DATA AUGMENTATIONS AND TRANSFER LEARNING FOR PHYSIOLOGICAL TIME SERIES

Harald Vilhelm Skat-Rørdam* & Mia Hang Knudsen* & Simon Nørby Knudsen* & Sneha Das & Line Clemmensen
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Lyngby, Denmark
{s175393,s183998,s174479,Sned,lkhc}@dtu.dk

ABSTRACT

Physiological time-series signals e.g., measured through wearables have received increasing interest as biomarkers for sleep disorders, stress, anxiety, and other psychiatric disorders, or health conditions. However, open source datasets are scarce making it difficult to develop strong prediction models for new application areas without extensive prior data collection. We investigate the possibilities of using existing datasets as well as different simulation strategies to create a foundational model transferable to new applications. We evaluate transferability for four different tasks (open source data) and compare the performance of transfer learning and simulated data augmentations.

1 INTRODUCTION

Previous studies have shown that physiological signals are affected by stress, of which heart rate (HR) is the most prominent, but blood volume pressure (BVP), electrodermal activity (EDA), and to some extent temperature (TEMP) have also been related to stress (Giannakakis et al., 2022). In other domains, like computer vision, it is well known that transfer learning and augmented or simulated data in addition have the potential to improve model performance (Wang et al., 2019). (Tremblay et al., 2018) found that using only synthetic training data created by domain randomization and thereafter fine-tuning with real-life images performed better results than a model trained only on real-world data. In addition, recent studies (Minor et al., 2020; Alkhalifah et al., 2022) found that using raw simulated data without any augmentation had significantly worse performance compared to using an augmented version.

State-of-the-art models for classifying and labeling physiological data consist of several deep-learning as well as classic machine learning models. As we are interested in transfer learning, we focus on deep learning architectures. Hu et al. (2022) proposed a well performing deep learning model that combines a convolutional neural network (CNN), a transformer-based neural network, and a feedforward neural network (FNN) to classify arrhythmias based on ECG data. Deznabi & Fiterau (2023) created a framework, MultiWave, that can handle signals sampled at different frequencies, a common issue when dealing with multivariate time series data, and also E4 data. Their work uses wavelets to decompose signals into sub-signals handled by separate components of the model, where each output is combined in a gating mechanism. With this technique, they were able to obtain 'top performance in stress and affect detection from wearables'(Deznabi & Fiterau, 2023). A recent study (Theunissen et al., 2021) successfully applied both an autoencoder and a convolutional autoencoder to predict furnace blowback events from multivariate time series data. Their results seem promising for building an autoencoder, which can be used for event prediction for time series data.

Multiple studies have seen a gain in performance by applying pre-trained models for classifying time series data across multiple domains Zhang et al. (2022); McDermott et al. (2021); Yeh et al. (2023). In one study Zhang et al. (2022) an encoder is pre-trained using self-supervised learning, and thereafter fine-tuned to many different independent classification tasks. Using the pre-trained

model they achieve an average increase of 15.4% in F1-score compared to state-of-the-art baseline methods.

In this work, we investigate the *possibilities* of developing a foundational model, increasing performance, by using simulations for augmentations as well as transfer learning. We propose a CNN-based autoencoder to make a foundational model for physiological time series signals with event prediction and evaluate the event predictions on four different tasks (datasets): Stress under alcohol use disorder, stress while driving, stress while puzzling, and stress during speech.

2 METHODS

2.1 DATASETS

We use four open source datasets, which are ADARP (Sah et al., 2022), AffectiveROAD (Haouij et al., 2018), EmoPairCompete (this dataset is anonymized), and WESAD (Schmidt et al., 2018). The four datasets are selected, because they are physiological datasets with a stress factor, but each with a different stressor. This adds variability and expands the use of the model to a larger domain within stress prediction. Furthermore, for simplicity of the task and continuity, the four datasets used are all collected with the Empatica E4 wristband Empatica (2023b), which is described in appendix A.

ADARP: A Multi Modal Dataset for Stress and Alcohol Relapse Quantification in Real Life Setting: In (Sah et al., 2022) a proof-of-concept study was conducted at Washington State University, to examine, whether a wearable sensor can be used to detect stress in patients suffering from alcohol use disorder (AUD). 11 participants (10 women) were recruited and all received mental health and AUD treatment at a treatment agency in the state of Washington. Participants were on average involved in the study for 14 days. During the initial session, the participants were given an E4 wristband and told to wear it during the day and only take it off when sleeping or at times when the device could be damaged. The participants were instructed to press the event button when they felt "more stressed, overwhelmed, or anxious than usual" (Sah et al., 2022). In total 1698 hours of physiological data were collected using E4, with an average of 11.5 hours each day and the participants tagged 409 events in total.

