Heart Rate Variability Feature Selection using Random Forest for Mental Stress Quantification

Chang Su

A Thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master of Applied Science (Qualifying Program) at
Concordia University
Montréal, Québec, Canada

September 2020

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Chang Su

Entitled: Heart Rate Variability Feature Selection using Random Forest for Mental Stress Quantification

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. C. Wang

_____ External Examiner
Dr. H. Ge (BCEE)

_____ Internal Examiner
Dr. C. Wang

_____ Co-Supervisor
Dr. W.-P. Zhu

_____ Co-Supervisor
Dr. Y. Zeng (CIISE)

Approved by: _____
Dr. Y.R. Shayan, Chair
Department of Electrical and Computer Engineering

_____ 20___

_____
Dr. Mourad Debbabi, Interim Dean,
Gina Cody School of Engineering and
Computer Science

Abstract

Heart Rate Variability Feature Selection using Random Forest for Mental Stress Quantification

Chang Su

Mental stress is considered as an essential element that affects decision making. Apart from mental stress, cognitive workload, mental effort, attention, and cognitive engagement are also involved in the decision-making process. Ambiguities of these concepts lead to confusion in their applications.

One objective of this thesis is to explore the relationship between mental stress and stress-related concepts. By investigating the mechanisms for decision-making, the difference and correlation of mental stress and other concepts are disclosed.

Heart rate variability (HRV) is a common method to measure mental stress. By investigating the correlation between HRV and mental stress, it can be confirmed that HRV does respond to mental stress changes instead of other concepts. HRV features are used to assess whether there is a relationship between baseline HRV and mental stress. However, the extracted features usually contain a large amount of redundancy, which adds computational complexity to mental stress quantification while not contributing to quantification accuracy. Recently, researchers have resorted to the random forest as a tool for HRV feature selection.

Another objective of this thesis is to select significant HRV features to quantify the mental stresses using the random forest method.

In this thesis, an open-source data set, called the SWELL-KW data set, is used for mental stress measurement, where three labels are assigned according to different mental

stress conditions, i.e., neutral, time pressure, and interruption. A set of HRV features are proposed based on time domain and frequency domain analysis for mental stress measurement. Statistical analysis is performed to select the essential features that reflect mental stress.

The random forest algorithm of feature selection is then studied, and the accuracy in measuring mental stress is validated by comparing the extracted features of the training set and the testing set. In order to evaluate the random forest algorithm's performance, the comparisons with other related algorithms, including support vector machine (SVM), decision tree, gradient boosting decision tree (GBDT), k-nearest neighbor algorithm (KNN), and deep neural networks (DNN), are also conducted in terms of accuracy and time cost.

The optimal HRV feature subset is proposed for mental stress quantification, including median RR, mean RR, median REL RR, HR, pNN25, SDRR RMSSD, SDRR RMSSD REL RR, TP, SD2, and SDRR. It is shown that this subset of features gives a high feature importance score and thus has a significant effect on mental stress quantification.

Performing random forest analysis with a sufficient amount of labeled data shows that the optimal HRV feature subset yields high mental stress quantification accuracy by using random forest. Moreover, random forest always makes the best overall performance in feature selection compared with other algorithms in terms of accuracy and time cost. It also infers the potential relation between physiological responses and mental activities.

Acknowledgments

First of all, I would like to give my thanks to my supervisors, Dr. Yong Zeng and Dr. Wei-Ping Zhu, for their patience and insightful guidance in my research, also for their consistent support and encouragement through all stages of my Master study. Their guidance helped me in my research and thesis writing.

Secondly, I would like to thank my friends for their love and support when I was struggling. I will never forget the memory that they accompanied me when I was in trouble. I would like to thank Mr. Wentao Zhang for his help in modifying the HRV analysis program.

Finally, I want to say thanks to my parents. Without their courage and support, I would not have the opportunity to enjoy this adventure in a new culture and an exciting university. They have provided me an invaluable opportunity to gain experience in studying abroad. I love you 3000 times.

Contents

List of Figures **ix**

List of Tables **x**

List of Abbreviations **xi**

1 Introduction **1**

1.1 Background and motivation 1

1.2 Objectives 4

1.3 Contributions 5

1.4 Thesis organization 5

2 Mental stress in decision making: mechanism models and concepts **7**

2.1 The concept of mental stress 7

2.2 The elicitation of mental stress 9

2.3 Relations of mental stress with other relevant concepts 11

2.4 Mechanisms for decision making 12

3 Heart rate variability (HRV) **15**

3.1 The anatomy of the heart 15

3.2 The electrical activity of the heart 17

3.3 The relation between HRV and mental stress 22

3.4 The physiological and cognitive responses to stress-related phenomena . . . 24
3.5 Different measuring methods for mental stress . . . 27

4 Random forest and other related algorithms 30
4.1 The principle of random forest . . . 30
4.2 Random forest algorithm . . . 32
4.3 Other related algorithms . . . 33
4.3.1 SVM . . . 33
4.3.2 Decision tree . . . 34
4.3.3 GBDT . . . 37
4.3.4 KNN . . . 38
4.3.5 DNN . . . 39

5 Statistical analysis and results 41
5.1 SWELL-KW dataset . . . 41
5.1.1 Participants . . . 42
5.1.2 Design and tasks . . . 42
5.1.3 Procedure . . . 43
5.2 Preprocessing . . . 44
5.3 Tuning . . . 49
5.4 Feature selection . . . 51
5.4.1 Comparisons and model choices . . . 52
5.4.2 Feature importance . . . 55
5.4.3 Feature selection and comparison . . . 59
5.5 Summary . . . 62

6 Conclusion and future work 65
6.1 Conclusion . . . 65

6.2 Future work 66

Bibliography **68**

Appendix A Python code for HRV feature selection **76**

List of Figures

1 The relationship between performance and mental stress [58] 8
2 A mechanism for decision making in rational states 13
3 A mechanism for decision making in intuition states 14
4 The anatomy of heart [4] 16
5 The conduction system of heart [50] 19
6 A typical normal ECG waveform [59] 21
7 The physiological and cognitive responses to stress related phenomena . . . 26
8 The conceptual diagram of a random forest model [45] 31
9 The maximal margin hyperplane [6] 33
10 Sorting number using decision tree algorithm 35
11 An example of KNN [83] 38
12 A general model of DNN with N hidden layers [66] 39
13 The design process [43] 43
14 The sample size under three mental stress conditions 52
15 Test accuracy comparisons among different models 54
16 Feature importance in 99.98% accuracy 57
17 Feature importance in 80.65% accuracy 58
18 Top 10 features importance 59
19 Top 10 features accuracy in different algorithms 60

List of Tables

1 Components' inherent rates in the cardiac conduction system [37] 20

2 Time domain features [80] 23

3 Frequency domain features [80] 24

4 Functions of brain waves [57] 28

5 Feature extraction in time domain 46

6 Feature extraction in frequency domain 47

7 Accuracies of 31 features in different algorithms 53

8 Feature importance ranking 56

9 The test accuracy gap of 31 features and top 10 features in random forest . . 60

10 The test accuracy from 31 features and top 10 features in different models . 61

11 The training time of each method 62

List of Abbreviations

ANS Autonomic Nervous System

AV Atrioventricular

CART Classification and Regression Tree

CWT Color Word Test

DNN Deep Neural Networks

ECG Electrocardiogram

EEG Electroencephalogram

GBDT Gradient Boosting Decision Tree

GC Glucocorticoid

HF High Frequency

HPA Hypothalamic–Pituitary–Adrenal

HR Heart Rate

HRV Heart Rate Variability

IBI Inter-beat Interval

ID3 Iterative Dichotomiser 3

KNN K-Nearest Neighbors

LF Low Frequency

NASA-TLX NASA Task Load Index

PNS Parasympathetic Nervous System

PSD Power Spectral Density

RMSSD Root Mean Square of Successive Differences

SA Sinoatrial

SDNN Standard Deviation of Normal-to-Normal R-R intervals

SDRR Standard Deviation of R-R intervals

SNS Sympathetic Nervous System

SVM Support Vector Machine

SWELL-KW SWELL Knowledge Work

Chapter 1

Introduction

1.1 Background and motivation

It has been commonly and scientifically known that mental stress plays a significant role in human decision making. When people engage in activities such as driving, playing games, or giving a lecture, they must constantly balance the demand for an accurate decision against many parameters, e.g., time pressure. Based on qualitative research and experiments, the traditional literature generally concludes that mental stress mostly negatively affects the decision-making process [28]. The study by Giora K [38] offers the effects of stress on a critical phase of the decision-making process and makes individual consideration of alternative faulty. Another study [95] shows that the relationship between mental stress and performance is a bell-shaped curve. Nguyen T A and Zeng Y [44, 55, 89] proposed a theoretical framework to illustrate that the mental stress can be determined by the mental workload and the mental capacity, while the individual mental capacity can be defined by knowledge, skills, and affect. Increases in mental workload may trigger more mental stress and reduce individual performance. Consequently, the study of measuring and quantifying mental stress is essential if we are to reduce the harmful effect caused by mental stress and to achieve the best performance.

For developing mental stress quantification, it is essential to define mental stress. Stokes and Kite suggested that stress should be viewed as an agent, circumstance, situation, or variable that disturbs the 'normal' functioning of the individual, and stress is also seen as an effect—that is, the disturbing state itself [22]. While later, Contrada contended that stress is defined as a processing capacity of an organism, resulting in psychological and biological changes that may place persons at risk for disease [17]. Briefly speaking, the definition of stress includes internal or external stressors, perception of the organism's stimulation, and a physiological response [32, 52]. Since mental stress can generate a physiological response, several scholars attempt to use these reactions to substitute mental stress. However, some researchers contend that the measurement of physiological parameters cannot accurately explain the human stress response and does not necessarily represent mental stress [22]. An opinion is purposed in some biological stress responses that can only represent mental workload. TA Nguyen et al. [56] concluded that Heart Rate Variability (HRV) can quantify mental stress while the Electroencephalogram (EEG) energy can quantify mental effort.

However, there is widespread confusion about mental stress (arousal), workload, cognitive workload, mental effort, attention, and cognitive engagement. In the literature, these different concepts sometimes are used to describe the same phenomenon, while the same concept may be resorted to referring to different phenomena. For example, Roger Daglius Dias concluded that HRV analysis is a metric to assess cognitive workload [22]. However, other researchers asserted that mental stress influences HRV [75, 86]. Meanwhile, the mental effort can be measured both on NASA-TLX and EEG energy as per the authors of [13, 59].

Ambiguities of concepts often lead to confusion in their applications. Cognitive workload, attention, and cognitive engagement are also referred to in mental stress studies. To the author's best knowledge, very few works have addressed the differences between mental

stress, mental effort, cognitive workload, attention, and cognitive engagement. The clarification of these concepts will facilitate the effective and efficient applications of existing research to real-world problems.

The majority of current studies investigating mental stress quantification make use of different triggering methods and criteria, e.g., Color Word Test (CWT). Using this approach, TA Nguyen et al. indicated that the Low Frequency (LF) / High Frequency (HF) ratio in the HRV signal and EEG signal could quantify the mental stress [56]. Moreover, the previous studies classify mental stress measurement into electrophysiological measurement, subjective measurement, and biochemical measurement. Quantifying mental stress by its physiological feature is a field of research that received special and increasing attention. For example, HRV and EEG are reliable methods for quantifying mental stress. Since several measuring methods can be used to quantify mental stress, more and more studies tried to find out the best quantitative features. There's a wide spectrum of opinions on this issue. A challenging problem that arises in this domain is to select the appropriate features. In this thesis, we will investigate mental stress based on HRV.

The HRV data set used for quantification in this thesis is taken from the SWELL Knowledge Work (SWELL-KW) data set [43], which is provided by Koldijk S, et.al. They researched on stress and user modeling. Participants experienced typical work pressures in their experiment, such as receiving unexpected email interruptions and completing the work on time. During their experiment, the data set was collected by researchers, which is called the SWELL-KW data set [43]. SWELL-KW designed a mental stress experiment that 25 people participated in collecting and storing real-time R-R interval data, which were used to obtain HRV data. This thesis uses the HRV data set based on this experiment. Features used to quantify cardiovagal reactivity included time and frequency domain measures such as high frequency (HF) power, Standard Deviation of Normal-to-Normal R-R intervals (SDNN), and Root Mean Square of Successive Differences (RMSSD). By

analyzing HRV data using the random forest algorithm, the relationship between feature selection, physiological responses, and autonomic nervous system dynamics is verified. In addition to a longitudinal study method of analyzing data from all individuals at different stress conditions, feature selection results are validated from 31 features in 3 conditions (relaxed, stressed, and interrupted) based on their accuracy by using random forest. Five classification methods viz Support Vector Machine (SVM), Decision Tree, Gradient Boosting Decision Tree (GBDT), K-Nearest Neighbors (KNN), Deep Neural Networks (DNN) have been selected to compare with Random Forest in this study. The accuracy value is recorded for analysis.

This thesis is focused on the hypothesis that mental stress can be determined by mental workload and mental capacity, i.e., adjusting to a positive emotion can reduce stress. Furthermore, we hypothesized that the combination of mental effort, cognitive engagement, attention, and cognitive workload would induce mental stress, which would cause a positive or negative effect on performance. This performance can generate physiological response, i.e., decreased HRV and increased blood pressure.

1.2 Objectives

The objective of this thesis is two-fold. The first objective is to clarify mental stress, cognitive workload, mental effort, attention, and cognitive engagement by investigating the decision-making process's mechanisms. The second objective is to quantify mental stress by measuring and analyzing the HRV features using the random forest algorithm.

For the first objective, a mechanism for mental stress to be triggered in decision making is proposed to clarify stress-related concepts. It reveals the relationship between mental stress and HRV, which can quantify mental stress by analyzing variables.

Although HRV feature selection is useful in statistical analysis, the extracted features usually contain considerable redundant. In the second objective, an optimal HRV feature

subset is selected by using the random forest.

1.3 Contributions

This thesis focuses on finding the relations between mental stress and physiological measures, and the correlation between mental stress, stress-related concepts, and decision-making activities. The main contributions of this thesis include the following:

- Based on the literature review on stress-related concepts and phenomena, a mechanism of decision making is proposed to infer the causal relationships between different stress-related concepts.
- Critical HRV features to quantify and classify mental stress levels are identified by applying the random forest algorithm.
- The effectiveness of the random forest algorithm is validated by comparing it with other related algorithms and models.

1.4 Thesis organization

The remainder of this thesis is organized as follows: chapter 2 reviews relevant research in mental stress concepts, stress elicitation, and stress measurements. It also proposes a decision-making mechanism and figures out the connection between mental stress, cognitive workload, cognitive engagement, and mental effort. Chapter 3 describes the theoretical aspect of HRV and the superiority of HRV measurement. I also show the physiological and cognitive responses to stress-related phenomena. Chapter 4 presents the theoretical aspect of random forest and the related models and algorithms (SVM, decision tree, GBDT, KNN, DNN). Chapter 5 introduces the experimental setup and validation results for SWELL knowledge work under different stress conditions. It also focuses on the statistical analysis

of the HRV dataset, including preprocessing, feature extraction, and feature selection. The algorithm is also validated by carrying out comparative studies on other models. Chapter 6 summarizes the research results of this thesis and suggests some topics for future work.

All the experiment data in this thesis were provided by the SWELL-KW dataset, collected within the SWELL project. The collection of this dataset was supported by researchers at the Institute for Computing and Information Sciences at Radboud University.

Chapter 2

Mental stress in decision making: mechanism models and concepts

This chapter reviews relevant research in mental stress concepts, stress elicitation, and stress measurements. It also proposes mechanisms for mental stress to be triggered in decision making.

2.1 The concept of mental stress

Mental stress, as a common psychological phenomenon, is often encountered in our daily lives. Selyes [78] defines stress as 'the nonspecific result of any demand upon the body'. The definition of stress includes internal or external stressors, perception of the organism's stimulation, and a physiological response [32, 52]. It is generally asserted that psychological stress includes two traditional modes: stimulus-based and response-based [84]. The former assumes that certain environmental conditions, situations, or external events are expected to trigger stressful and considered as stressors (i.e., war, divorce, workload, heat and cold, et cetera.), ignoring the differences between individuals, circumstances appraisal, and emotional effects. The latter asserts that stress is a change response pattern of physical

function under stressors, and the variables are endogenous. Meanwhile, according to the intensity and duration of stressors, mental stress is generally divided into acute stress and chronic stress [21]. The effect and challenges arising from a mild stressor are temporary, often lasting from several minutes to hours.

In contrast, chronic stressor usually lasts for quite a few hours each day, sometimes up to several weeks or months, whose essential feature is persistent, repetitive, or high intensity. No matter what kind of stress can influence our brain, it affects physical, cognitive, affective, and behavioral aspects. Therefore, the current techniques and stress measurement methods and quantification are mostly derived from the response-based stress model.

