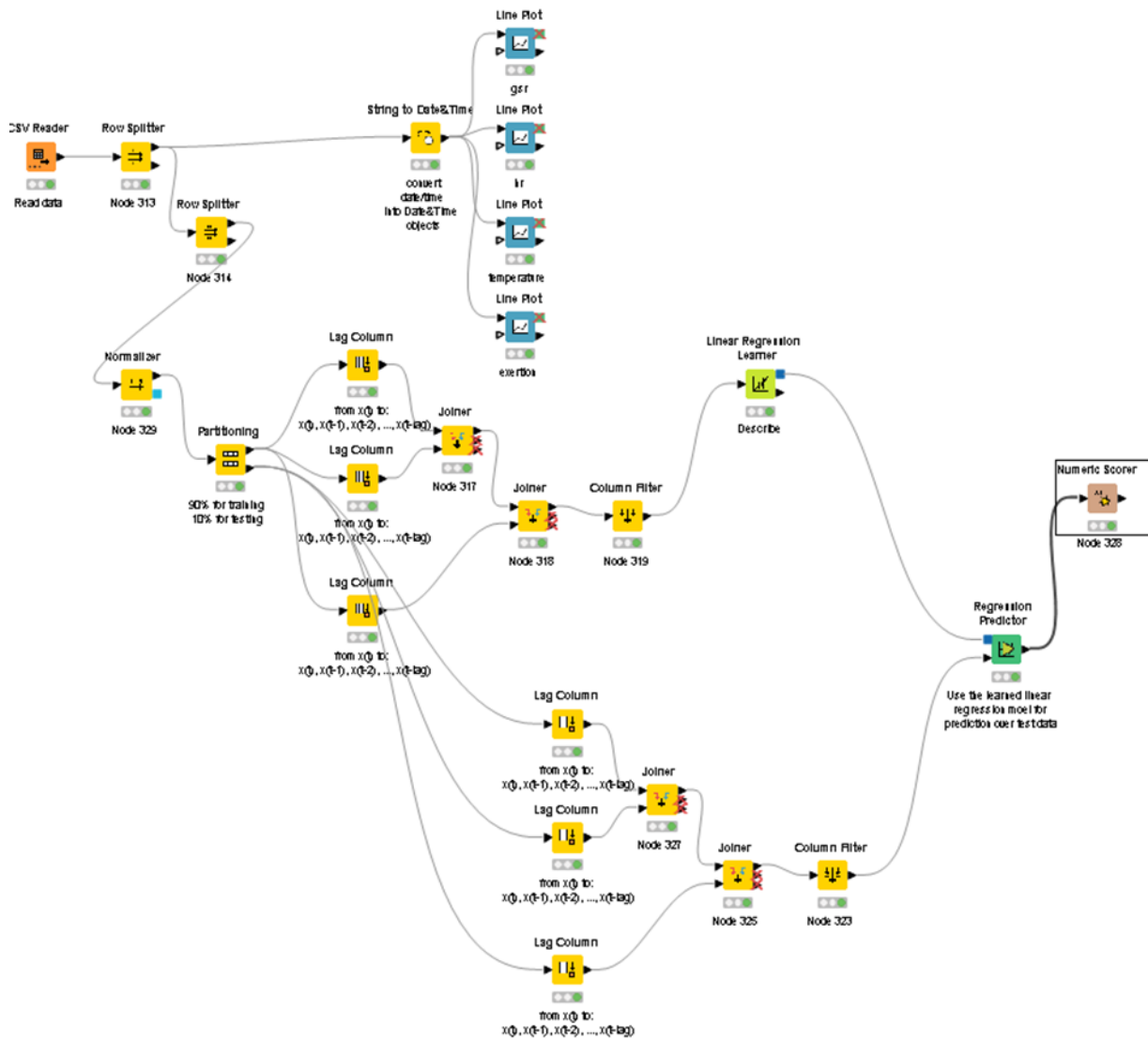


Figure 42: Fatigue estimation and prediction example workflow

Applying time series regression for each one of the research questions 2, 3, (estimation) and 4 (prediction) was characterised by some subtleties. First, it clearly made little sense to develop a uniform model for fatigue across all measurement subjects. Furthermore, the experiments setup was such that the length of the time series in most cases was somewhat short. This implies that a typical approach of using the first part of the time series for training a model, keeping the latter for testing it would only have limited value. Furthermore, the fatigue ground truth would only be available mostly for the start and the end of the experiments, rendering the values in-between not to be particularly useful. We therefore concluded that the time series setup would offer limited additional information on top of a classification model that is able to assign a work sequence activity for a given person to case of fatigue or not, as observed by the end of the experiment.

2.9 Summary

In this study, two models were evaluated for the task of fatigue exertion classification: the Random Forest classifier and the Feed-Forward Neural Network. Various configurations were tested for each model to find the models with the best performance.

For the Random Forest, different training attempts were conducted using different feature sets. These included using all available features from the dynamic and static data, excluding accelerometer data, using only dynamic data related to skin temperature, galvanic skin response, and heart rate, and using reduced features through PCA.

On the other hand, the Feed-Forward Neural Network utilised all features (both static and dynamic), with the exploration of different hyperparameters. Table 8 presents the results obtained by the models in the tested configurations.

Among the models tested, the Random Forest using all dynamic and static features achieved the highest accuracy on the test set, reaching approximately 65%. However, when considering the MSE within a fatigue exertion range of 1 to 8, the Random Forest with all dynamic and static features except accelerometers performed the best, yielding a value of 0.8274.

Table 8: Performance of models in various configurations

Model	Test accuracy	Test MSE
Random Forest using all features	65%	0.8615
Random Forest excluding accelerometer data	64%	0.8274
Random Forest using only heart rate, skin temperature and galvanic skin response	63%	0.8316
Random Forest using PCA	59%	1.0253
Forward Neural Network using all features	60%	0.9060

3 AI Models for Mental Stress Prediction

This section presents the approach used to develop a model for the mental stress prediction.

Since this work represents our first attempt in handling mental stress, we started with a simpler binary classification task, where the negative class “0” represents the no stress state, and the positive class “1” represents the presence of stress. Similarly to what presented in section 2, two machine learning models were implemented that could predict the instantaneous mental stress perceived by people in a stressful situation (see section 3.1). By leveraging the positive results obtained in using models like Random Forest and feed-forward neural network for the physical fatigue prediction task (see section 2), it was decided to apply the same approach for the mental stress prediction task as well. The models were trained following the same methodologies used in the physical fatigue experienced prediction task (i.e., models trained using all the available features; same data partitioning strategies; same models architecture; same evaluation metrics).

The results of the obtained models were then compared with a baseline model proposed in literature by Villani et al [REF-01]. The model proposed by Villani et al. allows for a binary determination of perceived mental stress by computing the RR interval within time windows of 2.5 minutes.

3.1 Data collection

To train the mental stress prediction model, a new dataset has been built from scratch. Two different sessions have been organised to collect data:

- SUPSI: the data collection involved 16 collaborators.
- Drachten: during the project GA, an additional data collection has been organised involving 5 project partners.

The limited number of people involved is mainly due to the need to wear the sensors, which makes it impossible to scale up with remote data collections.

During the data collection sessions, each participant was equipped with a combination of three sensors: an Empatica E4 on the left wrist, a Polar Verity Sense on their right forearm, and a Polar H10 on the chest. Given the preliminary nature of this study, we aimed to explore the use of different sensors positioned on various parts of the body. It should be noted that not all sensors capture the same set of metrics (a comprehensive list of the metrics captured by each sensor can be found in section 3.2). Our objective was to identify the most informative metrics and sensors for predicting mental stress.

