

## Article

# Stress Classification of ECG-Derived HRV Features Extracted from Wearable Devices

Kayisan M. Dalmeida <sup>1</sup> and Giovanni L. C Masala <sup>1, \*</sup>

<sup>1</sup> Department of Computing and Mathematics at Manchester Metropolitan University; kayisan.m.dalmeida@stu.mmu.ac.uk; g.masala@mmu.ac.uk

\* Correspondence: g.masala@mmu.ac.uk; Tel.: +44 (0)161 247 1407

**Abstract:** Stress has been identified as one of the major causes of automobile crashes which then lead to high rates of fatalities and injuries each year. Stress can be measured via physiological measurements and in this study the focus will be based on the features that can be extracted by common wearable devices. Hence the study will be mainly focusing on the heart rate variability (HRV). This study is aimed to develop a good predictive model that can accurately classify stress levels from ECG-derived HRV features, obtained from automobile drivers, testing different machine learning methodologies such as K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Multilayer Perceptron (MLP), Random Forest (RF) and Gradient Boosting (GB). Moreover, the models obtained with highest predictive power will be used as reference for the development of a machine learning model that would be used to classify stress from HRV features derived from HRV measurements obtained from wearable devices. We demonstrate that MLP was the ideal stress classifier by achieving a Recall of 80%. The proposed method can be also used on all applications in which is important to monitor the stress level e. g. in physical rehabilitation, anxiety relief or mental wellbeing.

**Keywords:** stress; wearable device; machine learning; smart watch; heart rate variability; electrocardiogram

## 1. Introduction

Stress can be defined as a biological and psychological response to a combination of external or internal stressors [1-2], which could be a chemical or biological agent or an environmental stimulus that causes stress to an organism [3]. Stress is, in essential, the body's coping mechanism to any kind of foreign demand or threat. At the molecular level, in a stressful situation the Sympathetic Nervous System (SNS) produces stress hormones, such as cortisol, which then, via a cascade of events, lead to the increase of available sources of energy [4]. This large amount of energy is used to fuel a series of physiological mechanisms such as: increasing the metabolic, increasing heart rate and causing the dilation of blood vessels in the heart and other muscles [5] while decreasing non-essential tasks such as immune system and digestion. Once stressors no longer impose a threat to the body, the brain fires up the Parasympathetic Nervous System (PSN) which is in charge of restoring the body to homeostasis. However, if the PSN fails to achieve homeostasis, this could lead to chronic stress; thus, causing a continual and prolonged activation of the stress response [6]. Conversely during acute stress, the stress response develops immediately, and it is short-lived.

Studies carried out in this field suggest that stress can lead to abnormalities in the cardiac rhythm, and this could lead arrhythmia [7]. Additionally, stress does not only have physical implications, but it can also be detrimental to one's mental health; in fact, chronic stress can enhance the chances of developing depression. For these reasons, it is important to develop a system that can detect and measure stress in an individual in a non-invasive manner in such way that stress can be regulated or relieved via personalised medical interventions or even by just alerting the user of their stressful state.

Furthermore, stress has been identified as one of the major causes of automobile crashes which then lead to high rates of fatalities and injuries each year [8]. As reported by Virginia Tech Transportation Institute (VTTI) and the National Highway Traffic Safety Administration (NHTSA), lack of attention and stress were the leading cause of traffic accidents in the US with a rate of ~80%. Therefore, being able to accurately monitor stress in drivers could significantly reduce the amount of road traffic accidents and consequently increase public road safety.

Given that stress is regulated by the Autonomous Nervous System, it can be measured via physiological measurements such as Electrocardiogram (ECG), Galvanic Skin Response (GSR), electromyogram (EMG), heart rate variability (HRV), heart rate (HR), blood pressure, breathe frequency, Respiration Rate and Temperature [9]. These are considered to be an accurate methodology for bio signal recording as they cannot be mask or conditioned by human voluntary actions. However, this study will be mainly focusing on HRV, which is controlled by PSN and SNS therefore, an imbalance in any functions regulated by these two nervous system branches will affect HRV [10]. HRV is the variation in interval between successive normal RR (or NN) intervals [11] it is derived from an ECG reading and it is measured by calculating the time interval between two consecutive peaks of the heartbeats [12]. As explained in [11] the RR intervals are obtained by calculating the difference between two R waves in the QRS complex.

HRV can be subdivided into time domain and frequency domain metrics as described in Table 1.

**Table 1.** Time and Frequency Metrics derived from HRV

Time Domain Metrics	
SDNN	Standard deviation of all NN intervals
SDANN	Standard deviation of the average NN intervals
AVNN	Average of NN intervals
rMSSD	Square root of the mean squared differences of successive RR intervals
pNN50	Percentage differences of successive RR intervals larger than 50ms
Frequency Domain Metrics	
TP	Total Power - total spectral power of all NN intervals up to 0.004 Hz
LF	Low Frequency – total spectral power of all NN intervals with frequency ranging from 0.04 Hz to 0.15 Hz
HF	High Frequency – total spectral power of all NN intervals with frequency ranging from 0.15 Hz to 0.4 Hz
VLF	Very Low Frequency – total spectral power of all NN intervals with frequencies > 0.004 Hz
ULF	Ultra-Low Frequency - total spectral power of all NN intervals with frequencies < 0.003 Hz
LF/HF	Ratio of low to high frequency

HRV is traditionally obtained from ECG and requires the use of computational software for calculation; this process is limited to laboratory or clinical settings and requires a certain degree of technical knowledge for interpretation and calculation. Thanks to the advancement of technology, however, commercially available portable devices and wearables have the capacity to monitor and record HRV measurements. [13] performed a systematic review and meta-analysis on the numerous studies that compared the quality of HRV measurements acquired from ECG and obtained from portable devices, such as Elite HRV, Polar H7 and Motorola Droid [13]. Twenty-three studies revealed that HRV measurements obtained from portable devices resulted in a small amount of absolute error when compared to ECG; however, this error is acceptable as this method of acquiring HRV is more practical and cost-effective as no laboratory or clinical apparatus are required [13].

Furthermore, the Apple Watch is one of the most best-selling and popular smartwatches in the market. Studies carried out by Shcherbina and colleagues [14], demonstrated that the Apple Watch was the best HR estimating smartwatch with one-minute granularity and with the lowest overall median error (below 3%) while Samsung Gear S2 reported the highest error. In addition, it is also important to validate the HRV estimation of the Apple Watch. Currently, the best way to obtain RR raw values from the Apple Watch is via the Breathe app developed by Apple. Authors in [15] conducted an investigation that validated the Apple Watch in relation to HRV measurements derived during mental stress in 20 healthy subjects. In this study, the RR interval series provided by the Apple watch was validated using the RR interval obtained from Polar H7 [15]. Successively, the HRV parameters were compared and their ability to identify the Autonomous Nervous System (ANS) response to mild mental stress was analysed [15]. The results revealed that the Apple Watch HRV measurements had good reliability and the HRV parameters were able to indicate changes caused by mild mental stress as it presented a significant decrease in HF power and RMSSD in stress condition compared to the relax state [15]. Therefore, this study suggests that the Apple Watch presents a potential non-invasive and reliable tool for stress monitoring and detection. In this study, raw RR intervals, from beat-to-beat measurements, obtained from the Breathe app are considered for stress classification.

This study is aimed to develop a good predictive model that can accurately classify stress levels from ECG-derived HRV features, obtained from automobile drivers, testing different machine learning methodologies such as K-Nearest Neighbour (KNN), Support Vector Machines (SVM), Multilayer Perceptron (MLP), Random Forest (RF) and Gradient Boosting (GB). Moreover, the models obtained with highest predictive power will be used as reference for the development of a machine learning model that would be used to classify stress from HRV features derived from HRV measurements obtained from wearable devices.

