Article

Biosignal-Based Driving Skill Classification Using Machine Learning: A Case Study of Maritime Navigation

Hui Xue [1,*] **, Bjørn-Morten Batalden** [1] **, Puneet Sharma** [2] **, Jarle André Johansen** [2] **and Dilip K. Prasad** [3]

1 Department of Technology and Safety, UiT The Arctic University of Norway, 9019 Tromsø, Norway; bjorn.batalden@uit.no

2 Department of Automation and Processing Technology, UiT The Arctic University of Norway, 9019 Tromsø, Norway; puneet.sharma@uit.no (P.S.); jarle.johansen@uit.no (J.A.J.)

3 Department of Computer Science, UiT The Arctic University of Norway, 9019 Tromsø, Norway; dilip.prasad@uit.no

* Correspondence: hui.xue@uit.no

Abstract: This work presents a novel approach to detecting stress differences between experts and novices in Situation Awareness (SA) tasks during maritime navigation using one type of wearable sensor, Empatica E4 Wristband. We propose that for a given workload state, the values of biosignal data collected from wearable sensor vary in experts and novices. We describe methods to conduct a designed SA task experiment, and collected the biosignal data on subjects sailing on a 240° view simulator. The biosignal data were analysed by using a machine learning algorithm, a Convolutional Neural Network. The proposed algorithm showed that the biosingal data associated with the experts can be categorized as different from that of the novices, which is in line with the results of NASA Task Load Index (NASA-TLX) rating scores. This study can contribute to the development of a self-training system in maritime navigation in further studies.

Keywords: biosignal; maritime navigation; classification; situation awareness (SA); neural network; maritime training

1. Introduction

The classic methodology for maritime training generally involves multiple sensors in addition to simulator for improving situation awareness (SA) in maritime navigation and seafaring skills [1]. The purpose of this study is to determine if a wearable sensor can be used to detect stress changes with skills during a maritime navigation task. We define stress as the task requirement for both experienced seafarers (experts) and novices (students). We collected the biosignal data of subjects for indicating the stress differences under the SA tasks during maritime navigation. Biosignal data including electrodermal activity (EDA), body temperature, blood volume pulse (BVP) and heart rate (HR) are some of the indicators to present the stress level, since stress is the body's reaction to pressure and a physical response to situations in which people feel threatened.

Safe maritime navigation in the Arctic region is challenging because there is less infrastructure, long distances between harbours and harsh weather conditions [2]. However, the safety of the Arctic route is of great significance to the economic development of Scandinavia, and at the same time has an impact on environmental protection and the safe growth of marine life. The sailing route in the Vessel Traffic Services area on the west coast of Norway, north of Bergen, is a typical route for training the seafarers because of its complexity and busy traffic, especially for training SA in the safety of the maritime navigation.

In maritime, the study of situation awareness (SA) has always been an important topic of discussion. Studies show that many ship collisions and groundings occur due to navigators' erroneous SA. Grech et al. [3] found that 71% of navigators' errors can be

attributed to SA-related problems. Therefore, training maritime students to improve their SA is one of the most important tasks in maritime education.

In maritime navigation, an experienced navigator can keep track of multiple tasks and deal with more complex situations without losing SA as compared to a novice. For a novice, managing multiple tasks required for navigation can be quite challenging [4]. As navigating a ship can be stressful, managing this stress can bring different results. In this study, we aim to investigate whether there are differences in the biosignals between the experts and novices for a given sailing task.

1.1. Related Research Work

There are only a few works related to performance assessment of SA objectively during maritime training navigation. In the existing literature, survey and interview are usually the common tools for assessing SA. From its conception, SA was defined by Endsley in 1988 as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, the projection of their status in the near future" [5,6]. In simple terms, it can be understood as "being aware of what is happening around you and understanding what information means to you now and in the future" [6]. In maritime operations, the awareness refers to the important information for sailing and being safe on board (particular job or goals in general), and only the situations that relate to the tasks are important to SA. For example, the navigator of a ship must be aware of other ships, the weather, the water, the grounding, and so on. When sailing a ship, during the different tasks, the navigator should be able to adjust and adapt the performance based on the current situation or the change of the situation. In order to improve the training system for SA within maritime navigation, several studies of sensor fusion technology have been used on simulators in the past few years [1,7]. However, as far as we know, use of bio-sensor data with the SA training system is not common. The main contribution of this paper is to provide a possibility for doing such research.

1.2. Objective and Contributions

The main objective of this study is to investigate the differences in the stress levels of experts and novices in SA experiment during a maritime navigation task. In this study, we investigate the following research hypotheses: First, the biosensor data from experienced experts can be distinguished from the biosensor data from students. Second, the stress levels obtained from the biosensor data show a correlation to the NASA-TLX rating results. Third, compared to the novices, the experts feel less stress during a navigation task.

The main contribution can be summarised as follows:

- Maritime transport requires safety and security. SA in maritime is the effective understanding of activity that could impact the security and safety. This study discovers that the stress level varies according to the experience of the seafarers, which matters in the performance of SA during the maritime navigation.
- SA training is common in the aviation domain and the maritime domain. While SA training-related stress level analysis is widely studied for aviation, SA training-related stress level analysis in maritime navigation is less studied. This paper is a pilot case study towards classifying stress level among the expert and novice seafarers. We have used a 'hybrid' convolutional neural network approach in combination with statistical, wavelet and higher-order crossing features to classify the stress level based on biosignals during the maritime navigation. We are first extracting features and then passing those features to a Convolutional Neural Network so we refer to is as a 'hybrid Convolutional Neural Network'.

2. Methodology

In order to study the SA-based stress level analysis, we used the Kongsberg K-Sim navigation platform. Both expert and novice drivers were given the same driving scenario and their biosignals were recorded using Empatica E4 band while they were driving on

the simulator. In the next sub-section we elaborate on the information about participants, materials and apparatus used for this study.

