A New Two-step Ensemble Learning Model for Improving Stress Prediction of Automobile Drivers

May Al-Nashashibi[1], Wa'el Hadi[2], Nuha El-Khalili[3], Ghassan Issa[4], and Abed Alkarim AlBanna[1]
[1]Computer Science, University of Petra, Jordan
[2]Information Security, University of Petra, Jordan
[3]Software Engineering, University of Petra, Jordan
[4]School of IT, Skyline University, UAE

Abstract: *Commuting when there is a significant volume of traffic congestion has been acknowledged as one of the key factors causing stress. Significant levels of stress whilst driving are seen to have a profoundly negative effect on the actions and ability of a driver; this has the capacity to result in risks, hazards and accidents. As such, there is a recognized need to determine drivers' levels of stress and accordingly predict the key causes responsible for high levels of stress. In this work, the objective is centred on providing an ensemble machine learning framework in order to determine the stress levels of drivers. Moreover, the study also provides a fresh set of data, as gathered from 14 different drivers, with data collection having taken place during driving in Amman, Jordan. Data was gathered via the implementation of a wearable biomedical instrument that was attached to the driver on a continuous basis in order to gather physiological data. The data gathered was accordingly categorised into two different groups: 'Yes', which represents the presence of stress, whilst 'No' represents the absence of stress. Importantly, in an effort to circumvent the negative impact of driver instances with a minority class on stress predictions, oversampling technique was applied. A two-step ensemble classifier was developed through bringing together the findings from random forest, decision tree, and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) classifiers, which was then inputted into a Multi-Layer Perceptron neural network. The experimental findings highlight that the suggested framework is far more precise and has a more scalable capacity when compared with all classifiers in relation to accuracy, g-mean measures and sensitivity.*

Keywords: *Ensemble learning, stress prediction, oversampling, data mining algorithms.*

Received June 23, 2020; accept January 6, 2021
https://doi.org/10.34028/iajit/18/6/9

1. Introduction

Globally, driving and road safety are among the current and leading problems. World Health Organization (WHO) conducted a report in 2018 on road safety around the world [44]. This report demonstrated that the number of annual traffic-related fatalities is 1.35 million people. In Jordan, Public Security Directorate (PSD) reported that, in 2017, 150, 226 traffic accident occurred, where in 10,446 of these accidents human injuries resulted [31]. As such, research work is conducted around the world to predict stress levels among drivers and ways to reduce it while driving. One of the standard tasks of Data mining comprises predicting a target variable (class) in otherwise unseen data [1, 2, 3, 9, 17, 19]. Affective Computing is acknowledged as being an interdisciplinary sector combining computer science and engineering with other areas, namely cognitive science, psychology and physiology [13, 15].

Emotions of human beings (affective states) are complex psychophysiological constructs composed of many underlying dimensions [24]. Currently, there is a critical and significant need for affective computing, especially across sectors such as healthcare and education [23]. Various study groups have been focused on providing solutions and different tools across the transportation sector. Accordingly, as has been highlighted in the work of [40], in line with the American Highway Traffic Safety Administration, drivers' responses can be negatively impacted by high stress levels, especially in more serious situations. This is acknowledged as one of the most pressing and fundamental factors underpinning road accidents, including aggression, intoxication and tiredness, amongst others. When driving a vehicle, driver's emotions, also named human affective state [24], needs to be monitored and is high-value in terms of giving insight into how traffic incidents may be avoided, whilst also ensuring driving is safe and comfortable.

When looking to predict the stress levels of drivers through data mining classifiers, the key aim is centred on labelling the instances. Two different labels, namely stress and non-stress, are used in order to deliver improved approaches to the reduction of traffic accidents [13]. Number of different data mining classifiers, as logistic regression, naïve bayes, support vector machines, decision trees, random forest and neural networks, have been utilised in order to create predictive models aimed at identifying the stress levels of drivers. This has been investigated in a number of different works that provided a comparative examination into such classifiers on various stress data sets [6, 12, 16, 33, 36, 38, 45]. There do remain opportunities to create improved predictive frameworks that offer a greater degree of robustness to identify the stress levels of drivers. Furthermore, as far as the author is aware, ensemble framework was not implemented for the classification of stress levels in any other works.

This work directs emphasis to developing an ensemble framework geared towards predicting occurrences of stress for drivers. As a result, various physiological features have been collected from drivers during a period of driving in Jordan, as well as other features for 14 different individuals. Using this data set, work presented here suggests an innovative two-step ensemble learning model for predicting the stress levels of drivers of vehicles in line with three widely recognised rule-based classifiers, specifically random forest, decision trees (J48), and RIPPER (JRip). The results of such classifiers are then inputted into a Multi-Layer Perceptron neural network with the aim of presenting a global framework comprising an ensemble of outputs. This subsequently improves the overall performance of the classification when it comes to predicting drivers' levels of stress and decreases any potential of noise and bias. The performances of such classifiers undergo in-depth comparison in order to establish the most optimal classification framework.

As a result, this paper makes the following contributions:

- Collecting a new dataset from 14 drivers in Amman, Jordan, using a real-time data collection system comprised of a smartphone application and a wearable biomedical device attached to the driver.
- Designing a new two-step ensemble learning framework for predicting automobile drivers' stress by integrating three well-known classifiers with Multi-Layer Perceptron neural networks.

The rest of this paper is organised as follows: section 2 summarises the literature of data mining achievements with respect to stress prediction. Section 3 discusses the proposed methodology; section 4 evaluates the experimental results; whilst section 5 summarises the main results from this study and concludes the work.

