2021 CMBEC44
Virtual Conference
CMBES/SCGB May 11-13, 2021

Human Mental State Monitoring in the Wild: Are We Better Off with Deeper Neural Networks or Improved Input Features?

Arthur Pimentel †[1], Abhishek Tiwari†[1], Shrikanth Narayanan[2] and Tiago H. Falk[1]

[1] INRS-EMT, Université du Québec, Montréal, Québec, Canada
[2] University of Southern California, Los Angeles, USA

Abstract— **Advances in wearable devices have allowed for the collection of multimodal biomedical data from hundreds of subjects in everyday environments (i.e., in the wild). This has enabled the development of real-time monitoring of various human mental states, such as stress and anxiety, in highly ecological settings. Within a hospital setting, for example, this allows for prediction of burnout within medical staff, as well as anxiety within the patient population, thus improving their quality-of-life. Long-term monitoring via wearables has allowed for large amounts of data to be collected – so-called big data – and thus has opened doors for new applications relying on data-heavy deep learning algorithms. One question that remains unanswered, however, concerns the benefits of blindly applying deep learning algorithms with the collected data versus spending some time and resources on feature engineering prior to machine learning. Feature engineering relies on domain knowledge to extract relevant parameters from the collected signals. In this paper, we aim to answer this question. In particular, we use a dataset collected from 200 hospital workers over a period of 10 weeks during their work shifts. We compare the advantages of using data directly from the wearable devices and applying them to deep learning algorithms versus carefully-crafted features applied to conventional machine learning algorithms. Experimental results are reported for stress and anxiety measurement from heart and breathing rate signals.**

Keywords— **Deep Learning, feature engineering, mental state monitoring, cardio-pulmonary signals, HRV.**

I. Introduction

There is no doubt that the COVID-19 pandemic has put an added burden on healthcare workers around the world, with reports of increasing burnout rates [1]. Moreover, record-high levels of stress and anxiety are being reported by the general population; in the USA alone, a recent survey showed that 8 in 10 adults reported significant increases in stress levels due to the pandemic [2]. As such, monitoring of stress and anxiety in real-time has become an extremely important topic.

Wearable technologies have played a key role in monitoring human mental states even prior to the pandemic [3]. Wearables allow individuals to move freely within the environment (i.e., "in the wild" compared to controlled laboratory conditions) and allows for long-term monitoring of numerous psychophysiological parameters. The resultant large amounts of data, in turn, have been crucial for the development of new monitoring and diagnostic applications relying on recent deep machine learning algorithms (e.g., [4–6]. The optimal depth of such "deep" algorithms is not known beforehand and is usually optimized on a per-case basis to trade-off feature generation and the non-linear mapping between input data (measured raw or pre-processed signals) and output labels (stress and anxiety levels in our case) [7].

An alternative to this fully-data-driven approach is the more classical approach that relies on domain knowledge to carefully craft features (termed "feature engineering") that lend high discriminatory power, and interpretability, based on prior psychophysiological insights. Such features can then be coupled with conventional machine learning algorithms. One question that remains unanswered is whether blindly applying data from wearable devices directly to deep neural networks is better than carefully crafting features and using conventional classifiers. Here, we aim to answer this question.

While deep neural networks can find optimal features and mappings in a data-driven way, in-the-wild measurements are known to generate many confounding factors, including movement artifacts, which could lead to erroneous decisions. Such limitations could be overcome with artifact-robust feature engineering. Moreover, it is known that deep neural networks have many hyper-parameters that need to be optimized, thus take a large amount of time and computational resources, thus leaving a large carbon footprint [8] with oftentimes just incremental improvements [9].

To help answer this question, we rely on a recently-collected dataset of 200 hospital staff who were monitored daily for a period of 10 weeks while wearing a smartshirt that collected cardiac, pulmonary, and activity information [10]. Self reported ratings of stress and anxiety were also collected daily. Two sets of features were computed. First, benchmark features measured by the device itself were used. These in-

†Equal Contribution

cluded several time- and frequency-domain heart rate variability (HRV) features measured from an electrocardiogram signal [11], as well as breathing rate and breathing depth [12]. Next, hand-crafted features that take the non-linear fractal dynamics of physiological signals into account were extracted [13, 14]. These features have been shown to decouple confounding factors, including movement artifacts [15–17]. We then compare the results of using benchmark features with deep neural networks of varying depths against the results achieved with feature engineering and conventional support vector classifiers.

