Using Denoising Autoencoder to Improve Photoplethysmographic Signal Quality for Automatic Stress Recognition

William Davies[1]

MSc Machine Learning

Supervisor: Dr Youngjun Cho

Submission date: 25 October 2021

[1] **Disclaimer:** This report is submitted as part requirement for the MSc Machine Learning at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged.

Code for this project can be found at the following repository:
https://github.com/william-davies/msc-project

Abstract

Background: Low cost, easily wearable (e.g. wristbased) photoplethysmography (PPG) sensors are a promising approach for everyday cardiovascular monitoring and mental stress assessment [16]. However, they are susceptible to noise, including motion artifacts, which can limit the usefulness of recorded signals. Furthermore, low cost sensors have low sampling rates, which limit the signal's information content. Unfortunately, current signal filtering methods require domain expertise to specify filter parameters, and can still fail to adequately denoise physiological signals.

Objective: We propose a novel framework using a denoising autoencoder module to denoise PPG signals before performing automatic stress classification.

Methods: We used a dataset of 17 participants participating in mental stress-inducing tasks. The participant's physiological signals and self-reported stress levels were recorded during the tasks and interposed rest periods. We manually annotated the dataset PPG signals into clean/noisy segments and will release the annotations to the research community. We downsampled the PPG signals to 16Hz and then trained a denoising autoencoder (DAE) to denoise PPG signals from an Empatica wristband sensor. We extracted heart rate variability (HRV) metrics from the denoised signal and evaluated their accuracy compared to HRV metrics extracted from ground truth Infiniti fingertip sensor signals. Finally, we used the extracted HRV features in a Gaussian Naive Bayes classifier to estimate stress. We compared the leave-one-subject-out cross-validation (LOSO-CV) performance of our DAE framework with that achieved using traditional filtering methods.

Results: First, we found that our DAE increased the signal quality of both clean (mean pSQI: from 0.786 to 0.985) and noisy (mean pSQI: from 0.536 to 0.987) PPG signals. Second, the DAE denoising resulted in more accurate heart rate (HR), SDNN and RMSSD compared to traditional filtering methods. However, for the pNN50 and LF/HF ratio metrics, traditional filtering methods were slightly better than DAE denoising (although both were worse than the raw Empatica signal). HRV metrics extracted from DAE denoised Empatica signals were also found to improve stress classification performance. Notably, using DAE denoised HRV metrics from the Empatica sensor produced better stress classification performance than the more reliable, higher sampling rate Infiniti sensor (Matthews Correlation Coefficient: 0.415 vs 0.378).

Conclusion: Our results demonstrate that DAE denoising can improve the accuracy of some HRV metrics compared to traditional filtering methods. Our proposed framework using a DAE denoising module also improves stress classification performance beyond traditional filtering. Importantly, DAE denoising of less reliable, low sampling rate, cheaper wristband sensor signals can produce better stress classification performance than using more reliable and expensive fingertip based sensors. As the former is applicable in many more scenarios, our findings are encouraging for mobile stress management systems in natural settings.

Acknowledgements

Thank you to my supervisor Youngjun Cho for being so generous with his time and for his invaluable insight. Thank you also to Jitesh Joshi for his very helpful advice.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The complexities and pressures of modern life have increased the mental stresses people experience. This has adverse effects not only in terms of health, but also productivity. Hence, there is a pressing need for stress management systems. People are often inept at monitoring their own stress levels, which underlines this need [77]. There currently exist systems that require users to complete self-report questionnaires of perceived stress [87]. These systems are limited because they disrupt the user's day and cannot measure stress continuously. Fortunately, stress can be measured automatically from physiological signals [57], which avoids the pitfalls of self-reporting stress. Automatic stress management systems that leverage physiological sensors are thus much more useful. Cardiovascular events have been found to be correlated with mental stress [28]. In particular, HRV is the variation in the time intervals between heartbeats and has been extensively used to measure mental stress levels [72, 47]. PPG is the technology used in most commercial continuous wrist-based blood volume pulse (BVP) monitors nowadays [5] and is the focus of this thesis.

In highly controlled clinical or specialist settings, PPG signals can be recorded with less noise. However, stress management systems that are useful in daily life must use low cost sensors to measure physiological signals in uncontrolled “in the wild” settings. Unfortunately, in such scenarios, measured signals are often noisy [63]. This can be caused by motion artifacts and changing ambient light levels. This noise can obscure the information content of physiological signals and make it difficult to estimate stress-level. In fact, noise particularly affects HRV metrics [103].

To denoise physiological signals, signal processing techniques such as bandpass filters are currently used. However, such methods often fail to accurately denoise the corrupted signal, by performing poorly against stationary noise [59] or suffering from phase shifts for example [32]. Another downside of current methods is they require parameters that depend on the target signal and require subject expertise, such as the passband range of bandpass filters [2].

As an alternative approach, machine learning (ML), and in particular artificial neural networks (ANNs), has been used to denoise physiological signals. Machine learning approaches do not require the specification of target signal specific parameters, as the model can learn the appropriate parameters itself from training data [2]. ML approaches have also been found to denoise better than traditional filtering methods [59, 2]. DAEs have been used to denoise electroencephalogram (EEG) [94, 111], electrocardiogram (ECG) [81, 109, 2, 98], electrodermal activity (EDA) [98], and electromyogram (EMG) [98] signals. With respect to PPG, [91] used an autoencoder to reconstruct BVP signals, [100] applied a DAE to BVP signals as a pretraining task, and [76] trained a convolutional neural network (CNN) to classify PPG signals as noise or not noise. However, to our knowledge, there have only been two applications of using DAEs to denoise PPG signals [59, 83].

[59, 83] both succeeded in denoising PPG. They used metrics such as HR [59] and number of clinically acceptable PPG beats to measure waveform features [83] to evaluate denoising quality. However, these simple metrics do not provide much information on the advantages of DAE for practical tasks such as automatic stress classification. As our novel contribution, we evaluate the effect of DAE denoising of PPG signals on i) accuracy of PRV metrics ii) performance on automatic stress classification. Our evaluation methods provide greater insight into the practical advantages possible with DAEs with respect to automatic stress classification.

We used a dataset of the PPG signals of 20 participants in an experiment designed to induce stress [85]. BVP was recorded with both an Infiniti finger sensor and an Empatica wristband sensor. As wristband sensors are applicable in far more situations than fingertip sensors, this thesis focusses on Empatica PPG signals. As an extension of this dataset, we manually annotated segments of the BVP signals as noisy or clean and provide these annotations to the research community. Low cost wearable sensors, which are easier to use in the wild, tend to have lower sampling rates [35]. As we want to support the use of ubiquitous low cost sensors, we downsampled PPG signals to 16Hz before conducting our experiments. To our knowledge, DAE denoising has never been tested on low sampling rate BVP signals.

Rather than trying to obtain state of the art stress classification performance, this thesis proposes a novel DAE framework to improve PRV metric extraction and stress classification performance. Empatica PPG signals were denoised through traditional filtering methods (the baseline) and by a DAE. The possible advantages of this DAE framework were evaluated in three ways. i) the signal pSQI improvement ii) the accuracy of extracted PRV metrics iii) LOSO-CV performance on stress classification. To our knowledge, the effect of DAE denoising of BVP signals on PRV metric accuracy and (LOSO-CV) stress classification performance has never been examined before.

This thesis has the following structure: Chapter 2 contains background information on automatic stress classification as well as traditional and machine learning techniques to denoise physiological signals. Chapter 3 describes our method for denoising BVP, extracting HRV, and automatically classifying stress. Chapter 4 presents our results. Chapter 5 is discussion. Chapter 6 is conclusion.

Chapter 2

Related work

2.1 Defining Stress and Motivating Automatic Stress Detection

Stress will be defined as strain caused by physical or psychological pressures [92]. Stress is a prevalent work-related health problem in Europe and can cause serious physical and mental health issues [43], as well as economic losses. Some examples of health issues stemming from stress include anxiety, depression, musculoskeletal disorders, chronic fatigue syndrome, and increased probability of infections [56, 1]. Such health problems can lead to absenteeism but also "presenteeism", where employees are at work, but are not performing to the best of their abilities. Absenteeism and presenteeism have been estimated to cost 272 billion EUR per year [36]. It is thus evident that stress has huge health and productivity related consequences and reducing stress is a useful goal.

The autonomic nervous system (ANS) controls stress-related physiological responses. The sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS) are the two main divisions of the ANS. The SNS provokes the fight-or-flight response [56] while the PNS regulates the rest and digest response. Increases in SNS activity change the body's hormone levels and cause responses such as muscle activation and faster heart rate [106]. Cognitive functions (such as logical thinking [74]) can also be affected, which may lead to mistakes.

In order to reduce stress, it is necessary to first detect it. Stress can be detected through self-report questionnaires. However, users must stop their current activity to complete such questionnaires, which affects their emotions. The possible biases of any form of self-reporting and the inherent difficulty of expressing emotions in writing are additional limitations of self-reports [74, 21]. Using physiological measurements to automatically detect stress is preferable as it is objective, continuous, and (depending on the invasiveness of the sensor) does not affect the user's emotion or greatly inconvenience the user.

Automatic stress detection could help manage and reduce stress. [37] measured driver stress using physiological signals. An application of such a technology could be to manage driver workload [8]. In stressful situations the navigation system could minimise alerts and phone calls could be automatically rejected. Conversely, in low-stress situations, the car could provide more entertainment possibilities. Computers could also leverage user stress information. For example, if a computer user is stressed, the computer could prevent unnecessary notifications [38]. More holistically, if a person could know that s/he was unusually stressed in a particular week, s/he could make behavioural changes, such as working out more, to minimise stress [38]. Alternative uses of stress detection could be to train people to handle stressful situations effectively. For example, personalized stressors of progressively increasing magnitude could be used to train emergency response workers [14]. The possibilities for automatic stress management systems far exceed the examples listed here. The main takeaway is that such interventions are only possible if stress levels can be monitored automatically.

2.2 Automatic Stress Detection from Physiological Signals

Photoplethysmography (PPG) is a non-invasive optical technology that measures the volumetric variations of blood circulation by using a light source and a photodetector. The light source emits light onto the skin and measures light reflected from (reflectance mode PPG) or transmitted through (transmission mode PPG) the tissue. These two modes are illustrated in figure 2.1. Infrared LEDs are the most commonly used light source in PPG. Because blood absorbs light more than the surrounding tissue, changes in blood flow cause changes in light intensity which can be measured by the photodetector. The signal recorded by PPG is called the blood volume pulse (BVP). BVP contains important cardiovascular information [97] and has repeatedly been used to predict stress [16, 61, 67, 88]. PPG technology is well used today because it is cheap, simple to operate, and comfortable to wear [95]. Because PPG is ubiquitous in wearable sensors and has been repeatedly used to predict stress, we focus on this technology.

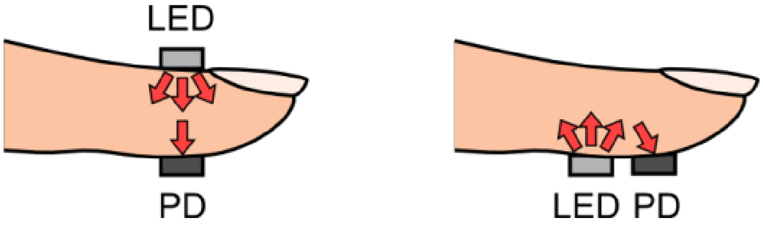

Figure 2.1: Transmission vs reflectance PPG [97]

Heart rate (HR) is the number of heartbeats in a minute. The time intervals between consecutive heartbeats are called interbeat intervals (IBIs). Heart rate variability (HRV) is the variation in IBIs [64]. Several studies have used HRV to measure mental stress [72, 47, 51]. Many authors [72, 24, 39, 107] have proposed several hand-crafted metrics, that are computed from IBIs, to describe HRV. HRV metrics are commonly calculated from the RR intervals of ECG measurements [6]. However, HRV metrics can also be calculated from the PP intervals of BVP measured using PPG [47]. The term pulse rate variability (PRV) can be used when working with PPG to clarify the distinction [90].

HRV metrics can be largely divided into the time-domain and frequency domain. Common time domain metrics are the standard deviation of the IBIs (SDNN), the root of the mean squared difference of consecutive IBIs (RMSSD), and the proportion of successive IBIs that differ by more than 50 milliseconds (pNN50). Let I be the sequence of IBIs in milliseconds and $\bar{I}$ be the mean IBI. Then, the equations for the time-domain features are given below. Although HR is not a HRV metric as such, the equation for calculating HR from IBI is shown for completeness.

$$
\begin{aligned}
\text{HR} &= 60/\bar{I} \\
\text{SDNN} &= \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(I_n - \bar{I})^2} \\
\text{RMSSD} &= \sqrt{\frac{1}{N-1}\sum_{n=1}^{N-1}(I_{n+1} - I_n)^2} \\
\text{pNN50} &= \frac{1}{N-1}\sum_{n=1}^{N-1}\mathbb{I}(|I_{n+1} - I_n| > 50)
\end{aligned}
\tag{2.1}
$$

Where $\mathbb{I}(\cdot)$ is the indicator function. To compute frequency-domain features, the spectrogram of the HRV can be used to separate the signal into its component rhythms that operate within different frequency ranges or bands. The periodogram can be integrated to find the power within a frequency band. This "area under the curve" is illustrated in figure 2.2. The low frequency (LF, 0.04Hz - 0.15Hz) and high frequency (HF, 0.15-0.40Hz) bands have been found to be related to SNS and PNSS activity [10]. Consequently, many studies have suggested using the LF/HF ratio as a sign of stress [72, 10]. Table 2.1 outlines the most important features computed from the IBI sequence.

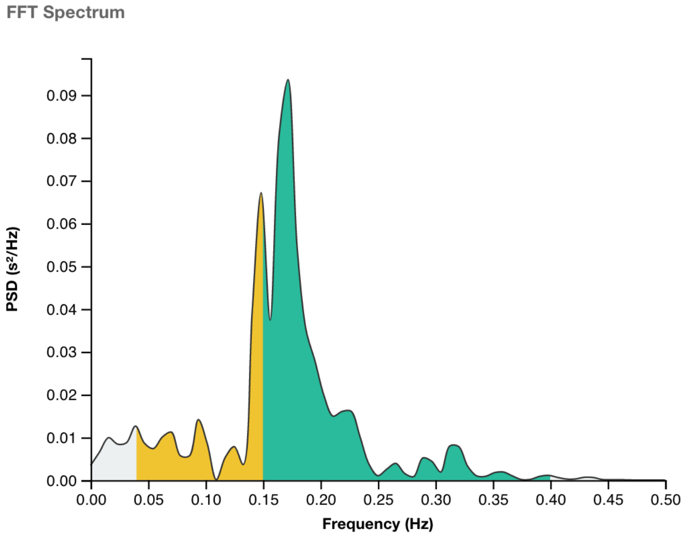

Figure 2.2: Graphical illustration of the LF (yellow) and HF (green) bands on a periodogram. Source: [42]

	Non-HRV	HRV
Time-domain	HR	SDNN, RMSSD, pNN50
Frequency-domain		LF, HF, LF/HF

Table 2.1: Key features calculated from IBI sequence. Adapted from [85]

2.2.1 Measuring Physiological Signals in the Wild

For use in the automatic stress detection systems described earlier, physiological signals must be measured in the wild, and not in highly controlled clinical settings. Additionally, wrist-based sensors, such as the Empatica wristband, are more practical than fingertip sensors, such as the Infiniti sensor. Unfortunately, physiological signals measured in the wild can be noisy [3]. Everyday activities and light exercise, as well as environmental noise, can cause inaccurate PPG recordings [11, 113]. Furthermore, wristband sensors tend to suffer from more noise than fingertip sensors because they are so sensitive to motion artifacts caused by hand movements [112]. This noise impacts the estimation of PRV from the PPG signal, which then impacts downstream stress classification. Poor signal quality can also cause inaccurate heart rate estimation [113]. Denoising PPG signals from low cost wristband sensors is thus an important goal for the research community.

