On the Generalizability of ECG-based Stress Detection Models

Pooja Prajod
Human-Centered Artificial Intelligence
University of Augsburg
Augsburg, Germany
pooja.prajod@uni-a.de

Elisabeth André
Human-Centered Artificial Intelligence
University of Augsburg
Augsburg, Germany
elisabeth.andre@uni-a.de

Abstract—**Stress is prevalent in many aspects of everyday life including work, healthcare, and social interactions. Many works have studied handcrafted features from various bio-signals that are indicators of stress. Recently, deep learning models have also been proposed to detect stress. Typically, stress models are trained and validated on the same dataset, often involving one stressful scenario. However, it is not practical to collect stress data for every scenario. So, it is crucial to study the generalizability of these models and determine to what extent they can be used in other scenarios. In this paper, we explore the generalization capabilities of Electrocardiogram (ECG)-based deep learning models and models based on handcrafted ECG features, i.e., Heart Rate Variability (HRV) features. To this end, we train three HRV models and two deep learning models that use ECG signals as input. We use ECG signals from two popular stress datasets - WESAD and SWELL-KW - differing in terms of stressors and recording devices. First, we evaluate the models using leave-one-subject-out (LOSO) cross-validation using training and validation samples from the same dataset. Next, we perform a cross-dataset validation of the models, that is, LOSO models trained on the WESAD dataset are validated using SWELL-KW samples and vice versa. While deep learning models achieve the best results on the same dataset, models based on HRV features considerably outperform them on data from a different dataset. This trend is observed for all the models on both datasets. Therefore, HRV models are a better choice for stress recognition in applications that are different from the dataset scenario. To the best of our knowledge, this is the first work to compare the cross-dataset generalizability between ECG-based deep learning models and HRV models.**

Index Terms—**Stress, Deep learning, Convolutional neural networks, Recurrent neural networks, Machine learning, Support vector machines, Physiology, Heart rate variability, Electrocardiography**

I. Introduction

Stress recognition research has become an important part of affective computing, especially in applications involving human-computer interaction [1]. Long-term stress has severe consequences and hence, there is a need for automatic stress recognition to detect stress early [1], [2]. Stress stimuli or stressors trigger physiological responses in people which can be detected through different bio-signals such as Electrocardiogram (ECG) and Electrodermal Activity (EDA) [1]–[3]. So, stress recognition research is further facilitated by the increasing popularity of wearable sensors that can unobtrusively collect real-time bio-signal data [4], [5].

ECG is one of the most common bio-signal used in stress and affect recognition [5], [6]. There are two popular approaches to detect stress from ECG - models based on handcrafted Heart Rate Variability (HRV) features [1], [2], [4], [7] and deep learning models [5], [6], [8]. HRV features and their relationship with stress have been studied thoroughly [9], [10]. They have also been validated as indicators of stress in different stressful conditions [1], [4], [11]. However, cleaning the ECG signal and computing the HRV features often require specific domain knowledge [5]. This paved the way for deep learning models, which typically have convolution layers for automatic feature extraction.

We say an ECG-based stress recognition model has good generalization capability if it performs well on samples collected using different sensor devices under different stress conditions. It is crucial to evaluate the generalizability of a model as it is not possible to collect stress data and train specialized models in every scenario. In some cases, the models have to be trained on an available dataset and deployed in a scenario different from the training dataset. For example, a neuro-rehabilitation use-case described in [12] employs an agent which adapts exercises by taking into account the stress level of the patient. Due to ethical considerations, it is difficult to collect a dataset by stressing the patients during a rehabilitation session. Another example to consider is stress recognition for special groups of people, like people with autism spectrum disorder (ASD), dementia, etc. Often, there is a lack of stress datasets that includes data collected from these groups of people. Moreover, there could be differences in the intensity or the characteristics of stress responses of the people belonging to these groups. For instance, one of the datasets we consider in this study is the WESAD dataset [1], which uses social evaluation as a stressor. But, in [13], the authors found that children with ASD had blunted physiological stress response to social evaluation stressor.