II. Materials and Methods

A. Participants

The TILES dataset we use in this study is from 200 participants (66 male, 134 female; age 38.6 ± 9.8 years) from a pool of employees (nurses and staff) of a large urban hospital in California. Two-thirds of the participants were nurses while one-third were hospital staff. Data were collected for a duration of 10 weeks. Participants consented to participate in the study, which received ethical board approval from the affiliated institutions. Complete details about this publicly available dataset can be found in [10]. Participants carried out their work day as usual but were asked to fill a brief phone-based daily survey that included information on levels of anxiety and stress on a 5-point scale. Participants were outfitted with multiple wearable sensors to collect a variety of biometric data, including audio features, heart rate, respiratory rate, and sleep quality. More specifically, a custom audiometric badge, a Fitbit Charge 2, and an OMsignal smartshirt were used. In this paper, only the cardiac and respiratory information measured by the OMsignal smartshirt are used.

B. Feature extraction and aggregation

Standard time- and frequency-domain HRV metrics were used as benchmark measures. These are typically extracted and used by smart devices. A complete list can be found in Table 1. For breathing, the following features directly provided by the OMSignal garment were used: mean and standard deviation of the instantaneous breathing rate (f_R), and breathing depth (b_D). These features have been shown in the literature to correlate with stress [12, 18] and anxiety [12, 15].

More recently, novel HRV and breathing features have been proposed to take into account non-linear behaviour of cardio-pulmonary signals and interactions, as well as to provide increased robustness against movement artifacts [16, 17]. Specifically, multi-scale permutation entropy HRV features have been shown to better quantify the complexity of the heartbeat time series at difference scales [14]. When coupled with motif features [19], improved robustness to artifacts was achieved. For breathing, complex dynamical behaviour [13, 20] can be modelled by extracting statistical and complexity features from three time series: inhale-to-exhale ratio, inter-breath interval, and amplitude envelope series. The interested reader is referred to [16, 17] for complete details.

All features were extracted over 5-minute long windows and further aggregated over an entire day using the following statistical functionals: mean, standard deviation, coefficient of variation, median, min, max, 1st and 3rd quartile, skewness and kurtosis. Additionally, the OMsignal smart garment provides a quality metric for its heart rate series measurement termed *RRPeakCoverage*. As such, several new quality-aware features have been developed, thus providing context to the task at hand. The features include quality weighted mean, standard deviation, and coefficient of variation. Overall, 250 breathing features (40 benchmark and 210 proposed) and 1036 HRV (182 benchmark and 854 proposed) features are available for analysis.

Table 1: Different groups of HRV features extracted

Time domain HRV features
mean, standard deviation, coefficient of variation, rmsdd, pNN50, mean of 1^{st} diff., standard deviation of absolute of 1^{st} diff., normalized mean of absolute 1^{st} diff
Frequency domain HRV features
High frequency power (HF), normalized HF, Low frequency power (LF), normalized LF, very low frequency power, HF/LF

C. Deep learning models

Deep neural networks have shown to provide state-of-the-art performance across numerous domains. By modifying the number of hidden layers and neurons, one can adjust the complexity of the internal feature generation and non-linear mapping between input signals and output labels of interest. Here, we explore three different network depths to gauge the benefits of increased model complexity, given benchmark features as input. These included a deep (DNN, 8 hidden layers), a medium (MNN,5 hidden layers), and a shallow (SNN,3 hidden layers) multilayer perceptron (MLP) network. Increasing the number of layers, in turn, also increases the number of parameters that need to be optimized and stored, as well as training time [7]. Here, all hidden layers used a *ReLU* activation function with a *sigmoid* activation for the output layer. Dropout was used for regularization for the first two layers, L2 weight regularization was used in all layers and *Adam* was

Table 2: Comparison of evaluated models

Architecture	No. parameters	L2 parameter	Dropout rate
SNN	14,337	0.001	0.1
MNN	16,449	0.001	0.1
DNN	19,617	0.0001	0.3

used as the optimizer. The networks have similar architectures, where the first hidden layer has 64 neurons, the last hidden layer has 16 neurons and the layers in between have 32 neurons each. The number of layers with 32 neurons depends on the total number of hidden layers. The model architectures along with the total number of parameters and regularization parameters are described in Table 2.

For hyper-parameter tuning, the data-set was split in an 80-20% split with a random seed. Dropout rate and L2 regularization parameter were then selected using grid search. Model hyper-parameters were optimized for the benchmark dataset. A learning rate of 0.01 and 100 epochs were chosen for both stress and anxiety.

D. Classification and feature selection

A 5-fold cross-validation setup was repeated five times with different random seeds to shuffle the data resulting in 25 (5-fold X 5) unique train and test set combinations for a robust evaluation of the prediction pipeline. Training traditional machine learning classifiers with a large feature set may lead to overfitting with many features also being highly correlated. As such, recursive feature elimination was performed with a step size of 10 using the Extra Trees Classifier. The top 100 features are then selected for classification at each cross validation step. For classification, a binary task was chosen and the stress and anxiety ratings typically fell into two clusters, i.e., high/low stress and anxiety levels.

