ARL-TR-9597 ● Oct 2022

DEVCOM ARMY RESEARCH LABORATORY

An Evaluation of Tabular Neural Network Approaches for Human Affective State Classification from Physiological Signals

by David Chhan and Vernon J Lawhern

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9597 ● OCT 2022

An Evaluation of Tabular Neural Network Approaches for Human Affective State Classification from Physiological Signals

David Chhan and Vernon J Lawhern
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
October 2022	Technical Report	October 1, 2021–September 1, 2022

4. TITLE AND SUBTITLE
An Evaluation of Tabular Neural Network Approaches for Human Affective State Classification from Physiological Signals

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
David Chhan and Vernon J Lawhern

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLH-FE
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9597

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID IDs: David Chhan, 0000-0003-2470-8663

14. ABSTRACT

Despite advances in machine learning approaches in application domains such as computer vision, natural language processing, and speech recognition, to name a few, there is some uncertainty regarding the viability of neural network approaches applied to tabular data—the type of data stored in a (row, column) table format. The standard approach for building machine learning classification methods on tabular data is in the form of decision trees (DTs). However, a recent study comparing different neural network architectures to DT-based methods found that neural network approaches, in many cases, outperformed DT-based methods when evaluated on 40 different tabular data sets, suggesting there are now viable neural network approaches for tabular data. In this report, we described our initial results evaluating tabular neural network approaches for human affective state classification. We used AutoGluon-Tabular, an open-source machine learning framework built for tabular data to develop models to predict high/low arousal or positive/negative valence using features extracted from electrocardiograms and galvanic skin responses obtained from three publicly available data sets. Our classification results were only marginally above random chance, suggesting that subject-independent cross-data set human affective state classification with peripheral physiological signals remains a significant challenge.

15. SUBJECT TERMS
Humans in Complex Systems, machine learning, tabular neural network, affective computing, cross–data set classification, wearable sensing, physiological signal processing, feature extraction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			David Chhan
Unclassified	Unclassified	Unclassified	UU	23	19b. TELEPHONE NUMBER (Include area code) (410) 278-5985

Contents

List of Figures iv

List of Tables iv

1. Introduction 1

2. Methods 3

2.1 Affective Computing 3

2.2 Approach 3

2.3 The Data 4

2.4 Physiological Signal Feature Extraction 5

2.5 ECG Features 5

2.6 GSR Features 6

2.7 AutoGluon-Tabular 7

3. Results 7

4. Discussion and Conclusions 9

5. References 11

List of Symbols, Abbreviations, and Acronyms 16

Distribution List 17

List of Figures

Fig. 1 Arousal-valence emotion distribution chart 4

List of Tables

Table 1 Summarized description of the data sets..... 5
Table 2 ECG-based features 6
Table 3 GSR-based features..... 6
Table 4 Within data set classification accuracy from different ML models by AutoGluon-Tabular..... 8
Table 5 Cross–data set classification accuracy from different ML models by AutoGluon-Tabular..... 8

1. Introduction

Machine learning approaches based on deep neural networks have made significant strides and perform at the state of the art across many challenging application domains including computer vision, natural language processing, speech recognition, and reinforcement learning, to name a few.[1,2] These results are often obtained using large labeled data sets trained with very deep neural networks, which learn highly nonlinear abstractions of raw data features in a hierarchical way.[2] In addition, these methods often incorporate inductive biases by way of neural architecture design to constrain the set of possible solutions. For example, convolutional neural networks (CNNs)[3,4] make extensive use of convolutions with small receptive fields to mimic, to some degree, the neural structure of the primate visual system. Indeed, it has been shown that representations learned by CNNs compare favorably to representations learned from the primate visual system.[5] As a result of these successes, neural network approaches are often treated as the de facto method to apply when building models in those domain areas.

Even with these advances across multiple application domains, there is some uncertainty regarding the viability of neural network approaches applied to tabular data. Tabular data consists of data stored in a (row, column) table format, where rows contain separate instances and columns contain distinct features. In addition, each column/feature in the table may have different possible data types (for example, binary vs. continuous vs. categorical), representing a highly heterogenous data format. A recent study from McKinsey & Company,[6] surveying more than 400 application domains across 19 different industries, showed that tabular data of this type is among the most common data format used in industry. The standard approach for building machine learning classification models on tabular data is generally in the form of decision trees (DTs),[7] which are a family of supervised machine learning models that build a tree-like graph with nodes representing the place where we pick a subset of features and propose a decision rule/threshold based on those features (e.g., if gender = male and age > 40 years, for a tabular data set containing gender and age as features); edges representing the result of this decision rule; and the leaves representing the output, which can be either another decision node with another set of input features or the predicted class label of interest.