In this work, we investigate if the models trained on one stress dataset can detect stress in another dataset. Specifically, among ECG deep learning models and HRV models, we determine if one group outperforms the other in detecting

This work is partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 847926 MindBot

stress samples from another scenario.

II. Related Work

Due to the health consequences of stress, there is extensive research on stress recognition. It is beyond the scope of this work to summarize the numerous works that improve stress recognition. So, we focus on works that compare various models or stress datasets to gain insights into trends pertaining to their performance.

There are multiple feature-based models proposed for stress recognition in various works. Bobade and Vani [2] compare the stress recognition performance of various machine learning models trained on hand-crafted features from various physiological signals. They use the WESAD dataset [1] to train K-Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA), Random Forest Classifier (RFC), Support Vector Machine (SVM), etc. They also propose a simple feed-forward Artificial Neural Network (ANN) trained on the same input. Their comparison shows that ANN achieves higher accuracy than other models.

As mentioned before, there are two main types of stress recognition models - deep neural networks and feature-based machine learning models. Naturally, questions arise on whether one type is better than the other. Zhang et al. [11] address this question by studying the performance of a deep neural network and feature-based models on a dataset they collected. They propose a stress recognition model consisting of both convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM). For comparison, they extract HRV features and train popular machine learning models like SVMs, RFC, Ada Boost, etc. The CNN-LSTM model takes $10\ s$ of raw ECG signal, whereas the other machine learning models use HRV features extracted from $60\ s$ of ECG data. Zhang et al. demonstrate that deep neural networks significantly outperform HRV-based models.

Dzieżyc et al. [14] compare various deep learning models on their performance in emotion recognition tasks (including stressful condition). An extensive study is performed on four different datasets, separately. They chose an input signal length of $50 - 60\ s$, which is longer than the typical input length for deep learning models. They note that CNN-based models tend to perform better than LSTM-based models.

All the above works train and test the stress recognition models on the same dataset. Cho et al. [15] consider two datasets differing in size and train ECG-based deep learning models to detect stress. They propose a transfer learning approach, which involves training a model on the bigger dataset and then fine-tuning it on the smaller dataset. They observe that the stress recognition on the smaller dataset improves through transfer learning. Other than the size, the datasets were very similar (e.g. same ECG sensor and configuration). The authors note that when data from other datasets are used, their model shows high bias to the type of stressor and a dependency on the sensor used. In line with this observation, Liapis et al. [16] demonstrate that a high stress recognition accuracy on one dataset does not necessarily translate to high accuracy in another dataset. To this end, they extract Skin Conductance (SC) features from the WESAD dataset [1] and train four machine learning models for stress recognition. These models achieve high accuracy while testing on the WESAD dataset. However, they did not achieve good results on input signals from a different dataset (UX evaluation dataset). Since the UX evaluation dataset is annotated primarily for emotion and not stress, it is difficult to conclude about the generalizability of the models. Nevertheless, their observation highlights the need for cross-dataset evaluations and assessing the generalizability of the stress recognition models.

As a first step towards combining stress datasets for developing generic models, Baird et al. [17] evaluate three datasets on their ability to predict cortisol values. Cortisol values are considered the ground truth for stress response. As they note, the scales of cortisol values of the datasets are incompatible and thus, a cross-dataset evaluation is not feasible. However, all three datasets were collected through similar Trier Social Stress Test (TSST) procedures. So, the responses in each condition of the test are expected to be similar and therefore, the trends in predicted cortisol values can be compared. To this end, they extract features from the speech signals in the datasets and train models for each dataset. They highlight the feasibility of using speech signals from one dataset as predictors of stress in another dataset.