2.1. Participants

The trial was performed with 10 healthy male participants. In order to compare the performance and emotion between the experts and novice in maritime navigation tasks, both experts and novices were invited to participate in the experiment. There were five navigators with extensive experience (mean age = 41.8 years, standard deviation = 14.0 years) and five second-year students from a nautical science program with little experience (mean age = 22.8 years, standard deviation = 1.2 years). The experts had an average of 9.7 years' experience as navigators with the longest period being 18 years and the shortest being one year.

2.2. Materials and Apparatus

A maritime navigation task was designed for testing the relationship between the navigating experience and stress. The maritime navigation task was performed on a 240° view simulator. It is equipped with the K-sim Navigation software from Kongsberg Digital. The vessel-model used in the experiment is called BULKC11 (overall length of 90 m and a moulded beam of 14 m). The task consists of two part, one part is sailing, the other part is filing the The Situation Awareness Global Assessment Technique (SAGAT) queries when the simulator screen is frozen. Each participant sailed a 40-minute voyage with four stops. Each section of the sailing lasts approximately 8 to 12 min. During the sailing section, participants had to complete the SAGAT queries in around 15 min (4 stops with an average of 4 min to answer the SAGAT queries). The whole experiment takes approximately 55 min. Figure 1 shows a participant sailing on the simulator.

Figure 1. One of the participants was sailing on the simulator.

During the experiment, each participant wore a wearable device for collecting the biosignal data. In this study, among the diversity of wearable sensors, a medical-grade wearable device, Empatica E4 Wristband (see Figure 2), was chosen for recording the real-time physiological data to conduct in-depth analysis. EDA and PPG sensors were equipped in the E4 Wristband that can simultaneously enable the measurement of sympathetic nervous system activity and heart rate [8]. Following is the description of the sensors in the E4 Wristband:

- **PPG Sensor:** Measures blood volume pulse (BVP), from which heart rate variability can be derived [8];

- **Infrared Thermopile:** Reads peripheral skin temperature [8];
- **EDA Sensor (GSR Sensor):** Measures the constantly fluctuating changes in certain electrical properties of the skin [8];
- **3-axis Accelerometer:** Captures motion-based activity [8].

The data from E4 Wristband such as Electrodermal Activity (EDA), body temperature, blood volume pulse (BVP) and heart rate (HR) are collected and used in the analysis.

Figure 2. Empatica E4 wristband [8].

3. Experiment

In this experiment, we used NASA Task Load Index (NASA-TLX) as a reference. The rating result of NASA-TLX is a subjective measurement evaluated by the participants themselves. The result show that there are different workload and stress level between experts and students. In the light of this result, we hypothesize that it is possible to classify the biosignal data we collected during the sailing task. Hence, we extracted the features of the data and analyzed it by using convolutional neural network(CNN) in deep learning.

3.1. NASA Task Load Index (NASA-TLX)

The NASA Task Load Index (NASA-TLX) was used as an assessment tool to rate the perceived workload in order to assess the performance of the participants [9,10]. The six categories were required to be rated from low to high level, namely, Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration Level. The rating was transferred to a ten-point scaler scores. All the participants were given NASA-TLX after the experiment.
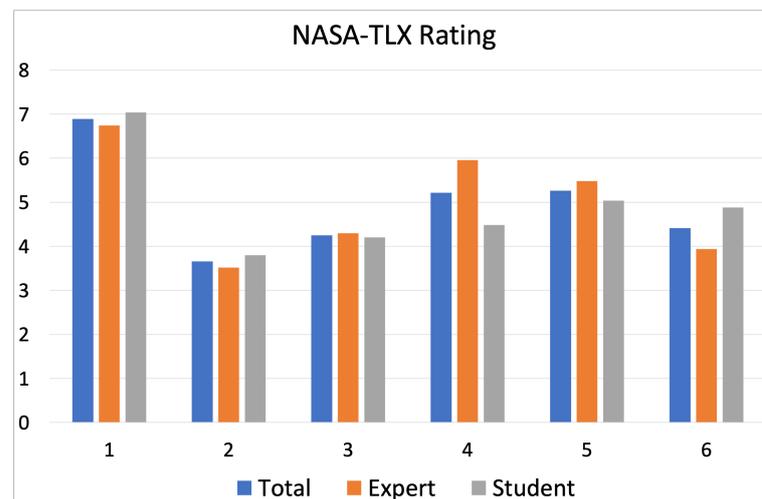
NASA-TLX Rating Results

There are two ways of analysing the NASA-TLX scores: one is a two-step process that needs participants to give both scores for each item and a pairwise comparison score between each pair of items (there will be in total 15 pairwise comparisons); another way is simple and convenient—calculating the average score of the six items for each participant [11,12]. In this study, only the "Raw TLX" was used [13] and the average score was calculated in order to keep the experimental validity [12]. When using the "raw TLX", individual subscales may be dropped if less relevant to the task [14].

Figure 3 shows the average scores from the experts and students. Students considered the workload was higher than expected by experts, however, it was not highly significant. Figure 4 shows the comparison of the raw scores for students, experts and overall. The results show that both students and experts felt that the task was low in physical demand and temporal demand, while it was a highly mentally demanding. There is not much difference between the students and experts in the rating results.
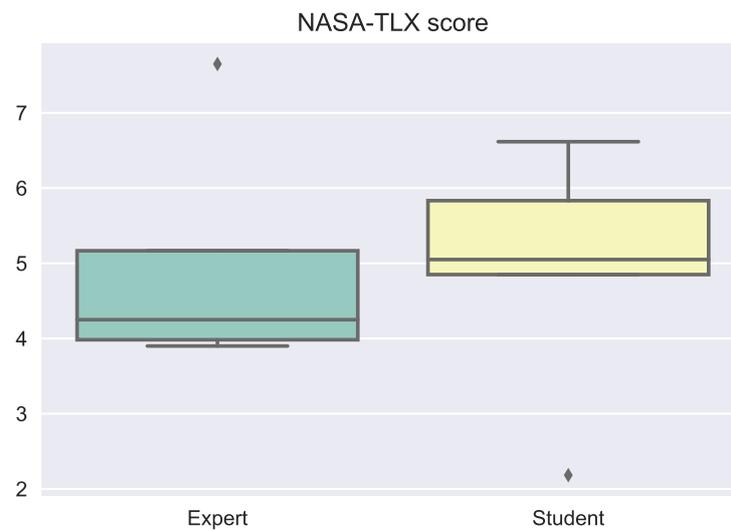
Figure 3. Comparison of the NASA-TLX rating raw score for students, experts.

