NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

REDUCING AVIATION FATALITIES BY MONITORING PILOTS' COGNITIVE STATES USING PSYCHOPHYSIOLOGICAL MEASUREMENTS

by

Yi-chung Lin

June 2021

Thesis Advisor:	Ruriko Yoshida
Second Reader:	Thomas W. Lucas

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2021	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
REDUCING AVIATION FATALITIES BY MONITORING PILOTS' COGNITIVE STATES USING PSYCHOPHYSIOLOGICAL MEASUREMENTS

5. FUNDING NUMBERS

6. AUTHOR(S) Yi-chung Lin

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

Airplane accidents are usually catastrophic, and the majority of flight-related accidents are caused by a lack of situational awareness during flight. To improve flight safety, we built a model to detect the cognitive states of pilots from their psychophysiological signals so that the aviators can be warned before falling into a dangerous mental state, including channelized attention, diverted attention, and startle/surprise. The research is composed of time series analysis and classification. We used seasonal decomposition, exponential smoothing, and autoregressive integrated moving average models to analyze the numerical psychophysiological measurements of 18 pilots and utilize such measurements to distinguish their cognitive states by classification methods, such as random forest, support vector machine, and logistic regression. The results can be a part of the risk management mechanism to alert pilots when necessary. The deliverables include a classification model of the problem and an analysis of the solutions obtained from the model. These models are written in R so that anyone can run calculations in real time to monitor the cognitive states of pilots and to support follow-on/future analysis work.

14. SUBJECT TERMS
channelized attention, diverted attention, startle/surprise, airplane state awareness, electroencephalogram, electrocardiogram, galvanic skin response

15. NUMBER OF PAGES
83

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

REDUCING AVIATION FATALITIES BY MONITORING PILOTS' COGNITIVE STATES USING PSYCHOPHYSIOLOGICAL MEASUREMENTS

Yi-chung Lin
Major, Taiwanese Armed Forces
BE, R.O.C. Air Force Academy, 2007

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2021

Approved by: Ruriko Yoshida
Advisor

Thomas W. Lucas
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Airplane accidents are usually catastrophic, and the majority of flight-related accidents are caused by a lack of situational awareness during flight. To improve flight safety, we built a model to detect the cognitive states of pilots from their psychophysiological signals so that the aviators can be warned before falling into a dangerous mental state, including channelized attention, diverted attention, and startle/surprise. The research is composed of time series analysis and classification. We used seasonal decomposition, exponential smoothing, and autoregressive integrated moving average models to analyze the numerical psychophysiological measurements of 18 pilots and utilize such measurements to distinguish their cognitive states by classification methods, such as random forest, support vector machine, and logistic regression. The results can be a part of the risk management mechanism to alert pilots when necessary. The deliverables include a classification model of the problem and an analysis of the solutions obtained from the model. These models are written in R so that anyone can run calculations in real time to monitor the cognitive states of pilots and to support follow-on/future analysis work.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

ARIMA	Auto-Regressive Integrated Moving Average
ASA	airplane state awareness
AUC	area under the curve
CA	channelized attention
CAST	Commercial Aviation Safety Team
DA	diverted attention
ECG	electrocardiogram
EEG	electroencephalogram
ETS	Exponential Smoothing
GBM	Gradient Boosting Machine
GSR	galvanic skin response
IID	independently and identically distributed
KNN	K-Nearest Neighbors
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MSE	Mean Squared Error
NASA	National Aeronautics and Space Administration
R	respiration
ROC	Receiver Operating Characteristics

rpart	Recursive Partitioning and Regression Trees
SA	situational awareness
SC	skin conductance
SE	safety enhancement
SS	startle / surprise
STL	Seasonal Decomposition
SVM	Support Vector Machine

Executive Summary

A lack of situational awareness (SA) during flight is the major cause of flight-related accidents (Moroze and Snow 1999). That is, pilots who are distracted, asleep, or in other risky mental states fail to maintain their focus during flight effectively. The Commercial Aviation Safety Team (CAST) discovered that 13 out of 18 fatal events from 2003 to 2012 could be attributed to the loss of control in-flight were caused by lack of SA among flight crews (Rosenkrans 2015). In other words, these flight crews all experienced mental diversion. Therefore, analyzing the cognitive states of pilots is important to avoid distraction for flight safety.

Two years ago, Kaggle, a subsidiary of Google, hosted a competition named "Reducing Commercial Aviation Fatalities," which challenged data scientists and machine learning practitioners around the world to build a model capable of monitoring a pilot's mental state in real time by measuring his or her psychophysiological data, including electroencephalogram (EEG) recordings, electrocardiogram (ECG) signal, respiration (R), and galvanic skin response (GSR) (Kaggle 2019).

Subsequently, the Electronics and Communication Engineering CMR Institute of Technology in India analyzed the dataset from the Kaggle competition and published an article in the Second International Conference on Smart Systems and Inventive Technology (Mishra et al. 2019). The same dataset was also analyzed in 2019 in a master's degree thesis by a student in the Netherlands (Crijnen 2019). These two analyses revealed the importance and complexity of the dataset and how valuable the results could be for the aviation.

Additionally, a similar study conducted by the National Aeronautics and Space Administration (NASA) Langley Research Center also used EEG, ECG, GSR, and R signals from 13 participants as input features to predict seven different pilot cognitive states consisting of diverted attention, channelized attention, low workload, high workload, confirmation bias, startle/surprise, and rest (Harrivel et al. 2017). NASA's research showed that correctly identifying pilots' mental states is feasible by analyzing psychophysiological data.

In this research, our objective is to build a model using the dataset from the Kaggle competition to detect pilots' mental states (defined as either "safe" or "dangerous") to give

an immediate warning if they fall into hazardous mental states so that they will have more time to regain SA before the altered state becomes irreparable.

The performance of the classification model indicates that the Random Forest method outperforms all other classifiers, including Logistic Regression, Support Vector Machines, K-Nearest Neighbors, and Recursive Partitioning and Regression Trees, with approximately 90% reliability to discern pilots' mental states. Furthermore, according to the result of the Feature Importance assessment from the Random Forest algorithm, we can detect whether a pilot is distracted approximately 90% of the time by simply measuring his or her ECG, GSR, and R values.

In an attempt to predict pilots' mental states before they fall into a dangerous state, we also performed a time series analysis under the presumption that pilots' psychophysiological values can be predicted in terms of time. We interpreted the time series data using Naïve, Seasonal Decomposition, Exponential Smoothing, Auto-Regressive Integrated Moving Average (ARIMA), and Ensemble models. The forecasting performance shows that ECG, GSR, and R signals are predictable, and the ARIMA model performs the best in those three modalities.

Flight safety is crucial not only to mission success but civilian air travel. Therefore, we encourage researchers from around the world with interests in both commercial and military aviation safety to join the research in this field in order to reduce aviation mishaps caused by human factors. With our effort, flight accidents can be forestalled, and many lives can be saved.

References

Crijnen J (2019) Predicting a pilot's cognitive state from physiological measurements. Master thesis, Tilburg University, Tilburg, The Netherlands, http://arno.uvt.nl/show.cgi?fid=149399.

Harrivel AR, Stephens CL, Milletich RJ, Heinich CM, Last MC, Napoli NJ, AbrahamN, Prinzel LJ, Motter MA, Pope AT (2017) Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing. AIAA Information Systems-AIAA Infotech@Aerospace, 1135 (AIAA.org).

Kaggle (2019) Reducing commercial aviation fatalities. Booz Allen Hamilton, Accessed April 22, 2020, https://www.kaggle.com/c/reducing-commercial-aviation-fatalities.

Mishra A, Shrivastava KK, Anto AB, Quadir NA (2019) Reducing commercial aviation fatalities using support vector machines. 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), 360–364 (IEEE).

Moroze ML, Snow MP (1999) Causes and remedies of controlled flight into terrain in military and civil aviation. Technical report, Air Force Research Lab Wright-Patterson AFB OH, Human Effectiveness Directorate.

Rosenkrans W (2015) Airplane state awareness. Flight Safety Foundation, https://flightsafety.org/asw-article/airplane-state-awareness/.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

First of all, I want to express my gratitude to Professor Yoshida and Professor Lucas, who served as my thesis advisors. Their keen eyes and knowledgeable observations have made a huge difference in the consistency of this research.

Also, I am thankful to the entire Naval Postgraduate School staff, including my teachers, program officer, academic associate, classmates, and all NPS personnel. I am so fortunate to have had the chance to complete my master's degree in such a special environment.

Finally, and most notably, I am indebted to the Armed Forces of the Republic of China, Taiwan, for supporting my education. I hope that the partnership between the United States and my country keeps improving, and in the future, I will be able to contribute to my community by implementing what I have learned at NPS.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

1.1 The Origin of This Research

Two years ago, the data science community and subsidiary of Google, Kaggle, sponsored a competition named "Reducing Commercial Aviation Fatalities" (Kaggle 2019) that challenged data scientists and machine learning practitioners around the world to build a model capable of monitoring a pilot's mental state in real time through psychophysiological data, such as electroencephalogram recordings, electrocardiogram signal, respiration, and galvanic skin response, of the aviator. According to Kaggle (2019), many lives will be saved in the future if pilots can be alerted when they fall into an abnormal mental state while operating the aircraft.

Due to the Kaggle competition rules (Kaggle 2019), the solution would not be available even when the competition was over. Yet, such research is valuable for both commercial and military aviation because it is still not possible to eliminate aviation fatalities entirely due to human factors as long as there is a need for human pilots onboard. Thus, I decided to use this topic as my thesis research in order to contribute to aviation safety.

1.2 Cognitive States

Before starting this research, two questions needed to be answered. The first problem is to analyze cognitive states of pilots, such as channelized attention, diverted attention, and startle/surprise for the purpose of reducing aviation fatalities. In military aviation, channelized attention is proven to be one of the most important factors attributed to insufficient situational awareness (SA) of pilots (Moroze and Snow 1999). Moroze and Snow (1999) also mention that some experts in this field have observed that the absence of SA of a flight crew is the most widely recognized factor that leads to aviation mishaps.

In addition, the research to comprehend a flight crew's psychological states, as mentioned previously, has been suggested as a safety enhancement (SE) to remedy pilots' deficiency of airplane state awareness (ASA) because the Commercial Aviation Safety Team (CAST)

discovered that 13 out of 18 commercial flight fatal events from 2003 to 2012 were attributed to the loss of control in-flight caused by flight crew loss of SA, and mental diversion was involved in every one of these instances (Rosenkrans 2015). Therefore, analyzing the cognitive states of pilots is imperative in order to enable pilots to overcome distractions and to ensure flight safety.

The second question is whether aviators' cognitive states can be classified and predicted by simultaneously measuring their psychophysiological signals. Recent research by National Aeronautics and Space Administration (NASA) Langley Research Center (Harrivel et al. 2017) used pre-processed electroencephalogram (EEG), galvanic skin response (GSR), electrocardiogram (ECG), and respiration (R) signals from 13 participants as input features to predict seven different pilot cognitive states (diverted attention, channelized attention, low workload, high workload, confirmation bias, startle/surprise, and rest). NASA's study provided the overall best accuracy—area under the curve (AUC) 0.95 with EEG, R, and GSR, and AUC 0.93 with ECG and GSR. The outcome of NASA's research shows that correctly identifying pilots' mental states is feasible and looks promising through the analysis of psychophysiological signals (EEG, ECG, GSR, and R).

1.3 Research Objectives

The dataset analyzed in our thesis is from the Kaggle competition (Kaggle 2019), which contains the same four modalities (EEG, ECG, GSR and R) as in NASA's research, but is from 18 participants. Furthermore, instead of seven mental states, four states (baseline, channelized attention, diverted attention, and startle/surprise) are included. As we know from the aforementioned information, those dangerous cognitive states contribute to flight accidents. As a result, our objective is to build a model to detect pilots' mental states in dichotomy (non-detriment and danger) for the purpose of immediately informing pilots when they fall into a hazardous mental state so that they will have more time to regain situational awareness.

In addition, we are trying to predict the patterns in terms of time pilots would have before losing "airplane state awareness." In this way, we can further generalize the time series patterns in which pilots experience distractions, which would induce one of the three cognitive states in an attempt to provide an effective early warning to the flight crew to

execute any proper measures for flight safety.

1.4 Thesis Structure

Chapter 2 reviews previous works related to our research analyzing similar datasets. Chapter 3 describes the methodology and workflow of this research, including information of the data, time series, and classification methods. Chapter 4 demonstrates our results from the overall analysis. Chapter 5 provides the research conclusion and recommendations for further improvements.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:
Background

In this chapter, we review previous works related to our research, focusing especially on one article (Mishra et al. 2019) published in the Second International Conference on Smart Systems and Inventive Technology (ICSSIT 2019), and one master's degree thesis (Crijnen 2019) from the Cognitive Science and Artificial Intelligence Department of Tilburg University in the Netherlands. Both applied different approaches to analyze datasets from the Kaggle Competition.

2.1 Literature Review: Reducing Commercial Aviation Fatalities Using Support Vector Machines

Mishra and his team published an article in the Second International Conference on Smart Systems and Inventive Technology titled "Reducing Commercial Aviation Fatalities Using Support Vector Machines" (Mishra et al. 2019). In this research, they claim that if there is no effective method to decrease the pilots' inaccurate decision-making due to distraction, the flight accident frequency will be positively correlated to the growing number of activities in the sky. This is because there are plenty of factors contributing to flight crews' intense workload and that influences pilots' concentration. Such factors include departure and approach operations, dealing with heavy traffic in aerospace, and poor meteorological conditions. Similarly, the Safety Enhancement recommendation made by CAST, "Airplane State Awareness – Training for Attention Management," places an emphasis on the limitations of human performance during flight (Rosenkrans 2015).

In an effort to reduce aviation accidents, Mishra and his team managed to build a model capable of detecting 400 pilots' cognitive states by their physiological data, using a Support Vector Machine (SVM) as a classifier. They also found a non-linear separable quality in the data by applying data visualization, so the Gaussian Kernel function was implemented in their model optimization. Moreover, using EEG, ECG, GSR, and R signal values as inputs, they created a user interface webpage for identifying mental states of humans.

While reviewing this research paper, we noticed one description inconsistent with the data

source Mishra and his team cite (Kaggle 2019). In the introduction section, they state that the physiological data from 18 participants is provided for algorithm training purposes. However, their data for training the model is a sample dataset of 400 pilots, which is, in fact incompatible with the dataset from the Kaggle competition. Besides, their final result of 95.5% accuracy is generated from 382 rows of data out of 400 rows. We assume they are using a much bigger database that consisted of 400 pilots for training the algorithm and randomly sampled one row of data from each pilot to compose a test dataset of 400 pilots. If so, the data generation must be a huge process because the scale of the experiment is substantial in contrast with what NASA performed with 24 participants.

Moreover, Mishra and his team suggest that machine learning models can be applied to understand pilots' current cognitive condition prior to each flight as a means of preventing flight accidents. As a matter of fact, channelized attention, diverted attention, and startle/surprise are the expected reactions with respect to different scenarios. We also know that the situations change rapidly during flight because of the velocity of movement. Knowing the captain is currently in a state of channelized attention before boarding does not mean he or she will be distracted during the mission. A more realistic approach would detect flight crews' mental states in the moment of operating the aircraft and providing an instant warning signal to alert pilots when they might be entering a dangerous state so that pilots would be able to fix the problem before actually neglecting necessary procedures of operation as a result of interruption.