The paper is organised as follows. Section 2 provides a discussion of related work conducted in the literature. Section 3 describes the experimental methodology of the study, including a description of the dataset, pre-processing, hyperparameter tuning and the design protocol used for the development of a simple stress detection web application based on Apple Watch derived data. Section 4 presents the experimental results and Section 5 an intensive discussion of the results obtained. Lastly, Section 6 provides the concluding remarks of the study, as well as proposed future work.

## 2. Related Work

As stress level changes so does the HRV and it has been proven that HRV decreases as stress increases [11]. This is possible because HRV provides a measure to monitor the activity of the ANS and therefore, can provide a measure of stress [16]. Authors in [16] explored the interaction between HRV and mental stress. Here they took ECG recordings during rest and mental task conditions, which was meant to reflect a stressful state. Linear HRV measures were then analysed in order to provide information on how the heart

responds to a stressful task. The results demonstrated that the mean RR interval was significantly lower during mental task than in the rest condition [16]. This difference was significant only when time domain parameters (pNN50) and the mean RR interval were analysed; while the frequency domain measure did not show a significant difference although there was an elevated LF/HF in the stressed condition [16]. As LF is associated with the SNS and HF with PNS, the increased LF/HF ratio does suggest that there is a higher sympathetic activity in the stress condition compared to the resting state [16]. Furthermore, investigations have been carried out in order to accurately classify stress in drivers via HRV measurements. For example, authors in [17] aimed to classify ECG data using extracted parameters into highly stressed and normal physiological states of drivers. In this study they extracted time domain, frequency domain and nonlinear domain parameters from HRV obtained by extracting RR intervals from QRS complexes. These extracted features were fed into the following machine learning classifiers: K Nearest Neighbor (KNN), radial basis function (RBF) and Support Vector Machine (SVM); more on machine learning classifiers in the next subsections. The results showed that SVM with RBF kernel gave the highest results, with 83.33% accuracy, when applied to time and nonlinear parameters while giving an accuracy of 66.66% with frequency parameter [17]. This was in concordance with the result obtained by [16] as the frequency domain parameters did not give a significant difference between rest and mental tasks.

### 3. Materials and Methods

#### 3.1. Datasets

The first part of this study consists in the development of a good stress predictive model from ECG-derived HRV measurements. The dataset used was collected at Massachusetts Institute of Technology (MIT) by Healey and Picard [18] which is freely available from PhysioNet [19]. The dataset consists of a collection of multi-parameter recordings obtained from 27 young and healthy individuals while they were driving on a designated route in the city and highways around Boston, Massachusetts. The driving protocol involved a route that was planned to put the driver through different levels of stress, specifically, the drive consisted of periods of rest, highway driving and city driving which were presumed to induce low, medium and high stress respectively [18]. This investigation measured four types of physiological signals: ECG, EMG, GSR and respiration. The dataset is available in the PhysioNet waveform format containing 18 *.dat* and 18 *.hea* files with a *.txt* metadata file. Each bio-signal *.dat* file contains the original recording for ECG, EMG, GSR, HR and Respiration. As the aim of this study is to classify stress based on HRV metrics, a beat annotation file was created from a beat annotation file were created from *.dat* files by using the WQRS tool that works by locating QRS complexes in the ECG signal using and gives an annotation file as the output [20]. The annotation file serves the purpose of extracting RR intervals together with its corresponding timestamp using the PhysioNet HRV toolkit. HRV features were extracted from the RR intervals by splitting the dataset in windows of 30 seconds. Time domain features were calculated using a C implementation that connects Python to the PhysioNet HRV toolkit and by calling the *get\_hrv* method which returns the HRV metrics. While frequency domain metrics were obtained by applying the Lomb Periodogram which determines the power spectrum at any given frequency [21]. GSR signals were used to determine and label the stress states in drivers, as the marker in the dataset was mainly made of missing values. The median GSR values were used as the cut-off point, thus, values above the media were labelled as stress while the values below the median were labelled as no stress. For clarity reasons, this dataset will be referred to as 'original-dataset'.

The second portion of this investigation aimed to develop machine learning model that would classify stress from HRV features derived from HRV measurements obtained from the Apple Watch. For this purpose, data was collected from 4 Apple Watch users, who were asked not to exercise or intake caffeine before and on the day of the experiment. The volunteers were subjected to 2 different conditions. Firstly, they performed total

relaxation of 15 minutes while listening to relaxing lo-fi music. Immediately after this task, the volunteers were to take 5 HRV measurements using the Apple Breathe App available on their Apple Watch. Finally, the subjects were asked to take HRV measurement immediately after 8 hours of work; this was performed the day after the first task was carried out. Thanks to the Breathe App it was possible to obtain raw RR intervals from beat-to-beat measurements and all the data was accessible from the user's Personal Health Record, which can be exported in XML format via Apple's Health App. The beat-to-beat measurements of interests, mapped into the <InstantaneousBeatsPerMinute> tag, were extracted from the XML file using the Python *ElementTree* module from the *xml* library that allows to parse XML data which can then be saved into panda DataFrame. Successively, the raw RR intervals (in seconds) were derived from the beat per minute (bpm) readings using the following equation:

$$RR = \frac{60}{bpm} \quad (1)$$

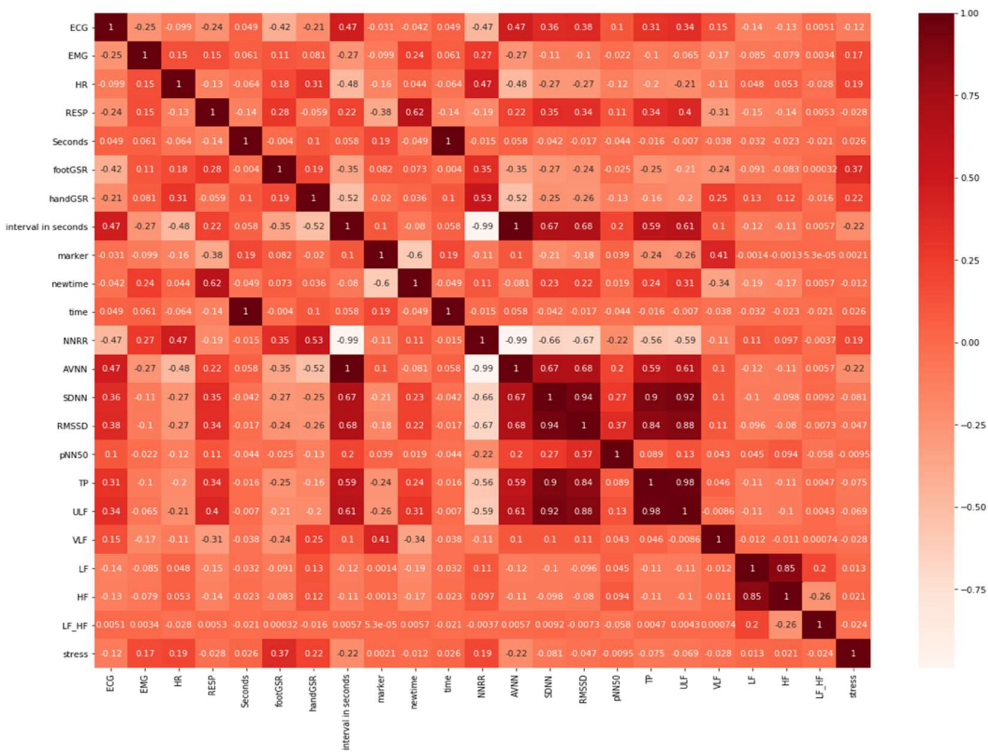
Moreover, HRV features were extracted from the calculated RR intervals utilising *NumPy* library, for time domain, and the *python* library, specifically the *frequency\_domain* module and the Welch's Method for frequency domain features [22]. This dataset will be used as a blind test for the obtained classifier, in order to measure its predictive power on unseen data; hence, this dataset will be referred to as 'blind-dataset' throughout this paper.

