Mental Stress Detection via Heart Rate Variability using Machine Learning

Bimo Sunarfrri Hantono, Lukito Edi Nugroho, and Paulus Insap Santosa

Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
{bhe, lukito, insap}@ugm.ac.id

Abstract: Mental stress is an undesirable condition for everyone. Increased stress can cause many problems, such as depression, heart attacks, and strokes. Psychophysiological conditions possible use as a reference to a person's mental state of stress. The development of mobile device technology, along with the accompanying sensors, can be used to measure the psychophysiological condition of its users. Heart rate allows measured from the photoplethysmography signal utilizing a smartphone or smartwatch. The heart rate variability is currently one of the most studied methods for assessing mental stress. Our objective is to analyze stress levels on the subjects when performing tasks on the smartphone. This study involved 41 students as respondents. Their heart rate was recorded using a smartphone while they were doing the n-back tasks. The n-back task is one of the performance tasks used to measure working memory and working memory capacity. In this study, the n-back task was also used as a stressor. The heart rate dataset and n-back task results are then processed and analyzed using machine learning to determine stress levels. Compared with three other algorithms (neural network, discriminant analysis, and naïve Bayes), the k-nearest neighbor algorithm is most appropriate to use in the classification of time and frequency domain analysis.

Keywords: Classification Algorithm; Heart Rate Variability; Mental Stress; Stress Detection

1. Introduction

Many people have investigated mental stress in various fields because of detrimental effects on daily routines. Mental stress can manifest itself as many different symptoms and signs, ranging from physiological to psychological and behavioral manifestations. Mental stress in humans has been investigated by Castaldo using several cognitive stressors in the laboratory or real-life scenarios [1]. Some physiological signals can be used to detect stress, such as galvanic skin response, blood pressure, electroencephalogram, respiration rate, and electrocardiogram [1]. However, the Heart Rate Variability (HRV) is currently one of the most studied methods for assessing mental stress. HRV describes the variation in the interval between successive peaks of heartbeat waves. HRV itself is far more sensitive in measuring stress than just using the heart rate itself. HRV analysis is a widely accepted tool for the noninvasive assessment of the autonomic nervous system (ANS) [2]. HRV analysis and use are becoming increasingly common since it is sensitive to both physiological and psychological changes [3].

HRV possible analyzed in the time, frequency, and non-linear domain. HRV analysis can perform on a 24-hour nominal record (long-term HRV analysis), 5-minute recording (short-term HRV analysis), or shorter records [2]. Short-term HRV analysis makes it possible to monitor individual stress levels continuously and in real-time. This analysis is essential in several circumstances or occupations, such as for a surgeon or a driver. In this paper, an ultra-short-term HRV analysis is an analysis conducted on HRV in less than 5 minutes.

At present, researchers had carried out some research on the real-life stress detection through short-term HRV analysis. Over the past decade, the development of the mobile device has grown exponentially. The existence of smartphones and smartwatches makes demands for short-term HRV analysis to monitor the healthcare and wellbeing status of individuals continue to increase [4]. Some smartphones have a heart rate sensor, which will give an estimate of the average heart

Received: March 17th, 2020. Accepted: July 17th, 2020
DOI: 10.15676/ijeei.2020.12.3.3
rate obtained from pulse signals recorded from the fingertips. They are based on photoplethysmography (PPG). PPG is using a light source for illuminating the skin and a photodetector for measuring the intensity of the reflected light. The pulsatile changes based HR estimates in reflected light induced by fluctuations in blood flow every heartbeat.

In this context, several studies had conducted on this topic. For example, research conducted on healthy and unhealthy subjects using HRV in the time and frequency domains resulted in the classification of tachycardia and bradycardia values [5]. Apple Watch can detect the relaxed and mental stress conditions of participants by using statistical analysis of HRV in the time and frequency domains [6]. Machine learning is implemented on ultra-short HRV features to identify the conditions of rest and stress in a person [7]. The use of the one-dimensional convolutional neural network (1D CNN) in a single pulse PPG can distinguish a person’s valence and arousal conditions [8].

This study’s objective is to compare mental stress detection using four machine learning methods on PPG signals obtained from smartphones. The n-back task is used as a stressor and as a performance measurement to validate the measurement results based on HRV. The first stage is pre-processing PPG signals, followed by feature extraction, and the last stage to implement four machine learning methods.