AffectiveROAD Dataset (ROAD): The AffectiveROAD dataset (Haouij et al., 2018) is from an experiment with 14 car drives, from 10 different participants. All participants drove the same route on three different types of roads. Each participant wore two Empatica E4 wristbands, one on each wrist, together with various other sensors, which are not relevant in this study. We limit our use to the E4 wristband on the left wrist since for most experiments the participants were asked to wear the wristband on their nondominant hand. Over the full drive, the driver has been exposed to different stressful driving situations. Additionally, the drivers were observed and subjectively evaluated on a stress metric between 0 and 1, where 0 was not stressed, and 1 was extremely stressed. These scores were validated by the driver after the session. We used two classes; with 0.75 as our threshold for stress, resulting in only classifying high stress periods, inspired by Bustos et al. (2021).

EmoPairCompete[1]: The data is collected through an experiment designed to study emotion and frustration through prosocial and competitive behaviors. The dataset is comprised of 28 participants in a semi-controlled stress-inducing task as well as rest periods, each of 5 minutes. We will classify signals into these two settings as stress or no-stress. During the experiment, all participants wore an Empatica E4 wristband on their nondominant hand.

WESAD: Multimodal Dataset for Wearable Stress and Affect Detection: WESAD (Schmidt et al., 2018) is another dataset that combines physiological data with stress exposure. This dataset contains data from 15 participants. Each participant wore, among other sensors, an Empatica E4 on their non-dominant hand. The data was collected over two hours, in which the participants went through different scenarios. In the stressful scenario, they were exposed to the Trier Social Stress Test (TSST) (Kirschbaum et al., 1993). Here, they had to give a 5-minute speech, which the participants were told could boost their career options. Thereafter, they needed to count down from 2023 in steps of 17 and start over if they made a mistake. Data from this situation are labeled as stress. Furthermore, there was a baseline scenario, in which participants were sitting/standing at a

[1]Paper currently submitted to ICLR 2024 workshop TS4H, with paper ID: 47

table and provided with magazines containing neutral content. Additionally, they had to watch a set of funny video clips in the amusing scenarios. The baseline scenario and the amusement scenario are combined into a non-stress category. In Schmidt et al. (2018) they, among other things, performed a binary classification task. With their classification, they obtained an accuracy of around 0.88 and an F1 score of 0.86 using physiological data from the wrist in a random forest model.

2.2 SIMULATION STRATEGIES

To simulate the four physiological signals (HR, BVP, EDA, TEMP) from the E4 wristband, two Python packages were used, Neurokit2 (Makowski et al., 2021) and JOS-3 (Takahashi et al., 2021). Neurokit2 can generate EDA (electrodermal activity), ECG (electrocardiography), and PPG (photoplethysmogram) from which we derived BVP and HR. JOS-3 simulates thermal physiology data and we used it to generate human skin temperatures. It is important to note that each physiological signal is generated independently, and therefore the correlation between the signals is lacking. Furthermore, we did not find documentation regarding age and physical traits which could be used as proxies during simulation, to create some of the correlation structures between the individual physiological signals. For BVP, EDA, and HR we had two different ways of simulating the data (Plain and Fragmented).

Plain: The BVP and HR signals are connected, thus the HR can be calculated based on the BVP value rather than simulating an unrelated HR signal. Using the interbeat interval, we can transform it into beats per minute, subsequently, we will down-sample the data to 1 Hz to mimic the data from the E4 wristband. After extracting the heart rate we scaled the BVP signal such that it has a mean of 0 and a max of a randomly selected value, as seen in Appendix 4. We simulated EDA data with a sampling rate of 1000 Hz. The next step was to scale the data such that it had a maximum value above 0 and a minimum below 20. The final step was to downsample the data to 4Hz using resampling. For simulating TEMP data, we used JOS-3. The package provides a detailed model, which can be used to simulate thermal physiology data of human skin temperature at the wrist.

Fragmented: Simulating 3 seconds signal and varying input parameter for the Neurokit's simulate functions. To create real-world-like PPG data, we chose to first simulate electrocardiogram (ECG) data using ECGSYN described by McSharry et al. (2003) to obtain HR for a 9-second segment. Using these HR segments, we further divided it into 3-second intervals. The HR value in the first second of each segment was used as HR input to simulate 3 seconds of BVP data at 64Hz. This process continued by concatenating each interval together until we had simulated the desired interval. Lastly, we scaled the BVP data in the same way as we did in the plain method. A visualization of the process can be seen in Appendix 4.