III. Approach

Deep learning models trained directly on the ECG signal typically outperform hand-crafted HRV features on a given dataset [11]. However, it remains unexplored if these deep learning models perform equally well in cross-dataset evaluations. To investigate this, we train 5 stress recognition models - two deep learning models using ECG signals as input, and three models based on hand-crafted HRV features. First, we train and evaluate the stress models on the same dataset using leave-one-subject-out (LOSO) cross-validation. We perform this evaluation on two different datasets. Then, we evaluate the LOSO models trained on dataset A using samples from the other dataset B (cross-dataset evaluation) to assess their generalization capabilities. Baird et al. [17] note that machine learning models can benefit from combining stress datasets as it increases the data available for training. It has not been investigated if this holds true if the datasets are vastly different, especially in terms of the stressors, the intensity of stress experienced, and the brand of sensors used. So additionally, we train the models on a combined dataset (merging samples from the two datasets) and evaluate them using LOSO validation.

A. Datasets

1) WESAD: WESAD [1] is a multimodal dataset that contains motion (ACC) and physiological (ECG, EDA, etc.) signals, which were collected using chest-worn RespiBan and wrist-worn Empatica E4 devices. The data was collected from 15 participants under three conditions: baseline, stress, and amusement. Stress was elicited using the Trier Social Stress Test (TSST) involving public speaking and mental arithmetic (counting down from 2023 by steps of 17) tasks. The stress condition lasted for about 10 minutes. The amusement condition was around 6.5 minutes long, where the participants watched funny video clips. In this work, we use the ECG data from the chest-worn device, sampled at 700 Hz. To be consistent with the labels used by the authors, we consider baseline and amusement conditions as the no-stress class.

2) SWELL-KW: SWELL knowledge work dataset [3] contains data collected from 25 participants who did typical office tasks (writing reports and making a presentation) under three conditions - neutral, email interruptions, time pressure. During the email interruption session, 8 emails were sent - many were irrelevant, and some required a reply. In the time pressure session, the participants had to complete their tasks in $2/3$rd of the time allotted for the neutral session. The neutral and email interruption sessions lasted for around 45 minutes each, whereas the time pressure session lasted for around 30 minutes. We use the ECG signals (sampled at 2048 Hz), which were collected using a TMSI Mobi device. The participants did not report feeling stressed in any of the conditions. However, they indicated higher temporal demand (they felt time pressure due to the pace of the task) during the time pressure session. In a subsequent study [4], the authors labelled the data from email interruptions and time pressure sessions as stress and the neutral session as no-stress for a binary stress classification task. Hence, we also consider the data belonging to email interruptions and time pressure sessions as stress samples.

B. Classification Models

We describe our stress detection models and their training parameters below:

1) Deep ECGNet: This is a CNN-LSTM stress detection model proposed in [8]. The idea of CNN-LSTM networks is to use the CNN layers as feature extractors and train the LSTM layers to learn temporal patterns in the extracted features. The model has an initial convolution block containing a 1D convolutional layer, a pooling layer, a dropout layer, and a batch normalization layer. The activation function for the convolution layer is rectified linear unit (ReLU). The 1D convolution layer has 50 filters with a kernel size corresponding to 0.6 s. The pooling layer has a size equivalent to 0.8 s of data. For a 256 Hz input, kernel size is 154, and pooling size is 205. The convolution block is followed by two LSTM layers and a final prediction layer. The first LSTM layer has 32 units and the second one has 16 units. We add a dropout layer and a batch normalization layer between the two LSTM layers. The activation function for the LSTM layers is Tanh, and for the prediction layer is Softmax. We use a dropout rate of 0.2 for both dropout layers.