Each session required the participant to complete two cognitive tests, namely the *stroop test* and the *3-back test*. The stroop test shows to the user several words saying different colours, written possibly in different colours. For each trial, the user is required to answer with the colour of the displayed word. For instance, Figure 43 shows the word “blue” written in green, thus the right answer is green.

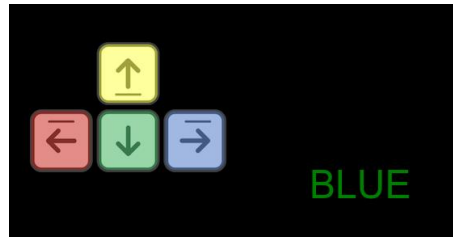


Figure 43: An example from the stroop test. The right answer is green (key down).

The 3-back test presents a a sequence of stimuli, one-by-one. The user has to decide if the current stimulus is the same as the one presented 3 trials in advance. Figure 44 shows a sequence of 6 stimuli, where the fourth and sixth represent right answers.

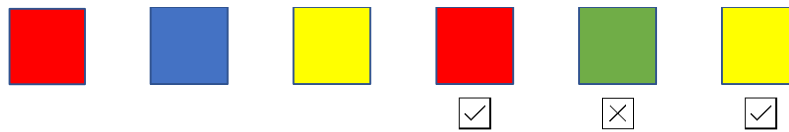


Figure 44: A short sequence from the 3-back test. Right answers are marked by a checkmark.

Both tests lasted for about 7 minutes, with the user required to provide an answer within 2 seconds, for each sample or stimulus.

At the end of the two tests, the participants have been asked to fill out the NASA Task Load Index (TLX) questionnaire.

During the experiment, dynamic data and quasi-static data were collected using the same methodology as for the physical exertion prediction model (see section 2.1). However, no labels have been collected related to perceived level of mental stress; instead, an additional questionnaire (i.e., the NASA Task Load Index) was given at the end of the experiment, asking the participants to judge different kinds of load experienced during the experiments.

Since no labels were collected during the experiments, new labels have been included in the dataset by means of an assumption. Specifically, the assumption assigns the non-stress label (negative class 0) to the observations collected while the participant was resting; conversely, the stress label (positive class 1) was assigned to all the observations collected while the participants were actively participating in the stress test. It is important to note that the participants were equipped with sensors prior to the start of the games, and therefore, sensor data acquired during this initial phase were labelled with the negative class (no stress state).

3.2 Dynamic data pre-processing

3.2.1 Data cleaning

The participants were equipped with a combination of three different sensors (i.e., Empatica E4, Polar Verity Sense, and Polar H10), collecting the following metrics:

- Empatica E4: RR interval, photoplethysmography, galvanic skin response, skin temperature, and 3-axis accelerometer data (x, y, and z).
- Polar Verity Sense: heart rate, photoplethysmography (components 0, 1, 2, and ambient), PP interval, 3-axis accelerometer data (x, y, and z), 3-axis gyroscope data (x, y, and z), 3-axis magnetometer data (x, y, and z).

- Polar H10: heart rate, RR interval, electrocardiography, and 3-axis acceleration data (x, y, and z).

The pre-processing and cleaning of the sensor data followed a similar approach to that described in section 2.2.1. Standard data cleaning procedures were applied, including removal of duplicates, and filtering out missing data.

The concept of sessions has been introduced, as done for the physical fatigue exertion detection task. In this case, a new session is identified every time a gap of 1 minute is detected within the sensor data (the gap is shorter this time, due to the shorter duration of the whole experiment).

Finally, linear interpolation has been applied to align observations with different timestamps (and acquisition frequencies), while noisy data (e.g., accelerometer, gyroscope, and magnetometer) have been smoothen with a filtering function (i.e., the savgol filter).

3.2.2 Features extraction

The methods described in section 2.2.2 were applied to expand the feature set. Mean deviation was extracted from heart rate, skin temperature, galvanic skin response, electrocardiography, photoplethysmography, RR interval, and PP interval. For these metrics, a 60-second windowing approach with overlapping was implemented, computing the mean value and standard deviation within each window. In this case, the maximum and minimum values were not considered, as they were found to be highly correlated with the mean value and standard deviation, as mentioned in section 2.5. Therefore, they were excluded during the feature selection stage.

Additionally, the total acceleration was calculated from the three components (x, y, z) of the accelerometer. The same procedure was applied to the gyroscope and magnetometer data. Considering these three metrics (accelerometer, gyroscope, and magnetometer), a 5-second windowing approach with overlap was used to compute the mean, standard deviation, and integral within each window.

3.3 Data from cognitive demanding tests

During the cognitive tests, different types of data have been collected. Data related to the games themselves (e.g., time to answer, users’ answers, details about the trial) are shown in Table 9 and Table 10, for the stroop test and 3-back test respectively. At the end of the two tests, the user has been asked to complete the NASA TLX questionnaire related to both the test. The collected features are listed in Table 11.

Table 9: Features collected during the stroop test

Field	Notes
id	Internal id
answer	The answer given by the user
colour	The colour of the displayed word
congruent	1 when the word displayed indicates the colour of its ink, 0 otherwise
corrAns	The correct answer
keypress	The ISO 8601 instant the user provided an answer
timestamp	The ISO 8601 instant the trial was displayed
word	The word displayed
userid	Hashed UUID

Table 10: Features collected during the 3-back test

Field	Notes
id	Internal id
answer	The answer given by the user
timestamp	The ISO 8601 instant the trial was displayed
colourtest	The displayed trial
corresp	The correct answer
keypress	The ISO 8601 instant the user provided an answer
userid	Hashed UUID

Table 11: Features collected with the NASA TLX questionnaire

Field	Notes
Timestamp	Acquisition date ISO 8601
userid	Hashed UUID
stroop.mental_demand	How mentally demanding was the task? (0 = Low; 100 = High)
stroop.physical_demand	How physically demanding was the task? (0 = Low; 100 = High)
stroop.temporal_demand	How hurried or rushed was the pace of the task? (0 = Low; 100 = High)
stroop.performance	How successful were you in accomplishing what you were asked to do? (0 = Perfect; 100 = Failure)
stroop.effort	How hard did you have to work to accomplish your level of performance? (0 = Very Low; 100 = Very High)
stroop.frustration	How insecure, discouraged, irritated, stressed, and annoyed were you? (0 = Very Low; 100 = Very High)
stroop.rank	Rank the 2 activities (1 = Less demanding; 2 = Most demanding)
3_back.mental_demand	How mentally demanding was the task? (0 = Low; 100 = High)
3_back.physical_demand	How physically demanding was the task? (0 = Low; 100 = High)
3_back.temporal_demand	How hurried or rushed was the pace of the task? (0 = Low; 100 = High)
3_back.performance	How successful were you in accomplishing what you were asked to do? (0 = Perfect; 100 = Failure)
3_back.effort	How hard did you have to work to accomplish your level of performance? (0 = Very Low; 100 = Very High)
3_back.frustration	How insecure, discouraged, irritated, stressed, and annoyed were you? (0 = Very Low; 100 = Very High)
3_back.rank	Rank the 2 activities (1 = Less demanding; 2 = Most demanding)

3.4 Static data from the questionnaire

For the mental stress detection, static data have been collected via questionnaire filled out by the participants. The same questionnaire used for the physical exertion prediction has been adopted, with just removing all the questions related to the work experience and physical strength and pains. The resulting set of features in listed in Table 12).

