Comparison of Several Machine Learning Classifiers for Arousal Classification: A Preliminary study

Ekin Can Erkuş
Biomedical Engineering
Middle East Technical University
Ankara, Turkey
eerkus@metu.edu.tr, ekincanerkus@hotmail.com

Fikret Arı
Department of Electrical and Electronics Engineering
Ankara University
Ankara, Turkey
fari@eng.ankara.edu.tr

Vilda Purutçuoğlu
Department of Statistics
Middle East Technical University
Ankara, Turkey
vpurutcu@metu.edu.tr

Didem Gökçay
21. Yüzyıl Ürünleri Co. Ltd.
Ankara, Turkey
didemgokcay@gmail.com

***Abstract*— Detection of arousal intervals, especially stress detection via a human-machine interface is a trending topic. Stress detection algorithms with high accuracy can be used in many fields such as criminal interrogations or a variety of stress-related experiments. There are many indicators of the stress on the human body, especially on the face area, such as galvanic skin response (GSR), pupil diameter, heart rate (HR), and electromyography (EMG). Hereby, the measurement of such physiological data in stressful, joyful and non-stressful cases can reveal the effects of the stress on the body signals.**

This preliminary study aims to compare several machine learning approaches, namely, linear discriminant analysis (LDA), k-nearest neighbour (k-NN), Naïve Bayes, support vector machines (SVM) and coarse tree algorithms in a classification study. To perform the analyses, the pupil data are collected from a total of 9 subjects while the subject was watching three types of movies, independently. The classifications are performed among the labelled data with multivariate features such as mean, median, maximum to minimum difference and variance, and their univariate versions in order to observe their independent discrimination performances. Moreover, the preprocessed raw data are also used in classification, independently. Here, the movies are selected such that they include either annotated positive, negative or neutral scenes, which may indicate the stressful, joyful and non-stressful intervals, respectively. Therefore, the classification results of these algorithms for the annotated labels in each channel separately are found to observe their effectiveness in detection of arousal intervals. Hence, the main aim is to contribute to the stress detection literature by providing a comparison between both the classification algorithms, features and raw data classification.

Keywords — Stress detection, pleasantness, machine learning, classification, pupil diameter, feature extraction.

I. INTRODUCTION

Stress can be considered as one of the physiological response of the organism to the environmental or internal conditions causing arousal in the body [1]. Herewith the revelation of the conditions to cause change the stress levels can be analysed to predict the stressful intervals of the subjects during the measurements [2]. Besides, the stress detection procedure for humans can be used in many situations such as in forensic interviews from verbal [3] or visual data [4], during vehicle driving to warn the driver for safe travel [5], for the environmental or workspace surveillance [6] or for the diagnosis of disorders [4]. Hence, the detection of the stress intervals may aid to obstruct the unexpected or unwanted actions or to identify the conditions the subject is in [7].

Stress recognition in human is a vast study field as the number of different biomedical data modalities, the diversity of the subject conditions, the variety of the environmental factors and the stress detection algorithms are numerous [7], [8]. Moreover, as the data type changes, the proper tools to process and analyse the data also varies. Hence, choosing the appropriate method for the analysis of the biomedical data can be a major problem for the stress detection studies [9]. Hereby, the comparison of the analysis methods takes place in most of the preliminary studies for the stress detection [10]. Therefore, as being a preliminary study, this study focuses on the comparison of some machine learning classification methods such as linear discriminant analysis (LDA), k-nearest neighbour (k-NN), Naïve Bayes, support vector machines (SVM) and coarse tree algorithms [11].