2. Literature Review

A stress prediction approach was developed in the study of [45], using physiological signals. This approach developed an emotion identification system involving three key phases, namely experimental setup for physiological sensing, signal pre-processing for the extraction of affective features and affective recognition using a learning system. A total of four signals (Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), Pupil Diameter (PD) and Skin Temperature (ST)) undergo monitoring and examination in an effort to separate affective states amongst computer users. The approach applied Support vector machine method in order to carry out the supervised categorisation of affective states between 'relaxed' and 'stressed'. Their findings have shown that:

1. The physiological signals monitored have a clear alignment with changes in emotional state as exhibited by the study sample when the interaction environment was subject to stress stimuli;
2. The pupil diameter was the most valuable affective state signal when contrasted alongside the other three physiological signals monitored [45].

In the work of [8], the objective was to develop a straightforward device that could be worn, utilising a non-invasive physiological parameter-based sensors to facilitate the identification of stress levels and tiredness amongst drivers. Signals pertaining to skin conductance and the oximetry pulse of drivers were detailed across a number of different states as tiredness levels, with those aspects gathered then applied in order to develop multilayer perceptron neural network to fetch a high-value set of performance measures. A multilayer perceptron neural network, with two-state, classifier underwent analysis and examination through the application of Receiver Operating Characteristics (ROC). The Classifier performance was examined through the adoption of the ROC approach and independent validation method. The link between tiredness with Skin Conductance (SC) and Oximetry Pulse (PO) was established in the study.

In the research carried out by [11], a stress identification approach was discussed in line with fuzzy logic and two physiological signals, namely Galvanic Skin Response and Heart Rate. Rather than delivering a global stress categorisation method, their approach established an individual stress template, collecting insight into the behaviour of people in situations with varying levels of stress. The suggested approach has the ability to identify stress at a rate of 99.5%, with assessment carried out across 80 different subjects. The findings were further improved upon from other methods in the literature, as well as from different learning approaches such as Support Vector Machine, K-Nearest Neighbour (K-NN), Gaussian Mixture Modelling (GMM) and Linear Discriminant Analysis. Their suggested approach was recognised as well-aligned with real-time applications.

In the work carried out by [12], a feature selection approach was suggested in line with Principal Component Analysis (PCA), with their effectiveness assessed in regards correct rate and computational time through the adoption of five categorisation approaches, namely linear discriminant function, c4.5 induction tree, support vector machine, Naïve Bayes and K-NN. Their work highlighted the value of feature selection and the overall importance of approaches applied in precisely categorising levels of stress.

Moreover, in the case of [23], the random forest-based approach was implemented across physiological functional variables in an effort to take stress levels of drivers and accordingly categorise them. From a methodological aspect, this work's contributions are centred on considering physiological signals as functional variables, decomposed on wavelet basis and providing an approach to variable selection. On the applied side, the suggested approach delivers a 'blind' approach to the categorisation of stress levels of drivers, performing as the expert-based study in terms of misclassification rate. It also provides the ranking of physiological factors in line with their value in stress level categorisation. The findings secured imply that heart rate signals and electromyogram are not overall pertinent when contrasted alongside electrodermal and respiration signals.

The overall efficiency with which upcoming stress levels can be predicted with consideration to a number of different features-notably current driving behaviour, current stress levels, and the shape of the road-was assessed in the study by Munoz-Organero and Corcoba-Magana [29]. The researchers utilised various features, including the severity of the road curve and the positive kinetic energy, in order to predict the evolvement of stress levels in upcoming minutes. Data was captured from four different drivers with three different car models and a motorbike, with a total of more than 220 test drivers. Then, data was used by the researchers, who subsequently arrived at findings that support upcoming stress levels as being able to be accurately predicted on a single user basis (correlation r=0.99 and classification accuracy 97.5%). However, when evaluating different users at one time, the accuracy of prediction was found to be more limited (correlation r=0.92 and classification accuracy 46.9%).

3. Methodology

The methodology implemented across this study is considered in this section in mind of developing the prediction framework for drivers' stress levels. Our approach is a combination of the most widely applied data mining classifiers, referred to as Cross-Industry Standard Process for Data Mining (CRISP-DM) [37]. Figure 1 provides a block diagram of our approach. As also highlighted in the figure, the methodology can be seen to be made up of five key phases, namely problem understanding, data understanding, data preparation, modelling, and evaluation. The key phases of this approach are also explained in the following subsections.

3.1. Problem Understanding

Stress can be defined as the reaction of a person to environmental stimulus [5]. There are three types of stress:

- Acute stress, which is the reaction of the body to a temporary stress stimulus.
- Episodic acute, which is a classification of an acute stress that occurs frequently.

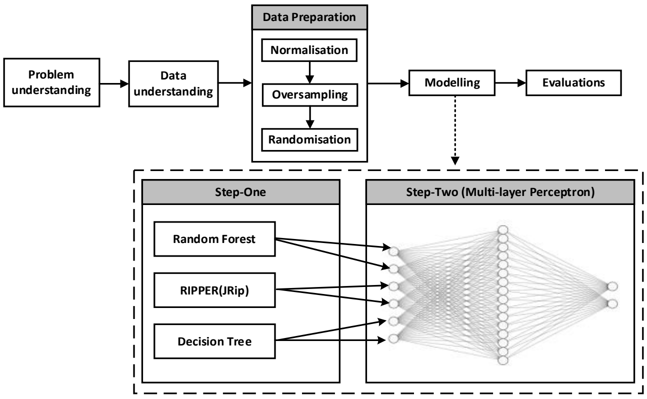

Figure 1. Proposed methodology.

- Chronic stress, which is a long-term stress that is caused by long-term factors.

The stress stimulus causes the activation of the Autonomic Nervous System (ANS) in the Peripheral Nervous System (PNS) to prepare the body to respond to the stimulus. Physiologically, this means the rise of adrenaline and cortisol levels [5]. This is why chronical stress has an effect on cardiovascular health, as well as on the immune system of the body [13].