In addition, given the fact that Mishra and his team do not present their final model in this article, other researchers will find it difficult to continue with this research. Therefore, we provide our model as an initial point for future researchers who are interested in this field, so they can continue refining the algorithm in an attempt to better predict pilots' cognitive states using machine learning methods or any newly developed computation technique.

2.2 Literature Review: Predicting a Pilot's Cognitive State from Physiological Measurements

A thesis, titled "Predicting a Pilot's Cognitive State from Physiological Measurements" by J.A. Crijnen in 2019 analyzes the same dataset from the Kaggle competition "Reducing Commercial Aviation Fatalities" (Crijnen 2019). In that thesis, the author asserts that

although machine intelligence can do the majority of the work for pilots during flight, there are still some critical decisions that need to be made by the flight crew themselves. Thus, this thesis mainly focuses on developing a better strategy to make pilots concentrate when it is necessary to resolve problems on their own in the cockpit.

The thesis research conducted by Crijnen (2019) begins by investigating the feasibility of using physiological data to forecast the mental states of a pilot and distinguished the engineering predictors attributing to the accuracy of the model. Crijnen (2019) also mentions that pilots perceiving themselves in a dangerous mental state do not necessarily remove the threat because their decision to recover from an abnormal state relies on the personality traits of the individual pilot, such as self-satisfaction and over-optimism. For example, the captain may choose to ignore the warning if he/she is too confident in his/her skills to follow standard procedures. Hence, the accurate prediction of cognitive states and an unignorable alarm which pilots must react to will make the decision-making process in the cockpit safer.

Due to the policy of the Kaggle competition, researchers cannot acquire the complete test dataset. So, we used the training dataset and separated it into training, validation, and test sub-datasets. One previous research (Marcel and Millan 2007) regarding EEGs states that the pattern of brainwaves is unique to each person so that the EEG can be a means for biometric recognition. Another article (Saechia et al. 2005) also mentions that every human heart is different. For this reason, using EEG data as inputs to characterize individuals is viable. Under those circumstances, it might not be a great idea to strictly separate training, validation, and test datasets by crew number. Even though the brainwaves can be generalized by similarities to every person while receiving the same stimulation (Jahangir and Pirouz 2020), there is still uniqueness in each individual's bio-electricity. Therefore, it is possible to lose some important correlations without considering individual cases.

Instead of using only the original data measured from the sensor, Crijnen (2019) also uses psychophysiological and statistical methods to extract features from each modality as additional inputs. In Crijnen's thesis research, such methods as sliding window, frequency domain analysis, and longitudinal bipolar method are used. Even if it is possible to obtain a deeper understanding from those descriptive statistical features, the increase in model complexity (Castrounis 2021) might cause other problems, including overfitting and losing fidelity of the data itself.

Thus, in Crijnen's research, two major tasks are conducted. The first one is the classification of cognitive states. In this task, a Gradient Boosting Machine (GBM) classifier is implemented because of its flexibility and ability to adapt to outliers and unbalanced datasets. Unlike NASA, which used AUC, Crijnen (2019) claims the balanced F-score, or more simply the F1 score, is a more appropriate measure of accuracy for unbalanced numbers of response variables. An F1 score of 0.55 is generated in this task, and overfitting is discovered because of the significant drop in F1 score when it is compared to training and validation results. The second task is to determine a switch in mental states. Four extracted features are used as inputs for training a logistic regression model due to their statistical significance. The AUC score of the result is 0.53, meaning the performance of this model is similar to binary random selection.

Based on the previously mentioned computational experiment, Crijnen (2019) provides the following conclusions.

- As to engineering features, GSR, ECG, and respiration data provide the most significant contributions, which is consistent with our own experimental result.
- Even though the performance of cognitive states classification does not outperform NASA's outcome, the accuracy is similar for every participant in the test dataset.
- The AUC value equals 0.53, meaning the cognitive states change and thus can distinguish whether the change in mental states fails to be detected when extracted features from the raw data are used as inputs.
- A model's ability to classify cognitive states differs according to the individual as evidenced by the results that show different performances among different pilots. In other words, physiological qualities rely upon the human. Responses to the distinctive psychological states differ among individuals as well.

Based on the findings just described, it would be a better strategy to train the data using all participants' signals simultaneously in order to gain sufficient information by taking human factor relationships into consideration. The main purpose of our model is to alert pilots when they are in a dangerous state intellectually, so the model is designed to determine whether a pilot is either focusing on flying or preoccupied by any external stimulus.

CHAPTER 3:
Methodology and Modeling

3.1 Methodology Overview

Figure 3.1 shows the workflow of our thesis research. This workflow is described in the main steps that follow.

- **Raw Data**: Two datasets were downloaded from the Kaggle competition "Reducing Commercial Aviation Fatalities" webpage (Kaggle 2019): the training dataset and the test dataset. Due to the Kaggle competition policy, we were not able to acquire the complete test set. Therefore, we decided to use only the training dataset and split the data into three parts for training, validation, and testing purposes.
- **Exploring Data Analysis**: Data exploration is a preliminary and necessary step in data analysis (Shelby 2018). It allows us to understand the basic structure and characteristics of the data, such as distributions, correlations, types, and number of variables, before it is processed. In this step, we chose data visualization to help us gain deeper insight.
- **Data Processing**: After having the basic picture of what the data looks like, we needed to adjust the data to the appropriate form in order to initiate future modeling. In our case, we conducted time series analysis because all the records are based on a time sequence. In addition, our objective is to alert pilots when they fall into adverse mental states. To this end, we further converted the multi-class responses into binary variables.
- **Modeling**: In our time series analysis, we found the predictability of electrocardiogram (ECG), galvanic skin response (GSR), and respiration (R) data. Thus, we applied those predictions to the classification model intending to check the feasibility of predicting future mental states through current psychophysiological data. Also, we managed to train a model to distinguish current intellectual states by using all the predictors except time.
- **Performance Evaluation**: This was a crucial and recursive step for verifying our model's effectiveness. If the result was not effective, then we needed to choose

another method of classification. The accuracy from the confusion matrix, AUC, and F1 score were adopted in our experiment as the criteria for evaluating performance.

Each of the steps depicted in Figure 3.1 is expanded upon in the following sections.

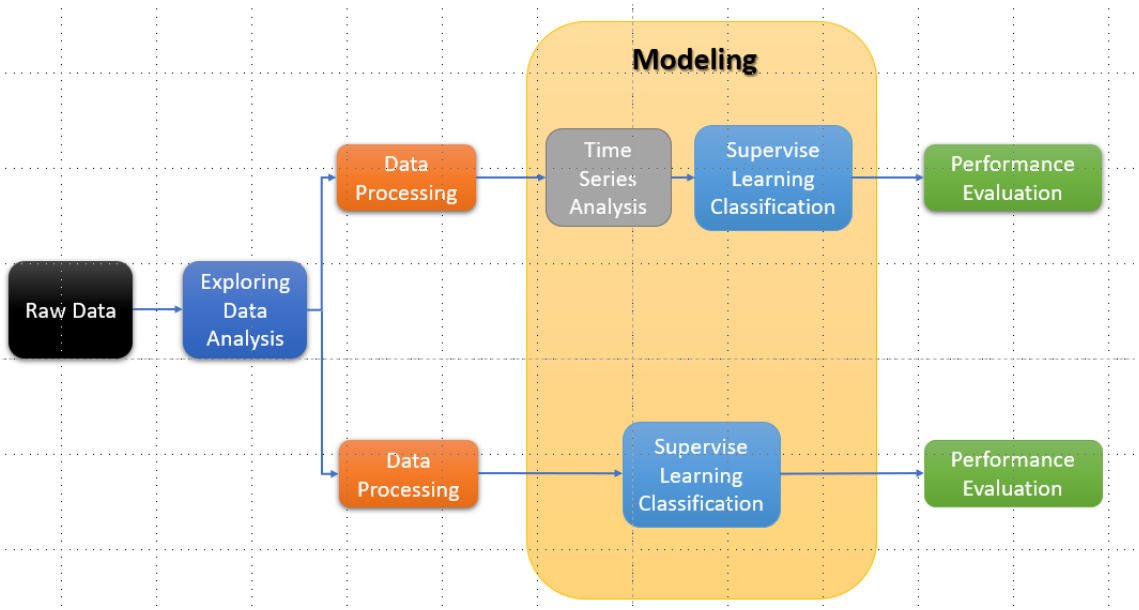

Figure 3.1. Workflow of this research process.

3.2 Data Description

We used actual psychophysiological evidence from 18 pilots in this dataset who were exposed to multiple disruptive activities. Outside of a flight simulator, the benchmark training set comprised a series of managed tests performed in a non-flight scenario. The pilots underwent distractions meant to induce one of the following three cognitive conditions (Kaggle 2019):

- **Channelized Attention (CA)** is defined as the condition that occurs when an individual is too concentrated on the mission to notice other important information. The bench-marking experiments were designed by making the participants play an enjoyable video game that included a puzzle.
- **Diverted Attention (DA)** is defined as the condition of an individual's mind being distracted by decision-related behavior or thinking processes. This state was induced

through a screen-tracking activity done by the participants. Regularly, before switching to the tracking mission, a math question came up that had to be answered.

- **Startle/Surprise (SS)** is defined as a response triggered by making the participants watch videos that included pop-up scares.

A set of two pilots, each assigned a unique identification, were tracked over time in each trial and exposed to the experiment, which was designed to induce the psychological states of channelized attention (CA), diverted attention (DA), or startle / surprise (SS). The dataset includes three experiments (one for each state) where only one of the mental states was triggered in the pilots. In other words, the pilots would be in either a safe mental state or a dangerous mental state in each experiment. For every timestamp in the dataset, the aim is to predict the pilots' actual response from each experiment.

Each sensor worked at a sample rate of 256 Hz. As this is physiological data from actual humans, the data may include noise and untruthful values.

3.2.1 Data Variables

The data variables for this dataset included the following:

- **crew**: A special identity for a pair of pilots. A total of nine crews are present in this dataset.
- **experiment**: One of CA, DA, and SS. The training dataset includes these three studies.
- **time**: The duration in seconds of the evaluation.
- **seat**: A pilot in the seat on the left is (0) or right is (1).
- **eeg** (prefix): A total of 20 different electroencephalogram signals (Figure 3.2). Blocka (2018) recommends that an EEG examination is used to look at the brain's electrical function. Brain cells interact through electric signals. An EEG scan, which captures brainwave forms, may help recognize issues connected with certain operations. As shown in Figure 3.2, 10 to 20 tiny metal disks are connected to the scalp with cables. The electrodes measure the electrical activities of the brain and transmit messages to the device. The resulting traces behave like wavy lines of ups and downs, and these lines help physicians diagnose irregular trends easily. An irregularity can be triggered by a neurological disease.

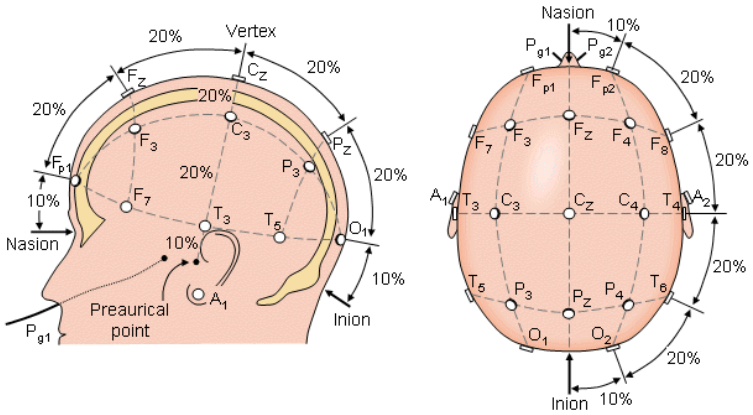

Figure 3.2. In order to document the unpremeditated EEG, the globally accepted 10-20 system is usually used. Source: Malmivuo et al. (1995).

- **ecg**: Three-point Signal electrocardiogram (Figure 3.3). A resolution/bit of 0.012215 μV and a scale of –100 mV to +100 mV were present in the detector. The data was gathered in micro-volts (mV). It is one of the easiest and quickest methods used to measure the heart. The electrodes are attached to the ECG machine by leads. Electromagnetic waves are generated, evaluated, and printed out. No energy goes through the body using this method. Instead, normal electrical signals regulate the various areas of the heart to maintain blood pumping. An ECG tracks these impulses to demonstrate how quickly the heart is pounding, how it is pumping (steadily or irregularly), and how quickly and when it is beating (normal or fast). Differences in an ECG may signify several heart-related problems (Saechia et al. 2005).

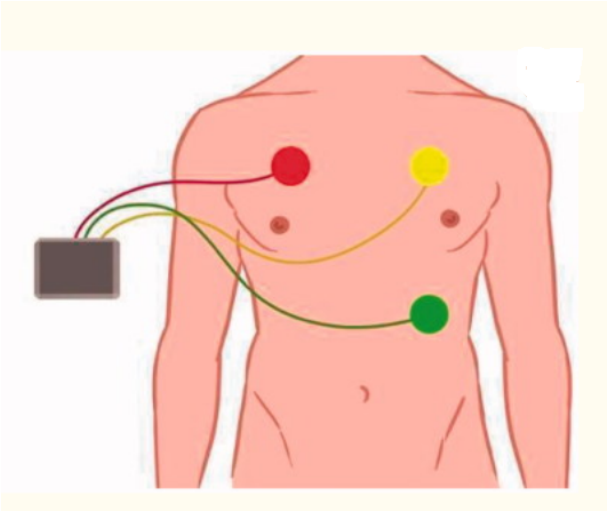

Figure 3.3. Portable three-lead ECG monitor. Source: Kristensen et al. (2016).

- **r**: An indicator of the rise and fall of the chest. The sensor had a 0.2384186 μV resolution/bit and a spectrum of –2.0 V and +2.0 V. The data is described in microvolts.
- **gsr**: An indicator of electrodermal function (Figure 3.4). The sensor had a 0.2384186 μV resolution/bit and a spectrum of –2.0 V and +2.0 V. The data is described in microvolts. The galvanic skin response (GSR), also recognized as skin conductance (SC), relates to shifts in the behavior of the sweat gland, which indicates the magnitude of the emotional condition or emotional arousal of the participants (Shi et al. 2007).

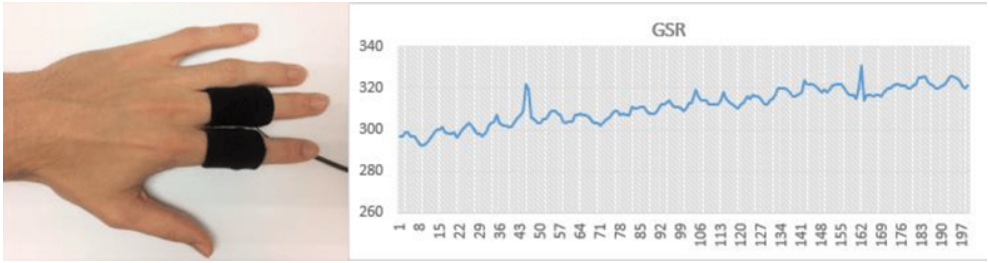

Figure 3.4. Galvanic skin response sensor. Source: Myroniv et al. (2017).

