AMERICAN UNIVERSITY OF BEIRUT

EXPLAINABLE MODELS FOR EMOTION RECOGNITION

by

DALIA RAGHED JABER

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
to the Department of Electrical and Computer Engineering
of the Maroun Semaan Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
December 2020

AMERICAN UNIVERSITY OF BEIRUT

EXPLAINABLE MODELS FOR EMOTION RECOGNITION

by

DALIA RAGHED JABER

Approved by:

Prof. Hazem Hajj, Associate Professor — Advisor
Electrical and Computer Engineering Department

Prof. Zaher Dawy, Professor — Member of Committee
Electrical and Computer Engineering Department

Prof. Wassim El Hajj, Associate Professor — Member of Committee
Computer Science Department

Prof. Fadi Maalouf, Associate Professor — Member of Committee
Psychiatry Department

Date of thesis defense: December 1st, 2020

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: _____ Jaber_____Dalia_____Raghed_____

Last First Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

☑ As of the date of submission

☐ One year from the date of submission of my thesis.

☐ Two years from the date of submission of my thesis.

☐ Three years from the date of submission of my thesis.

_____

Signature  Date 01/14/2021

ACKNOWLEDGEMENTS

I would like first to express my sincere gratitude and thanks to my supervisor, Prof. Hazem Hajj for always being available to help, for his support, guidance, and motivation. I would also like to thank Prof. Wassim el Hajj, Dr. Fadi Maalouf and Prof. Zaher Dawy for taking the time to be part of my thesis committee and for their helpful feedback and advice. Finally, I would like to express my appreciation and thanks to my colleagues and family for always supporting me throughout my academic journey.

ABSTRACT
OF THE THESIS OF

Dalia Raghed Jaber for Master of Engineering
Major: Electrical and Computer Engineering

Title: Explainable Models for Emotion Recognition

Despite significant advancements in artificial intelligence (AI), most machine learning (ML) solutions remain black boxes with little to no explanation of how decisions are made. To build trust in AI applications in health care, it is crucial for practitioners and patients to understand the reasons behind decisions made by ML models. In particular, there is a need for explainable AI systems for mental health. While there has been significant progress in developing stress prediction models, those models provide no explanation how they determine a prognosis. In this work, we propose a new design for an explanatory AI report of the results of automated stress assessment based on wearable sensors. Because medical practitioners and patients are likely to be familiar with blood test reports, we modeled the look and feel of the explanatory AI on those of a standard blood test report, in which the rows indicate the different physiological sources being tested, and the columns indicate the test results and associated parameters. The physiological measurements used by the AI model to generate the stress report include electrocardiogram, electromyography, electrodermal activity, respiration, and body temperature data. The test indicator results, reflecting the AI explanation, include the following indicators: the predicted stress probability, reference intervals for normal range of values for each physiological signal, warning flags that indicate results in the abnormal stress ranges, and the impact of each physiological signal to the overall stress prediction. The stress prediction and impact measures were derived using ML explainable models that show the contributions of individual features to the overall result of the model. The reference intervals and flags were then derived from those contributions. Historical studies in psychology were used to form ground truth explanations for the physiological signals. The AI explanation reports were then evaluated for usefulness and effectiveness using documented real stress and physiological study data from 14 users. The confidence in the predicted stress was reflected by the accuracy of the used ML prediction model, which came at F1-binary score of 0.78. The contributions of each physiological signal to the stress prediction were shown to correlate with ground truth. The reference intervals for stress versus non-stress were quite distinctive with little variation. In addition to these quantitative evaluations, a qualitative survey by an expert in psychiatry confirmed the confidence and effectiveness of the explanation report in understanding the different aspects of the AI system: result of stress prediction and which physiological (vital) signs were related to stressful episodes. The report also provided a source of additional medical insights into the patient's mental health.

TABLE OF CONTENTS

ILLUSTRATIONS

Figure

TABLES

Table

CHAPTER 1

INTRODUCTION

Although stress is a regular part of daily life, long-term stress can have severe consequences on health. Chronic mental stress can cause cardiovascular disease, depression, and increased susceptibility to infection [2]. The ability to detect when a person is stressed might therefore be very useful in efforts to prevent health problems, especially in patients with suicidal thoughts [1]. Several artificial intelligence (AI) systems have been proposed for early automatic stress detection using physiological measurements such as electrocardiogram (ECG) and electromyography (EMG) taken from wearable devices [2, 3, 4]. The practical use of AI systems is limited, however, because people do not always trust the automated solutions. The primary reason for the lack of trust is a lack of transparent explanations of the results produced by AI models. Because the impact of wrong diagnosis is high, health professionals and patients are reluctant to adopt technologies that are not well understood. We are interested in developing an AI-based stress evaluation model that automatically produces a report that explains the results of the evaluation in a way that is understandable and useful to human users.

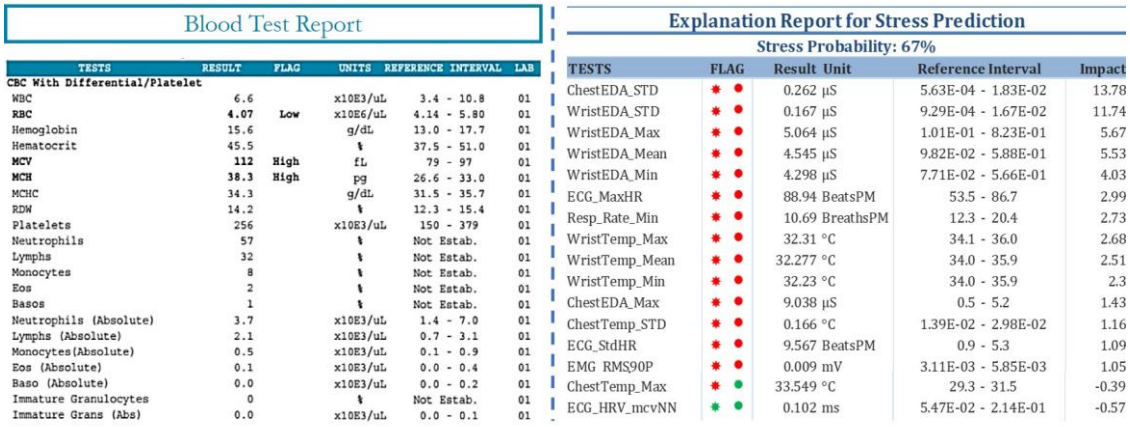
Understanding the reasons behind AI models' predictions has become so crucial that the European Union developed new data privacy rules in 2018 that state that companies that are use AI are obliged to provide either detailed explanations of individual AI algorithms or general information about how the algorithms make decisions when working with personal data [5]. Recently, there have been increasing efforts to develop explainable or interpretable AI systems, which make predictions and

behave in ways that humans can understand [6]. Despite those efforts, there are still no standard criteria by which to evaluate the interpretability of an AI system, nor is there even a clear definition of interpretability. Simple machine-learning (ML) models like decision trees, rule-based algorithms, and linear regression models may be considered interpretable, because they show the direct relationships between features and predictions. For more complex ML models, several approaches have been proposed to show the relationships, depending on the type of black-box model and the type of input data [6]. Some proposed approaches are model agnostic and can explain the outcome of any black-box model with any type of input [7, 8], whereas others focus specifically on deep neural networks used for image classification [9, 10, 11] or more general types of input [12]. One major limitation of previous interpretable AI approaches is that they fail to provide a user-centric explanation but instead focus on the mathematical relationships between features and predictions. In medicine, deep learning methods were used to create heatmaps to explain the predictions of AI systems that use medical images such as magnetic resonance images or X-ray images [31, 32, 34]. Other models were used to explain medical diagnoses by analyzing the influence of specific features on the diagnoses [33, 34]. No interpretable AI has yet been developed for stress prediction.

To address the lack of explainable AI systems for stress prediction, we propose a new design for an explainable AI system that evaluates stress using data from wearable devices. The explanatory component is inspired by medical blood test reports, which are already familiar to health care providers and patients. The predictive component is based on the findings of previous scientific studies in psychology. **Figure 1** shows a sample of the proposed AI report for stress evaluation alongside a typical blood test

report. The stress evaluation report includes the different physiological attributes that influence the overall probability that the subject is stressed and the reference ranges for each attribute. The stress evaluation report and the blood test report share several key aspects, including: the names of the individuals features that are measured directly, the measured values of the features and the corresponding units, the range of normal values for the features, and flags that indicate any abnormal values. The abnormal values on the stress report are related to stress. In addition to those attributes, the stress report gives an overall probability that the individual is in a state of stress and a quantitative measure of the influence of each measured feature, referred to as the 'IMPACT', on the overall stress probability. The overall stress probability and the IMPACT scores are expressed as percentages.

Figure 1 - Samples of a Blood Test Report [25] and a Stress Evaluation Report



Blood Test Report

TESTS	RESULT	FLAG	UNITS	REFERENCE INTERVAL	LAB
CBC With Differential/Platelet					
WBC	6.6		x10E3/uL	3.4 - 10.8	01
RBC	4.07	Low	x10E6/uL	4.14 - 5.80	01
Hemoglobin	15.6		g/dL	13.0 - 17.7	01
Hematocrit	45.5		%	37.5 - 51.0	01
MCV	112	High	fL	79 - 97	01
MCH	38.3	High	pg	26.6 - 33.0	01
MCHC	34.3		g/dL	31.5 - 35.7	01
RDW	14.2		%	12.3 - 15.4	01
Platelets	256		x10E3/uL	150 - 379	01
Neutrophils	57		%	Not Estab.	01
Lymphs	32		%	Not Estab.	01
Monocytes	8		%	Not Estab.	01
Eos	2		%	Not Estab.	01
Basos	1		%	Not Estab.	01
Neutrophils (Absolute)	3.7		x10E3/uL	1.4 - 7.0	01
Lymphs (Absolute)	2.1		x10E3/uL	0.7 - 3.1	01
Monocytes(Absolute)	0.5		x10E3/uL	0.1 - 0.9	01
Eos (Absolute)	0.1		x10E3/uL	0.0 - 0.4	01
Baso (Absolute)	0.0		x10E3/uL	0.0 - 0.2	01
Immature Granulocytes	0		%	Not Estab.	01
Immature Grans (Abs)	0.0		x10E3/uL	0.0 - 0.1	01

Explanation Report for Stress Prediction

Stress Probability: 67%

TESTS	FLAG	Result Unit	Reference Interval	Impact
ChestEDA_STD		0.262 µS	5.63E-04 - 1.83E-02	13.78
WristEDA_STD		0.167 µS	9.29E-04 - 1.67E-02	11.74
WristEDA_Max		5.064 µS	1.01E-01 - 8.23E-01	5.67
WristEDA_Mean		4.545 µS	9.82E-02 - 5.88E-01	5.53
WristEDA_Min		4.298 µS	7.71E-02 - 5.66E-01	4.03
ECG_MaxHR		88.94 BeatsPM	53.5 - 86.7	2.99
Resp_Rate_Min		10.69 BreathsPM	12.3 - 20.4	2.73
WristTemp_Max		32.31 °C	34.1 - 36.0	2.68
WristTemp_Mean		32.277 °C	34.0 - 35.9	2.51
WristTemp_Min		32.23 °C	34.0 - 35.9	2.3
ChestEDA_Max		9.038 µS	0.5 - 5.2	1.43
ChestTemp_STD		0.166 °C	1.39E-02 - 2.98E-02	1.16
ECG_StdHR		9.567 BeatsPM	0.9 - 5.3	1.09
EMG_RMS90P		0.009 mV	3.11E-03 - 5.85E-03	1.05
ChestTemp_Max		33.549 °C	29.3 - 31.5	-0.39
ECG_HRV_mcvNN		0.102 ms	5.47E-02 - 2.14E-01	-0.57

We evaluated our proposed approach with a set of qualitative and quantitative experiments. The qualitative experiments focus on the different tests and features included in the report. The qualitative assessment was based on inputs from expert psychiatrists and meant to determine whether the report provides adequate explanation for the AI decisions. In the quantitative assessments, we examined the validity of the

overall stress probability, the ranges of normal values presented in the report, the IMPACT scores, and the FLAGS. We evaluated the accuracy of the stress evaluation using leave-one-user-out cross-validation. We assessed the reliability of the range of normal values, or REFERENCE INTERVAL, by checking if the range changes when different subsets of patient data are used as the baseline. To assess the reliability of the IMPACT scores, we identified the physiological attributes that were affected by stress and their relative values during a non-stressful state in previous studies to form a ground truth reference. Finally, we assessed the accuracy of the FLAGS by checking how consistent the two FLAGS in the stress evaluation report are in indicating the same stressful state.