2) ECG Emotion Recognition Model: This CNN model is proposed in [6] for emotion recognition on various datasets, including WESAD and SWELL-KW. The model has three convolution blocks, each block consisting of two 1D convolution layers and a pooling layer. The convolution layers belonging to a block have identical parameters such as kernel size and the number of filters. From block 1 to block 3, the number of filters are 32, 64, and 128, whereas the kernel sizes are 32, 16, and 8. The pooling layers are of size 8 with strides of 2. The convolution blocks are followed by two fully connected layers with 128 nodes each. We add a dropout layer (dropout rate $=$ 0.6) after each fully connected layer. Finally, the model is connected to a stress prediction layer with Softmax activation. All the convolution layers and the fully connected layers have ReLU activation.

3) Multi-Layer Perceptron: This is a simple neural network with an input layer, two hidden layers, and a prediction layer. The hidden layers have 12 and 6 nodes. The activation function for hidden layers is ReLU and for the prediction layer is Sigmoid. To prevent over-fitting, we add a dropout layer (dropout rate $=$ 0.2) after the input layer.

4) RFC: This is an ensemble classifier that trains a certain number of decision trees on various subsets of the training set and uses their output to make a final prediction. This reduces over-fitting and improves the overall performance, even if individual classifiers are weak. Similar to [1], the number of decision trees (or estimators) is set to 100 and the minimum number of samples for splitting a node is set to 20.

5) SVM: This is a commonly used supervised learning model. Similar to [4], we use SVM with Radial basis function (Rbf) kernel.

We use Tensorflow to train neural networks and Scikit-learn to train other machine learning models. For all the models, we use a weighted loss to tackle class imbalance in the training dataset. For the neural network models, we use the Adadelta optimizer (learning rate $=$ 1.0) and cross-entropy loss. We train them for 200 epochs with a batch size of 128.

C. Evaluation metrics

We use *F1-score* and *Accuracy* metrics for evaluation. Accuracy is the ratio of number of correctly predicted samples to total number of samples in the test set. F1-score is computed as the harmonic mean of precision and recall. Precision is the number of correctly predicted samples of a class out of all the samples predicted to belong to the class. Recall is the number of correctly predicted samples of a class out of all the samples belonging to the class. To tackle the class imbalances in the datasets, we compute macro f1-score, i.e., compute f1-score for each class and average them. We perform within-dataset LOSO evaluation as it determines the generalizability of a model on data from unseen users. However, this is not an indicator of generalizability of the models on data collected using a different sensor or a different stressor. Hence, we evaluate the models using cross-dataset validation. This validation involves training a model on a dataset A and evaluating it using samples from another dataset B.

D. Pre-processing

The data collected in the two datasets have different sampling rates. This is not a concern for HRV features, but the ECG-based deep learning models require the input lengths to be the same. To keep the data consistent for all models, we down-sample the ECG signals in both datasets to 256 Hz.

There are various sources of noise in an ECG signal, including baseline wander, powerline interference, and EMG noise [18], [19]. Baseline wander is a low-frequency noise $(0.5 - 0.6\ Hz)$ that causes the signal to drift up and down. It is typically removed using a high-pass filter [6], [18], [19]. Powerline interference is caused by the electromagnetic interference of the power source of the sensor device. A common technique to remove this noise is using a band-stop or notch filter with a notch frequency of 50 or 60 Hz (depending on the device) [18], [19]. EMG noise is a high-frequency noise due to muscle contractions and the subject's movement. This noise can be reduced by using moving average [18].

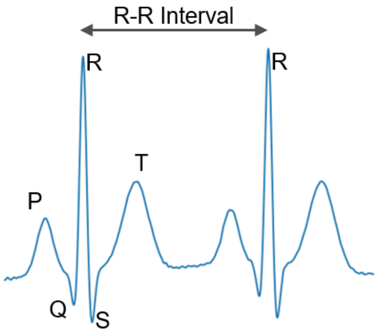

Fig. 1. An example ECG signal with P-wave, QRS complex, and T-wave.

As illustrated in Figure 1, a beat of ECG signal consists of P-wave, QRS complex, and T-wave. For stress recognition, we are mostly interested in the QRS complex. Elgendi et al. [20] propose a frequency band of $8 - 20\ Hz$ for the best signal-to-noise ratio on QRS components. We apply a second-order Butterworth band-pass filter with the proposed frequency band. We note that this filter removes most of the noises described above as their frequencies are outside the chosen band.