- **event**: The pilot's condition at a specified time: one of A = baseline, B = SS, C = CA, D = DA.

3.3 Time Series Analysis

Based on the result of data exploration for each numerical variable (EEG, ECG, GSR, and R) with respect to time, we can observe that there are potential patterns in this dataset, as can be seen from Figures 3.5 to 3.7. In an effort to further discover their relationships and forecasting power regarding time, we conducted a time series analysis.

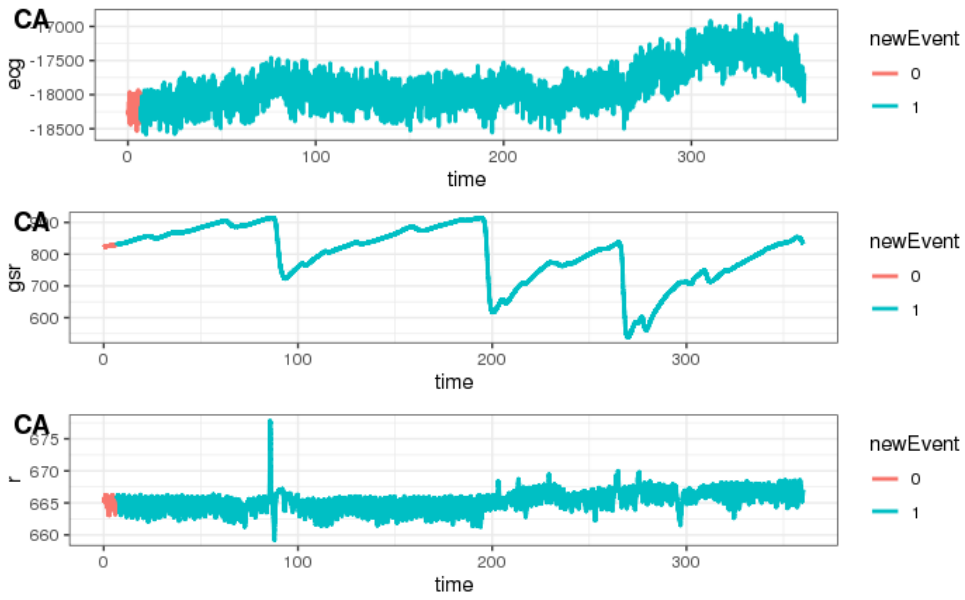

Figure 3.5. The variation of ECG, GSR, and respiration recordings during a channelized attention (CA) experiment; safe (red), dangerous (blue-green).

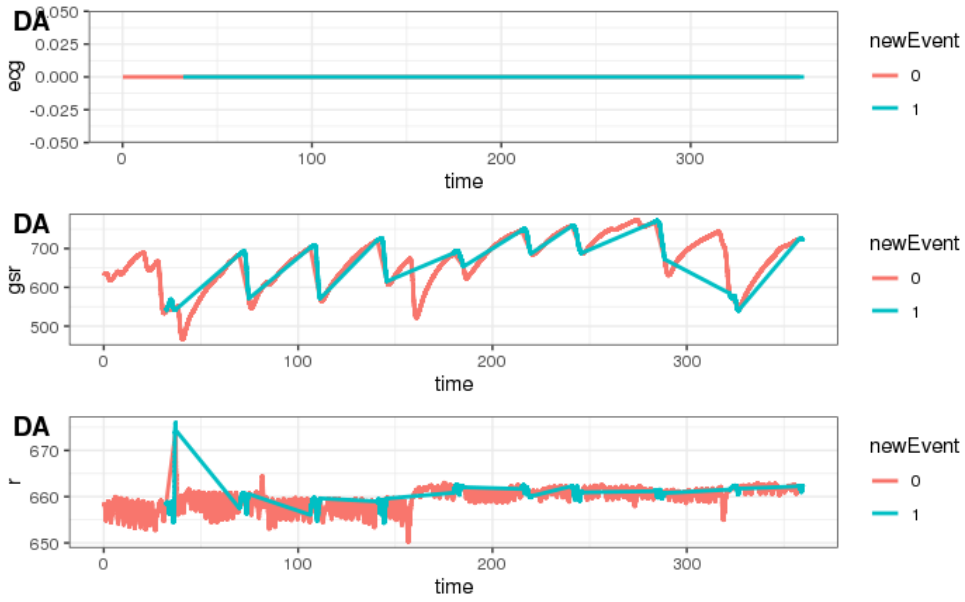

Figure 3.6. The variation of ECG, GSR, and respiration recordings during a diverted attention (DA) experiment; safe (red), dangerous (blue-green).

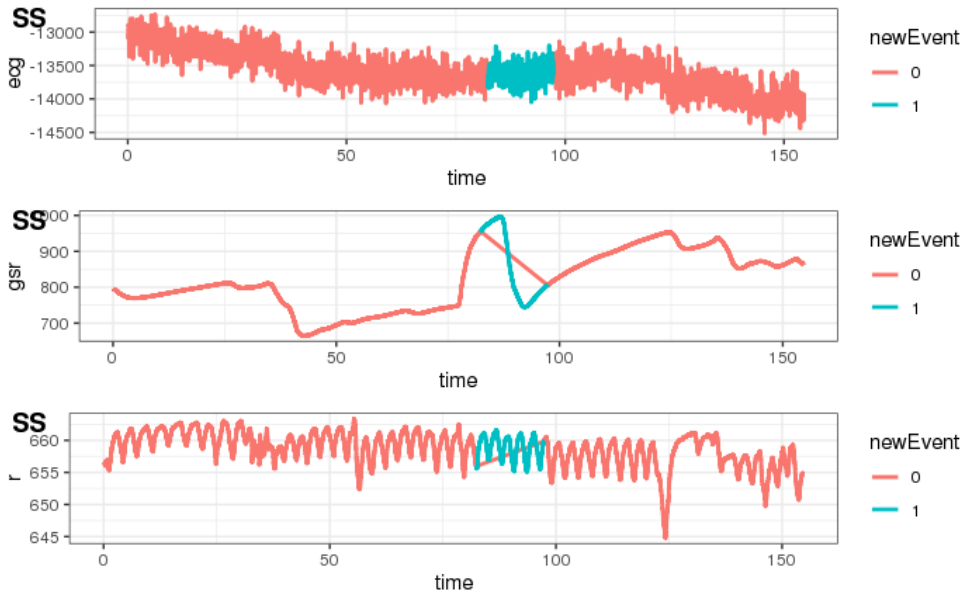

Figure 3.7. The variation of ECG, GSR, and respiration recordings during a startle / surprise (SS) experiment; safe (red), dangerous (blue-green).

According to Brockwell and Davis (2016), a time series is a dataset indexed by time-

ordered data points. Analysis of a time series involves different models to evaluate time series data with a view to derive useful information and other data characteristics. Time series forecasting inputs the observed historical data into a mathematical model to estimate the possible values during a specific period in the future.

Assuming that those psychophysiological signals are predictable through time, we can combine the predictions from the time series model with the classification methods to further forecast the mental states of pilots in an attempt to prevent a loss of SA of the flight crews in advance.

3.3.1 Data Processing

Before starting the time series modeling, we needed to restructure the data. The original training set contains more than 4.8 million rows. It occupies 1.15 gigabytes of computer memory without starting programming for analysis. Due to computational limitations, we condensed the dataset to one row per 0.1 second (Figure 3.8).

	time	ecg	gsr	r	newEvent2
1	0.1054688	-12883.3	794.674	656.464	0
2	0.1093750	-12883.3	794.674	656.464	0
3	0.1132812	-12884.8	794.499	656.434	0
4	0.1171875	-12884.8	794.499	656.434	0
5	0.1210938	-12884.8	794.499	656.434	0
6	0.1250000	-12962.7	794.456	656.422	0
7	0.1289062	-12962.7	794.456	656.422	0
8	0.1328125	-12962.7	794.456	656.422	0
9	0.1367188	-13049.4	794.286	656.411	0
10	0.1406250	-13049.4	794.286	656.411	0

	Series.Times	mecg	mgsr	mr	newEvent2
1	0.1	-12958.89	794.4464	656.4283	0
2	0.2	-10446.64	753.0771	659.4441	0
3	0.3	-10488.00	751.2092	660.0056	0
4	0.4	-10454.71	751.0686	659.9968	0
5	0.5	-10407.35	750.9743	660.2607	0
6	0.6	-10416.28	750.6458	660.0969	0
7	0.7	-10490.87	750.1897	659.9962	0
8	0.8	-10446.90	749.7234	659.8727	0
9	0.9	-10387.81	749.2448	659.5668	0
10	1.0	-10431.88	749.0862	659.2674	0

Figure 3.8. Time Series Data Restructure; the upper table shows the data before restructuring by averaging the column values within each 0.1 second; the lower table shows the restructured data after averaging the column values within each 0.1 second.

We also used the mean value of each numerical variable according to the new time intervals and focused on the data with a single pilot (crew = 1 and seat = 0) who was the captain of Crew 1. Because some psychophysiological data are unique to each person, such as EEG (Marcel and Millan 2007) and ECG data (Saechia et al. 2005), we considered it a reasonable approach to sample the data from a specific participant as part of our inputs for time series

analysis.

In addition, we converted the four states (CA, DA, SS, and Baseline) into binary responses (baseline and abnormal) in accordance with our objective, which is to distinguish whether a pilot falls into a dangerous mental state.

3.3.2 Time Series Models

In this section, we describe the time series models in order of complexity, from least to most complex.

Naïve

Using the Naïve model, we set all estimates to be just the value of the last observation for Naïve forecasts (Yoshida 2020).

This is expressed as

$$\hat{Y}_{t+1} = Y_t. \tag{3.1}$$

Hyndman and Athanasopoulos (2018) claim that, for several estimations relating to financial time series, this approach works surprisingly well. Since a Naïve forecast is ideal as the data matches a random walk (current value as a next-period forecast), it is also referred to as a random walk forecast.

The prediction for time $t + 1$ can be written as

$$\hat{Y}_{t+1} = Y_{t-k}, \tag{3.2}$$

where k is the seasonal lag (Yoshida 2020).

Seasonal Decomposition (STL)

Data from a time series may exhibit a few patterns, and splitting a time series into different element components is always useful for further analysis (Hyndman and Athanasopoulos 2018).

The main time series components are the following (Yoshida 2020):

- **trend**: When there is a long-term rise or decrease in the dataset, we identify it as a *trend*. A trend can be nonlinear.
- **seasonal (periodic) patterns**: When seasonal factors (e.g., a specific date of the year) can influence a time series, we call it a *Seasonal Pattern*. It always appears periodically; for instance, every 60 days or every week.
- **cycle**: When data shows fluctuations in an irregular manner, we call it a *cycle*. A similar example is the volatile nature of the bull and bear stock markets.
- **noise**: This is the remaining, unidentified variance from the previously described data components.

STL is a flexible and stable tool for time series decomposition. It is an acronym for "Seasonal and Trend decomposition using Loess," and loess refers to a nonlinear and nonparametric relationship estimation procedure. STL is also stable for outliers, and hence, predictions of the trend, cycle, and seasonal elements are not influenced by occasional irregular observations (Hyndman and Athanasopoulos 2018).

In our analysis, we used an **stl** function in the R programming language with seasonal window (**s.window**) to decompose a time series dataset to observe the seasonal patterns, trend, and irregular components. By analyzing the time series decomposition, we may further forecast future psychophysiological values. According to Hyndman and Athanasopoulos (2018), seasonal window (**s.window**) is the number of consecutive time stamps to be included in the seasonal component to approximate each value.

Exponential Smoothing (ETS)

Exponential Smoothing is an obvious extension of the moving average method (Yoshida 2020). Hyndman and Athanasopoulos (2018) mention that the weighted means of past observations are the forecasts generated by exponential smoothing, with the weights decreasing exponentially as the observations become older. Specifically, the more recent the observation is, the larger the weight is. This structure provides efficient and precise predictions over a broad variety of time series. Equation 3.3 shows the basic idea of weighted average, where $0 \leq \alpha \leq 1$ is the smoothing parameter.

$$\hat{Y}_{T+1|T} = \alpha Y_T + \alpha(1-\alpha)Y_{T-1} + \alpha(1-\alpha)^2 Y_{T-2} + \cdots . \tag{3.3}$$

The function **ets** in the R programming language was used in this analysis. Through optimizing the likelihood function, this model not only measures both the initial states and smoothing parameters, but also scans over a limited parameter space to ensure the final model has predictive power (Hyndman et al. 2008).

Auto-Regressive Integrated Moving Average (ARIMA)

Another method of time series forecasting is provided by ARIMA models. The two most commonly employed techniques to time series forecasting that have complementary approaches to the issue are the exponential smoothing and ARIMA models. Although exponential smoothing models focus on the concepts of trend and seasonality, the auto-correlations in the data are represented by ARIMA models (Hyndman and Athanasopoulos 2018).

An ARIMA model is composed of the following three components (Yoshida 2020):

- **Auto-Regressive Component**: For the next time periods, the AR(p) component references the prior time periods as predictors.
- **Moving Average Component**: Using an error regression technique, the MA(q) component minimizes the residual errors.
- **Integrated Component**:
  - The first order difference is denoted by AR(1), i.e., $Y_t - Y_{t-1}$.
  - Generally, only the variations between the first and second order are considered.
  - In an effort to make the time series stationary, this component is used to eliminate trends.

The following three numbers could well summarize a non-seasonal ARIMA model (Yoshida 2020):

- p: The number of auto regressive terms.
- d: The number of non-seasonal differences.
- q: The number of moving-average terms.

This is called an ARIMA(p,d,q) model.

The **auto.arima** function in the forecast package of R was used in this thesis. It enables the possibility to automatically optimize the ARIMA(p,d,q) model that provides the best fit over the training period, including engaging with the seasonal effects (Yoshida 2020).

Ensemble

The Ensemble model is a mixture of various models in the time series (Yoshida 2020). In our analysis, we constructed an Ensemble model by averaging the values predicted by the following models:

- Seasonal Decomposition (STL);
- Exponential Smoothing (ETS);
- Auto-Regressive Integrated Moving Average (ARIMA).

3.3.3 Performance Evaluation

Hyndman and Athanasopoulos (2018) maintain that when using genuine forecasts, it is important to determine forecast accuracy. Therefore, the scale of the residuals is not a good indicator of how large the true forecast errors would be. It is only important to assess the precision of predictions when considering how well a model does on additional data that has not been utilized when fitting the model.

It is a standard practice to divide the dataset into two groups, training and test data, while selecting models, where the training data is used to approximate the parameters of a forecast model and the test data is used to measure its accuracy. Since the test data is not used to assess the predictions, a reliable indicator should be given about how accurately the model is able to predict new data (Hyndman and Athanasopoulos 2018).