The Yerkes-Dodson law [95] shows the bell-shaped curve relationship between mental arousal (stress) and performance, as given in Figure 1.

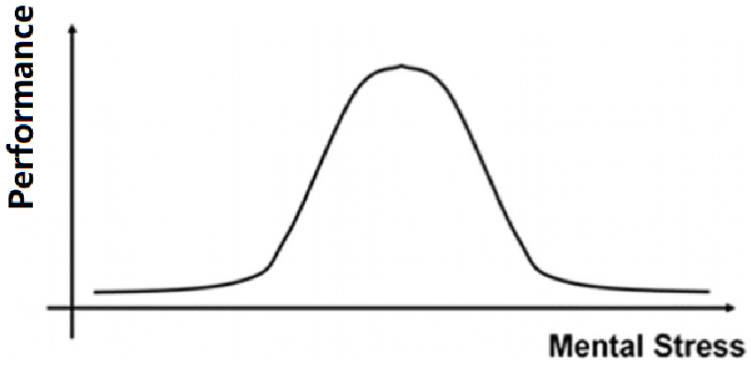

Figure 1: The relationship between performance and mental stress [58]

As shown in Figure 1, the performance may increase with mental stress, up to a point, but it will then show a negative correlation with mental stress if it is further increased.

According to the Yerkes-Dodson law, mental stress plays an important role in individual performance. However, Yerkes-Dodson law didn't specify the factors influencing mental stress. Many studies have been carried out on this topic. A paper relevant to this research was published by Nguyen T A and Zeng Y [58, 60, 61]. They gave a theoretical framework

for mental stress, which is expressed as:

$$\sigma = \frac{W_p}{(K+S) * a} \tag{1}$$

where W_p represents perceived mental workload, K is knowledge, S represents skill, a means affect, specifically emotion , and σ represents mental stress.

The stress performance model defines the factors influencing mental stress. It can be seen from Equation (1) that knowledge, skills, and emotion can define the individual mental capacity. The mental workload and mental capacity can determine mental stress. Therefore, the stress-performance model can be used to clarify and quantify these concepts.

2.2 The elicitation of mental stress

Before 1993, several researchers had elicited mental stress by using some laboratory tasks, such as the cold pressor test, the Stroop test, public speaking, etc. Much of the research in stressor distinguish in recent decades has divided the elicitation of mental stress into five terms based on stressors.

1 Working memory

Working memory refers to the brain system that provides temporary storage and manipulation of the necessary information for complex cognitive tasks [5]. Higher working memory individuals use simpler (and less efficacious) problem-solving strategies under high-pressure conditions and suffer from performance accuracy. A slice of researchers used working memory as a stressor for measuring mental stress. By way of illustration, the CWT is the classic working memory case that is widely used in the elicitation of mental stress [27, 67].

2 Reaction time

To a certain extent, reaction time indicates stress. When performing mental tasks at a satisfactory level of performance, it often encounters some complications, which can be due to many happenings and mistakes. Therefore, measuring reaction time is vital for monitoring and evaluating mental stress [7].

An example of this is the study carried out by Wolf Langewitz et al. in which a reaction time task is used to trigger individual mental stress [44]. By comparing the blood pressure at rest and under mental stress, they found that decreased parasympathetic nerve control leads to sympathetic and parasympathetic cardiovascular control disorders during hypertension.

3 Selective attention

Selective attention is directing our consciousness to relevant stimuli while ignoring irrelevant ones in the environment. This phenomenon is that people can focus on the process of particular aspects while ignoring irrelevant objects in the environment for a certain period.

The CWT is a typical case in this area. Vanitha L et al. claimed that HRV parameters are sensitive to working memory demands during the CWT test, thus sensitive to mental stress [89].

4 Physical pressure

Physical discomfort has also been used as a stress-inducing protocol. The typical case of physical pressure is the cold pressor test [90, 93]. The cold pressor test requires the subject to immerse the hand into an ice water container to trigger changes on blood pressure and heart rate of healthy participants [55]. Cold pressor stress indicated that acute stress undermines working memory performance, which is Secretory immunoglobulin A and cardiovascular reactions to mental arithmetic and cold pressor.

5 Social stress

At last, social stress also plays a principal role in measuring mental stress. It is generated based on relationships with others and a unique social environment. Public speaking is a representative case of social stress. Schubert et al. [77] reported that using speech task to induce stress, Standard Deviation of R-R intervals (SDRR) in HRV showed a discordant increase due to a slow respiration rate and a relative reduction in ventilation.

2.3 Relations of mental stress with other relevant concepts

Mental stress is associated and very often confused with concepts such as workload, cognitive workload, mental effort, attention, and cognitive engagement. Ambiguities of concepts will lead to confusion in their applications. Therefore, clarifying these concepts will facilitate the effective and efficient applications of existing research to real-world problems.

Several studies led to the definitions of mental stress, mental effort, cognitive engagement, cognitive workload, and attention. Beginning with mental stress, due to its wide range of applications, scholars in different research fields have given various definitions. In the aspect of mental effort, Heemstra stated that mental effort could be defined as the total use of cognitive resources [34]. Sun and Yao found that mental effort is positively related to design novelty and quantity [85]. Nguyen and Zeng verified that mental effort is the lowest at a high-stress level, and there is no significant difference in mental effort between medium-stress level and low-stress level [59]. Unlike mental effort, cognitive engagement is defined as the degree to which students are willing and able to immerse themselves in taking on the task at hand [34, 85]. The definition of cognitive workload is the measurable level of mental effort that an individual presents in response to one or more cognitive tasks, which is not the task but a property of the individual [68].

In short, mental stress is typically regarded as an essential influence factor leading to different cognitive degrees of later results (i.e., mental effort, cognitive engagement, and cognitive workload). Several scholars have attempted to use these results (e.g., cognitive

workload) to substitute mental stress. However, how mental stress affects cognition and behaviors of humans is unclear.

In order to better understand the variables that allow measuring the levels of mental stress, the definitions that are easily confused with psychological stress, such as mental effort, cognitive engagement, and cognitive workload, are worth distinguishing. They are interdependent and mutually motivate.

2.4 Mechanisms for decision making

Decision-making performance is related to the decision-makers' mental stress. Decision making is a process in which the stressor activates the individual cognitive system and creates the emotions, behaviors, and stress. Decision making will help to make more deliberate, thoughtful decisions by organizing relevant information and evaluating performances.

Stress-related concepts include mental stress, workload, cognitive workload, mental effort, attention, and cognitive engagement. Based on the performance-stress model, we try to draw this diagram to figure out the decision-making process.

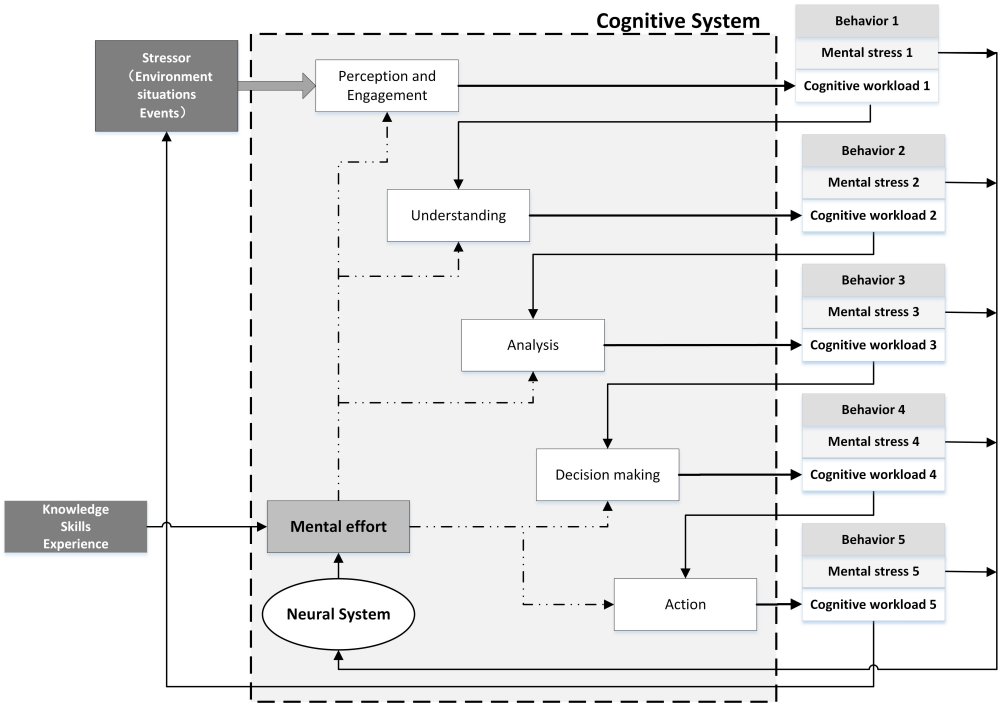

Figure 2: A mechanism for decision making in rational states

Mental stress generation in the mechanism for decision making will create these phenomena via the cognitive system. The mental stress can be determined in either a rational or the opposite situation. Figure 2 introduces the recursive cognitive process under mental stress in a rational situation. The white components indicate different cognition stages, which we consider as a cognitive system, including perception and engagement, understanding, analysis, decision making, and action. In the beginning, stressor causes perception and engagement. From this stage, it creates cognitive workload, as well as mental stress and behavior. This cognitive workload acts in the next stage. We can see that the cognitive workload updates and generates a new one in the next stage. In order to distinguish them, we marked them as 1 and 2. As the process progresses, we repeatedly update the cognitive workload, the same as mental stress and behavior. It can be noticed that mental stress and behavior can also affect the individual neural system, which will participate in the cognitive process by generating a mental effort and can be used when using knowledge

and skills to figure out the mental workload. The whole process is recursive.

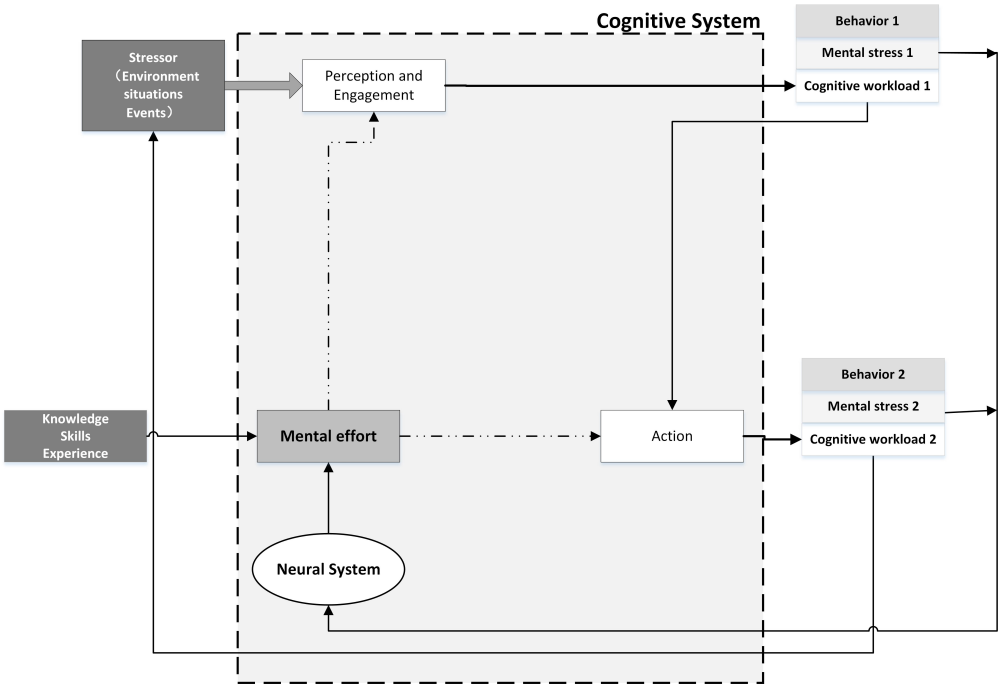

Figure 3: A mechanism for decision making in intuition states

The above cognitive processes are all in a rational situation. We can also encounter such a situation where people depend on conditional reflection or intuition to act. This cognitive process under mental stress is shown in Figure 3. After perception and engagement, people create actions directly. The difference between Figure 2 and Figure 3 is since individual knowledge and skills are different. In section 2.1, the stress- performance model defines the factors influencing mental stress. Moreover, mental effort creates mental stress by acting on the cognitive system.

The stress-performance model factors are related to the concepts, including mental stress, workload, cognitive workload, mental effort, attention, and cognitive engagement. Therefore, decision-makers such as skills, knowledge, and affect come from mental capability. The mechanism for decision making can infer the casual connections between stress-related concepts.

Chapter 3

Heart rate variability (HRV)

HRV is the physiological phenomenon in which the time interval between consecutive heartbeats changes. This chapter describes the theoretical aspect of HRV and the superiority of HRV-based mental stress measurement by comparing it with other measurement methods.

3.1 The anatomy of the heart

The heart is a muscular pump with its rhythmic contractions and allows a constant flow of blood through all tissues ensuring a regular exchange of gasses, nutrients, and waste products. The heart is wrapped with a thin membrane called the pericardium. It is located in the central part of the chest above the diaphragm (muscle barrier which divides the abdomen from the chest). The heart's size is about a closed fist; the weight varies from 300-350 grams for men and 250-300 grams for women. The heart consists of two atria and two ventricles. The atria (the right atrium and the left atrium) receive blood. Afterward, it transmits the blood into the two lower chambers called ventricles (the right ventricle and the left ventricle), as shown in Figure 4.

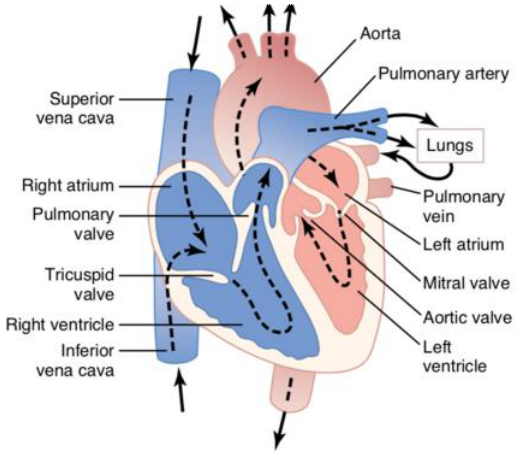

Figure 4: The anatomy of heart [4]

The atria and ventricles on two sides of the heart are separated by the wall called the septum, which prevents the mixing of the blood of the heart's left and right side. The wall dividing the right and left atria is called the inter-atrial septum, while the part dividing right and left ventricular is called the inter-ventricular septum. The right atrium delivers the deoxygenated blood to the right ventricle, pushing the deoxygenated blood to the lungs. After releasing the carbon-dioxide and taking on oxygen, the oxygenated blood comes to the left atrium. The left ventricle takes the oxygenated blood from the left atrium and pushes it to the rest of the body.

The heart's pulsation is a product of rhythmic contractions and relaxations of the heart muscle, which is called the myocardium. During the contraction phase, the wall of the atrium or ventricle contracts, increasing the pressure within the heart and ejecting blood out of the closed chamber. Subsequently, the atrial or ventricular wall relaxes and is ready to receive a new amount of blood.

The Autonomic Nervous System (ANS) controls the heart contractions. ANS is divided into the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS). The SNS and PNS work antagonistically. The SNS prepares the human body to respond to stressful situations. That response is known as the fight or flight response [12].

Simultaneously, the PNS controls the human body's free functions in a normal basal condition, popularly called rest and digest system [11, 12]. The ANS's SNS part is activated in response to a stressful situation, while challenging physical activity, or when we feel angry or are frightened. The following [49] are the most common facts related to the SNS:

1) HR can increase from 70 to 150 bpm in 3 seconds.

2) The blood pressure can double in 10 seconds.

The heart can contract without outside innervation. However, the power of the heart contraction is controlled by the ANS. Under the effect of the SNS part, the HR and the power of the heart contraction are increased. While under the control of the PNS part of the ANS, HR and cardiac contractions are decreased.

3.2 The electrical activity of the heart

Since the first human Electrocardiogram (ECG) recording was published in 1887 by Augustus Waller[91], the ECG signal has been used widely in many fields. Researchers detect and quantify human activities and responses by monitoring the electrical activity of heart rate.

It is expected that all heart activities have electrical impulses. The electrical impulse causes the heart muscle contraction. The formation and transmission of electrical impulses depend on the characteristics of the heart's cells.

Bio-electricity represents the ability of biological tissues to generate electricity without external excitation. The first research regarding bio-electricity was published by Luigi Galvani [30]. He discovered that the muscles of dead frogs' legs twitched when struck by an electrical spark.

