sensors

MDPI

Article

Heart Rate Variability-Based Subjective Physical Fatigue Assessment

Zhiqiang Ni [1,2], Fangmin Sun [1] and Ye Li [1,*]

[1] Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; zq.ni@siat.ac.cn (Z.N.); fm.sun@siat.ac.cn (F.S.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: ye.li@siat.ac.cn

Abstract: Accurate assessment of physical fatigue is crucial to preventing physical injury caused by excessive exercise, overtraining during daily exercise and professional sports training. However, as a subjective feeling of an individual, physical fatigue is difficult for others to objectively evaluate. Heart rate variability (HRV), which is derived from electrocardiograms (ECG) and controlled by the autonomic nervous system, has been demonstrated to be a promising indicator for physical fatigue estimation. In this paper, we propose a novel method for the automatic and objective classification of physical fatigue based on HRV. First, a total of 24 HRV features were calculated. Then, a feature selection method was proposed to remove useless features that have a low correlation with physical fatigue and redundant features that have a high correlation with the selected features. After feature selection, the best 11 features were selected and were finally used for physical fatigue classifying. Four machine learning algorithms were trained to classify fatigue using the selected features. The experimental results indicate that the model trained using the selected 11 features could classify physical fatigue with high accuracy. More importantly, these selected features could provide important information regarding the identification of physical fatigue.

Keywords: heart rate variability; physical fatigue; feature selection; machine learning

Citation: Ni, Z.; Sun, F.; Li, Y. Heart Rate Variability-Based Subjective Physical Fatigue Assessment. *Sensors* **2022**, *22*, 3199. https://doi.org/10.3390/s22093199

Academic Editor: Andrea Facchinetti

Received: 21 March 2022
Accepted: 20 April 2022
Published: 21 April 2022

1. Introduction

With the improvement of human living standards, more and more people realize the importance of exercise to health and engage in regular physical exercise to keep healthy. However, proper exercise is good for health, while excessive exercise may bring harm to the body(e.g., muscular and skeletal injuries [1], overtraining syndrome [2], atrial fibrillation [3] and immune function reduction). Excessive exercise is defined as a relative term, implying that a bout(s) of exercise is fine for some individuals while excessive for other individuals due to the differences in a variety of factors, such as physical fitness and genetics [4]. Physical fatigue, as a common physiological phenomenon during exercise, is a direct reflection of the degree of exercise. Accurate detection and evaluation of physical fatigue levels can effectively prevent excessive exercise and further reduce physical injury caused by excessive exercise.

Fatigue is a term used to describe a subjective feeling of tiredness or lack of energy. Objectively evaluating fatigue is a challenging task as a person's subjective feelings cannot be easily assessed by other people [5]. The rating of perceived exertion (RPE) [6] is a measure of fatigue that has been widely used for fatigue assessment in sport science, specifically in running research [7–9]. As the RPE has advantages, such as being noninvasive, unobtrusive, noninterruptive and easy to use, many previous studies on fatigue assessment have used it as the ground truth. Moreover, previous studies [10,11] indicated that the RPE represented feedback from cardiovascular, respiratory and musculoskeletal systems and it provides an overall fatigue state assessment of a subject, while a single biomechanical or physiological

parameter usually provides very limited information. So, in this study we collected the RPE of subjects at each experiment stage and used it as the physical fatigue ground truth for model training and testing.

With the development of wearable devices [12], more and more physiological parameters could be easily collected. Many researchers have been trying to use various physiological parameters collected by wearable devices to assess physical fatigue. One of the physiological signals for physical fatigue detection is an electromyogram (EMG) [13–16]. EMGs can reflect the electrical activity of local muscle, which is related to the physical fatigue of the local muscle. Physical fatigue not only reduces the body's exercise ability but also causes neurological function decline. So, electroencephalograms (EEG) are another physiological index used for physical fatigue assessment [17,18]. Unfortunately, both EMGs and EEGs are weak bioelectric signals and are easily disturbed by many kinds of noise. Additionally, their acquisitions require professional operations to paste the acquisition electrodes to relevant body parts, so they are not widely used.

Heart rate variability (HRV), a tiny time variation between adjacent heartbeats, was demonstrated to have a relationship with autonomic nervous activities [19]. In recent years, machine learning algorithms based on HRV signals have become a research hotspot in various applications, such as noise detection [20], cuff-less blood pressure measurement [21], mental fatigue evaluation [22] and exercise-induced physical fatigue evaluation [23]. Compared with exercise-induced physical fatigue evaluation, there have been more achievements in mental fatigue evaluation. For HRV-based mental fatigue evaluation, one research team used the kernel principal component method to select important HRV features which have a strong relationship with fatigue states [24]. Their study results show that selected features can easily distinguish between normal samples and fatigue samples. Another research team concentrated on using neural networks and HRV analysis with a power spectral density algorithm to build a driver fatigue detection model, and an accuracy of 90% was achieved [25]. Moreover, decision trees, support vector machines and K-nearest neighbor classifiers have been also proposed to quantify mental fatigue [26].

As a comparison, a few recent efforts have been made towards HRV-based physical fatigue assessment. Ramos et al. [27] combined EMG features and HRV features to build a binary SVM classifier. By analyzing HRV from blood volume pulse signals, Cosoli et al. [28] evaluated the performance of two machine learning algorithms in distinguishing between nonfatigue and fatigue conditions and presented a fatigue-related index to quantify the physical fatigue. Guan et al. [29] proposed a bidirectional long- and short-term memory neural network to classify physical fatigue. The model used HRV features and inertial sensor signals as inputs and achieved 80.55% accuracy. Nevertheless, most of these models either used multiple kinds of signals or coarsely classified physical fatigue into two levels. Therefore, the purpose of this study was to investigate whether the machine learning method combined with HRV, which has been proven to be useful in mental fatigue assessment, is also effective for continuous and real-time monitoring of physical fatigue during exercise. Furthermore, the study also aimed to investigate which HRV features were the most significant in the classification. The selection and analysis of relevant features may also be important for improving the interpretability of the physical fatigue assessment model.

Based on the results of previous studies and the advancements in machine learning technology, we proposed a novel method for the automatic and objective classification of physical fatigue. First, we proposed a feature selection method to remove useless HRV features that have a low correlation with physical fatigue and redundant HRV features that have a high correlation with the selected features. Then, four machine learning algorithms were trained to classify fatigue using the selected features. Experimental results for 80 healthy subjects indicate that the model trained using the selected features could classify physical fatigue with a high accuracy of 85.5%.

