TAASR: Temporally-Aware Affective State Recognition with Attention-Augmented CNNs

by

Daniel Abramow

(Under the Direction of Tianming Liu)

Abstract

Affective state recognition (ASR) involves using the body's physiological signals to extract useful information about ones mental state. ASR systems are often implemented in highly controlled environments with cumbersome chest sensors and intrusive facial expression monitoring setups, making it difficult to translate their performance to real environments. Recently, the widespread adoption of wrist-worn wearables has highlighted a need for further research into practical ASR with commercially available devices. In this paper, we propose TAASR, an InceptionTime based end-to-end learning architecture augmented with channel attention and global feature fusion for three-class ASR (*baseline* vs. *stressed* vs. *amused*), and TAASR-MT, a multi-task version of TAASR that uses mental health self-assessments to improve basic ASR performance. For practicality, we train these architectures primarily with wrist-based signals and report a best classification accuracy and F_1-score of 81.16% and 70.02, demonstrating noticeable improvements upon InceptionTime and prior works that employ simpler classification approaches.

Index words: Affective State Recognition, Mental Health, Wearables, Time Series Classification, Machine Learning, Deep Learning, Convolutional Neural Networks

TAASR: Temporally-Aware Affective State Recognition with Attention-Augmented CNNs

by

Daniel Abramow

B.S., University of Georgia, 2021

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Master of Science

Athens, Georgia

2023

TAASR: Temporally-Aware Affective State Recognition with Attention-Augmented CNNs

by

Daniel Abramow

Major Professor: Tianming Liu

Committee: Jaewoo Lee
Jennifer Gay

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2023

Acknowledgments

I would like to begin by thanking Dr. Tianming Liu for his constant support and guidance throughout my time at the University of Georgia Institute for Artificial Intelligence. I would also like to express gratitude to Dr. Jennifer Gay and Dr. Jaewoo Lee for serving on my advisory committee - this paper could not have been completed without their valuable knowledge and expertise.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In recent years, many have demonstrated interest in wearable devices that shrink the form factor of a smartphone down to that of a wrist watch. In fact, between 2016 and 2019, the number of wearables in circulation nearly doubled from 325 million to 722 million, and it is forecast that this total surpassed one billion devices at the end of 2022 (Laricchia, 2022). Adapa et al. (2018) attribute the rapid adoption of wrist-worn devices to their overall usability, water resistance, battery life, technological novelty, and most importantly, the availability of useful fitness tracking features. Common fitness tracking applications that work in tandem with smartwatches include: irregular heartbeat detection,[1] sleep cycle tracking and identification,[2] and activity recognition.[3] Many of these applications rely on machine learning models and are continually being improved through industry research.

Today, commercially available wearables allow for the accurate measurement and analysis of a wide variety of physiological signals that were once exclusively monitored in a laboratory setting. While many use these devices for their intuitive form factor, look, and feel, wrist-worn fitness trackers can serve as a motivational tool for those that wish to improve their quality of life via exercise and positive habit building (Lyons et al., 2014). Furthermore, devices like the *Apple Watch Series 7* and *FitBit Versa* track their user's activities, daily step count, heart rate, and temperature, enabling the average consumer to quantify their lifestyle choices and overall physical health on a daily basis.

Many have developed wearables that monitor sleep quality and cardiovascular health, whereas few have explored the idea of a device or software solution that monitors mental health. Since the beginning of the COVID-19 pandemic, reports of depression and anxiety increased at alarming rates (Hayward, 2022). It is essential that recent increases in instances of anxiety and depression, as well as future influxes of mental health disorders, be addressed on a broad scale.

[1] Apple provides FDA approved, native support for irregular heartbeat detection with the *Apple Watch*.

[2] For a list of smartwatch enabled sleep tracking applications, visit https://www.nytimes.com/wirecutter/reviews/best-sleep-tracking-app/

[3] A third-party application for FitBit activity recognition is detailed at https://github.com/andresquintanilla/fitbit-activity-recognition.