Apart from brain imaging techniques to gather data from brain activities, stress detection studies use data modalities which are often collected from the face area since most stress-related physiological responses also occur in face [7]. A data type example for such physiological responses to stressful stimuli can be pupil diameter which increases with a stressful stimulus in dilation process [12]. There are large number of stress detection studies which claims to detect stress by solely using the pupil diameter data [13][14]. Another data modality example can be the heart rate variability (HRV) which can be collected from the temple areas of the face to record the rate of the pulses of heart [15]. During stressful conditions, the HRV data amplitudes are expected to rise as a result of sympathetic nervous system activities [16]. There are several facial stress detection studies based on the heart rate measurements [17][18]. The third example for the data types can be galvanic skin response (GSR) which is the conductance of the skin [15]. The GSR data is expected to increase in amplitude during arousal conditions such as stress [19]. As many facial stress studies implement GSR data analyses, it is shown that GSR data show high susceptibility to stress [20]. The fourth example for the data types which are used in stress detection is the electromyography (EMG) which refers to the data imaging of the electrical activities of the muscle cell [21]. As the stress levels increase, EMG data amplitude is also

978-1-7281-8073-1/20/$31.00 ©2020 IEEE

expected to increase [22]. EMG data modality is often used with other data modalities for stress detection as it is affected by a variety of physiological situations [4][8][14]. Aside from the abovementioned data modalities, the stress detection literature includes many other data types, collected from facial area, such as body temperature, blink frequency, eye tracking and video recordings [23][12]. However, as its being a preliminary study, only the pupil diameter data collected during an earlier workshop [24] are used in the analyses of this study.

Machine learning is a trending area in clustering and classification, barring the usage in artificial intelligence fields [25]. As machine learning classifiers use the different properties of the features of the data, their performances depend mostly on some statistical assets of the data such as the distributions, trends and periodicity of the samples [18]. Moreover, classification methods are often useful in developing algorithms for diagnosis and detection purposes in biomedical data studies [26]. Along with the generally used single-classifier detection purposes, there are several studies which compare several machine learning classifiers for their correctly labelling the data into the subject conditions [27][28][29]. Hereby, to test the detectability of the stress from facial regions, several classification methods, namely, linear discriminant analysis (LDA), k-nearest neighbour (k-NN), Naïve Bayes, support vector machines (SVM) and coarse tree algorithms are used in this study. The selection of abovementioned methods mostly depends on their accessibility from a single toolbox (Statistics and Machine learning Toolbox [30] for classifiers) to obtain a set of standardized results which provide comparability. Moreover, 4 independent features, namely, mean, median, maximum-to-minimum difference and variance are extracted from the data, as they represent the basic statistical properties of the data behaviour. The multivariate feature extracted classification scenarios are performed by using those features.

The rest of the paper is composed of several sections: In the second section, namely, Background and Methods, the brief information about the methods such as data and classification methods which are used in this study can be found. Moreover, their parameters and usages in this study are also described in the second section. In the third section, the results of the experimentation processes are reported and the results are discussed. Finally, in the fourth section, conclusions about the experimental results in comparison with the literature, as well as the possible future works can be found.

II. Background and Methods

A. Data Collection

Pupil diameter is a measure of the pupil's size, which can be measured by image processing on the video recording of an eye [31]. Obviously, unless the disease or other factors that cause malfunctions in the pupil functions, the pupil diameter increases in dim light to capture more light and decreases during bright light [32]. Hence, the background light is needed to be adjusted during the experiments. Besides the effect of light, the pupil can dilate by the sympathetic nervous system and can constrict due to the parasympathetic system. Therefore, during the arousal intervals, the pupil diameter is expected to increase [12].

The pupil video is recorded using a camera, the analog to digital conversion and sampling is performed with a sampling rate of 30Hz in this study. The eye area of the subject is illuminated by a LED light with 940nm wavelength to provide better contrast between pupil and iris. Then, the pupil is recorded through a camera where the signals are transferred to the Raspberry Pi SBC board. Here, the image processing steps take place, where dark pupil shape is detected from the frames of images. The system details of this custom system and software details can be found in [33],[34], [35].