DT-based approaches have many benefits, including they are highly interpretable in their basic form (e.g., by tracking the hierarchical flow of decision nodes), which is an important concern in many real-world applications, and they are computationally fast to train. However, DT methods also have several downsides: 1) they are prone to overfitting, causing poor generalization, 2) they can be very

sensitive to small perturbations in the input data (potentially learning a very different tree when trained on different subsets of the data), and 3) they have difficulty modeling very complex, highly nonlinear decision rules. These downsides open up the possibility of using neural network approaches due to their ability to model highly nonlinear relationships and their improved robustness to minor deviations in the input data, as demonstrated in several other application domains. However, because previously proposed neural network architectures are not well suited for tabular data, the lack of appropriate inductive bias often causes them to fail to find optimal solutions for tabular data.

There has been increased interest in building neural network approaches for tabular data over the past several years.[8,9] In particular, a recent study comparing several different neural network architectures with DT-based methods found that the neural network approaches were competitive with DT-based methods when evaluated on 40 different tabular data sets with varying amounts of instances (690–418,000) and features (5–2000).[9] In many cases, these neural network approaches outperformed DT-based methods, suggesting that there are now viable neural network approaches that can be applied to tabular data.

In this report, we describe our initial results evaluating tabular neural network approaches for human affective state classification (e.g., stress, arousal) using wearable physiological sensor technologies such as electrocardiogram (ECG) and galvanic skin response (GSR). Affective state classification is a growing area of interest in the human–computer interaction (HCI) community, as the ability to model and predict human affective state opens new research directions focusing on improving how humans interact and team with autonomous, intelligent systems.[10,11]

One of the biggest challenges in affective state classification is robust performance across individuals, where there is not much prior work in this area.[12] In this domain, however, it is difficult to collect large labeled data sets due to the need for human subjects testing, as well as the significant degree of variability in the underlying affective state across individuals. We hypothesize that, in the low labeled data regime, that tabular neural network approaches trained on pre-extracted features could be a viable alternative to fully end-to-end training (i.e., without any a priori feature processing) with deep neural networks that traditionally require large labeled data sets to train effectively. To test this hypothesis, we use the model framework AutoGluon-Tabular,[13] which is an easy-to-use and highly accurate Python library for building neural networks for tabular data. This report summarizes our findings using three publicly available data sets: Cognitive Load, Affect and Stress Cognitive Load, Affect and Stress (CLAS),[14] ASCERTAIN,[15] and AMIGOS.[16] Our initial analysis focuses on inter-subject classification within

and across these three data sets, as this remains one of the key challenges in affective state classification.

2. Methods

2.1 Affective Computing

The seminal work by affective computing pioneer Rosalind Picard[17] over two decades ago has opened up an entirely new field of study that aims to recognize and understand human emotion. Since then, many studies have focused on developing methods and algorithms capable of recognizing one's emotional state from various input signals, such as neural signals with the use of the electroencephalogram (EEG),[18] to peripheral physiological signals such as ECG,[19,20] respiration,[21] plethysmography (PPG), electrodermal activity (EDA),[22] to behavioral signals such as facial expression,[23] eye movement,[24,25] and others (e.g., speech[26,27]).

Numerous approaches have been employed and developed in emotional recognition work from manipulating the input signals by means of features extraction both in time (linear and nonlinear) and frequency domains, to the use of learning algorithms. Broadly summarized, there are two categories of methodological frameworks: the application of classical machine learning methods on pre-extracted features[14,16,28] and the use of neural network-based algorithms for end-to-end learning on the raw signals.[29–32] While the majority of prior work focused on intra-subject within-data set classification, there are only a few works that have attempted to do subject-independent cross-data set prediction.[12,33–35] The closest prior work to ours is that of Siddharth et al.,[12] who trained a neural network in an end-to-end fashion from raw signals such as EEG, GSR, and others, for cross-data set user affective state classification.

Our work is different from this prior work in that it aims to evaluate classification performance of a tabular neural network approach for subject-independent cross-data set emotion recognition using extracted features from two input signals: ECG and GSR. The reason we focus on these two signals is the potential of their application in that many wearable devices widely available on the market have sensors that continuously measure and collect both signals. In addition, all three data sets used as part of this analysis contain both GSR and ECG for all participants.