The electrical charges in the tissue originate from the ions. Therefore, in cells, there are two kinds of electrical potentials: static potential and action potential.

The cell membrane changes from the static potential to the action potential. Stimulation can change the cell electrical potential, open the sodium ion (Na^+) channel, and allow many Na^+ to enter the cell. This process is called polarization. When the excitation is higher than the threshold, it opens the ionic channel, and the positive Na^+ come into the heart cell, causing the change of the electrical potential. This process is called depolarization. The repolarization is the descending process towards the static potential. It represents the change of the difference of the electrical potential inside the cell. Potassium ions (K^+) begin to fall along the electrochemical gradient. With the removal of potassium from the cell, its potential decreases and approaches its resting potential again. The sodium-potassium pump has been working continuously during this process. At the peak action potential, K+ channels open, and the cell becomes hyperpolarized. The K^+ are maintained at high concentrations within the cell. At the same time, Na^+ are maintained at high concentrations outside of the cell in neurons.

In general, the repolarization and depolarization represent the foundation of the heart's electrical activity, which allows the heart to work. The activity of action potential in the heart can be recorded to generate an ECG.

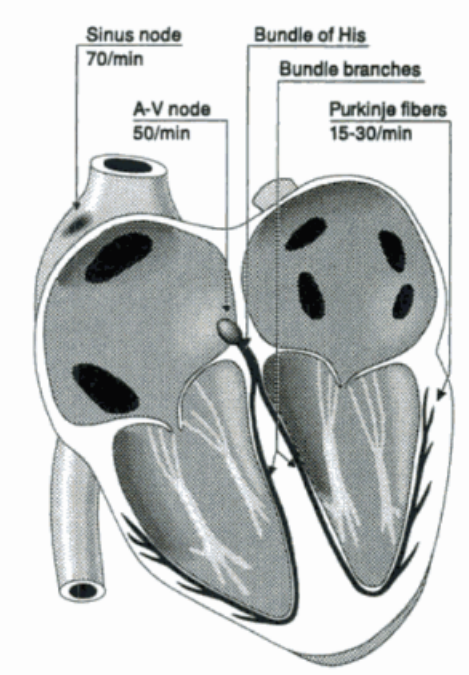

Figure 5: The conduction system of heart [50]

As shown in Figure 5 above, the cardiac conduction system consists of the following five components:

- The Sinoatrial (SA) node: This cell is found within the right atrium of the heart.
- Atrioventricular node: This cell can be found within the border of the right atrium and the right ventricle.
- Atrioventricular (AV) bundle: This cell is found within the right atrium of the heart.
- Right and left bundle branches: Both of which are located along the interventricular septum, the left bundle branch is further divided into the left anterior fascicles and the left posterior fascicles.
- Purkinje fibers: These fibers can be found in the inner ventricular walls of the heart. They receive conductive signals originating at the AV node and simultaneously activate the left and right ventricles by directly stimulating the ventricular myocardium.

All components show different inherent rates in the cardiac conduction system, as illustrated in Table 1 below.

Table 1: Components' inherent rates in the cardiac conduction system [37]

Component	Inherent rate (BPM)
SA node	60-100
AV node	40-60
Bundle of His	40-60
Right and left bundle branches	20-40
Purkinje fibers	20-40

It is essential to know that the ECG records the heart's electrical activity, in which each heartbeat is displayed as repeatedly multiple waveforms characterized by peaks and valleys.

Generally speaking, the frequency range of the ECG signal is from $0.05Hz$ to $100Hz$, and the dynamic range is from $1mV$ to $10mV$. The ECG signal is characterized by five peaks and valleys, Einthoven [26] identified the five deflections, which can be marked with the letters P, Q, R, S, T, respectively [69]. ECG also includes a U wave; however, the typical normal ECG may not show it. The normal ECG waveform is shown in Figure 6 below. This figure shows the electrical activity of the heart rate during a heart rate cycle. An ECG signal is a composite recording of all the action potential produced by myocardial nodes and cells. Each wave of the ECG corresponds to the cardiac electrical cycle event, as shown in Figure 6.

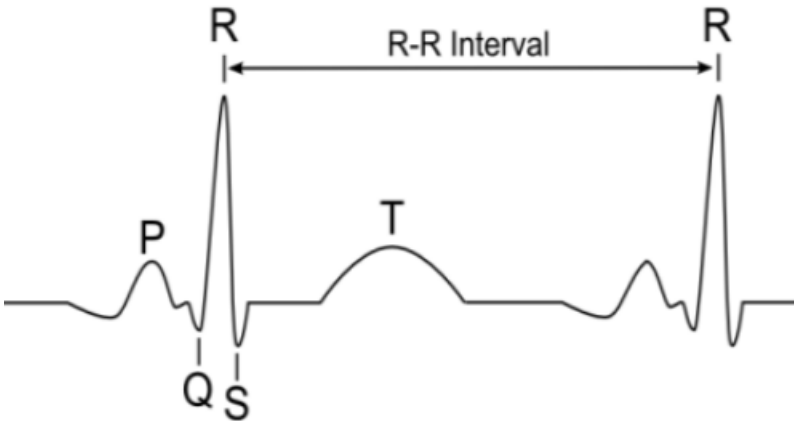

Figure 6: A typical normal ECG waveform [59]

In the ECG system, the P wave, T wave, and the QRS complex should be concerned. The P wave represents the activation of the upper chamber of the heart and the atrium, while the QRS complex wave and T wave represent the excitation of the ventricle or the lower chamber of the heart. The T wave reflects the repolarization of the ventricles. The QRS complex represents the ventricular contraction. The detection of the QRS complex is one of the most critical tasks in ECG signal analysis. Once the QRS complex is identified, more detailed information such as Heart Rate (HR), and HRV can be obtained [54, 76].

The RR interval, which is often used to monitor mental health, is the time between QRS complexes. The instantaneous heart rate can be calculated from the time between every two QRS complexes. The RR interval shows the connection between the power of HRV and the nervous system. It is different from the heart rate, which averages the number of beats per minute.

3.3 The relation between HRV and mental stress

HRV signal is a non-stationary signal, which describes the variations between consecutive heartbeats. Its changes can be interpreted as a current or upcoming disease and psychological activity.

The ANS can generate significant components of the stress responses in the physiological model. The ANS will create physiological responses such as HR, HRV, blood pressure, eye tracking, and skin conductance.

HRV is an objective measurement method that can be used to measure psychological stress. As shown in Figure 6, HRV is a fluctuation in the heartbeat interval controlled by the original part of the ANS. It can regulate our heart rate, blood pressure, breathing, and digestion.

HRV related researches commonly use features to measure. These features are extracted from the time domain and the frequency domain. HRV analysis for mental stress measurement is usually classified into two domains: time domain and frequency domain. Time-domain measurement can measure RR intervals directly or measure from the differences between RR intervals [25]. Researches about HRV experiments commonly use these metrics to measure: mean of the interval between successive RRs (RR), SDRR, the mean and standard deviation of HR.

In comparison, the frequency domain uses Power Spectral Density (PSD) to estimate the HRV signal. In the frequency domain, features can discriminate between the sympathetic and parasympathetic contents of the HRV signal. Commonly, the HF, LF, and VLF bands and the ratio of LF and HF bands power spectral density (LF/HF) are used as the frequency domain features of the RR interval signal [15]. In this thesis, several standard features of HRV in both time domain and frequency domain are shown in Table 2 and Table 3, respectively.

Table 2: Time domain features [80]

Parameter	Description
SDNN	Standard deviation of normal RR intervals
SDRR	Standard deviation of RR intervals
pNN50	Percentage of successive RR intervals that differ by more than 50 ms
RMSSD	Root mean square of successive RR interval differences

As shown in Table 2 above, SDNN and SDRR can measure RR intervals directly, while pNN50 and RMSSD measure the differences between RR intervals [25, 62]. These features can be calculated as

$$SDNN = \sqrt{\frac{1}{N-1}\sum_{i}(RR_i - RR_m)^2} \tag{2}$$

$$pNN50 = \frac{\sum_{i=1}^{N}(|R_i - R_{i+1}| > 50ms)}{N-1} \tag{3}$$

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2} \tag{4}$$

RR_i represents the i-th RR interval, where N means the total number of heartbeats, and RR_m represents the mean of the RR intervals. Like SDNN, SDRR can also measure how these intervals change over time, but it includes fault or abnormal beats [80].

Table 3: Frequency domain features [80]

Parameter	Description
ULF power	Absolute power of the ultra-low-frequency band (0.003 Hz)
VLF power	Absolute power of the very-low-frequency band (0.0033–0.04 Hz)
LF peak	Peak frequency of the low-frequency band (0.04–0.15 Hz)
LF power	Absolute power of the low-frequency band (0.04–0.15 Hz)
HF peak	Peak frequency of the high-frequency band (0.15–0.4 Hz)
HF power	Absolute power of the high-frequency band (0.15–0.4 Hz)
LF/HF	Ratio of LF-to-HF power

It is well known that the spectral power in HF of the RR interval reflects the activity of the cardiac vagus nerve. On the other hand, the LF frequency band is related to both vagal and sympathetic systems [65]. Some researchers found that heightened mental stress was associated with lowered HRV, specifically with reduced parasympathetic activation. Reduced parasympathetic activation was seen as a decrease in RMSSD and HF power and an increase in the LF/HF ratio. Some previous studies also indicated that the activity diaries, in conjunction with HRV data, could analyze and isolate important individual events: sleep, exam, physical activity, and caffeine [15, 72, 87].

3.4 The physiological and cognitive responses to stress-related phenomena

This section shows that the physiological phenomena are mapped into the decision-making process perspicuously. It is widely known that the mental stress reflected on many factors and various systems of the body. Researches over the past years have clarified that the entire brain is involved in responding to stressors. Researches over the past years have

clarified that the entire brain is involved in responding to stressors. With brain imaging technology development in mammals and the remarkable progress in genetic studies, a new understanding of stress networks has been gained in recent years. Stress networks are a set of highly connected brain structures activated when the animals perceive from their surroundings or are exposed to various stressful life events [30].

The stress-related performances contain mental stress, workload, cognitive workload, mental effort, attention, and cognitive engagement during decision-making activities. It is all known that the phycological and cognitive responses will change and reflect the stress-related phenomena. As mentioned in section 2.2.4, several physiological and cognitive responses can be used as measurement metrics for mental stress.

The major components of the stress responses in the physiological model can be generated by the ANS, Hypothalamic–Pituitary–Adrenal (HPA) axis, and brain network. Meanwhile, different stress types, including acute stress and chronic stress, have different effects on cognition, decision-making, memory, and health [10]. The body and nervous system's organization and interactions reflect a high degree of complexity and multidirectional communication.

Mental stress and related phenomena can be monitored and measured from the physiological and cognitive responses. One clear neurobiological indicator of the stress reaction is the significant activation of two stress response systems, rapidly acting SNS and the slower HPA axis, which results in a cascade of neuroendocrine changes [19]. Brain network is also activated by mental stress, as shown in Figure 7 below.

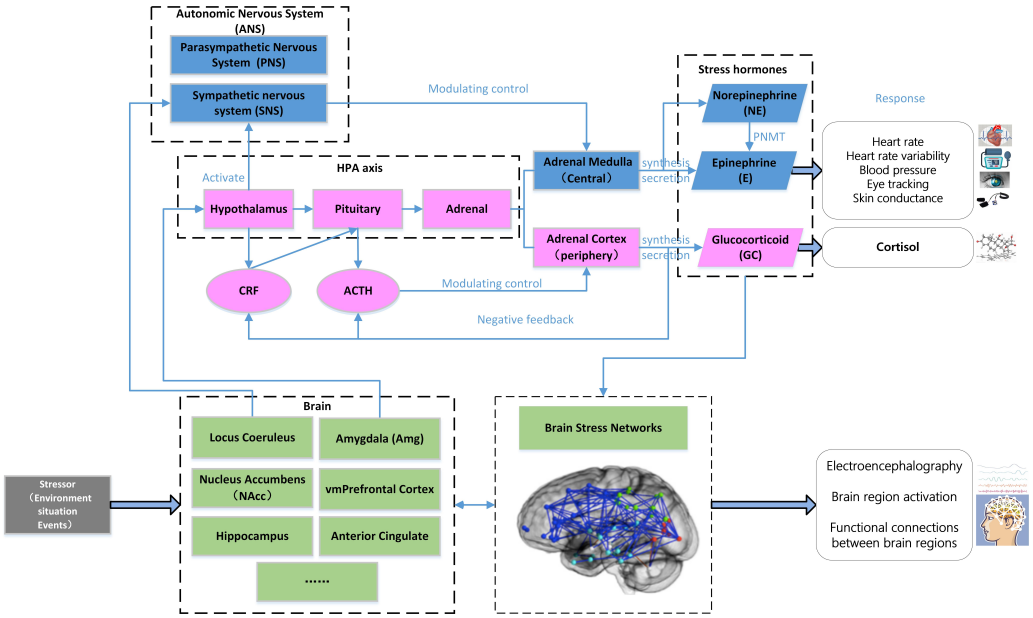

Figure 7: The physiological and cognitive responses to stress related phenomena

Figure 7 indicates the physiological and cognitive responses to stress-related phenomena. Based on the response systems and organs, these stress-related responses can be created from three approaches: the ANS, the HPA axis, and the brain network. As we mentioned in section 3.3, ANS includes SNS and PNS. Stressor stimulates SNS and modulates control on the Adrenal medulla. It creates the synthesis and secretion of norepinephrine and epinephrine. The blue components show that the ANS will create physiological responses such as heart rate, heart rate variability, blood pressure, eye tracking, and skin conductance. HRV analysis has been established as a quantitative measure of ANS activity related to mental stress [2].

Stress can also cause an increased cortisol output via the HPA axis activation, as shown in the pink component. When stressor reflects on the HPA axis, the hypothalamus creates CRF, which will stimulate the pituitary. Then the pituitary secretes ACTH, which will stimulate on Adrenal. The adrenal cortex is stimulated and secretes Glucocorticoid (GC), which gives negative feedback on previous parts. Cortisol is the most critical human GC,

which is known as the stress hormone. It increases blood sugar levels, enhances the brain's use of glucose under stress conditions. Besides, the green components show that the brain network will create EEG and brain region activation under mental stress.

It is well known that mental stress can be quantified from human bio-signals. Figure 7 links physiological and cognitive responses to stress-related phenomena. This figure validates the reliability of mental stress quantification based on physiological responses.

3.5 Different measuring methods for mental stress

In addition to HRV, many other modalities can be used for mental stress measurements such as EEG, cortisol, and NASA Task Load Index (NASA-TLX). Several measurement methods can be listed from subjective, biochemical, and psycho-physiological parameters separately.

These typical quantitative evaluations and acute mental stress techniques are introduced below, including the development, calculation methods, characteristics, and applicable scopes.

- EEG

EEG signals exhibit various characteristics in different brain waves. Some qualitative studies in the literature described how mental stress could be quantified from EEG signals.

The Electroencephalogram (EEG) signal is a non-stationary signal with different frequency elements at different time intervals. Recent research found that EEG reflects brain activity, and it is widely used in many fields, especially in mental stress [20, 39, 88]. All EEG channels are offline-referenced to the average of electrodes [3].

Based on frequency ranges, EEG signals can be classified into four bands: delta (1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), and beta (13-30 Hz). Each band represents a different function, as described in Table 4.

Table 4: Functions of brain waves [57]

Brain waves	Characteristics
Delta (1-4Hz)	Dominant when sleeping
Theta (4-8Hz)	When temporal and occipital lobes are relaxed, awaking state
Alpha (8-3Hz)	Mainly occipital and parietal lobes are relaxed, awaking state with eye closed
Beta (13-30Hz)	Dominant in frontal region during mental activity

By using the valence models of hemispheric specialization of emotion, Davidson et al. [20] stated that the left hemisphere is more involved in handling positive emotions and approaching-related behaviors. In contrast, the right hemisphere is more involved in handling negative emotions and withdrawal behaviors [88]. Recent research showed that EEG reflects brain activity and is widely used in many fields. R Khosrowabadi et al. [39] proposed a brain-computer interface for classifying EEG correlates of chronic mental stress.

- Cortisol

As a biochemical measurement, cortisol is one of the most common and popular biomarkers for quantifying stress in both animals and humans over the past several decades. It is widely believed that activation of the HPA axis during mental stress induces secretion of hormones, such as corticotrophin-releasing hormone and adrenal steroid hormones [35, 40].