The key contributions of our work are:

1. A user-centric stress report based on physiological measurements. We will use AI to automatically produce a report of a patient's stress probability based on ECG, electrodermal activity (EDA), EMG, respiration, and temperature data. The report will be given in a form that is familiar to medical experts and patients and will include the stress probability, the physiological factors on which the stress probability is based, the measured value and normal range for each factor, flags indicating abnormal values, and the level of contribution of each factor to the overall stress probability.

2. An assessment of our approach. We will qualitatively and quantitatively evaluate the validity of the stress report. For the qualitative assessment, expert psychiatrists will complete a questionnaire. For the quantitative assessment, each element of the report will be evaluated on the basis of the results of previous scientific studies.

The remainder of this report is organized as follows. Section 2 presents a literature review of existing explainable AI models and automated stress prediction systems. Section 3 covers our approach to explain the output of AI-based stress evaluation systems. Section 4 presents the implementation and evaluation of our approach. Section 5 summarizes our findings and plans for future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Explainable AI Models

Approaches to make complex AI prediction models understandable to humans generally focus on clarifying the input–output relationship. Different approaches have been proposed for different types of data and prediction models. One important approach to attempt to explain any black-box model is the Additive Feature Attribution method, in which the original black-box model is approximated with a simpler model that is easily explainable. The approximation is composed of a linear combination of binary variables, as shown in Eq. *1*

$$g(z') = \phi + \sum_{i=1} \phi_i \, z_i' \quad (1)$$

where $z' \epsilon 0,1^M$, with M as the number of simplified input features; and $\phi_i \epsilon R$, which represents the contribution of feature z_i to the model's prediction. In the simplified features vector, a feature with a value of '1' is present in the subject, and a feature with a value of '0' is absent in the subject.

Another approach that is commonly used to explain black-box models is Local Interpretable Model-Agnostic Explanations (LIMEs) [7]. In the LIME approach, the input data are perturbed, and the effects of the perturbation on the output are assessed. LIME then tries to approximate the machine learning (ML) model with another model that is easily interpretable. The interpretable model is a linear combination of the input variables with some simplifications and perturbations. The LIME model presents as an

output a list of explanations, reflecting the contributions of each variable to the results of the original ML model. A weak point of the LIME approach is the instability of the explanations, which can differ greatly which small changes in the input data.

The SHapley Additive exPlanations (SHAP) approach [8] combines LIME with Shapely values [20], a concept in cooperative game theory that was developed to distribute the gains from a cooperative game to players, or features. SHAP uses locality approximation and Shapely additive values to provide an explanation for any black-box model. The method uses three criteria: local accuracy; missingness, which does not give any importance to missing features; and consistency, which makes sure that even if a model changes, the feature impact will still have the same attribution assigned. To interpret the prediction of a Convolutional Neural Network (CNN), Zhou et al. [9] introduced the concept of class activation mapping (CAM), which indicates the discriminative image regions used by the CNN that impact target classification. CAM only works on CNNs that are composed of a Global Average Pooling (GAP) layer preceding a fully connected layer that produces the output. Deep Learning Important Features (DeepLIFT) [12] is another approach that uses back-propagation to explain a CNN model. DeepLIFT decomposes the output of a neural network for a specific input by backpropagating the contribution of every feature of the input. The Layer-wise Relevance Propagation (LRP) [11] method is equivalent to DeepLIFT with the reference activation of all neurons set to zero. The main idea behind the LRP algorithm is to explain a classifier's prediction specific to a given data point by using the topology of the learned model to attribute relevance scores to components of the input.

In medicine, explainable AI applications have been developed to interpret data from imaging studies. A recent study to detect COVID-19 using chest X-ray images

[31] introduced a technique called GSInquire that created heatmaps to confirm the diagnostic features learned by the proposed COVID-net model. To study the reliability of a CNN model designed to identify brain tumors in MRI images, Pereira et al. [32] used GradCAM, an improvement of CAM, to create heatmaps that show the factors that influenced the classification of features as tumors. For computed tomography (CT) imaging, a sensitivity analysis was applied to liver CT images to explain the segmentation of tumors [34]. The analysis was performed by maximizing the target neuron using gradient ascent. Another new ML system called Prescience was introduced [33] to interpret real-time predictions to prevent hypoxemia during surgery. The Prescience model uses SHAP attribution to analyze preoperative factors and in-surgery parameters. In another study [30], a framework was proposed for the design of an explanatory display to interpret the prediction of a pediatric intensive care unit in-hospital mortality risk model. The explanation was displayed in a user-centric manner and established using Shapely values.

Explainable models have not been applied to stress prediction based on physiological sensor data. Explainable AI systems for stress prediction need to augment their explanations with additional predictive models that provide descriptions of biological factors other than the stress state *per se*.

2.2 Stress Prediction Systems

There have been several attempts to create automatic stress prediction systems, each using different features to predict or detect stress. To reduce privacy concerns and power consumption, some approaches only use data from accelerometers. For example, Garcia-Ceja et al. extracted 34 features from the time and frequency domains of

accelerometer data and fed them into several classification models including Naives Bayes, decision tree, and random forest [13]. They were able to achieve an accuracy of 71% using decision trees. In addition to accelerometer data, Giakoumis et al. included GSR and ECG data and behavioral features to predict stress and found that prediction based on the physiological data and the behavioral features was more accurate than prediction based on physiological data alone [14]. Sun et al. were able to obtain an overall accuracy of 92.4% for 10-fold cross validation using GSR, ECG, and accelerometer data [15]. Carneiro et al. added a video camera and pressure-sensitive touchscreens to accelerometers and obtained an accuracy of 78% in classifying touches as stressed or not stressed using J48 tree [16]. Bomogolov et al. predicted stress with 72.39% accuracy using a random forest classifier based entirely on call logs, Bluetooth data, and SMS data from users' mobile phones [17]. When those data were combined with GPS and Wi-Fi information, the accuracy of stress prediction increased to 86% [18].

Although stress detection has been widely studied, it is still challenging to explain the results of the detection systems in a way that is easily understandable to humans. It is important for health care professionals and patients to understand the reasons behind decisions made by AI models, because the impacts of those decisions can be serious. Many of the models described in the literature to predict mental stress use complex algorithms to achieve accurate predictions; however, the interpretability of the models tends to decrease as the accuracy increases. Hence, there is a need for models that provide explanations and interpretations for complex stress prediction.

CHAPTER 3

PROPOSED METHOD

3.1 Problem Description

The objective of this work is to provide an explanation of the stress prediction conducted by AI systems that take as input the physiological signals listed in **Table 1**. The generated explanations need to be suitable for physician and patient comprehension.

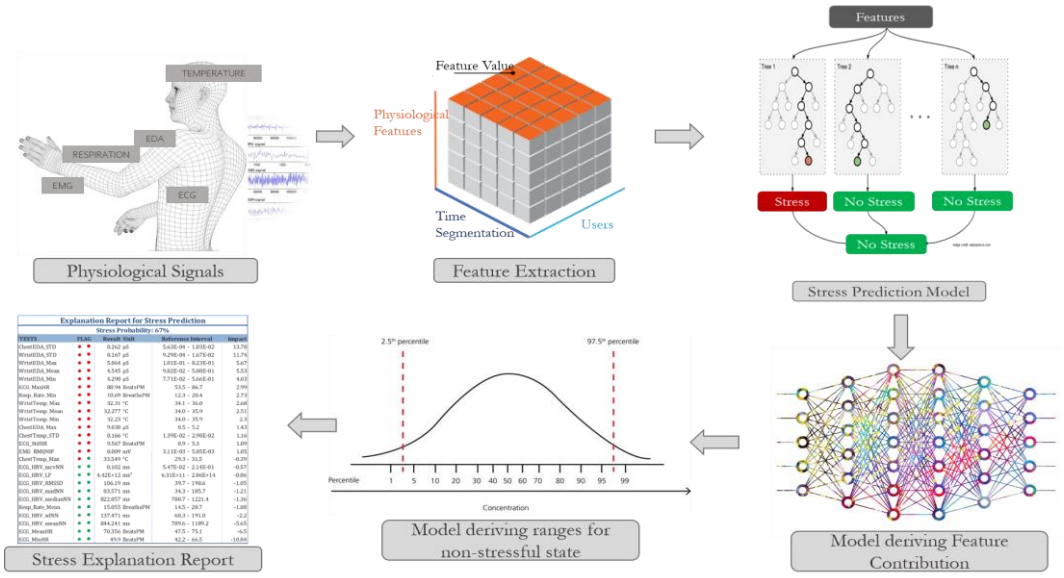
Table 1 - Physiological Measurements

Signal	**Measurement**
Electrocardiogram	Electrical activity of the heart
Electromyography	Electrical activity of muscles at rest and during contraction
Electrodermal activity	Wrist and chest skin conductance
Temperature	Wrist Temperature
Respiration	Respiration cycle and Respiration rate

There are several challenges that we aim to address. The first challenge is to determine what explanation should be displayed for physicians and patients and how the explanation should be presented. The second challenge is to develop models that can produce the necessary explanations. In order to produce the desired explanations, three models are needed (**Figure 2**). The first model extracts the desired physiological features by applying statistical signal processing to physiological data from ECG, EDA, EMG, respiration, and temperature sensors. The second model derives the contribution of each feature to the overall stress prediction using a separate, feature-based classifier

that takes as input the pre-processed features. The third model to determines the ranges of feature values that are indicative of a non-stressful state.

Figure 2 - The Proposed Solution



3.2 Proposed Explanations and Corresponding User Interface

Inspired by standard reports of blood test results, we propose to have the AI system automatically generate a report showing the measured values and normal ranges for each component of the stress assessment. The aim is to help patients and health care professionals understand which physiological factors are related to stressful episodes experienced by the patients.

3.2.1 Organization of the Report

For ease of reference, a sample blood test report is shown in **Figure 3**.