Research has shown that, in order to identify the existence of stress, there are three main methods. The first method depends on the person's self-reporting, meaning only the perceived stress would be measured, which is notably subjective and therefore differs from one person to the other. This method is potentially time-consuming and difficult during daily life activities. If performed after the activity has been finalised, it suffers from retrospection and rationalisation biases that degrade the recall accuracy [41]. The second method is through the use of biomedical measurements to detect levels of hormones and cortisol in the body through blood, saliva or urine. However, this method is not convenient in the case of daily life activities [5]. The third method centres on the use of physiological data, which indicates the activation of the sympathetic part of the ANS, such as the heart rate, blood pressure, respiration, skin conductance, muscle activation, and skin temperature [14]. The lateral method has been the most commonly used in studies that seek to detect stress.

Studies show that stress while driving causes high levels of arousal, which, in turn, leads to distraction and consequently poor driving performance [14]. The aim of this study is to investigate the factors that cause stress for drivers.

3.2. Data Understanding

Many studies in the literature have investigated one or more factor that may cause stress for drivers, whether related to the personality of the driver [8, 14, 34, 35], or otherwise related to external factors [25, 36]. The authors have conducted an extensive literature review in [12] pertaining to related studies, which forms the base of this study. As shown in Table 1, there are many factors recognised as affecting the driver during periods of driving, which can be classified into different sources, as follows:

- Long term-personal factors: These are characteristics of the driver that always affect the driving experience, such as the driver's style of driving, or the driver's experience.
- Short-term personal factors: These are factors related to the status of the driver temporarily, which vary from one trip to another, such as the driver's fatigue or stress status before driving.
- External factors, such as weather and time of the day.
- Road status and features.

In addition, there is a need to collect the following:

- Physiological data, which indicates the existence or absence of stress
- Demographical data about participants.

The above data is also different in terms of time space since data occurs in three-time spaces:

- Long-term time space, such as driving experience, style of driving, etc.
- Short-term space, such as weather, time of trip, stress and fatigue before the trip.
- Momentarily, such as road image, physiological data, self-report of stress during driving.

Thus, in order to evaluate the effect of all the previous factors on the driver, an instance of a driver's record must include all these features. This leads to the need to repeat the least frequent data with the rate of the most frequent data. The most frequent data in this set are the ECG and EMG data, which are measured with a frequency 2048 signals/second. This leads to huge datasets that require a cluster of computers to analyse. Therefore, the decision was made to use a window of 2 seconds as the frequency of the data; therefore, the physiological data frequency had to be lowered either by taking the average or root mean square of the values within a window. It was noticed that averaging the values affect the shape of the ECG and EMG data, hence root mean square was adopted using the following equation:

$$RMS\ (within\ 2\ seconds\ window) = \sqrt{1/n}\ \sum_{k=0}^{n} x_k^2 \qquad (1)$$

Where n is the total number of instances taken, and x is the value of each instance within the window.

Table 1. Data collected.

Long term personal factors	**Short-term personal factors**
• Years of driving experience • Daily driving duration (hours) • Driving experience • The total number of accidents. • Number of accidents last year • Driving concentration style: has 3 values and detects the directions that the driver concentrates on while driving, the values are 1- All sides of the road 2- Front and back 3- Front only • Illness: detects whether the driver has long term illnesses, and we limited the values to the following illnesses (Blood pressure, diabetes, Psychological diseases as Anxiety attacks) • Stress feeling frequency while driving operation. • Symptoms noticed during stressful feelings.	• Stress felt before driving. • Fatigue felt before driving. • Self-report of stress during driving.
Road status and features	External factors
• Longitude • Latitude • Road images • Car speed	• Weather condition • Time of the day • Distractions inside the car
Demographical data	Physiological data
• Gender • Age	• Electrocardiogram (ECG), (Heart Rate, HRV amplitude, HRV\LF, HRV\HF, LF/HF) • Electromyogram (EMG) (EMG amplitude, Median Freq) • Respiration amplitude (Resp.) • Skin Conductance SC\GSR

A .csv file is used to detail the stress data, which notably comprises 30 different features associated with stress prevalence across the subjects utilised in this study, notably comprising a total of 10,840 samples. This data is classified into two target classes: non-stress and stress: in the stress group, 417 instances were identified, whilst 10,423 were seen to make up the latter category. This presents a clear imbalance across the data, equal to approximately 4%-96%.

3.3. Data Preparation

Prior to undertaking the assessment of the classification frameworks on the gathered data, there is a need for the data to undergo various actions in order to enhance the classification process and quality of such [22]. Throughout the current study, a total of three different pre-processing stages are utilised, namely data normalisation, data oversampling, and data randomisation. Such stages are explained further below.

3.3.1. Data Normalisation

Normalisation is a fundamental stage in data mining and needs to be applied prior to getting to grips with any classification model [22]. All features are scaled to the same interval by normalization; this is done to ensure that all features cover the same value range and to avoid the impact of features with a wider value range. All numeric features are normalised in this work using the Min–Max normalisation described in Equation (2) [22]. All numeric features are scaled to the range [0, 1] by the use of normalisation.

$$Normalized\ (f_i) = \frac{f_i - F_{min}}{F_{max} - F_{min}} \qquad (2)$$

Where F_{min}, F_{max} are minimum and maximum of the feature f, respectively.

3.3.2. Data Oversampling and Randomisation

There are a variety of difficulties associated with binary categorisation (two classes) when dealing with imbalanced datasets [10, 27]. As such, oversampling has been chosen in order to mitigate the effect of any underlying vehicle driver samples with a smaller size on stress prediction [18]. Across most of the datasets recognised as imbalanced, sampling methods are recognised as able to improve classifier accuracy overall [10, 28]. However, it is important to note that oversampling is not considered to include any new data, and thus can result in overfitting, while under-sampling may exclude important samples from the learning process, implying that the most useful samples may be overlooked by the classifier [27, 28].