The subsequent sections address the process of measuring predictive performance and precision with the rolling horizon method. In addition, we further explain the accuracy measurements used to calculate the performance of the models that fit our time series dataset.

Time Series Cross-Validation / Rolling Horizon

Hyndman and Athanasopoulos (2018) also describe that time series cross-validation is an

advanced usage of training/test sets to verify the accuracy of the model (Figure 3.9). There are a number of test sets in this method, each one consisting of a single observation. The split training dataset only comprises the observations that happened before the test data was assessed. Thus, in building the prediction, no future observations should be included. As an accurate prediction dependent on a small training set cannot be achieved, the earliest observations are not treated as testing data.

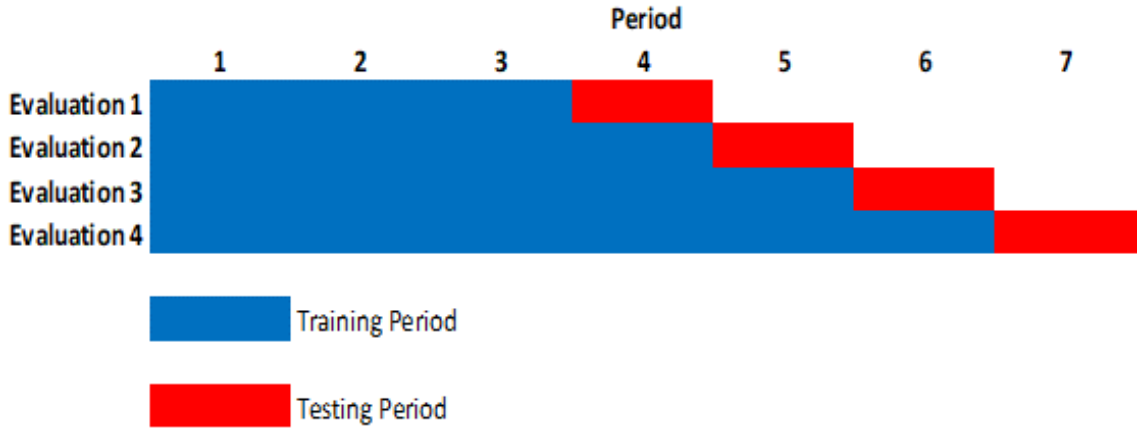

Figure 3.9. Rolling Horizons for time series model performance evaluation. We evaluate performance by repeatedly fitting the forecasting models to "rolling periods" and then measuring the performance in forecasting over the horizon that will be used in practice. This approach mimics the train/test dataset method used throughout the machine learning process. Source: Yoshida (2020).

We began to fit our models in our research to the first ten timestamps (per 0.1 second) or one second. We then extended the training set iteratively by an additional second until it surpassed 50 seconds, refitted the algorithm, and reassessed the accuracy of the model on the test set.

Accuracy Measurements

Yoshida (2020) asserts that there are multiple metrics that can be applied to assess the performance of time series models (including all the standard statistical approaches, such as Mean Squared Error (MSE)). Those tests, however, are not always scale-free (a desired property).

There are two methods that are now becoming effective forecasting assessment criteria:

- The Mean Absolute Percentage Error (MAPE) calculates the difference in prediction error and divides it by the real observation value,

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i - F_i|}{Y_i}. \tag{3.4}$$

- The Mean Absolute Scaled Error (MASE) measures the forecast error in contrast to the error of a Naïve model forecast (F),

$$MASE = \frac{\frac{1}{n} \sum_{i=1}^{N} |Y_i - F_i|}{\frac{1}{n} \sum_{i=1}^{N} |Y_i - Y_{1-10}|}. \tag{3.5}$$

The scale-free property of both MASE and MAPE implies that their values do not depend on the magnitude of the observations.

We defined two thresholds for our analysis that gave proof of the effectiveness and predictive power of a model:

- The 1-Step and 10-Step MAPE values are below 0.2. This suggests a difference between predicted and actual time series values of less than 20 percent.
- The 1-Step and 10-Step MASE values are less than 1. These could mean that in the time series, the model provided would forecast future observations more precisely than a simple Naïve model.

These thresholds allowed us to efficiently assess numerous time series models.

3.4 Classification Methods

Our research objective is to build a model to detect pilots' mental states (normal or dangerous) according to their real-time psychophysiological data (EEG, ECG, GSR, and respiration signals). To this end, we used classification methods to train our best model intending to become aware of pilots' mental states rapidly and accurately.

3.4.1 Data Processing

First, we looked at the distribution of the response variable. From Figures 3.10 to 3.13, we can easily notice the serious unbalance of observations with respect to the response variable. In order to make the data format consistent with our objective, we converted the response variable "event" into a binary one.

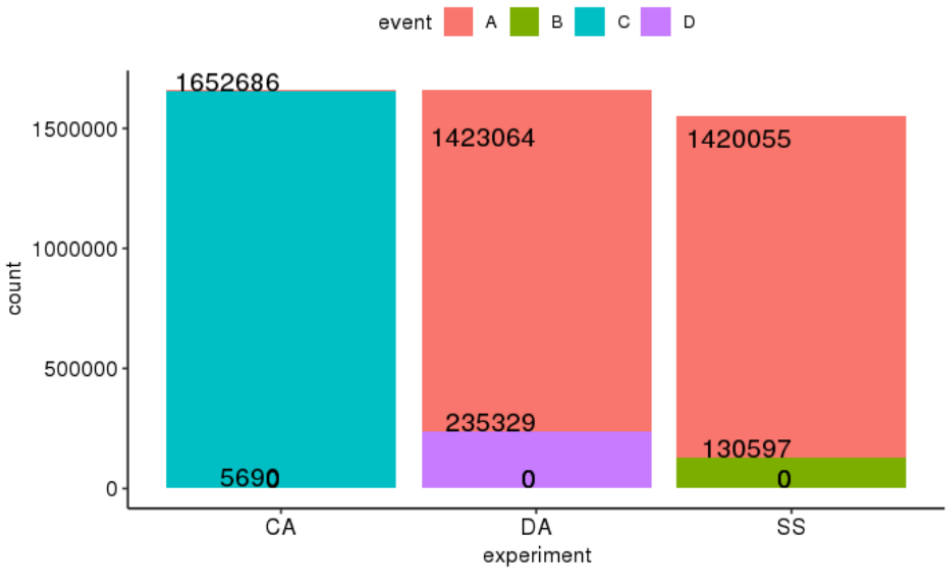

Figure 3.10. The distribution of events in each experiment before dichotomizing. A = baseline (red), B = SS (green), C = CA (blue-green), D = DA (purple).

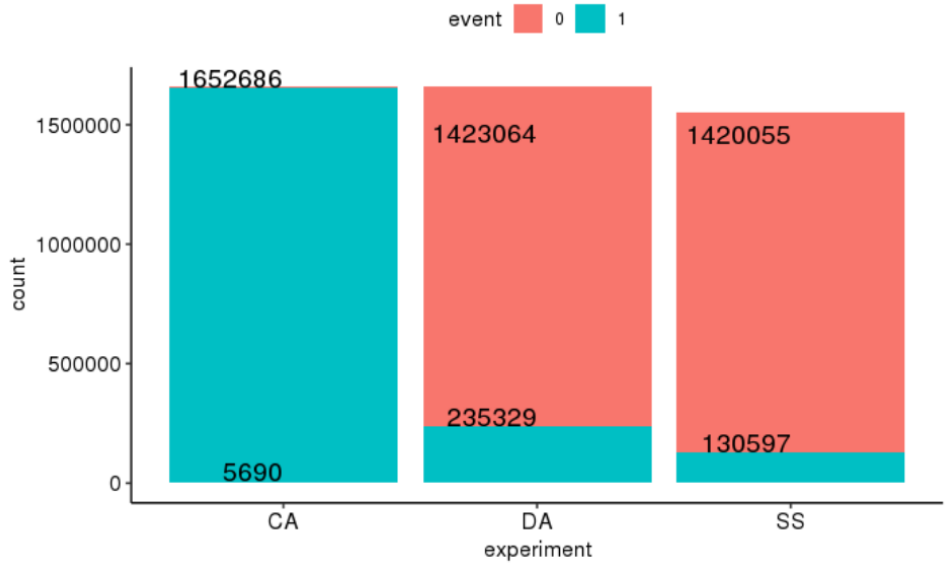

Figure 3.11. The distribution of the events with respect to each experiment after dichotomizing. 0 = safe (red), 1 = dangerous (blue-green).

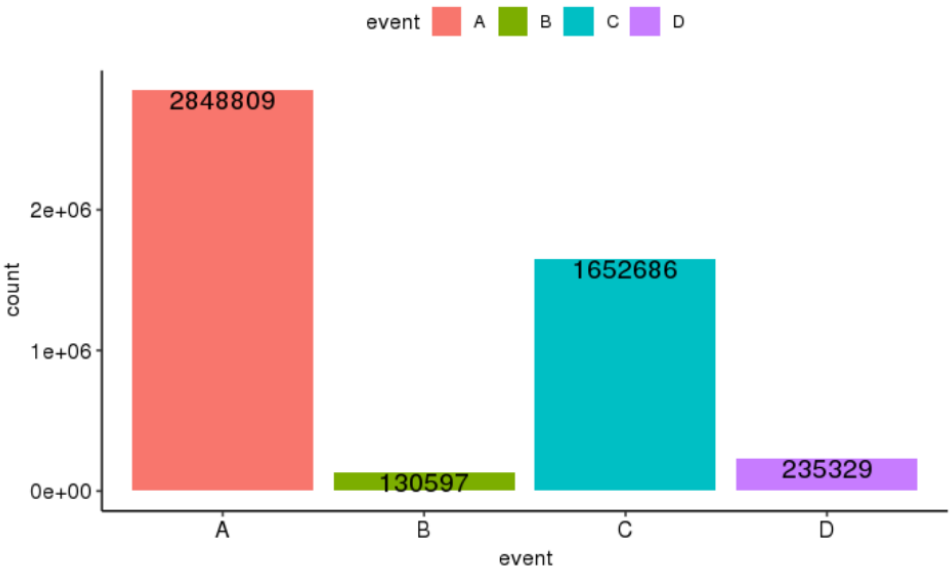

Figure 3.12. The distribution of events before dichotomizing. A = baseline (red), B = SS (green), C = CA (bluegreen), D = DA (purple).

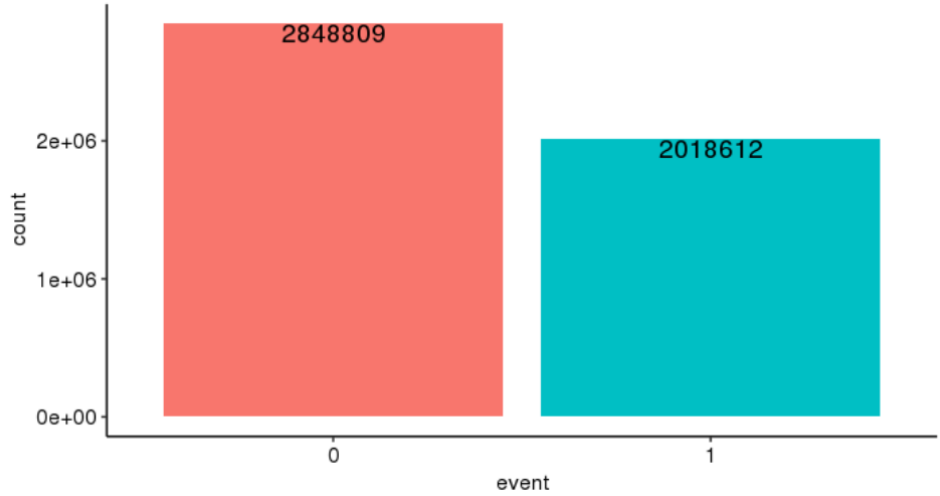

Figure 3.13. The distribution of events after dichotomizing. 0 = safe (red), 1 = dangerous (bluegreen).

Second, we dropped the column of "experiment," since it is obviously correlated to the "event." According to the data description (Kaggle 2019), every experiment would trigger only one response, so a participant would either stay at baseline or be distracted.

Third, we transformed "crew" and "seat" to be categorical because they represent a specific pilot.

Fourth, we randomly chose a sample size of 1/1000 with replacement from the original training dataset, so that we could have an independently and identically distributed (IID) dataset for training our algorithm (Stewart 2016) given that IID assumptions are required for most machine learning procedures (Nouretdinov et al. 2001).

Last, we divided the sample data into 80% (training) and 20% (validation) to calculate the initial performance. Then, we randomly sampled another dataset to be our test data which was not being used for training the model.

3.4.2 Classification Methods

The classification models implemented in this thesis are as follows.

Logistic Regression

Logistic regression is an algorithm used in machine learning for classification problems, such as estimating the probability that our dataset corresponds to one class over another (Yoshida 2020). For example, we are interested in forecasting the probability that a person, based on the amount of the balance owed, would default on repaying his or her credit card balance (Figure 3.14).

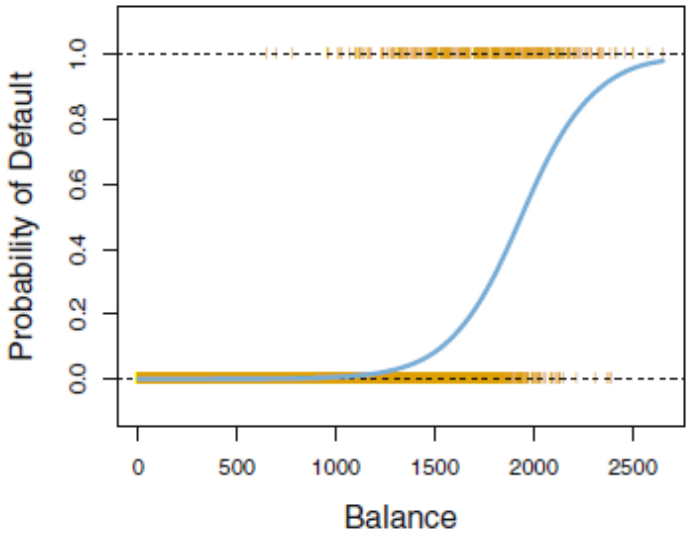

Figure 3.14. The predicted probabilities of whether a person will default on repaying their credit card given the amount of the balance owed, by logistic regression. All predicted values are between 0 and 1. Source: James et al. (2013).

We can think of the response as

$$P\ (\text{default} = \text{Yes} \mid \text{Balance}). \tag{3.6}$$

If the response variable is “Yes” or “No”,

let

$$f(X) = P(Y = 1|X), \tag{3.7}$$

and then the logistic regression model can be written as

$$f(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}. \tag{3.8}$$

where β_0 and β_1 are coefficients estimated given the data.
This is equivalent to

$$\frac{f(X)}{1 - f(X)} = \exp(\beta_0 + \beta_1 X). \tag{3.9}$$

If we take log at both sides of the equation, then we have

$$log(\frac{f(X)}{1 - f(X)}) = \beta_0 + \beta_1 X. \tag{3.10}$$

This is referred to as a logit function and it constructs a linear equation of X. In this thesis, we used the **glm** function in the R programming language for logistic regression (Yoshida 2020).