### 3.1.2. Data Pre-Processing

Firstly, missing values in original-dataset were replaced with the mean value of each column using the *SimpleImputer* module from scikit-learn library and the *replace* panda method. Then the data was further split into training and testing datasets with an 80:20 (training: testing) split. From this point onwards, the testing and training data were treated separately as different entities in order to prevent data overfitting and data leakage. Data normalisation was done separately on the training and testing set instead of the whole dataset that could leak information about the test into the train set. Normalisation was performed via the *MinMaxScaler* method from scikit-learn where continuous values are rescaled into a range between 0 and 1 with the aim to have all numeric columns in the same range as there are features that are in different ranges such as ECG, HR, EMG, seconds and HF.

### 3.1.2. Feature Selection

Features were selected based on their relevance to the classification task that this study proposed. This was accomplished using three techniques: Pearson's Correlation (Figure 1), Recursive Feature Elimination (RFE) [27] (Table 2) and Extra Tree Classifier [28] (Figure 2), used to estimate feature importance.



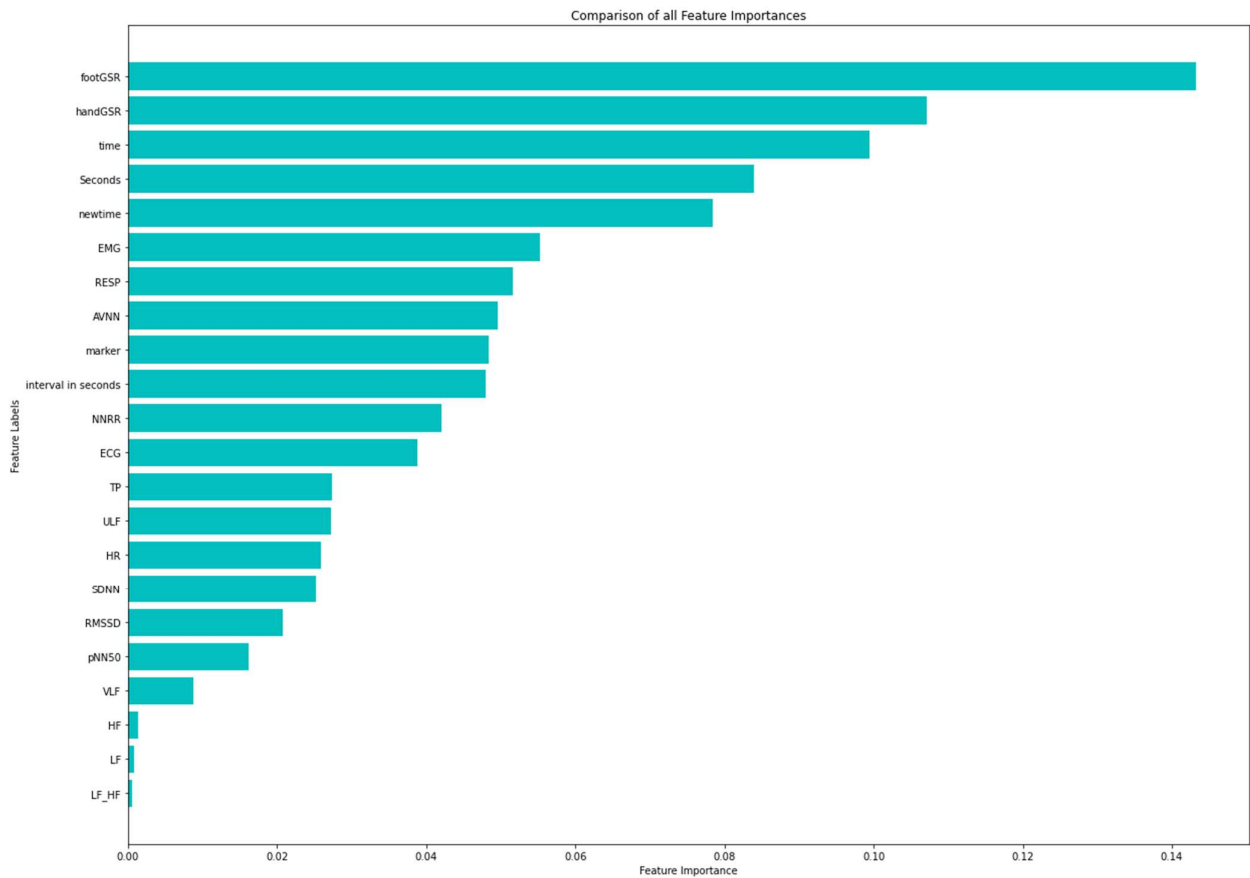
**Figure 1.** Heat map plot of Pearson’s Correlation Feature Selection performed on original-dataset.

**Table 2.** RFE feature importance score on original-dataset, the most relevant features have the lowest RFE score.

Feature	RFE score
EMG	1
HR	1
footGSR	1
handGSR	1
Interval in seconds	1
NNRR	1
AVNN	1
RMSSD	1
TP	1
ULF	1
RESP	2
marker	3
SDNN	4
LF_HF	5
LF	6
ECG	7
pNN50	8
HF	9



newtime	10
time	11
VLF	12
Seconds	13



**Figure 2.** Feature Importance of features from original-dataset using Extra Trees Classifier.

Feature selection scores from RFE, shown in Table 2, indicate that the most relevant features are those with the lowest score. This also shows that the best features (score of 1) were time domain HRV metrics such as RMSSD and AVNN, and frequency domain metrics like TP and ULF; followed by SDNN with a score of 4 (Table 2). Furthermore, Figure 2 illustrates a histogram of the feature importance scores based on the Feature Importance Extra Trees Classifier, which is an ensemble-based learning algorithm that aggregates the results of multiple decision trees to output a classification result [28]. In each decision, a Gini Importance of the feature is calculated which determines the best feature to split the data on based on the Gini Index mathematical criteria. Hence, Figure 2 show the Gini Importance of each feature.

The common features, from each method, that least contributed to classification or that had the lowest score were dropped from the dataset; these were LF\_HF, LF and HF. Additionally, GSR attributes were also dropped because they presented a very strong correlation with stress classification as these were used for stress labelling. Thus, in order to avoid data leakage and overfitting, they were eliminated. Moreover, intuitively redundant features were also dropped like the time related features, marker due to its high

number of missing values and EMG, given that it is irrelevant in the context of the smart watch.

As the second goal of this study was to develop a classification model that would classify stress from data obtained from wearable devices, 'modified-dataset' was created from tailoring original-dataset to present features that were purely relevant to the attributes calculated from the RR intervals recorded from the device. This also aimed to further test the classifiers' performance on a dataset resembling that generated from the wearable device. Therefore, the relevant features for modified-dataset were: HR, AVNN, SDNN, RMSSD, pMM50, TP and VLF.

### 3.2. Parameter tuning

In order to achieve the most efficient classification model, hyper-parameter tuning was performed on each algorithm used in this study to determine the best choice of parameters that would yield the highest performance. After generating the baseline for each classifier, where the parameters were set to their default values, *GridSearchCV*, a scikit-learn library [29] function that loops through a set of predefined hyperparameters and fit the model on the training set, was used to perform parameter tuning. The output from the grid search are the best parameter combinations that give the highest predictive performance which were then compared to their corresponding baseline models. All algorithms in this study were created with the scikit-learn library.

#### 3.2.1. K-Nearest Neighbour

K-Nearest Neighbour (KNN) performs classification based on the closest neighbouring training points in a given region [23], thus, the classification of new test data is dependent on the number of neighbouring labelled examples present at that given location. In order to obtain the best KNN classification model different values for  $k$  (number of nearest neighbours) and  $p$  value (the power parameter equivalent to the Euclidean distance) were investigated. The best parameter values resulted from the grid search are as follows:  $k = 25$  and  $p = 1$ .