2.3 MODEL

We use a semi-supervised CNN-based autoencoder with two frequency inputs, which are concatenated in the embedded latent layer. The model has two heads, one using a reconstruction loss (unsupervised) calculated with MSELoss, and the other a classification loss (supervised) calculated using BCEWithLogitsLoss, see Fig. 1. When pre-training and transferring a model, we use three of the datasets combined with either plain, fragmented, or no-simulated data. When we do not use a transfer model we only train using the target data and one of the three simulation strategies (Fig. 2 in Appendix). Code availability GitHub.

3 RESULTS

The results are summarised in Table 1 and 2. The tasks in ADARP, EmoPairCompete, and to some extent ROAD have better predictions when using 5-minute windows for prediction than 1-minute windows, whereas this seems to be reversed for WESAD. Even so, the results for WESAD are significantly improved when using transfer learning with a model pre-trained on the other three datasets.

In general, there is a performance improvement when using transfer learning, even when the tasks in the datasets are not directly related and the optimal time windowing differs for the tasks. The frag-

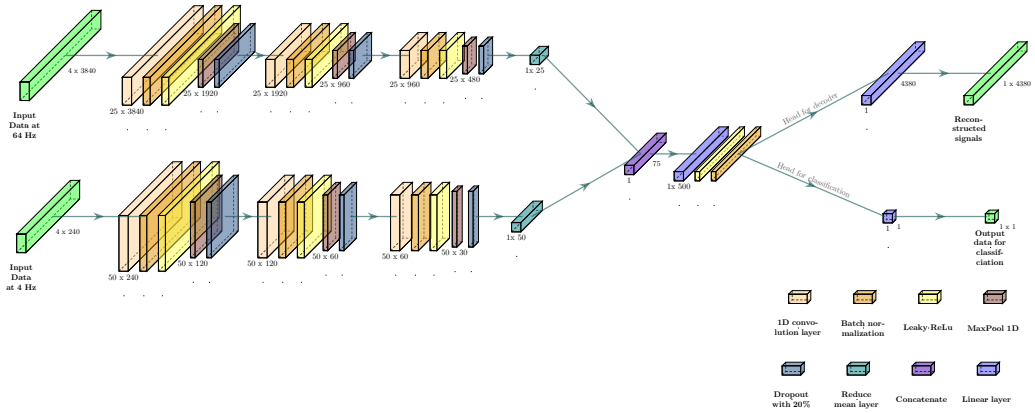

Figure 1: Model architecture for 1-minute data. The model takes input in both 4Hz and 64Hz, and each frequency has 3 convolutional layers followed by batch normalization, Leaky ReLu, max pooling, and dropout. Thereafter by concatenating the two frequency embeddings, we obtain our bottleneck of 75 nodes. Finally, the four reconstructed signals are outputted stacked together.

Transfer	Simulation	ADARP	EmoPairCompete	ROAD	WESAD
No	No	0.26 (0.10)	**0.53 (0.05)**	0.55 (0.08)	0.89 (0.04)
No	Plain	0.21 (0.06)	0.46 (0.07)	0.56 (0.09)	0.89 (0.04)
No	Fragmented	0.21 (0.07)	0.50 (0.05)	0.50 (0.17)	0.84 (0.06)
Yes	No	0.25 (0.00)	0.48 (0.07)	**0.67 (0.07)**	0.95 (0.01)
Yes	Plain	**0.28 (0.09)**	0.51 (0.07)	0.52 (0.09)	0.95 (0.01)
Yes	Fragmented	0.22 (0.06)	0.52 (0.03)	0.63 (0.03)	**0.96 (0.01)**

Table 1: Average F1-score of 10 runs. Standard deviation in parentheses. For 1-minute windows.

Transfer	Simulation	ADARP	EmoPairCompete	ROAD	WESAD
No	No	0.39 (0.15)	0.54 (0.13)	0.75 (0.16)	0.86 (0.10)
No	Plain	0.56 (0.13)	0.49 (0.22)	0.75 (0.08)	0.88 (0.06)
No	Fragmented	**0.63 (0.10)**	0.56 (0.21)	0.65 (0.18)	0.81 (0.10)
Yes	No	0.52 (0.09)	0.60 (0.04)	**0.79 (0.05)**	0.74 (0.15)
Yes	Plain	0.48 (0.18)	0.54 (0.07)	0.70 (0.04)	0.84 (0.04)
Yes	Fragmented	0.59 (0.17)	**0.73 (0.05)**	0.72 (0.06)	**0.93 (0.12)**

Table 2: Average F1-score of 10 runs. Standard deviation in parentheses. For 5-minute windows.

mented simulation strategy outperforms the plain strategy under transfer learning but is questionable or at least varying in performance when transfer learning is not included.

