Pamukkale Univ Muh Bilim Derg, 31(1), 79-85, 2025

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Performance evaluation of the machine learning algorithms for emotion classification on the CASE dataset

CASE veri seti üzerinde duygu sınıflandırma için makine öğrenmesi algoritmalarının performans değerlendirmesi

Emre Rifat YILDIZ[1*], *Yıltan BİTİRİM*[1]

[1]Department of Computer Engineering, Faculty of Engineering, Eastern Mediterranean University, Famagusta, North Cyprus, via Mersin 10, Türkiye.

emre.yildiz@emu.edu.tr, yiltan.bitirim@emu.edu.tr

Received/Geliş Tarihi: 13.12.2023 Revision/Düzeltme Tarihi: 27.03.2024 doi: 10.5505/pajes.2024.59321

Accepted/Kabul Tarihi: 02.04.2024 Research Article/Araştırma Makalesi

Abstract

Emotion classification using physiological signals is still a challenging task even the sensor technology and machine learning algorithms evolved within the decades. In this study, the performance of KNN, DT, RF, LR, and XGB algorithms on emotion classification was evaluated in terms of accuracy on the CASE dataset. Three sub-datasets namely Downsampled, Resampled-EM, and Resampled-VA were obtained from the original dataset. Then, hyperparameter tuning was applied to the smallest dataset and the algorithms were applied with the parameters that were obtained in hyperparameter tuning to the Resampled-EM, Resampled-VA, and original sets. As the results obtained, KNN, RF, and XGB provided higher accuracies on the Resampled-VA set when compared to the Resampled-EM set, where it was the contrary for the DT algorithm. XGB algorithm provided the highest accuracy of 97.44% among all the results. This study could be considered as the most comprehensive study that utilizes machine learning algorithms for emotion classification on the CASE dataset.

Keywords: Emotion recognition, Machine learning, Physiological signals, CASE dataset.

Öz

Sensör teknolojisi ve makine öğrenimi algoritmaları birkaç on yıl içinde evrimleşmiş olsa da fizyolojik sinyalleri kullanarak duygu sınıflandırması hala zorlu bir görevdir. Bu çalışmada, KNN, DT, RF, LR ve XGB algoritmalarının CASE veri seti üzerinde duygu sınıflandırmasındaki performansları değerlendirildi. Orijinal veri setinden Downsampled, Resampled-EM ve Resampled-VA olarak isimlendirilen 3 alt-veri seti elde edildi. Daha sonra, en küçük boyuta sahip veri setine hiperparametre ayarlaması uygulandı ve algoritmalar hiperparametre ayarlamasında elde edilen parametrelerle Resampled-EM, Resampled-VA ve orijinal setlere uygulandı. Elde edilen sonuçlara göre, KNN, RF ve XGB algoritmaları Resampled-VA setinde DT algoritmasına kıyasla daha yüksek doğruluklar sağladı. Bu durum Resampled-EM seti için tam tersi olarak gözlemlendi. XGB algoritması, %97.44 ile tüm sonuçlar arasında en yüksek doğruluğu sağladı. Bu çalışma, CASE veri setinde duygu sınıflandırması için makine öğrenimi algoritmalarını en kapsamlı şekilde kullanan çalışma olarak değerlendirilebilir.

Anahtar kelimeler: Duygu tanıma, Makine öğrenmesi, Fizyolojik sinyaller, CASE veri seti.

1 Introduction

Emotions have a much greater place in many areas of our lives than expected and are an integral part of human life. Our emotions change voluntarily or involuntarily in daily life, business life, or in response to the events we encounter. Emotions closely concern our interaction with the world around us [1], such as the decisions we make [2], our social relationships [3],[4], our productivity in business life [5], and our health [6],[7].

Emotion recognition and analysis have been an area where many studies have been conducted for several decades. It is known that emotional changes in a person trigger many behavioral and/or physiological changes [8]. In addition to causing physical changes in facial expressions or gestures, emotions also cause physiological changes in heart rate, blood flow, skin conductance, body temperature, and brain waves, and these changes can be observed through physiological sensors [9]. The use of sensors that emerged with the development of technology in emotion analysis has led computer science as well as psychology to work in this field [10]. The integration of physiological sensors with machine learning (ML) and/or deep learning (DL) algorithms [11],[12], emotion analysis is being carried forward day by day to examine human emotions in depth, going beyond the self-reporting techniques used in psychology.