The next steps in pre-processing are choosing input lengths and normalization. These steps differ depending on whether we use the filtered ECG signal as input or perform HRV feature extraction. Many studies have demonstrated that deep learning stress models can achieve good performance even on ultra-short-term ECG signals [6], [8], [11], [15]. Deep ECGNet and ECG Emotion models were designed and validated on 10 s segments of ECG signals [6], [8]. On the other hand, studies typically use 60 s of ECG data to extract reliable HRV features [1], [7], [9]. We use 10 s long filtered ECG data (without overlap) as input to the deep learning models. For HRV feature extraction, we use 60 s windows of data with 50 s overlap. We use this overlap to balance the number of training samples available for all the models.

Since the ECG devices used in the two datasets are different, the values would be recorded on different scales. Moreover, the individual stress responses could be different for different participants [6]. To circumvent these issues, we perform a user-specific Min-Max normalization. However, normalization will not eliminate the impact of using different sensor devices for recording ECG. For deep learning models, we perform normalization on the filtered ECG data. Whereas, for HRV features, we perform normalization for every feature. In real-time stress recognition applications, the entire data would not be available for normalization. Hence, we adapt the approach from [21] and use 5 minutes of baseline data to compute the normalization parameters (i.e. min and max values).

E. HRV features

To calculate HRV, we first have to find the peaks in the ECG signal. We use the algorithm proposed in [20] for finding the peaks, i.e., maximum amplitude in the QRS complex (see Figure 1). The algorithm utilizes the knowledge that for healthy adults (1) a beat has only one QRS complex and (2) the duration of QRS is $80 - 120\ ms$. We compute the interval between successive R-R peaks of an ECG signal to obtain the HRV signal. We calculate a total of 22 known HRV features from the time domain, frequency domain, and poincaré plots [1], [7], [9], [10], [22]. These features are listed in Table I along with their descriptions. We use NeuroKit2 [23] python library for computing these features.

TABLE I
LIST OF EXTRACTED HRV FEATURES

Feature	**Description**
HR	Number of R peaks in 1 minute
MeanNN	Mean of R-R intervals
MedianNN	Median of R-R intervals
MadNN	Median Absolute Deviation of R-R intervals
StdNN	Standard deviation of R-R intervals
CVNN	Ratio of StdNN to MeanNN
IQRNN	Inter-Quartile Range of R-R intervals
RMSSD	Root Mean Square of successive differences of R-R intervals
StdSD	Standard deviation of successive differences of R-R intervals
pNN50	% of successive differences of R-R intervals > 50 ms
pNN20	% of successive differences of R-R intervals > 20 ms
TINN	Triangular Interpolation of R-R intervals histogram
HTI	HRV Triangular Index
LF	Power of low frequency band $(0.04\ Hz - 0.15\ Hz)$ in HRV spectrum
HF	Power of high frequency band $(0.15\ Hz - 0.4\ Hz)$ in HRV spectrum
LF/HF	Ratio of LF to HF power
LFn	Normalized low frequency power, LF/total power
HFn	Normalized high frequency power, HF/total power
SD1	Spread of HRV on Poincaré plot perpendicular to the identity line
SD2	Spread of HRV on Poincaré plot along the identity line
SD1/SD2	Ratio of SD1 to SD2
S	Area of ellipse formed in the HRV Poincaré plot

IV. RESULTS AND DISCUSSIONS

In this section, we present the results of our evaluations. First, we evaluate the models using LOSO validation. We also compare their performance with other works on the same dataset. The results of LOSO evaluation on WESAD and SWELL-KW datasets are tabulated in Tables II and III, respectively.