In addition to suffering from noise, low cost wearable sensors have lower sampling rates [35]. This is partially because lower sampling rates have decreased power consumption, which increases battery life [22, 105, 62, 70, 19, 33]. Additionally, energy used in data transmission also decreases because there are fewer bytes to transmit. More generally, it is becoming increasingly difficult to store the increasing amount of digital health data [30]. The total global amount of digital health data is estimated to have exceeded 2 trillion gigabytes in 2020, while data storage requirements are quadrupling every two to three years [66, 99]. Therefore, it is desirable to trim data volumes while retaining clinically important information. Lower sampling rates are an effective way to achieve this decrease in storage requirements. Because of these advantages of low sampling rates, particularly with respect to low cost wearable sensors, this current work focuses on low sampling rate signals.

2.3 Traditional Signal Processing Techniques for Denoising PPG Signals

To address the aforementioned noise in PPG signals, several denoising signal processing algorithms have been developed. These include wavelet filters [49, 58, 78], filters centred on frequency analysis [54, 80, 113], and adaptive filters [69, 79]. These approaches have had some success in increasing the signal to noise ratio of corrupted PPG signals.

However, the current signal processing techniques do have some limitations. Frequency analysis struggles to denoise non-stationary noise and noise reference signals must be measured to use adaptive filters [59]. Furthermore, these traditional signal processing techniques can cause phase shifts in the PPG signal [32]. Another downside of current methods is they require parameters that depend on the target signal. The correct type of wavelet must be chosen for a particular PPG signal [58]. For example, the Haar wavelet is not appropriate for motion artifact reduction in PPG

[32]. Moreover, changes in blood flow over long recordings could mean that no single wavelet is appropriate for the entire signal [59]. Similarly, bandpass filters (which pass frequencies within a certain range and reject frequencies outside that range) require the frequency range of interest to be specified [2]. Specifying these parameters is time consuming and requires subject expertise.

2.4 Machine Learning Techniques for Denoising Physiological Signals

2.4.1 Brief Overview of Artificial Neural Networks

As an alternative to the traditional approaches in section 2.3, machine learning, and in particular ANNs, has been used to denoise physiological signals. Modern neural networks were introduced in the 1950s, with Rosenblatt's famous Perceptron model [82]. This was followed by Multilayer Perceptron (MLP) models in the 1970s [44, 45], where many perceptrons are stacked together in layers. Since then, more advanced ANN architectures such as the CNN [4], recurrent neural network (RNN) [84], bidirectional RNN [86], and long short-term memory (LSTM) [40] have been introduced.

2.4.2 Artificial Neural Networks for Denoising Physiological Signals

Machine learning approaches do not require the specification of target signal specific parameters, as the model can learn the appropriate parameters itself from training data [2]. It should be noted that the training data must represent the non-stationarity observed in real-world signals. ML approaches have also been found to denoise better than traditional filtering methods [59, 2]. CNNs have been used to denoise EEG signals [94]. [81] used a feed forward neural network to denoise ECG signals. [109] developed a multi-step pipeline to denoise ECG signals, where a wavelet transform first filters out most of the noise, and then a deep neural network denoising autoencoder removes any remaining noise. [111] developed a stacked sparse autoencoder to denoise EEG signals. [2] developed a deep recurrent neural network for ECG signal denoising. Interestingly, they pretrained the model with synthetic ECG data and fine-tuned on real data. [98] focused on predicting pain (which is not the same as stress but is an analogous task) from a multimodal input of EDA, EMG, and ECG physiological signals. They used a deep denoising convolutional autoencoder to fuse information from the different modalities.

ML approaches have also been applied to BVP. [91] used an autoencoder to reconstruct a BVP signal. However, the focus of their study was not on denosing. In fact, their BVP signals were very clean as subjects simply sat still and did not perform any tasks during measurements. Their study instead used the latent representation of the autoencoder to predict blood pressure. [100] used a denoising convolutional autoencoder on BVP to pretrain classifier weights for atrial fibrillation detection. They also used an auxiliary task (signal quality assessment) to leverage the benefits of multi-task learning. [76] trained a deep CNN to classify PPG signal as noise or not noise (but did not denoise the signal).

Expressly denoising a PPG signal has advantages over using an autoencoder as a pretraining task or to obtain a latent representation. This is because the BVP waveform itself is well studied and has many informative features. For example, the second derivative of the BVP waveform contains useful information on the distensibility of large arteries [46]. Additionally, many tools are designed to work on the BVP signal, not on an idiosyncratic latent representation. An example would be `heartpy` [103], the Python package used in this thesis to extract HRV metrics.

To our knowledge, there have only been two applications of denoising autoencoders to purposely denoise PPG signals [59, 83]. [59] trained a bidirectional recurrent denoising autoencoder (BRDAE) by adding synthetically generated noise to PPG recordings and training the BRDAE to reconstruct the original clean signal. The BRDAE was effective at denoising noisy PPG recordings collected in uncontrolled conditions. Comparison of HR calculated from the denoised PPG signal with HR calculated from reference ECG was used to evaluate denoising performance. [83] proposed a DAE denoising method that requires the detection of PPG beats. However, the detection of PPG beats may not be possible in situations of severe noise, which limits their method. To evaluate denoising quality the following assessments were used: correlation of denoised signal with clean signal, RMSE between clean and denoised signal, and signal-to-noise ratio (SNR) improvement after denoising. The number of acceptable PPG beats for measuring two clinical features (crest time, and systolic to diastolic peak height ratio) was also used to evaluate denoising quality.

To contribute on top of these early works, we investigate the usefulness of a DAE in extracting HRV metrics and as a module within a stress classification framework. The simple metrics, such as HR, used in the earlier studies do not contain much information regarding mental state or general physiology [89]. As mentioned before, HRV metrics are commonly used in stress detection but are generally clinically relevant markers too. Just to name some other applications, HRV has been used to detect autonomic neuropathy in diabetic patients [29], while reduced HRV has been associated with a higher risk of postinfarction mortality [108]. Additionally, existing work doesn't evaluate the effect of denoising on a practical downstream task, such as blood pressure estimation or mental stress classification. It would be advantageous to test how DAE denoising affects performance on such a task, which would hopefully pave the way to better employment of noisy PPG signals collected in the wild. This thesis extends previous work by evaluating the effect of DAE denoising on i) accuracy of extracted HRV metrics ii) LOSO-CV performance on automatic stress classification.

It should also be noted that [59] denoised 125Hz signals while [83] used 60Hz signals. As explained in section 2.2, we focus on denoising low sampling rate signals to support the use of low cost wearable sensors. [9] and [18] found that reliable PRV metrics could be calculated from PPG sampled at frequencies as low as 50Hz and 25Hz respectively. For this current study, we wanted to see if sampling rates pushed even lower would still yield informative insights about mental state. The heart rate of healthy adults typically does not exceed 2Hz [101]. With this ballpark value in mind, it was determined that 16Hz was a reasonable frequency to downsample the BVP signals. 16Hz would not be so low that it would completely fail to capture important cardiovascular information, but would be low enough to investigate the effect of DAE denoising on low sampling rate signals. To the best of our knowledge, no other work has evaluated the denoising performance of DAE on such low sampling rate BVP signals.

2.5 Summary

To summarise, stress is a serious problem with grave mental and economic consequences in modern society. In order to develop effective stress management systems, it is necessary to automatically recognise stress from physiological signals. Wearable PPG sensors are a promising low cost and convenient technology to obtain BVP signals. HRV metrics are informative features that can be calculated from BVP signals (or more specifically IBIs) and used to classify stress. Unfortunately, wearable PPG sensors in the wild suffer from noise artifacts, and traditional denoising methods have many limitations. ML DAE techniques have been applied to denoising BVP signals, but existing works have only used simple metrics such as HR to evaluate denoising quality, and have not examined the performance of DAEs on the low sampling rates common in low cost sensors. To address these shortcomings, our work proposes a novel framework using a DAE to denoise low sampling rate PPG signals before performing stress classification from extracted HRV metrics. We examine if DAEs can improve the accuracy of HRV metrics and LOSO-CV stress classification performance.

Chapter 3

Methodology

This chapter describes the dataset used in this thesis. We also describe the traditional signal processing baseline and proposed DAE method for BVP signal denoising. Finally, we describe our stress classification model. To highlight once again, we are proposing a novel framework using a DAE as a module to aid stress classification. Hence, our stress classification model itself is simple, but can be improved in future work.

3.1 Dataset

The dataset collected by [85] (Stress Dataset) was used in this thesis. Data was collected from 20 participants in an experiment designed to induce stress. Participants were asked to remain seated during the data collection phases, each of which lasted five minutes. Physiological data was collected from the participants from a number of contact sensors. The only contact sensors relevant to this current work are an Infiniti sensor on the tip of the index finger of the left hand, and Empatica E4 wristbands on each wrist. Both sensors used PPG to record BVP signals. In order to minimise artifacts, participants were requested to keep their Infiniti finger still. The collected data streams from the various contact sensors were roughly synchronised to within around one second.

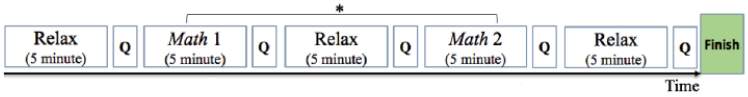

Figure 3.1: Dataset experiment protocol: Q = questionnaire; *the easy and difficult math tasks were counterbalanced. Source: [85]. Image originally adapted from [15].

A study protocol diagram is shown in figure 3.1. The experiment was divided into five sessions, each lasting five minutes. The participant relaxed in sessions 1, 3 and 5. In sessions 2 and 4, the participant performed some mental arithmetic. One of the maths tasks was easy and the other was more difficult. In this thesis we will refer to the sessions by the following terms: Rest 1:

initial rest session; Easy Maths, easy maths session; Hard Maths, hard maths session; Easy Rest, rest session immediately after easy maths session; Hard Rest, rest session immediately after hard maths session. Participants filled in a self-assessment questionnaire after each session (fig. 3.2).

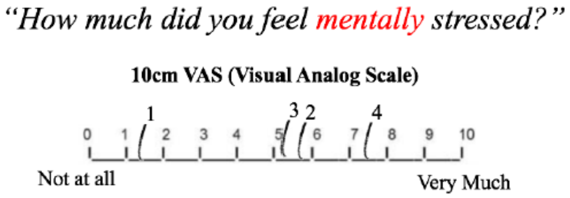
"How much did you feel mentally stressed?"
10cm VAS (Visual Analog Scale)
1
3 2
4
0
1
2
3
4
5
6
7
8
9
10
Not at all
Very Much

Figure 3.2: Self-assessment questionnaire [85].

For the self-assessment questionnaire after each session (denoted by "Q" in figure 3.1), participants were asked to subjectively rate "how much did you feel mentally stressed?" (figure 3.2). These self-report ratings, however, were not used in the current study.

For full details about the dataset collection, including the specific arithmetic task used, please see the original data collection thesis [85].

3.1.1 BVP Signals Used

The Stress Dataset contained physiological signal recordings for 20 participants. However, due to issues with missing data, we could only use the data from 17 participants.

This thesis uses the Infiniti BVP data and the BVP data collected by the Empatica wristband on the left arm. The left arm Empatica was chosen because, while trying to keep the Infiniti finger on the left hand still, it was thought that participants would also keep the left wrist still and thus there would be fewer motion artifacts in the Empatica left wristband data. Of course, the aim of this thesis is to denoise noisy signals so it is important to note that there were still plenty of noisy spans in the BVP signals collected from the left arm.

3.2 Annotating BVP Signals as Noisy vs Clean

In order to train a DAE, it was necessary to first manually annotate BVP signals as clean/noisy. The rational behind this is explained in section 3.4.

The typical BVP signal is shown in figures 3.3 and 3.4. The signal commonly rises quickly with the systolic contraction, and falls more slowly. Under some conditions, the diastolic contraction will cause a dicrotic notch (see figure 3.4) in the falling signal [60].

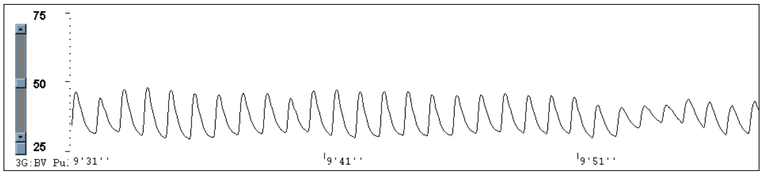

Figure 3.3: Typical BVP signal recorded by Infiniti sensor. Source: [31]

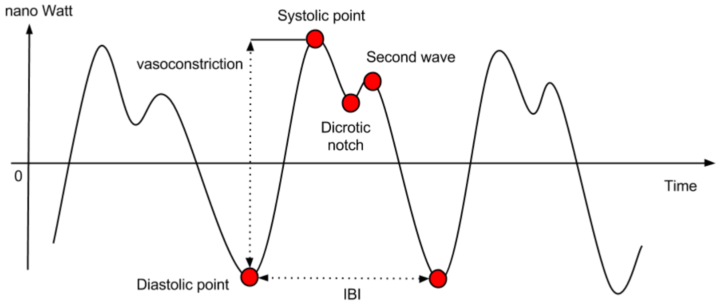

Figure 3.4: Typical BVP signal recorded by Empatica sensor. Source: [27]

We used the typical signals in figures 3.3 and 3.4 as a reference for clean signal. Thereby, we were able to annotate BVP signals as noisy if there seemed to be significant motion or other artifacts corrupting the true signal.

We manually annotated, as either clean or noisy, each 0.5 second span of the Infiniti BVP signal corresponding to each treatment (Rest 1, Easy Math, Easy Rest, Hard Math, Hard Rest) of each of the 17 participants. An example annotation is shown in figure 3.5.