Figure 3 - An Example of a Standard Blood Test Report

TESTS	RESULT	FLAG	UNITS	REFERENCE INTERVAL
CBC With Differential/Platelet				
WBC	6.6		x10E3/uL	3.4 - 10.8
RBC	**4.07**	**Low**	x10E6/uL	4.14 - 5.80
Hemoglobin	15.6		g/dL	13.0 - 17.7
Hematocrit	45.5		%	37.5 - 51.0
MCV	**112**	**High**	fL	79 - 97
MCH	**38.3**	**High**	pg	26.6 - 33.0
MCHC	34.3		g/dL	31.5 - 35.7
RDW	14.2		%	12.3 - 15.4
Platelets	256		x10E3/uL	150 - 379
Neutrophils	57		%	Not Estab.
Lymphs	32		%	Not Estab.
Monocytes	8		%	Not Estab.
Eos	2		%	Not Estab.
Basos	1		%	Not Estab.
Neutrophils (Absolute)	3.7		x10E3/uL	1.4 - 7.0
Lymphs (Absolute)	2.1		x10E3/uL	0.7 - 3.1
Monocytes(Absolute)	0.5		x10E3/uL	0.1 - 0.9

The key aspects of the blood test report include:

- TESTS: the different blood tests included in the report
- RESULT: the measured values of the different blood tests
- FLAG: indicators of normal/abnormal test results
- UNITS: the units of the measured values
- REFERENCE INTERVAL: the range of normal test values

To make our AI-generated stress prediction report compatible with what patients and health care professionals are used to seeing, we will use a similar organization. An example of how the stress prediction report will look is shown in **Figure 4.**

Figure 4 - An Example of a Stress Prediction Report

Explanation Report for Stress Prediction

Stress Probability: 98%

TESTS	Flags		Result	Unit	Reference Interval			Impact(%)
WristEDA_STD	*	●	0.167	µS	1.334E-03	-	1.556E-02	11.37
ChestEDA_STD	*	●	0.085	µS	9.592E-04	-	1.619E-02	10.42
ECG_MaxHR	*	●	108.627	BeatsPM	60.360	-	86.482	7.19
WristEDA_Mean	*	●	3.222	µS	1.158E-01	-	5.427E-01	7.06
WristEDA_Max	*	●	3.582	µS	1.231E-01	-	9.650E-01	6.66
ECG_HRV_meanNN	*	●	645	ms	784.109	-	1125.708	6.56
ECG_MeanHR	*	●	96.669	BeatsPM	52.704	-	74.884	6.1
WristEDA_Min	*	●	2.988	µS	0.108	-	0.533	5.43
ECG_HRV_medianNN	*	●	637.143	%	768.571	-	1150.714	5
ECG_MinHR	*	●	83.74	BeatsPM	45.327	-	67.857	4.38
WristTemp_Max	*	●	32.15	°C	34.070	-	35.910	3.28
WristTemp_Mean	*	●	32.137	°C	33.993	-	35.858	2.83
WristTemp_Min	*	●	32.11	°C	33.990	-	35.830	2.76
Resp_Rate_Mean	*	●	11.568	BreathsPM	14.509	-	19.314	2.52
Resp_Rate_Min	*	●	8.641	BreathsPM	12.356	-	18.276	2.24
ChestEDA_Max	*	●	9.662	µS	0.757	-	6.810	1.98
ChestEDA_Mean	*	●	9.491	µS	0.749	-	6.635	1.52
ChestTemp_STD	*	●	0.066	°C	1.676E-02	-	3.158E-02	1.18
ECG_StdHR	*	●	6.084	BeatsPM	1.520	-	5.213	1.07
ChestEDA_Min	*	●	9.336	µS	7.469E-01	-	6.743E+00	1.03
EMG_STD	*	●	0.014	mV	3.860E-03	-	5.792E-03	0.99
EMG_RMS50P	*	●	0.014	mV	3.860E-03	-	5.790E-03	0.9
EMG_RMS90P	*	●	0.014	mV	3.880E-03	-	5.821E-03	0.82
EMG_RMS	*	●	0.014	mV	3.860E-03	-	5.800E-03	0.79
Resp_Rate_Max	*	●	14.858	BreathsPM	16.259	-	20.413	0.71
ECG_HRV_sdNN	*	●	37.514	ms	56.468	-	158.901	0.65
EMG_Max	*	●	0.056	mV	1.430E-02	-	8.386E-02	0.49
ECG_HRV_pNN20	*	●	50	%	60.859	-	89.828	0.48
EMG_Min	*	●	-0.067	mV	-4.148E-02	-	-1.624E-02	0.45
Resp_Rate_Std	*	●	2.189	BreathsPM	1.418E-01	-	9.784E-01	0.3
ECG_HRV_LF	*	●	2.98E+11	ms^2	5.79E+11	-	5.18E+12	0.3
ECG_HRV_madNN	*	●	26.429	ms	31.429	-	116.786	0.29
ECG_HRV_RMSSD	*	●	28.027	ms	48.170	-	141.432	0.21
ChestEDA_SCLMean	*	●	-0.569	µS	-28.394	-	-1.362	0.2
ECG_HRV_pNN50	*	●	3.333	ms	5.480	-	71.714	0.19
ChestEDA_SCLSTD	*	●	1.182	µS	0.628	-	10.275	0.17
ECG_HRV_HF	*	●	3.51E+11	ms^2	5.39E+11	-	4.17E+12	0.14
ECG_HRV_mcvNN	*	●	0.041	ms	3.896E-02	-	1.346E-01	0.13
EMG_NumPeaks	*	●	5905	#	6234	-	7676	0.07
ChestEDA_SCRSTD	*	*	0.325	µS	1.146E-01	-	1.566E+00	-0.09
ChestEDA_SCRMean	*	*	0.569	µS	0.898	-	24.527	-0.09
ChestTemp_Min	*	*	31.968	°C	32.080	-	34.304	-0.13
ChestTemp_Mean	*	*	32.115	°C	32.127	-	34.421	-0.14
WristTemp_STD	*	*	0.012	°C	9.522E-03	-	1.841E-02	-0.2
ChestTemp_Max	*	*	32.368	°C	31.908	-	34.511	-0.44

- STRESS PROBABILTY: the stress level of the patient in percentage, varying between 'Not stressed' (0%) and 'Extremely stressed' (100%)
- TESTS: the different physiological signals included in the report, extracted from the physiological signals listed in **Table 2**.
- RESULT: the measured values of the physiological signals, typically presented as statistical measures (e.g., mean or median over a given interval) of the raw data.

- UNIT: the units of the measured values
- REFERENCE INTERVAL: the range of normal values for the physiological signals under non-stressful conditions
- IMPACT: the percentage contribution of each physiological signal to the overall stress prediction (section *3.3* shows how the impact is calculated).
- FLAGS: indicators of normal/abnormal physiological signals. Red indicates values associated with stress, whereas green indicates values not associated with stress. Star-shaped flags represent correspondence to the REFERENCE INTERVAL. Circle-shaped flags represent the IMPACT of the test result on the overall prediction.

3.2.2 Choice of TEST Signals

The physiological measurements included in the report are commonly used in experimental procedures to study the biological effects of stress [26]. Additional features that are crucial to the explanation of the stress prediction are shown in **Table 2**. Acerbi et al. extracted several EDA and ECG features and reported the values at baseline and during and stress [21]. They then performed t-tests to identify features whose values differed between stressful and non-stressful conditions. In another study, the same procedure was followed using only EMG signals [22]. For the temperature and respiration features, statistical features are extracted including the mean, maximum, minimum, and standard deviation.

Table 2 - Stress Explanation Features

Signal	**Features**	**Description**

ECG	μ_{HR}[1], σ_{HR}, Max_{HR}, Min_{HR}	mean, standard deviation, maximum, and minimum heart rate (bpm)
	HF_{HRV}[2], LF_{HRV}	variance in HRV in the high frequency range (.15–.40 Hz); variance in HRV in the low frequency range (.04–.15 Hz)
	$\lvert\mu\rvert_{NN}$[3], σ_{NN}, Mad_{NN}, Med_{NN}, MCV_{NN}	mean of the absolute values, standard deviation, median absolute deviation, median, and median-based coefficient of variation of the successive differences between the RR intervals (interval between two heart beats)
	$RMSSD_{NN}, PNN_{20}$, PNN_{50}	root mean square of the RR intervals, number of interval differences of successive RR intervals greater than 20 ms or greater than 50 ms
EMG	μ_{EMG}[4], σ_{EMG}, $Max_{EMG,} Min_{EMG}$	mean, standard deviation, maximum, and minimum values of EMG activity in the lower trapezius
	$\#Peaks_{EMG}, RMS_{EMG}$, $RMS50P_{EMG}, RMS90P_{EMG}$	number of peaks in signal, normalized root mean square value as a percentage of the reference contraction, 50th, 90th percentile of rank-ordered root mean square values
EDA	$\mu_{WristEDA}$[5], $\sigma_{WristEDA}$, $Max_{WristEDA}, Min_{WristEDA}$	mean, standard deviation, maximum, and minimum values of EDA connected to the user's wrist
	$\mu_{ChestEDA}$[6], $\sigma_{ChestEDA}$, $Max_{ChestEDA}, Min_{ChestEDA}$	mean, standard deviation, maximum, and minimum values of EDA connected to the user's chest
	$\mu_{ChestSCL}$[7], $\sigma_{ChestSCL}$, $\mu_{ChestSCR}$[8], $\sigma_{ChestSCR}$	means and standard deviations of the skin conductance level and skin conductance response
Respiration	$\mu R_{espRate}$[9], $\sigma_{RespRate}$, $Min_{RespRate}, Max_{RespRate}$	mean, standard deviation, minimum, and maximum of the respiration rate

[1] Hear Rate
[2] Heart Rate Variability
[3] NN-Intervals (time interval between R peaks)
[4] Electromyography
[5] Electrodermal activity recorded from the Wrist
[6] Electrodermal activity recorded from the Chest
[7] Skin Conductance Level recorded from the Chest
[8] Skin Conductance Response recorded from the Chest
[9] Respiration Rate

Temperature	$\mu_{WristTemp}$[10] , $\sigma_{WristTemp}$, $Max_{WristTemp}$, $Min_{WristTemp}$	mean, standard deviation, maximum, and minimum values of the temperature measured from the user's wrist

3.2.4 Ground Truth Data

We evaluated the results of our stress prediction model using ground truth data collected from experiments that tested the effects of stress on physiological measurements [21,22,23,24]. The ground truth data provide information about which physiological features can be used as stress indicators. We compared the list of stress indicators obtained experimentally to the list of features determined by our model to indicate stress.

The previous studies recorded the mean values and standard deviations of features measured during stressful and non-stressful conditions. They then used Kruskal–Wallis tests or Friedman tests to compare the data between the two conditions to identify significant differences ($p<0.05$). They found that the significant features were μ_{NN}, μ_{HR},σ_{HR}, $RMSSD_{HRV}$, $PNN50_{HRV}$, and μ_{EDA}. **Table 3** shows the normal ranges, stress ranges, and p-values of the significant features. In order to extract stress levels of subjects using the EMG signal of the upper trapezius muscle, an experimental procedure was performed in which the subjects were faced with three different stressful situations: a calculation task, a logical puzzle task, and a memory task. The EMG signal was found to be a meaningful feature to detect stress, as its amplitude was higher during stress than during relaxed conditions. The same was found for the EMG root mean square values. Therefore, on the basis of the experiments performed, we determined that the following features show elevated EMG amplitude during stressful situations: μ_{EMG},

[10] Temperature recorded from the Wrist

RMS_{EMG}, and $RMS50P_{EMG}$. The respiratory system's response to stress was reported in [23,24], showing that the respiration rate $\mu_{RespRate}$ increases during stress.

Table 3 - ECG Features Shown Experimentally to Indicate Stress

Physiological Feature	Range for No Stress	Range for Stress	p-Value*
$\mu_{NN}(ms)$	788±126	642±96	0.005
$\mu_{HR}(BPM)$	78.45±12.38	95.54±13.69	0.005
$\sigma_{HR}(BPM)$	6.43±1.15	10.48±3.88	0.001
$RMSSD_{HRV}(s)$	0.04±0.02	0.03±0.01	0.018
$pNN50_{HRV}(s)$	22.89±19.44	7.35±4.98	0.043

* Significant difference between groups ($p<0.05$)

3.2.5 Online Analytical Processing (OLAP) Customization

Our stress evaluation report allows for different levels of customization that are common with decision support systems. The detailed list of physiological measurements can be treated as a multi-dimensional OLAP data warehouse. Different levels of extracts and aggregations can be generated and customized to fit users' needs. For example, a simple aggregate custom report might include only heart rate, respiration, and body temperature.