As we expected, the deep learning models perform better than the other models in within-dataset evaluations. On the WESAD dataset, the performance difference is relatively small

TABLE II
LOSO EVALUATION OF ECG-BASED STRESS MODELS ON WESAD DATASET

Model	F1-score	Accuracy
LDA [1]	0.813	0.854
LDA [24]	-	0.887
CNN (Spectrogram) [25]	0.794	0.824
Transformer [5] (without fine-tuning)	0.697	0.804
Our RFC	0.813	0.863
Our SVM	0.832	0.871
Our MLP	**0.859**	0.895
Our ECG Emotion model	0.858	0.897
Our Deep ECGNet	0.857	**0.908**

TABLE III
LOSO EVALUATION OF ECG-BASED STRESS MODELS ON SWELL-KW DATASET

Model	F1-score	Accuracy
SVM [4]	-	0.589
Transformer [5] (without fine-tuning)	0.588	0.581
Our RFC	0.644	0.670
Our SVM	0.609	0.639
Our MLP	0.668	0.689
Our ECG Emotion model	0.627	0.709
Our Deep ECGNet	**0.688**	**0.755**

($< 5\%$), whereas it is higher on the SWELL-KW dataset. The authors of [11] had a similar observation between machine learning models and a CNN-LSTM model, both trained on a stress dataset they acquired. We also note that, among the models based on HRV features, the MLP model performs the best. This is in line with [2], where a simple feed-forward network is shown to perform better than machine learning methods (e.g., SVM, RFC) in a multimodal stress recognition task. Considering both F1-score and Accuracy, Deep ECGNet has the overall best performance in both WESAD and SWELL-KW within-dataset LOSO evaluations.

Next, we evaluate our models using cross-database validation. That is, models trained on the WESAD dataset are tested on SWELL-KW data and vice versa. The results of cross-dataset evaluations are presented in Tables IV and V.

TABLE IV
CROSS-DATASET EVALUATION OF WESAD MODELS ON SWELL-KW DATASET

Model	F1-score	Accuracy
Our RFC	0.467	0.483
Our SVM	**0.535**	**0.538**
Our MLP	0.478	0.49
Our ECG Emotion model	0.395	0.411
Our Deep ECGNet	0.391	0.418

The results of cross-dataset evaluations are the opposite of within-dataset evaluations. The deep learning models perform much worse than models trained on the HRV features. In cross-dataset validation of WESAD models, SVM achieves the best F1-score and Accuracy. Among SWELL-KW models, the RFC model has the overall best performance in cross-dataset evaluation, considering both F1-score and Accuracy.

TABLE V
CROSS-DATASET EVALUATION OF SWELL-KW MODELS ON WESAD DATASET

Model	F1-score	Accuracy
Our RFC	**0.581**	0.637
Our SVM	0.509	**0.647**
Our MLP	0.49	0.621
Our ECG Emotion model	0.342	0.385
Our Deep ECGNet	0.392	0.415

From Tables IV and V, it is clear that HRV-based models outperform deep learning models in predicting stress from a different dataset. This could be attributed to deep learning models learning dataset-specific features and not generic stress features. We note that stressors in the two datasets are different and thus, the stress responses may be different. Additionally, the sensors used for collecting ECG data are also different. All these factors could influence the low generalization capabilities of the deep learning models. More focused studies and datasets are required to improve the generalizability of the deep learning models. On the other hand, HRV features have been studied thoroughly and validated as indicators of stress across multiple datasets with different stressors. Moreover, HRV is computed based on the QRS peak position and thus, is not influenced by the difference in sensors.

Based on our observations, we suggest employing HRV models when the application scenario is different from the dataset. The deep learning models perform better than HRV models on both WESAD and SWELL-KW within-dataset evaluations. So, deep learning models are preferred when the input to the model is similar to its training data.

Finally, we investigate if combining the stress datasets lead to better stress recognition. Combining the WESAD and SWELL-KW datasets results in ECG data of 37 participants. We train and evaluate our models using the data from the combined dataset using LOSO validation.