However, there are many challenges and difficulties in measuring and quantified evaluating the level of stress using cortisol. First of all, not all types of acute negative stressors consistently activate HPA to trigger the cortisol changes [16, 51]. Second, even acute mental stressors trigger the adrenal cortex to release cortisol into the bloodstream by activating specific cognitive processes and their central nervous system. The cortisol levels can be influenced by numerous factors, such as gender, age, and caffeine [23, 41, 48].

- NASA-TLX

Numerous literature studies have confirmed that subjective measurement still plays an essential role in stress data collection. NASA-TLX, as a kind of subjective measurement, is widely used in stress measurement.

As a popular multidimensional metric, NASA-TLX is designed to obtain workload estimates immediately or after a task. Previous research on the subscale selection and weighted averaging methods has produced a tool that has proven to be reasonably easy to use and has reliable sensitivity to experimentally significant operations in recent decades [33].

Based on the principle of measuring self-reported stress, NASA-TLX calculates stress from six different dimensions: mental demands, physical demands, temporal demands, own performance, effort, and frustration. Through assessing the weight value of two factors out of six and evaluating the factor values, NASA-TLX can finally quantify the mental stress by calculating the total workload. NASA-TLX is more sensitive to low mental workloads [64].

Chapter 4

Random forest and other related algorithms

This chapter introduces the theoretical aspect of random forest and other related algorithms, including support vector machine (SVM), decision tree, gradient boosting decision tree (GBDT), k-nearest neighbor algorithm (KNN), and deep neural networks (DNN).

4.1 The principle of random forest

The random forest learning method [10] is presented by Breiman in his article Random forests. The random forest, as a classification algorithm, is a tree-based classifier. Its theoretical background rests on the concept of bagging and decision trees. This includes developing multiple trees from the random sampling subspace of the input features, using a randomly selected subset of training samples. Then it combines the results by voting or the maximum posterior rule output. The random forest is an ensemble learning algorithm that constructs a set of individual classifiers, also referred to as base learners.

Random forest is composed of many independent decision trees. During the classification task, each decision tree in the forest will be judged and classified separately when the

new input sample is entered. Each decision tree will get its classification result. Therefore, the random forest will choose the result which has the most voted classification as the final result.

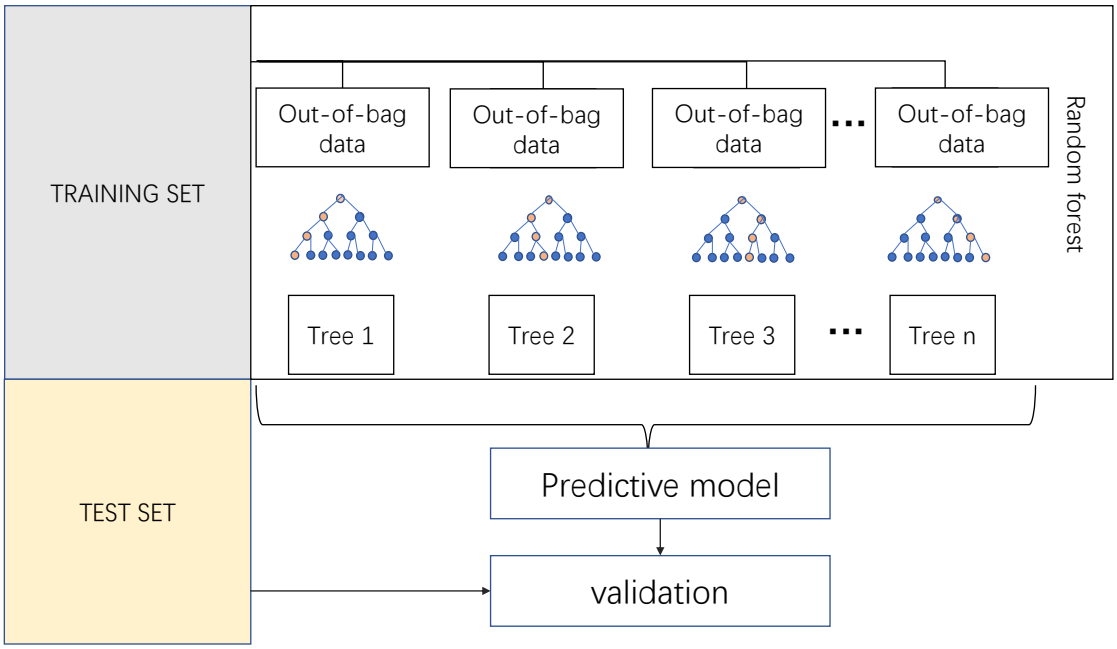

Figure 8: The conceptual diagram of a random forest model [45]

Figure 8 indicates the conceptual diagram of a random forest model. Samples taken from the training set can generate different decision trees. Then, all decision trees are used to form a single prediction. The prediction can be validated using the testing set.

Random forest classifies observations according to most of these learners' classification, which is often referred to as voting because observations are categorized based on decisions or votes made by most basic learners during classification [47].

4.2 Random forest algorithm

The construction of a random forest follows four steps:

1. If there are N samples for training, select N times from the N samples randomly with replacement. The selected N samples are used to train a decision tree as the samples at the root node of the decision tree.

2. When each sample has M attributes, randomly select m ($m << M$) attributes from these M attributes when each node of the decision tree needs to be split. Then select an attribute as the split attribute of the node from m attributes using some strategies.

3. During the decision tree generation, each node must be split according to step 2 until it can no longer be split.

4. Random forest occurs by following steps 1 to 3.

The decision tree is a classic weak model. When it tries to label data, no matter the distribution of the training data, it will always do better than accidentally [94]. In comparison, a random forest makes a massive development.

The random forest can judge the feature importance, determine the interaction between different features. Random forest is flexible and can increase the weak model (the decision tree) in terms of accuracy to a better extent. However, it may cost more massive computational resources.

In this section, the construction and the pros and cons of the random forest have been described. It is one of the supervised learning methods that are being applied and compared in this thesis. The following section describes the fundamentals of the other supervised learning methods, the SVM, decision tree, GBDT, KNN, and DNN.

4.3 Other related algorithms

4.3.1 SVM

SVM is a supervised machine learning algorithm that can be used for classification or regression. However, it is mainly used for classification problems. In the SVM algorithm, each data item is drawn as a point in an n-dimensional space, where n is the number of features we have, and the value of each feature is the value of a specific coordinate [96]. Then, we classify by finding a hyperplane that can distinguish the two categories. SVM defines the linear classifier with the most considerable interval in the feature space. The learning strategy of SVM is to maximize the interval, which is shown in Figure 9.

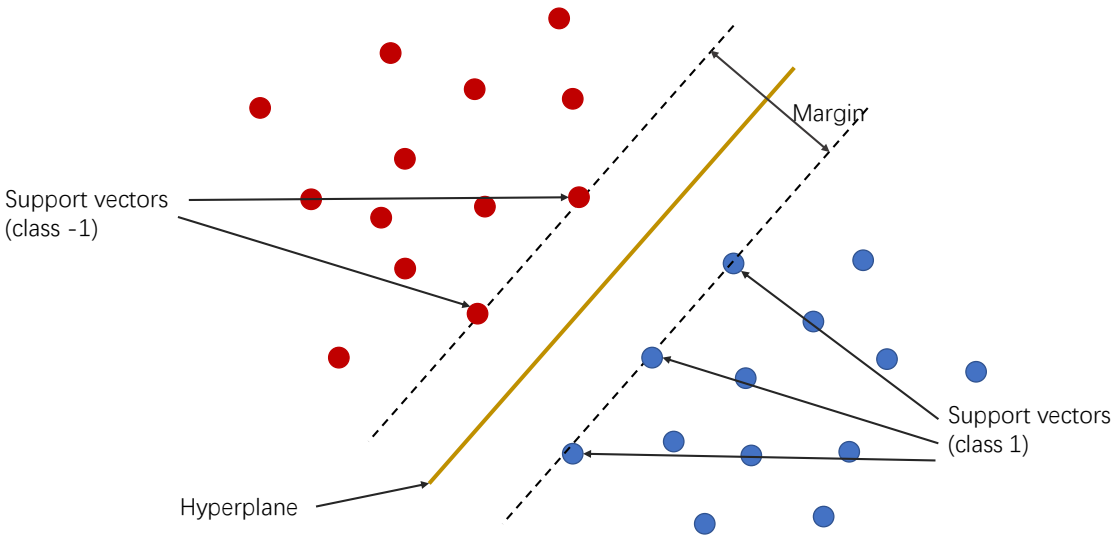

Figure 9: The maximal margin hyperplane [6]

In the binary classification case, the training observation can be divided into two different classes, usually expressed as -1 and 1. The margin represents the area within the two

hyperplanes. The support vector represents the support vectors closer to the hyperplane and influences the position and orientation of the hyperplane.

Using the maximal margin classifier is generally a successful way to classify when it is possible to find a separating hyperplane, though, there might be problems with overfitting the data in some cases [31]. It is worth noting that, commonly, there does not exist a hyperplane that can separate the two classes strictly.

4.3.2 Decision tree

A decision tree is a decision support tool that uses a tree-like model of decisions and possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements [74].

A decision tree contains nodes and directed edges, where the nodes can be classified into the root node, the internal node, and the leaf node [81]. Without a parent node, The root node represents the beginning node. The internal node represents a feature, while the leaf node represents a class. For example, the node of '$A > B$' in Figure 10 is the root node, the node of '$B > C$' in Figure 10 is the internal node, and the node of '$A > B > C$' in Figure 10 is the leaf node.

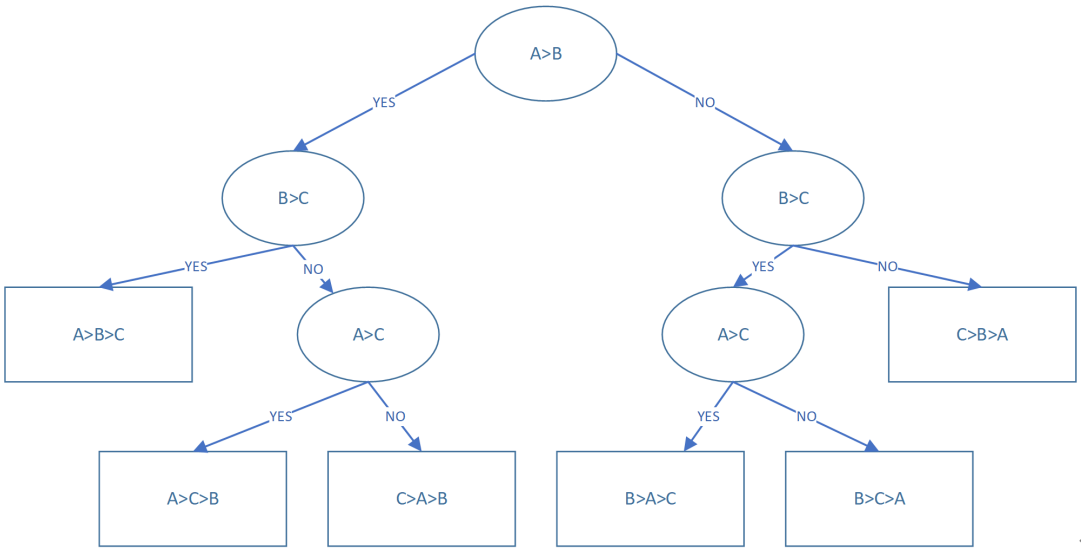

Figure 10: Sorting number using decision tree algorithm

A decision tree is a flowchart-like structure. Each internal node represents a judgment on an attribute in a decision tree, each branch represents an output of the judgment, and each leaf node represents a class label. The paths from the root to the leaf represent classification rules.

Figure 10 shows a straightforward application of the decision tree algorithm, which supposes that we want to sort three values, A, B, and C ($A \neq B \neq C$). To sort these values, firstly, this problem should be divided into smaller sub-problems. Then, try to figure out each sub-problem and repeat the classification step until getting the final result.

The decision tree algorithm is considered one of the best-supervised learning classification methods. The generation of the decision tree can be mainly divided into the following two steps:

1. when the attribute of a node cannot be judged, divide this node into N ($N \in Z, N \geq 2$) child nodes.

2. choose an appropriate threshold to minimize the training error.

The typical decision trees include Iterative Dichotomiser 3 (ID3), C4.5, and Classification and Regression Tree (CART).

ID3 uses the information gain to decide which feature goes into a decision node [70]. The information gain is expressed as:

$$g(D, A) = H(D) - H(D|A) \tag{5}$$

where $H(D)$ represents the entropy of set D, the $H(D|A)$ represents the conditional entropy of set D and feature A. The $g(D, A)$ represents the mutual information of set D and feature A. For a set of data, the smaller the entropy, the larger the information gain, the higher the impurity, the better the classification result will be. However, ID3 incurs some problems. As a smaller segmentation causes a smaller classification, ID3 may overfit the training data [24]. Moreover, the calculation of information gain depends on the size of the features.

In order to avoid this segmentation problem, C4.5 makes improvement based on ID3. C4.5 uses gain ratio to overcome the bias [70]. The gain ratio is express as :

$$GR(D, A) = \frac{g(D, A)}{H(A)} \tag{6}$$

where $g(D, A)$ represents the mutual information of set D and feature A, the $H(A)$ represents the entropy of feature A.

By dynamically defining discrete attributes, C4.5 reduces the restriction that features must be categorical [71]. CART is similar to C4.5, but it supports numerical target variables [9]. CART is a binary tree, which only classifies the parent node into two child nodes. The Gini impurity is the lost function being used in the CART method [70].

In the analysis, decision trees and closely related influence diagrams are used as visual and analytical decision support tools, in which the expected value (or expected utility) of

competitive alternatives can be calculated.

4.3.3 GBDT

Another decision tree learning is GBDT, which has been very successfully applied to many fields such as smart city [79], and its significant advantage is the ability to find nonlinear interactions automatically through decision tree learning with the minimum error.

The GBDT using an additive model classifies or regresses the data by reducing the residuals, which are generated during the training process. Each iteration creates a weak classifier through multiple iterations, and each classifier is trained based on the residuals of the previous classifiers. In conclusion, the GBDT algorithm has four steps:

1. Each iteration generates a new decision tree.
2. Before starting each iteration, GBDT calculates the first derivative and second derivative of the loss function at each training sample point.
3. GBDT generates a new decision tree through the greedy strategy and calculates the predicted value of each leaf node.
4. Add the new decision tree into the model.

The GBDT is generally regarded as one of the best out-of-the-box classifiers. It can generalize and can combine weak learners into a single strong learner. The GBDT has many nonlinear transformations and strong performances. There is no need to do complex feature engineering and feature transformation. However, the shortcoming of GBDT is still apparent. Since the boost is a serial process and is difficult to parallelize, GBDT has high computational complexity, and it is also not suitable for high-dimensional sparse features [46].

4.3.4 KNN

The KNN algorithm is one of the simplest classification algorithms and one of the most commonly used learning algorithms.

KNN is a nonparametric statistics method for classification and regression. The mechanism of KNN can be explained as follows: given a test document to be classified, the algorithm searches for the k-nearest neighbors among the pre-classified training documents based on some similarity measure, and ranks those K neighbors based on their similarity scores, the categories of the k nearest neighbors are used to predict the category of the test document by using the ranked scores of each as the weight of the candidate categories, if more than one neighbor belongs to the same category then the sum of their scores is used as the weight of that category, the category with the highest score is assigned to the test document provided that it exceeds a predefined threshold, more than one category can be assigned to the test document [1, 18].

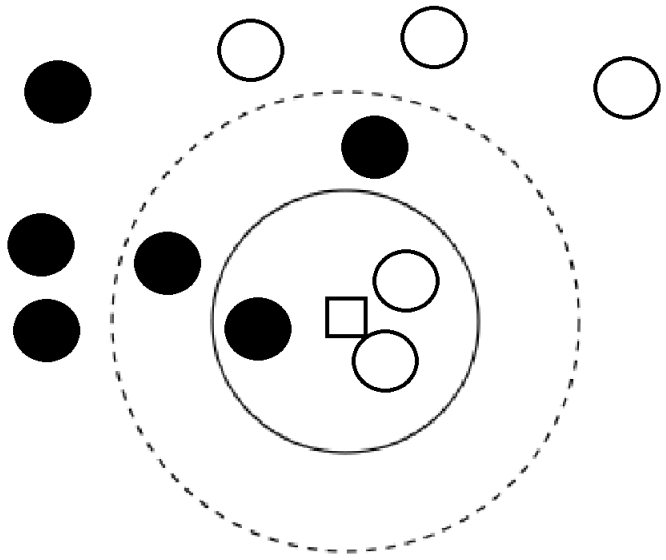

Figure 11: An example of KNN [83]

Figure 11 indicates an example of KNN. White circles and black circles represent two

different classes of sample data. The white square represents data pending for classification. Suppose K = 3, the white square's three nearest points are two white circles and one black circle. Based on statistical methods, this white square belongs to the class of white circles. However, suppose K= 5, the five points closest to the white square are two white circles and three black circles. Based on statistical methods, this white square belongs to the class of the black circles.