2.2 Approach

Our approach is to develop a pipeline to evaluate the applicability of neural network approaches applied to tabular data. For this, we use the open-source package

AutoGluon-Tabular,[13] which can be used to predict affective state (high/low arousal or positive/negative valence) using features extracted from peripheral physiological signals. We used three publicly available data sets that have common recordings of physiological signals (ECG and GSR) with similar stimuli to elicit emotional responses from their respective participants. For each stimulus, a "ground-truth" emotional state was assigned. Generally, emotions can then be mapped to a 2-D valence-arousal emotional space, as shown in Fig 1. The figure comprises examples of different emotions distributed in a high-low arousal and positive-negative valence dimensions.

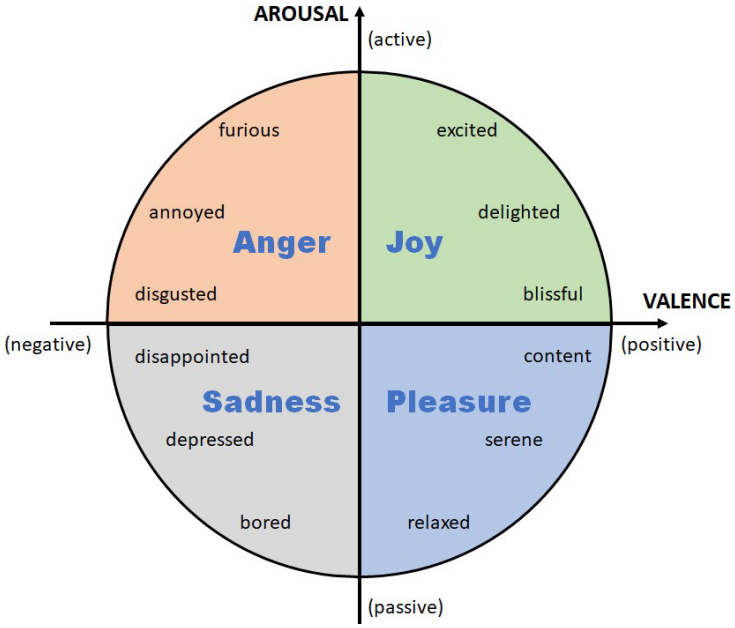

Fig. 1 Arousal-valence emotion distribution chart[36] *

2.3 The Data

Three publicly available data sets were used in this work: CLAS,[14] ASCERTAIN,[15] and AMIGOS.[16] The CLAS data set is a database for cognitive load, affect and stress recognition. The ASCERTAIN data set aims to investigate emotion and personality recognition using commercial sensors, while the AMIGOS data set examines relation between affect, personality, and mood in individuals and groups. The three data sets contain recordings of the two physiological signals of interests, namely ECG and GSR. In addition to the common peripheral physiological measures, all data sets used a similar experimental paradigm in which participants were asked to look at pictures or watch video clips found to elicit certain emotional responses. A summary of the data sets is described in Table 1.

Table 1 Summarized description of the data sets

Data set	Participants	Trials	Stimuli	Physiological signals
CLAS[14]	62	1,888	16 pictures and 16 videos	ECG, PPG, and GSR
ASCERTAIN[15]	58	2,088	36 videos	EEG, ECG, GSR, and visual
AMIGOS[16]	40	12,580	16 short and 4 long videos	Audio, visual, depth, EEG, ECG, and GSR

2.4 Physiological Signal Feature Extraction

Following the description provided by each data set, the raw ECG and GSR recordings as well as arousal/valence labeling were extracted and processed for each stimulus trial for each participant. For the CLAS data set, a total of 1888 trials were extracted from 59 subjects (each subject watched 16 pictures for 20 s each and 16 videos for 60 s each, totaling 32 trials per subject; three subjects were excluded because of missing data). A total of 2088 trials were extracted from 58 subjects for the ASCERTAIN data set (each subject watched 36 video clips of duration between 51 and 127 s long). The AMIGOS data yielded a total of 12580 trials from 37 subjects (340 20-s clips/trials per participant: 94 corresponding to the 16 short and 246 to the 4 long videos experiment; three subjects do not have labeling data). As each data set contained different sets of stimulus types (pictures vs. videos), we elected to combine trials across stimulus types for each data set to maximize the total amount of data available for modeling purposes.