The results of LOSO validation on the combined dataset is shown are Table VI. The F1-score and Accuracy of every model are significantly worse than the corresponding WESAD models (see Table II). All the models achieve slightly lower F1-score and Accuracy than the corresponding SWELL models (see Table III). Despite the increase in training data, combining these two datasets does not improve the individual dataset or overall stress recognition. So, combining the datasets is not beneficial for either of the datasets; even detrimental in the case of the WESAD dataset.

V. CONCLUSION

Due to the health benefits of detecting and mitigating stress early, there is a need for accurate and robust stress recognition models. The stressor and intensity of stress experienced by people are different for different stressful conditions. This coupled with the ethical challenges of collecting stress data, especially for special groups like people with autism, escalates the need for stress models with good generalization capabilities. Using two publicly available stress datasets (WESAD and

TABLE VI
LOSO EVALUATION OF MODELS ON COMBINED WESAD AND SWELL-KW DATASETS

Model	WESAD subjects		SWELL-KW subjects		All subjects	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
Our RFC	0.758	0.793	0.647	0.671	0.692	0.720
Our SVM	0.732	0.796	0.605	0.633	0.657	0.699
Our MLP	**0.758**	**0.813**	0.657	0.677	**0.698**	**0.732**
Our ECG Emotion model	0.609	0.677	0.593	0.683	0.599	0.681
Our Deep ECGNet	0.692	0.711	**0.695**	**0.739**	0.694	0.728

SWELL-KW), we assessed the generalizability of five stress recognition models - two ECG-based deep learning models and three HRV feature-based models. We first evaluated the models using within-dataset LOSO validation, followed by a cross-dataset evaluation. We found that ECG-based deep learning models outperform the HRV-based models on both stress datasets. However, HRV-based models were significantly better at recognizing stress in cross-dataset evaluations. So, the HRV-based stress recognition models seem to be the better option when the model is deployed in a scenario that is considerably different than the training data acquisition. We also investigate if the stress recognition improves when the models are trained on a combined dataset. The stress recognition on SWELL-KW subjects did not improve by combining datasets. On the other hand, this led to significantly lower performance of all five models on the WESAD dataset.

The datasets we considered in this paper differ in many aspects including the type of stressor, the intensity of stress experienced, and the brand of the ECG sensor. In the future, we plan to extend our work by considering more datasets and comparing them by controlling some of the aspects (e.g. sensor device). This will help us gain insights into the impact of specific factors on the generalizability of stress models.

REFERENCES

[1] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing wesad, a multimodal dataset for wearable stress and affect detection,” in *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 400–408, 2018.

[2] P. Bobade and M. Vani, “Stress detection with machine learning and deep learning using multimodal physiological data,” in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 51–57, IEEE, 2020.

[3] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, “The swell knowledge work dataset for stress and user modeling research,” in *Proceedings of the 16th international conference on multimodal interaction*, pp. 291–298, 2014.

[4] S. Koldijk, M. A. Neerincx, and W. Kraaij, “Detecting work stress in offices by combining unobtrusive sensors,” *IEEE Transactions on affective computing*, vol. 9, no. 2, pp. 227–239, 2016.

[5] B. Behinaein, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad, “A transformer architecture for stress detection from ecg,” in *2021 International Symposium on Wearable Computers*, pp. 132–134, 2021.

[6] P. Sarkar and A. Etemad, “Self-supervised ecg representation learning for emotion recognition,” *IEEE Transactions on Affective Computing*, 2020.

[7] S. Sriramprakash, V. D. Prasanna, and O. R. Murthy, “Stress detection in working people,” *Procedia computer science*, vol. 115, pp. 359–366, 2017.

[8] B. Hwang, J. You, T. Vaessen, I. Myin-Germeys, C. Park, and B.-T. Zhang, “Deep ecgnet: An optimal deep learning framework for monitoring mental stress using ultra short-term ecg signals,” *TELEMEDICINE and e-HEALTH*, vol. 24, no. 10, pp. 753–772, 2018.