In this research, we utilise Synthetic Minority Oversampling Technique (SMOTE) to overcome the effect of imbalanced datasets [18, 27]. The implementation of the SMOTE algorithm, which is recognized as the most widely used oversampling solution, is seen in this work [18] to better ensure that this downside is overcome. The K-NN approach is used in this case, which selects K nearest neighbours, joins them, and then defines the synthetic instances of the space. The algorithm then considers the feature vectors, as well as its nearest neighbours, and calculates the inter-vector distance: the difference is multiplied by a random number between (0, 1), which is then integrated back into the feature [27].

Finally, so as to circumvent the problems of overfitting, the dataset is then exposed to an entirely random filter method, which shuffles in the case of instances formed during the process.

3.4. Modelling

The first step in the modelling phase is to choose which candidate classifiers will be utilised in the investigation [9, 32]. This would encompass a review of previous related works and accordingly determine the frequently used classifiers that have previously been successful.

The proposed classification model is depicted in Figure 1, in which three different well-known classifiers are built from the training data, namely Random Forest, Decision Tree (J48), and RIPPER (JRip). The selection of Random Forest, J48, and JRip is because they produce if-then rules that are easy for the decision maker to interpret and understand. Such classifiers were frequently applied in the past literature, with the implementations of these classifiers publicly available. The predictions from these classifiers are fed to a Multi-Layer Perceptron neural network in an effort to produce a global model that contains of an ensemble of outputs. This not only enhances the classification performance of predicting stress level of automobile drivers, but also decreases any possibility of biased and noised decisions. Following multiple computations, six input layers, including the class outputs from selected classifiers and Multi-Layer Perceptron neural network, along with sixteen hidden layers, is found to provide better predictive performance. This choice of the number of the hidden layers is set based on the following formula: (input features number+output classes number)/2 [35].

A two-step ensemble framework has been developed through combining Random Forest, J48, and JRip with Multi-Layer Perceptron neural networks. This was decided through various considerations. First and foremost, in order to validate the predictive ability of the ensemble framework, in addition to the objective to improve the overall categorisation performance. Importantly, Multi-layer Perceptron neural networks are viewed as being the most suitable approach to the creation of the two-step framework, with their combined predictive performance seen to be comparably improved when contrasted alongside individual classifiers [30]. It is also important to take into account the fact that such networks are superlative classifiers in ensemble architecture, gathering knowledge from other previously created frameworks [39, 43]. In mind of all factors, in the present work, a Multi-Layer Perceptron neural networks framework is developed through bringing together the three classifiers.

Finally, when it comes to investigating the performance of the proposed model, three different standard classifiers were evaluated, namely, Random Forest, J48, and JRip. The objective of this phase is to find the best classifier with the highest overall efficiency.

3.5. Evaluations

The most used assessment metrics in the literature for biomedical applications such as accuracy, specificity, precision, g-mean, sensitivity and Area Under Curve (AUC) [4, 42] are used to measure the overall performance of the proposed classification system.

This stage's objective is to establish which categorisation approach performs the best when it comes to predicting the levels of stress amongst drivers.

The confusion matrix is widely recognised as one of the simplest approaches to performance measurement across any binary classifier [41]. As can be seen detailed in Table 2, two different output labels are utilised in the dataset, namely stress and no-stress, with other potential outcomes detailed as shown below:

Table 2. Confusion Matrix for stress prediction.

Actual class	**Predicted class value**	
	stress	no-stress
Stress	True positive (TP)	False negative (FN)
no-stress	False positive (FP)	True negative (TN)

- True Positive (TP): Number of instances predicted as class stress that have class stress.
- False Positive (FP): Number of instances predicted as class stress that have class no-stress.
- True Negative (TN): Number of instances predicted as class no-stress that have class no-stress.
- False Negative (FN): Number of instances predicted as class no-stress that have class stress.

The five evaluation measurements detailed below are further adopted in order to assess our stress prediction framework:

- Specificity, also known as true negative rate, is defined as the percentage of instances classified as class no-stress out of all instances that truly have class no-stress [42] and is calculated using Equation (3).

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3}$$

- Sensitivity, also known as true positive rate, is defined as the percentage of cases classified as class stress among all cases that truly have class stress [44] and is calculated using Equation (4).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{4}$$

- Accuracy is one of the most commonly recognised approaches to performance categorization, and is referred to as the percentage of correctly classified instances to the total number of instances [41, 42], calculated in line with Equation (5).

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{5}$$

- In a classification test, the geometric mean score, also known as g-mean or g-measure, can be applied as an indication of the balance between the majority and minority classes [41]. A low g-mean value is a signifier underpinning a weak performance in the categorisation of stress instances, even when there is the correct categorisation of no-stress instances [4, 26], calculated in line with Equation (6):

$$\text{g} - \text{mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{6}$$

AUC makes reference to the area under ROC curve. This presents an approach to establishing how well a parameter may differentiate between two groups (stress/no-stress) [2, 42]. The AUC value ranges between [0, 1]. A value of 1 is seen to indicate ideal/perfect performance whilst 0.5 showcases random (50/50) performance. The AUC value may be calculated at different points; thus, the final AUC value is not biased by a single threshold [20].

4. Experimental Results

In order to evaluate the performance of the proposed classification model described in Figure 1, two sets of experiments on the stress dataset were conducted as follows:

- First experiment: three different classifiers, namely Random Forest, J48, and JRip have been trained and evaluated with proposed model without the inclusion of an oversampling process.
- Second experiment: the oversampling process is applied. Next, the same three classifiers and the proposed model are used. The aim of this experiment is to examine the effects of an oversampling process in classification model.

The authors applied a standard 10-folds cross-validation method to reduce the overfitting and increase the stability of the classifiers evaluation in all the experiments (all the instances are used for training and testing) [4, 19]. The final result is then estimated by averaging the ten testing results. It should be noted that all the experiments have been conducted in WEKA tool [21].