2.5 ECG Features

ECG is a measure of electrical activity of the heart that has been shown not only as an indicator of the function of the heart but also as being controlled by the autonomic nervous system, which is the interplay between the sympathetic and parasympathetic branches of the nervous system. Therefore, heart rate (HR)-based features, especially changes in variability in the RR intervals of the ECG signals, or so-called the heart-rate variability (HRV), are known to associate with the physiological state, such as stress level, as well as overall health and well-being. Following BioSPPy,[37] the biosignal processing Python package, the 256-Hz ECG signals from the three data sets were filtered between 3 and 45 Hz and R peaks were detected. For each trial of the ECG signals, a time series of RR interval or the inter-beat-interval (IBI) were extracted to compute HR-based time- and frequency-domain features (Table 2). Some of the features were previously used in Kalinkov et al.[38] for arousal recognition.

Table 2 ECG-based features

	ECG-based features (67)
Statistical features	Mean HR; derivative of the HR (dHR); root mean square of successive differences (RMSSD); standard deviation of NN or R-R intervals (SDNN); standard deviation of the differences between adjacent RR intervals (SDSD); percentage of the differences between adjacent RR intervals that are greater than 20 ms (SD20); percentage of the differences between adjacent RR intervals that are greater than 50 ms (SD50); Poincaré plot standard deviation perpendicular the line of identity (SD1); Poincaré plot standard deviation along the line of identity (SD2); ratio of SD1-to-SD2 (SD1/SD2); variance of the RR intervals
Frequency-domain features	Low-frequency (LF) total HRV power [0.01,0.08]; mid-frequency (MF) [0.08,0.15] total HRV power; high-frequency (HF) [0.15,0.5] total HRV power; ratio of the LF/HF HRV power; 8 spectral components of the LF HRV; 8 spectral components of the MF HRV; 36 components of the HF HRV

2.6 GSR Features

GSR, or EDA, is a measure of skin conductivity produced by the sweat glands that are controlled by the sympathetic branch of the autonomic nervous system. An increase in skin conductance has been shown to associate with an elevated level of arousal. A total of 55 GSR-based features were extracted from the GSR measurements for the three data sets. Before extraction of the relevant features, the GSR signals were processed with 1-Hz lowpass filter. Phasic and tonic components of the filtered and Z-scored GSR signals were computed using cvxEDA Python algorithms.[39,40] In addition, the statistical time- and frequency-domain features used in Soleymani et al.[41] and Kalinkov et al.[38] were also computed, as seen in Table 3.

Table 3 GSR-based features

	GSR-based features (55)
Statistical features	Mean; derivative; variance; standard deviation; derivative of the negatives; ratio of the negative to total length of the signal; number of minima and maxima; rise time; peaks; median absolute deviation (MAD); minimum; maximum; energy; slope; intercept; entropy; interquartile range; autoregressive coefficients; skewness; kurtosis; index of maximum
Time-domain features	Phasic and tonic components of the filtered GSR signal (magnitude, number of peaks, average width of the peaks, area under the curve); number of zero crossings and rate of zero crossing of the filtered GSR signal with a cutoff of 0.08 Hz (very LF [VLF]) and 0.2 Hz (LF)
Frequency-domain features	Power in the band [0,2.4] Hz; peaks of the VLF and LF; 13 spectral components of the GSR signal between 0 and 2.4 Hz

2.7 AutoGluon-Tabular

With the recent development of neural network approaches that can perform well on tabular data, we are interested to see if these approaches can be used to predict individual emotional state (arousal or valence) with features extracted from ECG and GSR signals. For this, we use AutoGluon-Tabular,[13] an open-source machine learning framework that is easy to use and enables quick prototyping of neural networks and classical solutions on tabular data. It leverages automatic hyperparameter tuning, model selection/ensembling, as well as architecture search to achieve optimal solutions. The AutoGluon-Tabular Python package fits a total of 13 different models across a family of approaches that includes gradient boosting machines[42] (LightGBM, LightGBMXT, LightGBMLarge, CatBoost,[43] XGBoost[44]), random forests (RandomForestGini, RandomForestEntr), extra trees (ExtraTreesGini, ExtraTreesEntr), K-nearest neighbors (KNeightborsDist, KNeightborsUnit), and neural networks (NeuralNetFASTAI, NeuralNetMXNet). AutoGluon-Tabular then fits a weighted ensemble method that combines predictions from each model equally; this model is often used as the final model for predictions. Details about each modeling approach can be found in Erickson et al.[13]