3.3 Model to Derive the Contributions of each Feature to The Stress Probability

An important aspect of the stress evaluation report is the IMPACT, or indication of how much each factor contributes to the overall stress probability. To calculate the impact for each factor, we customized the SHAP model, where the total probability of stress $P_X(Stress)$ for each set of TEST measurements X is computed as the sum of the mean probability $P_{Avg}(Stress)$ and the individual contributions of each TEST feature:

$$P_X(Stress) = P_{Avg}(Stress) + \sum_{i \in F_1,\ldots,F_N} \phi_i \quad (2)$$

where F represents the choice of physiological feature, and N represents the number of features for observation X. $P_{Avg}(Stress)$ represents the probability of a random person being stressed. The ϕ_i, also known as the SHAP value, is used to derive the percentage contribution of each feature. A positive value indicates that the feature reinforces the prediction of stress, whereas a negative value indicates a negative contribution, which is an indication of non-stress. Those contributions indicate deviation from the average probability of stress $P_{Avg}(Stress)$.

The SHAP ϕ_i values for each feature i can be calculated using any ML classifier by removing (nullifying) the features i one at a time and then computing the resulting predictions. In our model, we used a random forest classifier. Mathematically, the ϕ_i is computed as follows:

$$\phi_i = \sum[f_{(S \cup \{i\})}(x_{(S \cup \{i\})}) - (f_s(x_s)] \quad \left(\frac{|S|!\,(|M| - |S| - 1)!}{|M|!}\right) \qquad (3)$$

where S is a set of indexes in z' (as seen in Eq. 1), M is the set of all input features, x_S represents the values of the input features in the set S, and $f_{(\,)}$ represents the hypothesis function for the classifier. To obtain the SHAP values, a model f_s is trained with the feature i withheld, and another model $f_{(S \cup \{i\})}$ is trained with that feature present. Then, the predicted values from both models are compared to the current input x_S.

The IMPACT measure is calculated as the percentage of the features' contributions ϕ_i as follows:

$$IMPACT_{i,X}(\%) = \left(\frac{\phi_{(i,X)}}{\sum_{Features}|\phi_X|}\right) \times 100 \qquad (4)$$

The $P_{Avg}(Stress)$ can be computed from historical training data by computing the percentage of individuals who are stressed, or the average of the stress probability:

$$P_{Avg}(Stress) = Mean(y_{train}) \qquad (5)$$

where y_{train} represents true labels of stress predictions for individuals available in historical training data.

The authors of SHAP also proposed KernelSHAP and TreeSHAP and provided many global interpretation methods. KernelSHAP is an approach to estimate Shapely values inspired by local surrogate models, which are interpretable models used to explain the predictions of any black-box ML model. With KernelSHAP, it will be possible to use any classification model to provide the stress prediction. As for the TreeSHAP, it provides interpretation for any tree-based model and has a faster implementation than KernelSHAP. TreeSHAP reduces the computational complexity from $O(TL2^M)$, the complexity in KernelSHAP, to $O(TLD^2)$, where T is the number of trees, L is the maximum number of leaves in any tree, and is D the maximal depth of any tree. In addition to being computationally faster, TreeSHAP allows the creation of different visualizations that can help users understand the interpretation. Therefore, we used TreeSHAP as the model that assigns the feature contribution and the, tree-based, random forest classifier as our prediction model.

Random Forest Classifier for Stress Prediction

The measurements in the RESULTS column of the stress evaluation report are used as inputs to the random forest classification model, which indicates if the user is

stressed or not stressed according to each measurement. The random forest is an ensemble method used for classification or regression. It is trained using a bagging method, which consists of randomly selecting a subset of the training set, fitting a decision tree to each subset, and finally combining the results. For classification, the random forest uses the majority votes for the class prediction; because each tree provides one vote, the final vote can be the mode or the most frequent class predicted by each tree. When working with an imbalanced dataset, a version of the random forest classifier known as the 'balanced random forest' is highly useful. The balanced random forest model randomly under-samples each bootstrap sample to balance the labels. Finally, a leave-one-user-out cross-validation scheme is employed where the data of one user are held out for testing while the data of the rest of the users are used for training.

3.4 Model to Extract the Stress Ranges and Reference Intervals

To determine if the measurements are within a non-stressful range, our model provides ranges for each TEST that are related to stress and non-stress, respectively. Such ranges are useful to show what the normal values are for each feature and when the measurements might indicate stressful conditions.

We derive the ranges using the IMPACT values generated for each observation in the training dataset. First, we separate the feature values by their assigned IMPACT values. Then, we group the ones with positive values in a 'Stress Group' and the ones with negative values in a 'No Stress Group'. We then perform a t-test to make sure that there is a significant difference between the two groups of values. Then, similarly to how many laboratory tests define the Reference Interval, we use a non-parametric approach and take the values falling at the 2.5 and 97.5 percentiles in the No Stress

Group as the lower and upper limits of the REFERENCE INTERVAL, respectively. For the ’stress interval’, we use the values falling at the 2.5 and 97.5 percentiles in the Stress Group.

CHAPTER 4

RESULTS AND DISCUSSIONS

We evaluated our explainable AI design for a stress evaluation report through a set of qualitative and quantitative experiments.

Figure 5 - Sample of proposed explainable AI report for stress

Explanation Report for Stress Prediction

Stress Probability: 98%

TESTS	Flags		Result	Unit	Reference Interval			Impact(%)
WristEDA_STD	*	●	0.167	µS	1.334E-03	-	1.556E-02	11.37
ChestEDA_STD	*	●	0.085	µS	9.592E-04	-	1.619E-02	10.42
ECG_MaxHR	*	●	108.627	BeatsPM	60.360	-	86.482	7.19
WristEDA_Mean	*	●	3.222	µS	1.158E-01	-	5.427E-01	7.06
WristEDA_Max	*	●	3.582	µS	1.231E-01	-	9.650E-01	6.66

The quantitative assessments aimed to evaluate the reliability and accuracy of the following aspects of the explainable AI report, a sample is shown in **Figure 5**:

- STRESS PROBABILITY: To test this aspect, we used a standard ML evaluation approach as described in section 4.2.1.
- REFERENCE INTERVAL: To determine how robust the REFERENCE INTERVAL is to changes in the input data, we compared the REFERENCE INTERVALs created using two different subsets of test results, as described in section 4.2.2.
- IMPACT: To assess the accuracy of the IMPACT values, we examined the correlations between the IMPACT values and other stress indicators obtained from studies that examined what physiological measurements are affected by stress. The results are described in section 4.2.3

- FLAGS: To assess the accuracy of the FLAGS as indicators of whether the measurements for a particular factor are indicative of a stressful state, we tested how consistently the two FLAGS for each feature indicated the same stressful state. The results are described in section 4.2.4

The above evaluations were performed using a 4-fold cross validation to ensure balanced subsets of data with sufficient observations. The accuracy of this model, known as the system's confidence level, is the accuracy of the stress prediction model which can be considered as a historic accuracy based on historic data.

The qualitative assessment aimed to determine whether the report provides adequate explanation for the decisions of the AI. In the qualitative assessment, expert psychiatrists were asked the following questions:

1. Does the psychiatrist have the components needed to extract the explanation as captured by the report components: STRESS PROBABILITY, TESTS, REFERENCE INTERVALS FLAGS, and IMPACT?

2. Are any data missing or are more features needed?

3. Does the report provide the AI explanation needed for psychiatrists with examples?

4. Does the report provide the AI explanation needed for patients with examples?

5. Is the report user-friendly from the perspectives of experts and patients?

6. Can the explainable reports be useful for additional medical applications such as tracking patients' stress over time or providing other medical insights about the relationships between physiological signals and stress?

The details of the qualitative assessment section are described in section 4.3.

4.1 Dataset and Explainable Feature Extraction

The WESAD (Wearable Stress and Affect Detection) dataset [26] consists of different physiological measurements recorded during stressful and relaxed conditions. It contains physiological and motion data recorded from wrist-worn and chest-worn devices. The devices used are the RespiBAN Professional[11], which is placed around the subject's chest, and the Empatica E4[12], which is worn on the subject's non-dominant hand. The modalities include EDA and temperature data from an Empatica C4 device. The RespiBAN device provides data on respiration; ECG; EDA recorded on the rectus abdominis, considering that the abdomen has a high density of sweat glands; EMG recorded on the upper trapezius muscle on both sides of the spine; and temperature recorded on a sensor placed on the sternum.

Data were collected from 15 graduate students in a laboratory setting. Each subject experienced three conditions:

1. Baseline: Users were provided neutral reading material (e.g., magazines).
2. Amusement: Users watched a set of funny videos.
3. Stress: Users were exposed to the Trier Social Stress (TSST), which is used to induce stress in participants. The TSST generally includes three phases: an anticipatory speech preparation, speech performance, and verbal arithmetic.

[11] http://www.biosignalsplux.com/en/respiban-professional
[12] http://www.empatica.com/research/e4/

We eliminated faulty measurements, such as missing data caused by failures in signal recording. The features listed in **Table 2** were extracted from the different physiological raw signals using the numpy, Neurokit [27], and Biosppy [28] libraries in Python. Neurokit is a Python toolbox for statistics and signal processing of data from ECG, EDA, EMG, and EEG. Biosppy is a Python toolbox for bio-signal processing. We extracted data for 42 features, each with 1640 measurements taken over 90-second intervals. The data had an imbalance with 19.7% stress labels. The F_1 score was used as the evaluation metric.

4.2 Quantitative Evaluation

This section details the quantitative evaluation of the STRESS PROBABILITY, REFERENCE INTERVAL, IMPACT, and FLAGS components of the stress report. All the evaluations were performed using a 4-fold cross validation to ensure balanced subsets of data with sufficient observations.

4.2.1 Evaluation of the STRESS PREDICTION

The STRESS PREDICTION is made using the balanced random forest classifier. To evaluate the classifier, the data was divided into four subsamples and a 4-fold cross-validation approach was followed. Because the dataset was imbalanced, with 19.7% of the labels representing the class 'stress', we chose the F_1 binary score metric, which only reports results for the stress labels. The results of the cross validation are shown in **Table 4**. The average F_1 binary score was 0.75, which is an indication of high accuracy but less than what was achieved in the literature using different input features [8].

Table 4 - Evaluation of the Balanced Random Forest Classifier on the Binary Classification Task: Stress vs. Non-Stress

4-Fold Validation	F1-Binary Score
Fold 1	0.93
Fold 2	0.63
Fold 3	0.91
Fold 4	0.64
Average Score	**0.78**

4.2.2 Evaluation of the REFERENCE INTERVAL

The REFERENCE INTERVAL is defined by the range of values in healthy, non-stressed individuals. The STRESS INTERVAL, on the other hand, includes the test results of stressed individuals. The intervals were determined using the method described in section 3.4. We followed a statistical approach to create the REFERENCE INTERVAL from the No-Stress Group. The 42 features along with their intervals are shown in **Table 5**. We evaluated the REFERENCE INTERVAL by 1) validating that the Stress and No-Stress Groups, separated by the sign of the IMPACT, were independent, belonging to two different distribution and 2) evaluating the robustness of the REFERENCE INTERVAL.

To check if the values assigned to the Stress Group and No-Stress Group belonged to two different distributions with two independent ranges, we performed a t-test for each feature in the training dataset. The p-values obtained for the features are shown in **Table 5**. For all tests, the p-value was less than 0.05, which confirmed that the measured values for each feature were significantly different between the stressful condition and the non-stressful condition.