Table 12: List of features available as static data

#	Field	Unit	Acquisition method	Notes
1	Timestamp	ms	Read from system	Acquisition date
2	Userid		Manually assigned	Hashed UUID
3	Sex		Questionnaire	male / female
3	Weight	kg	Questionnaire	
4	Height	cm	Questionnaire	
5	Smoker		Questionnaire	yes / no
6	weekly_trainings		Questionnaire	
7	training_intensity		Questionnaire	
8	anxiety_past_week		Questionnaire	
9	depressed_past_week		Questionnaire	
10	light_work_hour		Questionnaire	
11	walk_hour		Questionnaire	
12	Household		Questionnaire	
13	weekly_shopping		Questionnaire	
14	night_sleep		Questionnaire	
15	Age		Computed	age = current year – year of birth (removed)

Features from #6 to #13 have been collected by answering the same questions listed in Table 4 in section 2.4 (indexed by the same feature names).

3.5 Data quality assessment

Experiments such as the ones described in this report bring in complications arising from mixing qualitative and soft data (questionnaires) with sensor-based data acquisition. While a couple of models are presented in the following (see sections 3.7 and 3.8), these complications introduced some quality issues within the new built dataset. Some of the issues have been already fixed, while some other require further investigation (both for understanding the source, and for implementing a mitigation strategy that minimises the information loss). Some examples of data quality issues detected in the dataset include: (a) null values recorded in for certain samples; (b) attributes with invalid values (semantic data inaccuracy, e.g., "strop.demand" attained invalid values of as high as 105, exceeding the maximum of 100); (c) stroop test data featuring invalid timestamp, where the user seems to answer a test before it is displayed. These are some examples of issues identified in early data collection experiments. This type of data quality assessment enabled remedying actions to ensure appropriate data quality feeding into the next data processing steps.

3.6 Data analysis and feature selection

At this stage, the same analyses were conducted as described in section 2.5. First, highly correlated features were identified. Figure 45 illustrates the correlation matrix of the numerical features. It is evident from the matrix that certain features exhibit correlation with each other; however, it was decided not to discard them.

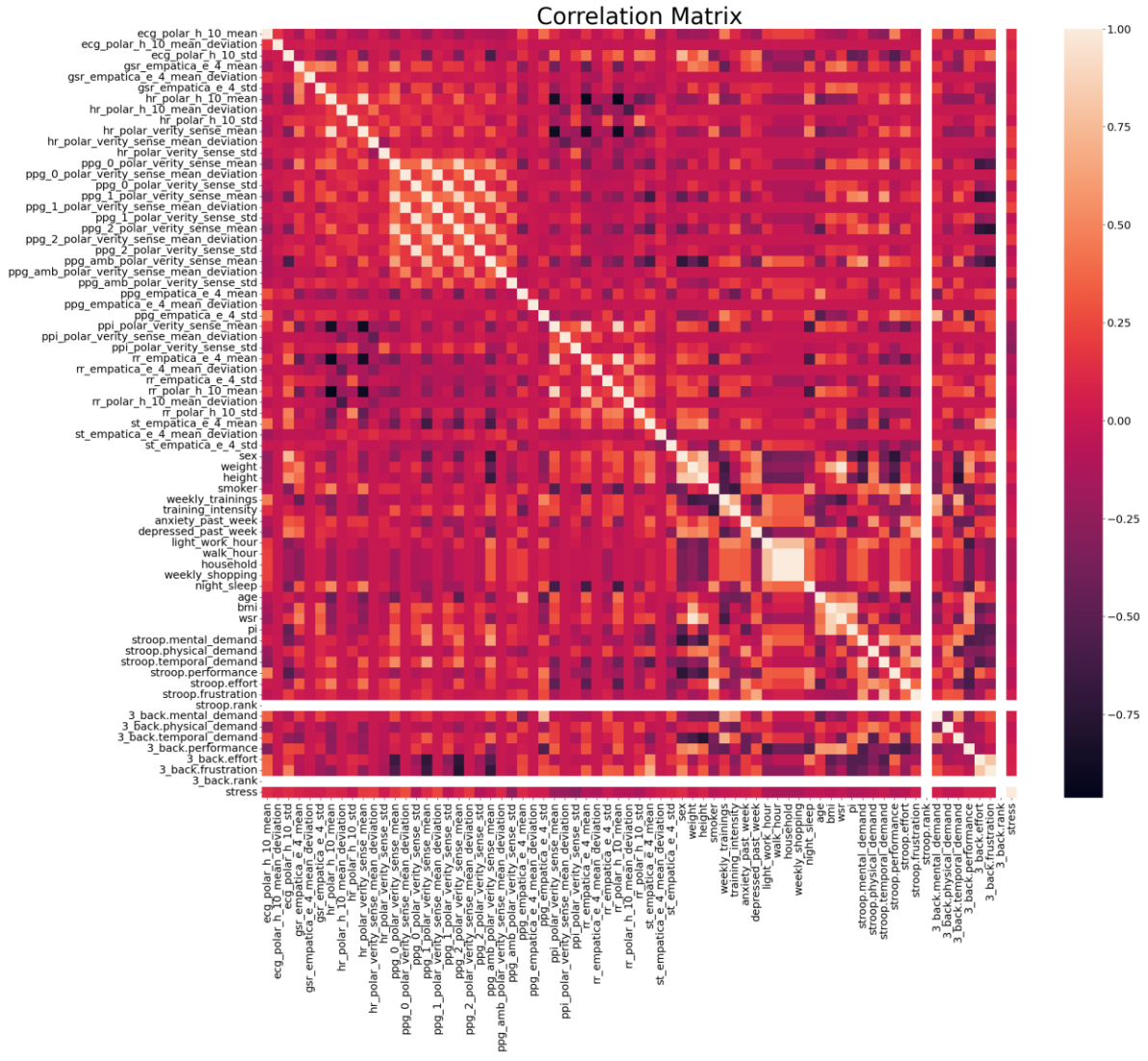


Figure 45: Correlation matrix showing the correlation between numerical features

Correlations between the features and the target variable, referred to as “stress”, were then analysed. Figure 46 and Figure 47 illustrate the correlations. It can be observed that the features most correlated with the target variable are those related to photoplethysmography. However, the correlation between the features and the target variable is not particularly strong. Even the most correlated feature (in terms of absolute value) has a correlation coefficient of less than 0.5.