B. Experimental Setup

The experimental design consists of data collection during video watching sessions. The subjects sit in front of a monitor in order to watch video clips. There are 3 types of video clips, namely, positive, negative and neutral arousing. The positive video clip is taken from the film: Dumb and Dumber I, final scene (Approximately 2 minutes and 04 seconds), where the overall clip is funny and the subject is expected to have joy and amusement to cause a positive arousal during some scenes while watching. Secondly, the negative video clip is taken from the film: The Taking of Pelham 123, first hostage shooting scene (Approximately 7 minutes and 41 seconds), in which the subject is expected to feel aroused by the negative scenes in this clip. Finally, the neutral clip, which is expected to cause no emotional response in subject, is consist of packaging instructions (Approximately 2 minutes and 58 seconds). Several segments of the video clips are annotated by the researchers and the most commonly annotated scenes are selected. The starting points of these scenes are taken as the beginning of impulses used to depict impulse responses in the data (i.e., stimuli, and the physiological response and recovery time of pupil dilation, which is around 3 seconds [36]). Hereby, 3 scenes are annotated from the positive and negative clips in the form of stimulus effects, whereas the neutral clip is taken as a single neutral scene. Moreover, the initial 20 seconds are cropped from the beginning of both positive and negative movies to form baseline neutral scenes for the positive and negative, respectively (Such baseline scenes are labelled as neutral for the rest of the study).

A total of 9 subjects participated in this experiment. The subjects filled emotion surveys PANAS, BDI and STAI and their current mood, depression level and anxiety are recorded. Negative mood and high levels of depression and anxiety were used as exclusion criteria. All subjects passed the threshold for exclusion, so no subject was excluded. The demographic information of the subjects can be found in Table I.

TABLE I. SUBJECT DEMOGRAPHICS

	Mean	Standard deviation
Age	27.11	5.34
BDI	5.88	4.62
PANAS(+)	46.22	3.95
PANAS(-)	28	5.37
STAI(S)	36.55	9.64
STAI(T)	40.33	8.97

C. Data Analyses

1) Preprocessing

The outlying data samples are extracted in several steps. First, the meaningful bands for the data measures of overall measures are observed by roughly investigating the data means and behaviour of the data. Hence, the pupil measurements within the values of 2 to 7 mm are cropped and the rest of the samples are discarded and replaced with the data mean for that subject not to cause time shift in the

measurements. Moreover, such correction also eliminates unexpected data behaviour such as blinks, where the data are padded as -1 by the recording device.

The second step is to crop the annotated scenes from the data, separately. Due to the different length of the video recordings, the scenes are temporally shifted in terms of data samples. Hence, the new correct locations of the annotated scenes in the data are found by multiplication of the time points in seconds by the total number of samples and dividing it to the total clip length in seconds. Hereby, the scenes are cropped and marked.

The third step is to eliminate the DC shifts of the cropped data by subtracting the first sample value from the samples of the cropped scene. This process is to prevent the effects of environmental changes such as light intensity and to reduce temporal effects such as trending of the data due to the measurement system or physiology.

The fourth step is also an optional step and used to standardize the data samples using the z-score. Apart from the fourth step, where the level shift is used to standardize the annotated data segments, this step is to eliminate the mean differences between the main data groups [37] and adjust them with respect to their standard errors.

Filtering in frequency domain is not operated as the sampling rate of the data modalities are relatively low (≤50Hz), which prevents both the significant noise effect of the hum noise around 50Hz and higher frequency noises (≥50Hz) [38].

2) Feature Extraction and Classification Matrix

The features, namely, mean, median, max-to-min value difference (i.e. range) and variance are extracted from the annotated data scenes. As the number of subjects is 9, number of classes is 3 and the number of scenes are 3, each feature is calculated 81 times. Hence, the classification matrix in size of 81x4 is computed by arranging the respective values of the features. Note that, the first 27 rows represent the features from the positive cases, next 27 are for the negative and the final 27 rows stand for the neutral cases. The classification models for feature extracted cases are computed by using this classification matrix.