#### 3.2.2. Support Vector Machine

The function of the Support Vector Machine (SVM) algorithm is to locate the hyperplane in the  $N$ -dimensional space (where  $N$  represents the number of features) that classifies the data instances into their corresponding class [24]. The performance of this algorithm is affected by hyperparameters such as the soft margin regularization parameter ( $C$ ) and kernel, a function that transforms low dimensional inputs space into a higher dimensional space making the data linearly separable.

For SVM classification model, different  $C$  values and kernels, such as Linear kernel, Polynomial (poly) kernel and Gaussian Radial Basis Function (RBF) kernel were tested. The best parameter settings resulted to be RBF kernel with  $\gamma = 10$  and  $C = 100$ .

#### 3.2.3. Multilayer Perceptron

Multilayer Perceptron (MLP) is a feedforward artificial neural network that was developed to circumvent the drawbacks and limitations imposed by the single-layer perceptron [25]. MLPs are made of at least 3 layers of nodes (input layer, hidden layer and output layer), where each node is connected to every node in the subsequent layer with a certain weight. MLP's performance, like other machine learning algorithms, is highly dependent on hyperparameter tuning of the following parameters: learning rate coefficient ( $h$ ), momentum ( $\mu$ ) and the size of the hidden layer.  $h$  determines the size of the weight's adjustments made at each iteration,  $\mu$  controls the speed of training and learning rate and size of the hidden layer correspond to the number of layers and neurones in the hidden layer. A configuration of  $h = 0.001$ ,  $\mu = 0.1$  and three hidden layers of 100, 10 and 400 nodes respectively, proved to be the optimal settings for the model following the grid search.



### 3.2.4. Random Forest

Random Forest (RF) is an ensemble-based learning algorithm consisting of a combination of randomly generated decision tree classifiers, which results are aggregated to obtain a better predictive performance [23]. Based on the parameter tuning grid search performed, the optimal configuration for this algorithm was when the number of trees in the forest (estimators) was set to 300 with the maximum number of features of set to the square root of the total number of features.

### 3.2.5. Gradient Boosting

Gradient Boosting (GB) is also an ensemble-based algorithm composed of multiple decision trees trained to predict new data and where each tree is dependent on one another. This model, which is trained in a gradual, sequential and additive manner, is highly dependent of the learning rate parameter that regulates the shrinkage of the contribution of each tree to the model. The optimal value for this parameter was found to be 0.14.

A Naïve Bayes probabilistic algorithm [23] was used as the baseline model for performance comparison between the other more complex algorithms. The configuration for this model were kept as simple as possible by utilising their parameters in their default values as presented by the *GuassianNB* python model.

## 4. Results

All results, related to original-dataset and modified-dataset, are described in terms of machine learning metrics such as Area Under Receiver Operator Characteristic Curve (AUROC), Recall/Sensitivity and F1 score [23] including the standard deviation of five-fold cross validation. Meanwhile, results from stress classification from data obtained from the apple watch are expressed in terms of prediction probability.

### 4.1. Stress Classification on Original-dataset

In this experiment, stress was classified from bio-signals obtained from subjects who drove under different stress conditions. The results obtained from hyperparameter tuning, illustrated in Table 3, showed that the 3 best models for the classification task imposed by this dataset, were MLP, RF and GB which yielded a AUROC of 83%, 85% and 85% respectively. Thus, meaning that the models have more than 83% probability of correctly classifying data instances.

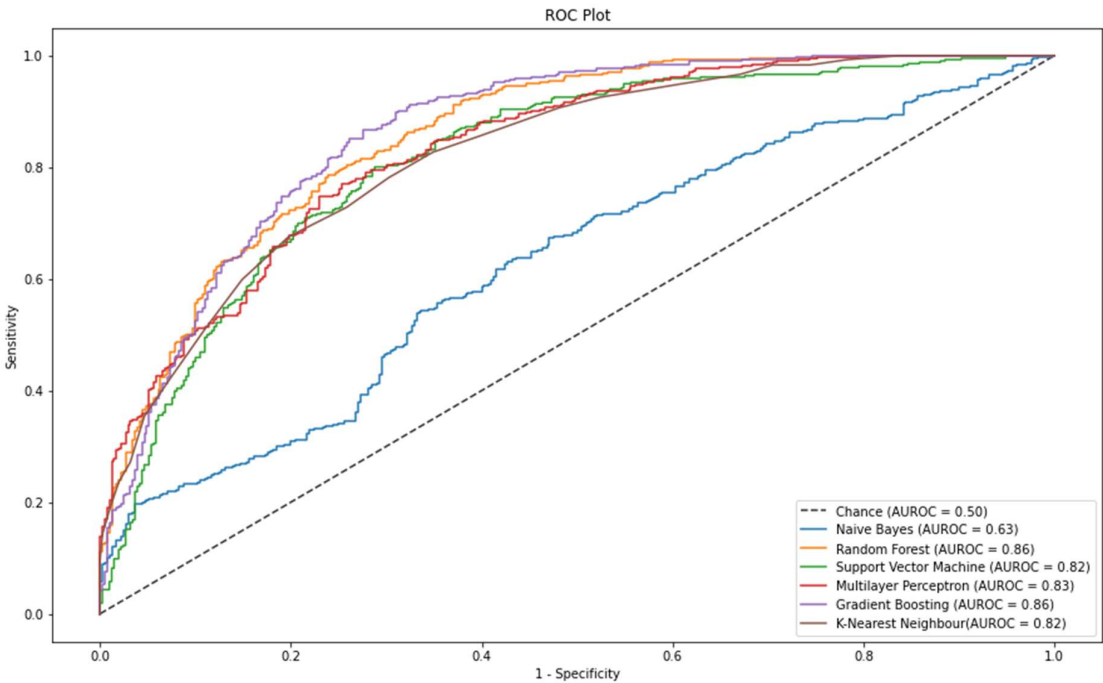
**Table 3.** Comparison of the predictive performance of the best classifiers obtained from the grid search (trained on original-dataset).

Algorithm	AUROC	Recall	F1 score
NB	0.60 ± 0.02	0.63 ± 0.04	0.61 ± 0.02
KNN	0.80 ± 0.01	0.76 ± 0.02	0.74 ± 0.01
SVM	0.81 ± 0.01	0.79 ± 0.03	0.77 ± 0.01
MLP	0.83 ± 0.01	0.81 ± 0.07	0.77 ± 0.02
RF	0.85 ± 0.01	0.81 ± 0.03	0.78 ± 0.02
GB	0.85 ± 0.01	0.80 ± 0.02	0.79 ± 0.01

NB, Naïve Bayes; KNN, K Nearest Neighbour; SVM, Support Vector Machine; MLP, Multilayer Perceptron; RF, Random Forest; GB, Gradient Boosting. NB represent the baseline model used as means of comparison for the other complex machine learning algorithm.

Moreover, MLP and RF presented a Recall of 81% while GB 80% (Table 3); this indicates that at least 80% of the predicted True Positive instances are actual positives. Therefore, at least 80% of the instances predicted to be in the stress class have been correctly classified as such. Finally, the F1 scores for MLP, RF and GB are 77%, 78% and 79%

respectively; thus, implying that the model has at least 77% accuracy on the dataset. Figure 3 illustrates the Receiver Operating Characteristics (ROC) curve for all the classifiers investigated in this study.



**Figure 3.** ROC curve plot of each classification model trained on original-dataset. ROC curve was calculated using the Sklearn roc\_curve [29] method, while the mean AUROC values were calculated using roc\_auc\_score.