Support Vector Machine (SVM)

The aim of the SVM algorithm is to locate a hyperplane in a multi-dimensional space that correctly classifies the data points (Yoshida 2020).

Support vectors are data points located closest to the hyperplane that impact the hyperplane's direction and orientation (Figure 3.15). We optimize the margin of the classifier through these support vectors (Yoshida 2020).

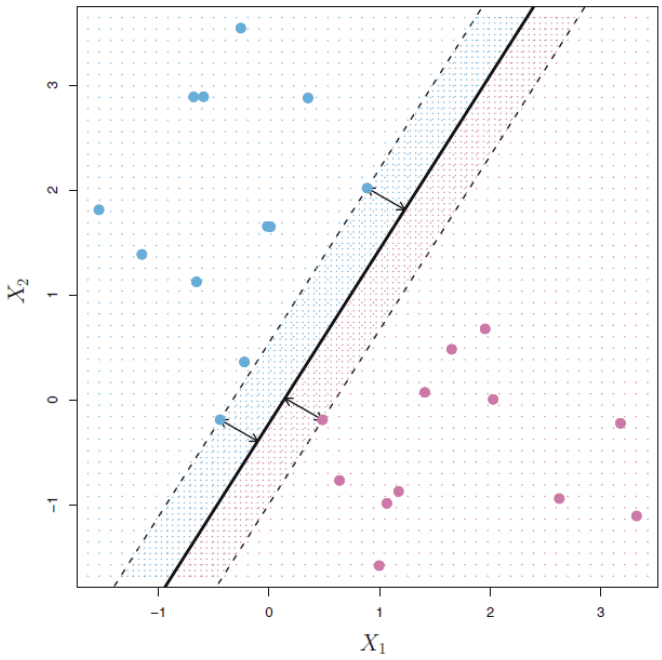

Figure 3.15. There are two types of observations that are either blue or purple. The hyperplane of the maximum margin is seen as a solid line. The gap between the solid line to one of the dotted lines is the margin. An example of the support vectors is displayed where the two blue points and the purple point intercept the dotted lines. The arrows signify the distance from those points to the hyperplane. The purple and blue grid displays a classifier's judgment centered on the separate hyperplane. Source: James et al. (2013).

In this analysis, we used the **svm** function in the R programming language from the e1071 package to fit the SVM model (Yoshida 2020).

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a method that preserves every accessible case and classifies new data or cases based on a measure of distance (Figure 3.16). It is often used to assign a data point to a certain category because of the similarity of its surroundings (Yoshida 2020).

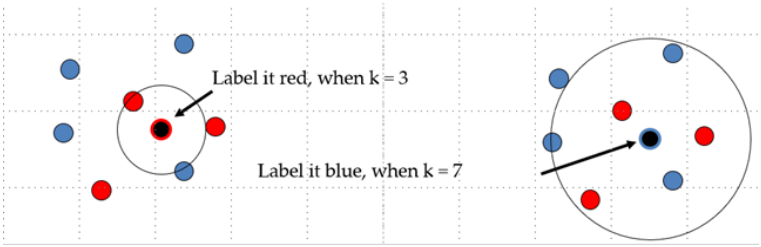

Figure 3.16. A new point is now given the most common mark of its K-Nearest Neighbors. Source: Yoshida (2020).

Similarity is described between two data points by a distance metric (Yoshida 2020).

Here, we used the **knn** function in the R programming language to fit the model for classification (Yoshida 2020).

Recursive Partitioning and Regression Trees (rpart)

Decision trees classify data points partitioning the data set as a whole (Figure 3.17). Decision trees are simple, straightforward, and user friendly, but not always precise (James et al. 2013).

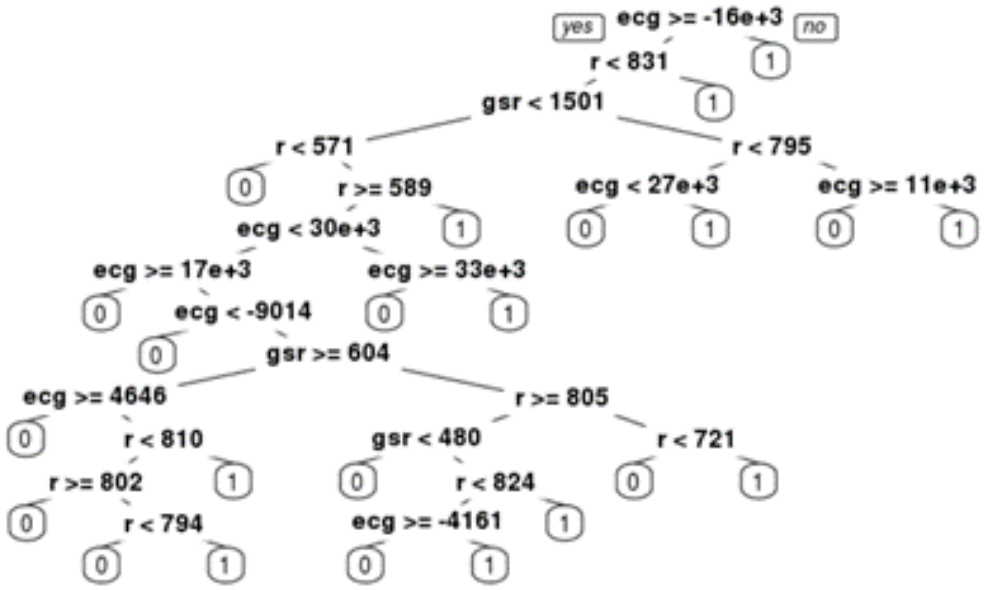

Figure 3.17. Decision tree of our training dataset by the **rpart** function.

The **rpart** algorithm in the R programming language, which is implemented in the functions in this thesis by recursively separating the dataset, implies that the subsets resulting from a split are further split before a predetermined termination criterion is achieved. The separation is focused on the independent variable at each point, resulting in the greatest possible reduction in the dependent variable's heterogeneity (Yoshida 2020).

Random Forests

Random Forests are an ensemble approach that involves averaging and bootstrapping. This approach uses multiple similar distributed trees to bootstrap training data samples. When trees are sufficiently developed, it is also possible to reduce bias. In addition, the Random Forests algorithm decreases the variation by averaging noisy unbiased trees. By minimizing similarity between trees using data bootstrapped for each tree and sampling accessible variable-sets at each node, it maximizes the effects of variance reduction (Yoshida 2020).

For several types of datasets, the Random Forests approach is one of the most robust machine learning methods. In addition, without variable deletion, it can manage thousands of inputs, and provides an analysis of what variables are significant in the classification process (Figure 3.18). More importantly, even when a substantial percentage of the data is missing, this method remains efficient at retaining high accuracy (Yoshida 2020).

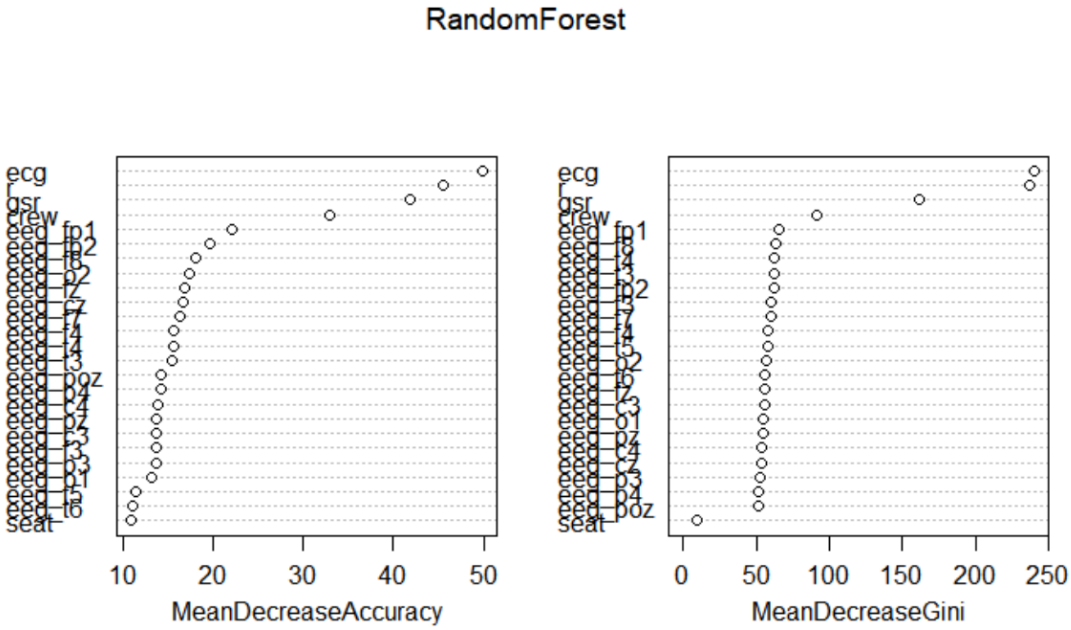

Figure 3.18. Feature Importance graphs from the Random Forests method; X-axis: Accuracy(Left); Gini(Right); Y-axis: The importance of predictors in descending order.

In this thesis, the **randomForest** function in the R programming language from the randomForest package was used to build the Random Forests model (Yoshida 2020).

3.4.3 Performance Evaluation

There are three evaluation criteria used in this analysis. We explain them in the following paragraphs. Figure 3.19 is a two-by-two confusion matrix, which is the most important tool for evaluating binary classification performance.

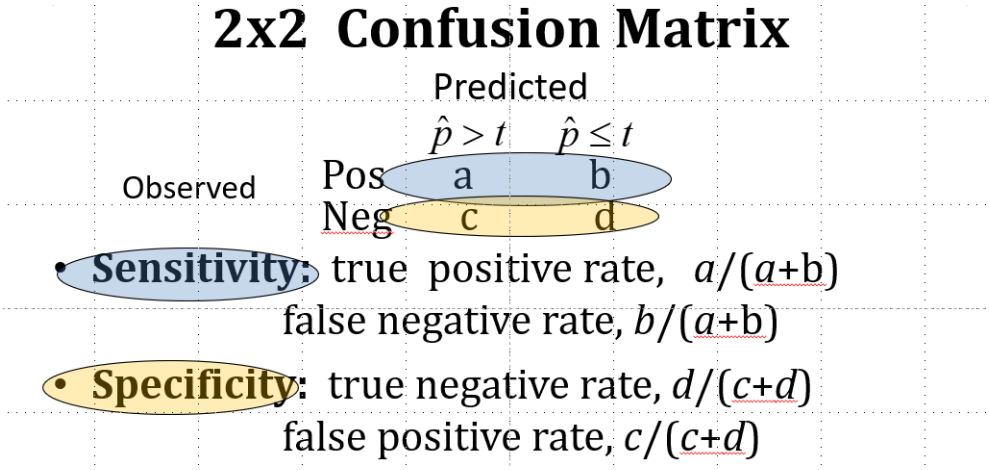

Figure 3.19. Confusion matrix. Source: Yoshida (2020).

Accuracy

Accuracy evaluates how many observations were properly classified, whether positive or negative. On extremely unbalanced data, however, we should avoid employing accuracy as a criterion. In such a situation, by merely classifying all observations as the majority class, it is easy to get a high accuracy rate. Applying accuracy is generally a good start when the data is balanced as well as when every class is identically distributed (Yoshida 2020).

Hence,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{3.11}$$

where TP represents true positive; TN denotes true negative; FP indicates false positive; and FN means false negative.

Area Under the Curve (AUC)

AUC is a model performance measurement for the problems of classification at different threshold settings. The probability curve represents the Receiver Operating Characteristics (ROC), and the degree or capability of separability is expressed by AUC. This shows how well the model will distinguish various categories. The stronger the model is, the higher the

AUC. This technique is used to assess the performance of classification models, as shown in Figure 3.20 (Yoshida 2020).

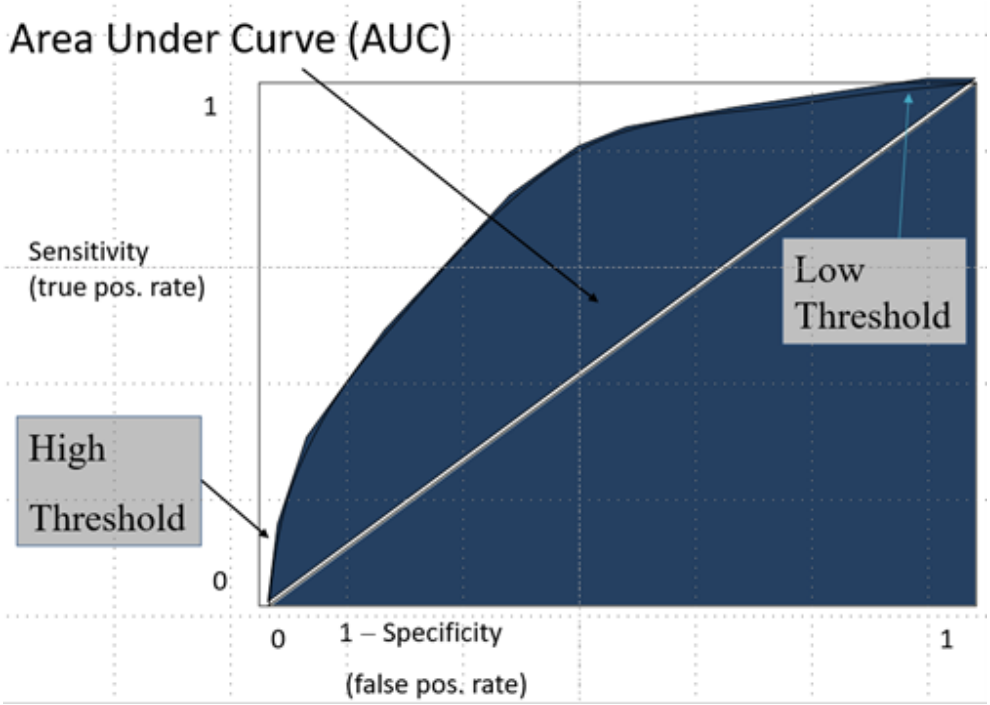

Figure 3.20. Area Under the ROC Curve; the AUC calculates the whole two-dimensional region below the entire ROC curve. X-axis: 1 – Specificity, also known as the false positive rate; Y-axis: Sensitivity, also known as the true positive rate. Source: Yoshida (2020).

F1 Score

The F1 score is the Precision and Recall Weighted Average. This score takes into consideration all false positives and false negatives. It is not as easy to explain as accuracy, but the F1 score is generally more useful than accuracy, especially for unequally distributed classes (Crijnen 2019).

CHAPTER 4:
Results and Analysis

4.1 Results of Time Series Analysis

By accurately predicting pilots' mental states in advance, we can give out a warning before they become distracted. In other words, pilots will be more cautious of loss of SA by knowing that they might be heading toward an unmitigated disaster. For this reason, we started this research using time series analysis.