Table 5 - Intervals and P-Values for the Values of each Feature under Stressful and Non-Stressful (Reference) Conditions

Feature	Stress Interval	Reference Interval	P-Value
$Max_{ChestEDA}$	11.27 ± 7.99	3.79 ± 3.03	1.35 E-110
$\mu_{ChestEDA}$	10.73 ± 7.5	3.7 ± 2.95	5.12 E-118
$Min_{ChestEDA}$	11.12 ± 7.84	3.75 ± 3	1.21 E-103
$\mu_{ChestSCL}$	-4.33 ± 4.11	-14.88 ± 13.52	2.18 E-128
$\sigma_{ChestSCL}$,	1.65 ± 1.5	5.45 ± 4.82	8.51 E-161
$\mu_{ChestSCR}$	15.54 ± 15.3	12.72 ± 11.82	6.45 E-57
$\sigma_{ChestSCR}$	5.65 ± 5.47	0.84 ± 0.73	9.69 E-139
$\sigma_{ChestEDA}$	0.15 ± 0.14	0.01 ± 0.01	8.9 E-180
HF_{HRV}	2.72 E+11 ± 2.57 E+11	2.35 E+12 ± 1.82 E+12	5.12 E-227
LF_{HRV}	3.03 E+11 ± 3.03 E+11	2.88 E+12 ± 2.3 E+12	4.87 E-226
Mad_{NN}	21.43 ± 15.72	74.11 ± 42.68	1.02 E-204
MCV_{NN}	0.03 ± 0.02	0.09 ± 0.05	4.28 E-196
$\lvert\mu\rvert_{NN}$	701.81 ± 82.84	954.91 ± 170.8	1.3 E-219
Med_{NN}	693.93 ± 80.36	959.64 ± 191.07	5.41 E-214
PNN_{20}	33.61 ± 28.55	75.35 ± 14.49	1.19 E-220
PNN_{50}	15.15 ± 15.15	38.6 ± 33.12	1.57 E-167
$RMSSD_{NN}$	28.54 ± 18.23	94.8 ± 46.63	2.69 E-230
σ_{NN}	39.89 ± 25.97	107.69 ± 51.22	2.92 E-218
Max_{HR}	105.74 ± 18.24	73.42 ± 13.06	4.72 E-229
μ_{HR}	91.57 ± 15.91	63.79 ± 11.09	7.57 E-229
Min_{HR}	80.45 ± 14.26	56.6 ± 11.27	5.56 E-215
σ_{HR}	9.04 ± 3.72	3.37 ± 1.85	8.91 E-228
Max_{EMG}	1.04 E-02 ± 8.33 E-02	4.91 E-02 ± 3.48 E-02	3 E-68
μ_{EMG}	1.215 E-07 ± 7.94 E-07	- 1.24 E-07 ± 6.41 E-07	4.04 E-91
Min_{EMG}	- 9.72 E-02 ± 7.28 E-02	- 2.88 E-02 ± 1.27 E-02	7.38 E-146
$\#\,Peaks_{EMG}$	6416.5 ± 866	6955 ± 721	1 E-49
RMS_{EMG}	1.09 E-02 ± 5.06 E-03	4.83 E-03 ± 9.7 E-04	1.16 E-226

$RMS50P_{EMG}$	1.09 E-02 ± 5.05 E-03	4.83 E-03 ± 9.7 E-04	3.82E-226
$RMS90P_{EMG}$	1.09 E-02 ± 5.04 E-03	4.83 E-03 ± 9.7 E-04	3.43E-225
σ_{EMG}	1.09 E-02 ± 5.06 E-06	4.83 E-03 ± 9.7 E-07	2.56 E-226
$Max_{RespRate}$	13.11 ± 3.36	18.34 ± 2.08	1.04 E-229
$\mu_{RespRate}$	11.68 ± 2.69	16.91 ± 2.4	4.95 E-231
$Min_{RespRate}$	9.82 ± 2.33	15.32 ± 2.96	7.3 E-232
$\sigma_{RespRate}$	2.08 ± 1.19	0.56 ± 0.42	3.62 E-225
$Max_{WristEDA}$	5.07 ± 4.05	0.55 ± 0.43	4.65 E-223
$\mu_{WristEDA}$	4.36 ± 3.69	0.33 ± 0.21	2.73 E-232
$Min_{WristEDA}$	4.23 ± 3.56	0.32 ± 0.21	2.78 E-232
$\sigma_{WristEDA}$	0.17 ± 0.16	0.01 ± 0.01	4.87 E-212
$Max_{WristTemp}$,	31.79 ± 2.12	34.99 ± 0.92	2.88 E-175
$\mu_{WristTemp}$	31.69 ± 2.1	34.93 ± 0.93	1.65 E-178
$Min_{WristTemp}$	31.66 ± 2.15	34.91 ± 0.92	5.33 E-173
$\sigma_{WristTemp}$	0.06 ± 0.04	1.39 E-02 ± 4.44 E-03	6.13 E-211

Because the REFERENCE INTERVAL is obtained using the existing observations, it is dependent on the data used. Therefore, it is important to determine if the range would be different if it were based on another set of observations. We evaluated the robustness by performing again a 4-fold cross validation. In each fold, the REFERENCE INTERVAL from each of the training and testing subsets are generated. We compare these intervals using the relative percentage different (RPD) method which evaluates the change in the REFERENCE INTERVAL. For each feature, we computed the RPD between the intervals generated using the respective subsets with Eq. 6:

$$RPD_{feature} = \frac{|\mu_{RI_A} - \mu_{RI_B}|}{2\mu_{RI_B}} \times 100 \qquad (6)$$

By computing the RPD for each feature, we obtained of 16.8% total difference from the cross-validation as seen in Table 6. between the intervals. Because that difference is relatively small, we concluded that the REFERENCE INTERVAL is robust to changes in the data used to calculate it and is therefore reliable.

Table 6 - Evaluating the robustness of the reference interval

4-Fold Validation	Total RPD (%)
Fold 1	20.5
Fold 2	15.4
Fold 3	15.4
Fold 4	16.2
Average RPD	**16.8**

4.2.3 Evaluation of the IMPACT

The IMPACT of each feature was generated using Eq. 1 and Eq. 3, which are based on the SHAP method. The accuracy and success of the SHAP method were proven outside of this paper [8]. The IMPACT can be positive or negative, indicating that the corresponding feature contributes to an increase or decrease in the overall stress probability, respectively. We evaluated the IMPACT parameter from two perspectives: its effectiveness as an indicator of stress and its ability to provide insights into the causes of stress in a given individual.

4.2.3.1 Effectiveness of the IMPACT values as indications of stress:

To demonstrate the ability of the IMPACT parameter to explain how each feature affects stress, we examined the correlation between the IMPACT values for the features in our report and the results of previous studies. The previous studies found that the

following features were affected by stress: μ_{HR},σ_{HR}, $RMSSD_{HRV}$, $PNN50_{HRV}$,μ_{EDA}, μ_{EMG}, RMS_{EMG}, $RMS90P_{EMG}$, and $\mu_{RespRate}$. Those studies recorded for some of the features the range of values that indicated a normal or non-stressful state. For those features, the experimental reference intervals provide insight on whether the test result is indicative of a stressful or normal state. We tested whether the IMPACT parameter could provide the same information by creating a contingency table showing the relationship between the test results that were assigned a positive or negative IMPACT value and the test results that fell within or outside the experimental reference interval. We then performed a Chi-squared test for each feature. We also used the 4-fold cross validation to create a contingency table for each subset of data. A sample of the results of one of the folds is shown in **Table 7**.

The experiment showed what the normal and stressful ranges were for the ECG and EDA features (**Table 3**); however, for the EMG features and respiration rate, they only specified if the feature values increased or decreased with stress, without providing normal ranges. Therefore, for those features, the REFERENCE INTERVAL used in the Chi-squared test was the one generated by our model, as shown in **Table 5**.

We performed the Chi-squared test on each feature of the testing data in each fold with the null hypothesis that the two categories separated on the basis of IMPACT and the REFERENCE INTERVAL were not correlated. A sample of the computed p-values and the contingency matrix are shown in **Table 7**. In each of the 4-folds, all of the tests resulted in a p-value < 0.05, indicating that the null hypothesis was not supported by the data. Therefore, we rejected the null hypothesis and confirmed a correlation between the results of using the REFERENCE INTERVAL and the

IMPACT, respectively, as stress indicators. Thus, the IMPACT was found to be an effective parameter to indicate stress.

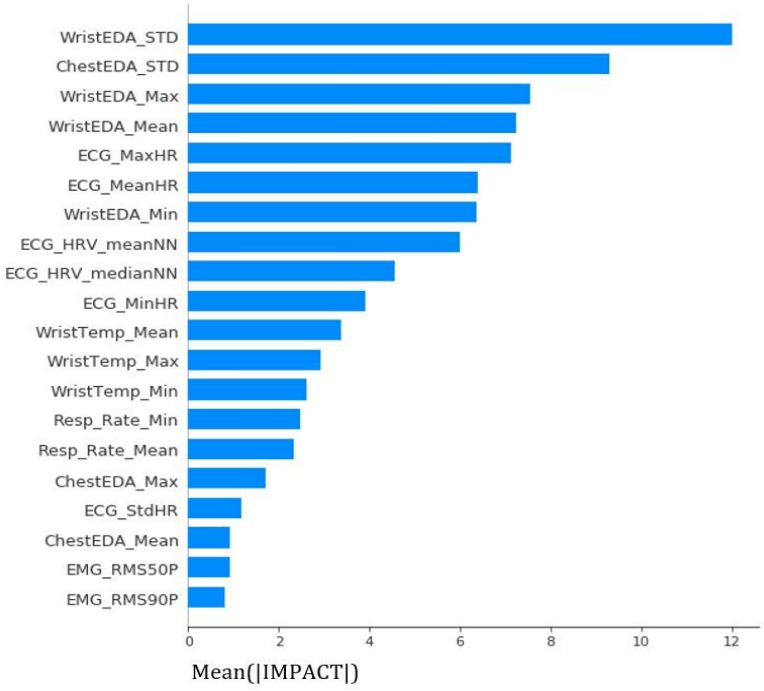
Table 7 - Results of Chi-Squared Tests for SHAP Evaluation of Stress Prediction

	Impact > 0	**Impact < 0**	**p-value**
$pNN50_{HRV} \notin$ 'Ref. Int.'	5	69	1.76E-07
$pNN50_{HRV} \in$ 'Ref. Int.	23	12	
$RMSSD_{HRV} \notin$ 'Ref. Int.'	2	57	2.86E-16
$RMSSD_{HRV} \in$ 'Ref. Int.'	48	13	
$\mu_{RespRate} \notin$ 'Ref. Int.'	72	2	2.86E-25
$\mu_{RespRate} \in$ 'Ref. Int.'	0	46	
$\mu_{WristEDA} \notin$ 'Ref. Int.'	25	0	2.18E-25
$\mu_{WristEDA} \in$ 'Ref. Int.'	1	94	
$\mu_{WristTemp} \notin$ 'Ref. Int.'	117	1	3.53E-11
$\mu_{WristTemp} \in$ 'Ref. Int.'	0	2	
$\mu_{NN} \notin$ 'Ref. Int.'	34	6	1.64E-21
$\mu_{NN} \in$ Ref. Int.'	0	80	
$\mu_{HR} \notin$ 'Ref. Int.'	17	25	1.06E-04
$\mu_{HR} \in$ 'Ref. Int.'	7	71	
$\sigma_{HR} \notin$ 'Ref. Int.'	39	26	7.60E-06
$\sigma_{HR} \in$ 'Ref. Int.'	53	2	
$RMS_{EMG} \notin$ 'Ref. Int.'	81	0	5.804e-26
$RMS_{EMG} \in$ 'Ref. Int.'	1	38	
$RMS50P_{EMG} \notin$ 'Ref. Int.'	80	0	4.75e-25
$RMS50P_{EMG} \in$ 'Ref. Int.'	2	38	
$\mu_{EMG} \notin$ 'Ref. Int.'	8	22	1.57E-03
$\mu_{EMG} \in$ 'Ref. Int.'	4	86	

4.2.3.2 Insights provided by the IMPACT

Figure 6 provides a summary of the mean IMPACT values assigned to each feature from all observations. The length of the bar represents the average impact of the feature on stress. The results show that the main physiological indicators of stress are related to the electrical heart activity and the skin conductance measured from the chest or the wrist.