Figure 3 consolidates the findings shown in Table 3, illustrating that the models with the greatest ROC area are GB, RF and MLP. It is also visible that this NB model serves as a good baseline model as its ROC curve suggest that its classification is nearly due to chance.

Furthermore, in order to determine whether there were statistical differences between the models and the baseline, a One-Way ANOVA statistical test with Tukey’s *post Hoc* comparison was performed on the mean AUROC scores (Table 4).

**Table 4.** Statistical Evaluation of the machine learning models.

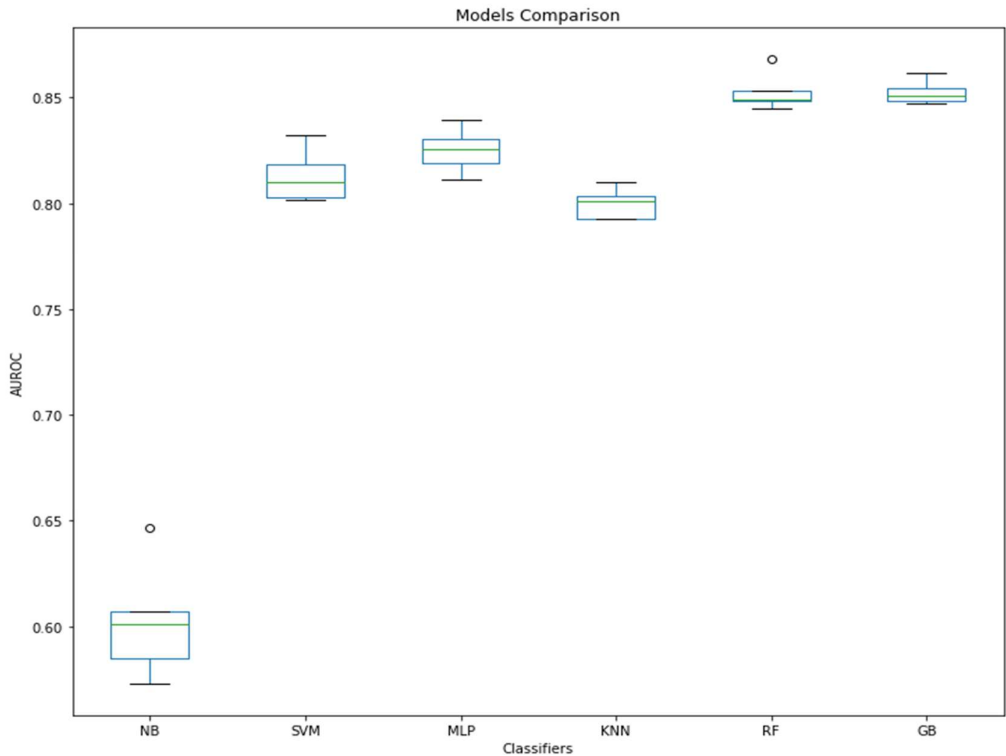
Model A	Model B	mean(A)	mean(B)	diff	se	p-tukey <sup>1</sup>
GB	KNN	0.8524596	0.8001984	0.0522612	0.00901323	<b>0.001</b>
GB	MLP	0.8524596	0.8251284	0.0273312	0.00901323	<b>0.03886793</b>
GB	NB	0.8524596	0.6026306	0.249829	0.00901323	<b>0.001</b>
GB	RF	0.8524596	0.852724	-0.0002644	0.00901323	0.9
GB	SVM	0.8524596	0.8131576	0.039302	0.00901323	<b>0.001</b>
KNN	MLP	0.8001984	0.8251284	-0.02493	0.00901323	0.076635008
KNN	NB	0.8001984	0.6026306	0.1975678	0.00901323	<b>0.001</b>
KNN	RF	0.8001984	0.852724	-0.0525256	0.00901323	<b>0.001</b>
KNN	SVM	0.8001984	0.8131576	-0.0129592	0.00901323	0.671270074
MLP	NB	0.8251284	0.6026306	0.2224978	0.00901323	<b>0.001</b>
MLP	RF	0.8251284	0.852724	-0.0275956	0.00901323	<b>0.03591592</b>
MLP	SVM	0.8251284	0.8131576	0.0119708	0.00901323	0.732202624
NB	RF	0.6026306	0.852724	-0.2500934	0.00901323	<b>0.001</b>
NB	SVM	0.6026306	0.8131576	-0.210527	0.00901323	<b>0.001</b>
RF	SVM	0.852724	0.8131576	0.0395664	0.00901323	<b>0.001</b>

<sup>1</sup>P values in bold represent statistical significance, where P < 0.05.

The null and alternate hypothesis formulated were:

- Null Hypothesis (H0): The mean AUROC score for the compared 2 models are equal.
- Alternative Hypothesis (H1): The mean AUROC score for the 2 compared models are not equal, at least AUROC value of one model is different from the other.

This showed that there was a statistical difference between the AUROC means of all hyperparameter-tuned models and the baseline (NB – AUROC = 60%) as the P < 0.05 (Table 4). This confirms that the parameter tuning did improve the model’s performance significantly and thus, H1 is accepted. Moreover, the Tukey’s comparison test showed that there is a statistically significant difference between the AUROC values of GB and MLP and between MLP and RF (P < 0.05). However, the difference between GB and Rf are not statistically significant (P = 0.9). Figure 4 summarises the results obtained during this experimental series, by illustrating the performance comparison between the hyperparameter-tuned models and the baseline NB model.



**Figure 4.** Model performance comparison of machine learning algorithms trained on original-dataset.

4.2. Stress Classification on Modified-dataset.

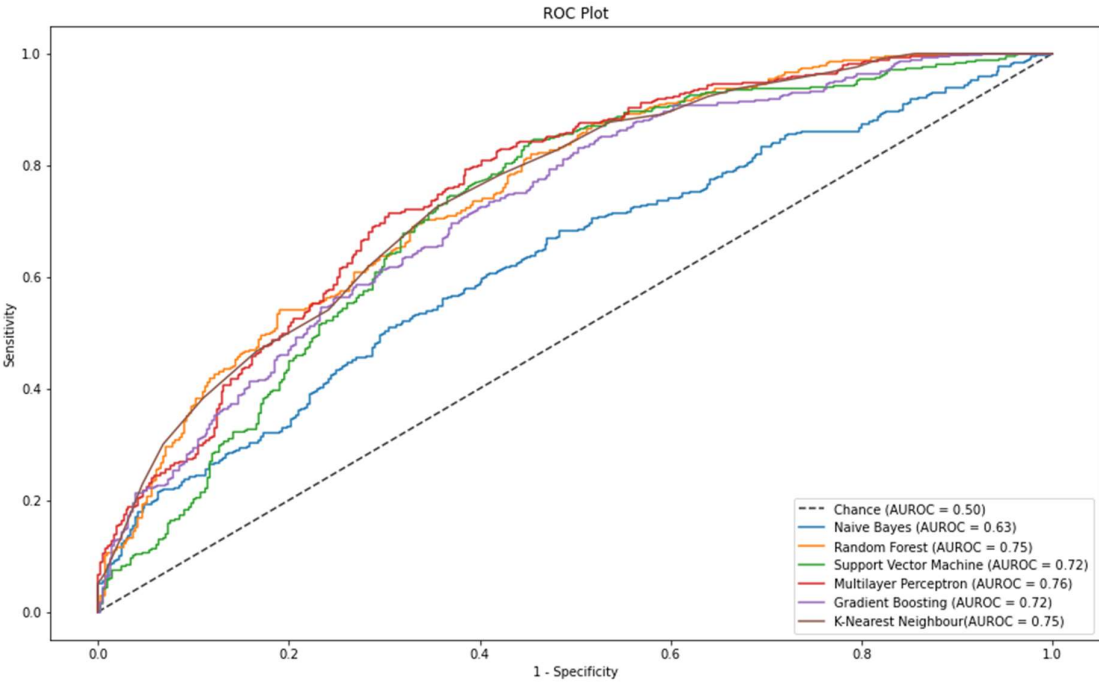
The other objective of this study was to develop a classification model that would classify stress from HRV data obtained from wearable devices. To achieve this, classifiers from Table 3, were used for stress classification of modified-dataset which is a modification of original-dataset but with features that mimics those obtained from the wearable device (as explained in the feature selection, paragraph 3.1.2). Table 5 shows the results obtained during the classification task.