4.3.5 DNN

DNN is the basis of deep learning, which is part of a broader family of machine learning methods based on artificial neural networks with representation learning [92].

To understand DNN, firstly, it is essential to understand the DNN model. Figure 12 shows a general model of DNN with two hidden layers.

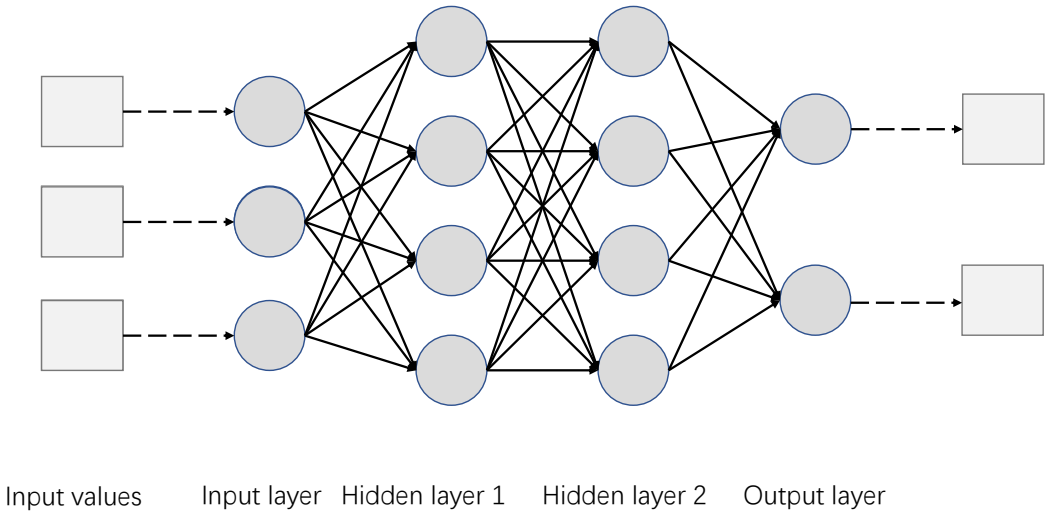

Figure 12: A general model of DNN with N hidden layers [66]

In Figure 12, DNN can be classified into three types of layers: the input layer, the

hidden layer, and the output layer. Usually, the first layer is the input layer, the final layer is the output layer, and the layers in the middle are all hidden layers [14]. A DNN consists of a succession of convolutional and max-pooling layers; the layers are fully connected. Each layer only receives connections from its previous layers.

More precisely, deep learning systems have a substantial credit assignment path (CAP) depth. The CAP is the chain of transformations from input to output. CAPs describe potentially causal connections between input and output [53].

Chapter 5

Statistical analysis and results

This chapter reviews the experimental design and setup of the SWELL-KW dataset. It then quantifies mental stress based on HRV features and selects significant features using the random forest method. To validate the random forest algorithm's performance, comparisons with other related algorithms, including SVM, GBDT, KNN, and DNN, are conducted. HRV data are imported into a Python-based program (see the source code of the program can be found in the Appendix).

5.1 SWELL-KW dataset

This section reviews the SWELL-KW related experimental design and setup. All the experiment data are provided by the SWELL-KW data set, which was collected within the SWELL project [43]. The collection of this data set was supported by researchers at the Institute for Computing and Information Sciences at Radboud University. It is an empirical study in the sense that it is based on real-world data.

In their experiment, they recorded many of the details regarding the data set. Therefore, the results regarding the actual meaning of the variables and classification were presented.

5.1.1 Participants

The collected data are from 25 subjects (seventeen males and eight females) with an average age of 25 [43]. All participants wrote their reports and presentations. They received a standard subject fee for experiment participation. To motivate the participants to do their best on the reports, they were told that the amount of the fee was dependent on their performance.

5.1.2 Design and tasks

In their experiment, Koldijk et al. [43] manipulated the following conditions under which the participants worked:

- Neutral 'No stress': the participants can engage in tasks for an unlimited time. After a maximum of 45 minutes, the participant was asked to stop and informed that enough 'normal work' data had been collected.
- Stressor 'Time pressure': the time to complete all tasks is 2/3 of the time required by the participant in a neutral state (up to 30 minutes).
- Stressor 'Interruptions': the participants received eight e-mails during the task. Some are related to a task, while others are irrelevant. Some e-mails require a reply, while others do not. For example, "Can you look up when Einstein was born?" or "I found many beautiful pictures for this website's demonstration."

Participants are asked to write reports and make presentations on predefined topics. Six topics are prepared, including three opinion topics and three information topics. In the opinion topics, participants need to perform Experience and opinion about 'stress at work', 'healthy living', or 'privacy on the internet'. At the same time, three information topics include describing 5 Tourist attractions in Perth (West Australia), planning a coast to coast road trip in the USA, and writing about the life of Napoleon.

All participants worked under all three conditions. The neutral condition was always the first condition in order to collect an uninfluenced baseline of normal working. The order of the two stressor conditions was counterbalanced, see Figure 13. The within-subject design included relaxation breaks in starting each condition in a well-rested state.

Order		Block1	Q		Block2	Q		Block3	Q
A	Relax	Neutral		Relax	Stressor interruptions		Relax	Stressor time pressure	
B		Neutral			Stressor time pressure			Stressor interruptions	

Figure 13: The design process [43]

In Figure 13, The neutral condition represents no stress situation, which is considered as the baseline of normal working. 13 participants use order A, while 12 participants use order B. The orders of two stressor conditions are balanced.

This data set focuses on high task load stress in-terms of mental demand, frustration, and temporal demand in working professionals [42]. The raw and preprocessed signals are available in the SWELL-KW data set.

5.1.3 Procedure

In order to record the stress response of the experiment and reduce the influence of other factors on the experiment, they instructed the participants not to smoke or drink caffeine 3 hours before the experiment. Before the experiment started, the experiment and records were explained, and all participants signed a consent form to confirm that the recorded data can be used for research purposes. The experiment used body sensors. When the experimenter checked the records, the participants read the experiment description and fill in the questionnaire.

As shown in Figure 13, the experiment is divided into three different blocks for different mental stress conditions. Each block lasts approximately 1 hour. Before starting each block, there exist 8 minutes of relaxation. In each block, participants are provided with two of the six topics selected randomly from the list. The two topics include one opinion topic and one information topic. Participants are asked to write two reports for both topics and choose one topic to make a presentation. Participants are provided with different topics in each block. In both stress conditions, participants were provided with a countdown clock to show the remaining time.

After completing the task, the participants are asked to fill out a questionnaire about the current block. Repeat the relaxation, task execution, and questionnaire process for blocks 2 and 3, as shown in Figure 13. The subjects were given a short rest between these two conditions, and the entire experiment took about 3 hours. After the experiment, participants need to report it.

5.2 Preprocessing

In the SWELL-KW data set, HRV features were computed as follows [63]:

1. An Inter-beat Interval (IBI) signal is extracted from the peaks of the ECG of each subject.

2. Each HRV index is computed on a 5-minute IBI array.

3. A new IBI sample is appended to the IBI array, while the oldest IBI sample is deleted. The new IBI array is used to compute the next HRV index.

This process is repeated until the end of the entire IBI signal.

The inputs of the SWELL-KW dataset were R-R intervals. In this thesis, the provided data sets are preprocessed. The SWELL-KW data set provides both processed training data

and test data, containing 32 features(one feature is deleted) and three different conditions. The training data has 369289 samples, and the test data has 41033 samples.

Generally, physiological signals used for analysis are often pigeonholed by a Non-stationary time performance. Hence, the features of time and frequency exemplifications are desirable. The feature extraction algorithm converts essential information of the original signal into a more condensed lower dimension feature vector. The extracted features explicitly give the stress index of the physiological signals.

This thesis measures mental stress by using the ECG signal. The ECG signal is directly assessed using a commonly used peak finder algorithm [73] to obtain the R-R interval. The power spectral density of the HRV features from the ECG signal extracted using the Welch algorithm dominates the stress detection. The raw ECG is further preprocessed using the window average method [43]. A total of 31 different features are identified for further classification of stress levels. All features are listed in Table 5 and Table 6 for classification.

Table 5: Feature extraction in time domain

number	name	Abbreviation
1	Mean RR	Mean R-R interval
2	Median RR	Median R-R interval
3	SDRR	Standard deviation of R-R interval
4	RMSSD	Root mean square of successive difference in distance
5	SDSD	Standard deviation of all interval of differences between adjacent RR intervals
6	SDRR_RMSSD	Ratio of SDRR over RMSSD
7	HR	Heart rate
8	pNN25	Percentage of number of adjacent RR intervals differing by more than 25 ms
9	pNN50	The ratio of NN50 to the total number of NNs
10	SD1	Short-term poincare plot descriptor of the heart rate variability
11	SD2	Long-term poincare plot descriptor of the heart rate variability
12	SKEW	Skewness of all RR intervals
13	MEAN_REL_RR	Mean of the relative RR
14	MEDIAN_REL_RR	Median of the relative RR
15	SDRR_REL_RR	Standard Deviation of the relative RR
16	RMSSD_REL_RR	Root mean square of successive difference in distance of the relative RR
17	SDSD_REL_RR	Short and long-term poincare plot descriptor of the relative RR
18	SDRR_RMSSD_REL_RR	Ratio of SDRR over RMSSD of the relative RR
19	KURT_REL_RR	Kurtosis of all relative RR intervals
20	SKEW_REL_RR	Skewness of all relative RR intervals

Table 6: Feature extraction in frequency domain

number	name	Abbreviation
21	VLF	Very low frequency power from 0.003 HZ to 0.04Hz
22	VLF_PCT	VLF as a percentage of total
23	LF	FLow frequency power from 0.04 HZ to 0.15Hz
24	LF_PCT	LF as a percentage of total
25	LF_NU	low frequency of HRV in normalized unit
26	HF	High frequency power from 0.15 HZ to 0.4 Hz
27	HF_PCT	HF as a percentage of total
28	HF_NU	high frequency of HRV in normalized unit
29	TP	Total HRV power spectrum
30	LF/HF	Ratio of LF to HF
31	HF/LF	Ratio of HF to LF

Here, we introduce some features defined by math equation. The REL_RR_i [63] can be expressed as

$$REL\_RR_i = 2[\frac{REL_RR_i - REL_RR_{i-1}}{REL_RR_i + REL_RR_{i-1}}] \tag{7}$$

The RMSSD is defined as

$$RMSSD = \frac{\sqrt{\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2}}{N-1} \tag{8}$$

The HF_NU [36] can be expressed as

$$HF\_NU = \frac{HF}{HF + LF} \tag{9}$$

which is similar to the LF_NU feature,

$$LF\_NU = \frac{LF}{HF + LF} \tag{10}$$

In addition, the average LF/HF ratio is defined by

$$R_A = \frac{\sum_{i=1}^{n}(t_i * r_i)}{\sum_{i=1}^{n} t_i} \tag{11}$$

During data analysis, the random forest algorithm is used for feature selection. The sklearn module in Python is used for feature selection. This module includes many machine learning algorithms and models, such as random forest, decision tree, GBDT, SVM, DNN, and KNN. Before using these models, tuning parameters are determined as an important part. The goal of tuning is to achieve a great harmony of deviation and variance of the overall models.

The following description of features gives a detailed overview of the features selected to predict the mental stress [82].

- HR Statistical Feature: Consider the statistical characteristics of the ECG signal. HR is the current rate of heartbeats per minute. The ECG signal's heart rate is calculated by calculating the duration between RR intervals and dividing it every minute. It includes HR.

- HRV Statistical Features: HRV is defined as the change in the time between consecutive sequences of heartbeat intervals. The RR interval is described as the period between two adjacent R waves. The HR and RR intervals are considered to be mutual. The unit of measurement for HR is beats per minute (BPM), and the RR interval in milliseconds (ms). HRV statistical Features include all features in the time domain except for HR.

- Frequency Domain Features of HRV: Bands of frequency are assigned to count the number of RR intervals that match each band. The non-parametric emission PSD analysis was studied using Welch's method. The spectral density of power indicates how power is distributed with frequency, as shown in Table 6.

  The spectral analysis is carried using the following procedure [82]:

  –The ECG signal is split into data segments, with overlapping segments of length (L/2).

  –The Hamming window is applied to the overlapped segments.

  –The task is calculated by Fast Fourier transform, and it is averaged, which results in an array of frequency and power.

The sklearn.preprocessing.StandardScaler class standardizes features by removing the mean and scaling to unit variance.

5.3 Tuning

Before making the comparison, it is essential to perform the hyperparameter optimization of each algorithm. The standard methods in python-sklearn are the GridSearchCV and the RandomizedSearchCV [8]. The principle of GridSearchCV is to select the best set by trying each set of hyperparameters one by one. Concerned about the cost of time, the RandomizedSearchCV is chosen for hyperparameter tuning in this thesis. In this chapter, 50 sets of hyperparameters are chosen randomly and validated by 3-fold cross-validation for all algorithms. After tuning, the optimal set of hyperparameters is used to train the training set and get the predictive model.

Considering that the parameter of each model has a great influence on the final results, the tuning parameter is necessary to do at the beginning. The hyperparameter optimization

results of the training data set regarding the random forest, SVM, decision tree, GBDT, KNN, and DNN are presented.

In the random forest case, the optimal number of features tried at each split is 3. The number of trees used when training the algorithm is 100. The maximum number of random forest features can try in the individual tree is 4. The selection criterion is the Gini impurity.

In the case of the SVM, it starts with the radial basis function kernel. The receiver operating characteristic is displayed as a function of two tuning parameters, including gamma and cost. After hyperparameter tuning, the optimal tuning gives a degree of 3, the optimal gamma value is 0.2575, and the optimal cost is 15.75.

For the decision tree, The optimal selection criterion is the Gini impurity. The two tuning parameters to consider are the depth of the tree and the maximum number of features. After tuning, the optimal depth of the tree is 50. The optimal maximum number of features is 31.

In the case of the GBDT, the optimal number of features tried at each split is 4. The loss function uses deviance loss. The criterion to measure the quality of a split is the mean squared error with an improvement score by Friedman [29]. The number of trees used when training the algorithm is 150. The maximum number of random forest features can try in the individual tree is 4. The depth of each tree is 7.

For the KNN, the optimal number of neighbors is 1. It uses the uniform weighting. The Manhattan distance calculates the distance between real vectors using the sum of their absolute difference in the Minkowski metric.

According to the DNN, it has three hidden layers with 25, 20, and 15 hidden units, respectively. The optimal activation function is relu. The parameter alpha for regularization is 0.0001.

5.4 Feature selection

In this section, each HRV data set is divided into a training set and a test set. We compare random forest predictive performance with five other classification methods (SVM, decision tree, random forest, GBDT, KNN, and DNN). The comparison with other models is based on the test accuracy during different label rates. It is necessary to define accuracy. As well known, the success of the predictive model is calculated based on the degree of the predictive model on the target variable or label of the test data set. The accuracy represents the correct predictions on the total, as shown in equation 11 below,

$$Accuracy = \frac{TP}{TP + FP} \tag{12}$$

where TP indicates true prediction when the predicted values match the actual values of the test dataset label, and FP represents the false prediction when the predicted values don't match the actual values of the test data set labels. The random forest shows better accuracy in feature selection in model comparisons. Hence, the random forest will be chosen as the feature selection model in this project, which makes feature selection from all the features and ranks the selected features based on the accuracy.

The HRV data set includes the training data set and the test data set. The training set has 369289 samples, and the test set has 41033 samples. The data set labels the samples based on three mental stress conditions in the SWELL-KW experiment. The label distribution for each condition is shown in Figure 14 below.