[9] T. Pham, Z. J. Lau, S. Chen, and D. Makowski, “Heart rate variability in psychology: A review of hrv indices and an analysis tutorial,” *Sensors*, vol. 21, no. 12, p. 3998, 2021.

[10] F. Shaffer and J. P. Ginsberg, “An overview of heart rate variability metrics and norms,” *Frontiers in public health*, p. 258, 2017.

[11] P. Zhang, F. Li, R. Zhao, R. Zhou, L. Du, Z. Zhao, X. Chen, and Z. Fang, “Real-time psychological stress detection according to ecg using deep learning,” *Applied Sciences*, vol. 11, no. 9, p. 3838, 2021.

[12] R. Arora, M. L. Nicora, P. Prajod, D. Panzeri, E. André, P. Gebhard, and M. Malosio, “Employing socially interactive agents for robotic neurorehabilitation training,” *arXiv preprint arXiv:2206.01587*, 2022.

[13] B. A. Corbett, R. A. Muscatello, and C. Baldinger, “Comparing stress and arousal systems in response to different social contexts in children with asd,” *Biological psychology*, vol. 140, pp. 119–130, 2019.

[14] M. Dzieżyc, M. Gjoreski, P. Kazienko, S. Saganowski, and M. Gams, “Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data,” *Sensors*, vol. 20, no. 22, p. 6535, 2020.

[15] H.-M. Cho, H. Park, S.-Y. Dong, and I. Youn, “Ambulatory and laboratory stress detection based on raw electrocardiogram signals using a convolutional neural network,” *Sensors*, vol. 19, no. 20, p. 4408, 2019.

[16] A. Liapis, E. Faliagka, C. Katsanos, C. Antonopoulos, and N. Voros, “Detection of subtle stress episodes during ux evaluation: Assessing the performance of the wesad bio-signals dataset,” in *IFIP Conference on Human-Computer Interaction*, pp. 238–247, Springer, 2021.

[17] A. Baird, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, L. Stappen, J. Konzok, S. Sturmbauer, E.-M. Meßner, B. M. Kudielka, *et al.*, “An evaluation of speech-based recognition of emotional and physiological markers of stress,” 2021.

[18] R. Kher, “Signal processing techniques for removing noise from ecg signals,” *J. Biomed. Eng. Res*, vol. 3, no. 101, pp. 1–9, 2019.

[19] H. Limaye and V. Deshmukh, “Ecg noise sources and various noise removal techniques: A survey,” *International Journal of Application or Innovation in Engineering & Management*, vol. 5, no. 2, pp. 86–92, 2016.

[20] M. Elgendi, M. Jonkman, and F. De Boer, “Frequency bands effects on qrs detection.,” *Biosignals*, vol. 2003, p. 2002, 2010.

[21] T. Luong, N. Martin, A. Raison, F. Argelaguet, J.-M. Diverrez, and A. Lécuyer, “Towards real-time recognition of users mental workload using integrated physiological sensors into a vr hmd,” in *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 425–437, IEEE, 2020.

[22] S. Betti, R. M. Lova, E. Rovini, G. Acerbi, L. Santarelli, M. Cabiati, S. Del Ry, and F. Cavallo, “Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments by using biological markers,” *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 8, pp. 1748–1758, 2017.

[23] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. Chen, “Neurokit2: A python toolbox for neurophysiological signal processing,” *Behavior research methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[24] A. Karan and A. Kaygun, “Time series classification via topological data analysis,” *Expert Systems with Applications*, vol. 183, p. 115326, 2021.

[25] L. Liakopoulos, N. Stagakis, E. I. Zacharaki, and K. Moustakas, “Cnn-based stress and emotion recognition in ambulatory settings,” in *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1–8, IEEE, 2021.