For comparisons, a conventional support vector machine (SVM) classifier with an RBF kernel and a 'balanced' setting is also explored. This setting uses the target value to automatically adjust weights inversely proportional to class frequencies in the input data [21]. As the data is imbalanced (% imbalance for stress : 0.582 and anxiety: 0.432), balanced accuracy (BACC), F1-score (F1) and Matthews correlation coefficient (MCC) [22] are used as figures-of-merit. Additional figures-of-merit include number of parameters that need to be stored and training time. The pipeline and evaluation metrics are implemented using scikit learn [21] and keras [23].

As we are interested in answering the question whether deeper models with benchmark features are better than conventional classifiers but with carefully-crafted features, our experiments focus on the use of benchmark features with the deep neural networks (and an SVM for comparisons) and the engineered features with the SVM.

Table 3: Performance comparison of HRV features (* represents significance ($p < 0.001$) compared to NN models)

	Stress			Anxiety		
Model	BACC	F1	MCC	BACC	F1	MCC
SNN	0.606	0.684	0.216	0.593	0.528	0.188
MNN	0.601	0.672	0.204	0.590	0.529	0.182
DNN	0.600	0.666	0.203	0.592	0.523	0.187
SVM-Bench	0.624*	0.657	0.245	0.601	0.545	0.203
SVM-Proposed	0.652*	0.681	0.300*	0.630*	0.582*	0.260*

III. Results and Discussion

Tables 3 and 4 show the results for stress and anxiety prediction using HRV and breathing features, respectively. Results for varying-depth neural network models and benchmark features, along with SVM models for both benchmark (SVM-Bench) and proposed (SVM-Proposed) features are shown. All models perform significantly better than a random voting classifier ($p < 0.01/6$, with bonferroni correction).

As can be seen, in all cases, the SVM-proposed pipeline significantly outperformed ($p < 0.01/10$, bonferroni correction) all the NN models in BACC and MCC metrics. For the F1 metric, in case of stress prediction, the performance is similar to NN models for both HRV and breathing, while being significantly better for anxiety. Compared to the best performing NN models, for HRV features, the SVM-Proposed pipeline shows an improvement of 7.5% BACC for stress and 6.2% for anxiety, while for breathing features, it shows an improvement of 8.4% for stress and 6.7% for anxiety over the best performing NN models.

The SVM-bench pipeline, in turn, performs similar to NN models in both the BACC and MCC metric in most cases, with significant improvement observed for BACC metric with HRV based stress prediction. For the F1-metric, the NN models consistently outperform the SVM-bench for stress in both HRV and breathing feature sets. While for anxiety, the SVM-bench pipeline is either similar (for HRV) or significantly better (for breathing) than the NN models. Similar findings have been reported recently for neuroimaging studies [24], suggesting that the non-linearities in data may not be exploitable at available sample sizes, and simpler algorithms may perform equally well. Finally, all three NN models (SNN, MNN and DNN) have comparable performance with no significance performance improvement with increasing depth.

Table 5 shows the average training and hyper-parameter tuning times across stress and anxiety for all the models for HRV and breathing features. Although the training times are comparable, the hyper-parameter tuning required for NN models leads to about 6-10 times the total time compared to SVM-proposed pipeline for HRV features and about 19-26 times the total time compared to SVM-proposed pipeline for breathing features while giving better performance.

Table 4: Performance comparison of breathing features (* represents significance ($p < 0.001$) compared to NN models)

	Stress			Anxiety		
Model	BACC	F1	MCC	BACC	F1	MCC
SNN	0.587	0.682	0.182	0.593	0.517	0.192
MNN	0.593	0.672	0.189	0.588	0.514	0.181
DNN	0.581	0.679	0.168	0.593	0.527	0.189
SVM-Bench	0.588	0.637	0.174	0.591	0.562*	0.180
SVM-Proposed	0.643*	0.680	0.283*	0.633*	0.592*	0.265*

Table 5: Time for hyper-parameter tuning and training (in minutes)

	HRV		Breathing	
Model	Tuning	Training	Tuning	Training
SNN	125.42	8.6	124.88	8.48
MNN	137.28	10.95	141.8	10.86
DNN	175.28	11.42	168.48	11.09
SVM-Bench	0	5.3	0	1.1
SVM-Proposed	0	18.3	0	6.75

IV. Conclusion

In this paper, we show that feature engineering methods combined with simpler machine learning pipelines are capable of outperforming benchmark features with different neural networks architectures for mental state monitoring applications for "in-the-wild" conditions. These results show the importance of feature engineering especially for data collected in highly ecological settings. Future work will explore the fusion of HRV and breathing features, the use of crafted features with deep neural networks, as well as other deep neural network architectures.