The remainder of this paper is organized as follows. Section 2 introduces the data collection experiment and the physical fatigue evaluation modeling methods. Section 3 provides the physical-fatigue-related HRV feature selection results and the physical fatigue

classification results. The obtained fatigue classification results with different machine learning methods are presented and discussed in Section 4. Finally, the study is concluded in Section 5.

2. Materials and Methods

2.1. Data Collection

A total of 80 healthy subjects were recruited for participation in the data collection experiments; the statistical anthropometric characteristics of all subjects are summarized in Table 1. The subjects were asked to perform a preset treadmill exercise, which was modified from the Bruce protocol [30], and during the test, their ECG data were collected. The experiment process is shown in Table 2; it started with a 5 min pre-rest, during which the subjects were asked to stand still on the treadmill. Following this was the exercise stage, during which the subjects began to run at a speed of 3 km/h, and the speed increased to the next preset value every 5 min until reaching the maximum preset speed, and the subjects would run at this maximum preset speed until they were physically exhausted. It was not necessary to reach the maximum speed during the exercise stage, and the exercise could be terminated at any time the participant signaled that he was exhausted.

Table 1. Participants' statistical characteristics.

Statistical Characteristic	Value
Number of Subjects	80 (42 Males, 38 Females)
Age (years)	29.1 ± 6.5
Height (cm)	168.0 ± 8.1
Weight (kg)	61.7 ± 11.2

Table 2. Protocol of the modified Bruce treadmill test.

Stage	Duration (min)	Speed (km/h)	Incline (%)
Pre-rest	5	0	0
Ex-1	5	3	5
Ex-2	5	5	5
Ex-3	5	6.4	5
Ex-4	5	7.8	5
Ex-5	5	10.2	5
Ex-6	Until exhausted	11.6	5

The data collection scenario is shown in Figure 1. The subjects were asked to wear a 12-lead ECG device (GE Medical System Information Technologies, INC, CardioSoft Cardiac Testing System). The ECG device used in our study was specially designed for exercise ECG monitoring, and hardware anti-noise design and software filtering algorithms were made to ensure the high quality of the collected exercise ECG signal. In addition, we further compared and analyzed the quality of 12-lead ECG signals, and finally selected a V3-lead ECG signal, which had the best signal quality for subsequent HRV extraction and fatigue evaluation.

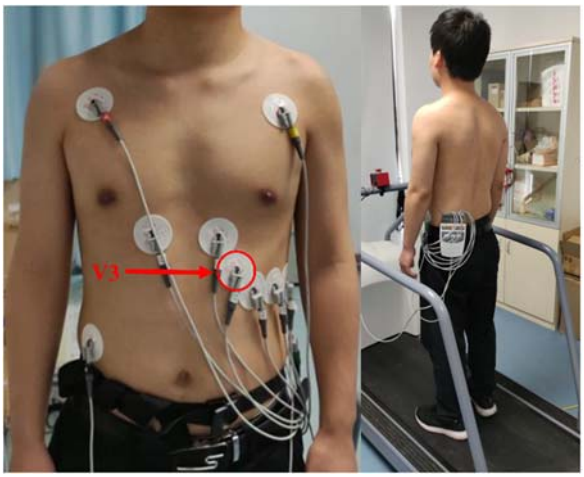

Figure 1. The scenario of the data collection experiment.

The RPE scale ranged from 6 to 20, and the subject was free to choose any integer value within this range. Prior to starting the run, we explained the RPE scale listed in Table 3 to each subject. During the running experiments, the fatigue states of subjects were recorded with the RPE. At the end of each stage, the subjects were asked to report their RPE. Three classes were defined based on the RPE values: (i) "Rested" for values from 6 to 10; (ii) "A bit tired" for values from 11 to 16; and (iii) "Tired" for values from 17 to 20.

Table 3. The RPE scale and its description.

Borg Rating	Description
6	Nothing
7 to 8	Very, very light
9 to 10	Very light
11 to 12	Fairly light
13 to 14	Somewhat hard
15 to 16	Hard
17 to 18	Very hard
19 to 20	Very, very hard

Each participant participated in at least 1 session and at most 3 sessions (with an interval of 1 week) of data acquisition experiments. Each session lasted 20 min to 60 min, and a total of 207 sessions were collected. Figure 2 shows the fatigue state distribution of the dataset collected in the 207 sessions of experiments. It can be seen that the perception of tiredness had individual differences. Under the same exercise intensity and time, e.g., at the EX-3 stage, participants reported being "Rested" in 20 sessions, "A bit tired" in 128 sessions and "Tired" in 59 sessions. In addition, all 80 subjects finished the first 3 exercise stages (to EX-3 stage), while from the Ex-4 stage onwards, there were subjects who stopped the exercise because of exhaustion.

The study was approved by the Institutional Review Board of Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. All subjects signed their written informed consent before the experiments.

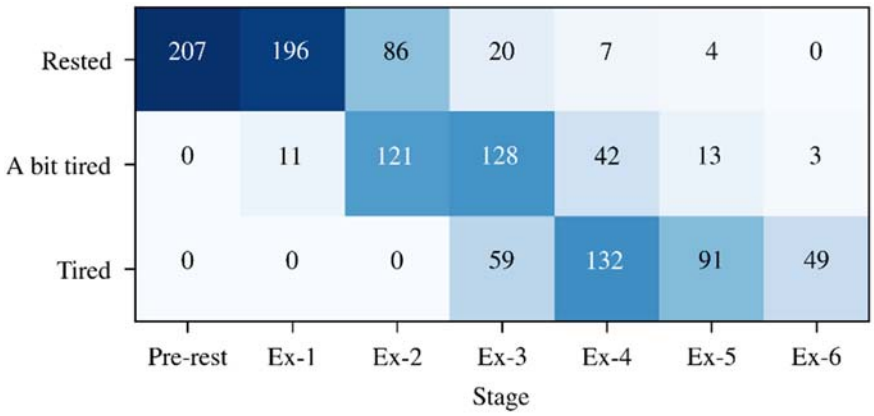

Figure 2. The distribution of the collected dataset.

2.2. Preprocessing and Feature Extraction

The sampling rate of ECG data was 200 Hz. In order to eliminate electronic noise and motion artifacts, the sampled signals were preprocessed using a Butterworth low-pass filter with a cutoff frequency of 50 Hz and a nine-level wavelet decomposition with the order 8 Daubechies wavelet. As the guideline [31] recommended that the ECG records used for HRV analysis should last for at least 5 min, we segmented the raw ECG with a 5 min sliding window without overlap and extracted HRV features from each segment.