**Table 5.** Predictive performance of machine learning classifiers on modified-dataset.

Algorithm	AUROC	Recall	F1 score
NB	0.60 ± 0.02	0.69 ± 0.04	0.63 ± 0.02
KNN	0.74 ± 0.02	0.76 ± 0.02	0.71 ± 0.02
SVM	0.74 ± 0.01	0.79 ± 0.02	0.74 ± 0.01
MLP	0.75 ± 0.01	0.80 ± 0.06	0.72 ± 0.02
RF	0.77 ± 0.01	0.74 ± 0.01	0.72 ± 0.01
GB	0.73± 0.01	0.70 ± 0.02	0.70 ± 0.01

As shown in Table 5, MLP seems to be the overall best performing classifier with 75% AUROC, 80% Recall and 72% F1 score.

In addition, Figure 5 illustrates the Receiver Operating Characteristics (ROC) curve for all the classifiers used for the classification of modified-dataset.



**Figure 5.** ROC curve plot of each classification model tested on modified-dataset. ROC curve was calculated using the sklearn-roc\_curve method, while the mean AUROC values were calculated using roc\_auc\_score.

Figure 5 also reaffirms the conclusion obtained from Table 5 as the AUROC curve from the MLP classifiers seems to be the furthest away from chance and to have the widest curve.

Additionally, a statistical analysis (One-Way ANOVA statistical test with Tukey’s *post Hoc*) of the top 3 best performing algorithms obtained from original-dataset and their corresponding algorithms from modified-dataset was performed in order to determine their statistical difference. Moreover, these results provided additional insight into which model would be best suited to be implemented in the stress detection application. As shown in Table 6, it is evident that the machine learning algorithms trained on original-dataset are statistically the better performing models ( $P < 0.05$ ) which is expected due to the fact that more information on the dataset is being fed to the model during training.

**Table 6.** Statistical Evaluation of the machine learning models. Numbers 1 and 2 correspond to original-dataset and modified-dataset respectively.

Model A	Model B	mean(A)	mean(B)	diff	se	p-tukey <sup>1</sup>
GB 1	GB 2	0.8524596	0.730697	0.1217626	0.00784991	<b>0.001</b>
GB 1	MLP 1	0.8524596	0.8251284	0.0273312	0.00784991	<b>0.01056949</b>
GB 1	MLP 2	0.8524596	0.7519522	0.1005074	0.00784991	<b>0.001</b>
GB 1	RF 1	0.8524596	0.852724	-0.0002644	0.00784991	0.9
GB 1	RF 2	0.8524596	0.7682086	0.084251	0.00784991	<b>0.001</b>
GB 2	MLP 1	0.730697	0.8251284	-0.0944314	0.00784991	<b>0.001</b>
GB 2	MLP 2	0.730697	0.7519522	-0.0212552	0.00784991	0.08765353
GB 2	RF 1	0.730697	0.852724	-0.122027	0.00784991	<b>0.001</b>
GB 2	RF 2	0.730697	0.7682086	-0.0375116	0.00784991	<b>0.001</b>
MLP 1	MLP 2	0.8251284	0.7519522	0.0731762	0.00784991	<b>0.001</b>
MLP 1	RF 1	0.8251284	0.852724	-0.0275956	0.00784991	<b>0.00951206</b>
MLP 1	RF 2	0.8251284	0.7682086	0.0569198	0.00784991	<b>0.001</b>
MLP 2	RF 1	0.7519522	0.852724	-0.1007718	0.00784991	<b>0.001</b>
MLP 2	RF 2	0.7519522	0.7682086	-0.0162564	0.00784991	0.31310346
RF 1	RF 2	0.852724	0.7682086	0.0845154	0.00784991	<b>0.001</b>

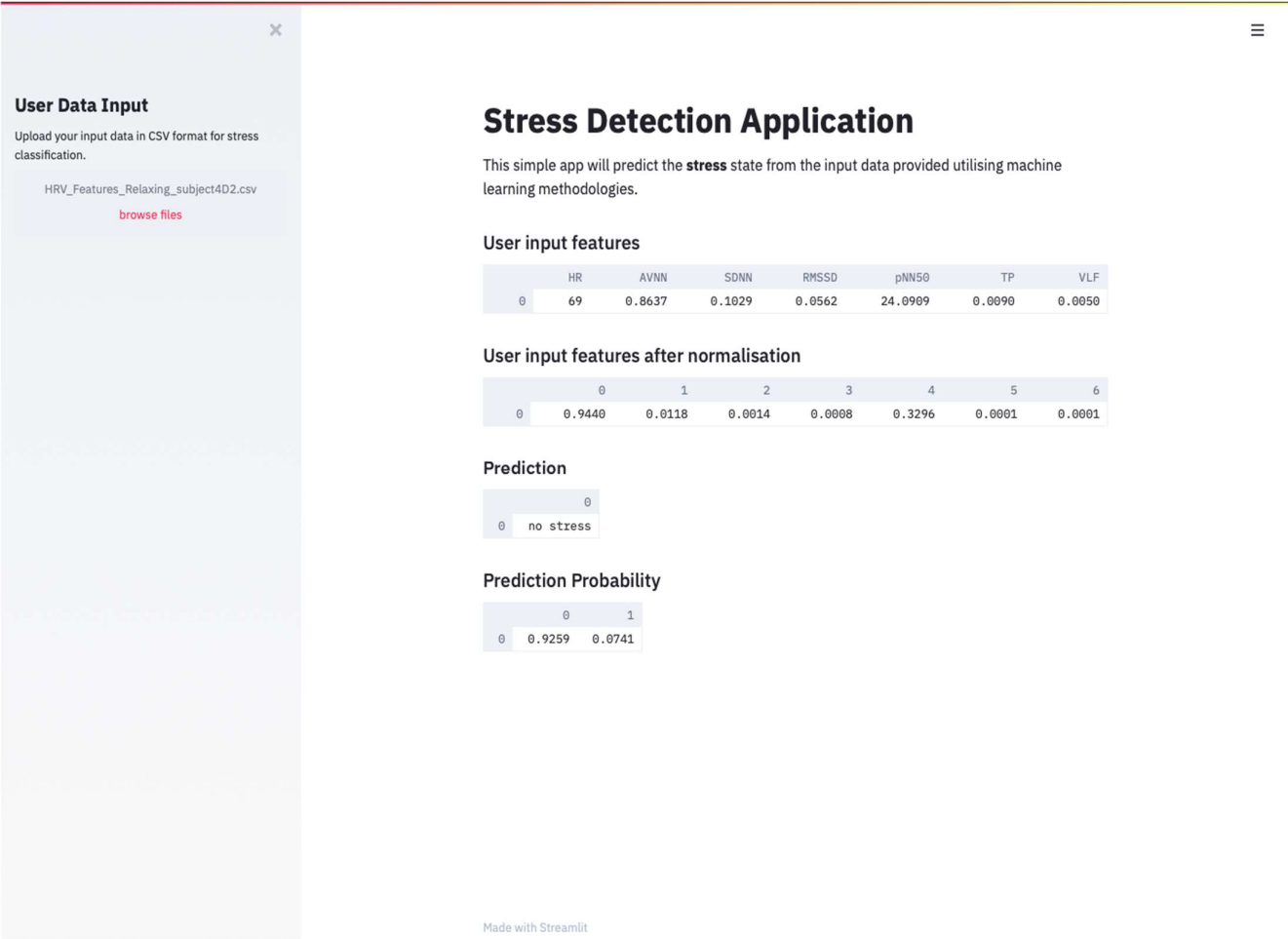
<sup>1</sup>P values in bold represent statistical significance, where  $P < 0.05$ .