4.1.1 Assumptions

We conducted the time series analysis based on the following assumptions:

- We can distinguish a pilot's mental state by processing his or her real-time psychophysiological signals. In this research, we used electroencephalogram, electrocardiogram, galvanic skin response, and respiration data.
- The psychophysiological values of a pilot can be predicted by time series models. Here, we assumed that the correlation in the time series will provide the predictive power to forecast future mental states of a pilot.

If the assumptions just stated hold, we can predict a pilot's state of mind beforehand and gain more time for pilots to adjust themselves to stay focused on flying.

4.1.2 Experiment Using Electroencephalogram Data

In the beginning, we chose the first column of the EEG series data, which is denoted as `eeg_fp1`, in order to test the predictability of the electroencephalogram through time. As mentioned in Chapter 3, an electroencephalogram is unique to each person (Saechia et al. 2005), so we minimized our data to exclude all but the `eeg_fp1` signals from the captain of the first crew.

Considering the efficiency of processing such a large amount of data, we restructured our data from approximately every 0.004 seconds per observation (Figure 4.1) to every 0.1

second per observation (Figure 4.2). To achieve this goal, we took the average value of all the `eeg_fp1` data within every 0.1 second window to be a new observation.

	time <dbl>	eeg_fp1 <dbl>
368531	0.1054688	0.295082
368533	0.1093750	0.476633
368535	0.1132812	3.684180
368537	0.1171875	0.109571
368539	0.1210938	-8.325240
368541	0.1250000	-9.906820

Figure 4.1. Before averaging the `eeg_fp1` within each 0.1 second;
time: The original time stamps;
`eeg_fp1`: The original `eeg_fp1` values.

Series.Times <dbl>	mfp1 <dbl>
0.1	-5.48856367
0.2	-11.39456786
0.3	-15.04297563
0.4	-6.95922835
0.5	0.09673909
0.6	2.38045522

Figure 4.2. After averaging the `eeg_fp1` within each 0.1 second;
Series.Times: The new time stamps (per 0.1 second);
mfp1: The mean `eeg_fp1` values within each new time stamp.

After the data processing just described, we began to train our time series models—such as Naïve, Seasonal Decomposition (STL), Exponential Smoothing (ETS), Auto-Regressive Integrated Moving Average (ARIMA), and Ensemble models—by using the observations from within the first 50 seconds.

Figure 4.3 shows the time series model performance in `eeg_fp1`. We used 1-step and 10-step MAPE and MASE to evaluate the error between the predictions and the true values. Nonetheless, the 10-step MAPE values are all greater than the 0.2 criteria, which means the difference between forecasting and true values is greater than 20% when using the Naïve model as a baseline.

Model Performance Comparison

Model	1-Step MAPE	10-Step MAPE	1-Step MASE	10-Step MASE
Naive	12	12	1	1
Seasonal Decomposition	0.029	0.39	0.45	0.9
Exponential Smoothing	4	8.6	0.44	0.89
ARIMA	2.7	4	0.4	0.58
Ensemble	3.4	5.8	0.41	0.74

Figure 4.3. Model performance comparison of `eeg_fp1` values.
Model order (from top to bottom): Naive, Seasonal Decomposition, Exponential Smoothing, ARIMA, Ensemble.
First column: MAPE for one step in time horizon.
Second column: MAPE for ten steps in time horizon.
Third column: MASE for one step in time horizon.
Fourth column: MASE for ten steps in time horizon.

This result tells us that `eeg_fp1` values are not predictable through time series models. The reason for this might be that the mean values of each 0.1 second interval would cause a signal distortion and make them lose their predictability.

Thus, we tried using the random forest algorithm to give us the importance comparison of our predictors in an attempt to find the top three significant variables. Then, we could use the time series models to re-examine the viability of forecasting psychophysiological signals by using those important variables as inputs. If the result was still not stable, we would be able to conclude that trying to find the predictability of psychophysiological signals would not be realistic in this research. Nevertheless, it would allow for future research opportunities on this subject.

4.1.3 Predictor Importance Assessment

From the Feature Importance graphs (Figure 4.4), through the Random Forest method, we can see that electrocardiogram, respiration, and galvanic skin response are the top three features contributing to the homogeneity of the nodes and leaves in the resulting random forest. Hence, we used these three signals to refit our time series models to verify whether it is achievable to predict the psychophysiological data through time.

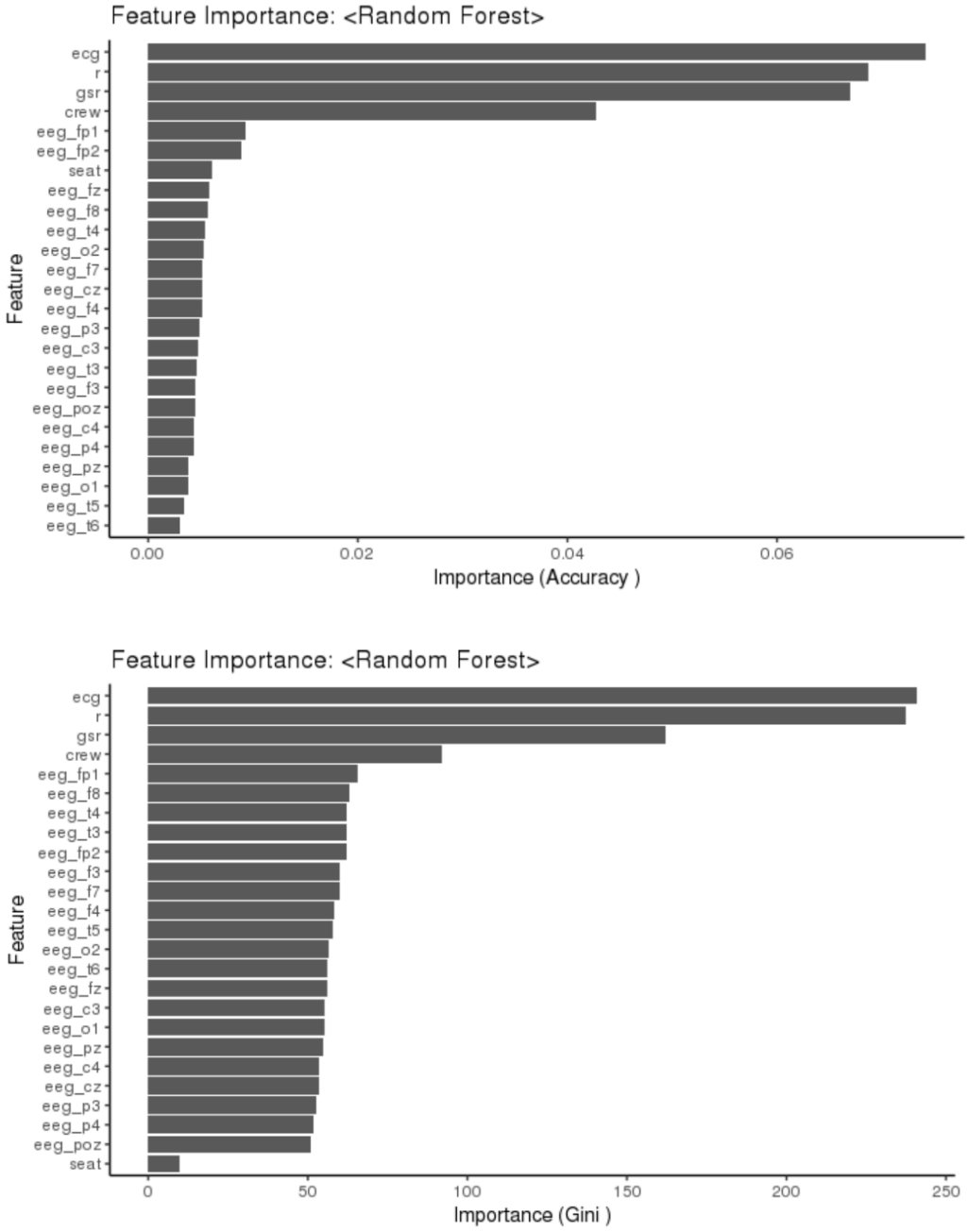

Figure 4.4. Feature Importance graphs from the Random Forest method. X-axis: Accuracy (upper); Gini (lower). Y-axis: The importance of predictors in descending order; the common top four predictors: ecg, r, gsr, crew.

4.1.4 Experiment Using Electrocardiogram Data

We generated the model performance comparison table for ECG data (Figure 4.5) by replicating the methods for fitting the `eeg_fp1` time series model.

Model Performance Comparison

Model	1-Step MAPE	10-Step MAPE	1-Step MASE	10-Step MASE
Naive	0.0062	0.0058	1	1
Seasonal Decomposition	0.0043	0.0049	0.68	0.85
Exponential Smoothing	0.004	0.0043	0.64	0.75
ARIMA	0.0037	0.0043	0.59	0.75
Ensemble	0.0038	0.0043	0.61	0.75

Figure 4.5. Model performance comparison of ECG values.
Model order (from top to bottom): Naive, Seasonal Decomposition, Exponential Smoothing, ARIMA, Ensemble.
First column: MAPE for one step in time horizon.
Second column: MAPE for ten steps in time horizon.
Third column: MASE for one step in time horizon.
Fourth column: MASE for ten steps in time horizon.

The result shows that all time-series models are in conformity with MAPE less than 0.2 and MASE less than 1 (except the Naïve model), which means that all models possess predictive power for this ECG data. Furthermore, the ARIMA model performs the best overall.

4.1.5 Experiment Using Galvanic Skin Response Data

The result for this data (shown in Figure 4.6) indicates that the galvanic skin response (GSR) can be forecast through time, and the ARIMA model offers the best in overall performance. In addition, all time series models are in conformity with MAPE less than 0.2 and MASE less than 1 (except the Naïve model), which means that all models possess predictive power for GSR values.

Model Performance Comparison

Model	1-Step MAPE	10-Step MAPE	1-Step MASE	10-Step MASE
Naive	0.0085	0.0085	1	1
Seasonal Decomposition	0.00089	0.0047	0.11	0.56
Exponential Smoothing	0.00036	0.0025	0.043	0.3
ARIMA	0.00033	0.0025	0.04	0.3
Ensemble	0.00045	0.0028	0.054	0.33

Figure 4.6. Model performance comparison of GSR values.
Model order (from top to bottom): Naive, Seasonal Decomposition, Exponential Smoothing, ARIMA, Ensemble.
First column: MAPE for one step in time horizon.
Second column: MAPE for ten steps in time horizon.
Third column: MASE for one step in time horizon.
Fourth column: MASE for ten steps in time horizon.

4.1.6 Experiment Using Respiration Data

Figure 4.7 shows that respiration signals can be predicted through time and that the ARIMA model still performs the best among all the time series models. Moreover, all time series models are in conformity with MAPE less than 0.2 and MASE less than 1 (except the Naïve model), which means that all models possess predictive power for Respiration data.

Model Performance Comparison

Model	1-Step MAPE	10-Step MAPE	1-Step MASE	10-Step MASE
Naive	0.0017	0.0017	1	1
Seasonal Decomposition	0.00026	0.0011	0.15	0.65
Exponential Smoothing	0.00013	0.0011	0.075	0.62
ARIMA	0.00012	0.00069	0.07	0.4
Ensemble	0.00015	0.00085	0.086	0.49

Figure 4.7. Model performance comparison of Respiration values.
Model order (from top to bottom): Naive, Seasonal Decomposition, Exponential Smoothing, ARIMA, Ensemble.
First column: MAPE for one step in time horizon.
Second column: MAPE for ten steps in time horizon.
Third column: MASE for one step in time horizon.
Fourth column: MASE for ten steps in time horizon.

4.1.7 Time Series and Prediction by Random Forest

From the previous experiments, we found that electrocardiogram, galvanic skin response, and respiration signals are predictive and that the ARIMA model provides the best performance in the three experiments. Consequently, we decided to combine the result from the ARIMA model with the Random Forest classification method to verify the usefulness of predicting future events (mental states) from 50 seconds to 100 seconds.

First, we converted the response variable to a binary factor and split the observations of the first 50 seconds into 80% training and 20% validation datasets. The model performance in the validation set is 93% Accuracy with 95% Sensitivity and 63% Specificity.

Second, we used the random forest classification method to generate predictions from 50.1 seconds to 100 seconds. Then, we applied the predictions as inputs for the Random Forest model to compare the predicted events versus the true values from the original dataset.

The result is shown in the following graph (Figure 4.8).

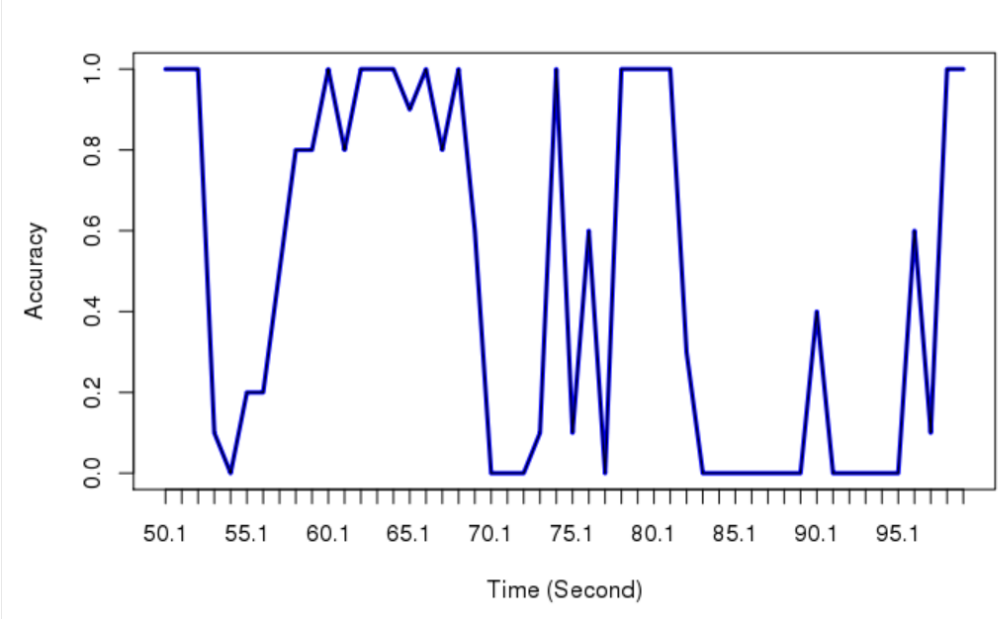

Figure 4.8. Accuracy of predictions from 50.1 seconds to 100 seconds. X-axis: Time in seconds; Y-axis: Accuracy of prediction.

When compared solely by their accuracy, we can see there is no specific pattern for this

irregular fluctuation, and the change of these accuracy rates also contradicts the principle of the time series forecast, which is that the longer we predict from the present time, the worse performance we have (Yoshida 2020).

Therefore, we conclude that using the time series to analyze psychophysiological data may not be useful for experiments of such short durations. Nevertheless, this finding allows for future research opportunities on this subject. In the next section, we mainly focus on the classification method by using present signals as inputs to evaluate whether there is a model that can accurately tell whether a pilot is falling into a dangerous mental state.

4.2 Results of Classification Methods

At this stage, we tried to find an algorithm that could accurately classify the pilots' mental states (safe or dangerous) through the processing of psychophysiological data (electroencephalogram, electrocardiogram, galvanic skin response, and respiration signals) so that we would be able to alert pilots who are gradually falling into a high-risk mental state (i.e., channelized attention, diverted attention, and startle / surprise).