4 Discussion and conclusion

In this work, we explored the potential of augmentation and transfer learning strategies toward developing a foundation model for physiological time-series data. As presented in the results, in all instances, except for EmoPairCompete 1-minute, the predictions are improved by including one or both of the simulated data and transfer learning methods. Additionally, similar to the results provided in (Minor et al., 2020; Alkhalifah et al., 2022), we find that using the fragmented simulation method generally outperforms the plain simulation method, as it augments the data to a greater extent. However, the general performance is not sufficiently high, and we observe considerable variability both in conclusions and between runs, indicating a need for continued research into strategies and methods for models robust to domain shifts and in-the-wild applications.

AUTHOR CONTRIBUTIONS

*These authors contributed equally.

REFERENCES

Empatica legal & compliance, 2024. URL https://www.empatica.com/legal. Accessed: 02-01-2024.

Tariq Alkhalifah, Hanchen Wang, and Oleg Ovcharenko. Mlreal: Bridging the gap between training on synthetic data and real data applications in machine learning. *Artificial Intelligence in Geosciences*, 3:101–114, December 2022. ISSN 26665441. doi: 10.1016/j.aiig.2022.09.002.

Wolfram Boucsein. *Electrodermal Activity*. Springer New York, NY, 2012. ISBN 978-1-4614-1125-3. doi: 10.1007/978-1-4614-1126-0. URL https://doi-org.proxy.findit.cvt.dk/10.1007/978-1-4614-1126-0.

Cristina Bustos, Neska Elhaouij, Albert Sole-Ribalta, Javier Borge-Holthoefer, Agata Lapedriza, and Rosalind Picard. Predicting driver self-reported stress by analyzing the road scene. (arXiv:2109.13225), 9 2021. URL http://arxiv.org/abs/2109.13225. arXiv:2109.13225 [cs].

Iman Deznabi and Madalina Fiterau. Multiwave: Multiresolution deep architectures through wavelet decomposition for multivariate time series prediction. (arXiv:2306.10164), June 2023. URL http://arxiv.org/abs/2306.10164. arXiv:2306.10164 [cs, eess].

Empatica. E4 data - bvp expected signal, 2023a. URL https://support.empatica.com/hc/en-us/articles/360029719792-E4-data-BVP-expected-signal. Accessed: 2023-10-02.

Empatica. E4 wristband, 2023b. URL https://e4.empatica.com/e4-wristband. Accessed: 2023-11-21.

M. Garbarino, M. Lai, D. Bender, R.W. Picard, and S. Tognetti. Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 EAI 4th International Conference on Wireless Mobile Communication and Healthcare (Mobihealth)*, pp. 39–42, November 2014. doi: 10.1109/MOBIHEALTH.2014.7015904.

Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460, 1 2022. ISSN 1949-3045. doi: 10.1109/TAFFC.2019.2927337.

Medical Device Coordination Group. MDCG 2021-24 guidance on classification of medical devices. URL https://health.ec.europa.eu/system/files/2021-10/mdcg_2021-24_en_0.pdf.

Neska El Haouij, Jean-Michel Poggi, Sylvie Sevestre-Ghalila, Raja Ghozi, and Mériem Jaïdane. Affectiveroad system and database to assess driver's attention. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, pp. 800–803, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351911. doi: 10.1145/3167132.3167395. URL https://doi.org/10.1145/3167132.3167395.

Rui Hu, Jie Chen, and Li Zhou. A transformer-based deep neural network for arrhythmia detection using continuous ecg signals. *Computers in Biology and Medicine*, 144:105325, 2022. doi: 10.1016/j.compbiomed.2022.105325. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482522001172.

Clemens Kirschbaum, Karl-Martin Pirke, and Dirk Hellhammer. The 'trier social stress test' – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28:76–81, 02 1993. doi: 10.1159/000119004.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, 2 2021. doi: 10.3758/s13428-020-01516-y. URL https://doi.org/10.3758%2Fs13428-020-01516-y.