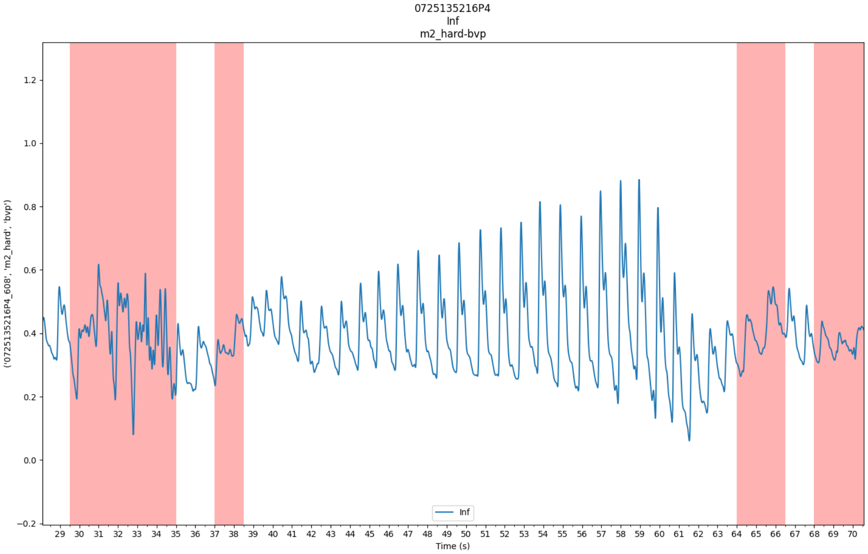

Figure 3.5: Annotated Infiniti BVP signal. Red spans are noisy signals.

The same annotation was performed for the BVP signal recorded by the Empatica wristband on the left arm.

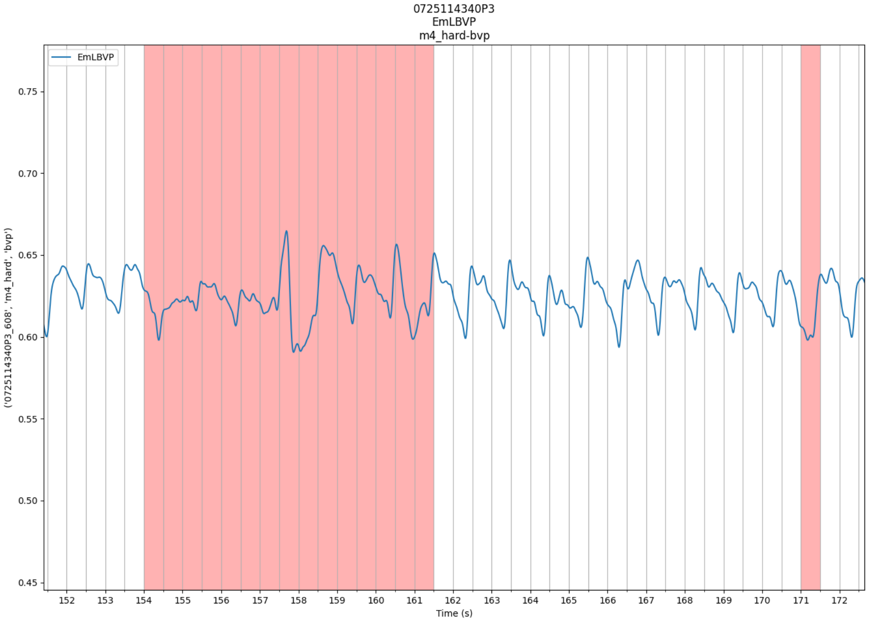

Figure 3.6: Annotated Empatica BVP signal. Red spans are noisy signals.

These highly granular manual annotations of the BVP signals are very time consuming. The annotations will be made available to the research community to complement the original Stress Dataset.

3.3 Preprocessing BVP Signals

We performed the BVP signal processing illustrated in figure 3.7 to i) perform traditional signal denoising as a reference ii) prepare BVP signal data for DAE training.

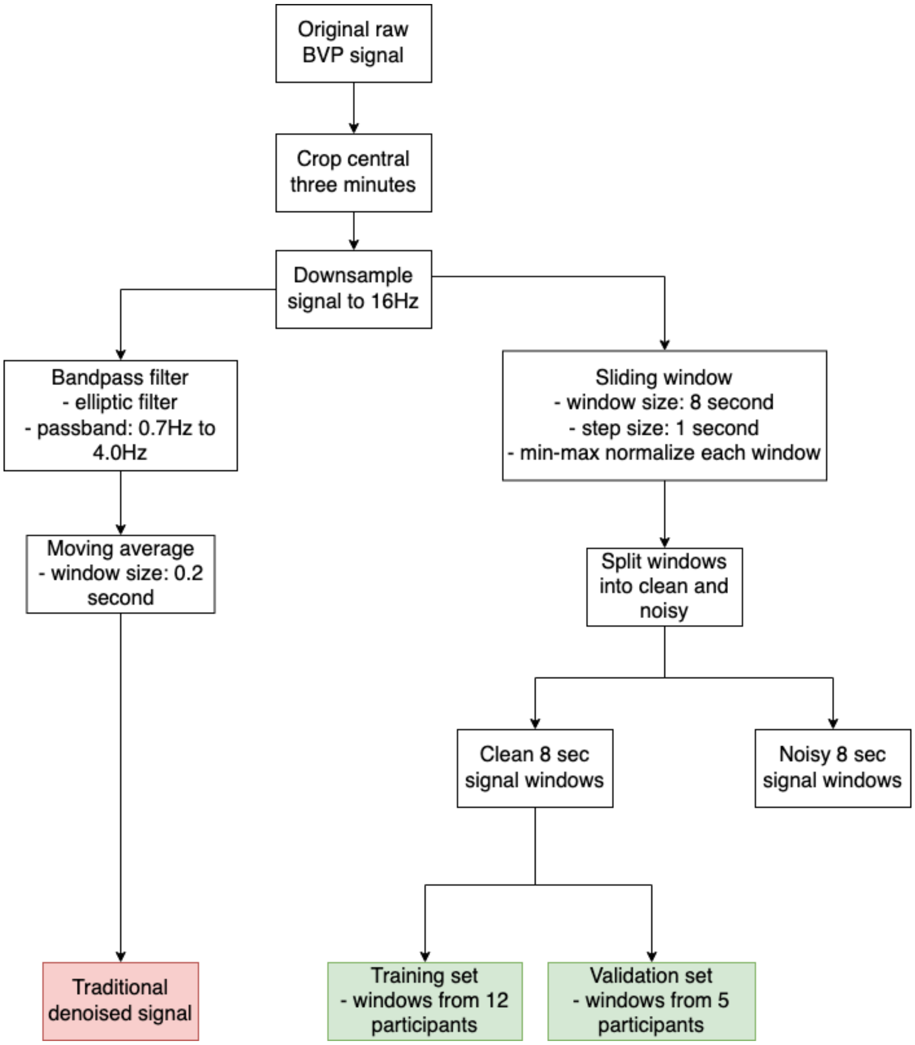
Original raw BVP signal
Crop central three minutes
Downsample signal to 16Hz
Bandpass filter
- elliptic filter
- passband: 0.7Hz to 4.0Hz
Sliding window
- window size: 8 second
- step size: 1 second
- min-max normalize each window
Moving average
- window size: 0.2 second
Split windows into clean and noisy
Clean 8 sec signal windows
Noisy 8 sec signal windows
Traditional denoised signal
Training set
- windows from 12 participants
Validation set
- windows from 5 participants

Figure 3.7: Data processing used to obtain i) traditional denoised signal ii) training and validation data sets for proposed DAE.

Our proposed framework is illustrated in figure 3.8. The "Stress classification model" module can be improved in future work.

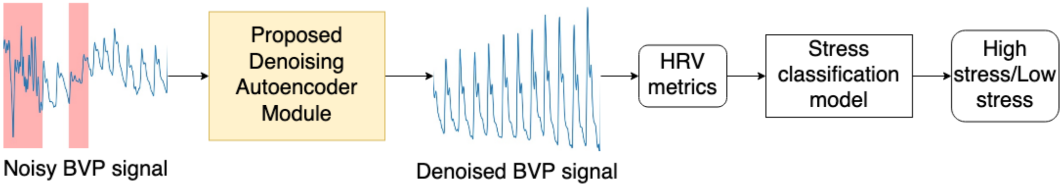

Figure 3.8: Outline of proposed framework using a DAE as a module within a stress classification system.

3.3.1 Preprocessing Common for Both Traditional Preprocessing and Denoising Autoencoder

Following [85], only the central three minutes of the BVP signal of each session was used. The first and last minute were cropped out. This gave participants time to acclimate to the emotion of the session, and avoided the potential problem of people giving up on the math task towards the end. Infiniti and Empatica signals were originally sampled at 256Hz and 64Hz respectively. However, as explained in chapter 2, we wanted to examine the ability of DAEs to denoise low sampling rate signals and downsampled the PPG signals to 16Hz.

3.3.2 Traditional Preprocessing Benchmark

We needed a baseline denoising method using traditional signal processing techniques. It should be noted that both the traditional denoising and DAE denoising were performed on the 16Hz downsampled signal. For this baseline denoising method we first bandpass filtered the signal to remove low and high frequency noise. We then further filtered the signal using a moving average. For the bandpass filter an elliptic filter was used [41]. The lower and upper cutoff frequencies were set to 0.7Hz and 4Hz respectively. This range was chosen as it will include the HRs of most healthy adults [101]. A 0.2 second window (empirically chosen) was used for the moving average filter.

3.3.3 Preparing Denoising Autoencoder Dataset

The Stress Dataset was too small to train a DAE on the three minute signals. Additionally, the computational requirements to train such a model (2880 input length at 16Hz) would have been prohibitively great. Consequently, a sliding window was used to generate more training examples and to reduce the input size. We used a sliding window with a window size of eight seconds (128 samples at 16Hz) and a step size of one second (16 samples at 16Hz). Eight seconds was deemed to be an appropriate window size as, with HR unlikely to drop below 0.7Hz, eight seconds would contain at least five PPG heart beats. This would allow the DAE to learn the shape of a PPG beat. We also min-max normalized each eight second window so that all samples would lie between 0 and 1 [73].

An autoencoder is a type of ANN that attempts to reconstruct its input [71]. For example, in our case we would input the 128 samples of the BVP signal, the DAE would encode the signal into a compressed latent representation, and then attempt to reconstruct the original input while minimizing reconstruction error. The weights of the DAE can be trained through backpropagation, as with any ANN, to learn this mapping [71]. If the hidden layer of the network contains more nodes than the input, then the model may learn the trivial identity function and is not particularly useful. However, by imposing a "bottleneck" by having fewer hidden layer nodes than input nodes, the autoencoder can learn a compressed, lower dimensional representation (often called a latent representation) of the input. If the input features were independent of each other, this compression would be very difficult. However, by learning and leveraging the structure or patterns that exist in the input data, compression is achievable.

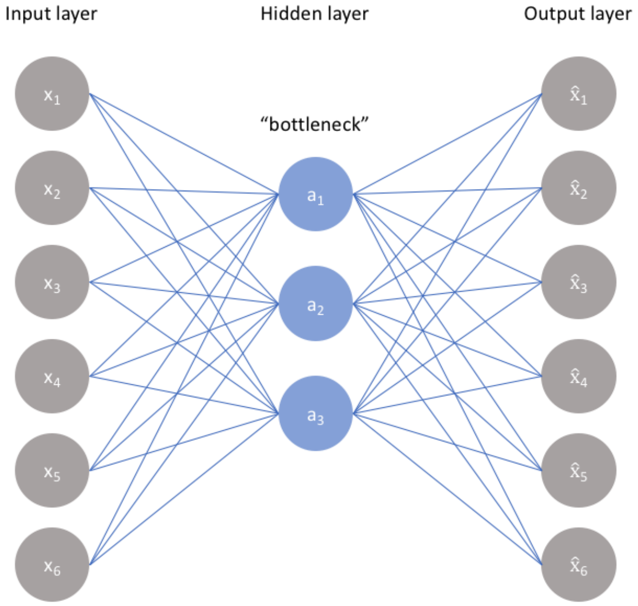

Figure 3.9: Illustration of a bottleneck in an autoendoer [48].

By training our autoencoder on clean BVP signals, the model will learn the patterns in clean signals. However, if we input a noisy signal into the trained model, the model may not have the capacity to reconstruct the strange noisy patterns, and could instead output a reconstructed signal that resembles a clean signal. In other words, the autoencoder can denoise the signal. Consequently, for training and validating our autoencoder, we only used clean signals. As described in section 3.2, we manually annotated BVP signals as clean or noisy. We were very strict in defining a clean window for model training purposes, and only used windows where the entire signal was clean (e.g. we did not include an entire eight second window even if only 0.5 seconds were noisy). This strictness was enforced to ensure that the autoencoder only learnt the patterns that exist in clean signals.

The clean signals were split into a training and validation set. The validation set consisted of all clean windows from five participants. The training set consisted of all clean windows from the remaining 12 participants. The dataset was split along participants (rather than just diving shuffled examples into a say 70% training 30% validation split) for two reasons. First, due to the use of a sliding window with such a large overlap, randomly splitting the data into training and validation would likely allow validation data to leak into the training set. Second, there is often large subject-to-subject variation in experiments on humans [110]. Therefore, splitting the dataset along participants is a more stringent way of testing model performance, as it assesses how well the model can generalize to unseen subjects.

3.4 Denoising Autoencoder Architecture and Training

A simple MLP architecture was used for the DAE. The MLP had an input layer, a single hidden layer, and an output layer. The input and output size of the DAE were both 128 (determined by the eight second window at 16Hz). We used one hidden layer of eight nodes (empirically chosen). The activation function used for the encoder was the Rectified Linear Unit and for the decoder was the sigmoid activation function. We used a batch size of 32. We trained the model with the Adam optimizer [52] for 1000 epochs and our loss function was the mean absolute error.

3.5 Denoising Entire Signal

As explained in section 3.3.3, the DAE did not have sufficient capacity to reconstruct entire three minute signals. The DAE was instead trained to denoise eight second windows of BVP signal. To denoise the entire three minute signal, we replicated the method used by [59]. Our sliding window had a window size of eight seconds and a step size of one second. These windows were passed through our trained DAE and the denoised windows were aligned and averaged at one second intervals to generate the denoised three minute signal (illustrated in figure 3.10).

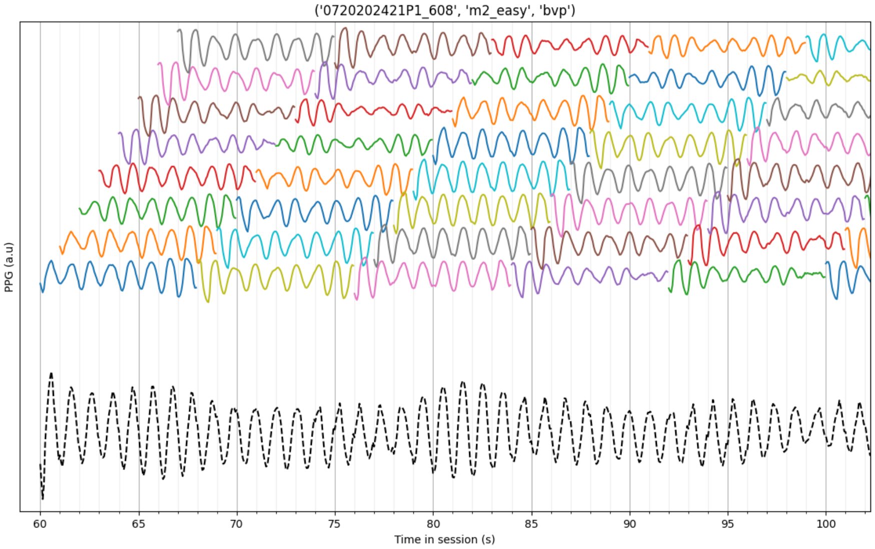

Figure 3.10: Illustration of the averaging of denoised windows to generate the final denoised three minute signal. The averaged signal is shown at the bottom as a dotted line. The averaged signal has been multiplied by 4 to emphasize waveform features. The many signals visible above are the denoised eight second windows. The denoised segments are displayed in different colors to distinguish them from each other. a.u: Arbitrary Unit.