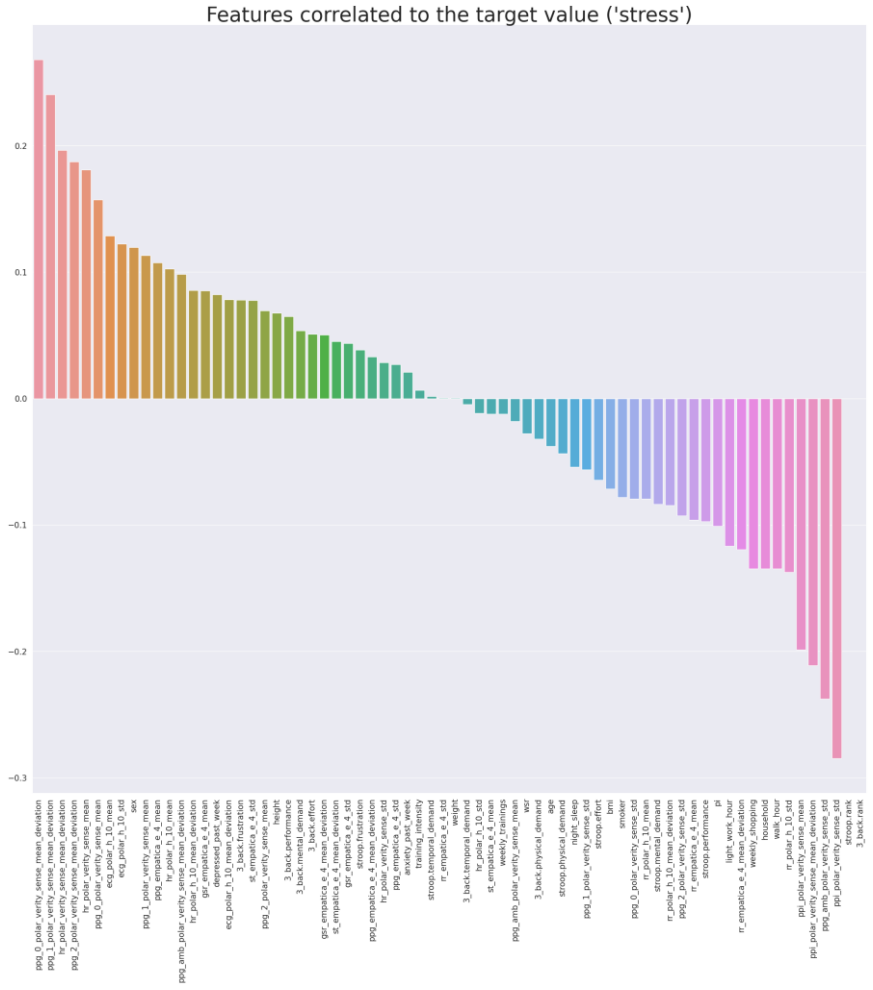


Figure 46: Feature correlation with the target variable

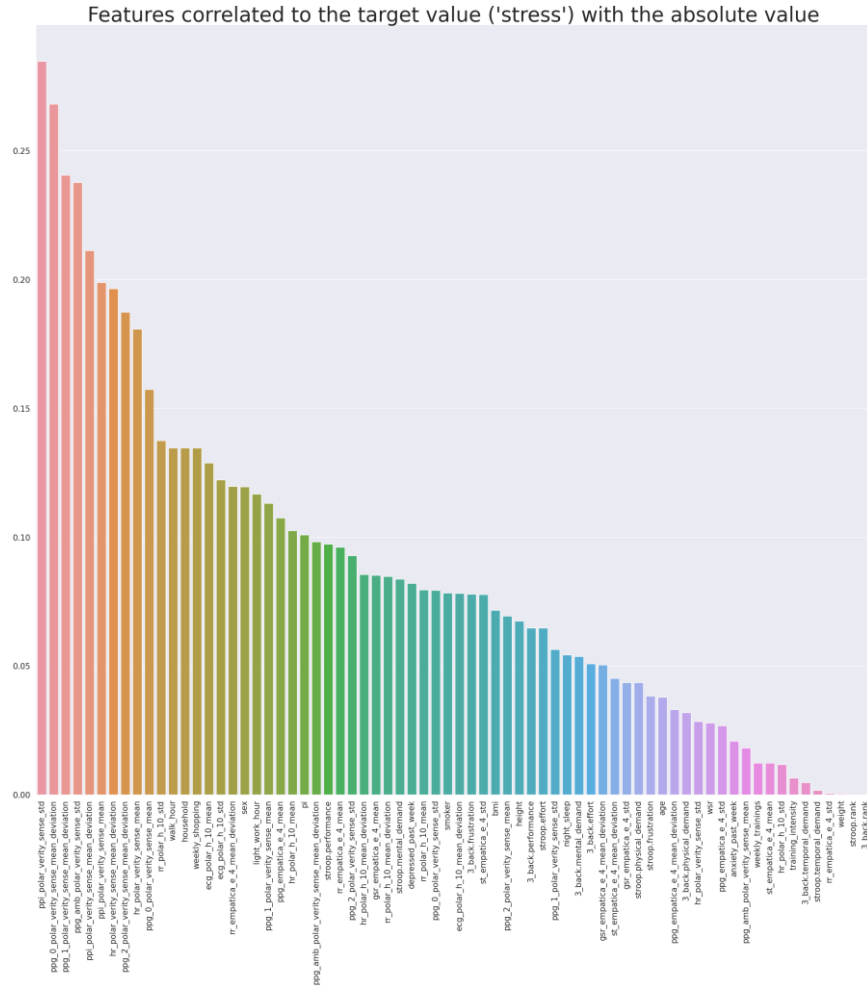


Figure 47: Feature correlation with the target variable (absolute value)

We then proceeded to identify features with a variance lower than 0.05. The following features were identified: "st_empatica_e_4_std", "wsr", "stroop.rank", "3_back.rank", "st_empatica_e_4_mean_deviation", and "gsr_empatica_e_4_std". Among these, it was determined that only the "stroop.rank" and "3_back.rank" features would be removed as they contained only one value. However, despite the variance of the other features falling below the threshold, it was decided to retain them. Previous tests indicated that in certain configurations, the models tend to utilise them to some extent.

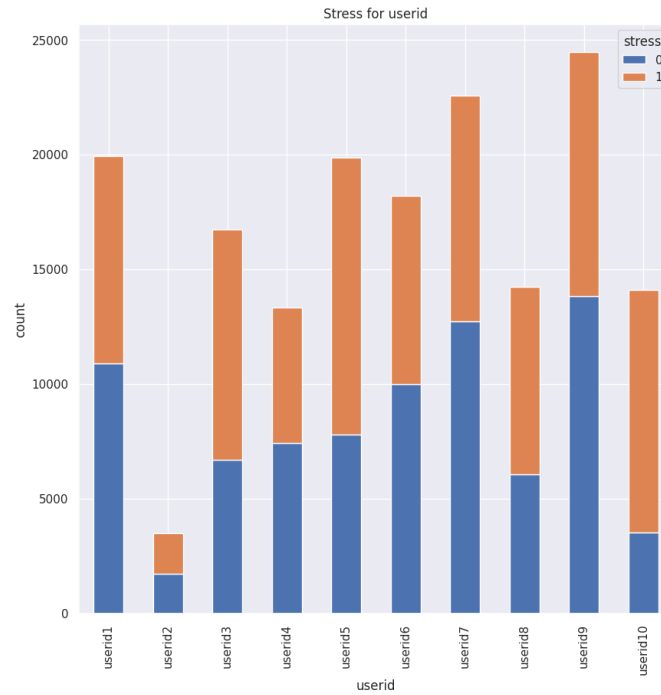


Figure 48: Observations available for a given mental stress label, grouped by user