Matthew McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. A comprehensive ehr timeseries pre-training benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, pp. 257–278, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383592. doi: 10.1145/3450439.3451877. URL https://doi.org/10.1145/3450439.3451877.

P.E. McSharry, G.D. Clifford, L. Tarassenko, and L.A. Smith. A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*, 50(3):289–294, 2003. doi: 10.1109/TBME.2003.808805.

Eric N. Minor, Stian D. Howard, Adam A. S. Green, Matthew A. Glaser, Cheol S. Park, and Noel A. Clark. End-to-end machine learning for experimental physics: using simulated data to train a neural network for object detection in video microscopy. *Soft Matter*, 16(7):1751–1759, 2020. ISSN 1744-683X, 1744-6848. doi: 10.1039/C9SM01979K.

P. Rautaharju and F. Rautaharju. *Heart Rate and Heart Rate Variability*, pp. 45–67. Springer London, London, 2007. ISBN 978-1-84628-481-6. doi: 10.1007/978-1-84628-481-6_4. URL https://doi.org/10.1007/978-1-84628-481-6_4.

Ramesh Kumar Sah, Michael McDonell, Patricia Pendry, Sara Parent, Hassan Ghasemzadeh, and Michael J. Cleveland. Adarp: A multi modal dataset for stress and alcohol relapse quantification in real life setting. (arXiv:2206.14568), June 2022. URL http://arxiv.org/abs/2206.14568. arXiv:2206.14568 [cs, eess].

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pp. 400–408, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356923. doi: 10.1145/3242969.3242985. URL https://doi.org/10.1145/3242969.3242985.

Yoshito Takahashi, Akihisa Nomoto, Shu Yoda, Ryo Hisayama, Masayuki Ogata, Yoshiichi Ozeki, and Shin ichi Tanabe. Thermoregulation model jos-3 with new open source code. *Energy and Buildings*, 231:110575, 2021. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2020.110575. URL https://www.sciencedirect.com/science/article/pii/S0378778820333612.

Carl Daniel Theunissen, Steven Martin Bradshaw, Lidia Auret, and Tobias Muller Louw. One-dimensional convolutional auto-encoder for predicting furnace blowback events from multivariate time series process data—a case study. *Minerals*, 11(10), 2021. ISSN 2075-163X. doi: 10.3390/min11101106. URL https://www.mdpi.com/2075-163X/11/10/1106.

Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1082–10828, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6100-0. doi: 10.1109/CVPRW.2018.00143. URL https://ieeexplore.ieee.org/document/8575297/.

Christiaan H. Vinkers, Renske Penning, Juliane Hellhammer, Joris C. Verster, John H. G. M. Klaessens, Berend Olivier, and Cor J. Kalkman. The effect of stress on core and peripheral body temperature in humans. *Stress*, 16(5):520–530, 2013. doi: 10.3109/10253890.2013.807243. URL https://doi.org/10.3109/10253890.2013.807243. PMID: 23790072.

Fei Wang, Hao Wang, Haichao Wang, Guowei Li, and Guohai Situ. Learning from simulation: An end-to-end deep-learning approach for computational ghost imaging. *Optics Express*, 27(18):25560, September 2019. ISSN 1094-4087. doi: 10.1364/OE.27.025560.

Chin-Chia Michael Yeh, Xin Dai, Huiyuan Chen, Yan Zheng, Yujie Fan, Audrey Der, Vivian Lai, Zhongfang Zhuang, Junpeng Wang, Liang Wang, and Wei Zhang. Toward a foundation model for time series data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pp. 4400–4404, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3615155. URL https://doi.org/10.1145/3583780.3615155.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency, 2022.

A EMPATICA E4

To ensure the data is ecologically valid, we need a method to capture data while participants are unconstrained by traditional laboratory settings. The benefit of collecting data in the wild is that it offers the opportunity to capture data from individuals in their natural environment and provides a more true-to-life basis. Therefore, wearable sensors such as wristbands and smartwatches present great opportunities for the non-invasive and convenient acquisition of multimodal signals. There exists a vast range of different brands and models of portable devices to collect signals with varying precision. Empatica's E4 wristband (Empatica, 2023b) (Garbarino et al., 2014) is classified as an IIa medical device (Group) in the European Union and approved by The US Food and Drug Administration (FDA)(Emp, 2024). This supports the E4 wristband as a non-invasive medical device to be able to collect data in naturalistic settings. For this thesis, we will only examine the data collected from an E4 wristband. The wristband has four different sensors:

1. A photoplethysmography (PPG) sensor to measure blood volume pulse (BVP), from where heart rate (HR) among others can be derived.
2. An electrodermal activity sensor to measure electrodermal activity (EDA).
3. An infrared thermopile sensor to measure skin temperature (TEMP).
4. A 3-axis accelerometer sensor, to capture movements of the wristband (ACC).