3. Results

Before training the models and fitting the data, we Z-score normalized each feature (67 ECG-based features + 55 GSR-based features = 122 total) for each subject within each data set. This helps ensure that all features are on the same scale, in addition to helping minimize variability in physiological signals across individuals. For within-data set arousal/valence prediction, we used leave-one-subject-out cross-validation procedure in which one subject was held out for testing, from the remaining data set where 70% of subjects were randomly selected to be in the training set, while the remaining 30% of the data set was used as the validation set. The random train-validation split was repeated five times to ensure enough coverage of the randomized composition of the train-validation subject pool sets. Prediction accuracy for each subject was obtained from averaging the five runs. Cross-subject prediction results reported in Table 4 were achieved by taking the mean and standard deviation of all the runs across all the subjects. For cross-data set prediction, we used two combined data sets with five randomized 70/30 training-validation splits and tested on the third data set. The prediction accuracy was reported in Table 5 as an average of the five runs.

Table 4 Within data set classification accuracy from different ML models by AutoGluon-Tabular

Data set	CLAS		ASCERTAIN		AMIGOS	
Arousal(A)/Valence(V)	**A**	**V**	**A**	**V**	**A**	**V**
LightGBM	0.58±0.10	0.56±0.11	0.54±0.08	0.51±0.07	0.54±0.05	0.52±0.05
LightGBMXT	0.59±0.11	0.55±0.11	0.56±0.07	0.53±0.07	0.56±0.05	0.54±0.06
LightGBMLarge	0.59±0.10	0.56±0.09	0.54±0.08	0.52±0.08	0.55±0.05	0.52±0.04
CatBoost	0.59±0.11	0.56±0.11	0.56±0.08	0.51±0.08	0.56±0.05	0.54±0.06
XGBoost	0.58±0.08	0.57±0.11	0.55±0.08	0.53±0.07	0.55±0.05	0.53±0.04
RandomForestGini	0.60±0.09	0.56±0.11	0.56±0.09	0.52±0.08	0.55±0.05	0.54±0.04
RandomForestEntr	0.58±0.08	0.58±0.11	0.56±0.08	0.52±0.08	0.56±0.05	0.54±0.05
ExtraTreesGini	0.59±0.09	0.57±0.11	0.55±0.08	0.53±0.08	0.55±0.05	0.54±0.05
ExtraTreesEntr	0.59±0.10	0.58±0.12	0.56±0.08	0.52±0.08	0.56±0.05	0.54±0.05
KNeightborsDist	0.54±0.08	0.54±0.09	0.52±0.08	0.51±0.08	0.52±0.03	0.52±0.02
KNeightborsUnif	0.54±0.08	0.54±0.09	0.52±0.08	0.51±0.08	0.52±0.03	0.52±0.02
NeuralNetFASTAI	0.56±0.07	0.55±0.08	0.54±0.06	0.50±0.05	0.55±0.05	0.54±0.04
NeuralNetMXNet	0.57±0.07	0.55±0.07	0.54±0.06	0.51±0.04	0.55±0.05	0.53±0.04
WeightedEnsemble L2	0.59±0.09	0.56±0.10	0.56±0.07	0.51±0.06	0.56±0.05	0.54±0.05

Table 5 Cross–data set classification accuracy from different ML models by AutoGluon-Tabular

Train:	ASCERTAIN+AMIGOS		CLAS+AMIGOS		CLAS+ASCERTAIN	
Test:	CLAS		ASCERTAIN		AMIGOS	
Arousal(A)/Valence(V)	A	V	A	V	A	V
LightGBM	0.50	0.52	0.51	0.52	0.52	0.52
LightGBMXT	0.52	0.53	0.52	0.51	0.52	0.52
LightGBMLarge	0.51	0.53	0.51	0.51	0.51	0.52
CatBoost	0.54	0.52	0.52	0.51	0.53	0.53
XGBoost	0.51	0.53	0.51	0.51	0.51	0.53
RandomForestGini	0.52	0.52	0.51	0.51	0.53	0.52
RandomForestEntr	0.54	0.53	0.52	0.53	0.53	0.52
ExtraTreesGini	0.53	0.52	0.51	0.51	0.54	0.52
ExtraTreesEntr	0.54	0.53	0.51	0.51	0.53	0.53
KNeightborsDist	0.53	0.51	0.51	0.50	0.51	0.51
KNeightborsUnif	0.53	0.51	0.51	0.50	0.51	0.51
NeuralNetMXNet	0.52	0.53	0.50	0.51	0.52	0.51
WeightedEnsemble L2	0.52	0.53	0.51	0.51	0.53	0.52