Figure 4. Comparison of the NASA-TLX rating results for students, experts and overall. The number from 1 to 6 presents Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration Level, respectively.

3.2. Data Pre-Processing

The dataset consists of four signal channels associated with EDA, body temperature (Temp.), BVP and HR. EDA is collected from electrodermal activity sensor measured in μS with frequency of 4 Hz. Body temperature data are measured on the Celsius (°C) scale with a frequency of 4 Hz [15]. BVP data are from photoplethysmograph (PPG) and sampled at 64 Hz [15]. HR data are the average heart rate values per second, derived directly from the BVP analysis [16]. All the signals were downsampled to 1 Hz for data analyses. The data are associated with ten participants, each participant has four sailing sections, the data can be split into forty samples (see Table 1). In order to compare the data with different resolutions, and to make it easier to analyse, we have normalized the downsampled data.

Table 1. The form of the data samples.

No. of Samples	No. of Participants	No. of Sailing Sections	Signal Channel 1	Signal Channel 2	Signal Channel 3	Signal Channel 4
Sample 1		Sailing section 1	EDA data	Temp. data	BVP data	HR data
Sample 2	Participant 1	Sailing section 2	..	..	..	..
.		Sailing section 3	..	..	..	..
.		Sailing section 4	..	..	..	..
.						
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
		Sailing section 1	..	..	..	..
	Participant 10	Sailing section 2	..	..	..	..
Sample 39		Sailing section 3	..	..	..	..
Sample 40		Sailing section 4	EDA data	Temp. data	BVP data	HR data

Normalization of the data is done by calculating the standard normal distribution. The standard normal distribution is the simplest case of the normal distribution when the data are standardized to have a mean of zero and a standard deviation of one [17]. Before calculating the standard normal distribution, the standardized value of the signal data is computed from the mean and standard deviation using the following formula (see Equation (1)) [18]:

$$z_i = \frac{x_i - \bar{x}}{S}, \quad i = 1, 2, \dots, n \tag{1}$$

where $z_i = \{z_1, z_2, \dots, z_n\}$ is standardization value of the sample, $x_i = \{x_1, x_2, \dots, x_n\}$ is the value of the downsampled signals, $\bar{x}$ is the mean of x_i on each file, S is the sample standard deviation, n is the number of the data on each file.

Finally, the standard normal distribution of the data was calculated using Equation (2) [19]:

$$f(z_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}}, \quad i = 1, 2, \dots, n \tag{2}$$

where $z_i = \{z_1, z_2, \ldots, z_n\}$ is standardization value of the sample and n is the number of the data.

All 40 samples were labelled into two categories: expert and novice.

3.3. Classification Features Extraction

The normalized signal data were split into 40 samples as we mentioned in Section 3.2. Next, the features vectors (FVs) were collected from each sample. Since each sample of data has four signal channels, the total number of rows of the data is 160. The FVs include statistical-based feature vectors (SFV), wavelet-based feature vectors (WFV), and higher-order crossings (HOC)-based feature vectors.

3.3.1. Statistical-Based Features

For the statistical-based feature vectors (SFV), for each signal channel in each data sample, we calculated five types of vectors. These are mean vectors, the standard deviation vectors, variance vectors, median skewness vectors and kurtosis vectors. All SFVs were collected by the statistic feature vectors, i.e., (see Equation (3)):

$$SFV = [\mu_{SFV}, \sigma_{SFV}, Var, Sk_{SFV}, Kur_{SFV}] \tag{3}$$

where μ_{SFV} are the mean vectors, σ_{SFV} are the standard deviation vectors, Var are variance vectors, Sk_{SFV} are the median skewness vectors, Kur_{SFV} are kurtosis vectors.

1. Mean vectors

   Mean vectors are a collection of the mean of the normalized sample in each signal channel and defined as :

$$\mu_{SFV} = [\bar{x}(i)_{EDA}, \bar{x}(i)_{temp}, \bar{x}(i)_{BVP}, \bar{x}(i)_{HR}]^T, \quad i = 1, 2, \ldots, N \tag{4}$$

   where $\bar{x}(i)_{EDA}$ is the mean of the EDA data for each sample, $\bar{x}(i)_{temp}$ is the mean of body temperature data for each sample, $\bar{x}(i)_{BVP}$ is the mean of the BVP data for each sample, $\bar{x}(i)_{HR}$ is the mean of heart rate data for each sample, and $N = 40$ is the number of the data samples.

2. Standard deviation vectors

   The standard deviation vectors σ_{SFV} are defined in the following (see Equation (5)):

$$\sigma_{SFV} = [S(i)_{EDA}, S(i)_{temp}, S(i)_{BVP}, S(i)_{HR}]^T, \quad i = 1, 2, \ldots, N \tag{5}$$

   where $S(i)_{EDA}$ is the standard deviation of the EDA data for each sample, $S(i)_{temp}$ is the standard deviation of body temperature data for each sample, $S(i)_{BVP}$ is the standard deviation of the BVP data for each sample, $S(i)_{HR}$ is the standard deviation of heart rate data for each sample, $N = 40$ is the number of data samples.

