Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Application of data mining algorithms for improving stress prediction of automobile drivers: A case study in Jordan

Wa'el Hadi [a,*], Nuha El-Khalili [b], May AlNashashibi [a], Ghassan Issa [c], Abed Alkarim AlBanna [c]

[a] Computer Information Systems, University of Petra, Jordan
[b] Software Engineering, University of Petra, Jordan
[c] Computer Science, University of Petra, Jordan

ARTICLE INFO

Keywords:
Data mining algorithms
Stress prediction
Feature selection
Oversampling

ABSTRACT

Driving daily through traffic congestion has been recognised as a major cause of stress. High levels of stress while driving negatively impact the driver's decisions which could potentially lead to accidents and other long-term health hazards. Accordingly, there is a great need to determine stress levels for drivers based on measuring and predicting the major causes (features or classes) that increase stress levels.

In this paper, the problem of predicting automobile drivers' stress levels, as experienced during actual driving, is investigated through the application of five different data mining algorithms, namely K-Nearest Neighbour (KNN), Decision Tree (J48), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN). An experiment was conducted on 14 drivers taking various routes in Amman – Jordan, with a wearable biomedical device attached to the driver to instantly collect physiological data. The collected data (dataset) is grouped into two different categories, namely 'Yes' to signify the presence of stress and 'No' to signify the absence of stress.

In order to efficiently apply data mining algorithms to the data set, oversampling was used to avoid the negative effect of driver samples with a lesser class on the prediction of stress.

The findings are evaluated in relation to stress prediction and accordingly contrasted alongside standard reference approaches that do not consider oversampling and/or feature selection using the Friedman rank test. The proposed approach, in combination with RF, was seen to surpass any others in terms of accuracy, AUC, specificity, and sensitivity. The accuracy, AUC, specificity, and sensitivity rates produced by RF utilising our proposed approach were 98.92%, 99.91%, 98.46%, and 99.36%, respectively.

1. Introduction

On a worldwide scale, road and driving safety are recognised as major issues. The World Health Organization (WHO) highlighted, in reports published in both 2015 and 2018, the increase in traffic-related fatalities from a total of 1.25 million in 2015 [1] to 1.35 million in 2018 [2]. The majority of these deaths were seen to be in low- and middle-income countries, where significant economic development has been seen to go hand-in-hand with a greater degree of motorisation and road traffic injury. Despite the high number of mortalities related to traffic accidents, most of these accidents are seen to be predictable and therefore preventable. Thus, there is a need for interventions that improve road safety. According to Direct Line & Brake Reports, the critical reason leading up to 94% of the crashes is the driver. Stress, using mobiles, tiredness, intoxication and aggressive driving are among the reasons behind road accidents [3]. Therefore, monitoring the human affective state can provide a wealth of valuable information to avoid traffic incidents whilst also delivering comfortable and safe driving.

Affective Computing is the interdisciplinary field utilising sensitive computing systems to recognise and understand human emotions, to subsequently provide human requests with smart responses. The newly introduced wearable and non-intrusive sensor technology delivers real-time physiological tools to monitor the cognitive states and human affective states. A number of research groups have sought to deliver tools and solutions to drivers to monitor their stress level [4–6]. In this work, a number of different physiological, physical and contextual features have been collected from automobile drivers during driving in Jordan, captured through a system that consists of a biomedical device and a

[*] Corresponding author. University of Petra Amman, Jordan.
E-mail address: whadi@uop.edu.jo (W. Hadi).

https://doi.org/10.1016/j.compbiomed.2019.103474
Received 25 July 2019; Received in revised form 14 September 2019; Accepted 26 September 2019
Available online 28 September 2019

mobile application, in an effort to predict stress levels. Physiological signals, including Electrodermal Activity (EDA), Electrocardiogram (ECG), Electromyogram (EMG) and Respiration (RESP), were captured for a total of 14 driving experiences. Furthermore, a number of other features were captured, including demographic data, time and weather condition data.

Since data was collected within Jordanian context, it is important to highlight that Jordan is among the high countries in the world in traffic accidents due to aggressive driving. It was reported by Public Security Directorate (PSD) that, in 2017, a total of 150,226 road traffic accidents were recorded, 5.6% of them caused fatalities, and 10% caused serious injuries. 97% of accidents were caused by human errors. Road accident fatalities account for 51% of fatalities in Jordan. These accidents are seen to incur costs equal to approximately 308 million dinars [7].

By completing an analysis in this work, the questions presented below can be answered:

1. Are data mining techniques suitable for automobile drivers' stress investigation?
2. Is the overall categorisation accuracy achieved through the algorithms seem to improve after the application of the feature selection methods?
3. In examining automobile drivers' stress and its prediction, what are the features identified as being the most important in predicting stress?

In an effort to answer the research questions, five commonly implemented data mining categorisations are assessed on an experimental basis, namely KNN, J48, SVM, ANN, and RF [8–10], evaluating the data captured from automobile drivers in Jordan. More detailed information about the data and the experimental findings can be seen in Sections 3 and 4, respectively.

With the questions above in mind, this work presents a number of valuable contributions, as below:

- Establishes features that estimate automobile drivers' stress.
- Explores the more commonly utilised data mining algorithms in relation to real stress data, as captured from Jordanian automobile drivers.
- Evaluates more widely implemented data mining algorithms on the same dataset in pursuing the objective to investigate the overall suitability of such adoption in estimating the stress levels of automobile drivers.
- Completes a thorough and in-depth experimental study pertaining to stress data.

The paper is organised as follows: Section 2 provides an overview of other works carried out in the field of stress analytics and the application of data mining so as to predict the stress levels of automobile drivers; Section 3 presents a discussion on the proposed methodology surrounding the prediction model and approaches implemented; Section 4 considers the results; whilst Section 5 presents the conclusions and future work to be carried out in this field.

2. Related works

Studies measuring stress have either used psychological measurements using surveys or physiological measurement of stress [11]. Questionnaires used to measure the perceived stress vary greatly and there is no agreement on a specific one as a standard. On the other hand, many studies in the last 30 years consider salivary cortisol level as the biomarker for measuring stress. This test requires the collection of saliva samples to be analysed in laboratory [11–13], which is not a feasible method of stress detection while driving. Therefore, physiological measurements such as blood pressure, heart rate, respiration, skin conductance and muscle activation (EMG) have been used in many studies to detect stress, since they are easier to measure, and they indicate the existence of stress.

A recent review covered the literature of measuring drivers' stress published between 1990 until 2017 has shown the variety of captured data, environmental settings of capturing them, and data analysis approaches used. Data collected were classified into three modalities: physiological, physical (e.g. facial expressions, vehicle dynamic data), and context (i.e. environmental parameters). The conclusion of this critical review is two folds. First, future work in this area need to consider the integration of different modalities to create a more robust stress detection system. Secondly, it is important to research methodological issues related to data collection and analysis when building such systems [14]. The authors addressed the second conclusion in Ref. [15]. And are addressing the first conclusion in this paper. Thus, the dataset collected for this work included most of the features that were reported in literature to detect drivers' stress level; and we implemented and compared five common data mining algorithms.

3. Methodology

The Methodology section provides an overview of the five phases of the proposed methodology for this work (see Fig. 1): Data collection,

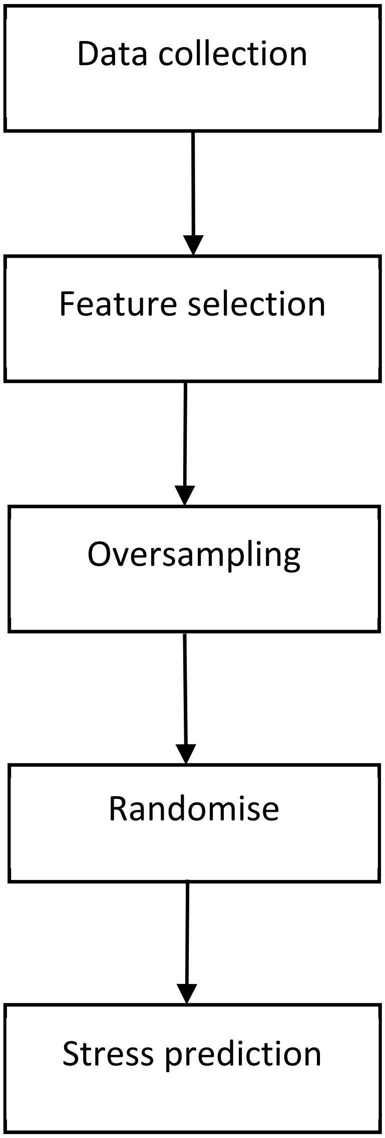

Fig. 1. Proposed methodology.