2. Methods

This study involved 41 students who volunteered to participate. The number of respondents is considered sufficient as has been done in other similar studies that use between 12 to 42 respondents [6]–[11]. All the participants are in good health and write down their agreement when participating. Before data collection, participants were asked not to consume energy drinks, alcohol, coffee, or smoking and assume they had an enough sleep. Participants are given information on the purpose of data retrieval and instructions on the tasks that they will complete before starting the assignment. Participants will work on the n-back task together with their heart rate recorded using the PPG sensor found on the smartphone. The n-back task is a continuous performance task that is commonly used as an assessment in cognitive neuroscience to measure a part of working memory and working memory capacity [12]. The subject is presented with a sequence of stimuli, and the task consists of indicating when the current stimulus matches the one from \( n \) steps earlier in the series. The stimulant can be anything, for example, pictures, colors, and alphabet. Figure 1 illustrates the stimulus used in this study in alphabetical form. The load factor \( n \) can be adjusted to make the task less or more difficult.

![Figure 1. Illustration of the n-back tasks](image)

Participants sit comfortably as possible during the data collection process while holding a smartphone and place their index finger on the smartphone’s heart rate sensor. They must be quiet to reduce body movements, especially the hand, while holding the smartphone, to avoid weak signal readings. Figure 2 shows the experimental procedure. The experimental protocol consisted of three test sessions, each session consisting of three n-back task stages. Participants rested for one minute, in the beginning, ending, and between each session, but the fingers remain at the heart rate sensor, to measure the resting conditions.
The heart rate measurement results obtained will be pre-processed before finally being analyzed based on the HRV features. The HRV parameters divided into time and frequency domain. Since both are calculated in different ways, the gaps also are addressed differently. The time-domain analysis will derive the HRV parameters from R-R intervals. We proposed eight parameters of the time domain, i.e., mRR, RR, SDHR, SDRR, CVRR, RMSSD, pRR20, and pRR50. Four parameters of the frequency domain were proposed, namely ultra-low frequency (ULF), very low frequency (VLF), low frequency (LF), and high frequency (HF). Table 1 describes the HRV features in the time and frequency domain.

The smartphone device used in this study is the Samsung Galaxy Note 5 with the Android operating system 7.0 (Nougat). The smartphone uses an Exynos 7420 processor, 4GB RAM, 32GB internal storage and has a heart rate (PPG) sensor. The PPG (photoplethysmogram) sensor of the smartphone is validated using two commercial heart rate measuring devices in the form of a chest belt, namely Garmin HRM3-SS and Magene Mover 2. The results of the Pearson correlation test for the two devices are 0.9884 and 0.9355. Values that are close to number one indicate that there is a very strong relationship between the two variables being compared.

<table>
<thead>
<tr>
<th>HRV</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRR</td>
<td>ms</td>
<td>Mean of RR intervals</td>
</tr>
<tr>
<td>SDRR</td>
<td>ms</td>
<td>The standard deviation of RR intervals</td>
</tr>
<tr>
<td>mHR</td>
<td>BPM</td>
<td>Mean of heart rate</td>
</tr>
<tr>
<td>SDHR</td>
<td>BPM</td>
<td>The standard deviation of heart rate</td>
</tr>
<tr>
<td>CVRR</td>
<td></td>
<td>Coefficient variation of RRI</td>
</tr>
<tr>
<td>RMSSD</td>
<td>ms</td>
<td>Root mean square of the successive differences</td>
</tr>
<tr>
<td>pRR20</td>
<td>%</td>
<td>The ratio of the difference in consecutive RRIIs over 20 ms</td>
</tr>
<tr>
<td>pRR50</td>
<td>%</td>
<td>The ratio of the difference in consecutive RRIIs over 50 ms</td>
</tr>
<tr>
<td>Frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ULF</td>
<td>ms^2</td>
<td>Power of ultra-low frequency band (0 – 0.003 Hz)</td>
</tr>
<tr>
<td>VLF</td>
<td>ms^2</td>
<td>Power of very-low-frequency band (0.003 – 0.04 Hz)</td>
</tr>
<tr>
<td>LF</td>
<td>ms^2</td>
<td>Power of low-frequency band (0.04 – 0.15 Hz)</td>
</tr>
<tr>
<td>HF</td>
<td>ms^2</td>
<td>Power of high-frequency band (0.15 – 0.4 Hz)</td>
</tr>
<tr>
<td>LF/HF</td>
<td></td>
<td>The ratio of LF and HF band powers total</td>
</tr>
</tbody>
</table>