3. Variance vectors

   Variance is the average of the squared differences from the mean. It is the square of the standard deviation. In a similar manner as above, variance vectors are calculated and represented as Var. It can be calculated using the following (see Equation (6)):

$$Var = \sigma_{SFV}^2 = [S(i)_{EDA}^2, S(i)_{temp}^2, S(i)_{BVP}^2, S(i)_{HR}^2]^T, \quad i = 1, 2, \ldots, N \tag{6}$$

   where σ_{SFV} is the standard deviation of each signal for each sample defined in Equation (5).

4. Median skewness vectors

Median skewness is also called Pearson's second skewness coefficient. It is defined as [20]):

$$Sk_2 = \frac{3(\mu - \nu)}{\sigma} \qquad (7)$$

where Sk_2 is the Pearson's second skewness coefficient, μ is the mean of the data, ν is the median, and σ is the standard deviation of each signal in each file. This formula compares the mean to the median in a precise way and shows how many standard deviations apart they are [21]. The median skewness vectors Sk_{SFV} are the collection of median skewness of each signal channel for each sample:

$$Sk_{SFV} = [Sk_2(i)_{EDA}, Sk_2(i)_{temp}, Sk_2(i)_{BVP}, Sk_2(i)_{HR}]^T, \quad i = 1, 2, \ldots, N \quad (8)$$

where $Sk_2(i)_{EDA}$ is the median skewness of the EDA data for each sample, $Sk_2(i)_{temp}$ is the median skewness of body temperature data for each sample, $Sk_2(i)_{BVP}$ is the median skewness of the BVP data for each sample, $Sk_2(i)_{HR}$ is the median skewness of heart rate data for each sample, $N = 40$ is the number of samples.

5.　Kurtosis vectors

Kurtosis is an important descriptive statistic of data distribution. It describes how much the tails of a distribution differ from the tails of a normal distribution. Is is defined as the fourth central moment divided by the square of the variance [22,23]:

$$Kurtosis(X) = E[(\frac{X - \mu}{\sigma})^4] = \frac{\mu_4}{\sigma^4} \qquad (9)$$

where X is the dataset, $Kurtosis(X)$ is the kurtosis value of the normal distribution for the dataset, μ_4 is the fourth central moment, μ is the mean (defined in Equation (4)), and σ is the standard deviation of each signal in each sample defined in Equation (5).

The kurtosis vectors are the collection of kurtosis of each signal channel for each sample:

$$Kur_{SFV} = [Kur(i)_{EDA}, Kur(i)_{temp}, Kur(i)_{BVP}, Kur(i)_{HR}]^T \quad i = 1, 2, \ldots, N \quad (10)$$

where $Kur(i)_{EDA}$ is the kurtosis of the EDA data for each sample, $Kur(i)_{temp}$ is the kurtosis of body temperature data for each sample, $Kur(i)_{BVP}$ is the kurtosis of the BVP data for each sample, $Kur(i)_{HR}$ is the kurtosis of heart rate data for each sample, $N = 40$ is the number of the samples.

3.3.2. Wavelet-Based Features

Wavelet Transform is a powerful tool for analysing and classifying the time series signal data. Daubechies wavelets was selected because it is the most commonly used set of discrete wavelet transforms [24]. Among the extremal phase wavelet of Daubechies family, db4 wavelet was chosen in this study, where the number 4 refers to the number of vanishing moments [25].

In this study, the signal from each sensor collected during each sailing section was subjected to wavelet decomposition into N levels, and the result of the decomposition is divided into two parts as one set (of the Nth level) of approximation coefficients (cA) and N set (from 1 to N^{th} level) of and detail coefficient (cD). The cA represents low-frequency signal and the cDs represent high-frequency signal. The original signal usually can be decomposed to several levels, and each layer decomposition coefficients are obtained from the previous decomposition. In other words, the original signal S is decomposed into (see Equation (11)):

$$S = cD1 + cD2 + \ldots + cDN + cAN \qquad (11)$$

where S is the dataset, $cD1, cD2, \ldots, cDN$ are high-frequency signal obtained by decomposition from the first layer, the second layer and the N layers respectively, cAN is the low-frequency signal obtained by decomposition of the Nth layer.

In the wavelet decomposition, the greater the gain, the more obvious the performance of the different characteristics of noise and signal is, and the more conducive to the separation of the noise and signal. On the other hand, the greater the number of decomposition levels, the greater the distortion of the reconstructed signal, which affects the final denoising effect to a certain extent. In this study, in order to handle the contradiction and choose an appropriate decomposition level, the highest six values of both cA and cD from the first level decomposition were selected. In addition, the mean, standard deviation, entropy of cD and cA are added into the feature vectors.

3.3.3. Higher-Order Crossings (HOC)-Based Features

The higher-order crossings (HOC) method is also often called zero-crossing and level-crossing method [26]. It counts the number of axis-crossing, i.e., the symbol changes in the dataset. In our dataset, we set the zero mean signal data from each sensor collected from each stop of each participant as a series $\{\vec{Z}\} = \{Z_t, t = 1, 2, \ldots, N\}$. The number of crossing of the horizontal axis, is denoted D_1, and it is the same as the number of sign changes in $\{\vec{Z}\}$ [26].

The higher-order crossings are defined by using the difference operator $\triangledown$, and $\triangledown Z_t$ is defined (see Equation (12) [27]):

$$\triangledown Z_t = Z_t - Z_{t-1} \tag{12}$$

For the second order of the difference, it is:

$$\begin{aligned}
\triangledown^2 Z_t &= \triangledown(\triangledown Z_t) \\
&= \triangledown(Z_t - Z_{t-1}) \\
&= Z_t - 2Z_{t-1} + Z_{t-2}
\end{aligned} \tag{13}$$

Higher orders can be computed in the same manner as above. In general, the kth order difference is (see Equation (14) [28]):

$$\triangledown^{k-1} Z_t = \sum_{i=1}^{k} C_{i-1}^{k-1}(-1)^{i-1} Z_{t+1-i}$$
$$with \quad C_{i-1}^{k-1} = \frac{(k-1)!}{(i-1)!(k-i)!} \tag{14}$$

where $k = 1, 2, \ldots$, and $\triangledown^0$ is the identity.