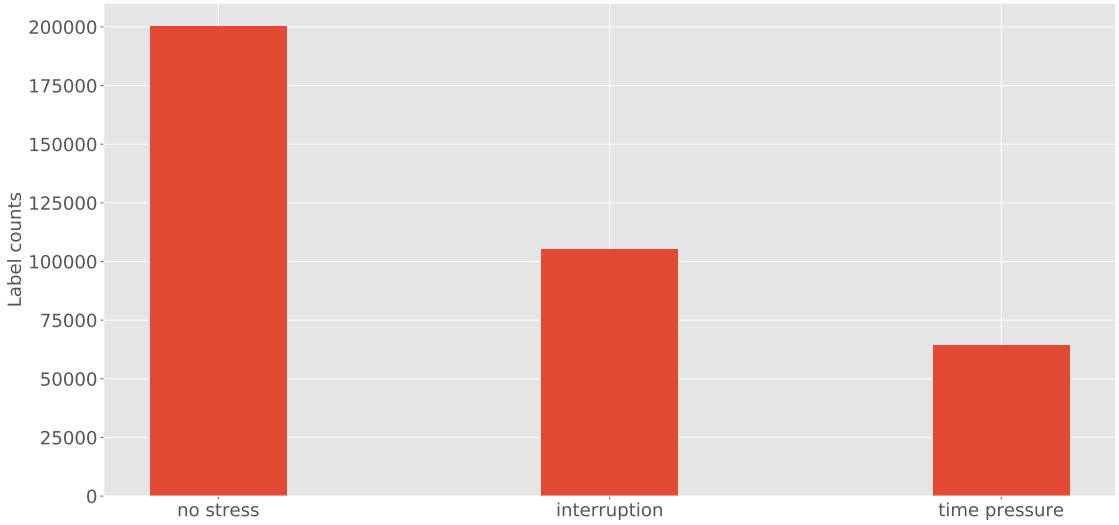

Figure 14: The sample size under three mental stress conditions

This figure shows the number of samples under the 3-label classification. The labels can be classified into no stress, interruption, and time pressure. The condition of no stress has 200082 samples, the interruption condition has 105150 samples, and time pressure has 64057 samples. It can be seen that nearly half of the entire data in the training set is labeled no-stress condition.

While in the test set, the label of no stress has 22158 samples, the interruption label has 11782 samples, and the label of time pressure has 7093 samples.

5.4.1 Comparisons and model choices

In this section, the performances based on the training data set of SVM, decision tree, random forest, GBDT, KNN, and DNN are compared. Then the reliability of random forest is validated.

After selecting and calculating 50 sets of hyperparameters randomly, the optimal parameters are chosen to train the training data set and get the optimal predictive model.

All algorithms perform high accuracy, which is almost close to 1. Since the high accuracy, the controlling label rate of data can observe the connection between features and various algorithms. To reduce the variance, we took the mean result of 5 experiments as the final result. The test accuracy of each model is recorded while increasing the training data's label rate, as shown in Table 7 below.

Table 7: Accuracies of 31 features in different algorithms

label rate (%)	0.01	0.1	0.5	1	2	5
SVM	**57.6**	80.4	**95.8**	**98.5**	**99.5**	99.9
Decision tree	53.1	66.1	86.1	91.7	95.7	98.2
Random forest	**56.3**	**80.7**	**96.1**	**98.4**	**99.5**	99.9
GBDT	**58.9**	**80.6**	95.3	98.0	99.4	99.9
KNN	56.2	**81.8**	**97.3**	**99.2**	**99.8**	99.9
DNN	54.0	71.7	91.3	94.9	97.5	99.0

It is seen from table 7, the relation between classification accuracy (percentage of correct classifications) based on the results on the test set and label rates of training data set, when the label rate is 0.01%, 0.1%, 0.5%, 1%, 2%, 5%, are found. The top 3 accuracies in the different label rates are blackened. As seen in Table 7, compared with SVM, decision tree, GBDT, KNN, and DNN, random forest performance has always been among the top three compared with other models in any label rate situation. Apart from this, GBDT has apparent advantages when the label rate is meager. For example, when the label rate is 0.01%, its accuracy is 58.9 percent, which is 2.3% higher than the SVM. On the other hand, with the increase in the label rate, KNN performance is getting better. The accuracy becomes 99.8% when the label rate is 2% using the KNN method. When the label rate increases to 5%, the decision tree can only achieve an accuracy of 98.2% due to limited fitting ability, while other algorithms can achieve an accuracy of more than 99%. It's worth

noting that random forest always shows great results under different label rates. In order to observe changes visually, the relation between accuracy and label rate in different models is illustrated in Figure 15.

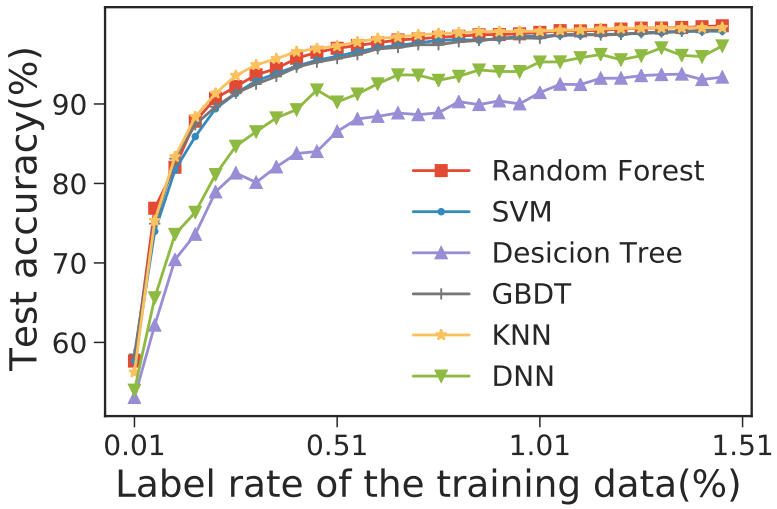

Figure 15: Test accuracy comparisons among different models

In order to see the relationship between the accuracy and algorithms more clearly, in Figure 15, the label rate of training data is from 0% to 1.51%. The interval is not fixed; the test accuracy is plotted as a function of the label rate of models. From this figure, random forest, SVM, decision tree, GBDT, and KNN are significantly better than other algorithms in terms of accuracy. When the label rate is 0.51%, random forest, SVM, and KNN perform better than others.

The results show that the GBDT performs well at a low label rate. Compared with GBDT, the random forest is faster because it can be trained in parallel. What's more, it can still reveal that the decision tree performs the worst among all the models. The reason for its poor effect may be that the model is too simple to describe the data. The DNN has a stronger fitting ability compared with other models. However, when the label rate is low, it is easy to overfit the training set, resulting in low generalization ability on the test set.

Random forest is an ensemble algorithm, and it is more robust than SVM. KNN is slower in real-time than the random forest as it has to keep track of all training data and find the neighbor nodes in the prediction process.

In summary, the random forest model is the optimal model to calculate the HRV feature set and quantify mental stress. Hence, the feature selection model will be based on the random forest.

5.4.2 Feature importance

Random forest builds multiple decision trees and analyzes them together to obtain a more accurate prediction. When training a tree, it can be computed how much each feature decreases the tree's weighted impurity. For a random forest, each feature's impurity decrease can be averaged, and the features are ranked according to this measure.

The random forest is used to calculate feature importance on the training data based on the labeled data set. After calculation, we still rank these features according to their feature importance. The feature importance ranking is plotted in Table 8 below.

Table 8: Feature importance ranking

Rank	Feature number	Importance	Rank	Feature number	Importance
1	feature 6	0.068172	17	feature 16	0.029605
2	feature 0	0.063425	18	feature 23	0.029540
3	feature 1	0.057316	19	feature 11	0.028006
4	feature 7	0.053397	20	feature 9	0.027981
5	feature 13	0.046859	21	feature 19	0.026895
6	feature 10	0.037176	22	feature 14	0.026170
7	feature 17	0.036852	23	feature 3	0.025288
8	feature 2	0.036189	24	feature 25	0.024653
9	feature 28	0.035526	25	feature 30	0.024238
10	feature 5	0.033625	26	feature 27	0.022790
11	feature 22	0.032298	27	feature 29	0.021220
12	feature 15	0.032083	28	feature 26	0.020247
13	feature 18	0.031695	29	feature 4	0.019831
14	feature 20	0.030795	30	feature 24	0.016323
15	feature 8	0.030513	31	feature 12	0.001612
16	feature 21	0.029678			

The feature number is consistent with the feature number in Table 5 and Table 6. It corresponds to the latter's feature name. This table displays the feature ranking, and it also reveals the importance of each feature. In order to understand the difference more intuitively and understand the range of each feature, we provide the feature importance scores in Figure 16.

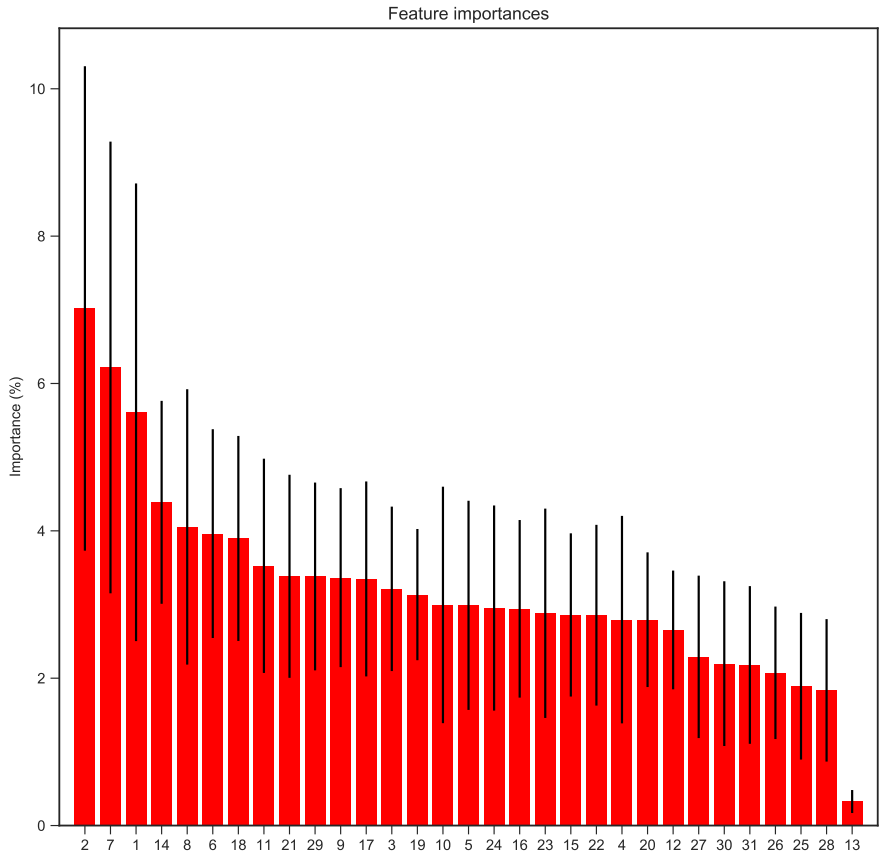

Figure 16: Feature importance in 99.98% accuracy

In Figure 16, the horizontal line represents the feature number, which corresponds to the feature name in Table 5 and Table 6. Each red bar represents the mean importance of each feature. The black line represents the standard deviation of each feature importance.

It is seen in Figure 16 that HR, mean RR, median RR, pNN25, and median REL RR perform better than others, while mean REL RR shows the worst feature importance. Based on the feature importance, this thesis trains the classifier on the training data. The accuracy of the model is also collected, which is up to 99.98% when the label rate is 5%.

The accuracy illustrates that the predictive model training by the training set performs well. There are three reasons for its high accuracy:

1. Data from both the training set and test set is large enough.

2. The quality of the data set is high. No outliers and missing values exist in the acquired data.

3. The proportion of the training set and the test set is relatively balanced.

Feature importance in Figure 16 is valuable because they show their roles in the whole features and indicate that each part of the condition is used for the same features.

Considering the effect of label rate on feature importance, we still rank these features according to their feature importance from 0.1% training data with 80.65%. The feature importance ranking is plotted in Figure 17 below.

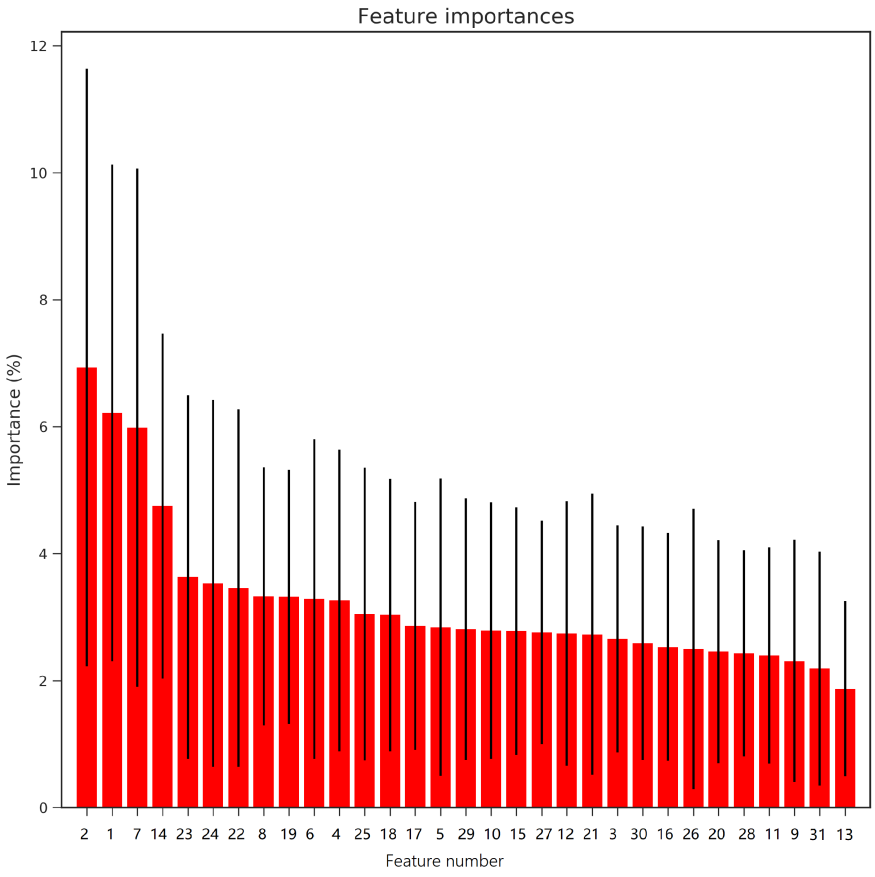

Figure 17: Feature importance in 80.65% accuracy

It is seen in Figure 17 that median RR, mean RR, HR, median REL RR, and LF perform better, while mean REL RR still shows the worst feature importance. In conclusion, median RR, mean RR, HR, and median REL RR show outstanding performance on feature importance in different accuracies. On the other hand, mean REL RR always shows the worst feature importance.

5.4.3 Feature selection and comparison

In this section, the feature importance score of the top 10 features is calculated under the model with 99.98% accuracy. From this importance, the top 10 features are selected to retain the model. Therefore, the feature importance of the top 10 features can be predicted.

The performance of the random forest using the top ten features (median RR, mean RR, median REL RR, HR, pNN25, SDRR RMSSD, SDRR RMSSD REL RR, TP, SD2, SDRR) from 5% training set is evaluated. In Figure 18, when categorizing observations, the top ten features are displayed from the perspective of feature importance according to the random forest algorithm.

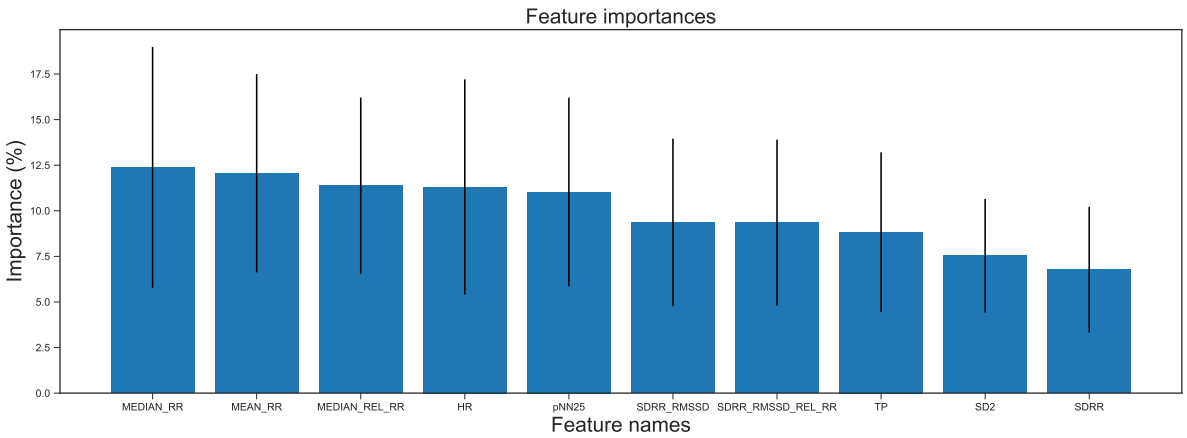

Figure 18: Top 10 features importance

From Figure 18, The feature importance of all ten features adds up to 1. The higher the important value of a feature, the more important the feature is in mental stress quantification

correctly on HRV. The difference between any two features in the top 10 features is not obvious. When using the top ten features only, the random forest still performs very well. Its accuracy is still up to 99.9%. Furthermore, a comparison is made between the random forest and other models using all the features shown in Figure 19.