Figure 6 - Average Impact of Physiological Features on Stress



4.2.4 Evaluation of the FLAGs

The two FLAG columns in the stress report inform the patient and health care professionals if any measures should be taken regarding the corresponding feature as it relates to stress. We evaluated the consistency between the two FLAG indicators. Then, we evaluated the insights provided by the FLAGS into the causes of stress.

4.2.4.1 Consistency between the two FLAGs

We extracted data for four factors from the sample report in **Figure 3** to illustrate the evaluation (**Figure 7**). The star-shaped FLAGs are associated with the REFERENCE INTERVAL, whereas the circle-shaped FLAGs are associated with the IMPACT. If the star-shaped FLAG is green, then the measured value of the feature is within the REFERENCE INTERVAL. Red star-shaped FLAGs indicate values that are outside the REFERENCE INTERVAL. If the circle-shaped FLAG is red, then the effect of the feature at the measured level is to increase stress. If the circle-shaped FLAG is green, then the effect of the feature at the measured level is to decrease stress. Because both FLAGS are supposed to indicate signs of stress, they should be consistent for each feature.

Figure 7 - Test Results Extracted from a Sample Report

Explanation Report for Stress Prediction

Stress Probability: 98%

TESTS	Flags		Result	Unit	Reference Interval			Impact(%)
ECG_MaxHR	✸	●	108.627	BeatsPM	60.360	-	86.482	7.19
EMG_Max	✸	●	0.056	mV	1.430E-02	-	8.386E-02	0.49
ChestTemp_Mean	✸	✸	32.115	°C	32.127	-	34.421	-0.14
WristTemp_STD	✸	✸	0.012	°C	9.522E-03	-	1.841E-02	-0.2

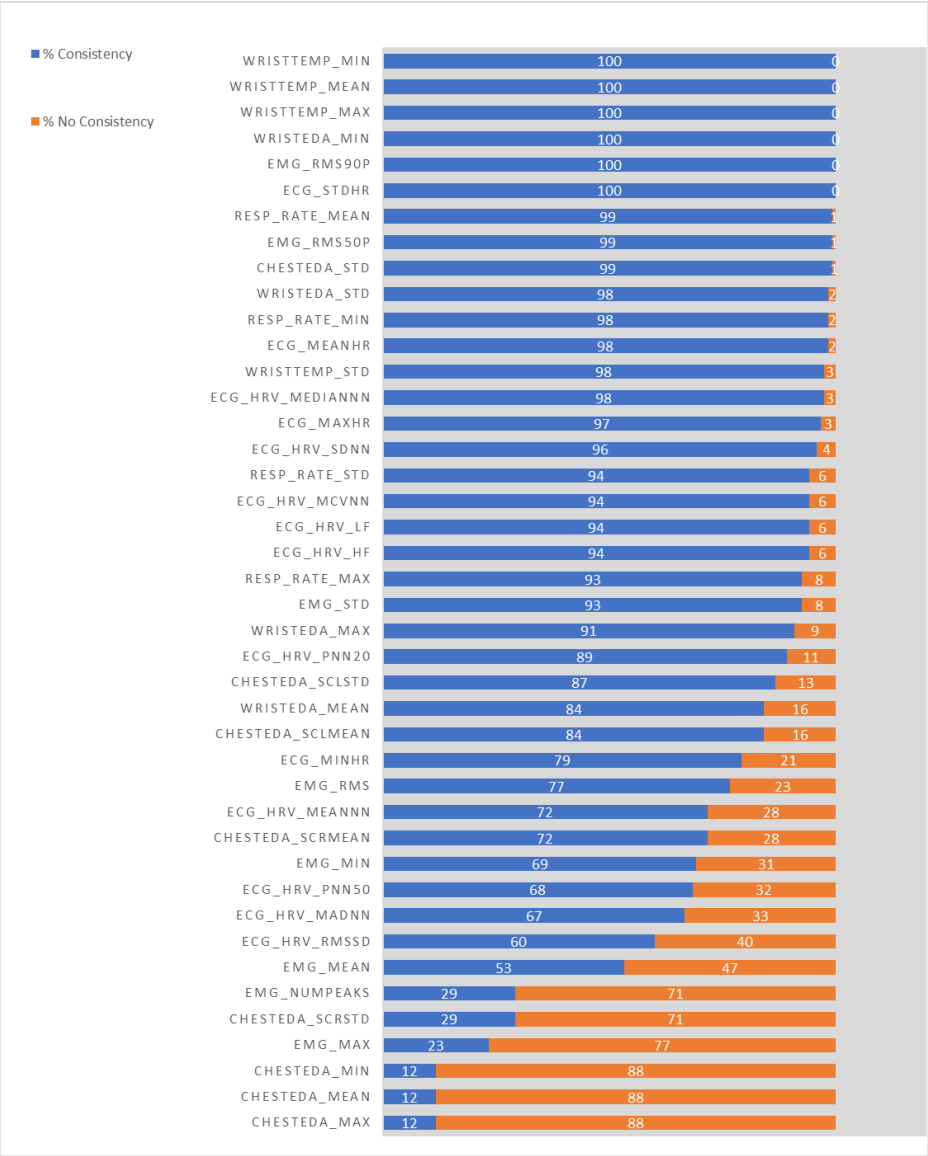
The FLAGS consistency was again performed by dividing the data into 4 subsets. In each subset, the report was generated per observation of data and the percentage of flag consistency was evaluated. The percentage of consistency per subset of data is represented in **Table 8.** Overall, there was 80% consistency between the two FLAGS.

Table 8 – Flag evaluation through consistency check

4-Fold Validation	Consistency (%)
Fold 1	81
Fold 2	79
Fold 3	80
Fold 4	80
Average Score	**80**

The bar charts shown in **Figure 8** show the percentages of reports in the training data which shows the consistency (blue bars) and inconsistency (orange bars) between the FLAGS for each feature. The results showed that the FLAGS with the most inconsistency were mainly associated with the features extracted from the chest EDA and the EMG signal. Four features had inconsistency greater than 50%. Features with high inconsistency would not be considered good stress indicators compared with other features with low inconsistency

Figure 8 - Consistency of the Two FLAGS



4.2.4.2 Insights provided by the FLAGS

The FLAGs in the stress prediction report might help explain the predicted stress probability. To illustrate that, we consider an example report generated for one individual (**Figure 8**). In that report, the model predicted that the user was stressed with a probability of 63%. The values registered for the ECG signal and chest EDA indicate a stressful state, which is represented by the positive IMPACT and the red FLAGS. The

37% model uncertainty is due to the features that had green FLAG indicators, which include the minimum, maximum, and mean values of the EDA signal recorded from the wrist.

Figure 9 – Test Results Extracted from a Sample Report showing insights provided by the FLAGS

Explanation Report for Stress Prediction

Stress Probability: 63%

TESTS	Flags		Result	Unit	Reference Interval			Impact(%)
WristEDA_STD	✸	●	0.036	µS	0.00133	-	0.01556	8.84
ChestEDA_STD	✸	●	0.054	µS	0.00096	-	0.01619	7.53
ECG_MaxHR	✸	●	120.798	BeatsPM	60.3603	-	86.4816	6.35
ECG_HRV_meanNN	✸	●	635.893	ms	784.109	-	1125.71	5.31
…	…	…	…	…	…	…	…	…
ChestTemp_STD	✸	●	0.026	°C	0.01676	-	0.03158	-0.45
WristEDA_Min	✸	●	0.322	µS	0.10772	-	0.53307	-10.33
WristEDA_Max	✸	●	0.458	µS	0.12307	-	0.96501	-13.81
WristEDA_Mean	✸	●	0.403	µS	0.11577	-	0.54269	-14.17

4.3 Qualitative Assessment

The qualitative assessment aimed to determine whether the psychiatrists and patients can understand the prediction of the AI system. A questionnaire was provided to an expert psychiatrist to provide their evaluation on the explanation report. The questionnaire was accompanied by instructions on how to interpret and read the report in addition to a description of each TEST in the report, found in Appendix A.

The following section will include a summary of the questionnaire's result, the full questionnaire and answers can be found in Appendix A-4.

4.3.1 Does the psychiatrist have the components needed to extract the explanation as captured by the report components?

The expert psychiatrist assessed the report parameters and physiological attributes used to be moderately to extremely useful to help them and the patient understand how the model is making its decision. One of the TESTS signals was labeled as not useful: "EDA recorded from the Chest" taking into consideration the practicality of collecting these measurements from the patients. However, in this work we are considering that the equipment to collect all of the measurements is available and focusing on making sure that the report provides the needed information for the psychiatrists to understand how the system provided its prediction.

4.3.2 Does the report provide the AI explanation needed for psychiatrists and patients?

The report was found to be providing the needed explanation for the psychiatrist and patient. In addition, the OLAP approach was found to be moderately important in providing a simpler explanation to the patient. However, the TESTS were thought to be neither clear nor unclear for the patients.

4.3.3 Are any data missing or are more features needed?

No additional features or components were found to be required.

4.3.4 Is the report user-friendly from the perspectives of experts and patients?

The evaluation of the report was provided as follows: the report's display and the instruction manual were easy to follow. As for the report's organization it was found to be a little confusing.

4.3.5 Can the explainable reports be useful for additional medical applications?

The expert psychiatrist considered the report to be very useful for a medical diagnosis. It could be successfully used to track patient's stress over time. In addition, the explanation report allows to study the relation between physiological signals and stress.

4.4 Discussion on Difference in Reference Intervals

The Reference Interval per physiological measurement, indicating the no-stress range, might be different between genders and more specifically it might be different per individual. In this section, we aim to study the difference in reference intervals generated by gender and then per individual for some features. For this analysis, we will consider the features that were assigned a positive impact higher than 5 % in the report of Figure 4. These features are: $\sigma_{ChestEDA}$, $\sigma_{WristEDA}$, Max_{HR}, $\mu_{WristEDA}$, $Max_{WristEDA}$, $|\mu|_{NN}$, and μ_{HR}. We aim to study if a significant difference is found between the reference intervals generated:

a. Based on gender
b. Per individual.

By generating the reference interval using the data of each individual separately, we found that the reference intervals of $\sigma_{Wrist_{EDA}}$, $\sigma_{Chest_{EDA}}$ and Max_{HR} showed the higher difference between individuals, compared to the other studied features.

However, if we compare the reference intervals of the same features by separating the subjects into Males and Females, we found that the main difference in reference intervals was in the $|\mu|_{NN}$, μ_{HR} and Max_{HR}. However since the data was collected from 3 Females and 11 Males and since we have few inputs per individual, we

cannot confirm our analysis as a larger dataset is required to draw much reliable insights.