4.1. First Experiment

In the first sets of experiments, three recognised classifiers (Random Forest, J48, and JRip), are used to investigate the classification performance of the proposed model against stress dataset (without the inclusion of an oversampling process).

The confusion matrix results of this experiment are shown in Table 3. The accuracy, sensitivity, specificity, g-mean, and AUC results are shown in Table 4.

Table 3. Confusion matrix results against stress dataset without oversampling.

Classifiers	**Actual class value**	**Predicted class value**	
		stress	**no-stress**
Random Forest	stress	179	238
	no-stress	13	10410
J48	stress	271	146
	no-stress	93	10330
JRip	stress	212	205
	no-stress	91	10332
Proposed model	stress	309	108
	no-stress	61	10362

Table 4. Results against stress dataset without oversampling.

Classifiers	Accuracy	Sensitivity	Specificity	g-mean	AUC
Random Forest	0.9768	0.429	0.999	0.655	0.977
J48	0.9780	0.650	0.991	0.803	0.873
JRip	0.9727	0.508	0.991	0.710	0.771
Proposed model	0.9844	0.741	0.994	0.858	0.972

After analysing Table 4, it was observed that our proposed model slightly outperformed Random Forest, J48, and JRip classifiers with reference to predictive accuracy measure. In fact, the proposed model outperformed Random Forest, J48, and JRip by 0.76%, 0.64%, and 1.17%, respectively. These results confirm that our proposed model is designed to improve the stability and the predictive accuracy of current machine learning classifiers. Additionally, all classifiers we considered provided excellent accuracy scores for classification; however, those excellent scores are misleading for datasets with imbalanced distributed classes [35]. In our dataset, a classifier that labels all the instances as class no-stress will achieve an accuracy score of 96.15%. The imbalance dataset makes accuracy, not a reliable performance metric to use. For this purpose, it is necessary to evaluate other metrics such as sensitivity and g-mean.

Regarding the sensitivity measure, which represents the stress class recall and, more importantly in our case, we found that our proposed model achieved the best results, while the Random Forest achieved the worst. To be more specific, our proposed algorithm achieved 23.3%, 9.1%, and 31.2% higher sensitivity rate than JRip, J48, and Random Forest classifiers. However, these results produced by all considered classifiers are an indication of a poor performance in classification of minority instances. As shown in Table 3, Random Forest, J48, JRip, and our proposed model misclassified 238, 146, 205, and 108 instances that have class stress, respectively, which indicate a low recall of the stress class. Hence, there is need for oversampling the minority class to enhance the performance of the classification models.

The G-mean measure confirms the superiority of our proposed model. More specifically, our proposed algorithm achieved 14.8%, 5.5%, and 20.3% higher g-mean rate than JRip, J48, and Random Forest classifiers.

In terms of AUC measure, the Random Forest algorithm outperformed the proposed model, J48, and JRip by 0.5%, 10.4%, and 20.6%, respectively. This notice is also confirmed by the specificity rates attained by all classifiers, which indicate a high recall of the no-stress class. In fact, as can be seen from Table 3, Random Forest misclassified only 13 instances that are have class no-stress.

In brief, from the first experiment results it is concluded that all considered algorithms performed well and are applicable for automobile drivers' stress identification.

4.2. Second Experiment

To improve the generalisation performance of the classification models, we used the SMOTE algorithm to oversample the stress dataset's minority class (stress class) to reduce the impact of smaller underlying automobile driver instances on the stress prediction with a lower size on the prediction of stress. Next, the proposed model and the same three classifiers are used. In this experiment, it was considered to investigate the impact of oversampling on the imbalanced stress dataset's classification.

In this study, the oversampling process is only used on the learning parts of the cross-validation process. This is to ensure a fair testing procedure for data that has been altered or misrepresented. Generally, for the imbalanced datasets, oversampling process are seen to improve the classification accuracy [4]. Notably, however, determining how often the minority class should be oversampled in advance is difficult. Hence, a total of nine several oversampling percentages were applied (100-500 in steps of 50) to achieve the best classification model as recommended in the previous research [4, 7]. Figures 2-6 display the evaluation results and illustrate the impact of the oversampling process on the performance of the classifiers.

Figure 2 displays the accuracy derived by the proposed model, Random Forest, J48, and JRip on the oversampled dataset.

Oversampling with SMOTE had no effect on the predictive accuracy of the J48 and JRip classifiers, as shown in Figure 2. At 350% oversampling percentage, the Random Forest classifier improved marginally in terms of classification accuracy with 0.74%. Our proposed model classifier increased classification accuracy by 0.74%, which reached 150% oversampling percentage. Furthermore, our proposed model is the best classifier in identifying stress with a classification accuracy of 98.95%.

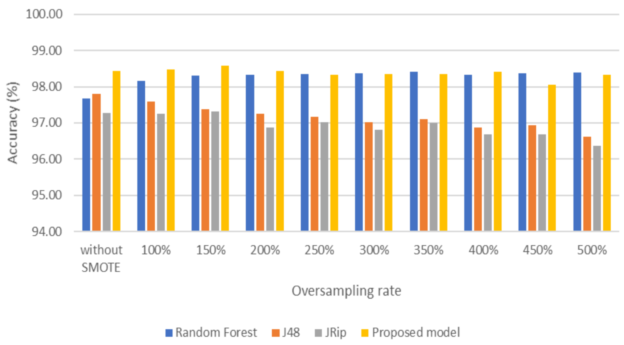

Figure 2. Classification accuracy rates of all considered classifiers using SMOTE technique.

We are interested in the sensitivity rates in Figure 3, which reflects the recall of the minority class. We can see that oversampling gradually improves the sensitivity of all classifiers. For Random Forest, the best oversampling rate is 500% with a sensitivity ratio of 72.4%. Further, Random Forest is the most improved classifier with 14.4%, which was reached at 500% oversampling rate. Also, our proposed model is still the best classifier in identifying stress with a ratio of 79.9% which was reached at 450% oversampling rate.