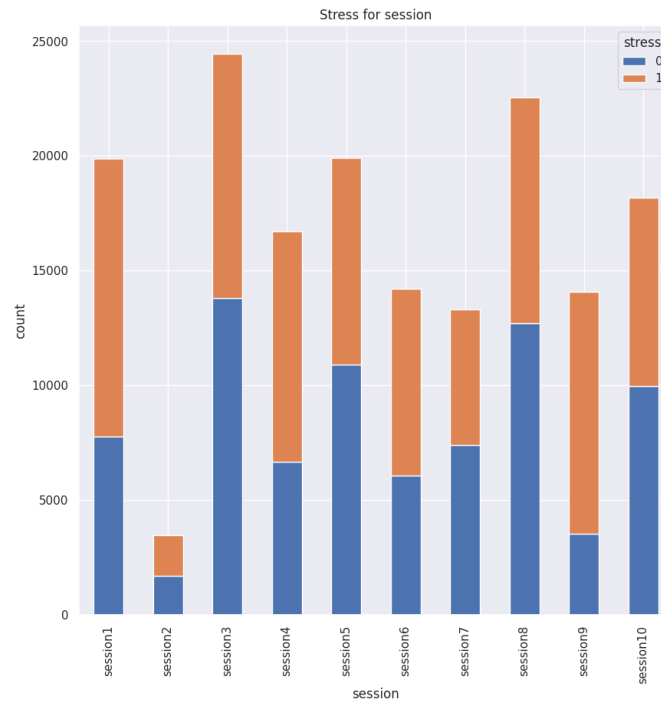


Figure 49: Observations available for a given mental stress label, grouped by session

Next, we analysed the distributions of mental stress labels per participant and per session. Plots were generated, as shown in Figure 48 and Figure 49, to illustrate the number of available observations.

3.7 Model 1: Random Forest

In all the models presented below, the accelerometer, gyroscope, and magnetometer data were excluded. This decision was made after training a model for the binary prediction of

mental stress that heavily relied on features from these three sensors. The observed reliance on these sensors could be attributed to the participants’ movement during breaks between games (they were moving arms/hands only while completing the tests).

3.7.1 Training excluding accelerometer, gyroscope, and magnetometer

In this trial, the training strategy is the same as reported in section 2.6.1, i.e., by exploiting all the available features (with no dimensionality reduction). We thus report only the data distributions (Figure 50), the accuracy and F1 scores (Figure 51), the confusion matrix (Figure 52), as the feature importance (Figure 53) as learnt by the best model (MSE: 0.0863).

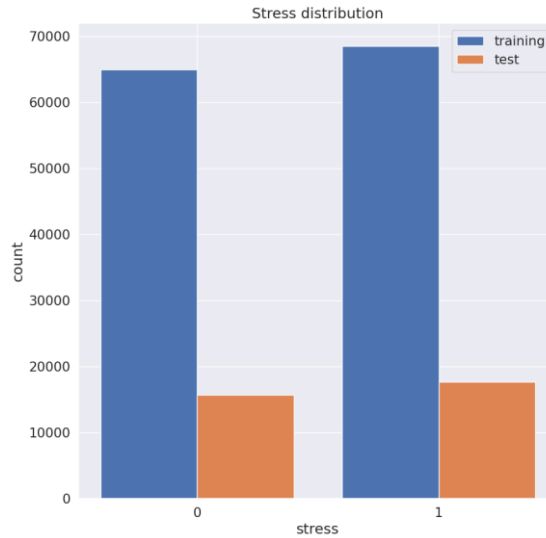


Figure 50: Number of observations available in training and test datasets, for each class

	precision	recall	f1-score	support
0	0.87	0.96	0.91	15705
1	0.96	0.87	0.91	17671
accuracy			0.91	33376
macro avg	0.92	0.92	0.91	33376
weighted avg	0.92	0.91	0.91	33376

Figure 51: Classification report of the obtained model

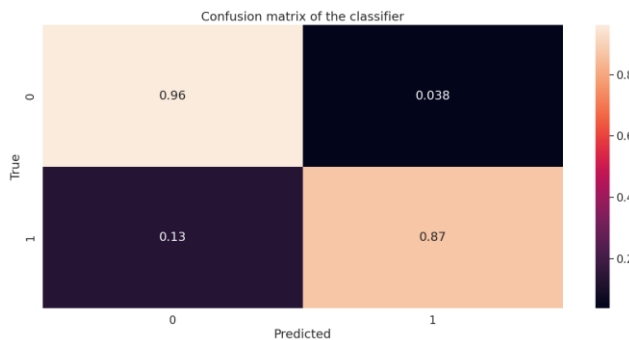


Figure 52: Confusion matrix of the obtained model

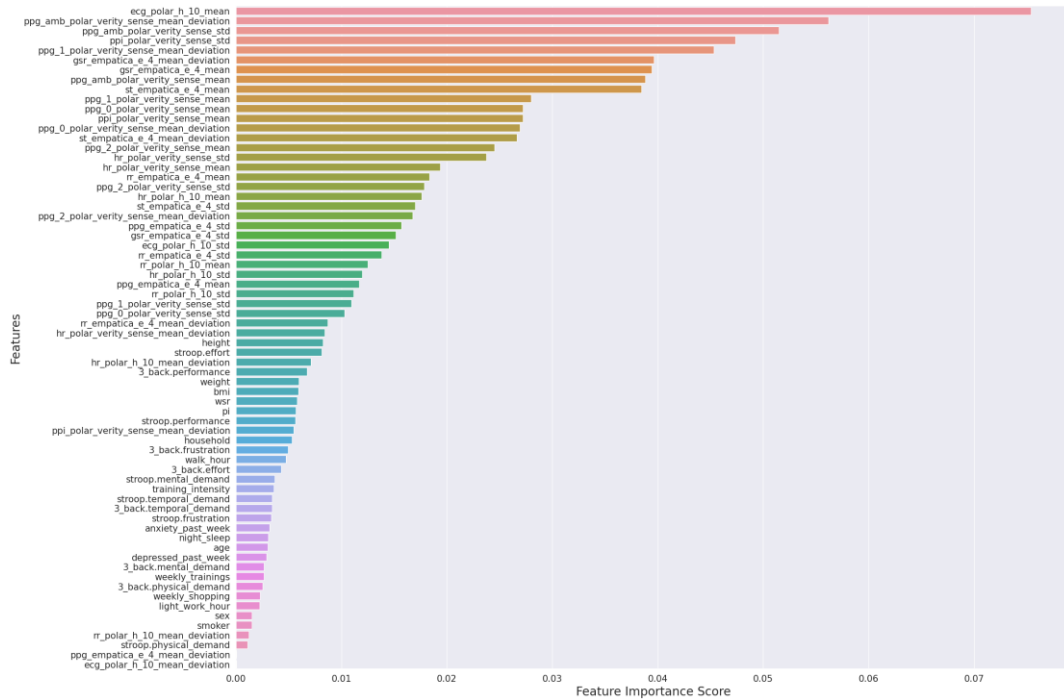


Figure 53: Feature importance according to the obtained model

The proposed model outperforms the baseline model by Villani et al, as depicted in Figure 54. The baseline has been executed by considering the RR values as returned by the different considered devices, i.e., the Empatica E4 and Polar H10.

ML model accuracy: 0.9136804889741131
 Empatica E4 natural affection based method accuracy: 0.52837368168744
 Polar H10 natural affection based method accuracy: 0.52837368168744

Figure 54: Comparison between the proposed ML model and the baseline model.

3.7.2 Training by grouping by user

In this trial, we tried to learn from some users and make prediction on the other. This strategy differs from the one presented in section 2.6.1; here, the dataset is split into train/test sets by retaining 7 participants within the training set and leaving the other 3 participants for the testing phase. We report the data distributions (Figure 55), the accuracy and F1 scores (Figure 56), the confusion matrix (Figure 57), as the feature importance (Figure 58) as learnt by the best model (MSE: 0.3768).