Tables 4 and 5, respectively, show within- and cross–data set arousal (A) and valence (V) classification accuracy produced by different models used in AutoGluon-Tabular. Overall, results are consistently within 5%–10% higher than chance, which is similar to prior works focusing on cross–data set prediction of user affective state, albeit with a different set of data sets.[12] Neural network architecture-based models performed a few percentage points lower than the DT-based

classifiers. Results reported here could not be directly compared with others, as ways of splitting the data for training, validating, and testing could yield widely different classification results. We employed a leave-one-out cross-validation as a way to build models that could perform subject-independent classification. Considering inter-individual variability in physiological and emotional responses, developing subject-independent models is still a significant challenge.

4. Discussion and Conclusions

In this report, we evaluated the feasibility of using AutoGluon-Tabular to classify arousal and valence for emotional recognition across individuals and data sets using ECG and GSR signals. We have also developed a pipeline to process physiological signals, extract meaningful features, and implement subject-independent classification, both within and across data sets. Our classification results using *AutoGluon-Tabular* suggest that subject-independent classification of affective state remains a significant challenge, with classification accuracies only marginally above random chance. Given this result, there are a few discussion points that we would like to highlight:

- Given the relatively small data sets used in this study (CLAS, ASCERTAIN, AMIGOS), we decided to aggregate trials from multiple stimulus types (pictures vs. videos). This introduces a potential confound in our analysis, as trials with pictures as stimulus were shorter than trials with videos as stimulus (see Section 2.4). Thus, this may impact the quality/robustness of the feature extraction process as these trials were of different lengths. It is certainly possible that this could degrade classifier performance in unexpected ways. Further investigation is needed to determine if this is the case.
- We were limited to using only GSR and ECG as features, since these were the only two signal modalities available across all three data sets. It is indeed possible that the omitted signal features (features found in only a subset of the three data sets) could provide the additional information needed to accurately model user affective state. For example, EEG, which measures electrical activity in the brain using small electrodes attached to the scalp, provides highly salient information for modeling user affective state[12] and was recorded in two of the three data sets (ASCERTAIN and AMIGOS). However, because it is fairly cumbersome to apply, is easily contaminated by noise sources called artifacts,[45] and is in general not readily available in the general consumer application domain, we opted not to include it in our evaluation, as our focus was on modeling user affective state with peripheral

physiological signals. This highlights one of the biggest challenges in multi-data set learning, which is the fact that not all data sets contain the same recording modalities, making learning across data sets challenging.

- Prior work focusing on user affective state classification across multiple data sets also reported accuracies for arousal/valence around approximately 62%,[12] which is not significantly different than what we observed in our results (56%). Interestingly, even when using EEG as the signal modality, the results were not that different (also around 62%), reflecting the significant challenge in modeling user affective state across data sets. We note that our results are not directly comparable to these, however, as Siddharth et al.[12] use a different combination of data sets and different set of recording modalities.
- Inter-individual variability in emotional responses plays a major role in determining one's physiological responses. Data normalization procedures such as Z-scoring were insufficient to normalize the data across individuals, suggesting other forms of data normalization are needed.
- Emotional response to each stimulus is a continuous and multidimensional space as opposed to a discrete 2-D distribution. Therefore, "ground truth" labels for arousal and valence assigned to each stimulus may not accurately represent actual emotional elicitation. This introduces a form of *label noise*,[46,47] the possibility that the labels assigned to the data are incorrect. This can hinder the ability to train machine learning approaches if the amount of label noise is high relative to the size of the data set. In addition, it is possible that only a handful of features are important, salient, and useful, thus modeling extra features could serve as additional noise that can degrade classifier performance. AutoGluon-Tabular is a fully automated machine learning approach that performs several optimizations including hyperparameter tuning, feature selection, and ensembling, which should be able to account for this phenomena. However, reducing the number of features used as input through a priori selection (e.g., from existing literature) could help classification performance.
- While AutoGluon-Tabular generally performed well on large tabular data containing more than 100K instances and over 1000 features, it might not work as well for physiological features containing less than 1 order of magnitude of the features and instances data. To the best of our knowledge, this is the first attempt at using tabular neural network approaches for user affective state classification from peripheral physiological signals. Further investigation across more data sets is needed to determine if tabular neural networks can be used for affective state classification.