This method posed some advantages over simply concatenating denoised eight second windows end to end. If a given eight second window is denoised poorly, the other overlapping denoised windows can help average out the poor denoised output. An example of this can be seen in figure 3.11.

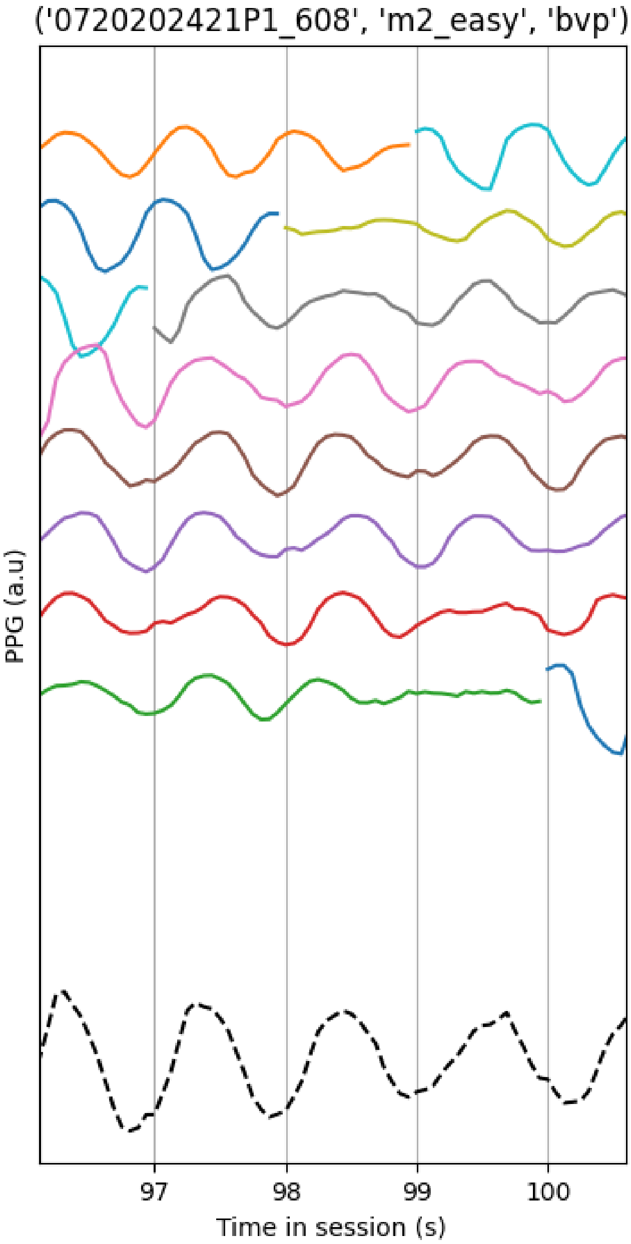

Figure 3.11: Please look at the window from 99s to 100s. Some of the denoised eight second segments (e.g. the green and the red lines) fail to reconstruct the PPG peak from 99s to 100s. However, the other overlapping segments succeed in reconstructing this peak. In the averaged signal, the peak is recovered. a.u: Arbitrary Unit.

3.6 Stress Classification using Naïve Bayes

3.6.1 Data Labelling

We developed a stress classification model to evaluate whether DAE denoising aided stress classification using HRV features as inputs. First, we had to label the data. Given the small dataset, we only used two levels of stress (following the work done by [85]). The two stress classes we used were "low stress" and "high stress".

The data from the "Hard Maths" sessions were labelled "high stress". The initial "Rest 1" was used as a baseline reading. The "Easy Maths", "Easy Rest", and "Hard Rest" sessions combined were labelled as "low stress". This labelling strategy was chosen because it led to the best stress classification performance on Infiniti BVP in [85].

3.6.2 Data Preprocessing

As in section 3.3, we only used the central three minutes of each session. Following a similar approach in [85], we divided each three-minute session into two two-minute sliding windows with one-minute overlap. HRV features were extracted from each two-minute window.

We used the first two-minute window of "Rest 1" as a baseline. HRV features were extracted, using the `heartpy` Python package [104], from the eight "data windows" of the remaining four sessions. Each human has a unique physiology, such as a different resting heart rate. To counteract the effects of such differences, we standardized the HRV features by dividing by the baseline feature (or subtracting the baseline feature for pNN50, which is a proportion). A comparable standardization approach was used by [85, 53]. 7 features (HR, SDNN, RMSSD, pNN50, LF, HF, LF/HF) were extracted from each data window.

3.6.3 Gaussian Naïve Bayes classifier

We used a Gaussian Naïve Bayes (GNB) classifier for stress classification because it has been shown to perform well on small imbalanced datasets [53]. Its performance was evaluated using LOSO-CV and examining the macro F1 average (averaged over both "low stress" and "high stress" classes) and Matthews Correlation Coefficient (MCC). For each fold of the LOSO-CV, the train and test datasets were normalized by subtracting the training set mean and dividing by the training set standard deviation.

The Naïve Bayes classifier "naively" assumes conditional independence between every pair of features given the value of the class variable:

$$P(y|x_1, ..., x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i|y)}{P(x_1, ..., x_n)} \tag{3.1}$$

$P(x_1, ..., x_n)$ is constant given the input. Therefore, the following classification rule can be used:

$$
\begin{aligned}
P(y|x_1, ..., x_n) &\propto P(y) \prod_{i=1}^{n} P(x_i|y) \\
\hat{y} &= \underset{y \in \{\text{low},\text{high}\}}{\operatorname{argmax}} P(y) \prod_{i=1}^{n} P(x_i|y)
\end{aligned}
\tag{3.2}
$$

The **Gaussian** Naïve Bayes classifier assumes the likelihood of the features to be Gaussian.

$$
P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_{iy}^2}} \exp(-\frac{(x_i - \mu_{iy})^2}{2\sigma_{iy}^2})
\tag{3.3}
$$

Putting it all together:

$$
\hat{y} = \underset{y \in \{\text{low},\text{high}\}}{\operatorname{argmax}} P(y) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{iy}^2}} \exp(-\frac{(x_i - \mu_{iy})^2}{2\sigma_{iy}^2})
\tag{3.4}
$$

Where x_i is the value of a feature, and y is the stress class. The mean (μ_{iy}) and standard deviation (σ_{iy}) of the ith component Gaussian are calculated using the values of x_i in the training set. The prior $P(Y = y)$ is computed by counting the frequency of class y in the training data. Although the simple GNB model was used in this work to propose our framework using a DAE as a module in stress classification, future works can use more complex classification models (such as ANNs).

Chapter 4

Results

In this chapter we evaluate the viability and performance of our novel DAE framework for stress detection. We achieve this by examining if DAE denoising increases signal quality and/or results in more accurate HRV features. Finally, we use LOSO-CV to evaluate stress classification performance following DAE denoising.

4.1 Graphical Inspection of Denoising Quality

First we qualitatively observed the denoising quality. As a first check, the DAE is able to acceptably reconstruct clean signals from the validation set (figure 4.1). More impressively, the DAE is able to better denoise noisy signals than traditional preprocessing (figure 4.2). However, there are also counter examples where the DAE does a worse job of denoising than traditional preprocessing (figure 4.3). There are also plenty of examples where the original signal is so noisy it is impossible to evaluate whether the DAE or traditional preprocessing has denoised well (figure 4.4). Not even our human eyes can gauge what the uncorrupted ground truth signal might be.

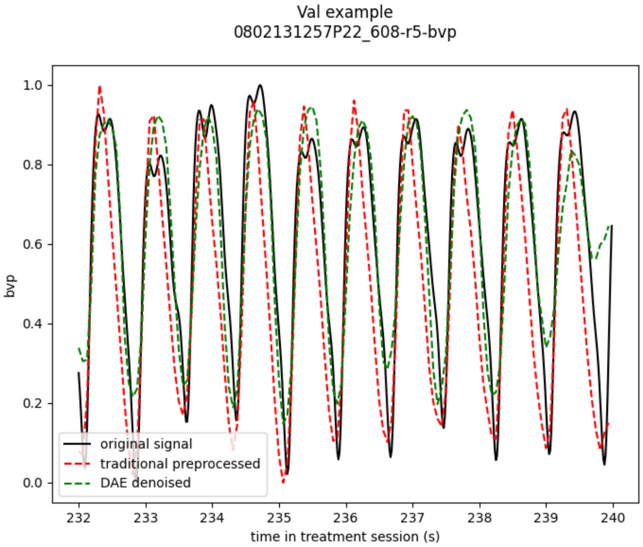

Figure 4.1: Plot of an eight second Empatica window in the validation set for training the denoising MLP. Both traditional processing and the DAE are able to reconstruct the clean signal.

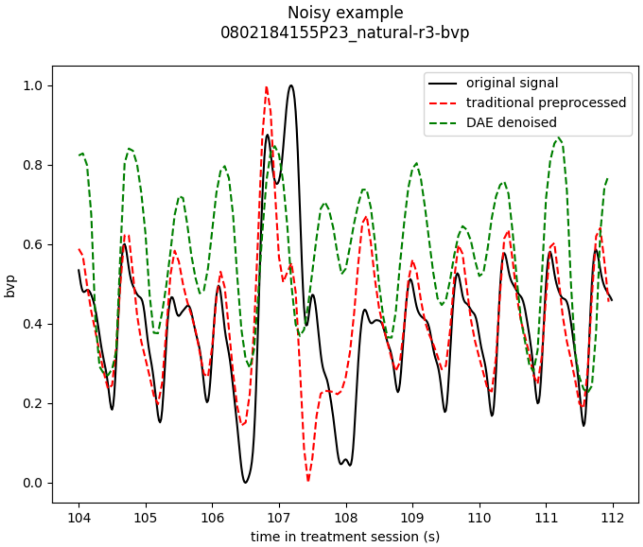

Figure 4.2: Plot of an eight second Empatica window in the noisy set. The traditional preprocessing seems to have lost the heartbeat between 107 and 108 seconds. The DAE however is able to recover that heartbeat.

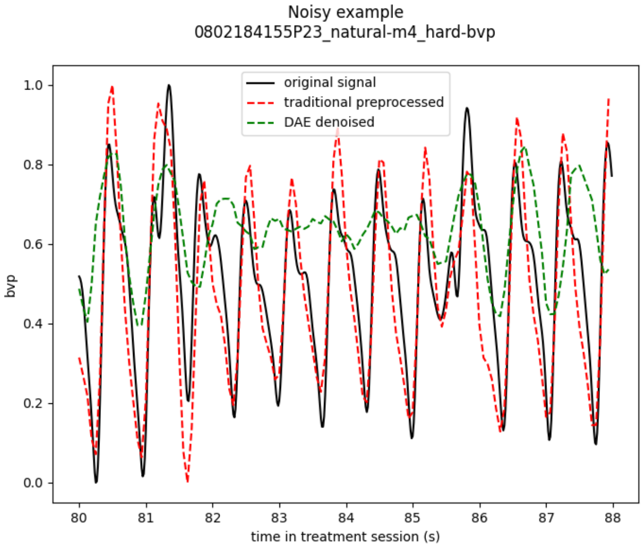

Figure 4.3: Plot of an eight second Empatica window in the noisy set. The DAE seems to have lost the clearly defined peaks from 82s to 85s. Traditional denoising however reliably accentuates these peaks.

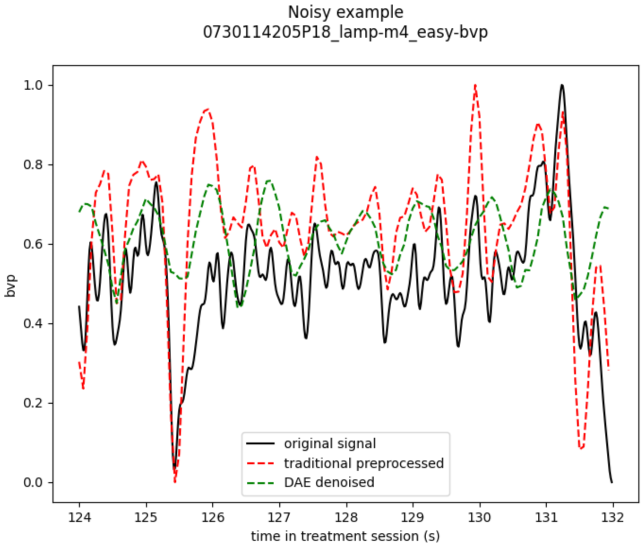

Figure 4.4: Plot of an eight second Empatica window in the noisy set. The original signal is so noisy it is impossible to gauge whether the DAE or the traditional preprocessing has denoised effectively.

4.2 Signal Quality Index Evaluation of Denoised Signal

To obtain a quantitative, high level evaluation of denoising quality, we used a metric to evaluate the quality of a BVP signal. This would allow us to assess if a denoising method improved signal quality. The relative power Signal Quality Index (pSQI) was chosen for this metric. The pSQI evaluates the intensity of a physiological signal in a frequency band of interest as a gauge of signal quality [17, 55, 20, 26]. The equation for the pSQI of BVP signals is given below (taken from [16]):

$$P(\hat{f}_{min} \leq f \leq \hat{f}_{max}) \cong \frac{\int_{\hat{f}_{min}}^{\hat{f}_{max}} PSD(f)df}{\int_{total} PSD(f)df} \tag{4.1}$$

where $0 \leq P \leq 1$, PSD is the power spectral density of the BVP signal, and $\hat{f}_{min}$, $\hat{f}_{max}$ are the lower and upper bound of expected HRs, respectively. The PSD of a signal describes its distribution of power into frequency components. For example, a 60 beats per minute (1 Hz) heart rate signal would have a peak in its power spectrum at 1Hz (figure 4.5). Following [16] we used 0.8Hz (48bpm) to 2.0Hz (120bpm) as the expected range of HRs, as this range will include the HRs of most healthy adults [102].

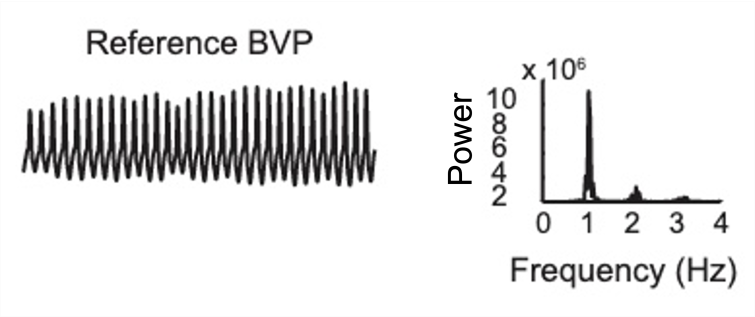

Figure 4.5: A BVP signal and corresponding PSD graph. Image adapted from [75].

Figures 4.6 and 4.7 show that both traditional preprocessing and DAE denoising are able to increase the pSQI of clean and noisy BVP signals. For both clean and noisy BVP signals, DAE denoising appears to increase pSQI more than traditional preprocessing. This is encouraging as it suggests that the DAE is able to denoise BVP signals effectively.