Over a specific time period, time domain features are just measurements of the mean and variability in the time interval between heartbeats, which is alternately referred to as normal-to-normal intervals (NN). The common time domain features include the mean of NN interval sequence (meanNN), the mean of heart rate sequence (meanHR), the standard deviation of NN interval sequence (SDNN) and the root mean square of successive differences in NN interval sequence (RMSSD). Another two features calculated by successive differences in NN interval sequence are the number of these differences greater than 50 ms (NN50) and the percentage of NN50 in total intervals (pNN50). By segmenting the long NN interval sequence into several nonoverlapping chunks with a chosen time window (1 min in this work), two types of HRV features, including the standard deviation of the averages of segmented chunks (SDANN) and the average of the standard deviations of segmented chunks (SDNNi), can be calculated. In addition to these statistical features, there are two geometric HRV features based on the NN interval sequence histogram with bins of 1/128 s. The HRV triangular index (HRVTi) is the ratio of the total number of all intervals to the height of the histogram. Additionally, the triangular interpolation of the histogram (TINN) is the baseline width of the minimum square difference triangular interpolation of the highest peak of the histogram.

For HRV frequency domain analysis, power spectrum density (PSD) was computed using the Lomb–Scargle method. Three main components were derived from the heart rate power spectrum, namely the very low frequency band (VLF) ranging between 0.0033 Hz and 0.04 Hz, the low = frequency band (LF) ranging between 0.04 Hz and 0.15 Hz and the high-frequency band (HF) ranging between 0.15 Hz and 0.4 Hz. The VLF component has been reported to be associated with arrhythmic death [32] and high inflammation [33]. The LF component appears to be sensitive to both sympathetic and parasympathetic activities, whereas the HF component is primarily mediated by the parasympathetic nervous activity [34]. Therefore, the LF/HF ratio has been regarded as a measure of physical workload and stress [35].

Aimed at the nonlinearity of heart rate signal, a number of nonlinear techniques have been applied to HRV analysis, which was thought to be an effective way to describe the changes in the biological signal. Three nonlinear methods were used for HRV analysis in this work, namely sample entropy (sampen), Poincare plot (SD1, SD2 and SD1/SD2) and detrended fluctuation analysis (α, α_1 and α_2).

Using MATLAB R2018a with the help of Physionet Cardiovascular Signal toolbox [36], a total of 24 features, including 10 time domain features, 7 frequency domain features and

7 nonlinear features, were computed for further processing. All the obtained HRV features are listed in Table 4.

Table 4. All HRV features.

Measures	Feature	Unit	Description
Time domain	meanNN	ms	Mean of NN interval sequence.
	meanHR	1/min	Mean of heart rate sequence.
	SDNN	ms	Standard deviation of NN interval sequence.
	RMSSD	ms	Root mean square of successive differences in NN interval sequence.
	NN50	count	Number of successive differences in NN interval sequences greater than 50 ms.
	pNN50	%	Percentage of NN50 in total intervals.
	SDANN	ms	Standard deviation of the averages of the segmented chunks.
	SDNNi	ms	Average of the standard deviations of the segmented chunks.
	HRVTi	-	Ratio of total number of all intervals to the height of the histogram.
	TINN	ms	Baseline width of the minimum square difference triangular interpolation of the highest peak of the histogram.
Frequency domain	aVLF	ms^2	Absolute powers of VLF band.
	aLF	ms^2	Absolute powers of LF band.
	aHF	ms^2	Absolute powers of HF band.
	LF/HF	-	Ratio of aLF/aHF.
	peakVLF	Hz	Peak frequency for VLF band.
	peakLF	Hz	Peak frequency for LF band.
	peakHF	Hz	Peak frequency for HF band.
Nonlinear domain	sampen	-	Negative natural logarithm of the conditional probability that two sequences remain similar at the next point.
	SD1	ms	Standard deviations along the major axis of the ellipse.
	SD2	ms	Standard deviations along the minor axis of the ellipse.
	SD1/SD2	-	Ratio of SD1 to SD2.
	α	-	Slope of a fitting line of the root mean square fluctuation of an integrated and detrended time series on a log–log scale.
	α_1	-	α on first linear region.
	α_2	-	α on second linear region.

2.3. Feature Selection

Feature selection is essential for training an effective model. There is no doubt that any unnecessary features, including unimportant features and redundant features, will increase the computational cost of model training and the risk of model overfitting, decrease the interpretability of the model and reduce the generalization performance of the model on the test set. Thus, dropping features with a weak correlation with physical fatigue (unimportant features) and removing highly redundant features are two steps needed during feature selection. The proposed feature selection method is described in Algorithm 1.

Algorithm 1 Feature selection.

Input: Original features Φ, input data $\{(X_1, y_1), \ldots, (X_N, y_N)\}$
Output: Selected features Φ'
Initialize thresholds r_1, r_2
Initialize number of repeats T
for f in Φ **do**
  Compute actual Gini importance $I_{\{f\}}$ from $\{(X_1, y_1), \ldots, (X_N, y_N)\}$ according to Equation (1)
end for
for $i = 1 \rightarrow T$ **do**
  Shuffle the labels $y_1, \ldots, y_N$, which is referred to as $y'_1, \ldots, y'_N$
  for f in Φ **do**
    Compute new Gini importance of $D^i_{\{f\}}$ from $\{(X_1, y'_1), \ldots, (X_N, y'_N)\}$ according to Equation (1)
  end for
end for
for f in Φ **do**
    Compute score of the feature according to Equation (4), which is referred to as $S_{\{f\}}$
end for
Select the features with a score lower than r_1, which is referred to as Φ_1
Delete the features Φ_1 from Φ
Compute the correlation matrix of features Σ according to Equation (5)
Select the features with less actual Gini importance in each pair of features with a correlation above r_2 which is referred to as Φ_2
Delete the features Φ_2 from Φ
Obtain the remaining features in Φ, which is referred to as selected features Φ'

2.3.1. Dropping Unimportant Features

Dropping low-correlated features requires the scores of all the features at first. The method based on feature importance of random forest (RF) and permutation importance [37] was proposed to score the importance of features. RF provides a Gini importance for the assessment of feature importance. Suppose that we have an input dataset with N instances $\{(X_1, y_1), \ldots, (X_N, y_N)\}$ where each $X_i = \left\{x_{\{f_1\}}, \ldots, x_{\{f_m\}}\right\}$ is a vector with m features and y_i is the corresponding label. First, RF uses the dataset to establish its model. Then, Gini importance of feature f is defined as the sum of the impurity improvement of all the nodes n in all trees S using the feature during the training phase, according to Equation (1):

$$I_{\{f\}} = \sum_S \sum_n \Delta\text{Gini}(n, S) \tag{1}$$

The decrease in Gini impurity resulting from optimal split $\Delta\text{Gini}(n)$ is defined as

$$\Delta\text{Gini}(n) = \text{Gini}(n) - p_l \Delta\text{Gini}(n_l) - p_r \Delta\text{Gini}(n_r) \tag{2}$$

$$\text{Gini}(n) = 1 - \sum_{k=1}^{K} p_{n,k}^2 \tag{3}$$

where $\text{Gini}(n)$ denotes Gini impurity at the node n; n_l and n_r denote the child nodes of n; p_l and p_r denote the ratio of the child nodes' sample size to the total sample size; and $p_{n,k}$ denotes the ratio of class $k = \{0, 1, \ldots, K\}$ in node n.