Feature selection, Oversampling, Randomise, and Stress prediction. The key components of each of the phases can be seen below.

3.1. Data collection

The data set used in this paper has been collected within a two-year project from 2017 till 2018, where drivers participating in this study drove around different routes in Jordan at different times of the day under various weather conditions. The experiments presented in this project were classified into four data sets. This paper analysed set 4, which contains 14 experiments following different free routes. Participants were 4 males and 10 females, with age range from 18 to 59. Three participants from 18 to 29, three participants from 30 to 39, two participants from 40 to 49 and six from 50 to 59. Two experiments were done by participants who had less than a year of experience in driving, while the rest had more than four-years of experience in driving. One experiment was during rain showers, while all the rest were at good weather conditions. Other demographic information about participants are included in Table 1.

A five-step procedure has been applied to all experiments performed in this study as follows:

First, take the participant's consent for using the collected information for research purposes, and their approval to capture face images. In addition, the participant vow to follow traffic rules and regulations.

Second, the participant fills the pre-experiment survey in Table 2, which collects demographic information and information about the trip circumstances as shown below:

Third, install the physiological device on the participant's body, and prepare the car for taking road images and face images (in case of the participant's approval) for details of the collection application, refer to Ref. [15].

Fourth, the driver drives his/her car on the chosen route. The maximum trip length in this set was 45 min.

Fifth, immediately after the end of the trip, the driver views the images of the road and annotate them. We choose not to use concurrent annotation while driving, since it adds more cognitive load on the driver and prevent the driver from practicing more natural activities during driving e.g. talking to other passengers. Meanwhile, we used the road images to improve the accuracy of the recall, and participants were asked to express if they felt stress at that moment, not to evaluate the situation. Thus, we minimized the bias of rationalization and recall that is reported in Ref. [16] about post evaluation technique in general.

Table 1
Demographic information.

Item	Category	Number of participants
Driving Skill	Excellent	9
	Good	5
Number of driving hours daily	Less than an hour	2
	from 1 to 2 h	9
	More than 4	3
Number of accidents in general	from 2 to 4	1
	from 1 to 2	6
	Zero	7
Illnesses	Nothing	11
	High Blood Pressure	1
	Diabetes	2
Symptoms when feeling stressed	Accelerated breath	5
	Shortness in breath	2
	Nothing	7
Feeling fatigue before the experiment	A little	3
	nothing	11
Feeling stressed before the experiment	A little	1
	Nothing	13
Distraction during the experiment	People in the car	5
	Children	4
	Nothing	5
Experiments time	6–9 am	3
	9-1 am	2
	1–4 pm	6
	4–8 pm	3

Table 2
Collected information in the pre-experiment survey.

Demographic and personal information	Experiment circumstances
• Gender • Age • Number of Driving Years • Driving Skills evaluation • Daily Driving Hours • Number of Accidents in general • Last Year's Number of Accidents • Frequency Of feeling Stress while driving • Stress Symptoms felt • Driving Concentration (front only, all sides)	• Experiment Time • Weather condition • Distractions in the car • Pre-Stress felt before Experiment • Pre-Fatigue felt before Experiment

Participants used their own cars, with no car alterations and minimal obstruction to the driver. A mobile was attached to the car windshield to capture the road status with images every 2 s. Nexus 10 device (Mind-Media, The Netherlands) was used to collect physiological data of the driver including (electrocardiogram (ECG), electromyogram (EMG) of Trapezius muscle, galvanic skin response/skin conductance (GSR/SC), respiration rate). After each experiment, self-evaluation report of stress level was collected from the drivers by showing them the road images and asking them to express whether they felt stress or not at this point. Descriptions on the 30 features are shown in Table 3. Also, for more details of the used data collection pipeline refer to Ref. [15].

The group of stress data is stored in a.csv file made up of 30 different features linked to the prevalence of stress amongst automobile drivers spanning 10,840 data instances. This stress data is seen to comprise two target classes, namely stress and non-stress: in the case of the former, a total of 417 data instances are detailed, whilst for the latter, 10,423 instances are detailed.

3.2. Feature selection

Prior to devising a predictive model, two of the key steps to be taken include reducing the data and choosing the features. A categorisation model devised applying the whole set of data features could present inconsistent findings; therefore, it would be preferable to choose the most relevant set of features that can contribute towards achieving a greater degree of true positive rate whilst decreasing the false positive rate. Accordingly, there are a number of different feature selection approaches that can be used in establishing the most valuable features; such approaches belong to embedded methods, filter methods and wrapper methods [21]. Accordingly, in this paper, the wrapper approach [22] is adopted in mind of choosing the most appropriate features. Primarily, the dataset undergoes a wrapper-based feature-selection algorithm, referred to as Boruta [23], the functional package of which can be found in the WEKA tool [24]. Importantly, randomness is incorporated into the data by the algorithm, which causes shuffled copies across all features. Such features then undergo training through the use of RF classifier [25] in an effort to determine the most vital and value measure for each feature. Importantly, the feature is seen to be more significant when the value measure is higher. This stage is then repeated, with an assessment performed by the algorithm in consideration to whether or not the original feature demonstrates a larger relevance score than those identified through the shuffled copies. Importantly, when the most important features are chosen, the algorithm process is concluded, with irrelevant features in the dataset then disregarded. As has been noted, the wrapper selection approach establishes the most optimal mix of data features, preserving the most ideal subset. When utilised on the stress dataset, the wrapper approach takes

Table 3
Stress dataset features.

Feature name	Type	Description
Sensor-B:EEG	Physiological	Electroencephalography (EEG) is one of the most commonly used neuroimaging modalities to study brain functions and conditions. EEG measures the fluctuations of electrical fields due to en-masse neuronal activity at millisecond resolution [17,18]
Sensor-E:SC/GSR	Physiological	GSR, is the measurement of the electrical conductivity of the skin. During stress, the skin conductance and transpiration is expected to rise and become more variable. It reflects the level of psychological or physiological arousal, elicited by cognition or emotions. In other words, GSR refers to changes in sweat gland activity that are reflective of the intensity of our emotional state, otherwise known as emotional arousal [19].
Sensor-H:RSP	Physiological	RSP can be measured as the rate or volume at which an individual exchanges air in their lungs. RSP rate and depth of breath (Respiration Amp.) are the most common measures of respiration. Emotional arousal increases respiration rate while rest and relaxation decreases respiration rate [20].
’[H] Respiration Rate’	Physiological	
’[H] Respiration Amp.’	Physiological	
’Sensor-C:EMG (raw)’	Physiological	EMG is the electro-physiological measure of a muscle, which has a direct relationship with strength at which a muscle is tensed [20].
’[C] EMG (20–500 Hz)’	Physiological	
’[C] EMG Amp. (20–500 Hz)’	Physiological	
’[C] EMG Median Freq.’	Physiological	
’[B] Heart Rate’	Physiological	Heart Rate Variability (HRV) is one of the most applied training methods for stress management. HRV is the variation in time between consecutive heartbeats. The acceleration and deceleration of the heart rate reflects the body’s ability to self regulate and maintain homeostasis. HRV changes under influence of health, age and our psychophysiological state like during stress, relaxation, exercise but also depends on health and age. HRV is often used to promote emotional and physical wellbeing and enhance relaxation.
’[B] HRV Amp.’	Physiological	
’[B] HRV-LF Power’	Physiological	
’[B] HRV-HF Power’	Physiological	
’[B] HRV-LF/HRV-HF′	Physiological	
etime	Environmental	Time of Experiment
eweather	Environmental	Weather status during the experiment
distraction	Environmental	Distractions inside the automobile
prestress	Environmental	Feeling stressed before the experiment
prefatigue	Environmental	Feeling fatigue before the experiment
gender	Personal	Gender of driver
age	Personal	Age range of driver
NofDrivingYears	Personal	Number of driving years
DailyDrivingHours	Personal	Number of driving hours daily
DrivingSkill	Personal	Personal evaluation of driving skills
NoGeneralAcc	Personal	Number of accidents in general
LastYearNoAcc	Personal	Number of accidents in the previous year
illness	Personal	Type of illnesses the driver has
FreqencyOfStress	Personal	Feeling stressed while driving in general
StressSymptoms	Personal	Symptoms when feeling stressed
DrivingConcentration	Personal	Driver’s method of concentration (style) while driving

away a total of 16 irrelevant features. The features of significance that are selected from the dataset (notably totalling 14) include the following: sensor_B_EEG, sensor_C_EMG, sensor_E_SC/GSR, sensor_H_RSP, EMG Amp., Heart Rate, HRV Amp., HRV_LF_Power, HRV_HF_Power, prestress, age, NofDrivingYears, FrequencyofStress, and DrivingConcentration.