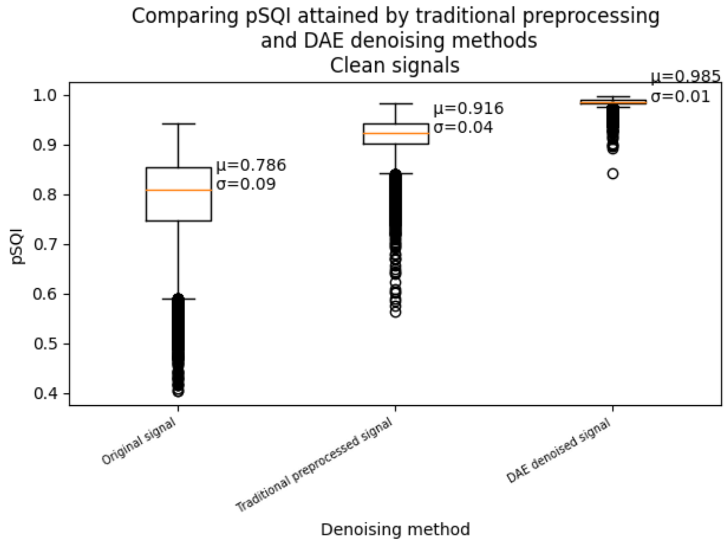

Figure 4.6: Box plot of pSQIs of clean eight second windows. Both traditional preprocessing and DAE increase the pSQI of clean signals. But DAE increases it more.

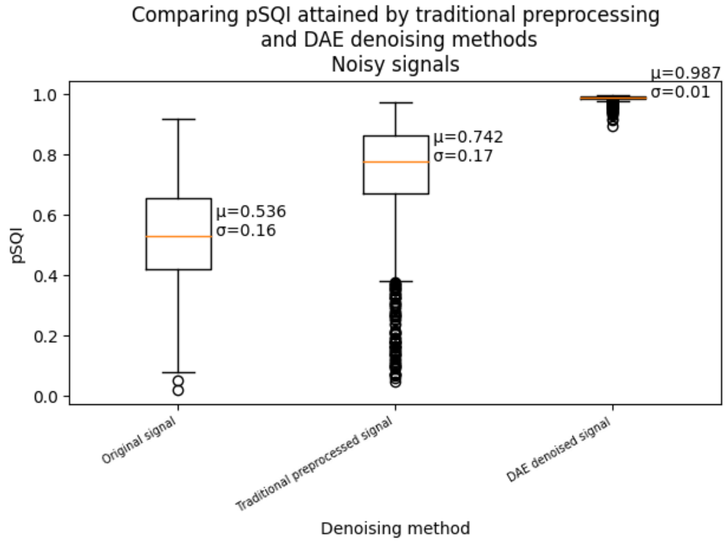

Figure 4.7: Box plot of pSQIs of noisy eight second windows. Both traditional preprocessing and DAE increase the pSQI of noisy signals. But DAE increases it more.

However, pSQI is not the perfect metric to evaluate denoising quality. pSQI just asesses the strength of a signal within the frequency band of interest (0.8Hz to 2.0Hz in our case). Therefore, if the DAE outputed a “denoised” BVP signal that had a strong strength in the frequency band of interest, but did not resemble the original signal at all (e.g. the peak positions were completely different), the pSQI of the denoised signal would be high but the practical use of the “denoised” signal would be limited.

4.3 Aside: Transfer Learning

This section is not central to the focus of the thesis, but some side experiments on using transfer learning for the DAE were conducted. A MLP DAE (same model architecture and training regime as explained in section 3.4) was trained on Empatica PPG signals and used to denoise Infiniti PPG signals. A MLP DAE was also trained on Inifiti PPG signals and used to denoise Empatica PPG signals. No fine-tuning on the target sensor signals was used in either case. The importance of a denoising method that can generalize to different sensors is explained in section 5.2.

Figures 4.8 and 4.9 show that training on Empatica and denoising Infiniti PPG still leads to reasonable denoising performance. The pSQI boxplots (figures 4.10 and 4.11) also show that at a higher level, the DAE trained on Empatica is able to increase the pSQI of Infiniti signals. The analogous figures 4.12, 4.13 and 4.14, 4.15 show that training on Infiniti and denoising Empatica PPG also leads to reasonable denoising performance.

Training on Empatica and Denoising Infiniti

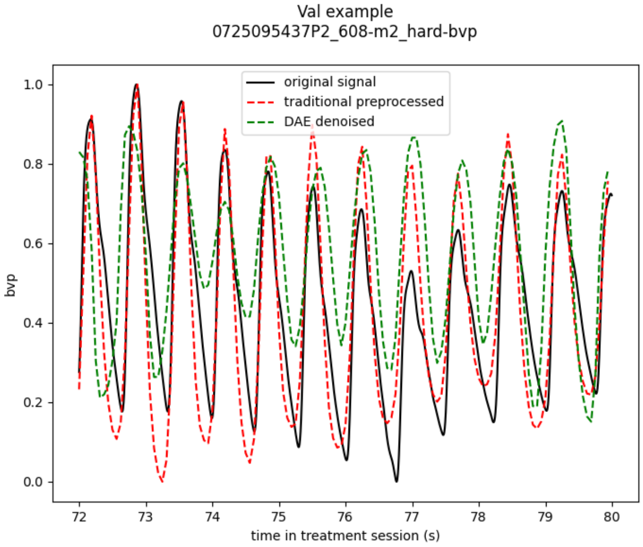

Figure 4.8: A DAE trained on clean Empatica signals and denoising a clean Infiniti signal.

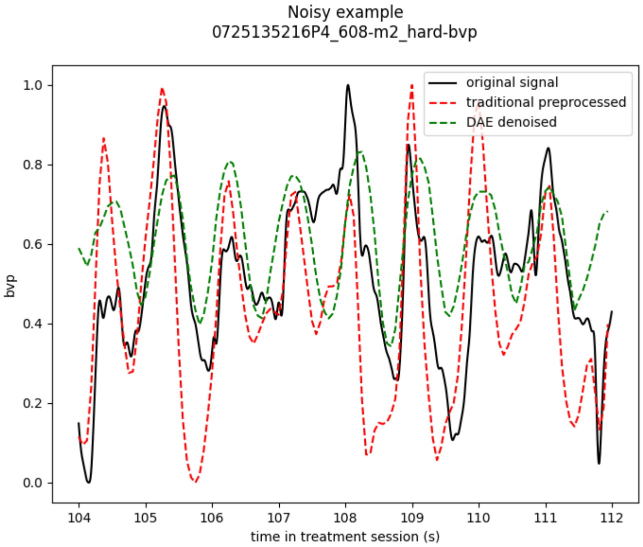

Figure 4.9: A DAE trained on clean Empatica signals and denoising a noisy Infiniti signal.

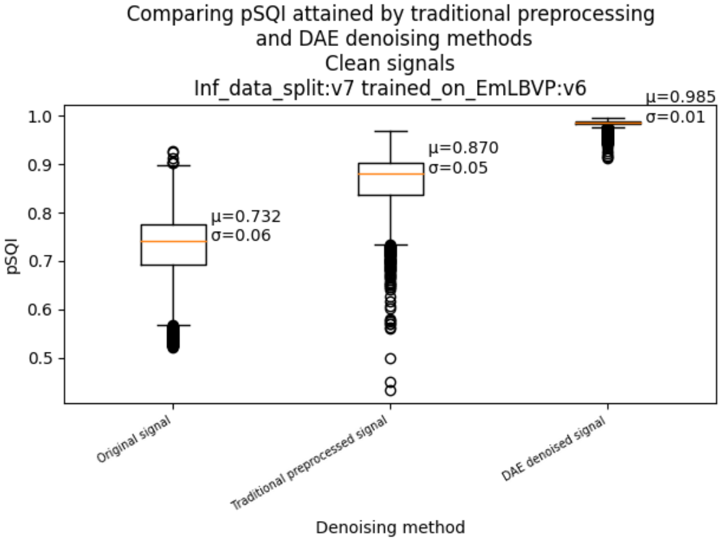

Figure 4.10: Box plot of pSQIs of eight second windows of clean Infiniti signals. The DAE used was trained on clean Empatica signals.

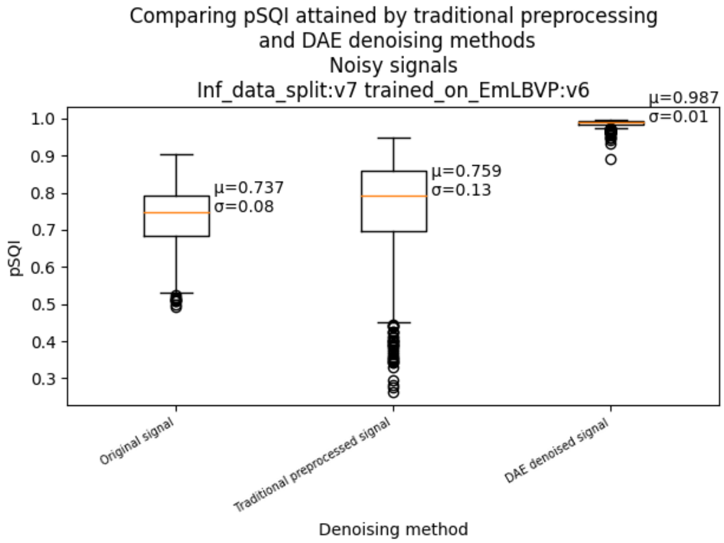

Figure 4.11: Box plot of pSQIs of eight second windows of noisy Infiniti signals. The DAE used was trained on clean Empatica signals.

Training on Infiniti and Denoising Empatica

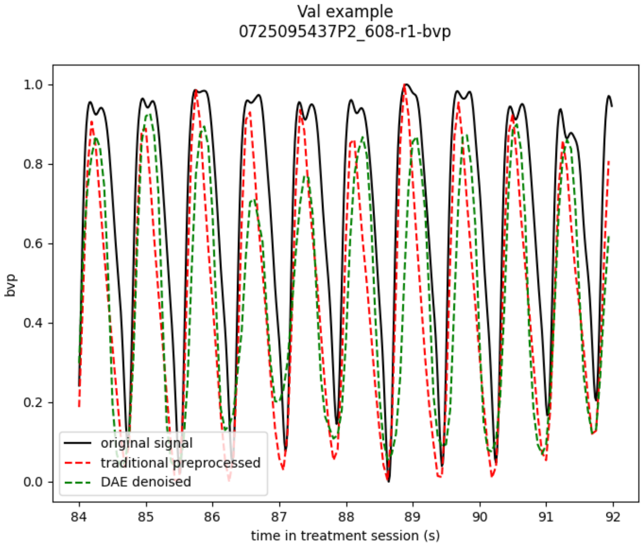

Figure 4.12: A DAE trained on clean Infiniti signals and denoising a clean Empatica signal.

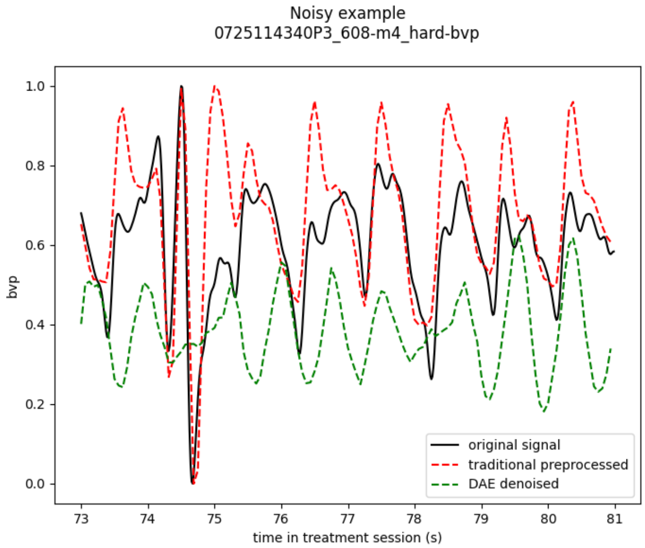

Figure 4.13: A DAE trained on clean Infiniti signals and denoising a noisy Empatica signal.

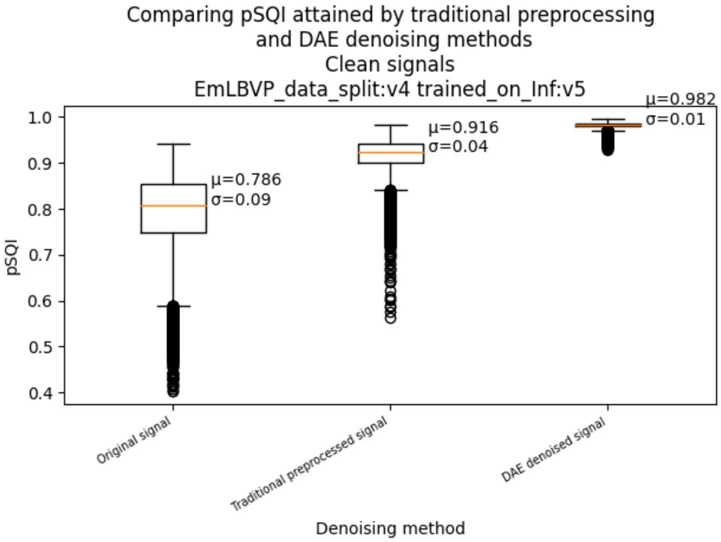

Figure 4.14: Box plot of pSQIs of eight second windows of clean Empatica signals. The DAE used was trained on clean Infiniti signals.

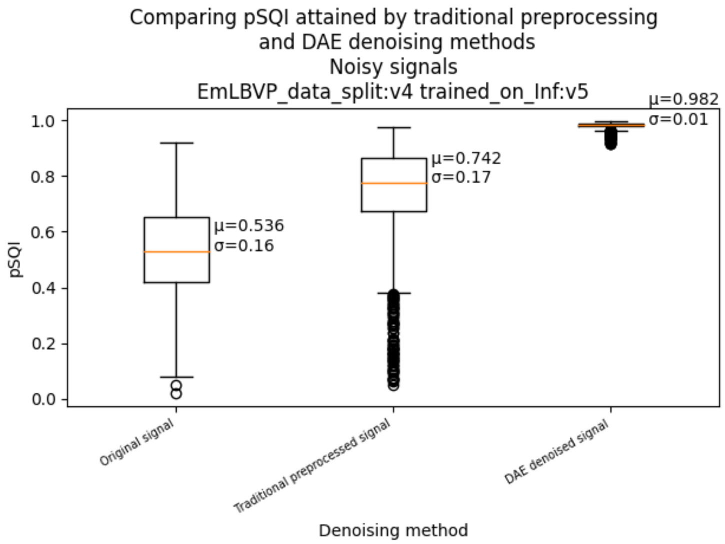

Figure 4.15: Box plot of pSQIs of eight second windows of noisy Empatica signals. The DAE used was trained on clean Infiniti signals.

4.4 Examining the Accuracy of Denoised Signal-Derived HRV

Due to the downsides of the pSQI metric mentioned in section 4.2, we needed another metric to evaluate denoising quality. The accuracy of HRV metrics extracted from denoised signals was used as such a metric. To achieve this, we needed ground truth HRV metrics.