Furthermore, Table 6 indicates that there is no significant difference in the AUROC values between RF2 and MLP2 ( $P = 0.31$ ); therefore, MLP2 can be chosen with confidence as the model that will be implemented in the stress detection web application. Additionally, another One-Way ANOVA statistical test with Tukey's *post Hoc* comparison was performed to determine whether there were statistical differences between the models and the Naïve Bayes baseline model. The results determined that there was a statistical difference between the AUROC means of the models and the baseline as the  $P = 0.001$  (not shown).

#### 4.3 Stress Classification from HRV measurements obtained from Apple Watch

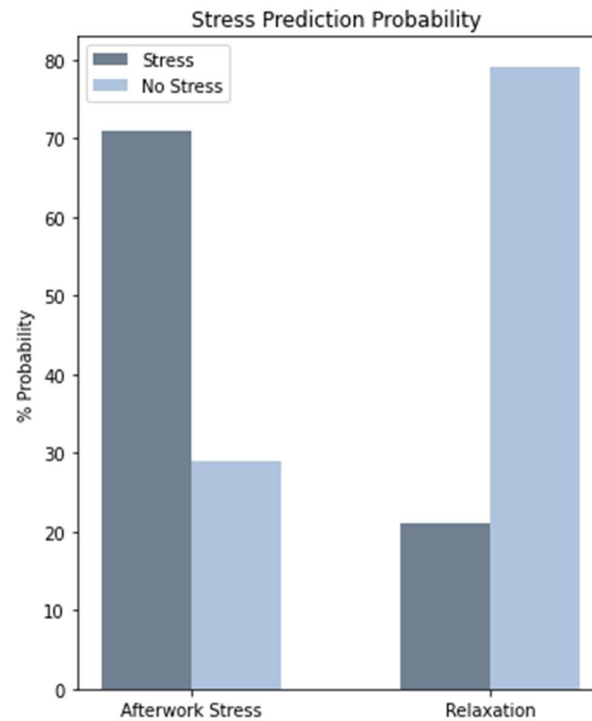
A simple web application that would perform stress classification on HRV data uploaded by the user (blind-dataset) was developed with the aim to analyse data extracted using wearable devices. The reference dataset for this application was modified-dataset which was used to validate the algorithms trained on the original-dataset. The aim of this process was to test the predictive power of the chosen model on data obtained from real participants. The application was developed in Python using the Streamlit framework and it is programmed in such way that the user can upload a csv format data, which will be first normalised and then classified as "stress" or "no stress" using the saved MLP model with Recall 80% and AUROC of 75%. Firstly, the application will prompt the user to insert the csv file in the side menu bar. Secondly, the backend code will normalise the input data, so all data instances are within the same range, and display the inserted and normalised data in a tabular format. Thirdly, the normalised data undergoes classification, and the results are displayed as Prediction Probability, shown in Figure 6.





**Figure 6.** User Interface of the Stress Detection Web Application developed using Streamlit.

After running the program with the input data derived from the volunteers, the prediction probabilities for the model to predict an instance as stress or no stress were recorded for the different stress scenarios. Figure 7 summarizes the results of this investigation in a bar chart presenting the mean prediction probabilities.



**Figure 7.** Mean Prediction Probability obtain from the stress detection app with volunteers input data, who were subjected to different stress conditions (after work stress and relaxation).

As displayed in Figure 7, the model was able to correctly classify a stress state with a prediction probability of  $71 \pm 0.1$  %. Additionally, it was able to achieve a prediction probability of  $79 \pm 0.3$  % when the model was presented with a relaxing situation.

## 5. Discussion

Stress has been identified to be one of the major causes of automobile crashes [8] and an important player in the development of cardiac arrhythmia [7]; therefore, it is important to be able to detect and measure stress in a non-invasive and efficient manner. In this study, to accomplish this, we address the stress detection problem by using traditional machine learning algorithms which were trained on ECG-derived HRV metrics obtained from automobile drivers [18-19].

In this paper, stress classification was performed mainly using HRV-derived features as studies have shown that HRV is impacted during changes in stress levels, given that it is highly controlled by the ANS [10]. Moreover, other investigations proved that RMSSD, AVNN and SDNN were evaluated as being the most reliable HRV metrics in distinguishing between stressful and non-stressful situations [25]. Those findings were also confirmed in this study as shown in Table 2, where AVNN, RMSSD, SDNN and were classified as the HRV features with the highest RFE feature importance score. Therefore, they were considered to be the features that contribute the most in the stress classification performance of the model.

Following hyperparameter tuning, we were able to produce stress classification models with high predictive power. As shown in Table 3, the best 3 model for the classification task imposed by original-dataset were MLP, RF and GB with AUROC of 83%, 85% and 85% respectively; thus, these classifiers have ~84% probability of successfully distinguishing between the stress and no stress class. In addition, MLP and RF gave Recall scores of 81% while GB of 80%; indicating that ~80% of the predicted positive instances are actual positives. Furthermore, these scores were statistically greater than the Naïve Bayes baseline model ( $P < 0.05$ ) as illustrated in Table 4.

There are very few studies performed on stress classification in drivers using HRV derived features [17-18], although each study took a different approach to the classification problem, the classification yielded similar results. For instance, [17] investigated KNN, SVM-RBF and Linear SVM as their potential classifiers for stress detection. Their results suggested that SVM with RBF kernel was the best performing model by giving an accuracy of 83% [17]. However, more extensive investigation is necessary to corroborate this finding by also considering other classification metrics.

In this paper we also aimed to develop a classification model that would classify stress from data obtained from wearable devices. For this purpose, the best classifiers trained on original-dataset were tested for the classification of modified-dataset which present features that mimics those derived from the wearable device. Table 5 demonstrates that the overall best model for the stress classification of HRV features derived from wearable-obtained RR intervals, is MLP with a AUROC of 75% and a Recall of 80%. This was determined based on the Recall score as in this stress classification task there is a high cost associated with False Negatives. For instance, if an individual's condition, which is actually stressed, is predicted as not stressed, the cost associated with this False Negative can be high especially in a medical or driving context which could then lead to a misdiagnosis or a car accident respectively. Therefore, it is imperative to select the model with the highest sensitivity.

Figure 6 shows the user interface (UI) of the simple stress detection web application. The purpose of this was simply to provide a visual UI to demonstrate how software functions. This could then be implemented into a mobile or car application where the user would be alerted what stress is detected and would prompt them to relax or take breaks.

The blind-dataset, obtained from the volunteers, served as a blind test for the MLP classifier in order to measure its predictive power on unseen data.

When classifying a stressful task, the web application was able to correctly classify stress conditions with a 71% prediction probability. Additionally, it was able to achieve a prediction probability of 79% when the model was presented with a relaxing state. However, it is important to further improve the model's performance by investigating multiple stress levels in order to obtain a more accurate stress detection.

## 6. Conclusions

In this paper, we developed a comparative study of different supervised machine learning models to determine which model can be used to analyse data extracted using wearable devices. The MLP model was considered to be ideal algorithm for stress classification due to its 80% sensitivity score. The predictive power of this classifier was found to be statistically greater compared to the baseline model created with the Naïve Bayes algorithm with a P value of 0.001. This model was then implemented in the stress detection application where stress can be classified from blind dataset of HRV features, extracted from real users using wearable devices under different stress conditions.