Unlike the common permutation importance, the label rather than the feature was permuted in this method. After shuffling the labels for the first time, the new dataset can be expressed as $\{(X_1, y'_1), \ldots, (X_N, y'_N)\}$, where $y'_1, \ldots, y'_N$ is a permutation of the actual labels. After training the RF model with the new dataset, we could compute the Gini importance of feature f, which is referred to as $D^1_{\{f\}}$. By repeating T times on permutation of labels at random, the null importance distributions $D_{\{f\}} = \left\{D^1_{\{f\}}, \ldots, D^T_{\{f\}}\right\}$ of various features were created to demonstrate how the model can make sense of a feature disregarding the original labels.

Randomly reordering labels could reduce the Gini importance of all features, because the input data no longer correspond to the real labels obtained in the real world. If the model relies heavily on a feature in its prediction, its importance is particularly affected. Thus, a metric to score a feature is calculating the percentage of the feature's null importance distribution as less than the actual importance. The formula for calculating this score of the feature f is given by Equation (4).

$$\text{score} = \frac{\text{count}(D_{\{f\}} < I_{\{f\}})}{T} \times 100\% \tag{4}$$

where count() denotes counting the elements that meet the criteria, $D_{\{f\}}$ is a set of null importance distributions of f with the number of repeats of T, and $I_{\{f\}}$ is the actual Gini importance of f.

From another viewpoint, the score is a kind of quantitative index adapted from the original feature importance. Compared with the original Gini importance, the score is more effective at selecting features with high importance. The features with a score lower than the threshold r_1 will be dropped as "unimportant features", which contribute little to the model.

2.3.2. Removing Redundant Features

The Pearson correlation coefficient was implemented to compute the correlation matrix and measure the redundancy between two features. Pearson correlation coefficient r provides an indicator to quantitatively evaluate the linear correlation between two variables [38]. It has a value ranging from -1 to +1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and +1 indicates a perfect positive linear relationship. The closer the absolute value of r to 1, the stronger the correlation. Given two variables, X and Y, the Pearson correlation coefficient r between X and Y is defined as Equation (5):

$$\text{r} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}} \tag{5}$$

where N is the number of the variable, X_i and Y_i are the values of X and Y for the i_{th} individual, and $\overline{X}$ and $\overline{Y}$ are the averages of X and Y.

In consequence, the larger the absolute value of the correlation coefficient between two features, the higher the mutual substitutability and the redundancy of the two features, and vice versa. As for each pair of features whose correlation is higher than a threshold r_2, the less important one will be regarded as the "redundant feature" and removed.

2.4. Physical Fatigue Classification

The last step is using machine learning algorithms to accurately classify physical fatigue levels. The classification model adopted four supervised machine learning algorithms, namely decision tree (DT), support vector machine (SVM), K-nearest neighbor (KNN) and light gradient boosting machine (LightGBM). These classification models were trained with the best features obtained by the feature selection method described in Section 2.3.

The performance metrics of the evaluation model were accuracy (ACC), precision, recall and F1 score (F1), and their definitions are listed in Equations (6)–(9)

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$F1 = \frac{2 \times \text{Precison} \times \text{Recall}}{\text{Precison} + \text{Recall}} \quad (9)$$

where TP refers to the number of correctly classified samples in a certain class, FP refers to the number of samples misclassified as a certain class when they belong to other classes, TN refers to the number of correctly classified samples in other classes, and FN refers to the number of samples belonging to a certain class that was misclassified as other classes. The average of these metrics among classes was calculated to obtain a final evaluation of the model's performance.

The 10-fold cross validation method was used to evaluate the performance of these models. In order to prevent a subject's data from being used partly for training and partly for testing, each iteration was trained with the data of 72 subjects and tested with the data of the remaining 8 subjects. The average performance of the 10 iterations was used as the final result.

3. Results

3.1. Optimal Feature Set

Using Equation (1), the scores of the 31 original features were calculated (t = 100) and are shown in Figure 3. When setting the threshold r_1 to 0, we selected 10 unimportant features and removed them. The remaining 14 important features include meanNN, meanHR, SDNN, NN50, pNN50, SDANN, HRVTi, TINN, aVLF, sampen, SD2, SD1/SD2, α and α1.

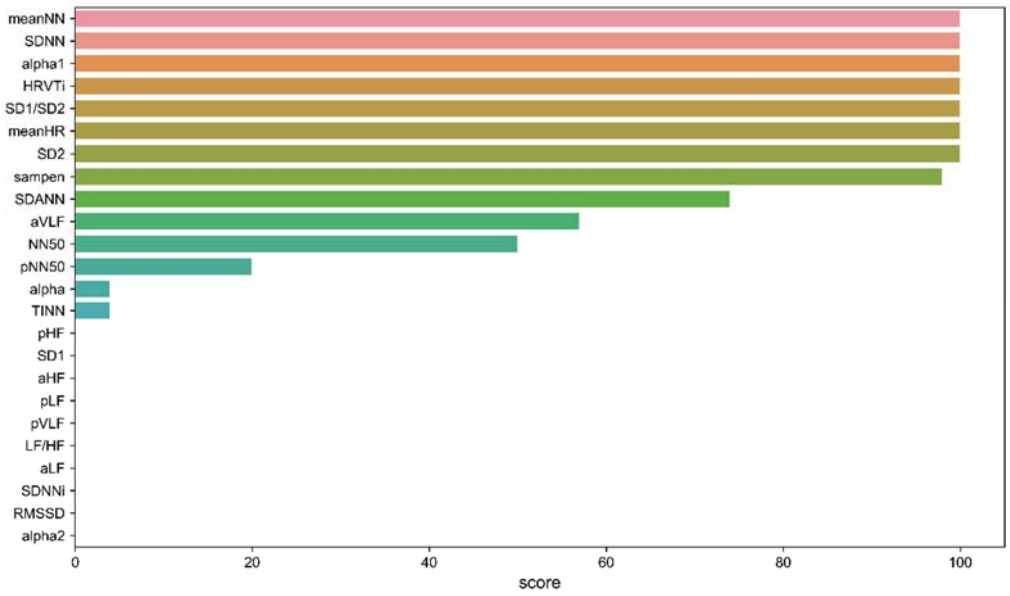

Figure 3. The scores calculated by Equation (4) of all the 24 original features.