3.3. *Oversampling*

There are various challenges associated with binary classification (two classes) when dealing with datasets that are class-imbalanced; when the distribution of the categories are overly skewed, machine learning approaches are more likely to create classifiers that show a preference for the majority class, attributing the most common label to the majority of the test samples. Despite the fact that it is commonplace for this to result in a significant degree of accuracy, nonetheless, it can induce poor decision-making in line with the minority class being more expected to induce higher misclassification costs when contrasted alongside most applications’ majority class [26–30].

Accordingly, the adoption of oversampling has been decided, in order to decrease the effect of underlying automobile driver samples with a lower size on the prediction of stress. For the majority of datasets that are seen to be imbalanced, sampling approaches are seen to enhance the overall accuracy of the classifier [29,31]. It is important to recognise, however, that oversampling is not seen to include any new data and can therefore result in overfitting, whereas important samples to the learning stage might be removed through under sampling, meaning the most important samples can be missed by the classifier [31,32].

In order to overcome this disadvantage, the SMOTE algorithm is implemented in this study [29], which is recognised as the most widely implemented oversampling process. In this regard, the KNN method is applied, which chooses K nearest neighbours, joins them, and accordingly establishes the space’s synthetic samples. The algorithm then considers the feature vectors along with its closest neighbours, and calculates the between-vector distance: the difference is multiplied by random number between (0, 1), and is then incorporated back into the feature [32].

Lastly, in an effort to avoid the issues of overfitting, the resulting dataset comes to be subjected to a completely random filter approaches, which completes shuffling in the case of those instances that have been identified during the process.

3.4. *Stress prediction*

This study adopts five of the most commonly implemented categorisation approaches [8–10], namely KNN, J48, RF, SVM, and ANN, all of which are recognised in terms of their value in stress prediction [33,34]. Furthermore, the study examines the effects of utilising the various datasets obtained with oversampling on stress prediction for all the algorithms.

4. Experiments

The proposed methodology has undergone assessment from an experimental perspective, notably utilising the stress dataset, as highlighted in SubSection 3.1. The outcome of this experimentation is to arrive at a methodology that is not just practical in the design of stress prediction models, but also can outperform other alternative commonly implemented methodologies. Accordingly, three different approaches were applied to the dataset followed by performance measurement as follows (Fig. 2 Depicts these three approaches):

The first approach is to devise stress prediction models (for KNN, J48, SVM, ANN, and RF) with the direct application of the dataset without the inclusion of a feature selection step (NFs-NOvrS).

The second approach, is to employ feature selection step (Fs-NOvrS)

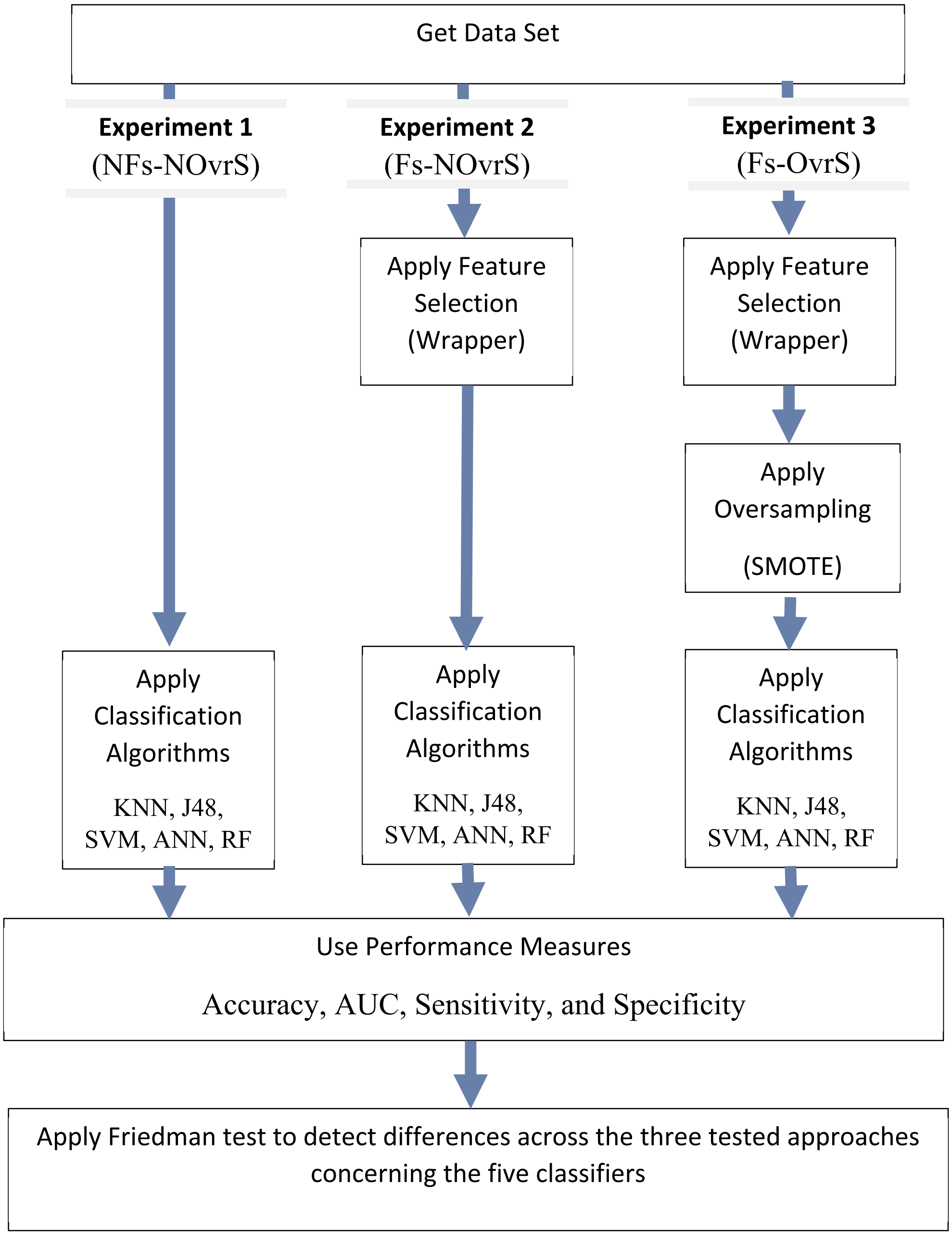
Get Data Set
Experiment 1 (NFs-NOvrS)
Experiment 2 (Fs-NOvrS)
Experiment 3 (Fs-OvrS)
Apply Feature Selection (Wrapper)
Apply Feature Selection (Wrapper)
Apply Oversampling (SMOTE)
Apply Classification Algorithms KNN, J48, SVM, ANN, RF
Apply Classification Algorithms KNN, J48, SVM, ANN, RF
Apply Classification Algorithms KNN, J48, SVM, ANN, RF
Use Performance Measures Accuracy, AUC, Sensitivity, and Specificity
Apply Friedman test to detect differences across the three tested approaches concerning the five classifiers

Fig. 2. Experiments diagram.

to identify the most valuable and important features in the prediction of stress amongst automobile drivers can be identified, with the effects of feature selection in stress dataset be examined in relation to classification.

Lastly, in the third approach, our own methodology is implemented (notably with the inclusion of feature selection and Oversampling). In other words, the stress dataset's minority class is oversampled with the application of the SMOTE algorithm (Fs-OvrS). Subsequently, the same five categorisation algorithms are applied. While working with this approach, however, it important to examine the effects of oversampling the stress dataset's class imbalance in categorisation.

In order to utilise the algorithms discussed in this work's experiments, the WEKA tool (Waikato Environment for Knowledge Analysis), as presented by Ref. [24]; was applied. WEKA is recognised as a landmark system in the field of machine learning and data mining and has been afforded much support and is therefore commonly used in business and academia. Furthermore, when it comes to studies based on data mining, WEKA is a widely implemented tool.