The Infiniti and Empatica BVP signals are roughly synchronized (to within 1 second) in the Stress Dataset. The Infiniti sensor is more accurate and suffers from fewer noise artifacts. This is because the Infiniti sensor is attached to the finger, while the Empatica wristband is worn on the wrist. Wristband sensors tend to suffer from more noise than fingertip sensors because they are so susceptible to motion artifacts from hand movements [112]. Therefore, due to the rough synchronisation and the greater accuracy of the Infiniti signal, HRV features extracted from clean Infiniti PPG signals were used as the ground truth HRV.

To conduct the HRV root-mean-square error (RMSE) evaluation we first selected Infiniti BVP signals from sessions where the signal had no noise during the central three minutes of the session. Then, we selected the corresponding original, traditional preprocessed, and DAE denoised Empatica BVP signals. We then extracted HRV metrics from the central three minutes of each of these selected BVP signals. We excluded any sessions where the Empatica data was so noisy that `heartpy` was unable to extract HRV metrics. Finally, we compute the RMSE of each metric across all selected sessions. These steps are illustrated in figure 4.16.

There was only one session where the original and traditional preprocessed Empatica signals were so noisy that `heartpy` failed to extract HRV metrics from the session. It is worth noting that the DAE denoised signal however was clean enough for `heartpy` to work.

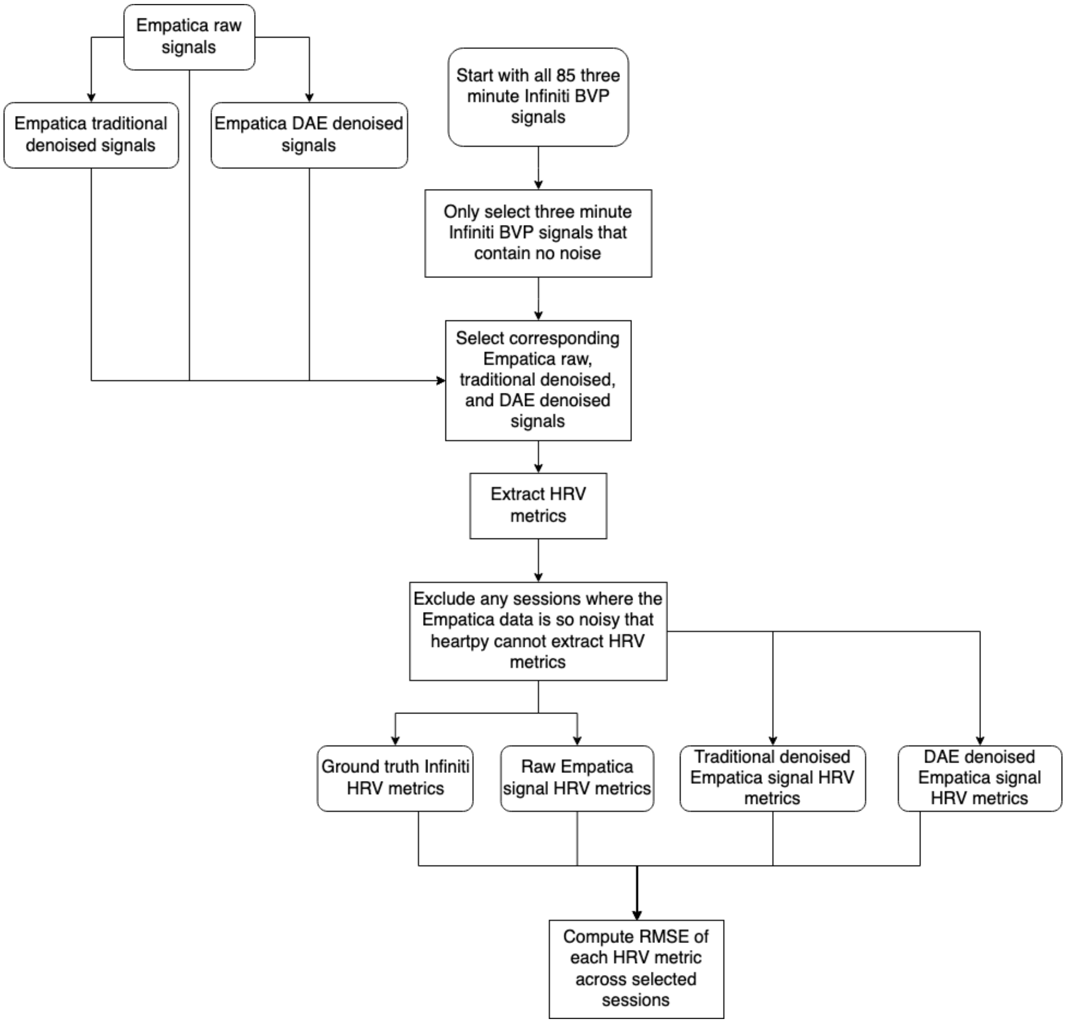
Empatica raw signals
Empatica traditional denoised signals
Empatica DAE denoised signals
Start with all 85 three minute Infiniti BVP signals
Only select three minute Infiniti BVP signals that contain no noise
Select corresponding Empatica raw, traditional denoised, and DAE denoised signals
Extract HRV metrics
Exclude any sessions where the Empatica data is so noisy that heartpy cannot extract HRV metrics
Ground truth Infiniti HRV metrics
Raw Empatica signal HRV metrics
Traditional denoised Empatica signal HRV metrics
DAE denoised Empatica signal HRV metrics
Compute RMSE of each HRV metric across selected sessions

Figure 4.16: Process to extract HRV metrics and compute RMSE.

	Empatica raw	Empatica traditional preprocessed	Empatica DAE denoised
HR	27.00	3.60	1.80
SDNN	47.20	29.40	18.60
RMSSD	89.30	51.20	40.60
pNN50	0.28	0.47	0.50
LF	487.00	1050.00	1.59E+18
HF	2330.00	1250.00	7.96E+18
LF/HF	2.29	3.01	3.24

Table 4.1: RMSE of each HRV metric. Comparing traditional preprocessing to DAE denoising.

It can be seen in table 4.1 that for HR, SDNN, and RMSSD, the DAE can denoise the Empatica PPG signal well. The DAE outperforms traditional signal processing for these HRV metrics. However, for the pNN50 and LF/HF HRV metrics, the DAE performs slightly worse than traditional preprocessing. Interestingly, for these two metrics, the HRV metrics extracted from the raw signal are actually closest to the "ground truth".

It should be noted that comparing the LF and HF band powers across preprocessing methods, through the RMSE, is not particularly useful. Denoising/filtering techniques such as bandpass filters can amplify frequencies within the passband for example. It is reasonable to however compare the relative power of the LF and HF bands (i.e. the LF/HF HRV metric). Therefore, the very large RMSE for the DAE for LF and HF is not significant as the LF/HF RMSE is much smaller. Moreover, LF/HF is a more useful feature for stress classification. [93] found that a high LF/HF ratio was closely related to mental stress.

4.5 Performance on Stress Classification Task

In figure 4.17, HR and all 6 HRV metrics are used as features for stress prediction. As might be expected, both traditional preprocessing and DAE denoising the Empatica signals before HRV extraction improves downstream stress classification. The DAE denoising improves stress classification even better than traditional preprocessing.

We tried combining HRV metrics extracted from the raw Empatica signal and the DAE denoised signal to see if that could push performance even higher. Bearing in mind the HRV RMSE from table 4.1, we saw that HR, SDNN, RMSSD were better extracted from DAE denoised signals, while pNN50, LF, HF, LF/HF were better extracted from the raw signal. Therfore, for our "combined" stress classification model, we used HR, SDNN, RMSSD extracted from the DAE denoised signal and pNN50, LF, HF, LF/HF from the raw Empatica signal. This improved stress classification performance for both F1 and MCC beyond that of the DAE denoised signal.

Most promising, the "combined" Empatica classifier even outperformed the raw Infiniti classifier for the MCC metric. The MCC metric has been shown to be more reliable compared to accuracy and F1 scores for binary classification [12, 13].

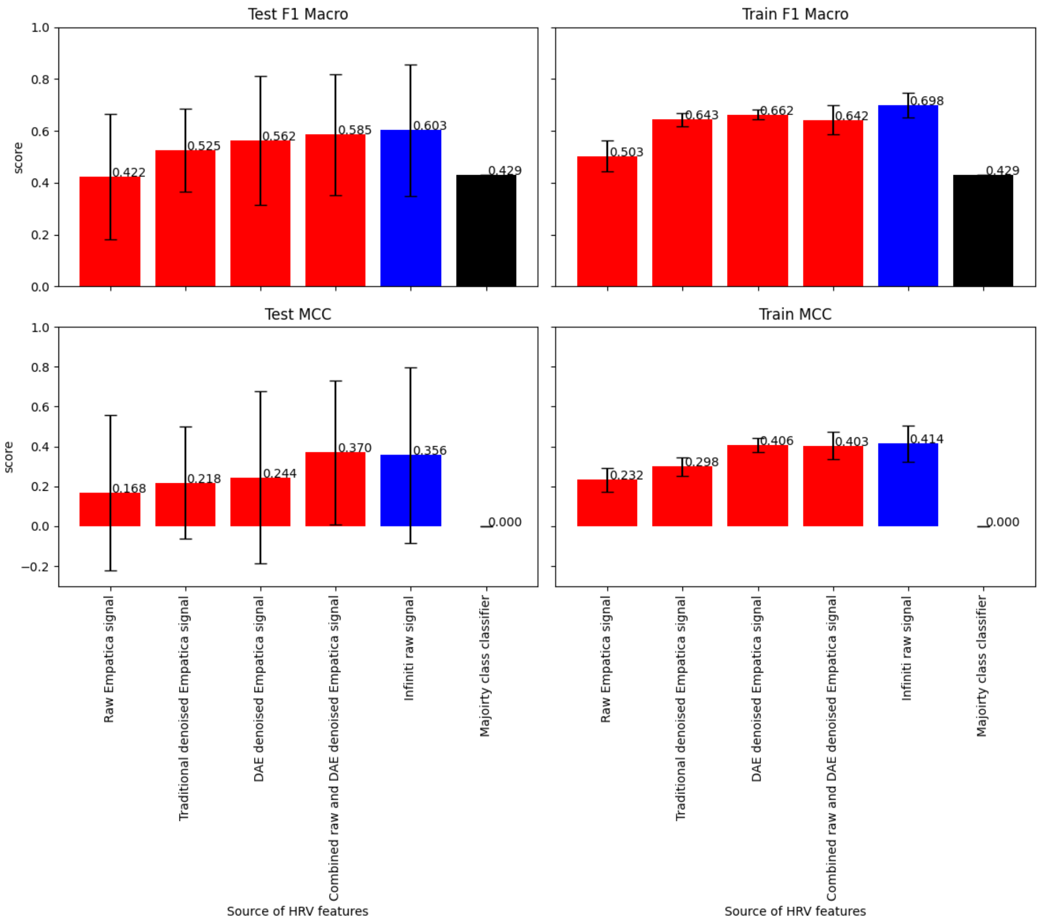

Figure 4.17: LOSO-CV evaluation of stress classification when HR and all 6 HRV metrics are used as features. As a reference, the “Majority class classifier” always predicts the most frequent label (low stress). Performance on the test set is shown on the left, while performance on the train set is shown on the right. The x-axis is: Raw Empatica signal, Traditional denoised Empatica signal, DAE denoised Empatica signal, Combined raw and DAE denoised Empatica signal, Infiniti raw signal, Majority class classifier. Error bars show one standard deviation.

As seen in table 4.1, the RMSE of the HF feature was high for HRV extracted from both the raw Empatica signal and the DAE denoised Empatica signal. Therefore, the HF feature was excluded and the stress classification experiment re-run with the new reduced feature set (results in figure 4.18). Dropping the HF feature improved combined performance for F1 macro but not MCC. However, dropping HF greatly increased DAE denoised MCC to the point that it exceeded the raw Infiniti reference. This non-clear cut behaviour only highlights that the DAE is particularly well suitable to improving certain HRV metrics and not others. Perhaps the DAE denoised HF feature is particularly poor, which is why performance increased by so much when the feature was excluded.

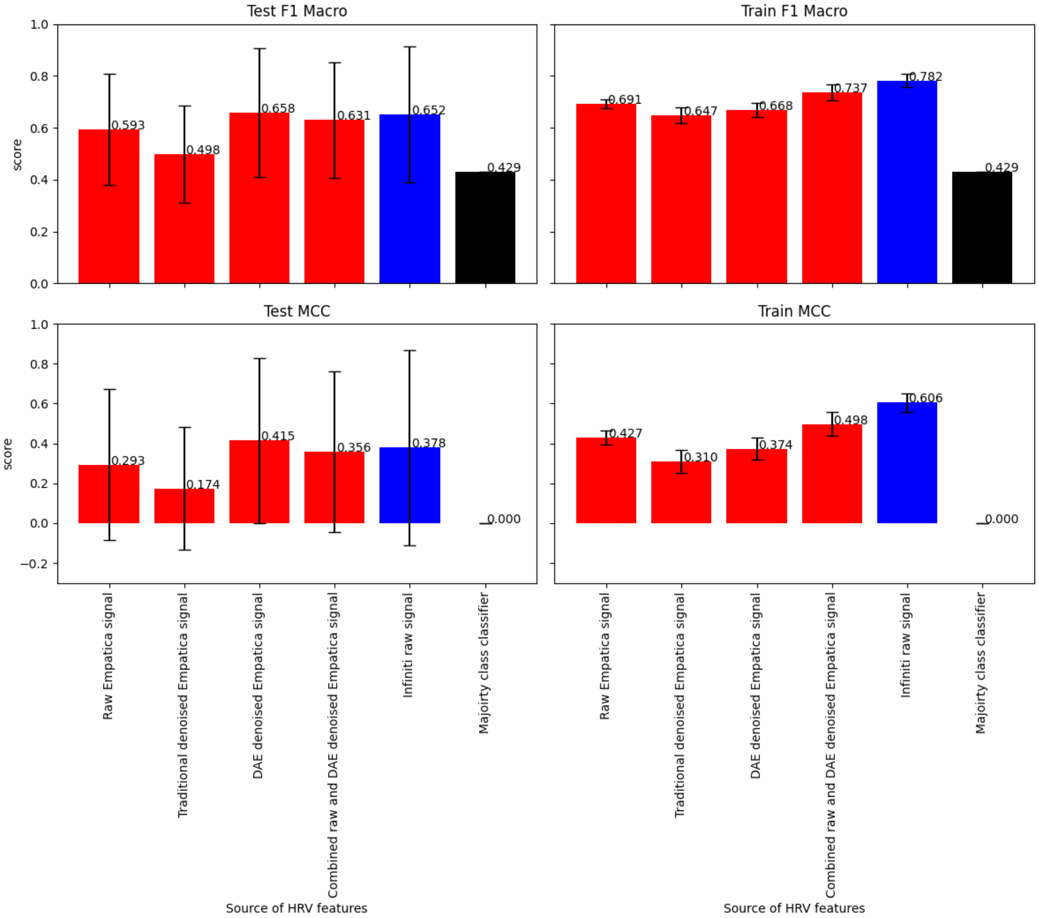

Figure 4.18: LOSO-CV evaluation of stress classification when the HF feature is excluded. As a reference, the "Majority class classifier" always predicts the most frequent label (low stress). Performance on the test set is shown on the left, while performance on the train set is shown on the right. The x-axis is: Raw Empatica signal, Traditional denoised Empatica signal, DAE denoised Empatica signal, Combined raw and DAE denoised Empatica signal, Infiniti raw signal, Majority class classifier. Error bars show one standard deviation.