In the literature, there are many datasets containing physiological sensor data recorded during emotional changes of individuals, and many studies that predict emotions by applying ML or DL algorithms to these datasets. In the study of Cui et al. [13], Electroencephalography (EEG) signals from DEAP and DREAMER datasets were used to recognize valence and arousal levels by applying the proposed Regional-Asymmetric Convolutional Neural Network algorithm. They obtained 95% accuracy on both valence and arousal classification tasks. Hassan et al. [14] employed Fine Gaussian Support Vector Machine and Deep Belief Network on the fused observations of physiological signals. Photoplethysmogram (PPG), Electrodermal Activity (EDA), and Zygomaticus Electromyography (zEMG) sensor data from the DEAP dataset were used to classify five different emotions and achieved 89.53% overall accuracy. In the study of Hasnul et. al [15], the proposed augmentation technique was applied to the Electrocardiogram (ECG) signals from DREAMER, AMIGOS, and A2ES (created for the corresponding study) datasets to classify emotions. Selected five ML algorithms that are K-Nearest

*Corresponding author/Yazışılan Yazar

Neighbor (KNN), Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and Multilayer Perceptron (MP) were employed with hyperparameter tuning on these three datasets. The classifiers were compared before and after the augmentation process and it was found that the KNN classifier provided the best results with classification accuracies above 90%. Hssayeni&Ghoraani [16] proposed a Convolutional Neural Network with multimodal data fusion methods and applied them to predict positive/negative affects and three emotions using the physiological data in the WESAD dataset. Also, the effect of physiological signal combinations on performance was explored and they obtained 78% F1-score and 79% accuracy for emotion classification. In the study of Bota et al. [17], seven ML algorithms were applied to the physiological data in publicly available five datasets that are ITMDER, WESAD, DEAP, MAHNOB-HCI, and EESD. The emotions were evaluated regarding low/high valence/arousal classification and the classifiers were compared in terms of Feature Fusion and Decision Fusion. For each dataset and for each sensor modality a different classifier provided the highest success with the accuracies varying from 40.74% to 96.68%.

To the best of our knowledge, two studies in the literature employed emotion classification on the CASE dataset. These studies were as follows. In Zhang et al.'s [18] study, they utilized deep learning, machine learning algorithms, and a proposed algorithm to classify valence and arousal levels using the CASE dataset. The classification task was divided into three parts: binary classification (high/low valence and arousal levels), 3-class classification (low/neutral/high valence and arousal levels), and 4-class classification (four quadrants of valence-arousal space). The proposed algorithm achieved recognition accuracies of 76.37% and 74.03% for valence and arousal, respectively, on the CASE dataset. Yıldız & Bitirim [19] conducted a study using the CASE dataset to assess the success of the KNN classifier in classifying valence-arousal levels and emotions based on physiological data. The dataset was initially Downsampled to create a balanced dataset, and the classification task was divided into three parts: EMG data, non-EMG data, and whole data classifications. The highest accuracy and F1-score were about 94% at k-value 1 for emotion classification in the whole data part.

In this study, the physiological data from the publicly available dataset CASE [20] were used and selected five ML algorithms were employed to explore the algorithm that provides higher accuracy. A preprocessing was applied to the dataset and two labels such that valence-arousal level and corresponding emotion were assigned to each data row. The emotion labels were obtained according to the circumplex model of emotions introduced by Russel [21]. Then, three sub-datasets namely Downsampled, Resampled-EM, and Resampled-VA were extracted from the original dataset in order to have more balanced emotion classes. Afterward, five ML algorithms that are KNN, DT, RF, Extreme Gradient Boosting (XGB), and Logistic Regression (LR) were applied to the Downsampled set with hyperparameter tuning. The parameters that provided the highest accuracy were also applied to the Resampled-EM, Resampled-VA sets, and the original dataset for each of the algorithms with 5-fold cross-validation.

The Downsampled set and preprocessed original dataset were taken from the study of Yıldız & Bitirim [19]. Additionally, the Resampled-EM and Resampled-VA sets were also created for our study. To the best of our knowledge, this study is the most comprehensive study that utilizes machine learning algorithms for emotion classification using the CASE dataset.

This paper is organized as follows: The next section describes the details of the dataset, the preprocessing steps and hyperparameter tuning details; the third section contains the classification results and discussion; and finally, the last section gives the concluding remarks and future work.