From $\{\vec{Z}\}$, a binary process, $\{\vec{X}\}$ is defined by (see Equation (15) [27–29]):

$$X_t = \begin{cases} 1, & Z_t \geq 0 \\ 0, & Z_t < 0 \end{cases} \tag{15}$$

Let

$$d_t = \begin{cases} 1, & X_t \neq X_{t-1} \\ 0, & otherwise, \end{cases} \tag{16}$$

where d_t is the indicator. It indicates that there is a symbol change in $\{\vec{X}\}$ when it is 1. The number of crossing of the horizontal axis, D_1, in $\{\vec{Z}\}$ is defined as [27,28]:

$$\begin{aligned}
D_1 &= d_2 + d_3 + \ldots + d_N \\
&= \sum_{t=2}^{N} [X_t - X_{t-1}]^2
\end{aligned}$$

where N is the length of the data.

For the kth order, the count of the symbol changes is:

$$D_k = d_2^{(k)} + d_3^{(k)} + \ldots + d_N^{(k)}$$
$$= \sum_{t=2}^{N} [X_t(k) - X_{t-1}(k)]^2$$

Above all, for the signal channel for each sample, the HOC-based feature vector, FV_{HOC}, is formed as follows (see Equation (17) [28]):

$$FV_{HOC} = [D_1, D_2, \ldots, D_L], \quad 1 < L \leq J \tag{17}$$

where J is the maximum order of the estimated HOC and L is the HOC order chosen in this study.

For our dataset, the FV_{HOC} was extracted from the four signals within a range of order $K = 1, \ldots, 50 (= J)$.

3.4. Deep Learning Model

In this study, deep learning algorithm was applied to classify the data. Convolutional Neural Network (CNN) was the approach employed for classification. Compared with traditional neural networks, the advantage of CNN was obvious, it has fewer parameters to learn for processing high-dimensional data, which helps to accelerate the training speed and reduce the chance of overfitting [30]. The steps of creating and training CNN are described below:

- First, load the dataset and separate the data into training and validation datasets. In this study, 80 percent of the data is used for training and 20 percent for testing. In order to protect against over fitting, cross-validation was applied. Cross-validation is 5-fold.
- Second, define the CNN architecture. For the first layer, the spatial input and output sizes of these convolutional layers are 32-by-32, and the following max pooling layer reduces this to 16-by-16. For the second layer, the spatial input and output sizes of these convolutional layers are 16-by-16, and the following max pooling layer reduces this to 8-by-8. For the next layer, the spatial input and output sizes of these convolutional layers are 8-by-8. The global average pooling layer averages over the 8-by-8 inputs, giving an output of size 1-by-1-by-4 times of initial number of filters. With a global average pooling layer, the final classification output is only sensitive to the total amount of each feature present in the input image, but insensitive to the spatial positions of the features. In the end, add the fully connected layer and the final softmax and classification layers.
- Third, specify the training options. We used Adam (adaptive moment estimation) optimizer, set the maximum number of epochs to 100, mini batch size to 128, and monitored the network accuracy during training by specifying validation data and validation frequency, shuffling the data every epoch, and plotting training progress [31].
- Fourth, train the network using the structure defined by layers, the training data, and the training options.
- Last, predict the labels of the validation data using the trained network, and calculate the final validation accuracy [31].

3.5. Optimization

The performance of CNN depends on an appropriate setting of hyper-parameters, including the batch size, learning rate, activation function, network structure, etc. [32]. Optimizing hyper-parameters yields better behaviour of the training algorithm, since hyper-parameters effect the performance of the training result for the model. Among the techniques of fine-tune machine-learning algorithms, automatic hyper-parameter tun-

ing is an effective and computational power saving method compared to manual grid search. In this process, the next parameter settings is dependent on the performance of previous configurations. Configurations are inferred and decided by the relation between the hyper-parameter settings and model performance [33]. Bayesian optimization for hyper-parameter automatic tuning is one of the frequently used automating tuning hyper-parameters, and we applied it on this dataset for finding a good optimum.

Bayesian optimization approaches use the results of previous configuration performance to constitute a probabilistic model. The probability is the scores given by the hyper-parameters [34]. This model is used as a surrogate function for the objective function for choosing the best hyper-parameters [35]. The surrogate can be easily modeled by Gaussian Process, and a set of hyper-parameters are selected to give the best performance on the surrogate function. These hyper-parameters are applied on the actual objective function. The surrogate model is updated and the previous steps are repeated until it is optimized [35].

Choose Variables to Optimize

In this study, four variables were chosen to optimize using Bayesian optimization, and their search ranges were specified. The four variables are: network section depth, initial learning rate, stochastic gradient descent momentum and L_2 regularization strength. The following is the illustration of the variables:

- Network section depth: Network section depth is the variable which controls the depth of the network.
- Initial learning rate: Select the best initial learning rate.
- Stochastic gradient descent momentum: Momentum adds inertia so that the network can update the parameters more smoothly and reduce the noise inherent in stochastic gradient descent [36].
- L_2 regularization strength: Choose a good value of regularization to prevent overfitting issues.

Optimization variables with properties are as below (see Table 2):

Table 2. Optimizing variable with properties.

Name	Possible Values	Type	Transform	Optimize
Section Depth	[1, 3]	integer	none	1
Initial Learn Rate	[0.01, 1]	real	log	1
Momentum	[0.08, 0.98]	real	none	1
L_2 regularization	$[1.00 \times 10^{-10}, 0.01]$	real	log	1

3.6. Results

The following section presents the results including feature selection and data classification.