### 3. Results

The training data used in the stress level classification comes from eight HRV features in the time domain and four HRV features in the frequency domain. The eight HRV features in the time domain used are mRR, SDRR, mHR, SDHR, CVRR, RMSSD, pRR20, and pRR50. While the four HRV features in the frequency domain used are ULF, VLF, LF, and HF. Reference grouping of HRV features used the time obtained at a specific value of accuracy and reaction time, which used as training data, and the result will be the input for the four classification models.
The neural networks classification model used has 1000 epoch and ten hidden layer parameters. From the input of eight HRV features in the time domain or four HRV features in the frequency domain, we will get five outputs, which are classes or labels given in the training data. The neural network classification model that uses optimizer is resilient backpropagation. The mean squared error (MSE) method is used as an error measurement during the training process. The results of the training of the neural networks classification model can be seen in the confusion matrix in Figure 3. Output class is a class of identification results from data, while the target class is the actual class of the data. From the confusion matrix, the training data in the time domain shows that the neural networks classification model has an accuracy rate of 74.2%. The correct classification values are found in the fifth class or rest conditions that are equal to 65.3%. While the least value in the fourth class or the highest stress level in this classification is 1.1%. Whereas in the frequency domain cannot classify with the correct value on data in the second, third, and fourth classes.

The number of $k$ in the k-nearest neighbor classification model used is one. The results of the classification of training data from the k-nearest neighbor classification model are shown in the

![Neural Network Training Confusion Matrix](image)

**Neural Network Training frequency Confusion Matrix**

![Neural Network Training frequency Confusion Matrix](image)

**Figure 3 Neural network’s data training confusion matrix**

b. Frequency domain
The confusion matrix shows the accuracy of the overall classification of training data with the k-nearest neighbor classification model of 88.6%. The correct classification values are found in the fifth class or rest conditions, which are 67.4%. The third class is the least value, which is 45%. The values obtained are also the same between the k-nearest neighbor classification model in the time or the frequency domain.

In the discriminant analysis classification model used is a linear type. The results of the classification of training data from the discriminant analysis classification model are shown in the confusion matrix in Figure 5. From the confusion matrix, the accuracy of the overall classification of training data with the discriminant analysis classification model in the time domain is 67.7%. The correct classification value is found in the fifth class or rest conditions equal to 61.5% and the lowest in the third class, which is equal to 0%. Whereas in the frequency domain, it has an accuracy of 70%. The correct classification value is found in the fifth class or rest conditions that are equal to 65.8% and the least value in the second class, which is equal to 0%. 

Figure 4 k-Nearest neighbor’s data training confusion matrix
The Naïve Bayes classification model used is a linear type. The results of the classification of training data from the discriminant analysis classification model are shown in the confusion matrix in Figure 6. From the confusion matrix, the accuracy of the overall classification of training data with the Naïve Bayes classification model in the time domain is 62.1%. The correct classification values are found in the fifth class or rest conditions that are equal to 58.1% and the lowest in the third class, which is equal to 0%. Whereas in the frequency domain, it was 17.3%. The correct classification values are found in the fifth class or rest conditions, which are 11.6% and the lowest in the second class, which is 0.4%.

**Figure 5. Discriminant analysis’ data training confusion matrix**
From the results of the training data classification above, they were then testing data from 20% of the existing respondents, namely as many as eight respondents. Figure 7 shows the classification accuracy of the test data and training data. Classification results in the time domain the highest accuracy is owned by the k-nearest neighbor model with a classification accuracy value of the test data of 82.8%. In comparison, the naïve Bayes model has the lowest accuracy value, with an amount of 60.8%. Figure 7 shows the similarity between the classification performance of test data with training data. Four classification models have difficulty in classifying data in the second and third classes. This condition also occurs in the classification of test data.

In the frequency domain classification result, the neural networks model can not classify data on the second, third, and fourth class. The k-nearest neighbor model as a whole classifies well. The discriminant analysis model has a slightly better performance than the neural networks model in conducting classifications in the second, third, and fourth class data class. However, the differences in the two models are not much, and the naïve Bayes model as a whole cannot classify with satisfactory results.
The naïve prediction is a model used to predict all data with the same value from the majority data. Classification in the time domain, the most data is data in the fifth class, then the accuracy value of naïve prediction is 67.5%. Thus, the lower limit of the model’s feasibility is if the model has an accuracy value above 67.5%, therefore the models that are declared eligible for classification are neural networks and k-nearest neighbor models with an accuracy value of 73% and 82.8%, respectively. Classification in the frequency domain, the most data is data in the fifth class, then the accuracy value of naïve prediction is 68.2%. From the lower limit of the model’s feasibility based on this method, the models that are suitable for classification are the neural networks, k-nearest neighbor, and discriminant analysis models.