Three signals are measured with the PPG sensor: BVP, HR, and inter-beat interval (IBI). Both HR and IBI are calculated from the BVP signal. Heart rate (HR) is extracted by peak detection on the BVP signal and then smoothing the signal with a window of 10 seconds according to the website of Empatica[2]. IBI is also derived based on BVP; however, with movement above 30% of the time, the variability of heart rate cannot be reliably calculated from the IBI signal. Due to the design and uncontrolled aspect of in-the-wild recording, IBI and heart rate variability will not be used in this thesis.

In addition to the four sensors, the wristband includes an event mark button to tag events based on the need and an internal real-time clock that timestamps the data, including a button press.

A.1 BLOOD VOLUME PULSE (BVP)

BVP is the change in the volume of blood over time and can be detected with a PPG sensor (Empatica, 2023a) and with E4 it is captured at a sample rate of 64Hz. Blood pressure can be described by different measures such as systolic blood pressure (SBP) and diastolic blood pressure (DBP), among others. In Giannakakis et al. (2022) 15 different studies have been examined and all of them have shown a significant increase in SBP and DBP during stress.

A.2 HEART RATE (HR)

HR is the average number of beats in some time interval. The E4 wristband provides this signal at a sampling rate of 1Hz, which is frequently referred to as beats per minute (BPM) (Rautaharju & Rautaharju, 2007). According to Giannakakis et al. (2022) and their review, HR is the most notable feature that significantly increases during stress. Here 23 different studies are examined, where 18 report a significant increase during stress, and the last 5 report no significant difference in HR during stress.

[2]https://support.empatica.com/hc/en-us/articles/360029469772-E4-data-HR-csv-explanation

A.3 ELECTRODERMAL ACTIVITY (EDA)

EDA is a measure of variation in the electrical properties of the skin. It is a general measure, that can be split into different features, the most commonly used are the tonic (electrodermal level) and phasic (electrodermal response) part (Boucsein, 2012). Multiple features of EDA have been reviewed in Giannakakis et al. (2022), where the studies found a significant and a non-significant increase during stress. Most studies (9) have examined the phasic part of EDA, which are peaks in the signal typically as a response to a stimulus. Of these 9 studies, 7 reported a significant increase during stress, and 2 reported no significant difference. With the E4 wristband, the coupling of the electrodes with the skin can take around 15 minutes and is measured at a sampling rate of 4 Hz.

A.4 SKIN TEMPERATURE (TEMP)

TEMP is the temperature of the skin. They are reported in degrees Celsius (°C). From the E4 wristband, TEMP is measured on either the left or right wrist with a sampling rate of 4Hz. Giannakakis et al. (2022) reviewed studies showing either a significant increase or decrease in skin temperature during stress conditions. These skin temperatures are reported for various locations on the body, but none of them specifically around the wrist. One study specifically looking at wrist TEMP and stress did not find significant differences in TEMP during stress (Vinkers et al., 2013).

Variable	Description	Number of classes
Age	All ages from 5 to 70 years old	66
Weather	Winter, Spring, Summer, Autumn, and indoors	5
Physical condition	Fit, normal, or poor shape	3
Gender	Female or male	2

Table 3: The variables we use to create a grid for simulating data for each unique combination of variables. In total, we have $66 \cdot 5 \cdot 3 \cdot 2 = 1980$ different combinations of conditions to simulate. These variables are realized based on the parameters presented in table 4.