In this case, a 10-fold cross-validation was applied in order to assess the algorithms discussed in these experiments. The experiments are conducted in the case of an I7 machine encompassing a 3 GHz processor with 16 GB main memory, with Windows 10 utilised as the main OS.

4.1. *Performance of classifiers*

To complete the assessment of all of the classifiers, namely KNN, J48, SVM, ANN, and RF, a total of four different well-known evaluation measures were applied: Accuracy, AUC, Sensitivity, and Specificity. From a more conventional perspective, the commonly applied performance measure in the categorisation problem is that of Accuracy. Nonetheless, it ignores the probability estimations of classification in favour of class labels [31,35]. In a number of different study areas, and specifically in the case of biomedical applications, three further performance measures based on the ROC (Receiver Operator Curve) and confusion matrix are more commonly adopted: AUC (area under the ROC curve), Sensitivity, and Specificity [31,36]. In one regard, the area under the ROC curve (AUC) demonstrates the degree to which a classification model is able to differentiate between two different diagnostic groups (stress/no-stress). From a practical perspective, however, it is common for the AUC to be applied when a representative measure of discrimination is recognised as necessary; this might even replace the overall Accuracy as a performance measure [36]. In a second regard, however, the sensitivity also known as true positive rate, is defined as the proportion of instances which were classified as class stress, among all instances which truly have class stress in the dataset. Finally, specificity also known as true negative rate, is defined as the proportion of the instances which were classified as class no-stress, among all instances which truly have class no-stress.

A total of three indicators were applied for all of these metrics, namely Mean, Standard deviation (Std) and Rank. Importantly, the Mean and Std were determined in consideration to the experimental results alongside the various classification method configurations (i.e. different thresholds for the five classifiers). As has been discussed in the work of [37]; rank which is obtained by a Friedman rank test; is carried out to draw a contrast between the performance results (AUC/Accuracy/Sensitivity/Specificity) across the three tested approaches concerning the five classifiers.

The mean classification accuracies attained from different three approaches of all considered classifiers are shown in Fig. 3. The RF classifier utilising Fs-OvrS approach exhibited the highest accuracy rate between others. On the other hand, SVM classifier utilising Fs-OvrS approach produced the worst performance at 60.77%. In general, the predictive accuracy rates achieved from most considered data mining techniques are acceptable, except the accuracy rates attained from SVM classifier utilising Fs-OvrS approach. Further, it is clear from Fig. 3 that the overall classification accuracy achieved through the considered classifiers seem to degrade after the application of the Fs-OvrS approach. On the other hand, the overall categorisation accuracy achieved through the algorithms seem to improve after the application of the Fs-NOvrS approach. However, for datasets with imbalanced class distribution, these acceptable rates are misleading [38,39]. For this reason, other metrics like AUC and sensitivity must be evaluated.

Fig. 4 provides mean AUC attained from different three approaches of all considered classifiers. Results clearly indicate that the classifiers derived by the Fs-OvrS approach has the highest AUC rates. In particular, the mean AUC rates attained using the five different classifiers (KNN, J48, RF, SVM, and ANN) built on Fs-OvrS approach revealed higher rates (0.9728, 0.9746, 0.9991, 0.6025, and 0.9392) compared to the other approaches. Further, the figure indicates that after removing several irrelevant features from the original stress dataset (NFs-NOvrS), the mean AUC rates improved. In fact, the mean AUC rates improved in most classifiers utilising Fs-NOvrS and Fs-OvrS approaches. Only the mean AUC rate of ANN on Fs-NOvrS approach declined from 0.8561 to 0.825. In addition, the classification model built by RF classifier on Fs-OvrS approach attained a higher mean AUC rate than all considered classifiers at 0.9991. On the other hand, the classification model built by SVM on Fs-OvrS approach produces the lowest AUC rate at 0.6025. These results give an indication that data mining techniques can successfully be applied in automobile drivers' stress investigation.

According to the specificity rates in Fig. 5, The RF classifier utilising NFs-NOvrS approach exhibited the highest accuracy rate between others. On the other hand, SVM classifier utilising Fs-OvrS approach produced the worst performance at 33.82%. In general, the specificity rates achieved from most considered data mining techniques are acceptable, except the specificity rates attained from SVM classifier utilising Fs-OvrS approach.

Regarding the sensitivity rates in Fig. 6 which represent the measurement that we are interested in because it reflects the recall of the minority class, we can notice that the Fs-OvrS approach extremely improves the sensitivity of all classifiers. The RF classifier utilising Fs-OvrS approach exhibited the highest accuracy rate between others. On the other hand, ANN classifier utilising NFs-NOvrS approach produced the worst performance at 8.23%. it is clear from the figure that the classifiers derived by the Fs-OvrS approach has the highest sensitivity rates. In particular, the mean sensitivity rates attained using the five different classifiers (KNN, J48, RF, SVM, and ANN) built on Fs-OvrS approach revealed higher rates (0.9586, 0.9757, 0.9936, 0.8667, and 0.9123) compared to the other approaches. These results give an indication that oversampling process can extremely improve performance in classification of minority class in the stress dataset.

In summary, it is clear from the figures that the RF classifier performs better than KNN, J48, ANN and SVM, and that SVM is the poorest classifier for predicting automobile drivers' stress levels. Further, Fig. 4 indicates that the classifiers derived by the Fs-OvrS approach has the highest AUC rates and, therefore, better average performance.

4.2. Non-parametric tests

The Friedman test [37] is a common non-parametric test utilised for evaluating the performances of the multiple classifiers using multiple

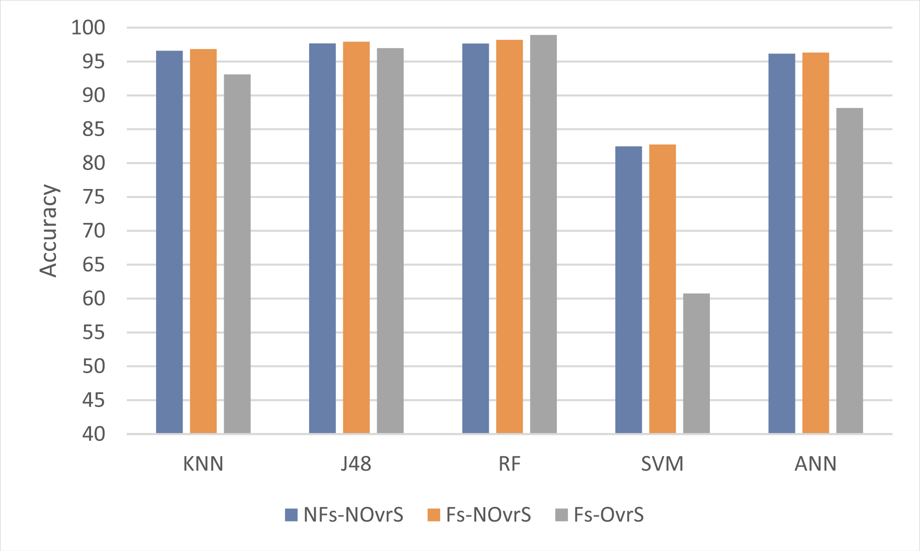

Fig. 3. Mean Accuracy results of all considered classifiers.

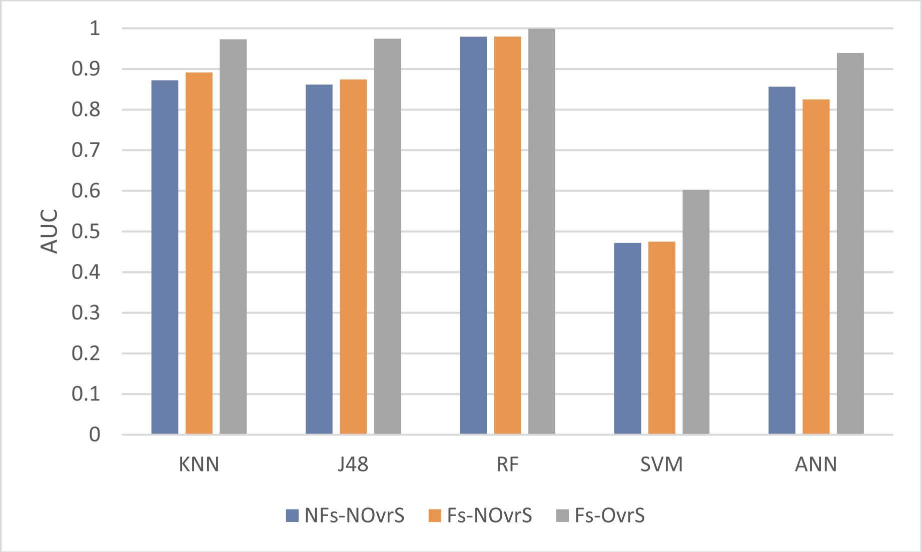

Fig. 4. Mean AUC results of all considered classifiers.