5. References

1. Schmidhuber J. Deep learning in neural networks: an overview. Neural Networks. 2015;61:85–117.

2. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.

3. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, editors, Advances in neural information processing systems, vol. 25. Curran Associates, Inc., 2012.

4. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770–778.

5. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences. 2014;111(23):8619–8624.

6. Chui M, Manyika J, Miremadi M, Henke N, Chung R, Nel P, Malhotra S. Notes from the AI frontier: insights from hundreds of use cases. McKinsey & Company; 2018 [accessed 2022 Oct 11]. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/notes%20from%20the%20ai%20frontier%20applications%20and%20value%20of%20deep%20learning/notes-from-the-ai-frontier-insights-from-hundreds-of-use-cases-discussion-paper.ashx.

7. Kotsiantis SB. Decision trees: a recent overview. Artificial Intelligence Review. 2013;39(4):261–283.

8. Arik SÖ, Pfister T. Tabnet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence. May 2021;35(8):6679–6687.

9. Kadra A, Lindauer M, Hutter F, Grabocka J. Well-tuned simple nets excel on tabular datasets. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, editors, Advances in Neural Information Processing Systems. 2021;34:23928–23941. Curran Associates, Inc.

10. Picard RW, Vyzas E, Healey J. Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001;23(10):1175–1191.

11. Breazeal C. Social interactions in HRI: the robot view. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2004;34(2):181–186.

12. Siddharth S, Jung T-P, Sejnowski TJ. Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing, 2019.

13. Erickson N, Mueller J, Shirkov A, Zhang H, Larroy P, Li M, Smola A. AutoGluon-Tabular: robust and accurate AutoML for structured data. arXiv preprint arXiv:2003.06505, 2020.

14. Markova V, Ganchev T, Kalinkov K. CLAS: a database for cognitive load, affect and stress recognition. BIA-2019. 2019;art. no. 8967457:171.

15. Subramanian R, Wache J, Khomami MA, Vieriu RL, Winkler S, Sebe N. Ascertain: emotion and personality recognition using commercial sensors. IEEE Transactions on Affective Computing. 2018;9(2):147–160.

16. Miranda-Correa JA, Abadi MK, Sebe N, Patras I. Amigos: a dataset for affect, personality and mood research on individuals and groups. IEEE Transactions on Affective Computing. 2021;12(2):479–493.

17. Picard RW. Affective computing. MIT Press, 1997.

18. Wu S, Xu X, Shu L, Hu B. Estimation of valence of emotion using two frontal EEG channels. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2017. p. 1127–1130.

19. Valenza G, Citi L, lanatà A, Scilingo E, Barbieri R. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. Scientific Reports. 05 2014;4:4998.

20. Mirmohamadsadeghi L, Yazdani A, Vesin J-M. Using cardio-respiratory signals to recognize emotions elicited by watching music video clips. 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP); 2016. p. 1–5.

21. Wu C-K, Chung P-C, Wang C-J. Representative segment-based emotion analysis and classification with automatic respiration signal segmentation. IEEE Transactions on Affective Computing. 2012;3(4):482–495.

22. Poh M-Zr, Swenson NC, Picard RW. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Transactions on Biomedical Engineering. 2010;57(5):1243–1252.

23. Zhang Y-D, Yang Z-J, Lu H-M, Zhou X-X, Phillips P, Liu Q-M, Wang S-H. Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. IEEE Access. 2016;4:8375–8385.

24. Lim JZ, Mountstephens J, Teo J. Emotion recognition using eye-tracking: taxonomy, review and current challenges. Sensors. 2020;20(8).

25. Tarnowski P, Kowlodziej M, Majkowski A, Rak R. Eye-tracking analysis for emotion recognition. Computational Intelligence and Neuroscience. 2020 Aug;2020:1–13.

26. Dai W, Han D, Dai Y, Xu D. Emotion recognition and affective computing on vocal social media. Information Management. 2015;52(7):777–788. Novel Applications of Social Media Analytics.

27. Majkowski A, Kolodziej M, Rak RJ, Korczyski Robert. Classification of emotions from speech signal. In 2016 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA). 2016;276–281.

28. Katsis CD, Katertsidis N, Ganiatsas G, Fotiadis DI. Toward emotion recognition in car-racing drivers: a biosignal processing approach. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 2008;38(3):502–512.