The initial step in finding the best classification algorithm is to randomly select a sample size of 1/1000 with replacement from the original training dataset so that we could have an independently and identically distributed subset for training our algorithm (Stewart 2016) given that IID assumptions are required for machine learning procedures (Nouretdinov et al. 2001).

Then, we divided the sample data into 80% (training) and 20% (validation) to evaluate the initial performance of the models according to Accuracy, Sensitivity, and Specificity from the confusion matrix provided by the caret package in the R programming language and AUC as the criterion to choose the best fit model.

After we had the preliminary best fit classifier, we tried to tune its parameters to achieve better performance by the chosen classification model.

The final step was to randomly sample another dataset as our test data, which was not used for training this model for the purpose of an unbiased evaluation of the final model fit. We not only calculated the Accuracy, but also used the F1 score to give us another perspective

on the overall performance, especially when we had imbalanced classes of the response variable.

4.2.1 Performance Comparison of Different Classifiers

Here, we summarize the respective performance of the Stepwise Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Recursive Partitioning and Regression Trees (rpart), and Random Forest models as shown in Figures 4.9 to 4.13:

Confusion Matrix (Stepwise Logistic Regression)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	566	394	960
Predicted: dangerous	3	11	14
Total	569	405	974
Performance Metrics			
Accuracy		0.5924	
Sensitivity		0.9947	
Specificity		0.0272	
AUC		0.5488	

Figure 4.9. The confusion matrix with the threshold value of 0.5 and the performance metrics for the Stepwise Logistic Regression model.

Confusion Matrix (SVM)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	569	405	974
Predicted: dangerous	0	0	0
Total	569	405	974
Performance Metrics			
Accuracy		0.5842	
Sensitivity		1	
Specificity		0	
AUC		0.5669	

Figure 4.10. The confusion matrix with the threshold value of 0.5 and the performance metrics for the SVM model.

Confusion Matrix (KNN)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	497	102	599
Predicted: dangerous	72	303	375
Total	569	405	974
Performance Metrics			
Accuracy		0.8214	
Sensitivity		0.8735	
Specificity		0.7481	
AUC		0.8791	

Figure 4.11. The confusion matrix with the threshold value of 0.5 and the performance metrics for the acKNN model.

Confusion Matrix (rpart)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	530	158	688
Predicted: dangerous	39	247	286
Total	569	405	974
Performance Metrics			
Accuracy		0.7977	
Sensitivity		0.9315	
Specificity		0.6099	
AUC		0.8192	

Figure 4.12. The confusion matrix with the threshold value of 0.5 and the performance metrics for the rpart model.

Confusion Matrix (Random Forest)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	549	119	668
Predicted: dangerous	20	286	306
Total	569	405	974
Performance Metrics			
Accuracy		0.8571	
Sensitivity		0.9649	
Specificity		0.7062	
AUC		0.9033	

Figure 4.13. The confusion matrix with the threshold value of 0.5 and the performance metrics for the Random Forest model.

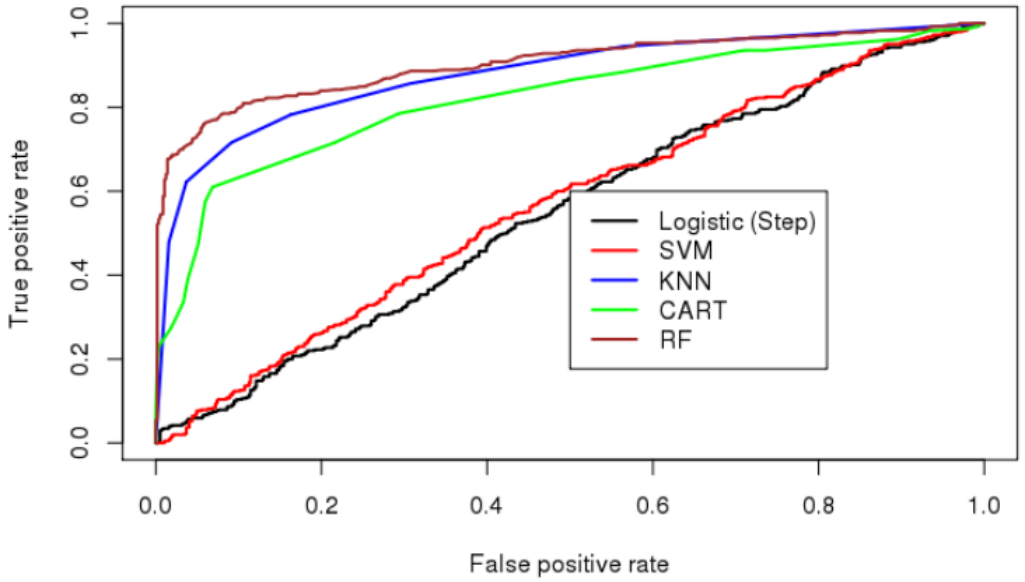

Figure 4.14. ROC comparison for different classifiers; Logit (black), SVM (red), KNN (blue), rpart (green), Random Forest (brown).

Model	Classifier	Accuracy	Sensitivity	Specificity	AUC
1	Logit	0.5924	0.9947	0.0272	0.5488
2	SVM	0.5842	1	0	0.5669
3	KNN	0.8214	0.8735	**0.7481**	0.8791
4	rpart	0.7977	0.9315	0.6099	0.8192
5	**Random Forest**	**0.8571**	**0.9649**	0.7062	**0.9033**

Figure 4.15. Performance comparison of different classifiers (Logit, SVM, KNN, rpart, Random Forest); the bold type denotes the most desired value.

From the ROC graph (Figure 4.14) and the performance comparison table (Figure 4.15), we can see that the Random Forest Classifier performs the best overall when using Accuracy, Sensitivity, Specificity, and AUC as the evaluation criteria. In Figure 4.15, the bold type denotes the most desired value.

In the next step, we focused on the Random Forest model with a different tuning length for the purpose of finding out whether the Random Forest model could perform even better.

4.2.2 Performance Comparison of Random Forest Models with Different Tuning Lengths

In this section, we tried a few different tuning lengths to test whether we could produce a higher performance using the Random Forest algorithm. We also added an F1 score as a supplement criterion because the response variable classes are not evenly distributed and an F1 score is useful in this situation (Crijnen 2019). The *tuneLength* enables the program to automatically tune the algorithm. For each tuning parameter, the *tuneLength* indicates the number of different values to explore (Kuhn 2015).

Confusion Matrix (Random Forest)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	549	119	668
Predicted: dangerous	20	286	306
Total	569	405	974
Performance Metrics			
Accuracy		0.8571	
Sensitivity		0.9649	
Specificity		0.7062	
AUC		0.9033	
F1 score		0.8876	

Figure 4.16. The confusion matrix with the threshold value of 0.5 and the performance metrics for the Random Forest model without tuning.

Confusion Matrix (Random Forest + tuneLength = 20)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	552	85	637
Predicted: dangerous	17	320	337
Total	569	405	974
Performance Metrics			
Accuracy		0.8953	
Sensitivity		0.9701	
Specificity		0.7901	
AUC		0.8801	
F1 score		0.9154	

Figure 4.17. The confusion matrix with the threshold value of 0.5 and the performance metrics for the Random Forest model with tuneLength = 20.

Confusion Matrix (Random Forest + tuneLength = 30)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	555	83	638
Predicted: dangerous	14	322	336
Total	569	405	974
Performance Metrics			
Accuracy		0.9004	
Sensitivity		0.9754	
Specificity		0.7951	
AUC		0.8852	
F1 score		0.9178	

Figure 4.18. The confusion matrix with the threshold value of 0.5 and the performance metrics for the Random Forest model with tuneLength = 30.

Confusion Matrix (Random Forest + tuneLength = 40)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	553	85	638
Predicted: dangerous	16	320	336
Total	569	405	974
Performance Metrics			
Accuracy		0.8963	
Sensitivity		0.9719	
Specificity		0.7901	
AUC		0.881	
F1 score		0.9154	

Figure 4.19. The confusion matrix with the threshold value of 0.5 and the performance metrics for the Random Forest model with tuneLength = 40.

Model	tuneLength	Accuracy	Sensitivity	Specificity	AUC	F1 score
1	none	0.8571	0.9649	0.7062	**0.9033**	0.8876
2	20	0.8953	0.9701	0.7901	0.8801	0.9154
3	30	**0.9004**	**0.9754**	**0.7951**	0.8852	**0.9178**
4	40	0.8963	0.9719	0.7901	0.881	0.9154

Figure 4.20. Performance comparison of the Random Forest model with different tuning lengths (none, 20, 30, 40); the bold type denotes the most desired value.

According to Figures 4.16 to 4.20, we found that the Random Forest model performs the

best when the tuning length equals to 30. This is our best classification model from the validation dataset by far.

4.2.3 Performance of the Random Forest Model on the Test Dataset

Due to the Kaggle competition rules (Kaggle 2019), we could not obtain the complete test dataset to evaluate our final model fit. Thus, we used a randomly selected and unbiased dataset that was also the same size as the data for the above evaluation procedure. In addition, we simulated this process 1,000 times and used the average Accuracy, AUC, and F1 score as our final model performance index (Figure 4.21).

Accuracy	AUC	F1 score
0.8976	0.8827	0.9173

Figure 4.21. The best model mean performance values from 1,000 test datasets.

Figure 4.21 shows the final performance of our best Random Forest model so far by averaging the results from 1,000 independent test datasets. These values indicate that this model is capable of successfully distinguishing the binary mental states of a pilot (safe or dangerous) approximately 90% of the time. Furthermore, the approximately 0.92 F1 score demonstrates that this model can do a fair job of identifying different mental states even though the response classes are imbalanced (2849.692 observations for the safe and 2017.308 observations for the dangerous by averaging 1,000 samples).

4.2.4 Importance of Predictors from the Random Forest Model

From the Feature Importance graphs (Figure 4.22), we can see that electrocardiogram, respiration, and galvanic skin response are the top three features that contribute to the accuracy and homogeneity of the nodes and leaves in the resulting Random Forest model. This result is not surprising because generally when we become nervous or shocked by a sudden event, our heartbeat rate and rhythm of breathing will change; additionally, some people even start to sweat. The variation pattern in the electroencephalogram, however, is not that intuitive from visualizing the signal values.

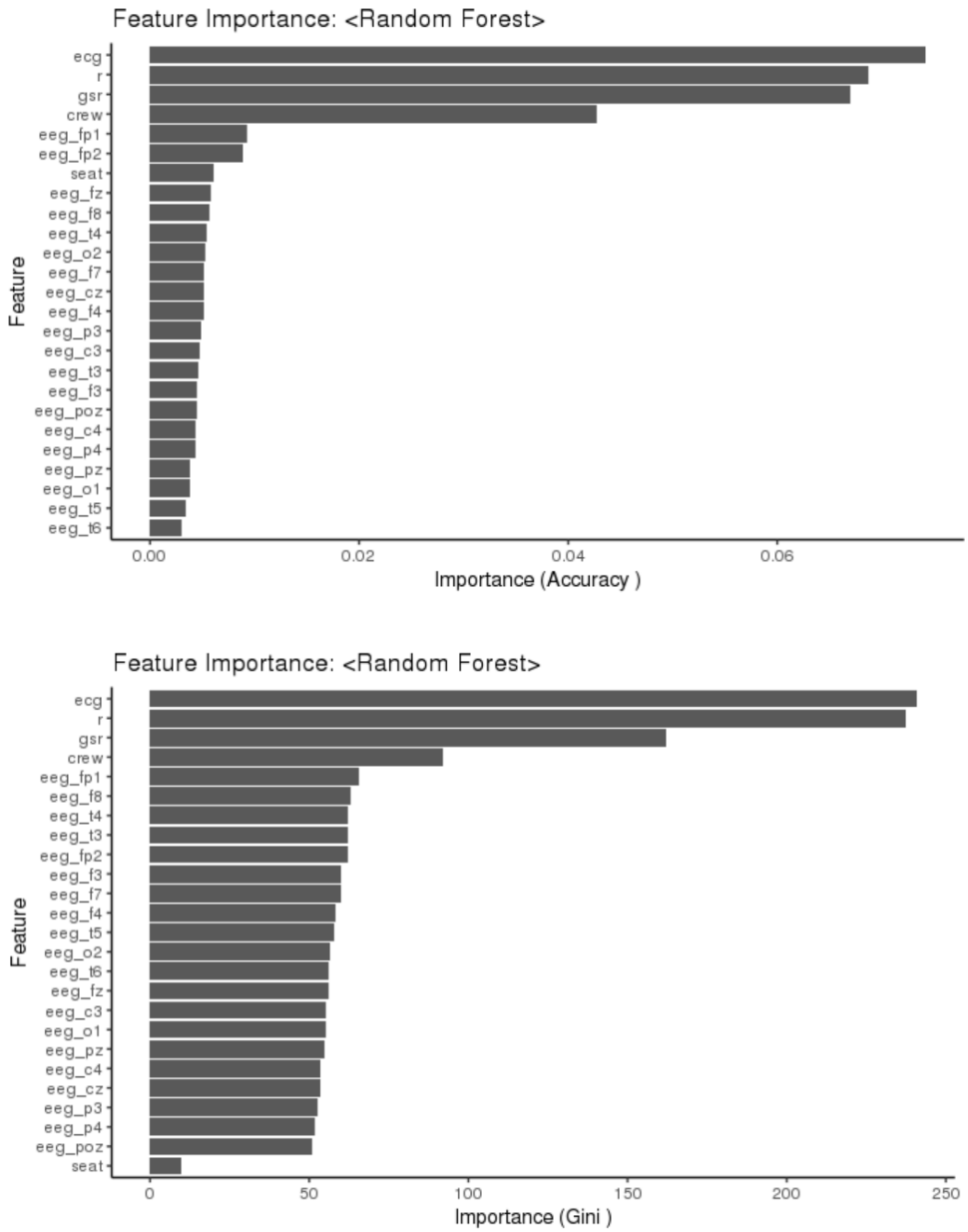

Figure 4.22. Feature Importance graphs from the Random Forest model. X-axis: Accuracy (upper); Gini (lower); Y-axis: The importance of predictors in descending order; the common top four predictors: ecg, r, gsr, crew.

Here, we tried to find out whether we could use only electrocardiogram, respiration, and galvanic skin response data to identify pilots' mental states. We also tried to add the fourth

predictor, "crew," which is also ranked as the fourth important from the Feature Importance graph, to test how the individual's electrocardiogram signal influences the algorithm based on a research conclusion that the electrocardiogram is unique to each person (Marcel and Millan 2007). Figures 4.23 to 4.24 show the results of this experiment.

Confusion Matrix (Random Forest ~ ecg + r + gsr)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	548	72	620
Predicted: dangerous	21	333	354
Total	569	405	974
Performance Metrics			
Accuracy		0.9045	
Sensitivity		0.9631	
Specificity		0.8222	
AUC		0.8927	
F1 score		0.9218	

Figure 4.23. The confusion matrix with the threshold value of 0.5 and the performance metrics for the Random Forest model with three predictors (ecg, r, gsr).