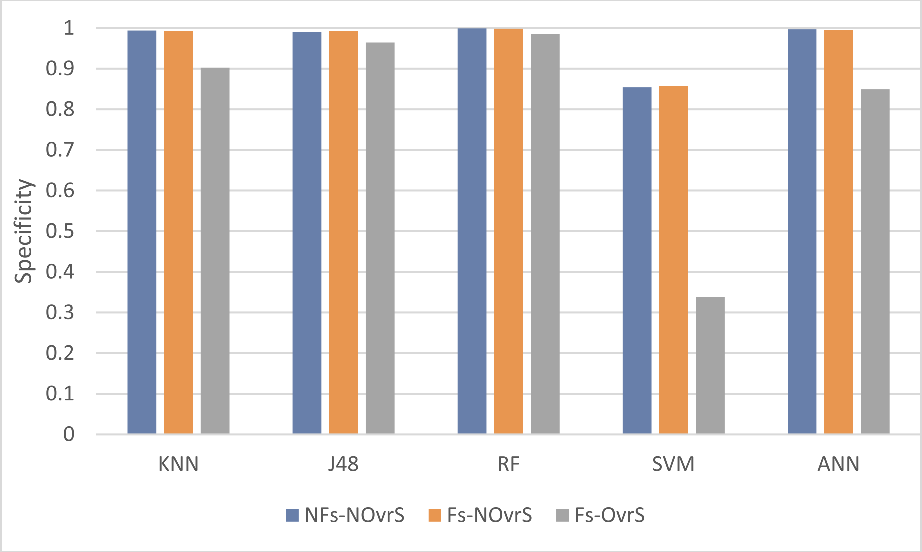

Fig. 5. Mean Specificity results of all considered classifiers.

datasets. Based on these performances, the algorithms are then ranked for each algorithm separately, the best performing algorithm gets rank 1, the second best gets rank 2, and so on. The Friedman test is calculated using equation (1):

$$X_F^2 = \frac{12N}{k\,(k+1)}\left[\sum_j^k R_j^2 - \frac{k\,(k+1)^2}{4}\right] \qquad (1)$$

Where N is the number of configurations used in the evaluation, k is the number of approaches to be compared, and R_j is the mean rank computed through the Friedman test for the ith approach.

As a result of the space restrictions and in an effort to facilitate the reader's analysis of the experiment, Table 4 shows the Mean of the attained experimental findings of the three approaches for each of the classifiers, i.e. KNN, J48, RF, SVM, and ANN. Notably, Appendix A provides a more comprehensive and in-depth breakdown of the simulation results.

As can be seen when examining the table, the adoption of AUC assessment measure and Mean and Std, Fs-OvrS approach, were found to deliver the most optimal results when compared alongside the other approaches.

In relation to the KNN classifier, a total of seven distinct neighbour configurations were applied through the experience (1–13 neighbours in a step of 2) whilst for each of the number of neighbours a total of 30 runs were carried out. In the case of the Friedman rank test, the average accuracy across all the considered configuration of neighbours for all of the applied approaches is represented. Using equation (1), $X_F^2 = 14$ was determined, with a p value = 0.000912 and a significance level of α = 0.05. As a result, the null equivalence hypothesis across all three approaches is then rejected. Thus, it can be concluded that there is a significant difference between the performances of all the approaches. When contrasting all of these approaches for a 5% significance level adopting the Nemenyi test [37], there was then the ability to arrive at CD = 1.2524, where CD is recognised as being the critical value for the difference of mean ranks between best and worst approaches [37]. As a result, the performance of Fs-OvrS (best approach) is significantly different than NFs-NOvrS (worst approach) because the value of CD is smaller than the difference between the best and worst approaches. In

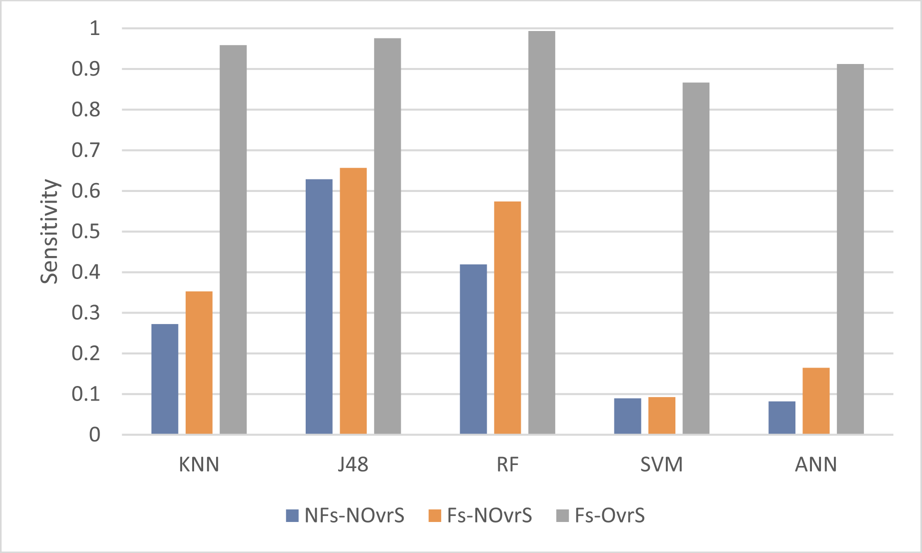

Fig. 6. Mean Sensitivity results of all considered classifiers.

Table 4
Achieved mean AUC results (Mean and Std) of all considered classifiers.

Classifier	NFs-NOvrS			Fs-NOvrS			Fs-OvrS		
	AUC	Std	Rank	AUC	Std	Rank	AUC	Std	Rank
KNN with different neighbours (1–13)	0.8723	0.03275	3	0.8914	0.02959	2	0.9728	0.00329	1
J48 with different confidence values (0.10–0.60)	0.8614	0.05115	3	0.8739	0.03574	2	0.9746	0.00507	1
RF with different trees (50–250)	0.9793	0.01092	2.8	0.9799	0.01169	2.2	0.9991	0.00041	1
SVM with different gamma values (0.10–0.50)	0.4719	0.05078	2.60	0.4749	0.04746	2.40	0.6025	0.07847	1
ANN with different hidden layer sizes (5–25)	0.8561	0.03288	2	0.8250	0.03498	3	0.9392	0.00694	1

relation to Table 4, Fs-OvrS approach was seen to achieve more optimal outcomes when compared with those approaches that did not utilise oversampling, which may be recognised as widely-used approaches. Furthermore, the Fs-NOvrS approach was seen to demonstrate greater performance when compared with the NFs-NOvrS approach.

Regarding the J48 classifier, the same examination was carried out, with the AUC results for the J48 classifier detailed in Table 4. Mean and Std were found to represent the most promising findings by all approaches associated with a particular threshold. Furthermore, Fs-OvrS approach was also found to demonstrate improved results when compared with the other approaches. In order to generate the value of the rank mean, each of the groups was seen to align with one of the six various number of confidence values utilised, i.e. 0.10–0.60 in a step of 0.10, and for each configuration, there was the completion of 30 different runs.

Once again, for this specific scenario $X_F^2 = 12$ underwent calculation, at a significance level of $\alpha = 0.5$, with a p value = 0.002479. As a result, the null hypothesis of equivalence across all approaches is not accepted. Thus, it can be concluded that there is a significant difference between the performances of all the approaches. When contrasting all of the approaches at a 5% significance level with the application of the Nemenyi test [37], the CD could then be attained, which equated to 1.3527. As a result, the performance of Fs-OvrS approach is significantly different than NFs-NOvrS approach. In regards Table 4, Fs-OvrS approach was seen to demonstrate greater performance when compared with the other approaches, utilising Rank, Mean, and Std as indicators.