D. Machine Learning Classification, Validation Algorithms and Analysis of Variance

In order to maintain the stability and comparability of the results, machine learning classifiers are required to be operated by using a single software toolbox, or custom scripts with same normalization processes. Hence, considering the ease of use and accessibility, all of the features are extracted and classified in MATLAB software, and the functions for classification algorithms are used from the Statistics and Machine Learning Toolbox [30] of MATLAB, R2020a, academic licence. Moreover, the algorithms are further selected based on their speed and the performances on preliminary empirical results with some toy data.

1) Linear Discrimination Analysis Classification Algorithm

In the training phase of linear discriminant analysis (LDA), under the assumption of multivariate Gaussian distributed samples of classes, the mean and covariance of the classes of samples are calculated. According to such statistical properties, a linear discriminate function is generated such that the probabilities of the samples on assigning to the correct classes are maximized [39]. Hence, by using the trained function, the new samples can be labelled. Hereby, in this study, the parameters (delta and gamma) which are required to use the related function are automatically set by the function itself.

2) k-Nearest Neighbour Classification Algorithm

k-nearest neighbour is a non-parametric classification algorithm where the test samples are assigned to the most repetitive class of the samples among *k* samples [40].

The *k* value is selected by following an empirical approach where the classification is performed for the set of k values which ranges from 1 to 8 as can be found in Table II. This pre-study is repeated 10 times in a Monte-Carlo simulation and the results are averaged.

TABLE II. CLASSIFICATION ACCURACIES FOR DIFFERENT NUMBER OF THE MAXIMUM NUMBER OF SPLITS PARAMETER FOR K-NN CLASSIFIER IN AN EMPIRICAL PRE-STUDY.

k value	1	2	3	4	5	6	7	8
100*Accuracy	50.6	63.0	44.4	54.3	62.7	58.0	60.5	60.5

As can be observed in Table II, the k value which yields the highest accuracy is k = 2. Hence, the k value is taken as 2 for the rest of the study. Moreover, the distance metric is taken as Mahalanobis although for 1 dimensional case, both Euclidean and Mahalanobis distance results are the same [41].

3) Kernel Naïve Bayes Classification Algorithm

By assuming that the features are independent, the samples are assigned to classes using the Bayes' theorem, maximizing the posterior probabilities of assigning samples to the correct classes [42]. In this study, Gaussian distribution is used as the kernel smoothing function. Moreover, the parameters required to run the function are automatically selected by the optimization tool as built-in in the function.

4) Support Vector Machines (SVM) Algorithm

Support vector machines (SVM) classification algorithm labels the sample points by drawing a visual hyperplane between the possible groups of samples such that the overall distances of the samples to that hyperplane is maximized [43]. Hence, it is an optimization model in which the samples fall in the correct classes. In this study, the parameters are optimized by the function itself and the computations are performed.

5) Coarse Tree Classification Algorithm

Tree classification is a decision algorithm where starting from the root of the tree, each sample is assigned to a branch depending on its features [40]. Hence, the branches are composed of samples with similar features. After all of the samples are assigned to a branch, the algorithm stops and the branches form the classes.

Among several tree classification algorithms, the coarse tree is selected as there are 3 labels to classify for this study. The maximum number of split parameter is empirically selected as a process of computing the classification rates by changing the parameter value from 1 to 8 by repeating the process 10 times in a Monte-Carlo simulation. The accuracies can be found in Table III.

TABLE III. CLASSIFICATION ACCURACIES FOR DIFFERENT NUMBER OF THE MAXIMUM NUMBER OF SPLITS PARAMETER FOR TREE CLASSIFIER IN AN EMPIRICAL PRE-STUDY.

Max no of Splits:	1	2	3	4	5	6	7	8
100*Accuracy	45.7	57.8	58.0	55.6	54.3	54.3	51.9	50.6

Hence, according to Table III, the parameter is selected as 3, since it yields the lowest error rate in classification.

6) Leave-one Out Cross Validation Algorithm

The process of parameter modelling is trained by the N-1 number of values in the classification matrix, it is tested for the remaining data and the performance for that iteration is recorded. This process is repeated N times, by ensuring that all the data are used as test samples once. Then, the overall performance is evaluated as the average of the iteration performances [39].