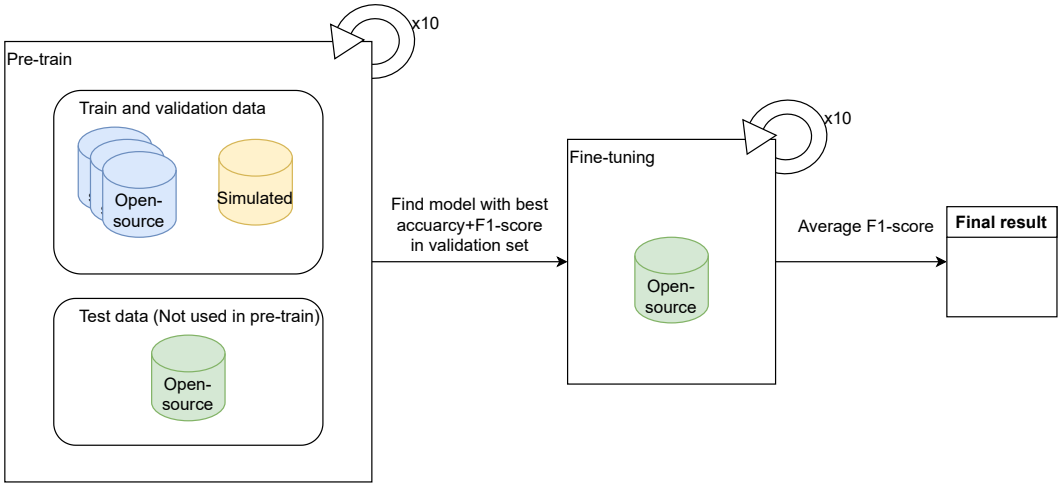

Figure 2: Validation strategy. One of the three simulation strategies is used for augmenting data in the training phase with either none, plain simulation, or fragmented simulation data. Three of the datasets are used for training when transfer learning is deployed. The target dataset is split in 80% for training, and 10% for both validation and testing.

	Parameter	Range/Values	Step size
BVP & HR			
	Standard deviation HR noise	[-1, 1]	0.1
	Physical condition effect on average HR	{good = -5 normal = 0, poor = 5}	
	Physical condition effect on standard deviation HR	{good = 10, normal = 8, poor = 6}	
	Genders effect on HR	{female = +8, male = 0}	
	The average HR for 5 to 9 years old	[70, 115]	1
	The average HR for above 10 years old	[60, 100]	1
	Scale value for BVP	[0.001, 500]	0.001
EDA			
	Drift for EDA	[-0.1, 0.1]	0.001
	SCR peak	True/False with a probability of 0.1 for true	
	EDA amplitude of laplace noise	[0.2, 2]	0.01
TEMP			
	BMI for good physical condition	[16, 27]	0.01
	BMI for normal physical condition	[18.5, 25]	0.01
	BMI for poor physical condition	[15.5, 18.5] $\cup$ [25, 40]	0.01
	Weight	BMI * height* height	
	Height for 5 to 9-year-olds in m.	[1.03, 1.39]	0.01
	Height for 10 to 14-year-olds in m.	[1.29, 1.7]	0.01
	Height for female above 15 year olds in m.	[1.5, 1.9]	0.01
	Height for male above 15 year olds in m.	[1.65, 2.05]	0.01
	Humidity based on weather	{winter = 89, spring = 85, summer = 78, fall = 76, indoor = 45}	
	Temperature for winter	[-18, 12.5]	0.5
	Temperature for spring	[-15, 25]	0.5
	Temperature for summer	[1.5, 32]	0.5
	Temperature for fall	[-6, 24]	0.5
	Temperature for indoor	[15, 28]	0.5
	I_{cl} for winter	[1.10, 1.5]	0.01
	I_{cl} for spring	[0.5, 1.1]	0.01
	I_{cl} for summer	[0.3, 0.8]	0.01
	I_{cl} for fall	[0.5, 1.1]	0.01
	I_{cl} for indoor	[0.35, 0.9]	0.01
	CI based on physical condition	{good = 4.2, normal = 3.5, poor = 2.8}	
	Wind speed indoor	0.1	
	Wind speed not indoor	[0, 25]	0.1
	Posture	[sitting, laying, standing]	
	Activity	[1, 4.4]	0.1

Table 4: The parameters for simulating data using Neurokit2 and JOS-3. The parameters are defined to realize our variables in 3.