When considering the RF classification model, the same examination was carried out, with the AUC results for the RF classifier detailed in Table 4. Mean and Std were found to represent the most promising findings by all approaches associated with a particular threshold. Furthermore, Fs-OvrS approach was also found to demonstrate improved results when compared with the other methods. In order to generate the value of the rank mean, each of the groups was seen to align with one of the five various number of trees utilised, i.e. 50–250 in a step of 50, and for each configuration, there was the completion of 30 different runs.

Once again, for this specific scenario $X_F^2 = 8.40$ underwent calculation, at a significance level of $\alpha = 0.5$, with a p value = 0.014996. As a result, the null hypothesis of equivalence across all approaches is not accepted. Thus, it can be concluded that there is a significant difference between the performances of all the approaches. When contrasting all of the approaches at a 5% significance level with the application of the Nemenyi test [37], the CD could then be attained, which equated to 1.4818. As a result, the performance of Fs-OvrS approach is significantly different than NFs-NOvrS approach. In regards Table 4, Fs-OvrS approach was seen to demonstrate greater performance when compared with the other approaches, utilising Rank, Mean, and Std as indicators.

Regarding the SVM classifier, the same examination was carried out, with the AUC results for the SVM classifier detailed in Table 4. Mean and Std were found to represent the most promising findings by all approaches associated with a particular threshold. Furthermore, Fs-OvrS approach was also found to demonstrate improved results when compared with the other approaches. In order to generate the value of the rank mean, each of the groups was seen to align with one of the five various number of gamma values utilised, i.e. 0.10–0.50 in a step of 0.10, and for each configuration, there was the completion of 30 different runs.

Once again, for this specific scenario $X_F^2 = 7.60$ underwent calculation, at a significance level of $\alpha = 0.5$, with a p value = 0.022371. As a result, the null hypothesis of equivalence across all approaches is not accepted. Thus, it can be concluded that there is a significant difference between the performances of all the approaches. When contrasting all of the approaches at a 5% significance level with the application of the Nemenyi test [37], the CD could then be attained, which equated to 1.4818. As a result, the performance of Fs-OvrS approach is significantly different than NFs-NOvrS approach. In regards Table 4, Fs-OvrS approach was seen to demonstrate greater performance when compared with the other approaches, utilising Rank, Mean, and Std as indicators.

Finally, regarding the ANN classifier, the same examination was carried out, with the AUC results for the ANN classifier detailed in Table 4. Mean and Std were found to represent the most promising findings by all approaches associated with a particular threshold. Furthermore, Fs-OvrS approach was also found to demonstrate improved results when compared with the other approaches. In order to generate the value of the rank mean, each of the groups was seen to align with one of the five various number of neurons values utilised, i.e. 5–25 in a step of 5, and for each configuration, there was the completion of 30 different runs.

Once again, for this specific scenario $X_F^2 = 10.0$ underwent calculation, at a significance level of $\alpha = 0.5$, with a p value $= 0.006738$. As a result, the null hypothesis of equivalence across all approaches is not accepted. Thus, it can be concluded that there is a significant difference between the performances of all the approaches. When contrasting all of the approaches at a 5% significance level with the application of the Nemenyi test [37], the CD could then be attained, which equated to 1.4818. As a result, the performance of Fs-OvrS approach is significantly different than Fs-NOvrS approach. In regards Table 4, Fs-OvrS approach was seen to demonstrate greater performance when compared with the other approaches, utilising Rank, Mean, and Std as indicators.

4.3. Discussion

Most of the extensive experimental results of all considered classifiers have proven that the Fs-OvrS approach preserved a higher predictive performance than Fs-NOvrS and NFs-NOvrS approaches on predicting automobile drivers' stress levels. The Fs-OvrS approach was capable of reducing the bias in the minority instances and helps shift the decision boundary towards the minority instances that are not easy to classify. This finding is consistent with previous studies [40,41], which have found that the oversampling approach improves the classification performance. Also, all the non-parametric tests have proven that the performance of Fs-OvrS approach is significantly different than Fs-NOvrS and NFs-NOvrS approaches.

In terms of sensitivity measure, the results indicated that the Fs-OvrS approach extremely improves the sensitivity of all classifiers. This is mainly because of the fact that the oversampling process improved the predictions of the true positive instances with a slight decrease in the true negative instances' predictions. This finding is consistent with previous studies [38,41], which have found that introducing oversampling methods improved the sensitivity rates of all considered classifiers with a slight reduce in the accuracy and specificity.

Similarly, after removing several irrelevant features from the original stress dataset, most of the results of all considered classifiers indicate that the Fs-NOvrS approach outputted performance rates are better than NFs-NOvrS approach in terms of AUC, accuracy, sensitivity, and specificity measures. Accordingly, the findings showcased here are aligned with those works carried out in the past [42–46], which have similarly suggested that classification AUC could be increased through feature selection when the irrelevant features are removed.

Furthermore, Table 4 highlights that the approach with feature selection (Fs-NOvrS and Fs-OvrS) showed greater performance than the approach without feature selection. More specifically, in the case of KNN utilising the approach Fs-NOvrS and Fs-OvrS approaches, higher AUC rates equating to 0.0191 and 0.1005, respectively, were arrived at when compared with the NFs-NOvrS approach. In the case of the J48 utilising dataset Fs-NOvrS and Fs-OvrS approaches, higher ACU rates equal to 0.0125 and 0.1132 were determined when compared with the NFs-NOvrS approach. Also, in the case of the RF utilising dataset Fs-NOvrS and Fs-OvrS approaches, higher AUC rates equal to 0.0006 and 0.0192 were determined when compared with the NFs-NOvrS approach. In the case of the SVM utilising dataset Fs-NOvrS and Fs-OvrS approaches, higher ACU rates equal to 0.003 and 0.1306 were determined when compared with the NFs-NOvrS approach. In addition, in the case of the ANN utilising dataset Fs-OvrS approach, higher ACU rates equal to 0.0831 was determined when compared with the NFs-NOvrS approach. However, when utilising dataset NFs-NOvrS approach, higher ACU rates equal to 0.0311 was determined when compared with the Fs-NOvrS approach.

Furthermore, as can be seen from all experiments, the conclusion can be drawn that all of the algorithms taken into account are seen to generate acceptable classification performance rates; the stress dataset relevance in relation to the features are therefore selected. Moreover, all algorithms may prove valuable and appropriate approaches to dealing with the problem of investigating the stress of automobile drivers.

Lastly, regardless of the works highlighted in Section 2.2 not directing attention to the same topic as this work's suggested approach, and owing to the fact to that direct contrast is therefore not possible, in this work, the best approach (Fs-OvrS approach), the accuracy values for all algorithms (KNN, J48, RF, SVM, and ANN) is recognised as falling within the same range as those results already documented in other literature. In the case of KNN, J48, RF, SVM and ANN, accuracy findings were documented as 90.74, 97.92, 95.83, 77.75 and 81.97, respectively. Moreover, when drawing a contrast between accuracy stable states across classifiers, RF was seen to be a better approach than KNN, J48, SVM, and ANN. Accordingly, it is stated in this work, supported by other works [47]; Calle et al., 2011; [10], that the most optimal classifier when it comes to accuracy is that of RF. This is mainly because of the fact that building of multiple decision trees during training step helps in better prediction.

5. Conclusions and future work

The problem pertaining to the investigation of stress levels as exhibited by vehicle drivers whilst driving is a key area for study, particularly in developing countries. Academics and researchers have applied a number of different algorithms when it comes to dealing with this issue, including SVM, Naïve Bayes, ANN, and RF. Throughout this paper, an approach taking into account feature selection and oversampling has been proposed in an effort to decrease the effects of underlying driver samples with lower class on stress prediction, with investigation directed towards performance alongside KNN, J48, SVM, ANN, and RF algorithms in predicting the stress levels of automobile drivers in Jordan, as a case study. Physiological data was gathered from a sample encompassing 14 drivers whilst driving, notably with the use of a biomedical device with the capacity to measure all physiological data and accordingly deliver a number of different insights as measured from the raw data. More specifically, a total of 30 features, including physiological, demographic and weather condition data, as examples, were gathered as the sample were driving.

The findings obtained were evaluated using accuracy, AUC, specificity, and sensitivity performance measures. In order to draw a contrast between the findings secured pertaining to the five applied approaches for all classifiers (KNN, J48, SVM, ANN, and RF), the Friedman rank test was employed, as suggested in the study by Ref. [37]. The approach proposed, in combination with the RF classifier, providing more optimal findings than the other two implemented approaches relating to all performance indictors previously highlighted, thus supporting the conclusion that this approach delivers greater performance rates and suitability in relation to design stress prediction models in the Jordanian context, in line with the characteristics discussed.