29. Machot FA, Elmachot A, Ali M, Al Machot E, Kyamakya K. A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors. Sensors. 2019;19(7).

30. Keren G, Kirschstein T, Marchi E, Ringeval F, Schuller B. End-to-end learning for dimensional emotion recognition from physiological signals. 2017 IEEE International Conference on Multimedia and Expo (ICME). 2017;985–990.

31. Gjoreski M, Gjoreski H, Lustrek M, Gams M. Deep ensembles for inter-domain arousal recognition. In: Hsu W, Yates H, editors, Proceedings of IJCAI 2018 2nd Workshop on Artificial Intelligence in Affective Computing. Proc Machine Learning Research (PMLR). 2020 July 5;80:52–64.

32. Khan AN, Ihalage AA, Ma Y, Liu B, Liu Y, Hao Y. Deep learning framework for subject-independent emotion detection using wireless signals. PLOS One. 2021 Feb;16(2):1–16.

33. Radhika K, Oruganti VRM. Deep multimodal fusion for subject-independent stress detection. In: 2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence); 2021. p. 105–109.

34. Arjun, RAS, Panicker MR. Subject independent emotion recognition using EEG signals employing attention driven neural networks. arXiv; 2021. arXiv:2106.03461.

35. Zhong P, Wang D, Miao C. EEG-based emotion recognition using regularized graph neural networks. arXiv; 2019. arXiv:1907.07835.

36. Yazdani A, Skodras E, Fakotakis N, Ebrahimi T. Multimedia content analysis for emotional characterization of music video clips. EURASIP Journal on Image and Video Processing. 2013.

37. Carreiras C, Alves AP, Lourenço A, Canento F, Silva H, Fred A, et al. BioSPPy: biosignal processing in Python. Instituto de Telecomunicacoes Revision; 2015–2018. https://biosppy.readthedocs.io/en/latest/.

38. Kalinkov K, Markova T, Ganchev T. Front-end processing of physiological signals for the automated detection of high-arousal negative valence conditions. Proc X National Conference with Intern Participation (ELECTRONICA-2019). 2019:1–4.

39. Greco A, Valenza G, Lanata A, Scilingo EP, Citi L. cvxEDA: A convex optimization approach to electrodermal activity processing. IEEE Transactions on Biomedical Engineering. 2016;63(4):797–804.

40. Ferdinando H, Alasaarela E. Emotion recognition using cvxEDA-based features. Journal of Telecommunication, Electronic and Computer Engineering. 2018;10:19–23.

41. Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing. 2012;3(1):42–55.

42. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001;1189–1232.

43. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. Catboost: Unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18; 2018. p. 6639–6649. Curran Associates Inc.

44. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug.

45. Lawhern V, Hairston WD, McDowell K, Westerfield M, Robbins K. Detection and classification of subject-generated artifacts in EEG signals using

autoregressive models. Journal of Neuroscience Methods. 2012;208(2):181–189.

46. Nettleton DF, Orriols-Puig A, Fornells A. A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial Intelligence Review. 2010;33(4):275–306.

47. Natarajan N, Dhillon IS, Ravikumar PK, Tewari A. Learning with noisy labels. In: Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors, Advances in neural information processing systems, vol. 26. Curran Associates, Inc.; 2013.

List of Symbols, Abbreviations, and Acronyms

2-D	two-dimensional
CNN	convolutional neural network
dHR	derivative of the HR
DT	decision tree
ECG	electrocardiogram
EDA	electrodermal activity
EEG	electroencephalogram
GSR	galvanic skin response
HCI	human–computer interaction
HF	high frequency
HR	heart rate
HRV	heart-rate variability
IBI	inter-beat-interval
LF	low frequency
MAD	median absolute deviation
MF	mid frequency
PPG	plethysmography
RMSSD	root mean square of successive differences
SD1	Poincaré plot standard deviation perpendicular the line of identity
SD2	Poincaré plot standard deviation along the line of identity
SD20	percentage of the differences between adjacent RR intervals that are greater than 20 ms
SD50	percentage of the differences between adjacent RR intervals that are greater than 50 ms
SDNN	standard deviation of NN
SDSD	standard deviation of the differences
VLF	very LF

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLD DCI TECH LIB
2 (PDF)	DEVCOM ARL FCDD RLH FE D CHHAN V LAWHERN