3.6.1. Dataset

The dataset consists of statistical based features (5 columns), wavelet-based features (12 columns) and HOC-based features (50 columns). The data collected from 10 participants with 4 sailing sections and 4 different signal channels (EDA, BVP, body temperatures and HR), i.e., the data comprises 160 rows and 67 columns in total.

3.6.2. Feature Selection

In this section, the results of the feature selection and the method are discussed.

From the original dataset, 73 features are extracted, out of which five are statistic-based features, eighteen are wavelet-based features, and fifty are HOC-based features. Among these features, seven different combinations can be made to give different accuracy results. After comparing the results by using the same learning algorithm from a different set of

features or combinations, the best results from the combination of the features will be chosen. The learning algorithms random forest and support vector machine (SVM) were chosen for comparing the results.

Table 3. Results of accuracy by using the random forest and SVM (kernel = 'linear') for different amount of features selection.

Feature Type	Feature Amount	Random Forest	SVM
Statistic-based features	5	0.51 (+/− 0.25)	0.54 (+/− 0.20)
Wavelet-based features	18	0.46 (+/− 0.24)	0.49 (+/− 0.17)
HOC-based features	50	0.64 (+/− 0.30)	0.58 (+/− 0.09)
Statistic-based + Wavelet-based features	23	0.54 (+/− 0.17)	0.53 (+/− 0.17)
Wavelet-based + HOC-based features	68	0.57 (+/− 0.39)	0.59 (+/− 0.06)
Statistic-based + HOC-based features	55	0.61 (+/− 0.40)	0.59 (+/− 0.21)
Statistic-based + Wavelet-based + HOC-based features	73	0.62 (+/− 0.21)	0.61 (+/− 0.09)

The accuracy of different combinations of feature selection by using random forest and SVM is listed on Table 3. It shows that HOC-based features give the highest accuracy by using the random forest algorithm, and combining all the 73 features gives the highest accuracy by using SVM algorithm. When using all of the 73 features, the accuracy by using random forest and SVM are very close to each other and the standard deviation of the accuracy is lowest among all the results. Therefore, selecting all the features to do the analyse will give a stable results. We propose a hybrid CNN approach for the classification task, where a combination of extracted features like statistical, wavelet and HOC features will be used instead of raw biosignals for CNN-based classification of stress level of user during maritime navigation.

3.6.3. Results from the Data Classification

The final result for the training accuracy is calculated by applying deep learning using Bayesian optimization on our optimal model. The results are given in Table 4. For better understanding, the histogram of results is shown in Figure 5 as below. The result shows that by selecting all the features, we got the highest accuracy which is 75.5%. The approximate 95% confidence interval (written as "testError95CI") of the generalization error rate are also given in the table. "testError95CI" is the interval resulting from Equation (18):

$$testError - 1.96 \cdot testErrorSE \leqslant testError95CI \leqslant testError + 1.96 \cdot testErrorSE \qquad (18)$$

where $testErrorSE$ is the standard error.

The standard error is calculated in Equation (19):

$$testErrorSE = \sqrt{\frac{testError \cdot (1 - testError)}{NTest}} \qquad (19)$$

where $Ntest$ is the number of elements in the array of label of testing data.

Table 4. Bayesian Optimization Results for CNN from different feature selections.

Feature Types	Validation Error	testError	testError95CI		Accuracy
Statistic-base + wavelet-based features (method 1)	0.357	0.357	0.193	0.522	64.3%
HOC-base features (method 2)	0.263	0.263	0.119	0.407	73.7%
Statistic-based + Wavelet-based + HOC-based features (method 3)	0.245	0.245	0.101	0.388	75.5%
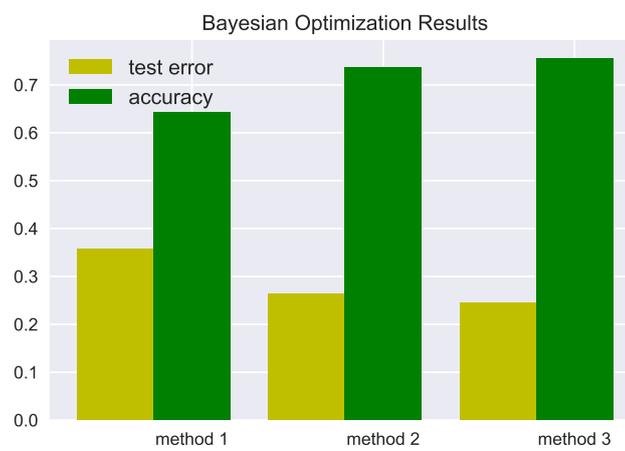
Figure 5. Histogram of Bayesian optimization results for CNN from different feature selections. In this figure, method 1 represents the selection of statistic-based feature and wavelet-based features, method 2 represents the selection of HOC-based features, and method 3 represents the selection of statistic-based feature, wavelet-based features and HOC-based features.

By increasing the number of features by applying CNN Bayesian Optimization, we obtained the highest accuracy. When adding the 50 HOC-based features, the prediction accuracy increased significantly. When using Wavelet-based features, the accuracy was poorer than using other features. Therefore, wavelet-based features do not affect much the results when combined with other features. However, it still gives the best results when adding all of the 73 features. The result presents that feature selection improves the classification accuracy. Nonetheless, we do not claim that it can apply to all types of data, since classification accuracy achieved with different feature reduction strategies is highly sensitive to the type of data [37].

4. Discussion

This study conducted an experiment for finding out the stress differences between experienced experts and students in SA in maritime navigation. The result of data analysis shows that the biosignal data from experts and students can be classified by a certain machine learning algorithm. The result of subjective measurement of workload shows that there is a difference between experts and students. The summary of findings is shown in Table 5.

Table 5. Summary of findings.