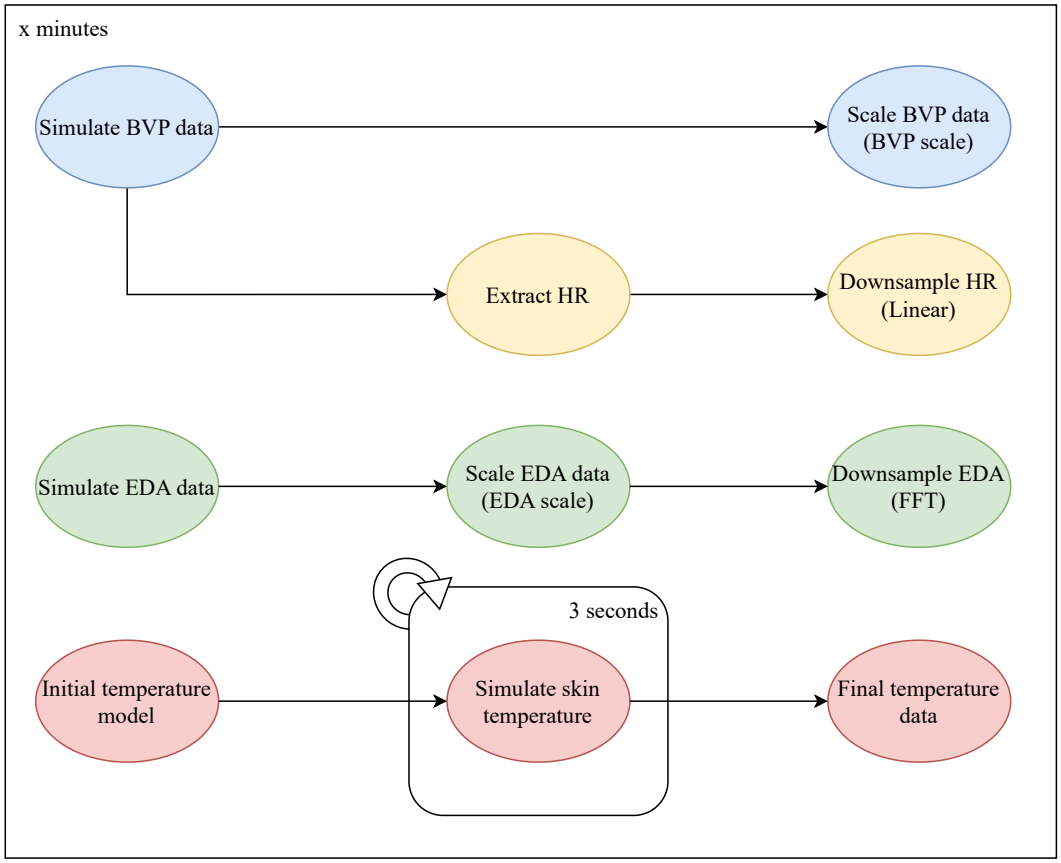

Figure 3: The plain method.

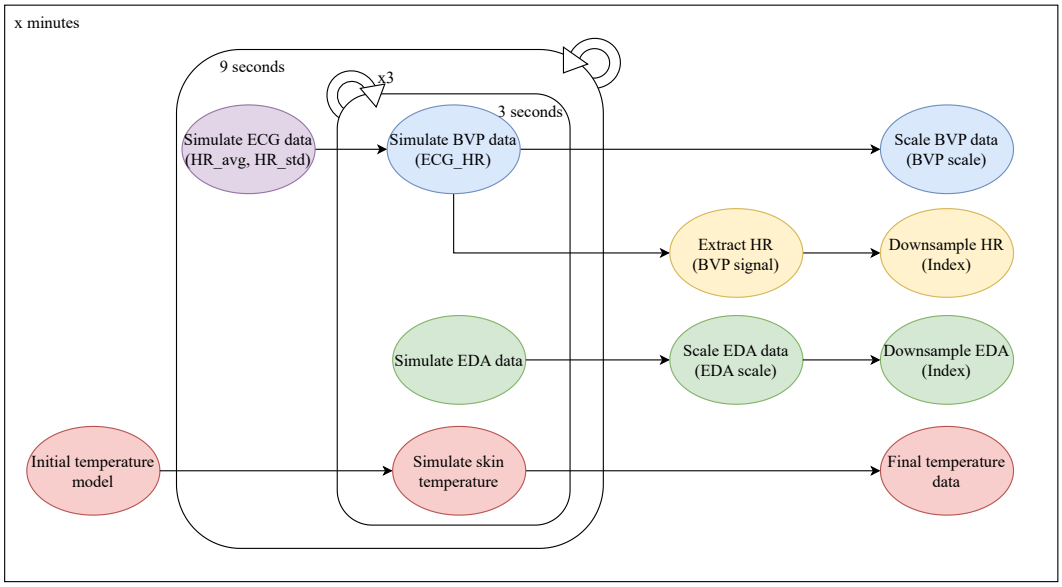

Figure 4: The fragmented method. Process of simulating BVP and HR data for a given interval of x minutes using the fragmented method. First, we simulate ECG signals for every 9 seconds. Then HR for each of the three 3-second intervals is used as input to simulate 3 seconds of BVP signals. When all x minutes of BVP signals are simulated the HR is extracted from BVP signals based on the peaks.