Confusion Matrix (Random Forest ~ ecg + r + gsr + crew)			
	Actual: safe	Actual: dangerous	Total
Predicted: safe	542	68	610
Predicted: dangerous	27	337	364
Total	569	405	974
Performance Metrics			
Accuracy		0.9025	
Sensitivity		0.9525	
Specificity		0.8321	
AUC		0.8923	
F1 score		0.9194	

Figure 4.24. The confusion matrix with the threshold value of 0.5 and the performance metrics for the Random Forest model with four predictors (ecg, r, gsr, crew).

Figure 4.25 shows that the model using only the electrocardiogram, respiration, and galvanic skin responses as predictors performs better overall than the models to which "crew" has

been added and that use all variables from the validation dataset. In addition, the model using electrocardiogram, respiration, galvanic skin response, and crew as predictors performs slightly better than the model that did not use the crew predictor in the test data.

From the results presented, we conclude that it does not matter whether we choose Model 1, 2, or 3, respectively, as shown in Figure 4.26. We can still have approximately 90% discriminability of pilots' mental states (safe or dangerous) by using the Random Forest model as a classifier. Those slight differences according to the result in Figures 4.25 and 4.26 might be the reason for the uniqueness of the electrocardiogram and some artifacts or noise that cannot be handled by the Random Forest model.

Model	predictors	Accuracy	Sensitivity	Specificity	AUC	F1 score
1	ecg, r , gsr	**0.9045**	**0.9631**	0.8222	**0.8927**	**0.9218**
2	ecg, r , gsr, crew	0.9025	0.9525	**0.8321**	0.8923	0.9194
3	all	0.9004	0.9754	0.7951	0.8852	0.9178

Figure 4.25. Performance comparison of Random Forest models with different combinations of predictors on the validation dataset; the bold type denotes the most desired value.
Model 1: The model with three predictors (ecg, r, gsr).
Model 2: The model with four predictors (ecg, r, gsr, and crew).
Model 3: The model with all predictors (20 eeg-prefix recordings, ecg, r, gsr, crew, and seat).

Model	predictors	Accuracy	AUC	F1 score
1	ecg, r , gsr	0.9082	0.8952	0.9253
2	ecg, r , gsr, crew	**0.9087**	**0.8965**	**0.9254**
3	all	0.8976	0.8827	0.9173

Figure 4.26. The mean performance values from 1,000 test datasets; the bold type denotes the most desired value.
Model 1: The model with three predictors (ecg, r, gsr).
Model 2: The model with four predictors (ecg, r, gsr, and crew).
Model 3: The model with all predictors (20 eeg-prefix recordings, ecg, r, gsr, crew, and seat).

Due to the limited information from the Kaggle competition, there are many unknowns in this dataset, such as the equipment for gathering the data and how accurately those data

were collected. Such factors absolutely affect our data analysis process and we might need multiple similar datasets to further verify our assumptions.

In comparison to recent similar studies, our research result is useful and straightforward. Based on the outcome of our experiments, we can gain an awareness that the pilot is being distracted by simply monitoring his or her electrocardiogram, respiration, and galvanic skin response, and such data collecting procedures are apparently easier and less burdensome for a pilot during flight than running an electroencephalogram, which would require attaching 20 sensors on the pilot's scalp.

CHAPTER 5:
Conclusion

5.1 Summary

In this chapter, we summarize the thesis in terms of time series analysis and classification methods and also provide recommendations for future research.

5.1.1 Time Series Analysis

In an attempt to predict pilots' mental states before they fall into a dangerous state, we conducted time series analysis under the presumption that the psychophysiological values of a pilot can be predicted in terms of time. We found that ECG, GSR, and respiration signals are predictable, and the ARIMA model performs the best in those three modalities within the first 50 seconds.

Nevertheless, the inconsistent accuracy of predictions in the range between 50 seconds to 100 seconds shows the mental states of pilots are not predictable. Thus, we conclude that using the time series to analyze psychophysiological data may not be feasible for experiments of such short durations. This finding, however, allows for future opportunities for research on this topic.

5.1.2 Classification Methods

Among all the classifiers we have tested, the Random Forest model provides the best performance, with an accuracy rate of 0.90, AUC of 0.88, and F1 score of 0.92 when we used all predictors (crew, seat, EEG, ECG, GSR, and respiration) as inputs.

According to the Feature Importance graphs from the Random Forest model (Figure 5.1), ECG, GSR, respiration, and crew are the top four most important features that contribute to accuracy and homogeneity of the nodes and leaves in the resulting random forest.

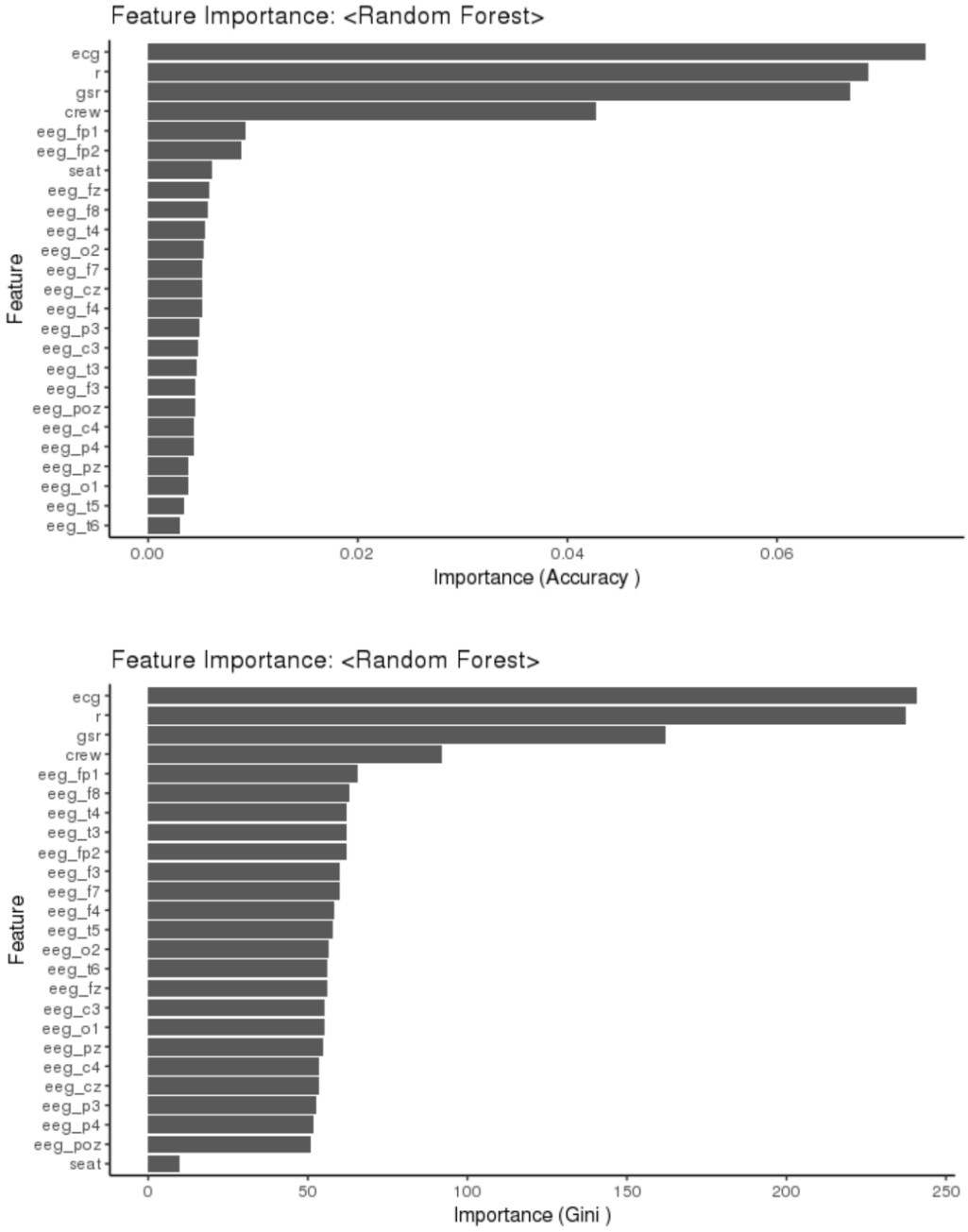

Figure 5.1. Feature Importance graphs from the Random Forest model. X-axis: Accuracy (upper); Gini (lower). Y-axis: The importance of predictors in descending order; the common top four predictors: ecg, r, gsr, crew.

Thus, we tried to refit the model with these top four features and test the model's performance. The average performance from 1,000 test datasets (Figure 5.2) shows that the model using ECG, GSR, respiration, and crew as predictors (Model 2) performs slightly better than the model that did not use the crew predictor (Model 1) and the model using all predictors (Model 3).

Model	predictors	Accuracy	AUC	F1 score
1	ecg, r , gsr	0.9082	0.8952	0.9253
2	ecg, r , gsr, crew	**0.9087**	**0.8965**	**0.9254**
3	all	0.8976	0.8827	0.9173

Figure 5.2. The mean performance values from 1,000 test datasets; the bold type denotes the most desired value.
Model 1: The model with three predictors (ecg, r, gsr).
Model 2: The model with four predictors (ecg, r, gsr, and crew).
Model 3: The model with all predictors (20 eeg-prefix recordings, ecg, r, gsr, crew, and seat).

We conclude that it does not matter whether we choose Models 1, 2, or 3 shown in Figure 5.2. We can still have an approximately 90% ability to discern pilots' mental states (safe or dangerous) by using the Random Forest model as a classifier. Those slight differences according to the results might be caused by the uniqueness of electrocardiogram data to each person (Saechia et al. 2005) and some artifacts or noise which cannot be entirely handled by the Random Forest model.

According to the results of our experiments, we can detect whether a pilot is distracted approximately 90% of the time, simply by measuring his or her electrocardiogram, respiration, and galvanic skin response signals. Keeping that in mind, our model is effective and easier to interpret that the methods presented in recent similar studies.

5.2 Recommendations for Future Research

First, while conducting the time series analysis, we were not able to use the entire dataset for the model training due to the limitation of computer capability. This kind of constraint can be overcome, however, when future analysts train the algorithm with a much more advanced and powerful computer or a more advanced parallel computing skill in programming.

Second, to achieve optimal performance of the model with multiple similar datasets it would be best to have datasets generated from similar experiments to test the reliability of using the Random Forest model as a classifier. If that method is still the best, then we can further apply this algorithm to the real flight environment for the ultimate goal of reducing aviation fatalities.

Third, although flight safety is crucial to mission success, we encourage researchers from around the world with interests in both commercial and military aviation safety to join the research in this field for the purpose of reducing aviation mishaps caused by human factors. With our effort, flight accidents can be forestalled, and with continued research and enhancements to our methods, many more lives can be saved.

List of References

Blocka K (2018) EEG (electroencephalogram): Purpose, procedure, and risks. Accessed April 22, 2020, https://www.healthline.com/health/eeg.

Brockwell PJ, Davis RA (2016) *Introduction to time series and forecasting* (Springer, Nature Switzerland AG.).

Castrounis A (2021) Machine learning: An in-depth guide – model evaluation, validation, complexity, and improvement. Overfitting, Accessed April 22, 2020, https://www.innoarchitech.com/blog/machine-learning-an-in-depth-non-technical-guide-part-3.

Crijnen J (2019) *Predicting a Pilot's Cognitive State from Physiological Measurements*. Master thesis, Tilburg University, Tilburg, The Netherlands, http://arno.uvt.nl/show.cgi?fid=149399.

Harrivel AR, Stephens CL, Milletich RJ, Heinich CM, Last MC, Napoli NJ, Abraham N, Prinzel LJ, Motter MA, Pope AT (2017) Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing. *AIAA Information Systems-AIAA Infotech@ Aerospace*, 1135 (AIAA.org).

Hyndman RJ, Akram M, Archibald BC (2008) The admissible parameter space for exponential smoothing models. *Annals of the Institute of Statistical Mathematics* 60(2): 407–426.

Hyndman RJ, Athanasopoulos G (2018) *Forecasting: Principles and Practice* (OTexts).

Jahangir M, Pirouz P (2020) Alpha wave. Brain waves, Accessed April 22, 2020, https://www.sciencedirect.com/topics/neuroscience/alpha-wave.

James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning*, volume 112 (Springer, New York).

Kaggle (2019) Reducing commercial aviation fatalities. Booz Allen Hamilton, Accessed April 22, 2020, https://www.kaggle.com/c/reducing-commercial-aviation-fatalities.

Kristensen AN, Jeyam B, Riahi S, Jensen MB (2016) The use of a portable three-lead ecg monitor to detect atrial fibrillation in general practice. *Scandinavian Journal of Primary Health Care* 34(3): 304–308.

Kuhn M (2015) A short introduction to the caret package. *R Found Stat Comput* 1.

Malmivuo J, Plonsey R, et al. (1995) *Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields* (Oxford University Press, USA).

Marcel S, Millan J (2007) Person authentication using brainwaves (eeg) and maximum a posteriori model adaptation. *IEEE transactions on pattern analysis and machine intelligence* 29(4):743–752.

Mishra A, Shrivastava KK, Anto AB, Quadir NA (2019) Reducing commercial aviation fatalities using support vector machines. *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 360–364 (IEEE).

Moroze ML, Snow MP (1999) Causes and remedies of controlled flight into terrain in military and civil aviation. Technical report, Air Force Research Lab Wright-Patterson AFB OH, Human Effectiveness Directorate.

Myroniv B, Wu CW, Ren Y, Christian A, Bajo E, Tseng YC (2017) Analyzing user emotions via physiology signals. *Data Sci. Pattern Recognit.* 1(2):11–25.

Nouretdinov I, Vovk V, Vyugin M, Gammerman A (2001) Pattern recognition and density estimation under the general iid assumption. *International Conference on Computational Learning Theory*, 337–353 (Springer, Berlin, Heidelberg).

Rosenkrans W (2015) Airplane state awareness. Flight Safety Foundation, https://flightsafety.org/asw-article/airplane-state-awareness/.

Saechia S, Koseeyaporn J, Wardkein P (2005) Human identification system based ECG signal. *TENCON 2005-2005 IEEE Region 10 Conference*, 1–4 (IEEE).

Shelby B (2018) *What Is Exploratory Data Analysis, Sisense* (Sisense Inc.), URL https://www.sisense.com/blog/exploratory-data-analysis/, accessed April 22, 2020, Publication Title: Sisense.

Shi Y, Ruiz N, Taib R, Choi E, Chen F (2007) Galvanic skin response (GSR) as an index of cognitive load. *CHI'07 extended abstracts on Human factors in computing systems*, 2651–2656.

Stewart B (2016) Week 3: Learning from random samples. *Handout].Soc500: Applied Social Science Statistics. Available at: http://scholar. princeton. edu/sites/default/files/bstewart/files/lecture3handout. pdf* .

Yoshida R (2020) Lecture, time series and classification, OA4106 advanced data analysis. July 2020, Department of Operations Research, Naval Postgraduate School, Monterey, CA.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California