7) Welch-Analysis of Variances (wANOVA)

Apart from the analysis of variances (ANOVA), Welch ANOVA handles the inhomogeneity of the variances in the groups [44]. Since the preliminary analyses show that the pupil data recordings of the groups have violation of the homogeneity of variances, wANOVA is used as it is a better alternative for such data. For the analyses, both the built-in MATLAB functions and SPSS software are used to perform wANOVA.

III. RESULTS AND DISCUSSION

The 3 classes of pupil diameter data are pre-processed independently from each other, some features are extracted from the data, and then classification and validation algorithms are applied on both the features and the raw data. Moreover, the Welch analysis of variances (wANOVA) is performed to investigate the pairwise relationship between the labels and to further validate the classification accuracies. The results for each experimental case are reported in this section.

A. Classification Results

Classifications for three labels, namely, positive, negative and neutral intervals are performed for the pupil diameter data in order to test the differentiability among the labelled annotations. The classification results are reported as overall accuracies of correctly labelling the samples in 10-fold cross validation step, in percentage. Moreover, as there are 3 independent labels, the successful classification rate (100*Accuracy) of 33.33% can be considered as the minimum viable value. Hence, Table IV represents the successful classification rates for univariate features, independently for the machine learning classifiers. Furthermore, the classification rate of those features are computed in multivariate case by using the classification matrix as described in Section 2, in order to compare the classifiers for the multivariate case. Moreover, the preprocessed raw data are also classified by the same classifiers to compare the effects of features instead directly classifying the raw data. However, for the validation of the raw data classification, 10-fold cross validation is used. The successful classification rate is calculated by multiplication of the 1-error rate by 100.

TABLE IV. CLASSIFICATION SUCCESS RATES (100*ACCURACY) OF CLASSIFIERS TO CLASSIFY 3 GROUPS: POSITIVE, NEGATIVE AND NEUTRAL.

Feature/Raw data		Classifier				
		LDA	k-NN	Naive Bayes	SVM	Coarse Tree
Multivariate Features		54.3	66.7	59.3	65.4	67.9
Univariate Features	Mean	48.2	46.9	35.8	39.5	37.3
	Median	44.4	61.7	38.3	37.0	43.2
	Max-min difference (range)	67.9	64.2	63.0	63.0	67.9
	Variance	55.6	49.4	53.1	54.3	45.7
Preprocessed Raw data		37.2	43.2	37.4	48.8	34.5

According to the results in Table IV, firstly, considering the overall results, classification using features provides better results than the raw data classification. Moreover, the multivariate classification gives better results except for the range feature. Secondly, comparing the univariate features, the range feature has the best discriminability for all the classifiers. Also, mean and median features seem to fail to discriminate the data into 3 groups successfully for the most of the cases. Finally, as the machine learning classifiers are compared, the results should be investigated separately for multivariate and univariate feature cases. Hereby, for the multivariate features, coarse tree, k-NN and SVM classifiers performed relatively better results than LDA and Naïve Bayes classifiers. On the other hand, k-NN and LDA results are better than the rest for the most of the features.

Although the analyses to achieve the classification accuracies and the number of groups to classify are different from each other, the studies in literature where only the pupil dilation data are used to classify into the groups represents a range of accuracies. Most of the studies in literature perform classification among 2 groups which provides them a baseline accuracy of 50% for equal numbers of samples or features to be classified in each group. However, classification into 3 groups reduces this baseline accuracy to 33.33%, provided that the number of samples or features are equal among the groups [39]. The studies in the literature for classification of stressful or arousal conditions using solely pupil dilation data show that the accuracy is generally around 60-85% for classification of two groups [14]. Herein, a similar study [12] show that the accuracy of distinguishing between the two groups is found to be 64.9% for decision tree and 72.7% for random forest classifiers. Our accuracy result for coarse tree algorithm is better than this rate with 67.9%, even for 3 group classification. Another study [45] investigates pupil dilation in an eye-tracking problem and reports the accuracy of classification between two groups as 72% using SVM classifier. Our SVM classification result is 65.4% for 3 groups, instead of 2 can be comparable with this result.