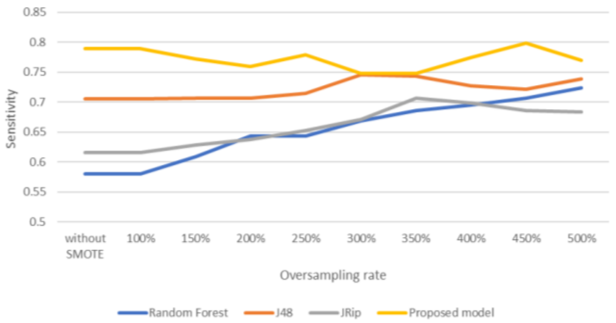

Figure 3. Sensitivity rates of all considered classifiers using SMOTE technique.

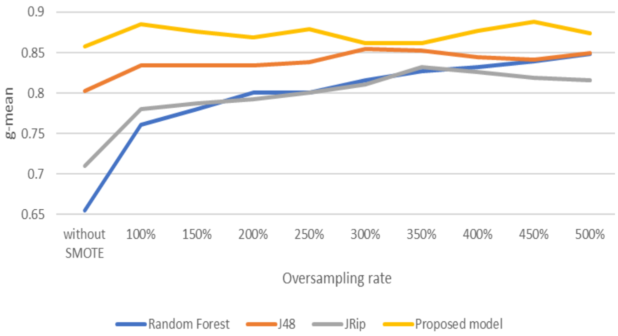

Figure 4. G-mean rates of all considered classifiers using SMOTE technique.

The same behaviour can be seen observed in Figure 4 for the g-mean ratio, with oversampling gradually improving the g-mean rates of all classifiers. Our proposed model is still the best classifier in identifying stress with a ratio of 88.8%, which was reached at 450% oversampling rate. Also, Random Forest is the most improved classifier with 19.3%, which was reached at 500% oversampling rate. These acceptable g-mean rates produced by all considered classifiers are an indication of an acceptable performance in classification of minority samples in the stress dataset.

According to the AUC rates in Figure 5, we can see that the oversampling process improves slightly the AUC rates for J48 and JRip classifiers with 0.1% and 7.2%, respectively. For Random Forest and our proposed model, they are produced stable and consistent results. These findings show that it is still possible for all classifiers to correctly classify both majority and minority classes in the dataset.

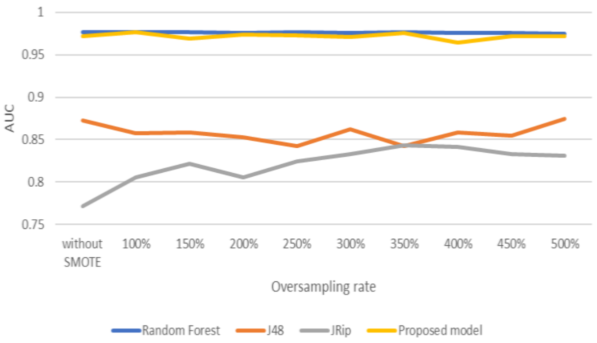

Figure 5. AUC rates of all considered classifiers using SMOTE technique.

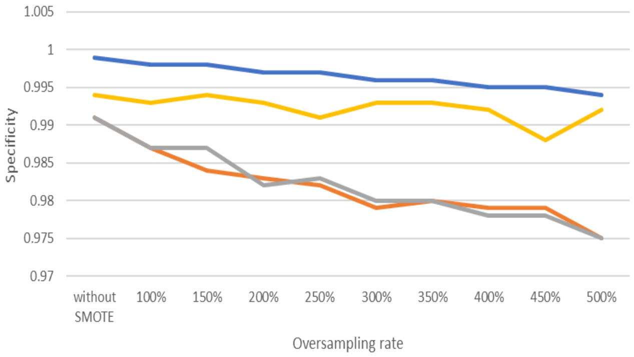

Figure 6. Specificity rates of all considered classifiers using SMOTE technique.

Finally, Figure 6 displays the specificity rates which reflect the recall of the majority class. Although it appears that the specificity rates are slightly decreased for all classifiers, the results of the specificity rates produced are above 97%. These results confirm that the oversampling process positively enhances the performance of classifiers to classify the minority classes without make a significant effect on the predictive to classify the majority classes.

Furthermore, the conclusion can be inferred from all experiments, all considered classifiers produce acceptable performance in classification; the significant and goodness of the features that are collected to identify the stress of automobile drivers. Moreover, all considered classifiers may demonstrate useful and suitable methods to address the stress prediction issue of automobile drivers.

5. Conclusions

The findings from this study can be seen across three phases: the first phase reflects the importance and evaluation of three different well-known algorithms toward prediction of drivers' stress; the next phase focuses on examining the effects of an oversampling process in classification model; and the last phase proposes a two-step ensemble model by integrating Random Forest, J48, and JRip with Multi-Layer Perceptron neural networks for predicting the automobile drivers' stress.

The initial phase of this study highlights the findings from the first experiment, this means that all of the algorithms considered performed well and are applicable for predicting stress in automobile drivers. However, regarding the sensitivity measure, which represents the recall of the stress class, which is most valuable in our case, the results indicate that all considered algorithms produced poor results (<66%). This because the gathered data extremely imbalanced with a ratio of approximately 4%–96%.

The results show that oversampling process improves steadily the sensitivity and g-mean measures of all considered classifiers. In other words, the oversampling process improves the prediction for instances with a minority class. This finding is consistent with previous experimental results in literature [4, 10, 27].

The third phase of this study proposes a two-step ensemble classifier model by integrating prediction results from Random Forest, J48, and JRip classifiers, fed into a Multi-layer Perceptron neural network. The findings show that the proposed model is much more accurate and more scalable than all considered classifiers with regards accuracy, sensitivity, and g-mean measures.