Figure 4 shows the correlations between 14 important features. As we can see, there were three pairs of features (meanNN and meanHR, pNN50 and NN50, SD2 and SDNN) with correlation magnitudes greater than the threshold r_2 which was set to 0.9. Then, we removed the redundant features, including meanNN, pNN50 and SDNN, which were least important features in each pair.

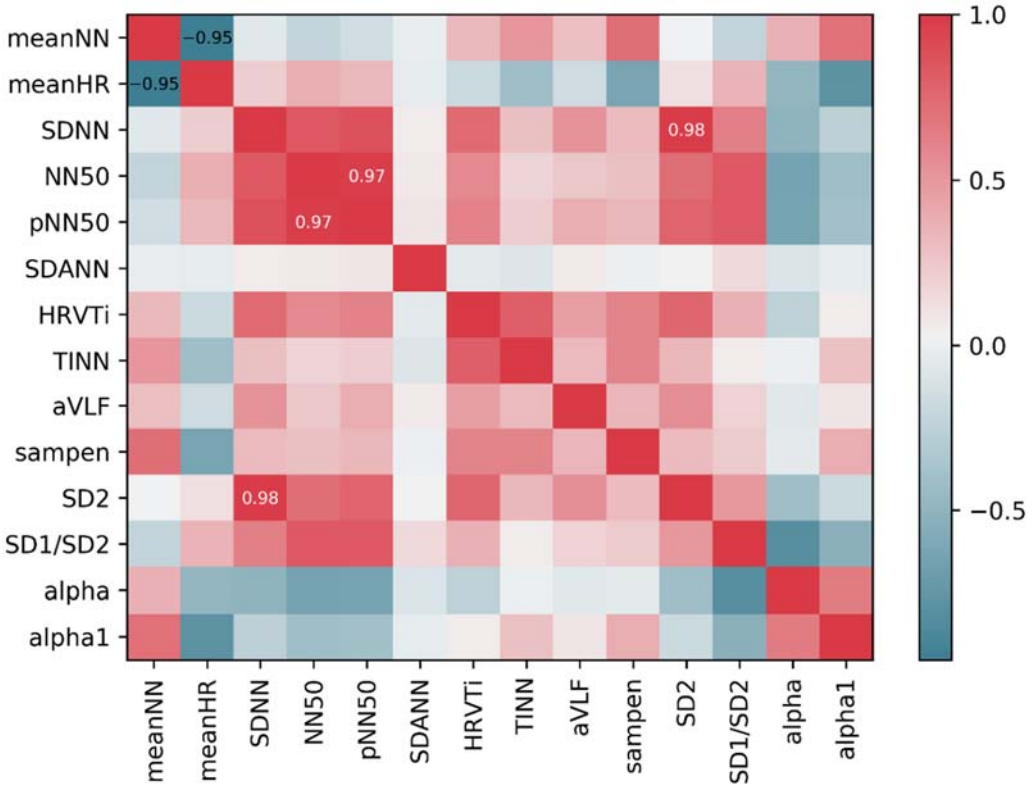

Figure 4. Correlations between 14 important features.

After the feature selection process, a total of 11 features with high importance and low redundancy were finally selected, resulting in an optimal feature set for modeling. The selected 11 features are given in Table 5.

Table 5. Optimal feature set.

Time Domain	Frequency Domain	Nonlinear Domain
meanHR	aVLF	sampen
NN50		SD2
SDANN		SD1/SD2
HRVTi		α
TINN		α_1

3.2. Classification Performance

Table 6 shows the average accuracy, precision, recall and F1 score of four machine learning models using different features in assessing physical fatigue. On the one hand, the average performance of models using selected features was superior to the performance of models using all features. For the DT, SVM, KNN and LightGBM models trained with selected HRV features, the average F1 score increased by 6.3%, 4.0%, 2.2% and 2.0%, respectively, when compared with corresponding models trained with all HRV features. On the other hand, the standard deviations of the models trained with selected HRV features were reduced, which means the models were more stable when the selected features were used. Therefore, it can be seen that both the performance and the stability of the models were increased by the selected features, which verifies the effectiveness of our proposed feature selection method.

Table 6. Performance of the four machine learning models using different features.

Model	Using All Features				Using Selected Features			
	Accuracy	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
DT	0.728 ± 0.043	0.646 ± 0.062	0.644 ± 0.053	0.642 ± 0.056	0.772 ± 0.030	0.711 ± 0.036	0.706 ± 0.034	0.705 ± 0.034
KNN	0.780 ± 0.045	0.712 ± 0.044	0.694 ± 0.041	0.696 ± 0.044	0.805 ± 0.020	0.754 ± 0.031	0.735 ± 0.038	0.736 ± 0.033
SVM	0.810 ± 0.038	0.752 ± 0.048	0.748 ± 0.053	0.747 ± 0.052	0.831 ± 0.037	0.780 ± 0.045	0.770 ± 0.040	0.769 ± 0.043
LightGBM	0.841 ± 0.030	0.811 ± 0.035	0.777 ± 0.041	0.781 ± 0.038	0.855 ± 0.015	0.829 ± 0.032	0.800 ± 0.031	0.801 ± 0.025

In addition, it was also shown that LightGBM outperformed other models in all the performance metrics, yielding an accuracy of 0.855 and an F1 score of 0.801. The performance demonstrated the possibility of using HRV for objective physical fatigue assessment.

4. Discussion

4.1. Performance Analysis

The overall confusion matrix of LightGBM in the 10-fold cross validation is shown in Figure 5. The row labels indicate the true classes for samples in each row, while the column labels indicate the predicted classes for samples in each column. The numbers labeled in each grid show the number of samples classified into the classes labeled in the row, and the labels shown in the row are the true classes of the samples. The color represents the proportion of the aforementioned samples to all samples in the same row.