Chapter 5

Discussion

In this chapter we discuss the aims of this thesis, the results of which were presented in chapter 4. Namely: i) evaluating the ability of DAEs to improve BVP signal quality; ii) evaluating accuracy of HRV metrics derived from DAE denoised signals; iii) LOSO-CV evaluation of stress classification following DAE denoising.

5.1 Denoising Quality Evaluation

5.1.1 Graphical Inspection of Denoising Quality

We used an DAE to denoise Empatica BVP signals. It is encouraging to see that the DAE can faithfully reconstruct clean signals. It is also promising that the DAE can recover peaks in noisy signals where traditional preprocessing fails to do so. It should be noted that the DAE often simplifies the original PPG waveform. Although it is able to recover peaks, it often loses more nuanced information such as the dicrotic notch (figure 5.1). This is not a problem in this current thesis because we focus on HRV metrics, which only depend on the time of heartbeats (i.e. peak locations).

However, these more nuanced waveform features can have other uses. For example the detection of the dicrotic notch can be used to calculate measures of arterial stiffness such as the large artery stiffness index [25, 65] and the augmentation index [68, 96]. However, PPG hardware is often designed for HR detection or pulse oximetry, and unfortunately may such lose nuanced PPG waveform features during in-built filtering and denoising. Consequently, an algorithm that can denoise while accentuating PPG waveform features would be particularly helpful for using PPG signals in personalized healthcare applications.

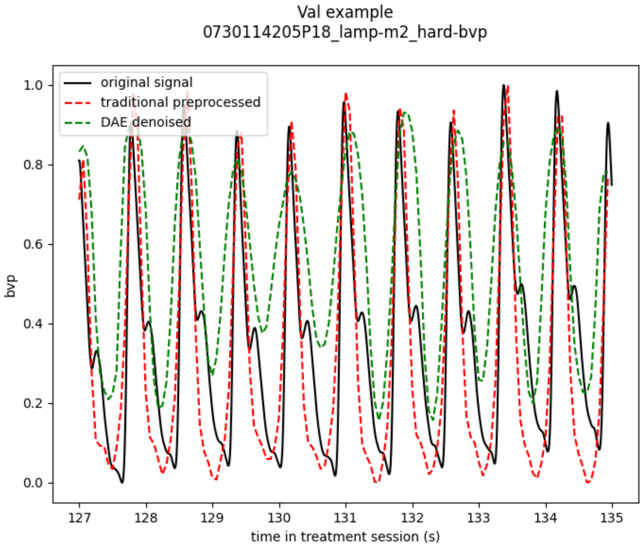

Figure 5.1: The DAE often loses the dicrotic notch when denoising.

5.1.2 Signal Quality Index evaluation of denoised signal

Both traditional preprocessing and the DAE were able to increase the pSQI of both clean and noisy PPG signals. This is expected in the traditional preprocessing case because the bandpass filter is explicitly designed to reject frequencies outside the range of interest. It is encouraging that the DAE was able to increase the pSQI more than traditional preprocessing. This may be because of the, deliberately, limited capacity of the DAE. The DAE was able to restore the primary HR signal but was unable to restore other less important frequencies in the PPG waveform. Additionally, the traditional preprocessing bandpass filter's range was 0.7Hz - 4.0Hz. This likely preserved some frequencies outside of the 0.8Hz to 2.0Hz range used for the pSQI calculation, which would decrease pSQI.

5.2 Aside: Transfer Learning

In the real world a myriad of different sensors are used to record physiological signals, including BVP. The Inifiti and Empatica sensors are just two examples. Sensors differ in many aspects, including: sensor placement (e.g. wrist vs fingerprint), signal acquisition protocol (e.g. transmittance vs reflection PPG), sampling rate, and any signal processing performed by the sensor itself. All these differences mean that the signals recorded by each sensor have their own characteristics, and are distinct from the signals recorded by other sensors. For a method applied to physiological signals to be very useful, it must be generalizable to different sensors and datasets. Unfortunately, most deep learning approaches in the literature have been unable to generalize to signals collected by different devices [34]. Therefore, it is encouraging that our method is generalizable and works across different sensors.

5.3 HRV Accuracy Evaluation of Denoised Signal

HR, SDNN, and RMSSD derived from DAE denoised Empatica signals were more accurate than those features derived from raw and traditional preprocessed Empatica signals. For pNN50 and LF/HF, traditional preprocessing was slightly better than DAE denoising. Interestingly, the raw Empatica signal resulted in the most accurate pNN50 and LF/HF features.

The signal peaks of the DAE denoised signal are i) rounded and ii) often slightly out of place compared to the true peak (see figure 5.2). Therefore, the peak locations calculated during HRV extraction may be slightly incorrect. This won't have a large impact on features such as HR, which considers the whole IBI sequence. However, for features like pNN50 which consider the individual differences between each successive IBI, these small disparities may have a significant effect. This may be the reason why the DAE performed better for HR and worse for pNN50. However, correcting this would be difficult. The mean absolute error was used as the loss function during model training. As long as the BVP peaks are somewhat in sync, this loss will be low. However, the specific location of the peak is not given any particularly great weight in this loss function. Additionally, rounded peaks, in the average, tend to perform better than sharply defined peaks for this loss function. The training of the DAE would need to be reconsidered to address these limitations.

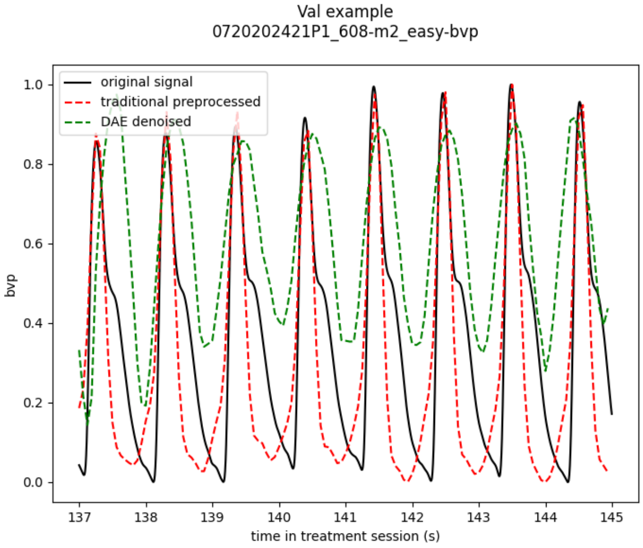

Figure 5.2: The peaks of the DAE denoised signal are i) rounded ii) often slightly out of sync with the true peaks.

5.4 Performance on Stress Classification task

Using HR and all 6 HRV metrics as features, the stress classifier trained on the DAE denoised signal outperforms the traditional denoised signal and raw signal for both F1 and MCC. Therefore, we have accomplished our aim of using DAE denoising to achieve better stress classification than when using traditional signal processing methods. The "combined" stress classification model is even more promising as it outperforms the raw Infiniti signal on the MCC. This further improved performance relates back to our findings in section 4.4 that DAE denoising helped improve some HRV metrics (HR, SDNN, RMSSD) but others were more accurate from the raw Empatica signal (pNN50, LF/HF).

Although finger sensors, like Infiniti, tend to be more accurate than wristband sensors like Empatica, they are not widely applicable in the wild. Therefore, achieving **better** stress classification on the Empatica signal than the Infiniti signal by using our DAE denoising module has very promising implications for future work on collecting PPG in natural settings.

By excluding the HF feature, our stress classifier trained on the DAE denoised signal outperforms classifiers trained on all other Empatica signals and on the raw Infiniti signal for both F1 and MCC. This feature selection result further highlights that our DAE denoising module can help in a stress classification framework.

5.5 Further work

Future work can try more advanced machine learning architectures for denoising (e.g. LSTM, bidirectional LSTM, CNN). PPG signals are cyclic and predictably sequential. For example, the diastolic peak (if present at all) always follows the systolic peak. A model that can understand these temporal dynamics (such as RNNs, and even 1-dimensional CNNs) would likely be better suited to denoising PPG signals than a simple MLP.

In this current work, a very simple GNB classifier was used for stress classification. A simple model was chosen just to demonstrate, as a proof of concept, the use of a DAE within our novel stress classification framework. Future work can use more advanced models (ANNs for example) to predict stress from HRV metrics.

In fact, stress could be predicted from the entire (denoised) signal itself, rather than just extracted HRV metrics. HRV metrics can grossly oversimplify the physiological nuances present in the BVP waveform [7, 50, 23]. This paradigm would make the denoising task very different as waveform features (e.g. dicrotic notch) could become important, while HRV metrics are based solely on peak locations. Conversely, other representations can be extracted from the PPG signal and then used for downstream stress classification. An example would be spectrograms, as used in [15], which are periodograms with an additional time axis (to show how the spectrum of frequencies of a signal changes over time). The possibilities are numerous, but the main focus for future work would be on the stress classification module.

Chapter 6

Conclusion

This thesis investigated if using a DAE to denoise PPG signals could improve automated stress classification performance. We found that our DAE improved the pSQI of BVP signals and the accuracy of the extracted HR, SDNN, and RMSSD HRV metrics. However, pNN50 and LF/HF HRV metrics extracted from the raw Empatica signal were more accurate than those extracted from the DAE denoised Empatica signal. We also manually annotated BVP signals as clean or noisy, and will be releasing our annotations to the research community.

Using HRV metrics extracted from DAE denoised Empatica PPG was found to improve participant-independent stress classification performance over using the raw Empatica PPG signal. To our knowledge, this is the first work to evaluate the use of DAE on improving HRV metrics and automatic stress classification. Of note, stress classification performance using DAE denoised Empatica signals exceeded that of raw Infiniti signals. This is particularly noteworthy because the Infiniti sensor is more accurate, however, it is far less usable as it must be attached to the finger. This contrasts with wristband sensors, such as Empatica, which can be used in far more scenarios. Therefore, it is encouraging that our denoising method can improve stress classification so significantly for widely applicable wristband sensors.

Additionally, our experiments were performed on PPG signals that were downsampled to 16Hz. Lower cost sensors often have such low sampling rates. The fact that our denoising technique works on low sampling rate signals is also promising for the use of low cost PPG sensors in the wild. Future work should look to using more advanced model architectures for the DAE, as well as more advanced models for automated stress classification.

Bibliography

[1] Ane Alberdi, Asier Aztiria, and Adrian Basarab. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of biomedical informatics*, 59:49–75, 2016.

[2] Karol Antczak. Deep recurrent neural networks for ecg signal denoising. *arXiv preprint arXiv:1807.11551*, 2018.

[3] H Harry Asada, Hong-Hui Jiang, and Peter Gibbs. Active noise cancellation using mems accelerometers for motion-tolerant wearable bio-sensors. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 2157–2160. IEEE, 2004.

[4] Les Atlas, Toshiteru Homma, and Robert Marks. An artificial neural network for spatio-temporal bipolar patterns: Application to phoneme classification. In *Neural Information Processing Systems*, pages 31–40, 1987.

[5] Brinnae Bent and Jessilyn P. Dunn. Optimizing sampling rate of wrist-worn optical sensors for physiologic monitoring. *Journal of Clinical and Translational Science*, 5(1):e34, 2021.

[6] George E Billman. Heart rate variability–a historical perspective. *Frontiers in physiology*, 2:86, 2011.

[7] George E Billman. The lf/hf ratio does not accurately measure cardiac sympatho-vagal balance. *Frontiers in physiology*, 4:26, 2013.

[8] PC Burns and TC Lansdown. E-distraction: the challenges for safe and usable internet services in vehicles, 2000.

[9] Szabolcs Béres and László Hejjel. The minimal sampling frequency of the photoplethysmogram for accurate pulse rate variability parameters in healthy volunteers. *Biomedical Signal Processing and Control*, 68:102589, July 2021.

[10] A John Camm, Marek Malik, J Thomas Bigger, Günter Breithardt, Sergio Cerutti, RJ Cohen, Philippe Coumel, EL Fallen, HL Kennedy, RE Kleiger, et al. 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. 1996.

[11] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, 4(4):195, 2018.

[12] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[13] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 14(1):1–22, 2021.

[14] Luca Chittaro. Anxiety induction in virtual environments: an experimental comparison of three general techniques. *Interacting with Computers*, 26(6):528–539, 2014.

[15] Youngjun Cho, Nadia Bianchi-Berthouze, and Simon J Julier. Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 456–463. IEEE, 2017.

[16] Youngjun Cho, Simon J Julier, and Nadia Bianchi-Berthouze. Instant stress: detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. *JMIR mental health*, 6(4):e10140, 2019.

[17] Youngjun Cho, Simon J Julier, Nicolai Marquardt, and Nadia Bianchi-Berthouze. Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging. *Biomedical optics express*, 8(10):4480–4503, 2017.

[18] A. Choi and H. Shin. Photoplethysmography sampling frequency: pilot assessment of how low can we go to analyze pulse rate variability with reliability? 38(3):586–600, February 2017. Publisher: IOP Publishing.

[19] Ahyoung Choi and Hangsik Shin. Photoplethysmography sampling frequency: pilot assessment of how low can we go to analyze pulse rate variability with reliability? *Physiological measurement*, 38(3):586, 2017.

[20] GD Clifford, J Behar, Q Li, and Iead Rezek. Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. *Physiological measurement*, 33(9):1419, 2012.

[21] Lisa J Crockett, John E Schulenberg, and Anne C Petersen. Congruence between objective and self-report data in a sample of young adolescents. *Journal of Adolescent Research*, 2(4):383–392, 1987.

[22] William R Dieter, Srabosti Datta, and Wong Key Kai. Power reduction by varying sampling rate. In *Proceedings of the 2005 international symposium on Low power electronics and design*, pages 227–232, 2005.

[23] Dwain L Eckberg. Sympathovagal balance: a critical appraisal. *Circulation*, 96(9):3224–3232, 1997.

[24] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996.

[25] Mohamed Elgendi. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews*, 8(1):14–25, 2012.

[26] Mohamed Elgendi. Optimal signal quality index for photoplethysmogram signals. *Bioengineering*, 3(4):21, 2016.

[27] Empatica. E4 data - bvp expected signal - empatica support.

[28] George S Everly, Jeffrey M Lating, and Melvin A Gravitz. *A clinical guide to the treatment of the human stress response.* Springer, 2002.

[29] David J Ewing, Christopher N Martyn, Robert J Young, and Basil F Clarke. The value of cardiovascular autonomic function tests: 10 years experience in diabetes. *Diabetes care*, 8(5):491–498, 1985.

[30] Quek Kia Fatt and Amutha Ramadas. The usefulness and challenges of big data in healthcare. *Journal of Healthcare Communications*, 3(2):21, 2018.

[31] Mark A. Finlayson. *ProComp Infiniti Hardware Manual.* Thought Technology Ltd.

[32] Jong Yong A Foo. Comparison of wavelet transformation and adaptive filtering in restoring artefact-induced time-related measurement. *Biomedical signal processing and control*, 1(1):93–98, 2006.