References

[1] Abu-Arqoub M., Hadi W., and Ishtaiwi A., "A New Associative Classification Based on RIPPER Algorithm," *Journal of Information and Knowledge Management*, vol. 20, no. 1, pp. 2150013, 2021.

[2] Aburub F. and Hadi W., "A New Associative Classification Algorithm for Predicting Groundwater Locations," *Journal of Information and Knowledge Management*, vol. 17, no. 4, pp. 1850043, 2018.

[3] AlAgha A., Faris H., Hammo B., and Al-Zoubi A., "Identifying β -Thalassemia Carriers Using A Data Mining Approach: the Case of the Gaza Strip, Palestine," *Artificial Intelligence in Medicine*, vol. 88, pp. 70-83, 2018.

[4] Al-Fayoumi M., Alwidian J., and Abusaif M., "Intelligent Association Classification Technique for Phishing Website Detection," *The International Arab Journal of Information Technology*, vol. 17, no. 4, pp. 488-496, 2020.

[5] Bakker J., Pechenizkiy M., and Sidorova N., "What's Your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data," *in Proceedings of 11th International Conference on Data Mining Workshops*, Vancouver, pp. 573-580, 2011.

[6] Barua S., Begum S., and Ahmed M., "Supervised Machine Learning Algorithms to Diagnose Stress for Vehicle Drivers Based on Physiological Sensor Signals," *Studies in Health Technology and Informatics*, vol. 211, pp. 241-248, 2015.

[7] Bogner C., Kuhnel A., and Huwe B., "Predicting With Limited Data-Increasing the Accuracy in Vis-Nir Diffuse Reflectance Spectroscopy by Smote," *in Proceedings of 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, Lausanne, pp. 1-4, 2014.

[8] Bundele M. and Banerjee R., "Detection of Fatigue of Vehicular Driver using Skin Conductance and Oximetry Pulse," *in Proceedings of the 11th International Conference on Information Integration and Web-based Applications and Services*, New York, pp. 739-744, 2009.

[9] Bunker R. and Thabtah F., "A Machine Learning Framework for Sport Result Prediction," *Applied Computing and Informatics*, vol. 15, no. 1, pp. 27-33, 2019.

[10] Chen B., Xia S., Chen Z., Wang B., and Wang G., "RSMOTE: A Self-Adaptive Robust SMOTE For Imbalanced Problems with Label Noise," *Information Sciences*, vol. 553, pp. 397-428, 2021.

[11] De Santos Sierra A., Sanchez-Avila C., Bailador del Pozo G., and Guerra Casanova J., "Stress Detection By Means of Stress Physiological Template," *in Proceedings of 3rd World Congress on Nature and Biologically Inspired Computing*, Salamanca, pp. 131-136, 2011.

[12] Deng Y., Wu Z., Chu C., and Yang T., "Evaluating Feature Selection for Stress Identification," *in Proceedings of IEEE 13th International Conference on Information Reuse and Integration*, Las Vegas, pp. 584-591, 2012.

[13] El-Khalili N., Alnashashibi M., Hadi W., Banna A., and Issa G., "Data Engineering for Affective Understanding Systems," *Data*, vol. 4, no. 2, pp. 52, 2019.

[14] Eyben F., Wöllmer M., Poitschke T., Schuller B., Blaschke C., Färber B., and Nguyen-Thien N., "Emotion on the Road-Necessity, Acceptance, and Feasibility of Affective Computing in the Car," *Advances in Human-Computer Interaction*, pp. 1-17, 2010.

[15] Fernández-Caballero A., González P., López M., and Navarro E., "Special Issue on Socio-Cognitive and Affective Computing," *Applied Sciences*, vol. 8, no. 8, pp. 1371, 2018.

[16] Ghaderi A., Frounchi J., and Farnam A., "Machine Learning-Based Signal Processing using Physiological Signals for Stress Detection," *in Proceedings of 22nd Iranian Conference on Biomedical Engineering*, Tehran, pp. 93-98, 2015.

[17] Hadi W., "Classification of Arabic Social Media Data," *Advances in Computational Sciences and Technology*, vol. 8, pp. 29-34, 2015.

[18] Hadi W., El-Khalili N., AlNashashibi M., Issa G. and AlBanna A., “Application of Data Mining Algorithms for Improving Stress Prediction of Automobile Drivers: A Case Study in Jordan,” *Computers in Biology and Medicine*, vol. 114, no. 7, pp. 103474, 2019.

[19] Hadi W., Issa G., and Ishtaiwi A., “ACPRISM: Associative Classification Based on PRISM Algorithm,” *Information Sciences*, vol. 417, pp. 287-300, 2017

[20] Hajian-Tilaki K., “Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation,” *Caspian Journal of Internal Medicine*, vol. 4, no. 2, pp. 627-635, 2013.

[21] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and Witten I., “The WEKA Data Mining Software,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, 2009.

[22] Han, J., Kamber M., and Pei J., *Data Mining: Concepts and Techniques*, Elsevier, 2012.

[23] Haouij N., Poggi J., Ghozi R., Sevestre-Ghalila S., and Jaïdane M., “Random Forest-Based Approach for Physiological Functional Variable Selection for Driver’s Stress Level Classification,” *Statistical Methods and Applications*, vol. 28, pp. 157-185, 2019.

[24] Harmon-Jones E., Gable P., and Price T., “The Influence of Affective States Varying In Motivational Intensity on Cognitive Scope,” *Frontiers in Integrative Neuroscience*, vol. 6, pp. 73, 2012.

[25] Healey J. and Picard R., “Detecting Stress During Real-World Driving Tasks Using Physiological Sensors,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156-166, 2005.

[26] Hido S., Kashima H., and Takahashi Y., “Roughly Balanced Bagging for Imbalanced Data,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 2, no. 5-6, pp. 412-426, 2009.