Table 2. Under the curve area in the time domain

<table>
<thead>
<tr>
<th>Model/Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>0.66</td>
<td>0.60</td>
<td>0.59</td>
<td>0.61</td>
<td>0.73</td>
</tr>
<tr>
<td>k-NN</td>
<td>0.84</td>
<td>0.73</td>
<td>0.72</td>
<td>0.8</td>
<td>0.899</td>
</tr>
<tr>
<td>DA</td>
<td>0.70</td>
<td>0.71</td>
<td>0.72</td>
<td>0.79</td>
<td>0.75</td>
</tr>
<tr>
<td>NB</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
<td>0.72</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Table 3. Under the curve area in the frequency domain

<table>
<thead>
<tr>
<th>Model/Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>0.51</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.686</td>
</tr>
<tr>
<td>k-NN</td>
<td>0.84</td>
<td>0.71</td>
<td>0.70</td>
<td>0.78</td>
<td>0.89</td>
</tr>
<tr>
<td>DA</td>
<td>0.63</td>
<td>0.66</td>
<td>0.67</td>
<td>0.76</td>
<td>0.68</td>
</tr>
<tr>
<td>NB</td>
<td>0.52</td>
<td>0.56</td>
<td>0.59</td>
<td>0.61</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Figure 8. Receiver operating characteristic in the time domain

- a. Neural Networks
- b. k-Nearest Neighbor
- c. Discriminant Analysis
- d. Naïve Bayes
Figure 9. Receiver operating characteristic in the frequency domain
Receiver Operating Characteristic (ROC) is the area of the curve in the true positive rate and false positive rate axis charts. This graph shows whether the model can distinguish classification data from one class to another. If the model can differentiate between classes, the ROC graph will lean towards the top left. Besides, if the model can distinguish between data classes and non-data classes, the area under the curve (AUC) is above 0.5. AUC values below 0.5 indicate that the model is unable to distinguish between data classes. AUC values close to 0.5 indicate that the model is less able to identify data classes (random guess). AUC values above 0.5 indicate the model can distinguish classes data class. They are shown in Figure 8, the ROC graph of the four models.

From Figure 8 and Table 1, it can be concluded that all models in the time domain have an AUC above 0.5. The lowest AUC value is owned by the neural networks model in the third class. This condition is also seen in confusion matrix neural network models that have low accuracy. Overall the lowest AUC in all four models is in the third class. Based on the receiver operating characteristic method and area under the curve, all models can be declared eligible for classification, but with a note that the model needs to improve classification performance, especially in the third class. Based on the two methods above, the model that is feasible to classify is the neural networks and k-nearest neighbor models.

From Figure 9 and Table 2, it can be seen that the neural network model has an AUC of 0 in the second, third, and fourth classes. So it can be said that the neural networks model cannot distinguish between second, third, and fourth class data with data in other classes. According to this method, models that are suitable for classification are the k-nearest neighbor model and discriminant analysis. The neural network model has an AUC below 0.5, while the naïve Bayes model has an AUC of close to 0.5 so that it can be declared unfeasible to classify the frequency domain.

4. Discussion

The training process and testing of four classification models, i.e., neural networks, k-nearest neighbor, discriminant analysis, and naïve Bayes can classify from eight features of the time domain and four features of the frequency domain. In the time domain classification model, the neural networks classification model, k-nearest neighbor, discriminant analysis, and naïve Bayes have classification accuracy on the test data with the following accuracy values as 73%, 82%, 66%, and 60%. The accuracy value is re-examined for the feasibility of the model using the naïve prediction method and the receiver operating characteristic method. The feasibility results of the classification model state that what is feasible to use in the classification process in the time domain are the neural networks and k-nearest neighbor models. The discriminant analysis classification model and the naïve Bayes model cannot be declared feasible for classification because these two models have a lower accuracy value compared to the accuracy values found in the naïve prediction.