Hypothesis Number	Description	Accepted/Rejected
H1	The bio-sensor data from experienced experts can be distinguished from the biosensor data from students	Accepted
H2	The stress levels obtained from the bio-sensor data show a correlation to the NASA-TLX rating results	Accepted
H3	Compared to the novices, the experts feel less stress during a navigation task	Accepted

Based on the results associated with the three hypotheses, we outline the following statements:

Statement 1. Biosignal data are considered as stress monitoring analysis, since stress can be a physical, mental or emotional reaction, and it causes hormonal, respiratory, cardiovascular and nervous system changes. For example, stress can make your heart beat faster, make you breathe rapidly, sweat and tense up. Based on the results of the classification of the biosensor data, we can see that the data from the experts and the students have different patterns. Accuracy of 75.5% is an acceptable result for distinguishing the data. This could have implication for stress difference with maritime navigating skills. Previous research shows that experts obtain better results in SA task [38]. When facing the same task, with different level of skills, the stress level is different.

Statement 2. NASA-TLX contains subjective data which was evaluated by the participants themselves. Results from the NASA-TLX show that experts and students had different evaluation of the workload, which is consistent with the results from classification of biosignal data, i.e., biosignal data show a different pattern for experts and students. This result is also consistent with other research suggesting that mental effort and anxiety are closely related to HRV [39].

Statement 3. The results of NASA-TLX rating show that experts have a smaller workload compared to students. Many research articles show that there is a high correlation between workload and stress. When there is the overload of the work, the stress is increasing [40,41].Therefore, there have implication that experts feel less stress than students. In addition, some other research also shows that higher levels of stress had a negative relationship with work SA [42]. Since experts have better SA score [38], they should be under less stress.

5. Conclusions

In this paper, we propose a deep learning approach using Bayesian optimization for classifying the biosignal data of navigators during the maritime operation. We extracted different types of the features to improve the prediction accuracy. We also compared the objective results to the subjective results, NASA-TLX rating results, that the two results are correlated.

We would like to highlight that number of samples were few in order to make any statistical claims of our findings. Based on our current study in the next step, we plan to experiment further with a larger set of population. Nevertheless, the results of our current data analysis as well as this study will contribute to auto-assessment system for evaluating the SA performance in maritime navigation.

Author Contributions: Conceptualization, H.X., P.S. and B.-M.B.; methodology, H.X., B.-M.B. and D.K.P.; software, H.X. and D.K.P.; validation, H.X., B.-M.B. and D.K.P.; formal analysis, investigation, resources and data curation, H.X., B.-M.B. and D.K.P.; writing—original draft preparation, H.X.; writing—review and editing, H.X., B.-M.B., J.A.J., P.S. and D.K.P.; supervision, B.-M.B. and D.K.P. All authors have read and agreed to the published version of the manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

BVP	Blood volume pulse
CNN	Convolutional Neural Network
EDA	Electrodermal activity
FV	Features vector
HOC	Higher order crossing
HR	Heart rate
NASA-TLX	NASA Task Load Index
SA	Situation awareness
SAGAT	Situation Awareness Global Assessment Technique
SFV	Statistical-base feature vector
SVM	Support vector machine
WFV	Wavelet-based feature vector

References

1. Sanfilippo, F. A multi-sensor fusion framework for improving situational awareness in demanding maritime training. *Reliab. Eng. Syst. Saf.* **2017**, *161*, 12–24. [CrossRef]
2. Marchenko, N.; Andreassen, N.; Borch, O.J.; Kuznetsova, S.; Ingimundarson, V.; Jakobsen, U. Arctic shipping and risks: Emergency categories and response capacities. *Transnav Int. J. Mar. Navig. Saf. Sea Transp.* **2018**, *12.1*. [CrossRef]
3. Grech, M.R.; Horberry, T.; Smith, A. Human error in maritime operations: Analyses of accident reports using the Leximancer tool. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; Sage Publications Sage CA: Los Angeles, CA, USA, 2002; Volume 46, pp. 1718–1721.
4. TheNauticalInstitute. Situational Awareness-The Sense of Everything, Feb 2020, Issue no. 23. Available online: https://www.nautinst.org/uploads/assets/8f9da438-3486-4b77-a4d79e1af61fe15c/Issue-23-Situational-Awareness.pdf (accessed on 21 August 2020).
5. Endsley, M.R. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society Annual Meeting*; SAGE Publications Sage CA: Los Angeles, CA, USA, 1988; volume 32, pp. 97–101.
6. Endsley, M.R.; Bolte, B.; Jones, D. G. *Designing for Situation Awareness: An Approach to User-Centered Design*; CRC press: Boca Raton, FL, USA, 2003; Chapter 2.
7. Van den Broek, A.; Neef, R.; Hanckmann, P.; van Gosliga, S.P.; Van Halsema, D. Improving maritime situational awareness by fusing sensor information and intelligence. In Proceedings of the 14th International Conference on Information Fusion, Chicago, IL, USA, 5–8 July 2011; pp. 1–8.
8. © 2019 Empatica Inc. E4 Wristband Real-Time Physiological Data Streaming and Visualization. 2020. Available online: https://www.empatica.com/en-int/research/e4/ (accessed on 30 November 2020).
9. Sharek, D. A useable, online NASA-TLX tool. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; SAGE Publications Sage CA: Los Angeles, CA, USA, 2011; Volume 55, pp. 1375–1379.
10. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183.
11. Rubio, S.; Díaz, E.; Martín, J.; Puente, J.M. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Appl. Psychol.* **2004**, *53*, 61–86. [CrossRef]
12. Bustamante, E.A.; Spain, R.D. Measurement invariance of the Nasa TLX. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; SAGE Publications Sage CA: Los Angeles, CA, USA, 2008; Volume 52, pp. 1522–1526.