2 Methodology

2.1 Dataset

For the emotion classification using physiological signals, the CASE dataset was used. This dataset generally contains physiological data and valence-arousal values obtained from 30 participants (15 male, 15 female). Each participant was equipped with a set of sensors and watched a set of categorized videos according to its content such as “Amusing” or “Relaxing”. While watching the videos, participants were asked to select valence-arousal levels using a joystick to indicate their emotions. The physiological data of the participants were acquired using EEG, ECG, BVP, EMG, GSR, Skin Temperature, and Respiration sensors. All the sensors and the joystick were acquiring data at 1000 Hz and 20 Hz, respectively. The physiological data and joystick annotations were saved in separate CSV files for each participant. In total, there were 30 files for physiological data and 30 files for annotations where the dataset contains 73,547,520 data rows and 81 valence-arousal levels.

A preprocessing was applied to the dataset to combine physiological data and the annotations. The details of the preprocessing were explained in the study of Yıldız & Bitirim [19]. When the dataset is represented as valence-arousal and emotion classes, it was observed that the dataset is imbalanced and it was decided to create a balanced sub-dataset from the original dataset. Figure 1(a) and Figure 1(b) illustrate the data distribution of the original dataset based on emotion labels and valence-arousal labels, respectively.

The Downsampled set was obtained and used as it was in the study of Yıldız & Bitirim [19]. The Downsampled set contains a total of 2.878.200 rows of data where each emotion class contains 221,400 rows of data and Figure 2(a) illustrates the data distribution of this set. In addition to this, two more sub-datasets were created to have more data with more balanced classes (compared to the original set) by aiming to have higher classification scores. These subsets were obtained as;

i. The average number of data in terms of emotion-labeled data was calculated for each participant and random data were taken from each class as the calculated average value and named as Resampled-EM,

ii. The average number of data in terms of valence-arousal labeled data was calculated for each participant and random data were taken from each class as the calculated average value and named as Resampled-VA.

For both of these sets, if a class contains data less than the average, all the data in the class were taken. The data selection was done for each participant separately. Hence, the Resampled-EM set consists of 36,064,126 rows of data and the Resampled-VA set consists of 27,904,992 rows of data. The number of samples in class for Resampled-EM, Resampled-VA, and original sets were as given in Table 1. Figure 2(b) and

Figure 2(c) illustrate the data distributions for Resampled-EM and Resampled-VA sets, respectively. The abbreviations "EM" and "VA" in Resampled-EM and Resampled-VA represent emotion and valence-arousal since the Resampled-EM set was created based on emotion labels and Resampled-VA set was created based on valence-arousal labels.

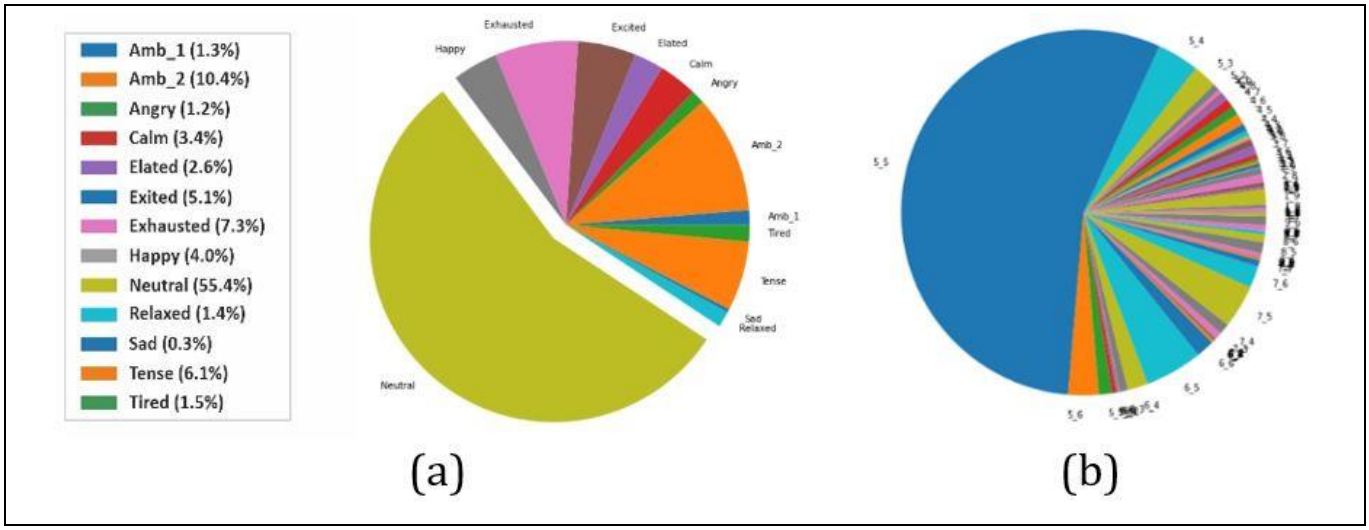

Figure 1. Data distribution of the original dataset. (a): Distribution based on emotion labels. (b): Distribution based on valence-arousal labels [19].