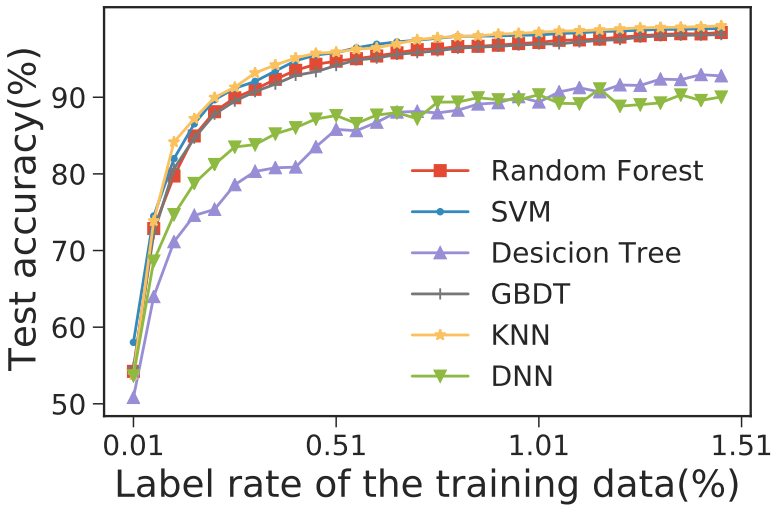

Figure 19: Top 10 features accuracy in different algorithms

Compared with Figure 15, Figure 19 shows that accuracy has no significant difference between the 31 features and 10 features. Table 9 shows the test accuracy from the top 10 features compared with all features in different label rates.

Table 9: The test accuracy gap of 31 features and top 10 features in random forest

label rate(%)	test accuracy (%)		
	31 features	10 features	Δ
0.1	80.7	76.6	4.1
0.5	96.1	93.9	2.2
1	98.4	97.1	1.3
2	99.5	99.2	0.3
5	99.9	99.9	0.0

where Δ represents the accuracy gap between 31 features and 10 features.

In table 9, first, the performance on the training set of the random forest using different

features is evaluated. Then, compared with using all features, the performance of using the top 10 essential features is not better when the label rate is low. For example, when the label rate is 0.1%, the Δ is 4.1%. However, when the label rate gets to a large enough level, the top 10 features can perform pretty well. For example, when the label rate is up to 5%, the Δ of the accuracy is 0. When the label rate is large, there is enough information, no matter the feature dimension's size. Otherwise, the feature dimension has a more significant impact on accuracy when the label rate is low.

The process of the feature dimensionality reduction must lead to information loss. However, if the amount of data is large, enough original information has been captured. It can reduce the error rate to a certain extent.

The accuracy of the top ten features using random forest and KNN outperform other models and algorithms. Table 10 shows the apparent differences between them.

Table 10: The test accuracy from 31 features and top 10 features in different models

label rate(%)	0.05		0.1		0.5	
	31 features	10 features	31 features	10 features	31 features	10 features
SVM	71.8	**72.3**	80.4	**80.1**	**95.8**	**95.7**
decision tree	60.2	64.4	66.1	68.7	86.1	84.8
random forest	**74.2**	**71.7**	**80.7**	**78.1**	**96.1**	**94.6**
GBDT	**72.8**	70.5	**80.6**	77.8	95.3	93.9
KNN	**726**	**71.9**	**81.8**	**81**	**97.3**	**95.6**
DNN	66.4	66.9	71.7	70.9	91.3	88.3

In Table 10, when the label rate is 0.05%, 0.1%, and 0.5%, the top 3 accuracies in the different label rates are blackened. The random forest performance has always been among the top three compared with other models in any label rate situation. When the label rate is up to 0.05%, The performances of SVM and decision tree in the top 10 features are better than in 31 features. Because using only ten features is equivalent to making feature selection, these algorithms can obtain significant features more quickly.

Considering that the time cost has an impact on the algorithms, the training efficiency is analyzed based on the top-10 features shown in Table 11.

Table 11: The training time of each method

models	training time (s)
SVM	47.8
Decision tree	0.28
Random forest	1.61
GBDT	30.2
KNN	6.34
DNN	9.8

In Table 11, the time reported here is the training time using sklearn on a CPU with 20 cores. For KNN and Random Forest, we set the parameter 'n_jobs' to -1, which means we use all cores to train them in parallel.

As we can see, the decision tree is the fastest method, and it is nearly 170 times faster than SVM. However, as we analyzed before, the decision tree is too easy to fit the training data well, and it gets low test accuracy. Among these methods, the random forest is the second fast, and it can also get high test accuracy. Compared with KNN, the random forest is hugely faster in the prediction process. Besides, the training time for random forest can be further reduced if we have more CPU cores. As a result, the random forest can get the most balanced trade-off between training costs and test accuracy. It validates the reliability of the random forest for mental stress quantification.

5.5 Summary

Reducing the cost of feature collection can decrease the time cost for model training and prediction. Another analysis of algorithms comparisons and model choices was conducted to find the reliability of random forest as an indicator of feature selection under mental

stress. The performances on the training data set of SVM, decision tree, random forest, GBDT, KNN, and DNN are compared. The accuracy is collected and analyzed by controlling the label rate.

According to the results, it was found that:

- Given different label rates, random forest performs better than decision tree, GBDT, and DNN in most cases.
- KNN is competitive to the random forest, but it is less used in practical applications due to long prediction time.
- GBDT performs well when there are few labeled data, but it cannot be trained in parallel.
- Compared with SVM, random forest is less sensitive to parameters because it is an ensemble-based method.

It is clear that the random forest model does a better job in HRV feature selection than other algorithms in mental stress quantification. Therefore, the feature selection is reliable based on the random forest method.

Concerning the feature importance, the top ten features are also used for feature selection. When using the top ten features, the random forest still performs very well. The difference between any two features in the top 10 features is not apparent because the top 10 features are all critical to the whole data set.

What still needs to be noted here is that when the data set has two (or more) correlated features, any of these correlated features can be used as the model's predictor, with no concrete preference. Nevertheless, once used one of them, others' importance is significantly reduced since the first feature already removes the impurity they can remove. As a consequence, they will have lower reported importance. It is not an issue when we want to use feature selection to reduce overfitting, since it makes sense to remove features that

are mostly duplicated by other features. However, when interpreting the data, it can lead to the incorrect conclusion that one of the features is a strong predictor while the others in the same group are unimportant, while actually, they are very close in terms of their relationship with the response label. When we have enough examples or when our accuracy requirements are not so high, we can take ten training features. It is also significant in practical applications because using fewer features can reduce the cost of collecting sample information, and it can significantly shorten the time for model training and prediction.

Chapter 6

Conclusion and future work

6.1 Conclusion

In this thesis, a comprehensive review of mental stress, HRV, and Random forest was conducted. Based on this, an inverse U shape relation between mental stress and performance was observed. Mechanisms for decision making is proposed to clarify and quantify mental stress, workload, cognitive workload, mental effort, attention, and cognitive engagement during decision-making activities.

In the design experiment, the designer's mental stress generally increases as his/her mental workload increases. For the correlation between HRV and mental stress, a series of HRV parameters (HR, mean RR, median RR, pNN25, median REL RR, etc.) were identified, which may be affected by mental stress. An analysis of identifying the relationship between mental stress and features is conducted with HRV data from the SWELL-KW program using random forest. The data is segmented based on three mental stress conditions.

According to the results, it was found that:

- The relation between mental stress and other related concepts is validated by the mechanisms for mental stress generation in decision making.

- When sufficient labels are available, the decreasing of feature dimension doesn't affect the accuracy much.
- The feature subset of median RR, mean RR, median REL RR, HR, pNN25, SDRR RMSSD, SDRR RMSSD REL RR, TP, SD2, and SDRR is the optimal subset for mental stress quantification.

The study of the difference and correlation of stress-related concepts indicates HRV does respond to mental stress changes instead of other concepts, which validates the correlation between HRV and mental stress.

The optimal HRV feature subset is proposed for mental stress quantification, which performs higher feature importance than other features. Moreover, when sufficient labels are available, the random forest algorithm using the optimal HRV feature subset yields a higher accuracy in mental stress quantification than other methods.

The decrease of feature dimension can reduce the cost of data collection. Besides, it can significantly shorten the time of model training and prediction. Feature selection is of great significance in real-world applications.

6.2 Future work

Throughout this thesis, machine learning depends on massive amounts of data. However, the collection of massive data costs lots of effort and money. Can we find the critical value that gives better results while using enough data? This research can facilitate the effective and efficient applications of existing research to real-world problems.

In order to reduce the cost of large-scale data collection, one conventional method is active learning. It allows labeling less data by selecting the most important samples from the learning process. This method aims to minimize the labeling cost and maximize the performance of the machine learning model. Since the unlabeled data is more comfortable to

obtain, semi-supervised learning can still be used in feature extraction and selection. When the label rate is low, the semi-supervised learning can mine a large amount of information based on unlabeled data. It can also improve accuracy and reduce costs. The research on improving accuracy under the given label rate is also essential.

Bibliography

[1] Riyad Al-Shalabi, Ghassan Kanaan, and M Gharaibeh. Arabic text categorization using knn algorithm. In *Proceedings of The 4th International Multiconference on Computer Science and Information Technology*, volume 4, pages 5–7, 2006.

[2] FM Al-Shargie, Tong Boon Tang, Nasreen Badruddin, and Masashi Kiguchi. Mental stress quantification using eeg signals. In *International Conference for Innovation in Biomedical Engineering and Life Sciences*, pages 15–19. Springer, 2015.

[3] Guzmán Alba, Jaime Vila, Beatriz Rey, Pedro Montoya, and Miguel Ángel Muñoz. The relationship between heart rate variability and electroencephalography functional connectivity variability is associated with cognitive flexibility. *Frontiers in human neuroscience*, 13:64, 2019.

[4] Christopher Alexander. *A pattern language: towns, buildings, construction*. Oxford university press, 1977.

[5] Alan Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[6] Hamid Reza Baghaee, Dragan Mlakić, Srete Nikolovski, and Tomislav Dragičević. Support vector machine-based islanding and grid fault detection in active distribution networks. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2019.

[7] Dalius Bansevicius, Rolf H Westgaard, and Chris Jensen. Mental stress of long duration: Emg activity, perceived tension, fatigue, and pain development in pain-free subjects. *Headache: The Journal of Head and Face Pain*, 37(8):499–510, 1997.

[8] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[9] L Breiman, JH Friedman, R Olshen, and CJ Stone. Classification and regression trees. 1984.

[10] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[11] Walter Bradford Cannon. *Bodily changes in pain, hunger, fear, and rage*. D. Appleton and company, 1915.

[12] Walter Bradford Cannon. The wisdom of the body. 1939.

[13] Alex Cao, Keshav K Chintamani, Abhilash K Pandya, and R Darin Ellis. Nasa tlx: Software for assessing subjective mental workload. *Behavior research methods*, 41(1):113–117, 2009.

[14] Dan CireşAn, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural networks*, 32:333–338, 2012.

[15] Hagit Cohen, Michael A Matar, Zeev Kaplan, and Moshe Kotler. Power spectral analysis of heart rate variability in psychiatry. *Psychotherapy and psychosomatics*, 68(2):59–66, 1999.

[16] Sheldon Cohen and Natalie Hamrick. Stable individual differences in physiological response to stressors: Implications for stress-elicited changes in immune related health. *Brain, behavior, and immunity*, 17(6):407–414, 2003.

[17] Richard Contrada and Andrew Baum. *The handbook of stress science: Biology, psychology, and health*. Springer Publishing Company, 2010.

[18] Padraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers–. *arXiv preprint arXiv:2004.04523*, 2020.

[19] Mary F Dallman and Dirk Hellhammer. Regulation of the hypothalamopituitaryadrenal axis, chronic stress, and energy: the role of brain networks. *The handbook of stress science: Biology, psychology, and health*, pages 11–36, 2011.

[20] RJSG Davidson. Frontal versus perietal eeg asymmetry during positive and negative affect. *Psychophysiology*, 16(2):202–203, 1979.

[21] Firdaus S Dhabhar. Effects of stress on immune function: Implications for immunoprotection and immunopathology. 2011.

[22] Roger Daglius Dias, Minhtran C Ngo-Howard, Marko T Boskovski, Marco A Zenati, and Steven J Yule. Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. *The British journal of surgery*, 105(5):491, 2018.

[23] Sally S Dickerson and Margaret E Kemeny. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychological bulletin*, 130(3):355, 2004.

[24] Thomas G Dietterich, Hermann Hild, and Ghulum Bakiri. A comparison of id3 and backpropagation for english text-to-speech mapping. *Machine Learning*, 18(1):51–80, 1995.

[25] Elias Ebrahimzadeh, Maede Kalantari, Mohammadamin Joulani, Reza Shahrokhi Shahraki, Farahnaz Fayaz, and Fereshteh Ahmadi. Prediction of paroxysmal atrial fibrillation: A machine learning based approach using combined feature vector and

mixture of expert classification on hrv signal. *Computer methods and programs in biomedicine*, 165:53–67, 2018.

[26] Willem Einthoven. Die galvanometrische registrirung des menschlichen elektrokardiogramms, zugleich eine beurtheilung der anwendung des capillar-elektrometers in der physiologie. *Archiv für die gesamte Physiologie des Menschen und der Tiere*, 99(9-10):472–480, 1903.

[27] JP Fauvel, N Bernard, M Laville, S Daoud, N Pozet, and P Zech. Reproducibility of the cardiovascular reactivity to a computerized version of the stroop stress test in normotensive and hypertensive subjects. *Clinical Autonomic Research*, 6(4):219–224, 1996.

[28] Rhona Flin, Eduardo Salas, Michael Straub, and Lynne Martin. *Decision-making under stress: Emerging themes and applications*. Routledge, 2017.

[29] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[30] Luigi Galvani. De viribus electricitatis in motu musculari. commentarius. *De Bonoiensi Scientiarum et Artium Intituo atque Academie Commentarii*, 7:363–418, 1791.

[31] James Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Robert. *An introduction to statistical learning: with applications in R*. Spinger, 2013.

[32] David S Goldstein and Bruce McEwen. Allostasis, homeostats, and the nature of stress. *Stress*, 5(1):55–58, 2002.

[33] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[34] ML Heemstra. An efficiency model of information processing. In *Energetics and human information processing*, pages 233–242. Springer, 1986.

[35] John Herbert. Fortnightly review: Stress, the brain, and mental illness. *Bmj*, 315(7107):530–535, 1997.

[36] Markad V Kamath, Mari Watanabe, and Adrian Upton. *Heart rate variability (HRV) signal analysis: clinical applications*. CRC Press, 2012.

[37] Arnold M Katz. *Physiology of the Heart*. Lippincott Williams & Wilkins, 2010.

[38] Giora Keinan. Decision making under stress: scanning of alternatives under controllable and uncontrollable threats. *Journal of personality and social psychology*, 52(3):639, 1987.

[39] Reza Khosrowabadi, Chai Quek, Kai Keng Ang, Sau Wai Tung, and Michel Heijnen. A brain-computer interface for classifying eeg correlates of chronic mental stress. In *The 2011 International Joint Conference on Neural Networks*, pages 757–762. IEEE, 2011.

[40] Clemens Kirschbaum, Brigitte M Kudielka, Jens Gaab, Nicole C Schommer, and Dirk H Hellhammer. Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary-adrenal axis. *Psychosomatic medicine*, 61(2):154–162, 1999.

[41] Clemens Kirschbaum, Stefan Wüst, and Dirk Hellhammer. Consistent sex differences in cortisol responses to psychological stress. *Psychosomatic medicine*, 54(6):648–657, 1992.

[42] Saskia Koldijk, Mark A Neerincx, and Wessel Kraaij. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Transactions on Affective Computing*, 9(2):227–239, 2016.

[43] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A Neerincx, and Wessel Kraaij. The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction*, pages 291–298, 2014.

[44] Wolf Langewitz, Heinz Rüddel, and Hartmut Schächinger. Reduced parasympathetic cardiac control in patients with hypertension at rest and under mental stress. *American heart journal*, 127(1):122–128, 1994.

[45] Qiang Li, Lei Chen, Xiangju Li, Xiaofeng Lv, Shuyue Xia, and Yan Kang. Prf-rw: a progressive random forest-based random walk approach for interactive semi-automated pulmonary lobes segmentation. *International Journal of Machine Learning and Cybernetics*, pages 1–15, 2020.

[46] Zhijun Liao, Yong Huang, Xiaodong Yue, Huijuan Lu, Ping Xuan, and Ying Ju. In silico prediction of gamma-aminobutyric acid type-a receptors using novel machine-learning-based svm and gbdt approaches. *BioMed research international*, 2016, 2016.