CHAPTER 5

CONCLUSION

In this work, we provided an explanation for stress prediction by AI systems based on physiological measurements. We aimed to address the limitation apparent in the stress prediction literature, which is the lack of sufficient explanation of the prediction to allow patients and health care professionals to trust the diagnosis. To make AI-based stress evaluation more user-friendly and medically beneficial, we propose a method to provide an explanatory report that is configurable on the basis of users' needs. Users would be able to know what biological features had the most influence on the results of the stress evaluation in addition to any health-related abnormalities.

Several challenges are addressed in this thesis. We determined what explanation to display for health care professionals and patients and also how to present the explanation. In addition, we developed AI models that can produce the necessary explanations. The explanations are presented in an explanatory report that lists the different physiological attributes that influence the stress probability, similarly to how the results of blood tests are presented. We used models that determine a mathematical relationship between the data and the prediction to provide information about the influence of each feature on the overall results of the stress evaluation. The physiological measurements used in the stress report include signals related to heart activity, muscle activity, body temperature, and skin conductance. The report uses the same physiological features that are commonly used in experiments to study the biological effects of stress.

The effectiveness of the report was evaluated using a quantitative and a qualitative assessment. The stress prediction accuracy was shown to be comparable to state of the art at an F1-score of 0.78. The contributions of each physiological signal to the stress prediction was shown to correlate with ground truth. The evaluation of the reference interval showed that the chosen intervals were reliable. In addition to these quantitative evaluations, a qualitative survey with a psychiatrist confirmed the confidence and effectiveness of the explanation report in understanding the stress prediction made by the AI system.

Our future work on interpretable AI-based stress evaluation will include the addition of more explanatory features related to the emotional states of the patient, such as sadness, relaxation, anxiousness, or happiness. In addition, the implementation of a user-study that takes into consideration all the missing parameters that would have been useful in our analysis, such as a larger dataset. Since it will allow separating the data of individuals based on gender and age group and obtain enough observations per user for a better analysis and more accurate results.

For the proposed user-study, the number of participants should be around 30 subjects, separated between 15 Males and 15 Females. We would separate them into 4 groups to compare between the normal and stress related physiological measurements based on gender and age : (1) Females between 18-25, (2) Females between 26-35, (3) Males between 18-25 and (4) Males between 26-35. The collected measurements will include the Respiration Signal, the ECG signal, the EMG signal collected from the Trapezius muscle and the EDA and Temperature measured from the wrist.

Two different experiments can be performed. The first experiment would be in a controlled environment which includes a series of relaxed and stressful tasks to be performed by the

subjects. For the stress conditions, the users will be exposed to the Trier Social Stress Test (TSST) as well as puzzles and logical tasks. The duration of the stress experiments will be 2 hours. As for the relaxed conditions, the subjects will be provided neutral reading materials such as magazines for 20 minutes, they will be required to watch a set of funny video clips for 15 minutes, for amusement and finally they will perform controlled breathing exercises after each stress experiment in the aim of returning to a close to neutral/no stress state. The duration is 7 minute and to be performed after each stress experiment. The labels will be collected using three self-reports: PANAS, STAI (State-Trait Anxiety Inventory) and Short Stress State Questionnaire (SSSQ).

Another type of experiment can be performed to collect the needed data in the "wild". The users will be asked to collect around 10 hours of measurements. More data is required in this experiment to make sure we have enough stress labels and since in the "wild" there are many factors that could lead to the collection of incorrect/faulty measurements (such as disconnected device, low battery, …). Every 45 minutes the user will be asked to fill out the PANAS and the STAI questionnaires,

REFERENCES

[1] R. Sioni and L. Chittaro, "Stress Detection Using Physiological Sensors," in Computer, vol. 48, no. 10, pp. 26-33, 2015.

[2] J. Wijsman, B. Grundlehner, H. Liu, H. Hermens and J. Penders, "Towards mental stress detection using wearable physiological sensors," Annual International Conference of the IEEE Engineering in Medicine and Biology Society,2011.

[3] M. Choi, G. Koo, M. Seo and S. W. Kim, "Wearable Device-Based System to Monitor a Driver's Stress, Fatigue, and Drowsiness," IEEE Transactions on Instrumentation and Measurement, vol. 67, no. 3, pp. 634-645, 2018.

[4] Ghaderi, J. Frounchi and A. Farnam, "Machine learning-based signal processing using physiological signals for stress detection," 2015 22nd Iranian Conference on Biomedical Engineering,Tehran, 2015, pp. 93-98

[5] "General Data Protection Regulation," European Commission, 2016. [Online]. Available: https://eur-lex.europa. eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. [Accessed 5 June 2020].

[6] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, "A survey of methods for explaining black box models," ACM computing surveys (CSUR), vol. 51, no. 5, p. 93, 2019.

[7] M. T. Ribeiro, S. Singh and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2016, pp. 1135-1144.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765-4774.

[9] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, Springer, 2014, pp. 818-833.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning deep features for discriminative localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921-2929.

[11] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PloS one, vol. 10, no. 7, p. e0130140, 2015.

[12] Shrikumar, P. Greenside and A. Kundaje, "Learning important features through propagating activation differences," in Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 3145-3153.

[13] E. Garcia-Ceja, V. Osmani and O. Mayora, "Automatic Stress Detection in Working Environments From Smartphones' Accelerometer Data: A First Step", IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 4, pp. 1053-1060, 2016.

[14] D. Giakoumis et al., "Using Activity-Related Behavioural Features towards More Effective Automatic Stress Detection", PLoS ONE, vol. 7, no. 9, p. e43571, 2012.

[15] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, “Activity-aware mental stress detection using physiological sensors,” 2nd Int. ICST Conf. Mobile Comput. Appl. Serv.,vol. 76, pp. 211–230,2012

[16] D. Carneiro, J. Castillo, P. Novais, A. Fernández-Caballero and J. Neves, "Multimodal behavioral analysis for non-invasive stress detection", Expert Systems with Applications, vol. 39, no. 18, pp. 13376-13389, 2012.

[17] Bogomolov, B. Lepri, M. Ferron, F. Pianesi and A. S. Pentland, "Pervasive stress recognition for sustainable living," 2014 IEEE International Conference on Pervasive Computing and Communication Workshops,Budapest, pp. 345-350,2014.

[18] G. Bauer and P. Lukowicz, "Can smartphones detect stress-related changes in the behaviour of individuals?," 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, Lugano, pp. 423-426,2012.

[19] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 (2018).

[20] Shapley, Lloyd S. “A value for n-person games.” Contributions to the Theory of Games 2.28 (1953): 307-317.

[21] G. Acerbi, E. Rovini, S. Betti, A. Tirri, J. F. Ronai, A. Sirianni, J. Agrimi, L. Eusebi, and F. Cavallo, “A Wearable System for Stress Detection Through Physiological Data Analysis, ” Italian Forum of Ambient Assisted Living, pp. 31-50, 2016.

[22] J. Wijsman, B. Grundlehner, J. Penders, and H. Hermens, “Trapezius muscle EMG as predictor of mental stress,” ACM Transactions on Embedded Computing Systems, vol. 12, no. 4, pp. 1–20, 2013.

[23] D. Widjaja, M. Orini, E. Vlemincx and S. Van Huffel, "Cardiorespiratory Dynamic Response to Mental Stress: A Multivariate Time-Frequency Analysis", Computational and Mathematical Methods in Medicine, vol. 2013, pp. 1-12, 2013. Available: 10.1155/2013/451857.

[24] [2]W. Suess, A. Alexander, D. Smith, H. Sweeney and R. Marion, "The Effects of Psychological Stress on Respiration: A Preliminary Study of Anxiety and Hyperventilation", Psychophysiology, vol. 17, no. 6, pp. 535-540, 1980. Available: 10.1111/j.1469-8986.1980.tb02293.x.

[25] K. Kloss, “Health and Wellness Testing Example Results,” Health Testing Centers, 04-Dec-2018. [Online]. Available: https://www.healthtestingcenters.com/health-and-wellness-testing-example-results/ .

[26] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K.Van Laerhoven. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection., International Conference on Multimodal Interaction, ACM, pp. 400-408. 2018.

[27] Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lesspinasse, F., Pham, H., Schölzel, C., & S H Chen, A.“NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing”(2020) [Online]. Available: https://github.com/neuropsychology/NeuroKit. [Accessed: 25-May-2020]

[28] Carreiras C, Alves AP, Lourenço A, Canento F, Silva H, Fred A, et al., “BioSPPy - Biosignal Processing in Python”(2015) [Online]. Available: https://github.com/PIA-Group/BioSPPy. [Accessed: 25-May-2020].

[29] E. PMC, “Defining laboratory reference values and decision limits: populations, intervals, and interpretations,” Europe PMC. [Online]. Available: http://europepmc.org/articles/PMC3739683. [Accessed: 31-Aug-2020].

[30] Barda, C. Horvat and H. Hochheiser, A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare, BMC Medical Informatics and Decision Making, vol. 20, no. 1, 2020. Available: 10.1186/s12911-020-01276-x.

[31] L. Wang and A. Wong, COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images, arXiv preprint arXiv:2003.09871, 2020.

[32] S. Pereira, R. Meier, V. Alves, M. Reyes, C.A. Silva, “Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment”, in Understanding and interpreting machine learning in medical image computing applications, Springer,2018, pp 106-114.

[33] S. Lundberg et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nature Biomedical Engineering, vol. 2, no. 10, pp. 749-760, 2018. Available: 10.1038/s41551-018-0304-0.

[34] V. Couteaux, O. Nempont, G. Pizaine, I. Bloch, “Towards Interpretability of Segmentation Networks by Analyzing DeepDreams”, in Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Springer, 2019, pp. 56–63.

APPENDIX

QUESTIONNAIRE FOR PSYCHIATRISTS

1. Instructions Manual: Description of Report Components

Explanation Report for Stress Prediction								
Stress Probability: 63%								
TESTS	**Flags**		**Result**	**Unit**	**Reference Interval**			**Impact(%)**
WristEDA_STD	✸	●	0.036	µS	0.001334	-	0.015556	8.84

In this section the different aspects of the report.

- **TESTS**: Type of physiological Signal. These measurements include attributes extracted from the physiological signals.
- **RESULT**: Current measurement of each physiological attribute in 'TESTS' represents the value of the features extracted from the raw recorded signals for this prediction during a period of 90 seconds. These features are typically statistical measures of the raw data.
- **UNITS**: Units of physiological signal
- **STRESS PROBABILITY**: The model prediction on how stressed the user is based on all the test results provided.
- **REFERENCE INTERVAL**: Range of normal physiological measure when patient is not stressed, provided by our model.
- **IMPACT**: The Impact provides indication of which physiological signals have more impact on the stress prediction by ranking the signals in descending order of impact. The idea is to calculate the feature's percentage contribution to the total probability of stress prediction. A feature with a negative impact decreases the stress probability and increases it otherwise.
- **FLAG**: Indicator of normal/abnormal signal in relation to stress. The Flag is 'red' if the test result is indicating stress and 'green' otherwise. The Star-shaped Flag is a representation of the correspondence to the non-stressful range. If the test result is inside of the normal/no stress range, this flag will be presented in Green and in red otherwise. The Circle-shaped Flag is a representation of the impact the test result has on the prediction: if the test impact is negative, this flag will be presented in green and red otherwise. The Red color is therefore an indication for stress.