With such findings, the conclusion can be drawn that data mining algorithms may prove to be a valuable and effective approach when it comes to managing the problem of investigating the stress levels of automobile drivers in the Jordanian context. The feature selection approach, when implemented alongside the stress dataset, eliminates a total of 16 irrelevant features. The fourteen significant features identified from the original dataset include the following: sensor_B_EEG, sensor_C_EMG, sensor_E_SC/GSR, sensor_H_RSP, EMG Amp., Heart Rate, HRV Amp., HRV_LF_Power, HRV_HF_Power, prestress, age, NofDrivingYears, FrequencyofStress, and DrivingConcentration.

When considering further work in this domain, there are two potential directions to be explored: the adoption of the suggested approach in line with other biomedical and non-biomedical categorisation issues; and the extension of this dataset so as to encompass additional countries and additional features, including the driving of vehicles on fixed routes, road type, and the surrounding region.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgment

The authors would like to acknowledge the support provided by the Deanship of Scientific Research at University of Petra for funding this work through project No. 4/3/2017.

Appendix A

Table 5
Achieved AUC results (Mean and Std) using KNN classifier with different neighbours (1–13)

Number of neighbours	NFs-NOvrS			Fs-NOvrS			Fs-OvrS		
	AUC	Std	Rank	AUC	Std	Rank	AUC	Std	Rank
1	0.7952	0.03847	3	0.8287	0.04213	2	0.9499	0.00434	1
3	0.8509	0.03939	3	0.8863	0.03144	2	0.9737	0.00368	1
5	0.8743	0.03126	3	0.8997	0.02682	2	0.9774	0.00327	1
7	0.8907	0.03268	3	0.9022	0.02709	2	0.9778	0.00292	1
9	0.8945	0.03132	3	0.9060	0.02730	2	0.9777	0.00291	1
11	0.8982	0.02938	3	0.9078	0.02633	2	0.9771	0.00297	1
13	0.9021	0.02673	3	0.9094	0.02601	2	0.9763	0.00297	1
Mean	**0.8723**	**0.03275**	**3**	**0.8914**	**0.02959**	**2**	**0.9728**	**0.00329**	**1**

Table 6
Achieved AUC results (Mean and Std) using J48 classifier with different confidence values (0.10–0.60)

Confidence values	NFs-NOvrS			Fs-NOvrS			Fs-OvrS		
	AUC	Std	Rank	AUC	Std	Rank	AUC	Std	Rank
0.10	0.8669	0.05863	3	0.8864	0.03999	2	0.9760	0.00482	1
0.20	0.8701	0.04975	3	0.8777	0.03266	2	0.9750	0.00498	1
0.30	0.8584	0.04912	3	0.8731	0.03559	2	0.9742	0.00512	1
0.40	0.8535	0.05079	3	0.8707	0.03497	2	0.9742	0.00517	1
0.50	0.8532	0.05136	3	0.8689	0.03625	2	0.9742	0.00525	1
0.60	0.8661	0.04723	3	0.8667	0.03499	2	0.9739	0.00505	1
Mean	**0.8614**	**0.05115**	**3**	**0.8739**	**0.03574**	**2**	**0.9746**	**0.00507**	**1**

Table 7
Achieved AUC results (Mean and Std) using RF classifier with different trees (50–250)

Number of trees	NFs-NOvrS			Fs-NOvrS			Fs-OvrS		
	AUC	Std	Rank	AUC	Std	Rank	AUC	Std	Rank
50	0.9780	0.01114	2	0.9754	0.01591	3	0.9990	0.00046	1
100	0.9798	0.01029	3	0.9805	0.01105	2	0.9991	0.00041	1
150	0.9797	0.01096	3	0.9812	0.01067	2	0.9992	0.00039	1
200	0.9796	0.01107	3	0.9811	0.01063	2	0.9991	0.00040	1
250	0.9795	0.01114	3	0.9814	0.01023	2	0.9991	0.00039	1
Mean	**0.9793**	**0.01092**	**2.8**	**0.9799**	**0.01169**	**2.2**	**0.9991**	**0.00041**	**1**

Table 8
Achieved AUC results (Mean and Std) using SVM classifier with different gamma values (0.10–0.50)

Gamma values	NFs-NOvrS			Fs-NOvrS			Fs-OvrS		
	AUC	Std	Rank	AUC	Std	Rank	AUC	Std	Rank
0.10	0.4597	0.05751	3	0.4758	0.05207	2	0.6061	0.07716	1
0.20	0.4735	0.06131	3	0.4806	0.03495	2	0.6112	0.07275	1
0.30	0.4812	0.03588	2	0.4757	0.05061	3	0.5935	0.07402	1
0.40	0.4653	0.04631	3	0.4836	0.04575	2	0.5936	0.09032	1

(*continued on next page*)

Table 8 (*continued*)

Gamma values	NFs-NOvrS			Fs-NOvrS			Fs-OvrS		
	AUC	Std	Rank	AUC	Std	Rank	AUC	Std	Rank
0.50	0.4797	0.05291	2	0.4586	0.05392	3	0.6079	0.0781	1
Mean	**0.4719**	**0.05078**	**2.60**	**0.4749**	**0.04746**	**2.40**	**0.6025**	**0.07847**	**1**

Table 9
Achieved AUC results (Mean and Std) using ANN classifier with different hidden layer sizes (5–25)

Number of neurons	NFs-NOvrS			Fs-NOvrS			Fs-OvrS		
	AUC	Std	Rank	AUC	Std	Rank	AUC	Std	Rank
5	0.8481	0.03215	2	0.8260	0.03651	3	0.9080	0.00807	1
10	0.8621	0.03134	2	0.8202	0.03553	3	0.9370	0.00722	1
15	0.8552	0.03538	2	0.8225	0.03955	3	0.9463	0.00570	1
20	0.8571	0.02981	2	0.8298	0.03084	3	0.9508	0.00630	1
25	0.8578	0.03573	2	0.8267	**0.03246**	3	0.9541	0.00743	1
Mean	**0.8561**	**0.03288**	**2**	**0.8250**	**0.03498**	**3**	**0.9392**	**0.00694**	**1**

References

[1] WHO, Global status report on road safety 2015, Retrieved from, https://www.who.int/violence_injury_prevention/road_safety_status/2015/en/, 2015.

[2] WHO, Global status report on road safety 2018, Retrieved from, https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/, 2018.

[3] Brake, Direct line, Safe driving Part Five: driven to distraction, Retrieved from, http://www.brake.org.uk/assets/docs/dl_reports/DLreport3-DISTRACTION-pt1-Dec11.pdf, 2012.

[4] S. Barua, S. Begum, M.U. Ahmed, Supervised machine learning algorithms to diagnose stress for vehicle drivers based on physiological sensor signals, Stud. Health Technol. Inform. 211 (2015) 241–248. Retrieved from, http://www.ncbi.nlm.nih.gov/pubmed/25980876.

[5] A. Ghaderi, J. Frounchi, A. Farnam, Machine learning-based signal processing using physiological signals for stress detection, in: 2015 22nd Iranian Conference on Biomedical Engineering (ICBME), IEEE, 2015, pp. 93–98. https://doi.org/10.1109/ICBME.2015.7404123.

[6] M. Munoz-Organero, V. Corcoba-Magana, Predicting upcoming values of stress while driving, IEEE Trans. Intell. Transp. Syst. 18 (7) (2017) 1802–1811. https://doi.org/10.1109/TITS.2016.2618424.

[7] PSD, The statistics of traffic accidents, Retrieved from, https://www.psd.gov.jo/images/traffic/traffic2017.pdf, 2018.

[8] C. Jacobé de Naurois, C. Bourdin, A. Stratulat, E. Diaz, J.-L. Vercher, Detection and Prediction of Driver Drowsiness Using Artificial Neural Network Models, Accident Analysis & Prevention, 2017. https://doi.org/10.1016/j.aap.2017.11.038.

[9] C. Jacobé de Naurois, C. Bourdin, A. Stratulat, E. Diaz, J.-L. Vercher, Detection and Prediction of Driver Drowsiness Using Artificial Neural Network Models, Accident Analysis & Prevention, 2017. https://doi.org/10.1016/j.aap.2017.11.038.