Apart from solely classifying the pupil diameter data, using multiple modalities such as galvanic skin response, skin temperature, blood pressure and heart rate variability along with pupil diameter yields expectedly better results, from 67% [46] up to 90.1% [13], [47] for SVM classifier. However, our preliminary study is only interested in the comparison of the classifiers for pupil diameter data by leaving the usage of multivariate data modalities and pursuing the best accuracy as future works.

B. wANOVA Results for Pupil Diameter

Apart from the classification analyses, posterior to the pooling the data samples of subjects, the pairwise relationships between the classes, namely, positive negative and neutral are investigated using the one-way wANOVA test for α=0.05. Hence, all pairs of groups among all experiments, show significant differences ($p<0.05$) for the wANOVA results. This result may show that, although the classification rates are relatively low compared to the literature, the data show statistical difference among groups.

IV. Conclusion

Among many machine learning classifiers, this preliminary study focuses on the classification performances of linear discriminant analysis (LDA), k-nearest neighbour (k-NN), Naïve Bayes, support vector machines (SVM) and coarse tree algorithms.

The results show that the relatively best overall classifier among the classifiers we test is k-NN. On the other hand, for the multivariate classification, coarse tree classifier has the best classification rate. Thereby, for the future studies, k-NN and coarse tree classifiers may be considered for use as the main classifier for the variety of purposes. On the other side, the range feature surprisingly has the best results among the features, and the performances of mean and median features are not well enough. Moreover, variance feature results average performance compared to other features. Furthermore, wANOVA results show that all pairs of groups have significantly different variances. Hereby, instead of classical statistical analyses to seek the mean differences, the differences between variances can mainly be used for the future studies.

The future work contains the several expansions of this preliminary study, with data collected from remarkably more number of subjects. The higher number of subjects may yield the significantly meaningful results and also provides the vision in the generation of the real-time stress detection algorithm as the first expansion. The second possible expansion of this study may include the comparison of many other classifiers in more details. Such a study may reveal the procedures for the optimal classifier selection depending on the data type and the variety of the analyses. On the other hand, the pre-processing part of this study could be improved by including the frequency domain analyses and new outlier detection algorithms.

Finally, in this preliminary study, it is assumed that the replication of the videos under the three conditions are equivalent apart from the brightness of the screen of the videos. Therefore, by substituting the first measured sample from the whole annotated part of the data, it is assumed and accepted that all the videos have the same normalized brightness. Although for practical purposes such types of assumptions are used [48], [49], they are statistically very strict. Because, the sources of noises can be different and some of these noises can be systematic and, thereby, can be eliminated via modelling in the normalization. Hence, similar to the certain biological signals [50] the amount of noises from different sources can be considered as a parameter to be estimated. In the extension of this study, we consider to construct a regression model whose predictors are the sources of variation and response in the measured signal. Then, we will estimate the predictors to discard the batch effects in replications so that they can be comparable.

Acknowledgment

The authors thank to the TUBITAK project (no: 117E650) for their support. The authors are indebted to the team that collected this data during the eNTERFACE workshop [24]: Fatih Ileri, Merve Balik, and Anil Berk Delikaya, as well as overseeing members Bilgin Avenoglu, Atil Ilerialkan, and Huseyin Haciahabiboglu.

References

[1] S. Folkman, "Stress: Appraisal and Coping," in *Encyclopedia of Behavioral Medicine*, 2013.

[2] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, 2005, doi: 10.1109/TITS.2005.848368.

[3] M. Hartwig and C. F. Bond, "Lie detection from multiple cues: A meta-analysis," *Appl. Cogn. Psychol.*, 2014, doi: 10.1002/acp.3052.

[4] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Trans. Affect. Comput.*, 2019, doi: 10.1109/TAFFC.2019.2927337.