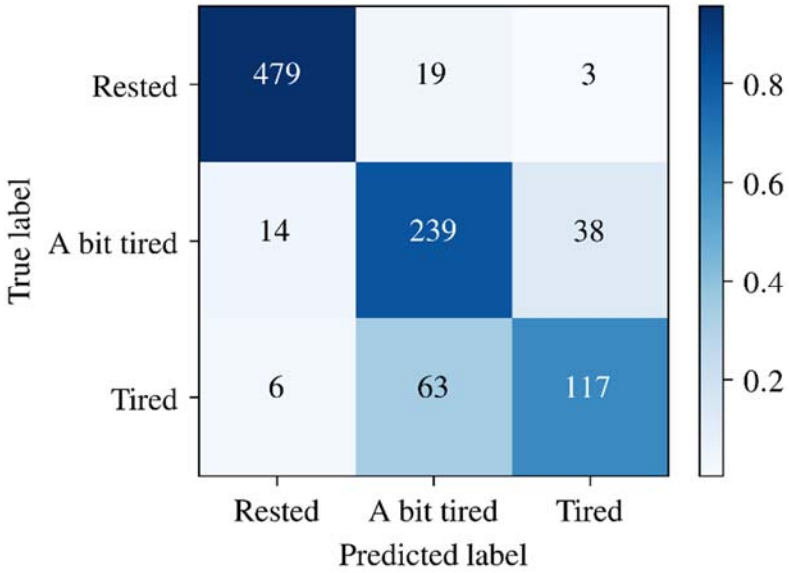

Figure 5. Overall confusion Matrix of LightGBM of the 10-fold cross validation.

From the results shown in Figure 5, we can see that the model performed best in predicting "Rested" samples and worst in predicting "Tired" samples, indicating the model's lack of sensitivity in discriminating "Tired" from "A bit tired". One of the factors affecting the model performance in distinguishing between the two labels is the imbalance of the dataset. In the collected dataset, there are more rested samples than tired samples. Due to the lack of sufficient tired samples, the classifier was insufficient in describing tired samples, so the trained model had a poor performance in generalizing the "tired" label.

In addition, unlike most of the previous studies, which coarsely classify fatigue states into "Tired" and "Non-tired", this study had one "Non-tired" state, namely "Rested", and two levels of tiredness, namely "A bit tired" and "Tired"; as there may be bias in individual perception of the two levels of tiredness states, the classification performance for these two levels of tiredness is relatively inferior. The accuracy of the model would be greatly improved if "A bit tired" and "Tired" were merged into one category.

4.2. Comparison with Related Works

Our results suggest that the meanHR, NN50, SDANN, HRVTi, TINN, aVLF, sampen, SD2, SD1/SD2, α and α_1 are the key HRV features for physical fatigue assessment. Inputting too few features or too many features may decrease the classification performance.

The HRV time domain features have been used in drivers' sleepiness detection. Abtahi et al. [39] conducted a variance analysis between groups and found that there was a statistically significant difference for these time domain features, including meanNN, SDNN, SDANN, SDNNi and NN50. According to their results, the meanNN and SDNN increased when drivers' mental state transformed from alert to severe sleepiness. The data analysis in [40] also showed that meanNN, SDNN and HRVTi were associated with drivers' mental stress levels.

Previous studies have explored the frequency domain features of HRV to detect drivers' mental fatigue. Some studies have shown that there is a significant rise in the LF/HF when the driver became drowsy [41]. While some studies suggest that the LF/HF has no significant changes when human states changed [39], other studies reported that the LF/HF even decreased with mental workload and mental stress increases [25,42]. Therefore, a study reported that the change direction and degree of HRV linear indexes may not be the same in different degrees of mental fatigue [26]. The results from our study show that the VLF was related to physical fatigue, which is consistent with the previous research results reported in [43].

The analysis in [44] showed that SD1 and SD2 decreased after a table tennis match, indicating activation of the sympathetic system and, simultaneously, deactivation of the parasympathetic system. Another study [45] pointed out that α_1 decreased when running at low intensity. They suggested that α_1 can provide the opportunity to track physiological status in real time to monitor exercise fatigue. Similar to the aforementioned studies, our study suggests that SD2, SD1/SD2, α and α_1 can help to assess physical fatigue.

4.3. Limitations

Although we obtained promising results for LightGBM using the selected features, there were still limitations. For example, the number of subjects involved in this study was small, which may affect the stability of the models. In addition, the results of our analysis verify that heart rate is an important indicator for the final evaluation of physical fatigue. However, heart rate can be affected by many factors, e.g., diseases (including hyperthyroidism [46], diabetes [47]), emotion [48] and age [49]. Future work should be carried out to study the influence of factors affecting heart rate on the proposed physical fatigue assessment model.

5. Conclusions

The application of HRV and machine learning algorithms in physical fatigue assessment was studied in this paper. First, 24 HRV features of four domains were computed from the original ECG. Then, a two-step feature selection method was proposed to select the best features. After feature selection, 13 original features were removed, and 11 optimal features were selected and used as the input of the model. These selected HRV features identified for physical fatigue detection are meanHR, NN50, SDANN, HRVTi, TINN, aVLF, sampen, SD2, SD1/SD2, α and α_1. Four machine learning algorithms, DT, SVM, KNN and LightGBM, were used to build classifiers that automatically detect the fatigue state. LightGBM achieved the best performance and had an accuracy of 0.855 and an F1 score of 0.801. The results verify the feasibility of using HRV to evaluate physical fatigue states. By using the features selected by our feature selection method, the proposed model achieved superior performance in assessing the physical fatigue state. Furthermore, the selected features can be applied to wearable ECG devices for physical fatigue assessment during exercise in real time.

In future works, more subjects with diseases (e.g, diabetes, hypertension, heart disease) and subjects with different levels of physical fitness and different ages will be included to in-

crease the reliability of physical fatigue assessment. Finally, other machine learning or deep learning models and features based on other physiological signals would be considered.