[27] Ishaq A., Sadiq S., Umer M., Ullah S., Mirjalili S., Rupapara V., and Nappi M., “Improving the Prediction of Heart Failure Patients’ Survival Using SMOTE and Effective Data Mining Techniques,” *IEEE Access*, vol. 9, pp. 39707-39716, 2021.

[28] Jiang Z., Pan T., Zhang C., and Yang J., “A New Oversampling Method Based on the Classification Contribution Degree,” *Symmetry (Basel)*, vol. 13, no. 2, pp. 194, 2021.

[29] Munoz-Organero M. and Corcoba-Magana V., “Predicting Upcoming Values of Stress While Driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1802-1811, 2017.

[30] Nagaraj K., Bhattacharjee B., Sridhar A., and GS, S. “Detection of Phishing Websites Using A Novel Twofold Ensemble Model,” *Journal of Systems and Information Technology*, vol. 20, no. 3, pp. 321-357, 2018.

[31] PSD. The statistics of traffic accidents. Retrieved from https://www.psd.gov.jo/images/traffic/traffic2017 Last Visited, 2021.

[32] Qabajeh I., Thabtah F., and Chiclana F., “A Dynamic Rule-Induction Method for Classification in Data Mining,” *Journal of Management Analytics*, vol. 2, no. 3, pp. 233-253, 2015.

[33] Rahman T., Zhang M., Voida S., and Choudhury T., “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*, Brussels, pp. 166-169, 2014.

[34] Rigas G., Goletsis Y., Bougia P., and Fotiadis D., “Towards Driver’s State Recognition on Real Driving Conditions,” *International Journal of Vehicular Technology*, vol. 2011, pp. 1-14, 2011.

[35] Różanowski K., Truszczyński O., Filipczak K., and Madeyski M., “The Level of Driver Personality And Stress Experienced As Factors Influencing Behavior on The Road,” *in Sustainable Development*, vol. 168, pp. 1009-1019, 2015.

[36] Schießl C., “Stress And Strain While Driving,” *in Proceedings of the Young Researchers Seminar-European Conference of Transport Research Institutes*, Brno, pp. 27-30, 2007.

[37] Shearer C., “The CRISP-DM Model: The New Blueprint for Data Mining,” *Journal Data Warehousing*, vol. 5, no. 4, pp. 13-22, 2000.

[38] Shiwu L., Linhong W., Zhifa Y., Bingkui J., Feiyan Q., and Zhongkai Y., “An Active Driver Fatigue Identification Technique Using Multiple Physiological Features,” *in Proceedings of International Conference on Mechatronic Science, Electric Engineering and Computer*, Jilin, pp. 733-737, 2011.

[39] Shu C. and Burn D., “Artificial Neural Network Ensembles and Their Application in Pooled Flood Frequency Analysis,” *Water Resources Research*, vol. 40, no. 9, 2004.

[40] Smart R., Cannon E., Howard A., Frise P., and Mann R., “Can We Design Cars To Prevent Road Rage?,” *International Journal of Vehicle Information and Communication Systems*, vol. 1, no. 1-2, pp. 44-55, 2005.

[41] Thabtah F., Hadi W., Abdelhamid N., and Issa A., “Prediction Phase in Associative Classification Mining,” *International Journal of Software Engineering and Knowledge*

Engineering, vol. 21, no. 6, pp. 855-876, 2011.

[42] Tharwat A., "Classification Assessment Methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168-192, 2021.

[43] Villaverde J., Godoy D., and Amandi A., "Learning Styles' Recognition in E-Learning Environments with Feed-Forward Neural Networks," *Journal of Computer Assisted Learning*, vol. 22, no. 3, pp. 197-206, 2006.

[44] WHO. Global status report on road safety 2018. Retrieved from https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/ Last Visited, 2021.

[45] Zhai J. and Barreto A., "Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables," *in Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1355-1358, New York, 2006.

May Al-Nashashibi received the PhD degree from the University of Bradford, Bradford, UK. She is currently an Assistant Professor in the University of Petra, Amman, Jordan. She started her research focusing on Arabic natural language processing and text mining. Currently her research interests are concentrated on applying data mining techniques in the fields of medicine, biology, and chemistry. She has published research papers in reputed international journals/conferences.

Wael Hadi is currently the Chair of Information Security at the University of Petra. He Holds a Ph.D. degree from the Arab Academy for Banking and Financial Sciences. His research interest in Data Mining, Machine Learning, and Big Data.

Nuha El-Khalili is the Dean of Faculty of Information Technology and the director of the E-learning center at University of Petra. She obtained her PhD from the School of Computing at University of Leeds in the United Kingdom. She has 19 years of experience in teaching Software Engineering courses. Her research interest includes: data engineering for data science, quality assurance for managing academic programs, and e-learning.

Ghassan Issa is a Professor of Computer Science. He received his M.S. and Ph.D. in Computer Science from Old Dominion University, Virginia, in 1987 and 1992 respectively. He was a faculty member and Department Chair of Computer Science at Pennsylvania College of Technology (Penn State), USA from 1992-1995. He also served as the Dean of Computer Science at the Applied Science University (Amman, Jordan) from 2003-2005, and the Dean of Information Technology at the University of Petra (Amman, Jordan) from 2008-2018. Currently he is a Professor and the Dean of the School of Information Technology at Skyline University (Sharjah, UAE). Professor Issa's research interest include Artificial Intelligence, Machine Learning and Deep Learning Fine Tuning, Case-Based and Analogical learning, and Associative Classification.

Abd Alkarim Albanna is the CTO and the co-founder of Jordanian startup called TAKALAM, a company that provides solutions for hearing and speech disorder. He is a member of the Leaders in Innovation Fellowship from the Royal Academy of Engineering (the UK, 2020) and PhD student at Loughborough University England.