[5] P. Napoletano and S. Rossi, "Combining heart and breathing rate for car driver stress recognition," in *IEEE International Conference on Consumer Electronics - Berlin, ICCE-Berlin*, 2018, doi: 10.1109/ICCE-Berlin.2018.8576164.

[6] C. Viegas, S. H. Lau, R. Maxion, and A. Hauptmann, "Towards independent stress detection: A dependent model using facial action units," in *Proceedings - International Workshop on Content-Based Multimedia Indexing*, 2018, doi: 10.1109/CBMI.2018.8516497.

[7] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *Journal of Biomedical Informatics*. 2016, doi: 10.1016/j.jbi.2015.11.007.

[8] A. Goyal, S. Singh, D. Vir, and D. Pershad, "Automation of Stress Recognition Using Subjective or Objective Measures," *Psychological Studies*. 2016, doi: 10.1007/s12646-016-0379-1.

[9] R. Kocielnik, N. Sidorova, F. M. Maggi, M. Ouwerkerk, and J. H. D. M. Westerink, "Smart technologies for long-term stress monitoring at work," in *Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems*, 2013, doi: 10.1109/CBMS.2013.6627764.

[10] P. Ren *et al.*, "Comparison of the Use of Blink Rate and Blink Rate Variability for Mental State Recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2019, doi: 10.1109/TNSRE.2019.2906371.

[11] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, doi: 10.1145/1390156.1390169.

[12] S. Baltaci and D. Gokcay, "Stress Detection in Human–Computer Interaction: Fusion of Pupil Dilation and Facial Temperature Features," *Int. J. Hum. Comput. Interact.*, 2016, doi: 10.1080/10447318.2016.1220069.

[13] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *Annual International Conference of the IEEE*

Engineering in Medicine and Biology - Proceedings, 2006, doi: 10.1109/IEMBS.2006.259421.

[14] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," *Comput. Methods Programs Biomed.*, 2012, doi: 10.1016/j.cmpb.2012.07.003.

[15] J. L. Semmlow, *Biosignal and Medical Image Processing*. 2008.

[16] C. Ottaviani, D. Shapiro, D. M. Davydov, and I. B. Goldstein, "Autonomic stress response modes and ambulatory heart rate level and variability," *J. Psychophysiol.*, 2008, doi: 10.1027/0269-8803.22.1.28.

[17] M. Chauhan, S. V. Vora, and D. Dabhi, "Effective stress detection using physiological parameters," in *Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, ICIIECS 2017*, 2017, doi: 10.1109/ICIIECS.2017.8275853.

[18] A. De Santos Sierra, C. Sánchez Ávila, J. Guerra Casanova, and G. Bailador Del Pozo, "A stress-detection system based on physiological signals and fuzzy logic," *IEEE Trans. Ind. Electron.*, 2011, doi: 10.1109/TIE.2010.2103538.

[19] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in *Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems*, 2013, doi: 10.1109/CBMS.2013.6627790.

[20] B. Alić, D. Sejdinović, L. Gurbeta, and A. Badnjevic, "Classification of stress recognition using Artificial Neural Network," in *2016 5th Mediterranean Conference on Embedded Computing, MECO 2016 - Including ECyPS 2016, BIOENG.MED 2016, MECO: Student Challenge 2016*, 2016, doi: 10.1109/MECO.2016.7525765.

[21] J. Webster, "Medical instrumentation: application and design, Fourth edition.," *John Wiley Sons, Inc. USA*, 2010.

[22] U. Lundberg *et al.*, "Psychophysiological stress and emg activity of the trapezius muscle," *Int. J. Behav. Med.*, 1994, doi: 10.1207/s15327558ijbm0104_5.

[23] F. Mokhayeri, M. R. Akbarzadeh-T, and S. Toosizadeh, "Mental stress detection using physiological signals based on soft computing techniques," in *2011 18th Iranian Conference of Biomedical Engineering, ICBME 2011*, 2011, doi: 10.1109/ICBME.2011.6168563.