13. Hart, S.G. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; Sage publications Sage CA: Los Angeles, CA, USA, 2006; Volume 50, pp. 904–908.
14. Colligan, L.; Potts, H.W.; Finn, C.T.; Sinkin, R.A. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *Int. J. Med. Inf.* **2015**, *84*, 469–476. [CrossRef] [PubMed]
15. © 2019 Empatica Inc. Data Export and Formatting from E4 Connect, January 2020. Available online: https://support.empatica.com/hc/en-us/articles/201608896-Data-export-and-formatting-from-E4-connect- (accessed on 12 May 2020).
16. © 2019 Empatica Inc. E4 Data-HR.csv Explanation, January 2020. Available online: https://support.empatica.com/hc/en-us/articles/360029469772-E4-data-HR-csv-explanation (accessed on 12 May 2020).
17. Patel, J.K.; Read, C.B. *Handbook of the Normal Distribution*; Dekker, M., Ed.; International Biometric Society: New York, NY, USA, 1982.
18. Bohm, G.; Zech, G. *Introduction to Statistics and Data Analysis for Physicists*; Desy: Hamburg, Germany, 2010; Volume 1, p.23.
19. Lane, D.M.; Scott, D.; Hebl, M.; Guerra, R.; Osherson, D.; Zimmer, H. *An Introduction to Statistics*; Rice University: Houston, TX, USA, 2017; pp. 261–268.
20. Weisstein, E.W. "Pearson's Skewness Coefficients." From MathWorld–A Wolfram Web Resource, 23 July 2020. Available online: https://mathworld.wolfram.com/PearsonsSkewnessCoefficients.html(accessed on 28 July 2020).
21. Doane, D.P.; Seward, L.E. Measuring skewness: A forgotten statistic? *J. Stat. Educ.* **2011**, *19*. [CrossRef]
22. Kokoska, S.; Zwillinger, D. *CRC Standard Probability and Statistics Tables and Formulae*; CRC Press: Boca Raton, FL, USA, 2000.
23. Pearson, K. IX. Mathematical contributions to the theory of evolution.—XIX. Second supplement to a memoir on skew variation. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character* **1916**, *216*, 429–457.
24. Akansu, A.N.; Haddad, P.A.; Haddad, R.A.; Haddad, P.R. *Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets*; Academic press: Cambridge, MA, USA, 2001.
25. Daubechies, I. CBMS-NSF regional conference series in applied mathematics. *Ten Lect. Wavelets* **1992**, *61*. [CrossRef]
26. Dickstein, P.; Spelt, J.; Sinclair, A. Application of a higher order crossing feature to non-destructive evaluation: A sample demonstration of sensitivity to the condition of adhesive joints. *Ultrasonics* **1991**, *29*, 355–365. [CrossRef]
27. Kedem, B. Higher-order crossings in time series model identification. *Technometrics* **1987**, *29*, 193–204. [CrossRef]
28. Petrantonakis, P.C.; Hadjileontiadis, L.J. Emotion recognition from EEG using higher order crossings. *IEEE Trans. Inf. Technol. Biomed.* **2009**, *14*, 186–197. [CrossRef] [PubMed]
29. Kedem, B.; Yakowitz, S. *Time Series Analysis by Higher Order Crossings*; IEEE Press: New York, NY, USA, 1994; p. 19.
30. Wu, J.; Chen, X.Y.; Zhang, H.; Xiong, L.D.; Lei, H.; Deng, S.H. Hyperparameter optimization for machine learning models based on Bayesian optimization. *J. Electron. Sci. Technol.* **2019**, *17*, 26–40.
31. © 1994–2020 The MathWorks, Inc. Create Simple Deep Learning Network for Classification, 2020. Available online: https://se.mathworks.com/help/deeplearning/ug/create-simple-deep-learning-network-for-classification.html (accessed on 25 November 2020).
32. Sameen, M.I.; Pradhan, B.; Lee, S. Application of convolutional neural networks featuring Bayesian optimization for landslide susceptibility assessment. *Catena* **2020**, *186*, 104249. [CrossRef]
33. Pham, V. Bayesian Optimization for Hyperparameter Tuning, 2016. Available online: https://arimo.com/data-science/2016/bayesian-optimization-hyperparameter-tuning/ (accessed on 5 May 2020).
34. Dewancker, I.; McCourt, M.; Clark, S. Bayesian optimization for machine learning: A practical guidebook. *arXiv* **2016**, arXiv:1612.04858.
35. Koehrsen, W. A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning, 2018. Available online: https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f (accessed on 5 May 2020).
36. © 1994–2020 The MathWorks, Inc. Deep Learning Using Bayesian Optimization, 2020. Available online: https://se.mathworks.com/help/deeplearning/ug/deep-learning-using-bayesian-optimization.html (accessed on 8 May 2020).
37. Janecek, A.; Gansterer, W.; Demel, M.; Ecker, G. On the relationship between feature selection and classification accuracy. In *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*; PMLR: McKees Rocks, PA, USA, 2008; pp. 90–105.
38. Xue, H.; Batalden, B.M.; Røds, J.F. Development of a SAGAT Query and Simulator Experiment to Measure Situation Awareness in Maritime Navigation. In Proceedings of the International Conference on Applied Human Factors and Ergonomics, New York, NY, USA, 24–28 July 2020; Springer: Cham, Switzerland, 2020; pp. 468–474.
39. Thayer, J.F.; Friedman, B.H.; Borkovec, T.D. Autonomic characteristics of generalized anxiety disorder and worry. *Biol. Psychiatry* **1996**, *39*, 255–266. [CrossRef]
40. Lundberg, U.; Hellström, B. Workload and morning salivary cortisol in women. *Work. Stress* **2002**, *16*, 356–363. [CrossRef]
41. Steptoe, A.; Cropley, M.; Griffith, J.; Kirschbaum, C. Job strain and anger expression predict early morning elevations in salivary cortisol. *Psychosom. Med.* **2000**, *62*, 286–292. [CrossRef] [PubMed]
42. Sneddon, A.; Mearns, K.; Flin, R. Stress, fatigue, situation awareness and safety in offshore drilling crews. *Saf. Sci.* **2013**, *56*, 80–88. [CrossRef]