[10] K. Nagaraj, B. Bhattacharjee, A. Sridhar, S. GS, Detection of phishing websites using a novel twofold ensemble model, J. Syst. Inf. Technol. 20 (3) (2018) 321–357. https://doi.org/10.1108/JSIT-09-2017-0074.

[11] S.J. Lupien, F. Seguin, How to measure stress in Humans?, Quebec, Canada. Retrieved from, https://www.stresshumain.ca/Documents/pdf/Mesures-physiologiques/CESH_howMesureStress-MB.pdf, 2007.

[12] D.H. Hellhammer, S. Wüst, B.M. Kudielka, Salivary cortisol as a biomarker in stress research, Psychoneuroendocrinology 34 (2) (2009) 163–171. https://doi.org/10.1016/j.psyneuen.2008.10.026.

[13] C. Kirschbaum, D.H. Hellhammer, Salivary cortisol in psychobiological research: an overview, Neuropsychobiology 22 (3) (1989) 150–169. https://doi.org/10.1159/000118611.

[14] M.N. RASTGOO, B. Nakisa, A. Rakotonirainy, V. Chandran, D. Tjondronegoro, A critical review of proactive detection of driver stress levels based on multimodal measurements, ACM Comput. Surv. 51 (5) (2018) 1–35. https://doi.org/10.1145/3186585.

[15] N. El-Khalili, M. Alnashashibi, W. Hadi, A.A. Banna, G. Issa, Data engineering for affective understanding systems, Data 4 (2) (2019) 52. https://doi.org/10.3390/data4020052.

[16] T. Rahman, M. Zhang, S. Voida, T. Choudhury, Towards accurate non-intrusive recollection of stress levels using mobile sensing and contextual recall, in: Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, ICST, 2014. https://doi.org/10.4108/icst.pervasivehealth.2014.254957.

[17] C.M. Michel, M.M. Murray, Towards the utilization of EEG as a brain imaging tool, Neuroimage 61 (2) (2012) 371–385. https://doi.org/10.1016/j.neuroimage.2011.12.039.

[18] L. Vézard, P. Legrand, M. Chavent, F. Faïta-Aïnseba, L. Trujillo, EEG classification for the detection of mental states, Appl. Soft Comput. 32 (2015) 113–131. https://doi.org/10.1016/j.asoc.2015.03.028.

[19] W. Boucsein, Electrodermal Activity, Springer US, Boston, MA, 2012. https://doi.org/10.1007/978-1-4614-1126-0.

[20] R.M. Stern, W.J. Ray, K.S. Quigley, Psychophysiological Recording, Oxford University Press, 2001.

[21] Fadi Thabtah, N. Abdelhamid, Deriving correlated sets of website features for phishing detection: a computational intelligence approach, J. Inf. Knowl. Manag. 15 (04) (2016) 1650042. https://doi.org/10.1142/S0219649216500428.

[22] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X.

[23] M.B. Kursa, A. Jankowski, W.R. Rudnicki, Boruta – a system for feature selection, Fundam. Inf. 101 (4) (2010) 271–285. https://doi.org/10.3233/FI-2010-288.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software, ACM SIGKDD Explor. Newsl. 11 (1) (2009) 10. https://doi.org/10.1145/1656274.1656278.

[25] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (1) (2012) 1063–1095. Retrieved from, http://dl.acm.org/citation.cfm?id=2503308.2343682.

[26] N. Abdelhamid, A. Ayesh, W. Hadi, Multi-label rules algorithm based associative classification. Parallel Processing Letters. https://doi.org/10.1142/S0129626414500017, 2014.

[27] A. Fernández, S. del Río, N.V. Chawla, F. Herrera, An insight into imbalanced Big Data classification: outcomes and challenges, Complex Intell. Syst. 3 (2) (2017) 105–120. https://doi.org/10.1007/s40747-017-0037-9.

[28] Wa'el Hadi, G. Issa, A. Ishtaiwi, ACPRISM: associative classification based on PRISM algorithm, Inf. Sci. 417 (2017) 287–300. https://doi.org/10.1016/j.ins.2017.07.025.

[29] S. Maldonado, J. López, C. Vairetti, An alternative SMOTE oversampling strategy for high-dimensional datasets, Appl. Soft Comput. 76 (2019) 380–389. https://doi.org/10.1016/j.asoc.2018.12.024.

[30] Fadi Thabtah, W. Hadi, N. Abdelhamid, A. Issa, Prediction phase IN associative classification mining, Int. J. Softw. Eng. Knowl. Eng. 21 (06) (2011) 855–876. https://doi.org/10.1142/S0218194011005463.

[31] M.S. Santos, P.H. Abreu, P.J. García-Laencina, A. Simão, A. Carvalho, A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients, J. Biomed. Inform. 58 (2015) 49–59. https://doi.org/10.1016/j.jbi.2015.09.012.

[32] He Haibo, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284. https://doi.org/10.1109/TKDE.2008.239.

[33] N. Keshan, P.V. Parimi, I. Bichindaritz, Machine learning for stress detection from ECG signals in automobile drivers, in: 2015 IEEE International Conference on Big Data (Big Data), IEEE, 2015, pp. 2661–2669. https://doi.org/10.1109/BigData.2015.7364066.

[34] V.C. Magaña, M.M. Organero, J.A. Fisteus, L.S. Fernández, Estimating the stress for drivers and passengers using deep learning, in: JARCA 2016, Almería, Spain, 2016, pp. 1–6.

[35] W. Hadi, Q.A. Al-Radaideh, S. Alhawari, Integrating associative rule-based classification with Naïve Bayes for text classification, Appl. Soft Comput. J. 69 (2018). https://doi.org/10.1016/j.asoc.2018.04.056.

[36] J. Huang, P.F. Miller, J.S. Wilson, A.J. de Mello, J.C. de Mello, D.D.C. Bradley, Investigation of the effects of doping and post-deposition treatments on the conductivity, morphology, and work function of poly(3,4-ethylenedioxythiophene)/poly(styrene sulfonate) films, Adv. Funct. Mater. 15 (2) (2005) 290–296. https://doi.org/10.1002/adfm.200400073.

[37] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[38] A.S. AlAgha, H. Faris, B.H. Hammo, A.M. Al-Zoubi, Identifying β -thalassemia carriers using a data mining approach: the case of the Gaza Strip, Palestine, Artif. Intell. Med. 88 (2018) 70–83. https://doi.org/10.1016/j.artmed.2018.04.009.

[39] I. Zliobaite, On the relation between accuracy and fairness in binary classification, in: The 2nd Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) at ICML'15, 2015. Retrieved from, http://arxiv.org/abs/1505.05723.

[40] G. Douzas, F. Bacao, F. Last, Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE, Inf. Sci. 465 (2018) 1–20. https://doi.org/10.1016/j.ins.2018.06.056.

[41] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, D. Khalili, The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes, Med. Decis. Mak. 36 (1) (2016) 137–144. https://doi.org/10.1177/0272989X14560647.

[42] S. Al-Harbi, a Almuhareb, a. Al-Thubaity, Automatic Arabic text classification, 9es J. Int. Analyse Statis. Données Textuelles (2008) 77–84.

[43] F. Thabtah, M.A.H. Eljinini, M. Zamzeer, W.M. Hadi, Naïve bayesian based on chi square to categorize Arabic data, in: Innovation and Knowledge Management in Twin Track Economies Challenges and Solutions - Proceedings of the 11th International Business Information Management Association Conference, IBIMA 2009 vols. 1–3, 2009.

[44] Fadi Thabtah, L. Zhang, N. Abdelhamid, NBA game result prediction using feature analysis and machine learning, Ann. Data Sci. 6 (1) (2019) 103–116. https://doi.org/10.1007/s40745-018-00189-x.

[45] G. Victo, S. George, V.C. Raj, A. Professor, Review on feature selection techniques and the impact of svm for cancer classification using gene expression profile, Int. J. Comput. Sci. Eng. Syst. 2 (3) (2011) 16–27. https://doi.org/10.5121/ijcses.2011.2302.

[46] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1997, pp. 412–420. Retrieved from, http://dl.acm.org/citation.cfm?id=645526.657137.

[47] X. Chen, M. Wang, H. Zhang, The use of classification trees for bioinformatics, Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 1 (1) (2011) 55–63. https://doi.org/10.1002/widm.14.