Author Contributions: Conceptualization, F.S. and Y.L.; methodology, Z.N.; software, Z.N.; data curation, Z.N.; writing—original draft preparation, Z.N.; writing—review and editing, F.S. and Y.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This study was sponsored by the Strategic Priority CAS Project under grant number XDB38040200; National Natural Science Foundation of China under grant number 62073310; the basic research project of Guangdong Province under grant number 2021A1515011838; Joint Fund of NSFC and Shenzhen under grant number U1913210; Joint Fund of NSFC and Guangdong province under grant number U1801261; and the National Key R&D Program of China under grant number 2018YFB1307005.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (protocol code SIAT-IRB-200715-H0510; 17 July 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used in this study are not publicly available because the institutional review board did not grant permission, but they are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Foschini, D.; Prestes, J.; Charro, M.A. Relationship between physical exercise, muscle damage and delayed-onset muscle soreness. *Rev. Bras. Cineantropometria Desempenho Hum.* **2007**, *9*, 101–106.
2. Budgett, R. Overtraining syndrome. *Br. J. Sports Med.* **1990**, *24*, 231–236. [CrossRef] [PubMed]
3. Goodman, J.M.; Banks, L.; Connelly, K.A.; Yan, A.T.; Backx, P.H.; Dorian, P. Excessive exercise in endurance athletes: Is atrial fibrillation a possible consequence? *Appl. Physiol. Nutr. Metab.* **2018**, *43*, 973–976. [CrossRef] [PubMed]
4. Smith, L.L. Overtraining, Excessive Exercise, and Altered Immunity. *Sports Med.* **2003**, *33*, 347–364. [CrossRef]
5. Ream, E.; Richardson, A. Fatigue: A concept analysis. *Int. J. Nurs. Stud.* **1996**, *33*, 519–529. [CrossRef]
6. Borg, G.A. Psychophysical bases of perceived exertion. *Med. Sci. Sports Exerc.* **1982**, *14*, 377–381. [CrossRef]
7. Coutts, A.J.; Reaburn, P.; Murphy, A.; Pine, M.; Impellizzeri, F. Validity of the session-RPE method for determining training load in team sport athletes. *J. Sci. Med. Sport* **2003**, *6*, 525.
8. Seiler, S.; Hetlelid, K.J. The Impact of Rest Duration on Work Intensity and RPE during Interval Training. *Med. Sci. Sports Exerc.* **2005**, *37*, 1601–1607. [CrossRef]
9. Scott, T.J.; Black, C.R.; Quinn, J.; Coutts, A.J. Validity and reliability of the session-RPE method for quantifying training in Australian football: A comparison of the CR10 and CR100 scales. *J. Strength Cond. Res.* **2013**, *27*, 270–276. [CrossRef]
10. Crewe, H.; Tucker, R.; Noakes, T.D. The rate of increase in rating of perceived exertion predicts the duration of exercise to fatigue at a fixed power output in different environmental conditions. *Eur. J. Appl. Physiol.* **2008**, *103*, 569–577. [CrossRef]
11. Venhorst, A.; Micklewright, D.; Noakes, T.D. Perceived Fatigability: Utility of a Three-Dimensional Dynamical Systems Framework to Better Understand the Psychophysiological Regulation of Goal-Directed Exercise Behaviour. *Sports Med.* **2018**, *48*, 2479–2495. [CrossRef] [PubMed]
12. Sun, F.; Yi, C.; Li, W.; Li, Y. A wearable H-shirt for exercise ECG monitoring and individual lactate threshold computing. *Comput. Ind.* **2017**, *92–93*, 1–11. [CrossRef]
13. Dong, H.; Ugaldey, I.; El Saddik, A. Development of a fatigue-tracking system for monitoring human body movement. In Proceedings of the 2014 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, Montevideo, Uruguay, 12–15 May 2014; pp. 786–791.
14. Khan, T.; Lundgren, L.E.; Järpe, E.; Olsson, M.C.; Viberg, P. A Novel Method for Classification of Running Fatigue Using Change-Point Segmentation. *Sensors* **2019**, *19*, 4729. [CrossRef] [PubMed]
15. Ražanskas, P.; Verikas, A.; Viberg, P.-A.; Olsson, M.C. Predicting physiological parameters in fatiguing bicycling exercises using muscle activation timing. *Biomed. Signal Process. Control* **2017**, *35*, 19–29. [CrossRef]
16. Yousif, H.A.; Rahim, N.A.; Bin Salleh, A.F.; Zakaria, A.; Alfarhan, K.A.; Mahmood, M. A Study of Lower Limb Muscles Fatigue during Running Based on EMG Signals. In Proceedings of the 2019 IEEE International Conference on Sensors and Nanotechnology, Penang, Malaysia, 24–25 July 2019; pp. 1–4.
17. Yang, Z.; Ren, H. Feature Extraction and Simulation of EEG Signals During Exercise-Induced Fatigue. *IEEE Access* **2019**, *7*, 46389–46398. [CrossRef]