In the frequency domain, neural networks, k-nearest neighbor, discriminant analysis, and naïve Bayes classification models have accuracy in the test data with the following accuracy values are 71.2%, 82.9%, 70.6%, and 16.4%. The lower limit of the naïve prediction method in the frequency domain classification is 68.2%. Therefore, the models that are declared feasible to classify the frequency domain are the neural networks model, k-nearest neighbor, and discriminant analysis. The models that can be declared feasible to do the classification are the k-nearest neighbor model and discriminant analysis, when the four models are examined by the receiver operating characteristic method. Neural networks and naïve Bayes models cannot be declared feasible because the neural networks model has an under the curve value that is below 0.5, indicating that the neural networks model cannot distinguish between data in certain classes with data in other classes. In contrast, the naïve Bayes model has the area under the curve value, most of which are in the range of values of 0.5. This condition indicates that the naïve Bayes model has not been able to distinguish clearly between data with certain classes with data in other classes; even the naïve Bayes model tends to use random guesses. Then it can be concluded
in this study that the models that are declared feasible to classify in the frequency domain are the k-nearest neighbor and discriminant analysis.

The k-nearest neighbor is a useful classification model in the time domain and the frequency domain to classify the stress level in this study. The receiver operating characteristic (ROC) graph k-nearest neighbor for classification in the time and frequency domain tends to be above the midline. Even in the class that shows the resting conditions of the ROC graph, the k-nearest neighbor has a high area under the curve value, which is 0.89. The under the curve graph indicates that there are apparent differences between the classes of data used to classify in the time domain and the frequency domain. Then it can be said that the subject’s behavior when working on n-back tasks has apparent differences. The difference in subject behavior is shown in the reaction time graph and the accuracy graph over a while. Subjects will experience memory space running out of cognitive architecture during n-back tasks so that the subject will experience an increase in stress levels. Besides, it can also be said that HRV data recorded during the n-back process with a 6000 ms moving window is data that can reflect the mental stress conditions of the subject. Reaction time and accuracy obtained from n-back data reflect the subject’s condition while working on the task. Reaction time and accuracy are used to validate models that classify time and frequency domain data.

5. Conclusion
In this study, the PPG heart rate sensing on a mobile device has been verified to work accurately and has provided the functionality needed. Detrending, filtering, and peak detection algorithms from PPG signals and calculation of the heart rate variability features in the time and frequency domains have been proven to work adequately in moderate motion artifacts. The answer n-back task's value is used as a label in training data and testing data in the machine learning. Ultra-short-term HRV analysis conducted on 41 participants showed promising results when implemented in the time domain. In contrast, the frequency domain analysis was still limited to distinguish the conditions of participants doing n-back tasks or resting conditions. This happens because the time used is relatively short, which is the lower limit for analysis in the frequency domain. The k-nearest neighbor algorithm is most appropriate for use in this study’s classification compared to four other algorithms used in the time or frequency domain. The neural network’s algorithm is also quite feasible in the time domain analysis, whereas the discriminant analysis algorithm is quite feasible for the frequency domain analysis. The reliable and valid measurement of mental stress will allow many areas of cognitive psychology to be explored further. The future work that needs to be done is to increase the number of participants, to involve medical or psychological experts to clarify the outcome of the work, and to predict the occurrence of stress.

6. References


**Bimo Sunarfrri Hantono** received a B.Eng. from Universitas Gadjah Mada, Indonesia at 2000 and a M.Eng. from Nanyang Technological University, Singapore at 2006. He has worked for Department of Electrical and Information Engineering at Universitas Gadjah Mada since 2002. His research interest includes human computer interaction, mobile sensing, and web-based application. He is currently pursuing a doctoral degree in Electrical Engineering, Universitas Gadjah Mada.

**Lukito Edi Nugroho** is a senior lecturer at the Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada. He joined as an academic staff after completing his undergraduate degree at the same Department. He obtained his M.Sc. degree from James Cook University in 1995 and his Ph.D. from Monash University, both in Australia. His research interests include software engineering, distributed and mobile computing, and application of ICT in education. He is particularly interested in how context-based scenarios can be specified in different types of applications.
Paulus Insap Santosa is a senior lecturer at the Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada. He joined as an academic staff after completing his undergraduate degree at the same Department in 1985. He obtained his M.Sc. degree from University of Colorado in 1991 and his Ph.D. from Nasional University of Singapore in 2006. His research interests include Human-Computer Interaction (User Experience), Web Application Development, Web Design, IT in Education.