Acknowledgements

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

1. Kannampallil Thomas G, Goss Charles W, Evanoff Bradley A, Strickland Jaime R, McAlister Rebecca P, Duncan Jennifer. Exposure to COVID-19 patients increases physician trainee stress and burnout *PloS one.* 2020;15:e0237301.
2. Association American Psychological, others . Stress in AmericaTM 2020: A national mental health crisis 2020.
3. Peake Jonathan M, Kerr Graham, Sullivan John P. A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations *Frontiers in physiology.* 2018;9:743.
4. Torres-Soto Jessica, Ashley Euan A. Multi-task deep learning for cardiac rhythm detection in wearable devices *NPJ digital medicine.* 2020;3:1–8.
5. Giannakakis G., Trivizakis E., Tsiknakis M., Marias K.. A novel multi-kernel 1D convolutional neural network for stress recognition from ECG in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*:1-4 2019.
6. Masood K., Alghamdi M. A.. Modeling Mental Stress Using a Deep Learning Framework *IEEE Access.* 2019;7:68446-68454.
7. Goodfellow Ian, Bengio Yoshua, Courville Aaron, Bengio Yoshua. *Deep learning*;1. MIT press Cambridge 2016.
8. Anthony Lasse F Wolff, Kanding Benjamin, Selvan Raghavendra. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models *arXiv preprint arXiv:2007.03051.* 2020.
9. Musgrave Kevin, Belongie Serge, Lim Ser-Nam. A metric learning reality check *arXiv preprint arXiv:2003.08505.* 2020.
10. Mundnich Karel, Booth Brandon M, l'Hommedieu Michelle, et al. TILES-2018: A longitudinal physiologic and behavioral data set of hospital workers *arXiv preprint arXiv:2003.08474.* 2020.
11. Camm AJMM, others . Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology *Circulation.* 1996;93:1043–1065.
12. Masaoka Yuri, Homma Ikuo. Anxiety and respiratory patterns: their relationship during mental stress and physical load *International Journal of Psychophysiology.* 1997;27:153–159.
13. Peng C-K, Mietus Joseph E, Liu Yanhui, et al. Quantifying fractal dynamics of human respiration: age and gender effects *Annals of biomedical engineering.* 2002;30:683–692.
14. Costa Madalena D, Peng Chung-Kang, Goldberger Ary L. Multiscale analysis of heart rate dynamics: entropy and time irreversibility measures *Cardiovascular Engineering.* 2008;8:88–93.
15. Wen Wan-Hui, others . Toward Constructing a Real-time Social Anxiety Evaluation System: Exploring Effective Heart Rate Features *IEEE Transactions on Affective Computing.* 2018.
16. Tiwari Abhishek, Albuquerque Isabela, Parent Mark, et al. Multi-Scale Heart Beat Entropy Measures for Mental Workload Assessment of Ambulant Users *Entropy.* 2019;21:783.
17. Tiwari Abhishek, Narayanan Shrikanth, Falk Tiago H. Breathing Rate Complexity Features for "In-the-Wild" Stress and Anxiety Measurement in *2019 27th European Signal Processing Conference (EUSIPCO)*:1–5IEEE 2019.
18. Castaldo Rossana, Melillo Paolo, Bracale Umberto, Caserta M, Triassi Maria, Pecchia Leandro. Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis *Biomedical Signal Processing and Control.* 2015;18:370–377.
19. Zanin Massimiliano, Zunino Luciano, Rosso Osvaldo A, Papo David. Permutation entropy and its main biomedical and econophysics applications: a review *Entropy.* 2012;14:1553–1577.
20. Vlemincx Elke, Van Diest Ilse, Bergh Omer. A sigh following sustained attention and mental stress: effects on respiratory variability *Physiology & behavior.* 2012;107:1–6.
21. Pedregosa Fabian, others . Scikit-learn: Machine learning in Python *Journal of machine learning research.* 2011;12:2825–2830.
22. Chicco Davide, Jurman Giuseppe. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation *BMC genomics.* 2020;21:6.
23. Gulli Antonio, Pal Sujit. *Deep learning with Keras*. Packt Publishing Ltd 2017.
24. Schulz Marc-Andre, Yeo BT Thomas, Vogelstein Joshua T, et al. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets *Nature communications.* 2020;11:1–15.