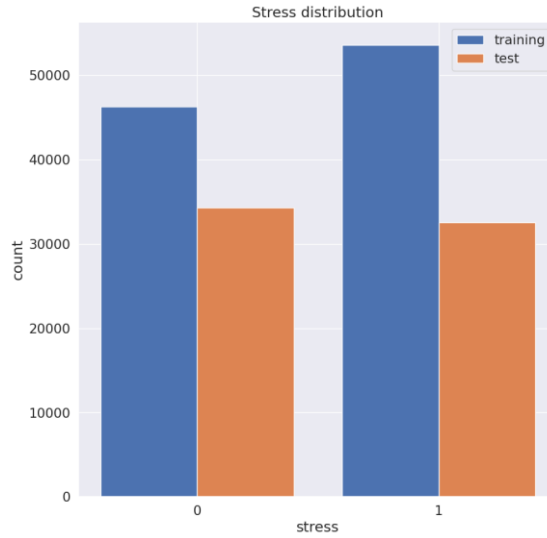


Figure 55: Number of observations available in training and test datasets, for each class

	precision	recall	f1-score	support
0	0.65	0.57	0.61	34319
1	0.60	0.68	0.64	32564
accuracy			0.62	66883
macro avg	0.63	0.62	0.62	66883
weighted avg	0.63	0.62	0.62	66883

Figure 56: Classification report of the obtained model

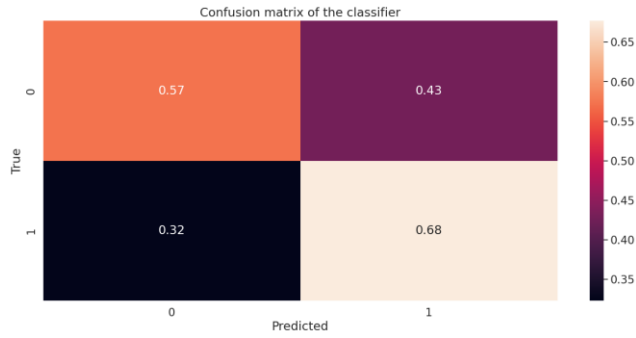


Figure 57: Confusion matrix of the obtained model

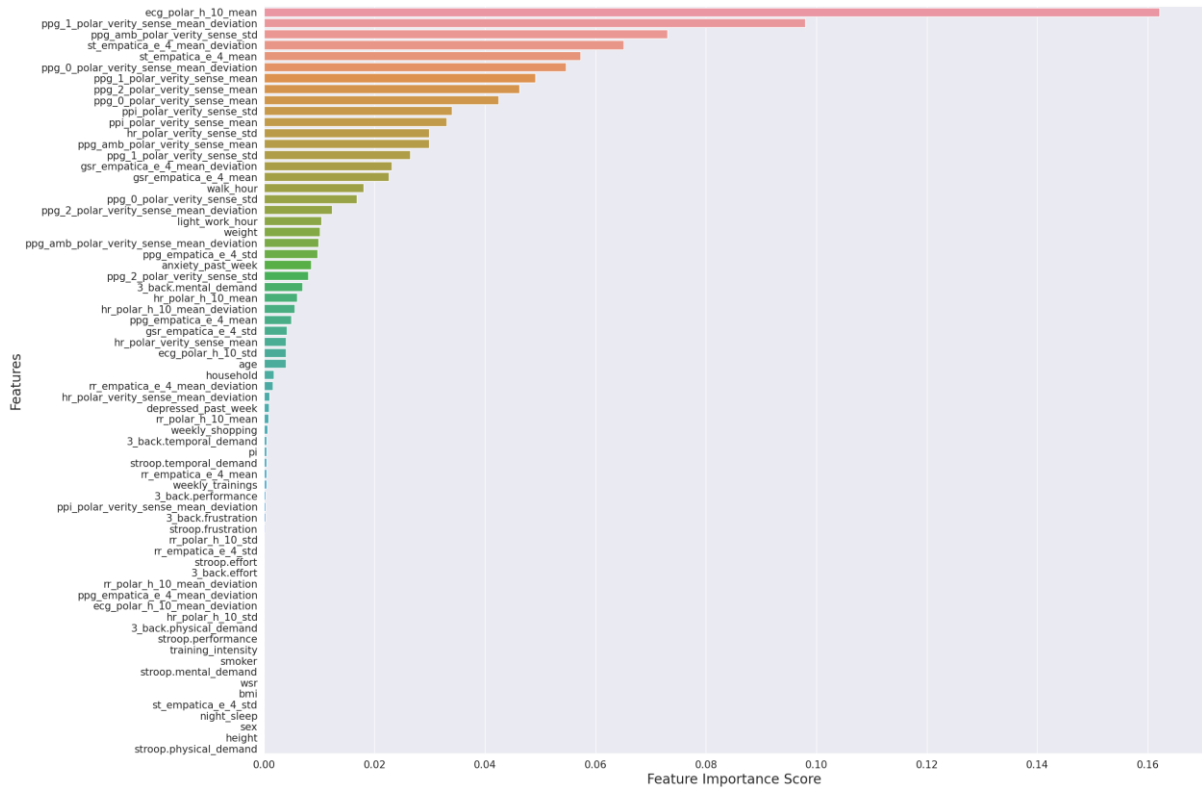


Figure 58: Feature importance according to the obtained model

Also in this trial, the proposed model outperforms the baseline model (as reported in Figure 59, considering the RR values from the Empatica E4 and the Polar H10).

ML model accuracy: 0.62311798214793
 Empatica E4 natural affection based method accuracy: 0.5392401656624254
 Polar H10 natural affection based method accuracy: 0.568799246445285

Figure 59: Comparison between the proposed ML model and the baseline model.

3.8 Model 2: Feed Forward Neural Network

In this section, the Feed Forward Neural Network models are presented. Features from the accelerometer, gyroscope, and magnetometer data were excluded as explained in section 3.7. The following models reuse the same architecture presented in section 2.7 (but with a different input size due to the different number of available features – from 50 to 70 features).

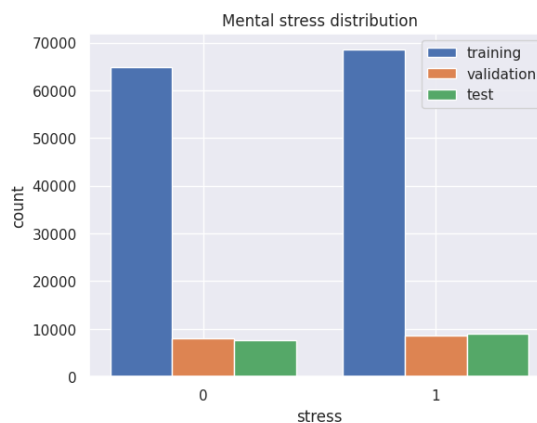


Figure 60: Number of observations available in training and test datasets, for each class

3.8.1 Results with the first version of the network

The following hyperparameters were used for the first version of the network:

- a) Optimiser: SGD
- b) Learning rate: 0.001
- c) Batch size: 1000
- d) Dropout: 0.1
- e) Epochs: 30

Results reported in Figure 61 and Figure 62 shows that the model learned well, without overfitting too much. The training loss continues to decrease, as well as the validation loss. Either taking as the benchmark metric the loss or the accuracy, the best performance was achieved at epoch 29 with a value of approximately 0.1122 for the loss, and 85.88% for the accuracy. The accuracy obtained is quite high, but it should be considered that the ongoing task is easier than the one address in section 2 (binary classification vs. multi-class classification). The confusion matrix (Figure 63) confirms that the model performs quite well.