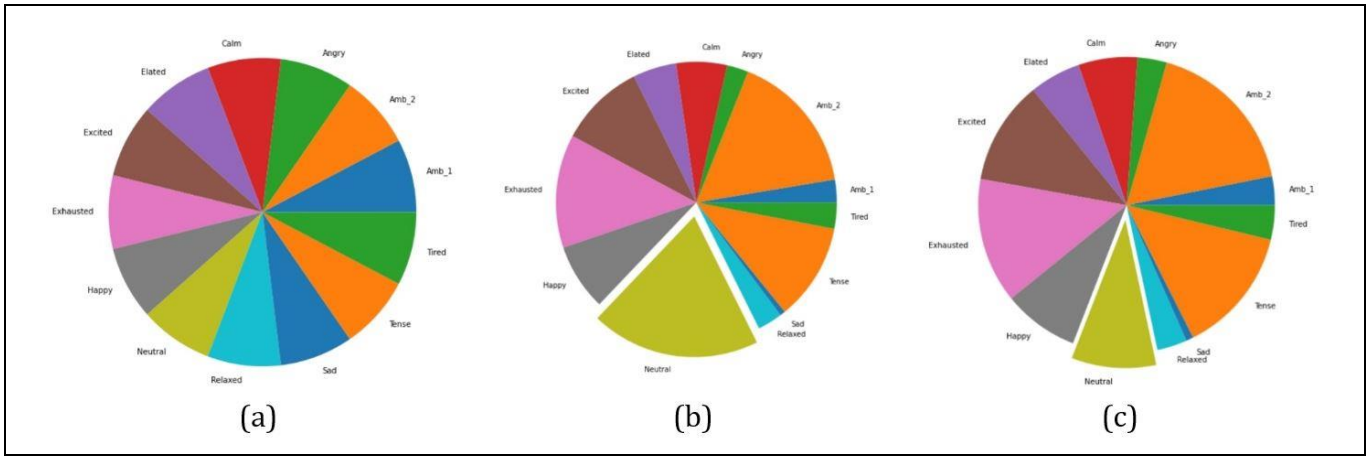

Figure 2. Data distribution of the sub-datasets. (a): Distribution of the Downsampled set. (b): Distribution of the Resampled-EM set. (c): Distribution of the Resampled-VA set.

Table 1. Number of samples for each subset and class.

	Original	Resampled-EM	Resampled-VA
Amb_1	947.600	947.600	875.511
Amb_2	7.628.300	5.905.868	4.874.788
Angry	888.750	888.750	886.761
Calm	2.481.850	2.112.339	1.797.234
Elated	1.917.800	1.805.333	1.556.682
Excited	3.746.884	3.547.036	3.157.668
Exhausted	5.394.800	4.702.477	3.814.546
Happy	2.940.500	2.764.796	2.304.913
Neutral	40.769.786	7.029.413	2.573.305
Relaxed	1.035.850	1.035.850	912.544
Sad	221.400	221.400	221.400
Tense	4.492.450	4.021.714	3.855.318
Tired	1.081.550	1.081.550	1.047.282
TOTAL	73.547.520	36.064.126	27.877.952

2.2 Classifications

For the classification, several ML algorithms were applied to be able to find the algorithm that provides the best classification scores. The classification was performed in two phases such that the first phase was applying hyperparameter tuning on the Downsampled set and the second phase was the classification on Resampled-EM, Resampled-VA, and the original set with the parameters that provided the highest classification score in the hyperparameter tuning phase.

The algorithms KNN, DT, RF, XGB, and LR were applied using Python and ran on the TRUBA infrastructure [22]. The used partition on TRUBA was "long" and the configuration of the remote computer was as follows; Intel Xeon E5/Core i7 CPU, number of nodes was 1, number of cores was 40, memory per core was 8000MB, installed Python version was 3.9.0 and installed sklearn version was 1.2.0.

In the second phase, the LR was not used since promising results could not be obtained with this algorithm. The data in the resampled sets and the original set was divided into 80% training and 20% test sets. The number of samples from each set was as follows: 28,851,296 samples for training and 7,212,830 samples for test from Resampled-EM set, 22,329,402 samples for training and 5,575,590 samples for test from Resampled-VA set, 58,838,020 samples for training and 14,709,500 samples for test from the original set, Five-fold cross-validation was applied to the training set for each of the sets.