[24] D. Gökcay *et al.*, "Preliminary Results in Evaluating the Pleasantness of an Interviewing Candidate Based on Psychophysiological Signals," in *eNTERFACE 19, Summer Workshop on Multimodal Interfaces*, 2019, pp. 45–49.

[25] D. Ellen, S. Day, and C. Davies, *Statistical and Machine-Learning Data Mining*. 2011.

[26] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device - in laboratory and real life," in *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, doi: 10.1145/2968219.2968306.

[27] D. Novak, M. Mihelj, and M. Munih, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interact. Comput.*, 2012, doi: 10.1016/j.intcom.2012.04.003.

[28] N. Y. Hammerla and T. Plötz, "Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition," in *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, doi: 10.1145/2750858.2807551.

[29] V. Bajaj and R. B. Pachori, "Detection of human emotions using features based on the multiwavelet transform of EEG signals," *Intell. Syst. Ref. Libr.*, 2015, doi: 10.1007/978-3-319-10978-7_8.

[30] MathWorks, "Statistics and Machine Learning Toolbox Release Notes," *MatLab*, 2015.

[31] N. Wade and B. Tatler, *The Moving Tablet of the Eye: The Origins of Modern Eye Movement Research*. 2010.

[32] M. A. Bouffard, "The Pupil," *CONTINUUM Lifelong Learning in Neurology*. 2019, doi: 10.1212/CON.0000000000000771.

[33] Y. Durna and F. Ari, "Design of a binocular pupil and gaze point detection system utilizing high definition images," *Appl. Sci.*, 2017, doi: 10.3390/app7050498.

[34] J. De Witt, "Open Source Eye Tracker Project." https://github.com/xef6/eyetracker.

[35] Jeremy A. Russell, "Open Source Eye Tracker Project." https://github.com/JeremyARussell/Gaze.

[36] S. Baltacı, "Fusion of pupil dilation and facial temperature features for detection of stress," Middle East Technica University, 2016.

[37] B. G. Tabachnick and L. S. Fidell, *Using multivariate statistics (6th ed.)*. 2012.

[38] B. Widrow *et al.*, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, 1975.

[39] B. D. Ripley, *Pattern recognition and neural networks*. 2014.

[40] A. R. Webb, *Statistical Pattern Recognition: Second Edition*. 2003.

[41] R. A. Johnson and D. W. Wichern, "Applied multivariate statistical analysis. Prentice Hall, Englewood Cliffs, NJ.," *Appl. Multivar. Stat. Anal. Prentice-Hall, Englewood Cliffs, NJ.*, 1992.

[42] *Handbook of Data Intensive Computing*. 2011.

[43] V. Kecman, "Support Vector Machines – An Introduction," 2005.

[44] A. J. Tomarken and R. C. Serlin, "Comparison of anova Alternatives Under Variance Heterogeneity and Specific Noncentrality Structures," *Psychol. Bull.*, 1986, doi: 10.1037/0033-2909.99.1.90.

[45] J. Jadue, G. Slanzi, L. Salas, and J. D. Velásquez, "Web user click intention prediction by using pupil dilation analysis," in *Proceedings - 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015*, 2016, doi: 10.1109/WI-IAT.2015.221.

[46] N. Sharma and T. Gedeon, "Hybrid genetic algorithms for stress recognition in reading," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, doi: 10.1007/978-3-642-37189-9_11.

[47] A. Barreto, J. Zhai, and M. Adjouadi, "Non-intrusive physiological monitoring for automated stress detection in human-computer interaction," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2007, doi: 10.1007/978-3-540-75773-3_4.

[48] M. E. Kret and E. E. Sjak-Shie, "Preprocessing pupil size data:

Guidelines and code," *Behav. Res. Methods*, 2019, doi: 10.3758/s13428-018-1075-y.

[49] W. W. Tryon, "Pupillometry: A Survey of Sources of Variation," *Psychophysiology*, 1975, doi: 10.1111/j.1469-8986.1975.tb03068.x.

[50] E. Wit and J. McClure, *Statistics for Microarrays*. 2004.