Model with best loss and accuracy (MSE: 0.1069)

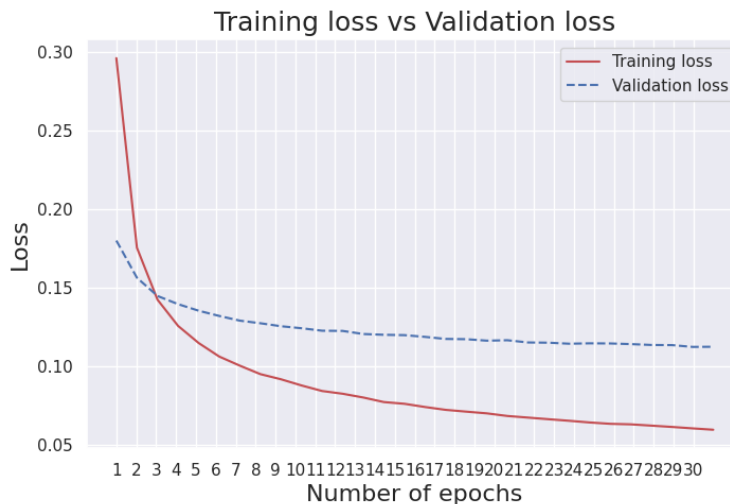


Figure 61: Training loss vs Validation loss

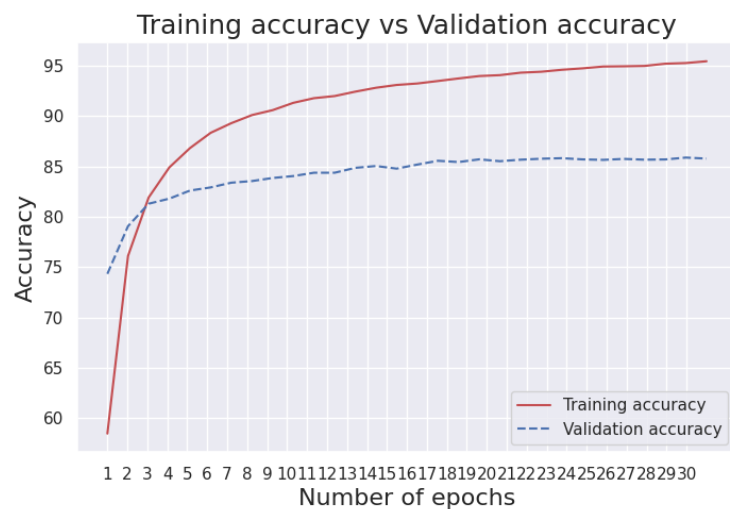


Figure 62: Training accuracy vs Validation accuracy

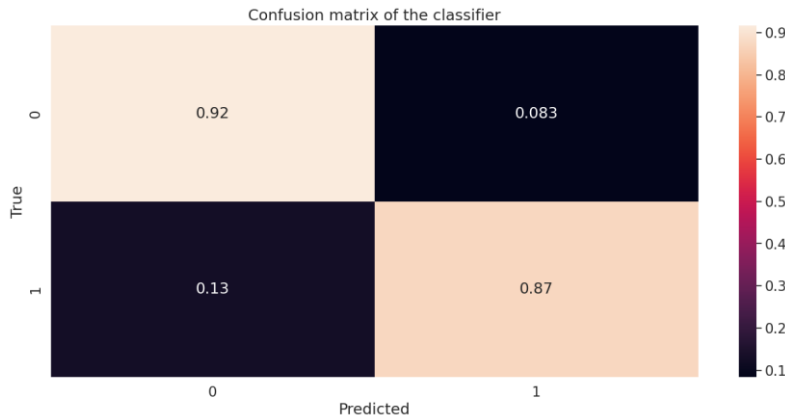


Figure 63: Confusion matrix of the best model selected through loss and accuracy

When compared with the baseline model, the proposed neural network (choosing the instance with the best loss) outperforms the baseline model (as reported in Figure 64), both when computing the affection-based method considering RR values from the Empatica E4 and the Polar H10.

ML model accuracy: 0.8930712059458163
 Empatica E4 natural affection based method accuracy: 0.540517861424119
 Polar H10 natural affection based method accuracy: 0.540517861424119

Figure 64 Comparison between the proposed NN (v1 – best loss/accuracy) and the baseline model.

3.8.2 Results with the second version of the network

In this case, the following hyperparameters were used to train the second version of the network:

- a) Optimiser: SGD
- b) Learning rate: 0.01
- c) Batch size: 1000
- d) Dropout: 0.2
- e) Epochs: 30

With this configuration, the performance slightly increases, even if the learning progress is not very smooth (Figure 65 and Figure 66 show a decreasing/increasing trend, with some oscillations). The best model is obtained after 30 epochs, scoring the best loss (0.0968) and the best accuracy (88.79%).

Model with best loss and accuracy (MSE: 0.1017)