2.2.1 Hyperparameter tuning

For the hyperparameter tuning, open-source Python libraries were utilized for the classifiers. Each library provided a collection of parameters and associated values for the classifier to be set. However, some of the parameters were selected since trying all the possibilities for each of the parameters increases the computational cost and requires a huge amount of resources. Since the size of the dataset was also a cost for the computations, the Downsampled set was used for this phase. To apply hyperparameter tuning, the GridSearchCV class was used from sklearn.model_selection library [23]. This class implements the given classifier along with the given grid of hyperparameters. It iterates over all the combinations and provides a set of results as a dictionary. Table 2 shows the number of fits and results as well as the total fit time (train), total score time (test), and total time (train&test) obtained from GridSearchCV.cv_results_ [23] for each algorithm. Total fit time and total score time were calculated using "mean_fit_time" and "mean_score_time" values provided by GridSearchCV.cv_results_, respectively.

Table 2. Computation time of the algorithms.

	No of fits	No of results	Total Fit Time (hr)	Total Score Time (hr)	Total Time (hr)
KNN	240	48	0.88	7.95	8.83
DT	675	135	2.58	0.26	2.84
RF	240	48	21.64	1.81	23.46
XGB	200	40	320.71	1.72	322.43
LR	400	80	77.87	0.07	77.94
TOTAL	1775	351	423.68	11.81	435.5

The Downsampled set was also divided into 80% (2,302,560 samples) for the training and 20% (575,640 samples) for test parts from 2,878,200 total samples with 6 different features. Each algorithm used the same data for each sub-dataset with 5-fold cross-validation, and accuracy was used to evaluate the success of classification. The 20% test data was used with the parameters that provided the best results during training. The details of the hyperparameter tuning steps for each algorithm are explained in the below subsections.

2.2.1.1 K-Nearest neighbor

For the KNN algorithm, sklearn.neighbors library [24] was used. The hyperparameter tuning was applied using three of the parameters that the library provides. These parameters were "n_neighbors", "weights" and "metric". The values for the parameters were as;

(i) n-neighbors: 1, 3, 5, 7, 9, 11, 13, 15,

(ii) weights: uniform, distance,

(iii) metric: minkowski, euclidean, manhattan.

Applying these values to the parameters produces 48 possible results and 240 fits for the training. The highest accuracy score was obtained as 91.97% with the parameter values "n_neighbors=3", "weights=distance" and "metric=manhattan".

2.2.1.2 Decision trees

For the DT algorithm, sklearn.tree.DecisionTreeClassifier library [24] was used. The library provides 12 parameters total and 4 of them were selected for the hyperparameter tuning of the DT classifier. These parameters were "criterion", "max_depth", "max_features" and "min_samples_split". The values for the parameters were as;

(i) criterion: gini, entropy, log_loss,

(ii) max_depth: 10, 50, 100, 200, 300,

(iii) max_features: auto, sqrt, log2,

(iv) min_samples_split: 2, 8, 10.

With the above parameter set, there were 135 possible results and a total of 675 fits for the training. According to the results that were obtained using the DT algorithm, the highest accuracy score was 91.37%. And this result was obtained with the parameter values "criterion=log_loss", "max_depth=50", "max_features=sqrt" and "min_samples_split=2".

2.2.1.3 Random forest

For the RF algorithm, sklearn.ensemle library [24] was used. This library contains 18 different parameters that can be adjusted before fitting the training set. For the hyperparameter tuning of the RF classifier, four of the parameters that are "criterion", "max_depth", "max_features" and "n_estimators" were selected. The applied values for the selected parameters were as;

(i) criterion: gini, entropy,

(ii) max_depth: 100, 150, 200, 300,

(iii) max_features: log2, sqrt, none,

(iv) n_estimators: 10, 30.

Applying these values to the parameters produces 48 possible results and 240 fits for the training. The highest accuracy score was obtained as 95.65% where the parameter values were "criterion=entropy", "max_features=None", "n_estimators=30" and "max_depth=100".

2.2.1.4 Extreme gradient boosting

For the XGB algorithm, an open-source library xgboost [25] was used. For the hyperparameter tuning of the XGB classifier, three parameters such that "n_estimators", "max_depth" and "learning_rate" were selected. The applied values for the selected parameters were as;

(i) n_estimators: 10, 30,

(ii) max_depth 3, 6, 10, 50, 100,

(iii) learning_rate: 0.1, 0.2, 0.3, 0.5.