[47] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[48] William R Lovallo, Thomas L Whitsett, Mustafa Al'Absi, Bong Hee Sung, Andrea S Vincent, and Michael F Wilson. Caffeine stimulation of cortisol secretion across the waking hours in relation to caffeine intake levels. *Psychosomatic medicine*, 67(5):734, 2005.

[49] Aleksandar Malinovic. Fast stress detection via ecg. Master's thesis, University of Waterloo, 2019.

[50] Jaakko Malmivuo, Robert Plonsey, et al. *Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields*. Oxford University Press, USA, 1995.

[51] Stephen B Manuck, Sheldon Cohen, Bruce S Rabin, Matthew F Muldoon, and Elizabeth A Bachen. Individual differences in cellular immune response to stress. *Psychological science*, 2(2):111–115, 1991.

[52] Bruce S McEwen. Protective and damaging effects of the mediators of stress and adaptation: Allostasis and allostatic load. 2004.

[53] Yajie Miao, Hao Zhang, and Florian Metze. Distributed learning of multilingual dnn feature extractors using gpus. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[54] Upasana Mishra and Love Verma. Noise removal from ecg signal by thresholding with comparing different types of wavelet. *International Journal of Application or Innovation in engineering and Management*, 3(3), 2014.

[55] Laurent Mourot, Malika Bouhaddi, and Jacques Regnard. Effects of the cold pressor test on cardiac autonomic control in normal subjects. *Physiological research*, 58(1), 2009.

[56] Thanh An Nguyen, Xu Xu, Yong Zeng, et al. Distribution of mental stresses during conceptual design activities. In *DS 75-7: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol. 7: Human Behaviour in Design, Seoul, Korea, 19-22.08. 2013*, pages 287–296, 2013.

[57] Thanh An Nguyen and Yong Zeng. Analysis of design activities using eeg signals. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 44137, pages 277–286, 2010.

[58] Thanh An Nguyen and Yong Zeng. A theoretical model of design creativity: Nonlinear design dynamics and mental stress-creativity relation. *Journal of Integrated Design and Process Science*, 16(3):65–88, 2012.

[59] Thanh An Nguyen and Yong Zeng. A physiological study of relationship between designer's mental effort and mental stress during conceptual design. *Computer-Aided Design*, 54:3–18, 2014.

[60] Thanh An Nguyen and Yong Zeng. Effects of stress and effort on self-rated reports in experimental study of design activities. *Journal of Intelligent Manufacturing*, 28(7):1609–1622, 2017.

[61] Thanh An Nguyen and Yong Zeng. A theoretical model of design fixation. *International Journal of Design Creativity and Innovation*, 5(3-4):185–204, 2017.

[62] Kizito Nkurikiyeyezu, Kana Shoji, Anna Yokokubo, and Guillaume Lopez. Thermal comfort and stress recognition in office environment. In *HEALTHINF*, pages 256–263, 2019.

[63] Kizito Nkurikiyeyezu, Anna Yokokubo, and Guillaume Lopez. The effect of person-specific biometrics in improving generic stress predictive models. *arXiv preprint arXiv:1910.01770*, 2019.

[64] Thomas E Nygren. Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human factors*, 33(1):17–33, 1991.

[65] Task Force of the European Society of Cardiology et al. Heart rate variability: standards of measurement, physiological interpretation and clinical use. *circulation*, 93:1043–1065, 1996.

[66] Josh Patterson and Adam Gibson. *Deep learning: A practitioner's approach*. " O'Reilly Media, Inc.", 2017.

[67] Radim Plhal, Jiří Kamler, and Miloslav Homolka. Faecal pellet group counting as a promising method of wild boar population density estimation. *Acta Theriologica*, 59(4):561–569, 2014.

[68] Marc Pomplun and Sindhura Sunkara. Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of the International Conference on HCI*, volume 273, 2003.

[69] Andrew Pullan, Martin L Buist, and Leo K Cheng. *Mathematically modelling the electrical activity of the heart: from cell to body surface and back again*. World Scientific Publishing Company, 2005.

[70] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[71] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[72] Robert Rauh, Michaela Burkert, Martin Siepmann, and Michael Mueck-Weymann. Acute effects of caffeine on heart rate variability in habitual caffeine consumers. *Clinical physiology and functional imaging*, 26(3):163–166, 2006.

[73] Deboleena Sadhukhan and Madhuchhanda Mitra. R-peak detection algorithm for ecg using double difference and rr interval processing. *Procedia Technology*, 4:873–877, 2012.

[74] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[75] Lizawati Salahuddin, Jaegeol Cho, Myeong Gi Jeong, and Desok Kim. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In *2007 29th annual international conference of the ieee engineering in medicine and biology society*, pages 4656–4659. IEEE, 2007.

[76] C Saritha, V Sukanya, and Y Narasimha Murthy. Ecg signal analysis using wavelet transforms. *Bulg. J. Phys*, 35(1):68–77, 2008.

[77] C Schubert, M Lambertz, RA Nelesen, W Bardwell, J-B Choi, and JE Dimsdale. Effects of stress on heart rate complexity—a comparison between short-term and chronic stress. *Biological psychology*, 80(3):325–332, 2009.

[78] Hans Selye. Stress without distress. In *Psychopathology of human adaptation*, pages 137–146. Springer, 1976.

[79] Ivana Semanjski and Sidharta Gautama. Smart city mobility application—gradient boosting trees for mobility prediction and analysis based on crowdsourced data. *Sensors*, 15(7):15974–15987, 2015.

[80] Fred Shaffer and JP Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, 5:258, 2017.

[81] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[82] S Sriramprakash, Vadana D Prasanna, and OV Ramana Murthy. Stress detection in working people. *Procedia computer science*, 115:359–366, 2017.

[83] Tavish Srivastava. Introduction to k-nearest neighbors: A powerful machine learning algorithm (with implementation in python & r). *Analyticsvidhya. com*, 2018.

[84] Mark A Staal. Stress, cognition, and human performance: A literature review and conceptual framework. 2004.

[85] Ganyun Sun and Shengji Yao. Investigating the relation between cognitive load and creativity in the conceptual design process. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 56, pages 308–312. SAGE Publications Sage CA: Los Angeles, CA, 2012.

[86] Joachim Taelman, Steven Vandeput, Arthur Spaepen, and Sabine Van Huffel. Influence of mental stress on heart rate and heart rate variability. In *4th European conference of the international federation for medical and biological engineering*, pages 1366–1369. Springer, 2009.

[87] Eleonora Tobaldini, Lino Nobili, Silvia Strada, Karina Rabello Casali, Alberto Braghiroli, and Nicola Montano. Heart rate variability in normal and pathological sleep. *Frontiers in physiology*, 4:294, 2013.

[88] Don M Tucker. Lateral brain function, emotion, and conceptualization. *Psychological bulletin*, 89(1):19, 1981.

[89] L Vanitha, GR Suresh, M Chandrasekar, and P Punita. Development of four stress levels in group stroop colour word test using hrv analysis. 2017.

[90] Ronald G Victor, WAYNE N Leimbach Jr, Douglas R Seals, B Gunnar Wallin, and Allyn L Mark. Effects of the cold pressor test on muscle sympathetic nerve activity in humans. *Hypertension*, 9(5):429–436, 1987.

[91] Augustus D Waller. A demonstration on man of electromotive changes accompanying the heart's beat. *The Journal of physiology*, 8(5):229, 1887.

[92] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092, 2015.

[93] Douglas L Wood, Sheldon G Sheps, Lila R Elveback, and Alexander Schirger. Cold pressor test as a predictor of hypertension. *Hypertension*, 6(3):301–306, 1984.

[94] Jerry Ye, Jyh-Herng Chow, Jiang Chen, and Zhaohui Zheng. Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 2061–2064, 2009.

[95] Robert M Yerkes, John D Dodson, et al. The relation of strength of stimulus to rapidity of habit-formation. *Punishment: Issues and experiments*, pages 27–41, 1908.

[96] Shaoda Yu, Peng Li, Honghuang Lin, Ehsan Rohani, Gwan Choi, Botang Shao, and Qian Wang. Support vector machine based detection of drowsiness using minimum eeg features. In *2013 International Conference on Social Computing*, pages 827–835. IEEE, 2013.

Appendix A

Python code for HRV feature selection

```
import numpy as np
    import pandas as pd
    from sklearn.ensemble import RandomForestClassifier
    import matplotlib.pyplot as plt
    import seaborn as sns
    import warnings
    warnings.filterwarnings('ignore')
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import StandardScaler
    from sklearn.metrics import accuracy_score
    from sklearn.model_selection import train_test_split
    from sklearn.svm import SVC
    from sklearn import tree
    from sklearn.neural_network import MLPClassifier
    from sklearn.neighbors import KNeighborsClassifier
    from sklearn.ensemble import GradientBoostingClassifier
    from sklearn.model_selection import RandomizedSearchCV
```

1.read data

train = pd.read_excel('data.xlsx', sheet_name='train')

test = pd.read_excel('data.xlsx', sheet_name='test')

train_values = train.values

test_values = test.values

full_data = np.concatenate((train_values,test_values),axis=0)

X = full_data[:,:-1]

y = full_data[:,-1]

np.save("X.npz",X)

np.save("y.npz",y)

2. data analysis

One_value_array = []

for i in range(33):

if len(train.iloc[:, i].value_counts()) == 1:

One_value_array.append(str(i))

delete 31 lines

X = np.delete(X, -1, axis=1)

scaler=StandardScaler()

model=scaler.fit(X)

X = model.transform(X)

3.modeling structure

X_train_val = X[:len(train)]

X_test = X[:len(test)]

y_train_val = y[:len(train)]

y_test = y[:len(test)]

print(len(X_train_val))

print(len(X_test))

4.label rate is 0.005

rate = 0.005

X_train, X_val, y_train, y_val = train_test_split(X_train_val, y_train_val,

test_size=1-rate, random_state=0, shuffle=True, stratify=y_train_val)

4.1 default accuracy

model = RandomForestClassifier(n_jobs=-1,random_state=0)

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test)

test_acc = accuracy_score(y_test,pred_test)

print('RF-test accuracy:', test_acc)

model = SVC()

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test)

test_acc = accuracy_score(y_test,pred_test)

print('SVM-test accuracy:', test_acc)

model = tree.DecisionTreeClassifier()

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test)

test_acc = accuracy_score(y_test,pred_test)

print('DT-test accuracy:', test_acc)

model = GradientBoostingClassifier(random_state=0)

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test)

test_acc = accuracy_score(y_test,pred_test)

print('GBDT:', test_acc)

model = KNeighborsClassifier(n_jobs=-1)

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test)

test_acc = accuracy_score(y_test,pred_test)

print('KNN:', test_acc)

model = MLPClassifier(solver='lbfgs', alpha=1e-5,hidden_layer_sizes=(20,10,5), random_state=0)

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test)

test_acc = accuracy_score(y_test,pred_test)

print('DNN:', test_acc)

4.2 tuning parameters

1RF

clf = RandomForestClassifier(n_jobs=-1,random_state=0)

list or distribution

param_dist = "max_depth": [3, None], list "n_estimators": [50,100,150],

"max_features": range(1, 11), distribution

"min_samples_split": range(1,5), distribution

"bootstrap": [True, False], list

"criterion": ["gini", "entropy"] list

RandomSearch+CV select Hyperparameters

n_iter_search = 50 random_search = RandomizedSearchCV(clf, param_distributions=param_dist,

n_iter=n_iter_search ,cv=3,scoring='accuracy', n_jobs = 8, iid=False, verbose=1) ran-
dom_search.fit(X_train, y_train)

best_parameters = random_search.best_estimator_.get_params()

for para, val in list(best_parameters.items()):

print(para, val)

model = RandomForestClassifier(n_jobs=-1,max_depth=best_parameters['max_depth'],n_estimators=

ters['n_estimators'],

max_features=best_parameters['max_features'],

min_samples_split=best_parameters['min_samples_split'],

bootstrap=best_parameters['bootstrap'],

criterion=best_parameters['criterion'],

random_state=0)

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test)

test_acc = accuracy_score(y_test,pred_test)

print('RF-test accuracy:', test_acc)

2)SVM

clf = SVC()

list or distribution

param_dist = 'C':np.linspace(3,20,5),'gamma':np.linspace(0.01,1,5)

RandomSearch+CV select Hyperparameters

n_iter_search = 50

random_search = RandomizedSearchCV(clf, param_distributions=param_dist,

n_iter=n_iter_search ,cv=3,scoring='accuracy', n_jobs = 8, iid=False, verbose=1) random_search.fit(X_train, y_train)

best_parameters = random_search.best_estimator_.get_params()

for para, val in list(best_parameters.items()):

print(para, val)

model = SVC(kernel='rbf', C=best_parameters['C'], gamma=best_parameters['gamma'], probability=True)

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test)

test_acc = accuracy_score(y_test,pred_test)

print('SVM-test accuracy:', test_acc)

3)DT

model = tree.DecisionTreeClassifier()

param_dist = "max_depth":[10,50,100,200,None],"max_features": [1,3,5,7,None]

n_iter_search = 50 random_search = RandomizedSearchCV(model, param_distributions=param_dist,

n_iter=n_iter_search ,cv=3,scoring='accuracy', n_jobs = 8, iid=False, verbose=1) random_search.fit(X_train, y_train)

best_parameters = random_search.best_estimator_.get_params() for para, val in list(best_parameters.it

print(para, val)

model = tree.DecisionTreeClassifier(max_depth=best_parameters['max_depth'], max_features=best_p

random_state=0)

model.fit(X_train,y_train)

pred_val = model.predict(X_val) val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test) test_acc = accuracy_score(y_test,pred_test)

print('DT-test accuracy:', test_acc)

4)GBDT

model = GradientBoostingClassifier(random_state=0)

list or distribution param_dist = "max_depth": range(5,10), list

"n_estimators": [50,100,150],

"max_features": range(1, 11), distribution

"min_samples_split": range(1,5)

n_iter_search = 50

random_search = RandomizedSearchCV(model, param_distributions=param_dist,

n_iter=n_iter_search ,cv=3,scoring='accuracy', n_jobs = 8, iid=False, verbose=1) ran-
dom_search.fit(X_train, y_train)

best_parameters = random_search.best_estimator_.get_params()

for para, val in list(best_parameters.items()):

print(para, val)

model = GradientBoostingClassifier(max_depth=best_parameters['max_depth'],n_estimators=best_pa

max_features=best_parameters['max_features'], min_samples_split=best_parameters['min_samples_

random_state=0)

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

```
val_acc = accuracy_score(y_val,pred_val)
pred_test = model.predict(X_test)
test_acc = accuracy_score(y_test,pred_test)
print('GBDT-test accuracy:', test_acc)
5)KNN
model = KNeighborsClassifier(n_jobs=-1)
list or distribution
param_dist =  'weights':['distance','uniform'],
'n_neighbors':[i for i in range(1,11)],
'p':[i for i in range(1,6)]
n_iter_search = 50
random_search = RandomizedSearchCV(model, param_distributions=param_dist,
n_iter=n_iter_search ,cv=3,scoring='accuracy', n_jobs = 8, iid=False, verbose=1) ran-
dom_search.fit(X_train, y_train)
best_parameters = random_search.best_estimator_.get_params()
for para, val in list(best_parameters.items()):
print(para, val)
model = KNeighborsClassifier(n_jobs=-1,weights=best_parameters['weights'],
n_neighbors=best_parameters['n_neighbors'],
p=best_parameters['p'],)
model.fit(X_train,y_train)
pred_val = model.predict(X_val)
val_acc = accuracy_score(y_val,pred_val)
pred_test = model.predict(X_test)
test_acc = accuracy_score(y_test,pred_test)
print('KNN:', test_acc)
```

6MLP

model = MLPClassifier(hidden_layer_sizes=(25,20,15), random_state=0)

list or distribution

param_dist = 'activation':['identity','logistic','tanh','relu'],

'alpha':[1e-3,1e-4,1e-5]

n_iter_search = 50

random_search = RandomizedSearchCV(model, param_distributions=param_dist,

n_iter=n_iter_search ,cv=3,scoring='accuracy', n_jobs = 8, iid=False,

verbose=1) random_search.fit(X_train, y_train)

best_parameters = random_search.best_estimator_.get_params()

for para, val in list(best_parameters.items()):

print(para, val)

model = MLPClassifier(solver='lbfgs', activation=best_parameters['activation'],

alpha=best_parameters['alpha'],

hidden_layer_sizes=(25,20,15), random_state=0)

model.fit(X_train,y_train)

pred_val = model.predict(X_val)

val_acc = accuracy_score(y_val,pred_val)

pred_test = model.predict(X_test)

test_acc = accuracy_score(y_test,pred_test)

print('DNN:', test_acc)