18. Lin, S.-Y.; Hung, C.-I.; Wang, H.-I.; Wu, Y.-T.; Wang, P.-S. Extraction of physically fatigue feature in exercise using electromyography, electroencephalography and electrocardiography. In Proceedings of the 2015 11th International Conference on Natural Computation (ICNC), Zhangjiajie, China, 15–17 August 2015; pp. 561–566.
19. Sztajzel, J. Heart rate variability: A noninvasive electrocardiographic method to measure the autonomic nervous system. *Swiss Med. Wkly.* **2004**, *134*, 514–522.
20. Ansari, S.; Gryak, J.; Najarian, K. Noise Detection in Electrocardiography Signal for Robust Heart Rate Variability Analysis: A Deep Learning Approach. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; Volume 2018, pp. 5632–5635.
21. Miao, F.; Liu, Z.-D.; Liu, J.-K.; Wen, B.; He, Q.-Y.; Li, Y. Multi-Sensor Fusion Approach for Cuff-Less Blood Pressure Measurement. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 79–91. [CrossRef]
22. Egelund, N. Spectral analysis of heart rate variability as an indicator of driver fatigue. *Ergonomics* **1982**, *25*, 663–672. [CrossRef]
23. Makivić, B.; Nikić Djordjević, M.; Willis, M.S. Heart Rate Variability (HRV) as a tool for diagnostic and monitoring performance in sport and physical activities. *J. Exerc. Physiol. Online* **2013**, *16*, 103–131.
24. Wu, Q.; Zhao, Y.; Bi, X. Driving Fatigue Classified Analysis Based on ECG Signal. In Proceedings of the 2012 Fifth International Symposium on Computational Intelligence and Design, Hangzhou, China, 28–29 October 2012; Volume 2, pp. 544–547.
25. Patel, M.; Lal, S.; Kavanagh, D.; Rossiter, P. Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert Syst. Appl.* **2011**, *38*, 7235–7242. [CrossRef]
26. Yue, Y.; Liu, D.; Fu, S.; Zhou, X. Heart Rate and Heart Rate Variability as Classification Features for Mental Fatigue Using Short-Term PPG Signals Via Smartphones Instead of ECG Recordings. In Proceedings of the 2021 13th International Conference on Communication Software and Networks (ICCSN), Chongqing, China, 4–7 June 2021; pp. 370–376.
27. Ramos, G.; Vaz, J.R.; Mendonça, G.V.; Pezarat-Correia, P.; Rodrigues, J.; Alfaras, M.; Gamboa, H. Fatigue Evaluation through Machine Learning and a Global Fatigue Descriptor. *J. Health Eng.* **2020**, *2020*, 6484129. [CrossRef] [PubMed]
28. Cosoli, G.; Iadarola, G.; Poli, A.; Spinsante, S. Learning classifiers for analysis of Blood Volume Pulse signals in IoT-enabled systems. In Proceedings of the 2021 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4. 0&IoT), Rome, Italy, 7–9 June 2021; pp. 307–312.
29. Guan, X.; Lin, Y.; Wang, Q.; Liu, Z.; Liu, C. Sports fatigue detection based on deep learning. In Proceedings of the 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 23–25 October 2021; pp. 1–6.
30. Bruce, R.A.; Kusumi, F.; Hosmer, D. Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease. *Am. Heart J.* **1973**, *85*, 546–562. [CrossRef]
31. Malik, M.; Bigger, J.T.; Camm, A.J.; Kleiger, R.E.; Malliani, A.; Moss, A.J.; Schwartz, P.J. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Eur. Heart J.* **1996**, *17*, 354–381. [CrossRef]
32. Bigger, J.T., Jr.; Fleiss, J.L.; Steinman, R.C.; Rolnitzky, L.M.; Kleiger, R.E.; Rottman, J.N. Frequency domain measures of heart period variability and mortality after myocardial infarction. *Circulation* **1992**, *85*, 164–171. [CrossRef]
33. Lampert, R.; Bremner, J.D.; Su, S.; Miller, A.; Lee, F.; Cheema, F.; Goldberg, J.; Vaccarino, V. Decreased heart rate variability is associated with higher levels of inflammation in middle-aged men. *Am. Heart J.* **2008**, *156*, 759.e1–759.e7. [CrossRef]
34. Berntson, G.G.; Thomas Bigger, J., Jr.; Eckberg, D.L.; Grossman, P.; Kaufmann, P.G.; Malik, M.; Nagaraja, H.N.; Porges, S.W.; Saul, J.P.; Stone, P.H. Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology* **1997**, *34*, 623–648. [CrossRef]
35. Wickens, C.D.; Gordon, S.E.; Liu, Y.; Lee, J. *An Introduction to Human Factors Engineering*; Pearson Prentice Hall: Upper Saddle River, NJ, USA, 2004; Volume 2.
36. Vest, A.N.; Da Poian, G.; Li, Q.; Liu, C.; Nemati, S.; Shah, A.; Clifford, G.D. An open source benchmarked toolbox for cardiovascular waveform and interval analysis. *Physiol. Meas.* **2018**, *39*, 105004. [CrossRef]
37. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef]
38. Mukaka, M.M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **2012**, *24*, 69–71.
39. Abtahi, F.; Anund, A.; Fors, C.; Seoane, F.; Lindecrantz, K. Association of Drivers' sleepiness with heart rate variability: A Pilot Study with Drivers on Real Roads. In *EMBEC & NBC 2017*; Springer: Singapore, 2017; pp. 149–152.
40. Yu, Y.J.; Yang, Z.; Oh, B.-S.; Yeo, Y.K.; Liu, Q.; Huang, G.-B.; Lin, Z. Investigation on driver stress utilizing ECG signals with on-board navigation systems in use. In Proceedings of the 2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV), Phuket, Thailand, 13–15 November 2016; pp. 1–6.
41. Rodriguez-Ibañez, N.; García-Gonzalez, M.A.; de la Cruz, M.A.F.; Fernández-Chimeno, M.; Ramos-Castro, J. Changes in heart rate variability indexes due to drowsiness in professional drivers measured in a real environment. In Proceedings of the 2012 Computing in Cardiology, Krakow, Poland, 9–12 September 2012; pp. 913–916.
42. Bhardwaj, R.; Natrajan, P.; Balasubramanian, V. Study to Determine the Effectiveness of Deep Learning Classifiers for ECG Based Driver Fatigue Classification. In Proceedings of the 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), Rupnagar, India, 1–2 December 2018; pp. 98–102.

43. Chang, T.-C. Detection of Exercise Fatigue Using Neural Network with Grey Relational Analysis from HRV Signal. In Proceedings of the 2019 IEEE International Conference on Computation, Communication and Engineering (ICCCE), Fujian, China, 8–10 November 2019; pp. 87–90.
44. Picabea, J.M.; Cámara, J.; Nakamura, F.Y.; Yanci, J. Comparison of Heart Rate Variability Before and After a Table Tennis Match. *J. Hum. Kinet.* **2021**, *77*, 107–115. [CrossRef]
45. Rogers, B.; Mourot, L.; Doucende, G.; Gronwald, T. Fractal correlation properties of heart rate variability as a biomarker of endurance exercise fatigue in ultramarathon runners. *Physiol. Rep.* **2021**, *9*, e14956. [CrossRef] [PubMed]
46. Galetta, F.; Franzoni, F.; Fallahi, P.; Tocchini, L.; Braccini, L.; Santoro, G.; Antonelli, A. Changes in heart rate variability and QT dispersion in patients with overt hypothyroidism. *Eur. J. Endocrinol.* **2008**, *158*, 85–90. [CrossRef] [PubMed]
47. Masaoka, S.; Lev-Ran, A.; Hill, L.R.; Vakil, G.; Hon, E.H.G. Heart Rate Variability in Diabetes: Relationship to Age and Duration of the Disease. *Diabetes Care* **1985**, *8*, 64–68. [CrossRef] [PubMed]
48. Quintana, D.S.; Guastella, A.J.; Outhred, T.; Hickie, I.; Kemp, A.H. Heart rate variability is associated with emotion recognition: Direct evidence for a relationship between the autonomic nervous system and social cognition. *Int. J. Psychophysiol.* **2012**, *86*, 168–172. [CrossRef]
49. Bonnemeier, H.; Wiegand, U.K.; Brandes, A.; Kluge, N.; Katus, H.A.; Richardt, G.; Potratz, J. Circadian profile of cardiac autonomic nervous modulation in healthy subjects: Differing effects of aging and gender on heart rate variability. *J. Cardiovasc. Electrophysiol.* **2003**, *14*, 791–799. [CrossRef] [PubMed]