2. Instructions Manual: How to Read and Interpret The Report

Figure 10 - Description of the Report's Attributes

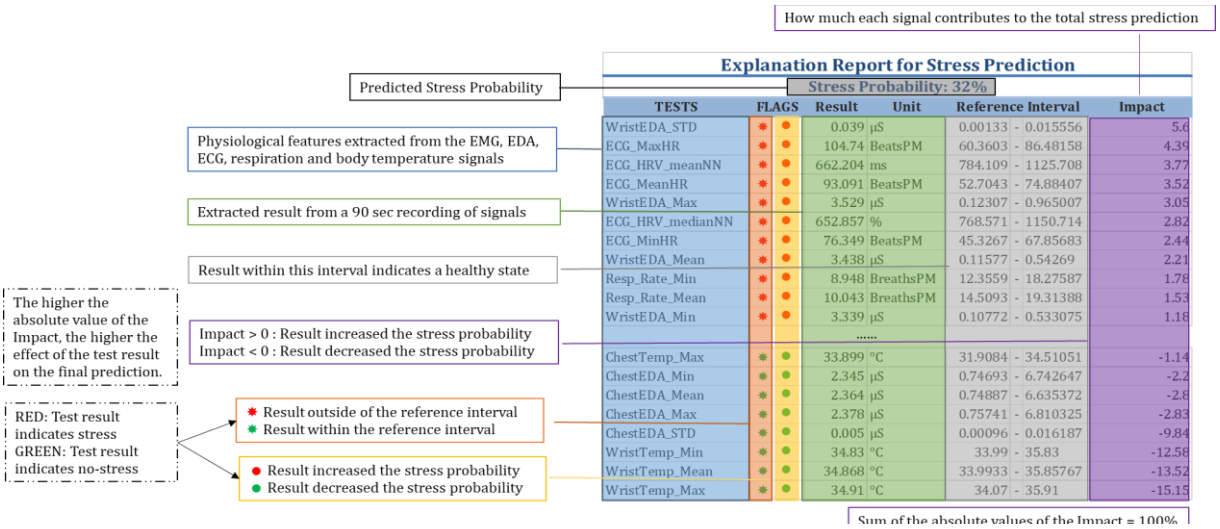
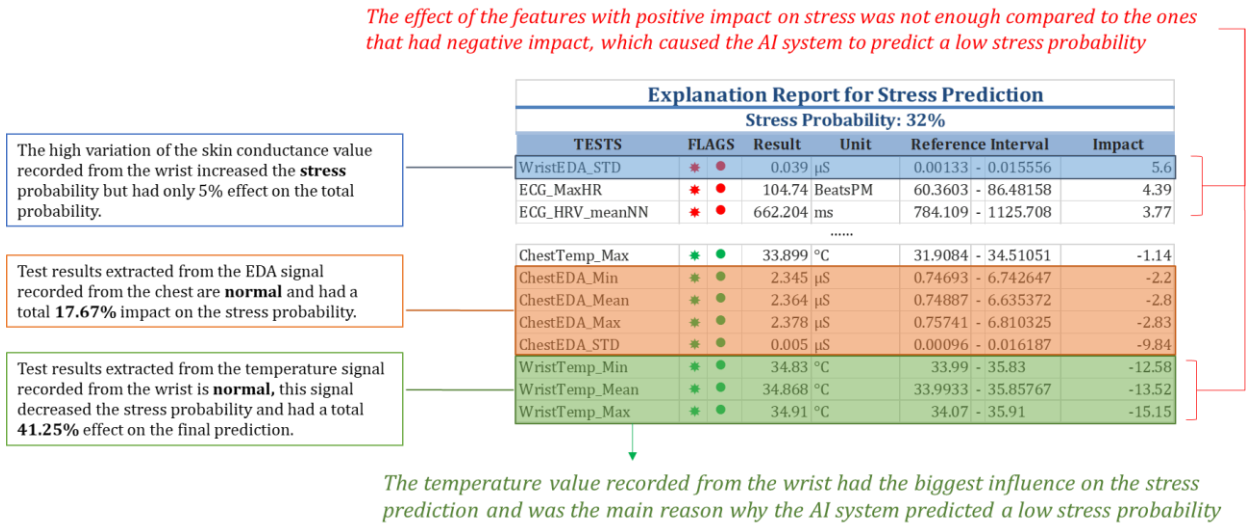


Figure 11 - Instructions on Report's Interpretation



3. Instructions Manual: Description of Physiological Attributes

Signal: Electrodermal Activity (EDA)			
EDA is neurological control of the rate of sweat gland production in the skin. It refers to the electrical changes measured at the surface of the skin. It captur electrical conductance across the skin, measured in microSiemens (µS)			
EDA SCL reflects the level of physiological arousal from the electrical skin conductivity. EDA SCR is an indication of autonomic nervous system arousal			
Recorded from the Chest	ChestEDA_Max	µS	Maximum value of electrodermal activity
	ChestEDA_Mean	µS	Mean value of electrodermal activity
	ChestEDA_Min	µS	Minimum value of electrodermal activity
	ChestEDA_SCLMean	µS	Mean value of the average level of skin conductance (SCL)
	ChestEDA_SCLSTD	µS	Measure of how far the values of the skin conductance level (SCR) differ from the mean
	ChestEDA_SCRMean	µS	Mean value of the average changes of the skin conductance level (SCL)
	ChestEDA_SCRSTD	µS	Measure of how far the values of the skin conductance response (SCR) differ from the mean,
	ChestEDA_STD	µS	Measures how far the EDA values vary from the mean EDA
Recorded from the Wrist	WristEDA_Max	µS	Maximum value of electrodermal activity
	WristEDA_Mean	µS	Mean value of electrodermal activity
	WristEDA_Min	µS	Minimum value of electrodermal activity
	WristEDA_STD	µS	Measures how far the EDA values vary from the mean EDA

Signal: Body Temperature			
Measure of the body temperature in Celsius (°C)			
Recorded from the Chest	ChestTemp_Max	°C	Highest temperature
	ChestTemp_Mean	°C	Average value of the temperatures
	ChestTemp_Min	°C	Lowest temperature
	ChestTemp_STD	°C	Measures how far the temperatures vary from the mean Temperature
Recorded from the Wrist	WristTemp_Max	°C	Highest temperature
	WristTemp_Mean	°C	Average value of the temperatures
	WristTemp_Min	°C	Lowest temperature
	WristTemp_STD	°C	Measure of the temperature's variation from the mean

Signal: Heart Rate Variability		
Heart rate variability (HRV) is the physiological phenomenon of variation in the time interval between heartbeats. It is measured by the variation in the beat-to-beat interval. Other terms used include: "cycle length variability", "RR variability" (where R is a point corresponding to the peak of the QRS complex of the ECG wave; and RR or NN is the interval between successive Rs)		
ECG_HRV_HF	ms^2	Variance in HRV in the High frequency (.15 to .40 Hz)
ECG_HRV_LF	ms^2	Variance in HRV in Low Frequency (.04 to .15 Hz)
ECG_HRV_madNN	ms	Median absolute deviation of NN intervals
ECG_HRV_mcvNN	ms	Median-based coefficient of variation of NN intervals
ECG_HRV_meanNN	ms	Mean value of the successive difference between the RR intervals which is the interval two heart beats.
ECG_HRV_medianNN	ms	Median value of the successive difference between the RR intervals which is the interval two heart beats.
ECG_HRV_pNN20	%	The proportion of number of pairs of successive NNs that differ by more than 20 ms divided by total number of NNs.
ECG_HRV_pNN50	%	The proportion of number of pairs of successive NNs that differ by more than 50 ms divided by total number of NNs.
ECG_HRV_RMSSD	ms	the square root of the mean of the squares of the successive differences between adjacent NNs
ECG_HRV_sdNN	ms	the standard deviation of NN intervals

Signal: Heart Rate		
the number of times the heart beats in the space of a minute, measure in beats per minute		
ECG_MaxHR	Beats per Minute	Highest heart rate recorded from
ECG_MeanHR	Beats per Minute	Average of the recorded heart rates
ECG_MinHR	Beats per Minute	Lowest heart rate recorded from
ECG_StdHR	Beats per Minute	Measure of the heart rate's variation from the mean

Signal: electromyography (EMG)		
The EMG signal is a biomedical signal that measures electrical currents generated in muscles during its contraction representing neuromuscular activities. The motor unit recruitment is reflected in the EMG signal amplitude This is a measure of the trapezius muscle.		
EMG_Min	mV	Minimum value of the EMG Signal
EMG_Max	mV	Maximum value of the EMG Signal
EMG_STD	mV	Standard deviation of the EMG Signal
EMG_RMS90P	mV	90^{th} percentile of the root mean square (RMS) of the EMG Signal
EMG_RMS	mV	The increase in RMS is related to the recruitment of additional motor units and an increased firing rate
EMG_RMS50P	mV	50^{th} percentile of the RMS of the EMG Signal

Questions & Results

How useful are the report parameters for the **physician** in understanding how the model is making its decision?

Report Stress Indicators	**Not at all useful**	**Slightly useful**	**Moderately Useful**	**Very Useful**	**Extremely Useful**	*Additional Comments*
Flags	☐	☐	☐	☒	☐	
Reference Interval	☐	☐	☐	☒	☐	
Impact	☐	☐	☒	☐	☐	
Stress factors (Tests)	☐	☒	☐	☐	☐	
Stress Probability	☐	☐	☐	☐	☒	

How useful are the report parameters for the **patient** in understanding how the model is making its decision?

Report Stress Indicators	**Not at all useful**	**Slightly useful**	**Moderately Useful**	**Very Useful**	**Extremely Useful**	*Additional Comments*
Flags	☐	☐	☒	☐	☐	
Reference Interval	☐	☐	☒	☐	☐	
Impact	☐	☒	☐	☐	☐	
Stress factors (Tests)	☐	☒	☐	☐	☐	
Stress Probability	☐	☐	☐	☐	☒	

How useful are each of the physiological signals for the **physician** in understanding how the model is making its decision?

Signals	**Not at all useful**	**Slightly useful**	**Moderately Useful**	**Very Useful**	**Extremely Useful**	Additional Comments
ECG Heart Rate	☐	☐	☐	☐	☒	
ECG HRV	☐	☐	☐	☒	☐	
Wrist Temperature	☐	☐	☐	☐	☒	
EMG	☐	☐	☒	☐	☐	
Respiration Rate	☐	☐	☒	☐	☐	

How useful is the usage of physiological attributes for the **physician** in understanding how the model is making its decision?

Not at all useful	**Slightly useful**	**Moderately Useful**	**Very Useful**	**Extremely Useful**	*Additional Comments*
☐	☐	☐	☒	☐	

Does the report provide the AI explanation needed for **psychiatrists**?

☒ Yes ☐No

Does the report provide the AI explanation needed for **patients**?

☒ Yes ☐No

If No, kindly specify what other features/signals can be included, or what other additional information should be added to the report.

1.
2.
3.

Part 2 – Assessment of Report listing specific tests

We integrated a report customization approach in the purpose of making the report clearer to patients and physician by allowing them to choose the tests they need and produce the corresponding explanation.

On a scale from 1 to 5 (1=not important, 5=very important), How much is this approach important in providing a more comprehensible and useful report to the patient?

☐ 1	☐ 2	☐ 3	☒ 4	☐ 5	Other comments:

On a scale from 1 to 5 (1=very clear, 5= not clear at all), taking into consideration the table describing each test, how clear could the test attributes be to the patient? (knowing that they can specify the tests used to generate the report)

☐ 1	☐ 2	☒ 3	☐ 4	☐ 5	Other comments:

Can this report be used to track patient's stress state over time?

☒ Yes ☐ No

Can this report be used to study the relation between physiological signals and stress?

☒ Yes ☐No

Part 3 – GUI Assessment

How clear is the report's display?

Hard to follow	**A little Confusing**	**Easy to follow**	*Additional Comments*
☐	☐	☒	

How clear is the report's organization?

Hard to follow	**A little Confusing**	**Easy to follow**	*Additional Comments*
☐	☒	☐	

How clear is the instruction manual that will be added to the report?

Hard to follow	**A little Confusing**	**Easy to follow**	*Additional Comments*
☐	☐	☒	