Applying these values to the parameters produces 40 possible results and 200 fits for the training. As the results show, the highest accuracy score was 95.98% with the parameter values "n_estimators=30", "max_depth=50" and "learning_rate=0.3".

2.2.1.5 Logistic regression

For the LR algorithm, sklearn.linear_model.LogisticRegression library [24] was used. Parameters "C", "penalty" and "solver" were selected for the hyperparameter tuning of the LR classifier. The values for the parameters were as;

(i) C: 0.1, 1, 10, 100,

(ii) penalty L1, L2, elasticnet, none,

(iii) solver: lbfgs, newton-cg, liblinear, sag, saga.

Applying these values to the parameters produces 80 possible results and 400 fits for the training. However, the accuracy scores were very low for the classification. The highest accuracy score was 18.76% with the parameter values "C=0.1", "penalty=L2", and "solver=liblinear".

3 Results and discussion

The classification successes were measured using accuracy metric and the results were given in Table 3. The table contains both the train and test results. "Train" in Table 3 indicates the arithmetic mean of 5-fold for each algorithm.

According to the results shown in Table 3, in the training part of the KNN algorithm 95.53% accuracy for the Resampled-EM set, 95.58% accuracy for the Resampled-VA set, and 96.49% accuracy for the original set was obtained. In the test part, the KNN algorithm provided 95.74% accuracy for the Resampled-EM set, 95.80% accuracy for the Resampled-VA set, and 96.61% accuracy for the original set. The accuracies of the Resampled-VA set were slightly higher than the accuracies of the Resampled-EM set and the accuracies of the original set were higher than the accuracies of the Resampled-EM and Resampled-VA sets for both training and test parts. The KNN algorithm provided the best accuracy as 96.61% for the original set in the test part.

In the training part of the DT algorithm, 94.87% accuracy for the Resampled-EM set, 94.60% accuracy for the Resampled-VA set, and 95.64% accuracy for the original set were obtained. In the test part, the DT algorithm provided 94.99% accuracy for the Resampled-EM set, 94.86% accuracy for the Resampled-VA set, and 95.67% accuracy for the original set. Contrary to KNN results, the accuracies of the Resampled-EM set in both training and test parts were higher than the accuracies of the Resampled-VA set. However, the accuracies of the original set for the DT algorithm were higher than the accuracies of the Resampled-EM and Resampled-VA sets' accuracies in both training and test parts. The highest accuracy using the DT algorithm was obtained as 95.67% in the test part of the original set.

The RF algorithm provided 96.74% accuracy for the Resampled-EM set, 96.92% accuracy for the Resampled-VA set, and 97.23% accuracy for the original set in the training part. In the test part, 96.83% accuracy for the Resampled-EM set, 97.02% for the Resampled-VA set, and 97.30% accuracy for the original set were obtained. Similar to the KNN algorithm, the accuracies of the Resampled-VA set were higher than the accuracies of the Resampled-EM set for both training and test parts. The accuracies of the original set for the RF algorithm were higher than the accuracies of the Resampled- EM and Resampled-VA sets' accuracies in both training and test parts. RF algorithm provided the best accuracy as 97.30% for the original set in the test part.

In the training part of the XGB algorithm, 96.96% accuracy for the Resampled-EM set, 97.13% accuracy for the Resampled-VA set, and 97.39% accuracy for the original set was obtained. XGB algorithm provided 97.05% accuracy for the Resampled-EM set, 97.22% accuracy for the Resampled-VA set, and 97.44% accuracy for the original set in the test part. Similar to the KNN and RF algorithms, the accuracies of the Resampled- VA were higher than the accuracies of the Resampled-EM set in both training and test parts of the XGB algorithm. The accuracies of the original set were higher than the accuracies of the Resampled-EM and Resampled-VA sets in both training and test parts of the XGB algorithm. The highest accuracy score for the XGB algorithm was obtained as 97.44% in the test part of the original set.

The accuracies of the Resampled-VA set were superior to the accuracies of the Resampled-EM set for KNN, RF, and XGB algorithms in both training and test parts. However, this was the opposite for the DT algorithm. All the algorithms provided the highest accuracies for the original set in training and test parts. The lowest accuracies were obtained using the DT algorithm for each set. XGB algorithm provided the best accuracies compared to KNN, DT, and RF algorithms in all parts. Furthermore, the highest accuracy for the training part was 97.39% and the highest accuracy for the test part was 97.44%, which were obtained using the XGB algorithm.