A benefit of this study is that there is a need for technologies that would monitor stress in drivers in order to reduce car crashes, as nearly 80% of road incidents are due to drivers being under stress. This project could be the initial steps for tackling this problem. In fact, the algorithm produced in this model could be implemented in smart cars. So, when drivers are experiencing episodes of stress, the automobile could switch to autopilot

as well as alert the driver of their state. This implementation could massively reduce traffic accidents as well as reduce the number of fatalities and injuries caused by car crashes.

However, the benefit of this study can also be extended to all applications in which it is important to monitor stress levels e. g. in physical rehabilitation post incident, in temporary or chronic anxiety, mental health disease as well as in many ageing conditions. The distribution of smart watches is growing in the population and people appreciate their functionalities. Therefore, wearable devices offer the big opportunity to extract health parameters without an uncomfortable and invasive approach.

We plan that future work should involve the improvement of the classification models by exploring a wider range of parameter values during the hyperparameter tuning process. Additionally, the Deep Learning approach could also be implemented in order to compare its performance in comparison to the supervised models used in this study.

Moreover, another future work we propose is the development a classifier that would be able to distinguish between different levels stress: high, medium and low. In addition to this, we suggest collecting new real-world ECG data, from which HRV features could be extracted, in order to gain a better insight on the predictive power of the models obtained in this study. This would also provide a more updated dataset compared to that used in this study, dated in 2005 [18]. As technologies have advanced, a more accurate ECG recording could be acquired; thus, making the classification more accurate and relevant to real world implementations.

Therefore, a natural evolution of this work will require the acquisition of a large dataset through smart watches and in an extensive number of tests involving human subjects e. g. through a driving simulator. Furthermore, will be important to test the model considering other domains focused to the elder and health care.

**Funding:** Authors of this paper acknowledge the funding provided by the EU Interreg 2 Seas Mers Zeeën AGEIn project (2S05-014) to support the work described in this publication.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Manchester Metropolitan University, UK, EthOS Application 24084, for the MSc thesis “Machine Learning for Stress Levels Classification from ECG Signals in Automobile Drivers”.

**Informed Consent Statement:** All subjects gave their informed consent for inclusion before they participated in the study and data have been anonymized.

**Data Availability Statement:** The main dataset used was public “PhysioNet [19]”, as described in the Materials and Methods section; only a few people were involved in the test to validate the feature extraction from apple watch in real condition and the modality of the experiments is reported in this work, therefore the tests can be replicated on further subjects.

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

## References

1. McLeod, S. A., 2010. What is the stress response. Simply Psychology. Available online: <https://www.simplypsychology.org/stress-biology.html> (accessed on 9/10/2020).
2. Moldagulova, A. and Sulaiman, R.B. Using KNN algorithm for classification of textual documents. In *8th International Conference on Information Technology (ICIT)*, IEEE, 2017, pp. 665-671.
3. Sato, T., Yamamoto, H., Sawada, N., Nashiki, K., Tsuji, M., Muto, K., Kume, H., Sasaki, H., Arai, H., Nikawa, T. and Taketani, Y. Restraint stress alters the duodenal expression of genes important for lipid metabolism in rat. *Toxicology* **2016**, 227(3), 248-261.
4. Salai, M., Vassányi, I. and Kósa, I. Stress detection using low-cost heart rate sensors. *Journal of healthcare engineering* **2016**.
5. Ahmed, M.U., Begum, S. and Islam, M.S., Heart rate and inter-beat interval computation to diagnose stress using ECG sensor signal. *MRTC Report* **2010**, 4.
6. Herman, J. Neural control of chronic stress adaptation. *Frontiers in behavioral neuroscience* **2013**, 7, p.61.
7. Bobade, Pramod, and M. Vani. Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data. In *Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, IEEE, 2020, pp. 51-57.

8. Rastgoo, M.N., Nakisa, B., Maire, F., Rakotonirainy, A. and Chandran, V. Automatic driver stress level classification using multimodal deep learning. *Expert Systems with Applications* **2019**, 138, p.112793.
9. Goel, S., Tomar, P. and Kaur, G. ECG Feature Extraction for Stress Recognition in Automobile Drivers. *Electronic Journal of Biology* **2016**, 12(2), pp.156-165.
10. Londhe, A.N. and Atulkar, M. Heart rate variability analysis: application overview. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018, pp. 1518-1523.
11. Reijmerink, I., van der Laan, M. and Cnossen, F. Heart rate variability as a measure of mental stress in surgery: a systematic review. *International Archives of Occupational and Environmental Health* **2020**, pp.1-17.
12. Khairuddin, M.K.N.B., Nakamoto, K., Nakamura, H., Tanaka, K. and Nakashima, S. Heart rate and heart rate variability measuring system by using Smartphone. In *5th Intl Conf on Applied Computing and Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSII-BCD)*, IEEE, 2017, pp. 47-52.
13. Dobbs, W.C., Fedewa, M.V., MacDonald, H.V., Holmes, C.J., Cicone, Z.S., Plews, D.J. and Esco, M.R. The accuracy of acquiring heart rate variability from portable devices: a systematic review and meta-analysis. *Sports Medicine* **2019**, 49(3), pp.417-435.
14. Shcherbina, A., Mattsson, C.M., Waggott, D., Salisbury, H., Christle, J.W., Hastie, T., Wheeler, M.T. and Ashley, E.A. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of personalized medicine* **2017**, 7(2), p.3.
15. Hernando, D., Roca, S., Sancho, J., Alesanco, Á. and Bailón, R., Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects. *Sensors* **2018**, 18(8), p.2619.
16. Taelman, J., Vandeput, S., Spaepen, A. and Van Huffel, S. Influence of mental stress on heart rate and heart rate variability. In *4th European conference of the international federation for medical and biological engineering*. Springer: Berlin, Heidelberg, **2009**, pp. 1366-1369
17. Munla, N., Khalil, M., Shahin, A. and Mourad, A. Driver stress level detection using HRV analysis. In *2015 international conference on advances in biomedical engineering (ICABME)*, IEEE, 2015, pp. 61-64.
18. Healey, J.A. and Picard, R.W. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* **2005**, 6(2), pp.156- 166.
19. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., Mietus, J.E., Moody, G.B., Peng, C.K. and Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, 101 (23), pp. e215-e220.
20. Zong W, Moody GB, and Jiang D. A robust open-source algorithm to detect onset and duration of QRS complexes. *Computers in Cardiology* **2003**, 30:737-740.
21. Li, K., Rüdiger, H. and Ziemssen, T. Spectral analysis of heart rate variability: time window matters. *Frontiers in neurology* **2019**, 10, p.545.
22. Gomes, P., Hugo Silva, and P. Margaritoff. pyHRV-Open-Source Python toolbox for heart rate variability, 2018.
23. Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification and scene analysis. Vol. 3. New York: Wiley, 1973.
24. Gandhi R. Support vector machine, *Introduction to machine learning algorithms*. Towards Data Science **2018**.
25. Amato, Flora, Nicola Mazzocca, Francesco Moscato, and Emilio Vivenzio. Multilayer perceptron: an intelligent model for classification and intrusion detection. In *2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pp. 686-691. IEEE, 2017.
26. Moridani, M. K., Z. Mahabadi, and N. Javadi. Heart rate variability features for different stress classification. *Bratislavske Lekarske Listy* **2020**, 121.9, 619-627.
27. Brownlee, J, 2020. Feature Selection in Python with Scikit-Lear. Machine Learning Mastery. Available online: <https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/> (accessed on 21/03/2021).
28. Sharaff A, Gupta H. Extra-tree classifier with metaheuristics approach for email classification. In *Advances in Computer Communication and Computational Sciences*. Springer: Singapore, **2019**, pp. 189-197.
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. *JMLR* **2011**, pp. 2825-2830.