[33] Daisuke Fujita and Arata Suzuki. Evaluation of the Possible Use of PPG Waveform Features Measured at Low Sampling Rate. *IEEE Access*, 7:58361–58367, 2019. Conference Name: IEEE Access.

[34] Nagarajan Ganapathy, Ramakrishnan Swaminathan, and Thomas Deserno. Deep Learning on 1-D Biosignals: a Taxonomy-based Survey. *Yearbook of Medical Informatics*, 27(01):098–109, August 2018.

[35] Rolandas Gircys, Agnius Liutkevicius, Egidijus Kazanavicius, Vita Lesauskaite, Gyte Damuleviciene, and Audrone Janaviciute. Photoplethysmography-based continuous systolic blood pressure estimation method for low processing power wearable devices. *Applied Sciences*, 9(11):2236, 2019.

[36] Juliet Hassard, Kevin Teoh, Tom Cox, M Cosmar, R Gründler, D Flemming, B Cosemans, and K Van den Broek. Calculating the cost of work-related stress and psychosocial risks. 2014.

[37] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

[38] Javier Hernandez, Pablo Paredes, Asta Roseway, and Mary Czerwinski. Under pressure: sensing stress of computer users. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 51–60, 2014.

[39] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, 92(1):84–89, 2004.

[40] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[41] MC Horton and RJ Wenzel. The digital elliptic filter-a compact sharp-cutoff design for wide bandstop or bandpass requirements. *IEEE transactions on Microwave theory and techniques*, 15(5):307–314, 1967.

[42] Elite HRV. What are hrv frequency measurements (lf, hf, lf/hf).

[43] O ISK. stress at work—facts and figures.

[44] Alexey Grigorevich Ivakhnenko. The group method of data of handling; a rival of the method of stochastic approximation. *Soviet Automatic Control*, 13:43–55, 1968.

[45] Alexey Grigorevich Ivakhnenko. Polynomial theory of complex systems. *IEEE transactions on Systems, Man, and Cybernetics*, (4):364–378, 1971.

[46] V Jayasree, T Sandhya, and P Radhakrishnan. Non-invasive studies on age related parameters using a blood volume pulse sensor. *Measurement science review*, 8(4):82, 2008.

[47] Ákos Jobbágy, Miklós Majnár, Lilla K Tóth, and Péter Nagy. Hrv-based stress level assessment using very short recordings. *Periodica Polytechnica Electrical Engineering and Computer Science*, 61(3):238–245, 2017.

[48] Jeremy Jordan. Introduction to autoencoders., Mar 2018.

[49] Greeshma Joseph, Almaria Joseph, Geevarghese Titus, Rintu Mariya Thomas, and Dency Jose. Photoplethysmogram (ppg) signal analysis and wavelet de-noising. In *2014 annual international conference on emerging research areas: Magnetics, machines and drives (AICERA/iCMMD)*, pages 1–5. IEEE, 2014.

[50] John M Karemaker. Heart rate variability: why do spectral analysis? *Heart*, 77(2):99, 1997.

[51] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry investigation*, 15(3):235, 2018.

[52] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[53] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[54] Rajet Krishnan, Balasubramaniam Natarajan, and Steve Warren. Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data. *IEEE transactions on biomedical engineering*, 57(8):1867–1876, 2010.

[55] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. Distanceppg: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5):1565–1588, 2015.

[56] Hindra Kurniawan, Alexandr V Maslov, and Mykola Pechenizkiy. Stress detection from speech and galvanic skin response signals. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 209–214. IEEE, 2013.

[57] Kalliopi Kyriakou, Bernd Resch, Günther Sagl, Andreas Petutschnig, Christian Werner, David Niederseer, Michael Liedlgruber, Frank H Wilhelm, Tess Osborne, and Jessica Pykett. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors*, 19(17):3805, 2019.

[58] CM Lee and Yuan Ting Zhang. Reduction of motion artifacts from photoplethysmographic recordings using a wavelet denoising approach. In *IEEE EMBS Asian-Pacific Conference on Biomedical Engineering, 2003.*, pages 194–195. IEEE, 2003.

[59] Joonnyong Lee, Sukkyu Sun, Seung Man Yang, Jang Jay Sohn, Jonghyun Park, Saram Lee, and Hee Chan Kim. Bidirectional recurrent auto-encoder for photoplethysmogram denoising. *IEEE journal of biomedical and health informatics*, 23(6):2375–2385, 2018.

[60] Chao Li, Jing Zhai, and Armando Barreto. Signal processing quantification of changes in the blood volume pulse (bvp) waveform due to exercise. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, volume 4, pages 3180–3183. IEEE, 2003.

[61] Fenghua Li, Peida Xu, Shichun Zheng, Wenfeng Chen, Yang Yan, Suo Lu, and Zhengkui Liu. Photoplethysmography based psychological stress detection with pulse rate variability feature differences and elastic net. *International Journal of Distributed Sensor Networks*, 14(9):1550147718803298, 2018.

[62] J Li, M Bhuiyan, X Huang, B McDonald, Todd Farrell, and EA Clancy. Reducing electric power consumption when transmitting ecg/emg/eeg using a bluetooth low energy microcontroller. In *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–3. IEEE, 2018.

[63] Suyi Li, Lijia Liu, Jiang Wu, Bingyi Tang, and Dongsheng Li. Comparison and noise suppression of the transmitted and reflected photoplethysmography signals. *BioMed research international*, 2018, 2018.

[64] Rollin McCraty and Fred Shaffer. Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. *Global advances in health and medicine*, 4(1):46–61, 2015.

[65] Sandrine C Millasseau, RP Kelly, JM Ritter, and PJ Chowienczyk. Determination of age-related increases in large artery stiffness by digital pulse contour analysis. *Clinical science*, 103(4):371–377, 2002.

[66] LB Minor. Harnessing the power of data in health. *Stanford Med. Heal. Trends Rep.*, 2017.

[67] P Madhan Mohan, V Nagarajan, and Sounak Ranjan Das. Stress measurement from wearable photoplethysmographic sensor using heart rate variability data. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, pages 1141–1144. IEEE, 2016.

[68] JOSEPH P Murgo, NICO Westerhof, JOHN P Giolma, and STEPHEN A Altobelli. Aortic input impedance in normal man: relationship to pressure wave forms. *Circulation*, 62(1):105–116, 1980.

[69] Navaneet K Lakshminarasimha Murthy, Pavan C Madhusudana, Pradyumna Suresha, Vijitha Periyasamy, and Prasanta Kumar Ghosh. Multiple spectral peak tracking for heart rate monitoring from photoplethysmography signal during intensive physical exercise. *IEEE Signal Processing Letters*, 22(12):2391–2395, 2015.

[70] Karan Nair, Janhavi Kulkarni, Mansi Warde, Zalak Dave, Vedashree Rawalgaonkar, Ganesh Gore, and Jonathan Joshi. Optimizing power consumption in iot based wireless sensor networks using bluetooth low energy. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 589–593. IEEE, 2015.

[71] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

[72] Massimo Pagani, Federico Lombardi, Stefano Guzzetti, Ornella Rimoldi, Raffaello Furlan, Paolo Pizzinelli, Giulia Sandrone, Gabriella Malfatto, Simonetta Dell'Orto, and Emanuela Piccaluga. Power spectral analysis of heart rate and arterial pressure variabilities as a marker of sympatho-vagal interaction in man and conscious dog. *Circulation research*, 59(2):178–193, 1986.

[73] S Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

[74] Klemen Peternel, Matevž Pogačnik, Rudi Tavčar, and Andrej Kos. A presence-based context-aware chronic stress recognition system. *Sensors*, 12(11):15888–15906, 2012.

[75] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010.

[76] Ming-Zher Poh, Yukkee Cheung Poh, Pak-Hei Chan, Chun-Ka Wong, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Daniel Wai-Sing Chu, and Chung-Wah Siu. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. *Heart*, 104(23):1921–1928, 2018.

[77] Nora Ptakauskaite, Anna L Cox, and Nadia Berthouze. Knowing what you're doing or knowing what to do: how stress management apps support reflection and behaviour change.

In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018.

[78] M Raghuram, K Venu Madhav, E Hari Krishna, and K Ashoka Reddy. Evaluation of wavelets for reduction of motion artifacts in photoplethysmographic signals. In *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*, pages 460–463. IEEE, 2010.

[79] M Raghu Ram, K Venu Madhav, E Hari Krishna, Nagarjuna Reddy Komalla, and K Ashoka Reddy. A novel approach for motion artifact reduction in ppg signals based on as-lms adaptive filter. *IEEE Transactions on Instrumentation and Measurement*, 61(5):1445–1457, 2011.

[80] K Ashoka Reddy, Boby George, and V Jagadeesh Kumar. Use of fourier series analysis for motion artifact reduction and data compression of photoplethysmographic signals. *IEEE transactions on instrumentation and measurement*, 58(5):1706–1711, 2008.

[81] Rui Rodrigues and Paula Couto. A neural network approach to ecg denoising. *arXiv preprint arXiv:1212.5217*, 2012.

[82] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[83] Monalisa Singha Roy, Rajarshi Gupta, Jayanta K Chandra, Kaushik Das Sharma, and Arunansu Talukdar. Improving photoplethysmographic measurements under motion artifacts using artificial neural network for personal healthcare. *IEEE Transactions on Instrumentation and Measurement*, 67(12):2820–2829, 2018.

[84] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[85] Jade Savage. Automatic stress recognition using plethysmographic signals obtained from webcam video images. Master's thesis, 2018.

[86] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[87] Silvia Serino, Pietro Cipresso, Andrea Gaggioli, Federica Pallavicini, Sergio Cipresso, Danilo Campanaro, and Giuseppe Riva. Smartphone for self-management of psychological stress: a preliminary evaluation of positive technology app. *Revista de Psicopatología y Psicología Clínica*, 19(3):253–260, 2014.

[88] Mert Sevil, Mudassir Rashid, Mohammad Reza Askari, Sediqeh Samadi, Iman Hajizadeh, and Ali Cinar. Psychological stress detection using photoplethysmography. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

[89] Fred Shaffer and Jay P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, page 258, 2017.

[90] Fred Shaffer, Rollin McCraty, and Christopher L Zerr. A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Frontiers in psychology*, 5:1040, 2014.

[91] Shota Shimazaki, Shoaib Bhuiyan, Haruki Kawanaka, and Koji Oguri. Features extraction for cuffless blood pressure estimation by autoencoder from photoplethysmography. In *2018 40Th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2857–2860. IEEE, 2018.

[92] Riccardo Sioni and Luca Chittaro. Stress detection using physiological sensors. *Computer*, 48(10):26–33, 2015.

[93] RP Sloan, PA Shapiro, E Bagiella, SM Boni, M Paik, JT Bigger Jr, RC Steinman, and JM Gorman. Effect of mental stress throughout the day on cardiac autonomic control. *Biological psychology*, 37(2):89–99, 1994.

[94] Weitong Sun, Yuping Su, Xia Wu, and Xiaojun Wu. A novel end-to-end 1d-rescnn model to remove artifact from eeg signals. *Neurocomputing*, 404:108–121, 2020.

[95] Nina Sviridova and Kenshi Sakai. Human photoplethysmogram: new insight into chaotic characteristics. *Chaos, Solitons & Fractals*, 77:53–63, 2015.

[96] Kenji Takazawa, Nobuhiro Tanaka, Masami Fujita, Osamu Matsuoka, Tokuyu Saiki, Masaru Aikawa, Sinobu Tamura, and Chiharu Ibukiyama. Assessment of vasoactive agents and vascular aging by the second derivative of photoplethysmogram waveform. *Hypertension*, 32(2):365–370, 1998.

[97] Toshiyo Tamura, Yuka Maeda, Masaki Sekine, and Masaki Yoshida. Wearable photoplethysmographic sensors—past and present. *Electronics*, 3(2):282–302, 2014.

[98] Patrick Thiam, Hans A Kestler, and Friedhelm Schwenker. Multimodal deep denoising convolutional autoencoders for pain intensity classification based on physiological signals. In *ICPRAM*, pages 289–296, 2020.

[99] Bill Tolson. Where should healthcare data be stored in 2018: and beyond. *Health IT Outcomes (https://www. healthitoutcomes. com/doc/where-should-healthcare-data-be-stored-in-and-beyond-0001)*, 2018.

[100] Jessica Torres-Soto and Euan A Ashley. Multi-task deep learning for cardiac rhythm detection in wearable devices. *NPJ digital medicine*, 3(1):1–8, 2020.

[101] H. Tsuji, F. J. Venditti, E. S. Manders, J. C. Evans, M. G. Larson, C. L. Feldman, and D. Levy. Determinants of heart rate variability. *Journal of the American College of Cardiology*, 28(6):1539–1546, November 1996.

[102] Hisako Tsuji, Ferdinand J Venditti Jr, Emily S Manders, Jane C Evans, Martin G Larson, Charles L Feldman, and Daniel Levy. Determinants of heart rate variability. *Journal of the American College of Cardiology*, 28(6):1539–1546, 1996.

[103] Paul van Gent, Haneen Farah, N Nes, and Bart van Arem. Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data. In *Proceedings of the 6th HUMANIST Conference*, pages 173–178, 2018.

[104] Paul van Gent, Haneen Farah, Nicole Nes, and B. Arem. Heart Rate Analysis for Human Factors: Development and Validation of an Open Source Toolkit for Noisy Naturalistic Heart Rate Data. June 2018.

[105] A Voss, N Wessel, A Sander, H Malberg, and R Dietz. Influence of low sampling rate on heart rate variability analysis based on non-linear dynamics. In *Computers in Cardiology 1995*, pages 689–692. IEEE, 1995.

[106] Jacqueline Wijsman, Bernard Grundlehner, Hao Liu, Julien Penders, and Hermie Hermens. Wearable physiological sensors reflect mental stress state in office-like situations. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 600–605. IEEE, 2013.

[107] Aaron Williamon, Lisa Aufegger, David Wasley, David Looney, and Danilo P Mandic. Complexity of physiological responses decreases in high-stress musical performance. *Journal of The Royal Society Interface*, 10(89):20130719, 2013.

[108] MM Wolf, GA Varigos, D Hunt, and JG Sloman. Sinus arrhythmia in acute myocardial infarction. *Medical Journal of Australia*, 2(2):52–53, 1978.

[109] Peng Xiong, Hongrui Wang, Ming Liu, Suiping Zhou, Zengguang Hou, and Xiuling Liu. Ecg signal enhancement based on improved denoising auto-encoder. *Engineering Applications of Artificial Intelligence*, 52:194–202, 2016.

[110] Ganggang Xu and Jianhua Z Huang. Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. *The Annals of Statistics*, 40(6):3003–3030, 2012.

[111] Banghua Yang, Kaiwen Duan, Chengcheng Fan, Chenxiao Hu, and Jinlong Wang. Automatic ocular artifacts removal in eeg using deep learning. *Biomedical Signal Processing and Control*, 43:148–158, 2018.

[112] Yangsong Zhang, Benyuan Liu, and Zhilin Zhang. Combining ensemble empirical mode decomposition with spectrum subtraction technique for heart rate monitoring using wrist-type photoplethysmography. *Biomedical Signal Processing and Control*, 21:119–125, 2015.

[113] Zhilin Zhang, Zhouyue Pi, and Benyuan Liu. Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on biomedical engineering*, 62(2):522–531, 2014.