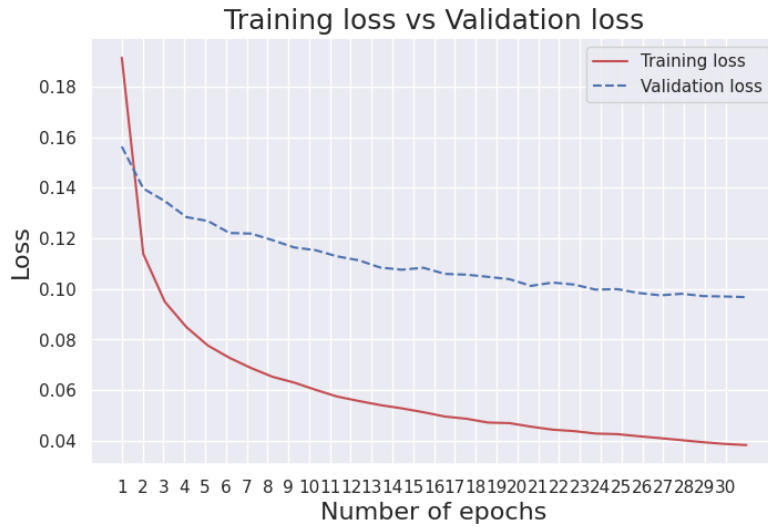


Figure 65: Training loss vs Validation loss

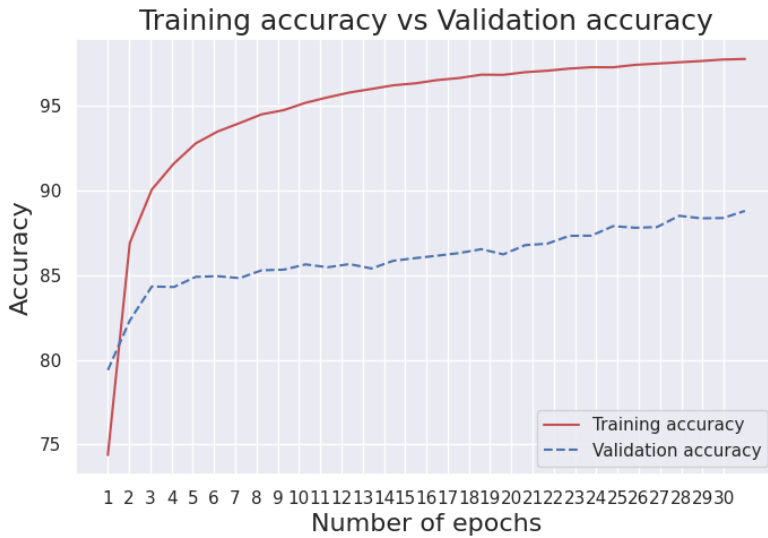


Figure 66: Training accuracy vs Validation accuracy

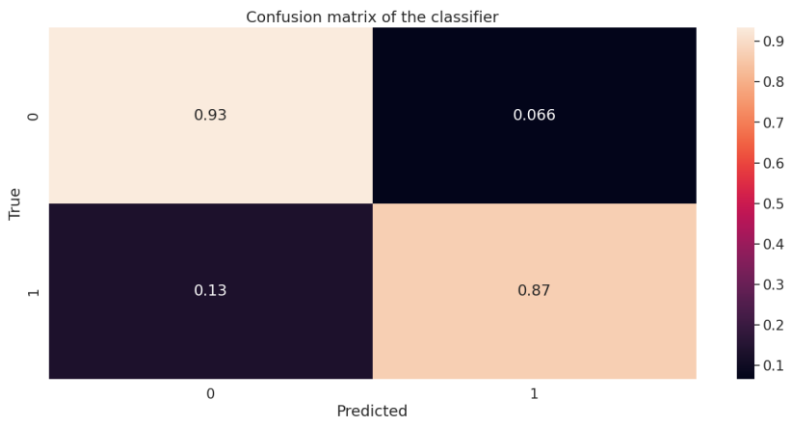


Figure 67: Confusion matrix of the best model selected through loss

The proposed neural network outperforms the baseline model (as reported in Figure 68, considering the RR values from the Empatica E4 and the Polar H10).

ML model accuracy: 0.8983457204507312
Empatica E4 natural affection based method accuracy: 0.540517861424119
Polar H10 natural affection based method accuracy: 0.540517861424119

Figure 68: Comparison between the proposed NN (v2 – best loss) and the baseline model.

3.9 Summary

In this study, two different models have been proposed to solve the task of mental stress detection: a Random Forest classifier, and a Feed-Forward Neural Network (with 2 different architectures). Different configurations and hyper-parameters have been tested to find the models with the best performance.

For both the models, different training attempts have been conducted by reusing the expertise built while building the fatigue exertion prediction model (presented in section 2), including training trials using all the available features, and the exploration of different hyperparameters. Table 13 presents the results obtained by the models in the various tested configurations.

Among the models tested, the Random Forest using all features (excluding accelerometer, gyroscope, and magnetometer) achieved the highest accuracy on the test set, reaching approximately 91%. The results obtained with the NN are comparable, obtaining an accuracy of about 90% (and a similar MSE).

Table 13: Performance of models in various configurations

Model	Test accuracy	Test MSE
Random Forest excluding accelerometer, gyroscope, magnetometer)	91%	0.0863
Random Forest (group by user)	62%	0.3768
Forward Neural Network using all features	89%	0.1069

4 FaMS module improvements

With the new ML models representing the most important improvements for the FaMS module, some other lighter improvements have been developed to make the module easier to adopt. These improvements mainly targeted a better integration with the HDT Core Infrastructure developed as part of T5.1. The HDT Core Infrastructure is a highly flexible infrastructure, able to support different configurations. Latest developments of FaMS pushed in the same direction, trying to make the module very easy to configure, both in terms of connections with the HDT Core Infrastructure, and choice of ML models to consider.

The FaMS module is release as a Docker image, which can be easily run in a Docker environment and setup using Docker environment variable. However, the dockerised application should account for flexibility, so that to allow users to change the configuration from the environment, without changing anything within the Docker image.

The new version of FaMS thus relies on Hydra,⁵ a framework enabling users in configuring complex applications. New configurations are made available, so that users can now quickly choose among different dynamic and static data sources, as well as different data brokers (e.g., MQTT, FIWARE Orion Context Broker). Also, the configuration of ML models is now external, meaning that the user can decide which model to execute, and pass their parameters directly from the environment (e.g., which features to load, for each considered model).

Finally, the code base underwent a huge refactoring, which makes it easier to find and solve bugs when found during the execution. Also, Hydra keeps a log trace of every execution, together with its configuration parameters, which helps the developers in finding and fixing bugs.

⁵ <https://hydra.cc/>

5 Conclusions

The objective of this project was to develop a machine or deep learning model capable of predicting perceived fatigue exertion in workers within a manufacturing context. Physiological data collected from wearable devices and static data capturing individual characteristics of the workers were utilised for this purpose. The aim was to identify fatigue exertion within a range of 1 to 10, where 1 represents minimum perceived fatigue exertion and 10 represents maximum perceived fatigue exertion. The project focused on inter-subject and inter-task predictions to ensure the model's applicability across different individuals and contexts.

While the valid fatigue exertion labels should range from 1 to 10, the available data only covered a range from 1 to 8. Future work intends to address this limitation by conducting new data collection campaigns to obtain data encompassing the entire fatigue exertion range from 1 to 10. Also, the project encountered some limitations due to the lack of very detailed information about the working environment (e.g., tasks, jobs). Additionally, information about the data source, including the types and placement of the wearable devices, may bring additional evidence that can improve the model performance. Future work will also cover these aspects.

With the available datasets, multiple attempts were made to develop models capable of predicting inter-subject and inter-task fatigue exertion. Overall, the Random Forest classifier outperformed the Feed-Forward Neural Network. Specifically, the models that provided the best results were those obtained by considering the time series for the data split using all the characteristics or excluding only the characteristics of the accelerometers, as well as using only the heart rate, skin temperature, and galvanic response characteristics of the skin. Among these models, no significant differences were observed, and the choice between them depends on specific requirements and needs. Some models struggled to distinguish fatigue exertion level 2, while others faced challenges with fatigue exertion level 4. When considering accuracy as the reference metric, the Random Forest model using all features achieved the highest accuracy of 65% on the test set. In contrast, evaluating performance based on MSE, the Random Forest model excluding accelerometers achieved the best results with an MSE value of approximately 0.8274 (within the fatigue exertion range of 1 to 8).

Furthermore, it was observed that the most important features for determining fatigue exertion were consistent across the different models. Skin temperature, galvanic skin response, and heart rate were generally identified as crucial dynamic data features, while grip strength, weight, and height were significant static data features. In future work, exploring additional model variants for predicting intra- and inter- subject as well as intra- and inter-task fatigue exertion would be valuable to assess the generalisation capability of the model.

Alongside the fatigue exertion level prediction, also new methods for detecting mental stress situations in workers have been designed and developed. A new dataset has been built, collecting data with ad-hoc designed experiments, aimed at stressing the participants with cognitive demanding tasks. Based on the new dataset, the same design approach as for the physical exertion prediction task has been exploited, leading to two new machine learning models for solving this task. The performance observed are higher if compared to models for physical exertion estimation, but it should be considered that the task (i.e., binary classification) is simpler than the other (i.e., multi-class classification). However, the limited amount of collected data does not allow us to draw solid conclusions, but the preliminary models represent a good starting point for further extensions and analysis. Possibly, new data

collections should be arranged in the future to mitigate some quality issues actually affecting the new dataset.

References

- [REF-01] Villani, V., Capelli, B., Secchi, C. et al. Humans interacting with multi-robot systems: a natural affect-based approach. *Auton Robot* **44**, 601–616 (2020). <https://doi.org/10.1007/s10514-019-09889-6>