Table 3. Classification results with the selected parameters.

		KNN	DT	RF	XGB
Train	DS	0.9197	0.9137	0.9565	0.9598
	EM	0.9553	0.9487	0.9674	0.9696
	VA	0.9558	0.9460	0.9692	0.9713
	ORG	0.9649	0.9564	0.9723	0.9739
Test	DS	0.9251	0.9140	0.9584	0.9616
	EM	0.9574	0.9499	0.9683	0.9705
	VA	0.9580	0.9486	0.9702	0.9722
	ORG	**0.9661**	**0.9567**	**0.9730**	**0.9744**

DS. Downsampled Set. **EM**: Resampled-EM set. **VA**: Resampled-VA set. **ORG**: Original dataset.

The following observations could be mentioned as well when the results in Table 3 are considered in general. It is noticeable that the Downsampled set has the lowest accuracies in all classifiers while it is contrary for the original set. While the difference in accuracies between these two sets reaches 4.52% (KNN), the accuracy achieved with the original set is 2.85% higher than the DS set on average. One of the main reasons for this difference in success can be said to be the number of data contained in the sets. Although the Downsampled set was balanced and the original set was unbalanced sets, the higher number of samples contained in the original set may have contributed to the success of the classifiers. When Resampled-EM and Resampled-VA sets are compared in this context, the highest difference in accuracies between them was 0.27%, while a mean accuracy difference of 0.15% was observed in all classifiers in both train and test. Furthermore, the accuracies of the Resampled-EM and Resampled-VA sets were higher than the Downsampled set and less than the original set. This was also the case with the number of samples in these sets. We could say that as the number of the sample increases the accuracy of the algorithms increases. When the performances of different algorithms on the same datasets are considered, it can be said that KNN, DT, RF, and XGB algorithms produced similar results and the results of LR were quite low compared to the others. The reason behind that could be the stochastic nature of the classifiers tends to produce different accuracies. To explore the aforementioned observations, a further study that examines the correlation between data distribution and ML algorithms in detail was planned to conduct.

4 Conclusion

In this study, five ML algorithms that are KNN, DT, RF, XGB, and LR were employed on the physiological data from the publicly available dataset CASE. Three sub-datasets were first created from the original dataset and the data were labeled with emotions derived from valence-arousal levels. The algorithms first applied on the Downsampled set with hyperparameter tuning to be able to find the best parameter settings. Then the algorithms were applied on Resampled-EM, Resampled-VA, and the original set with the parameters that were provided the highest scores in hyperparameter tuning. All the sets were divided into 80% training and 20% test sets, 5-fold cross-validation was applied to training data and the success was evaluated with accuracy.

According to the results, the accuracies of the Resampled-VA set for KNN, RF, and XGB algorithms were higher than the accuracies of the Resampled-EM set, for both training and test parts. However, it was the contrary for the DT algorithm. The accuracies for the original set in training and test parts were always higher than the other sets for all of the algorithms. XGB algorithm provided the highest scores for all the sets and both parts. The highest accuracy was obtained on the test part of the original set as 97.44% accuracy.

As the future work of this study, the correlation between data distributions and ML algorithm performances will be investigated, and publicly available datasets that contain similar physiological data will be included for emotion classification using the aforementioned algorithms as well as other popular ML algorithms.

5 Acknowledgement

This study has been funded by the Eastern Mediterranean University Scientific Research Budget with the Project Number BAPC-02-22-03.

The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

6 Author contribution statements

In this study, Emre Rifat YILDIZ performed the implementation and literature review; Yıltan BİTİRİM conceived the presented idea and supervised the work, Both the authors discussed the results and contributed to the final manuscript.

7 Ethics committee approval and conflict of interest statement

"There is no need to obtain permission from the ethics committee for the article prepared".

"There is no conflict of interest with any person/institution in the article prepared".

8 References

[1] Del Giudice M. "*The Motivational Architecture of Emotions*". Editors: Al-Shawaf L, Shackelford TK. The Oxford Handbook of Evolution and the Emotions, 1-39, Oxford, UK, Oxford University Press, 2021.

[2] Alsharif AH, Salleh NZM, Baharun R. "The neural correlates of emotion in decision-making". *International Journal of Academic Research in Business and Social Sciences*, 11(7), 64-77, 2021.

[3] Van Kleef GA, Côté S. "The social effects of emotions". *Annual review of psychology*, 73(1), 629-658, 2022.

[4] Tammilehto J, Kuppens P, Bosmans G, Flykt M, Peltonen K, Vänskä M, Lindblom J. "Attachment orientation and dynamics of negative and positive emotions in daily life". *Journal of Research in Personality*, 105, 1-12, 2023.

[5] Diener E, Thapa S, Tay L. "Positive emotions at work". *Annual Review of Organizational Psychology and Organizational Behavior*, 7(1), 451-477, 2020.

[6] Mazzocco K, Masiero M, Carriero MC, Pravettoni G. "The role of emotions in cancer patients' decision-making". *Ecancermedicalscience*, 13(1), 914-936 2019.

[7] Keller A, Litzelman K, Wisk LE, Maddox T, Cheng ER, Creswell PD, Witt WP. "Does the perception that stress affects health matter? The association with health and mortality". *Health Psychology*, 31(5), 677–684, 2012.

[8] Saxena A, Khanna A, Gupta D. "Emotion recognition and detection methods: a comprehensive survey". *Journal of Artificial Intelligence and Systems*, 2(1), 53–79, 2020.

[9] Gouizi K, Bereksi RF, Maaoui C. Emotion recognition from physiological signals". *Journal of Medical Engineering & Technology*, 35(6-7), 300-307, 2011.

[10] Pace-Schott EF, Amole MC, Aue T, Balconi M, Bylsma LM, Critchley H, Heath AD, Friedman BH, Gooding AEK, Gosseries O, Jovanovic T, Kirby LAJ, Kozlowska K, Laureys S, Lowe L, Magee K, Marin MF, Merner AR, Robinson JL, Smith RC, Spangler DP, Overveld MV, VanElzakker MB. "Physiological feelings". *Neuroscience & Biobehavioral Reviews*, 103(1), 267-304, 2019.

[11] Zhang J, Yin Z, Chen P, Nichele S. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review". *Information Fusion*, 59(1), 103-126, 2020.

[12] Rim B, Sung NJ, Min S, Hong M. "Deep learning in physiological signal data: A survey". *Sensors*, 20(4), 969-1008, 2020.

[13] Cui H, Liu A, Zhang X, Chen X, Wang K, Chen X. "EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network". *Knowledge-Based Systems*, 205, 106243-106252, 2020.

[14] Hassan MM, Alam MGR, Uddin MZ, Huda S, Almogren A, Fortino G. "Human emotion recognition using deep belief network architecture". *Information Fusion*, 51(1), 10-18, 2019.

[15] Hasnul MA, AbdulAziz NA, Abdulaziz A. "Augmenting ECG data with multiple filters for a better emotion recognition system". *Arabian Journal for Science and Engineering*, 48(1), 1-22, 2023.

[16] Hssayeni MD, Ghoraani B. "Multi-modal physiological data fusion for affect estimation using deep learning". *IEEE Access*, 9(1), 21642-21652, 2021.

[17] Bota P, Wang C, Fred A, Silva H. "Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet?". *Sensors*, 20(17), 4723-4740, 2020.

[18] Zhang T, El-Ali A, Wang C, Hanjalic A, Cesar P. "Cornet: Fine-grained emotion recognition for video watching using wearable physiological sensors". *Sensors*, 21(1), 52-77, 2020.

[19] Yıldız, E. R. & Bitirim, Y. "Performance Evaluation of KNN for Emotion and Valence-Arousal Classifications on CASE Dataset". *12th International İstanbul Scientific Research Congress on Life, Engineering, and Applied Sciences,* İstanbul, Türkiye, 21-23 January 2023.

[20] Sharma K, Castellini C, Van Den Broek EL, Albu-Schaeffer, A, Schwenker F. "A dataset of continuous affect annotations and physiological signals for emotion analysis". *Scientific Data*, 6(1), 196-209, 2019.

[21] Russell JA. "A circumplex model of affect". *Journal of Personality and Social Psychology*, 39(6), 1161-1179, 1980.

[22] Tübitak-Ulakbim. "Turkish National e-Science e-Infrastructure-TRUBA". https://www.truba.gov.tr/index.php/en/main-page/ (08.02.2024)

[23] Scikit-Learn. "sklearn.model_selection.GridSearchCV". https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (08.02.2024).

[24] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion, B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapean D, Brucher M, Perrot M, Duchesnay E. "Scikit-learn: Machine learning in Python". *The Journal of Machine Learning Research*, 12(1), 2825-2830, 2011.

[25] DMLC XGBoost. "Python API Reference". https://xgboost.readthedocs.io/en/stable/python/python_api.html (17.06.2023).