Currently, psychiatrists treat mental health issues with prescription medications and cognitive behavioral therapy (CBT), but many do not have the resources to seek professional help.[4] Still, there are many approaches to treating and identifying mental health disorders that are not reliant on access to medical professionals. The practice of mindfulness is a popular remedy for depression and anxiety (Staff, 2020), and research suggests that signals like heart rate variance (HRV) and respiration rate (Zhus et al., 2019) can be sound predictors of ones mood. Importantly, the physiological signals that directly respond to emotional changes can be easily monitored on everyday fitness trackers, chest sensors, and smartwatches, creating opportunities for the development of intuitive systems that can monitor the characteristics of ones affective state.

1.2 Objective

This paper aims to demonstrate that signals collected on widely available wearables can be used to help identify discrete emotional states. To accomplish this, we train various machine learning models on multimodal physiological data and evaluate their results based on accuracy and F_1-score in an effort to discover novel ASR approaches. Moreover, this body introduces the Temporally-Aware Affective State Recognition architecture (TAASR), an InceptionTime based (Fawaz et al., 2019) classification model for practical affective state recognition (ASR), and explores multi-task learning with discretized mental health self-assessments to augment our defined ASR objective. We also conduct a series of experiments to validate the inclusion of specific modules to our proposed architecture in an effort to justify TAASR as a viable ASR framework.

The next chapter of this paper covers mental health self-assessments and models of emotion, the intuition behind using physiological signals to identify emotional states, and approaches to affective state recognition with machine learning. Future chapters thoroughly explore the following:

- a multimodal data set introduced by Schmidt et al. (2018) for Wearable Stress and Affect Detection (WESAD)
- WESAD based affective state recognition tasks and learning algorithms that can perform them
- TAASR and its individual components
- implications and ethical concerns regarding automatic emotion classification systems and future research directions

[4] CBT is a type of question-and-answer based talking therapy that alleviates the symptoms of mental health disorders. It works by helping patients to learn healthy habits and thought patterns to better cope with stressful circumstances.

Chapter 2

Background

2.1 Models of Emotion and Mental Health Self-Assessments

Physicians often administer mental health self-assessments to their patients to help identify possible symptoms of mental health disorders. These assessments are crucial to modern psychiatric care, as emotions and mental health conditions often present themselves uniquely across individuals. By using an established set of questions centered around standardized measures of emotion, mental health self-assessments enable the medical community to evaluate mental health on a generalized scale, encouraging further research into emotion recognition and overall well-being.

One such model of emotion - the circumplex model (Russell, 1980) - formulates emotion as a two dimensional plane, where the x-axis (valence) represents an emotion's degree of positivty or negativity, and the y-axis represents an emotion's magnitude or severity (arousal) (Bestelmeyer et al., 2017). There

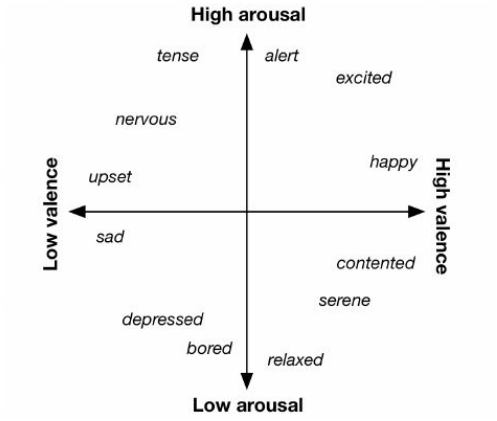

Figure 2.1: Circumplex model of emotion (Source: Russell, 1980)

are several dimensional abstractions of emotion, but the circumplex model, among others, is considered

to be the most widely accepted (Rubin and Talarico, 2009).[1] Furthermore, quantitative representations of emotion have inspired a wide array of mental health self-assessments, such as the Self-Assessment Manikin (SAM) and the Positive Affect - Negative Affect Schedule (PANAS), which place individuals on the circumplex plane and its variants.

Developed in 1994 by Bradley and Lang, the SAM asks subjects to simply rate their levels of valence, arousal, and dominance (level of control over an emotion) on a scale of 1 to 5. The results of this assessment can be visualized via the circumplex model, or its modified counterparts that handle more than two dimensions. Similarly, the PANAS assessment asks subjects to rate the degree to which they are experiencing a variety of emotions (e.g., interested, distressed, irritable, and active) from *very slightly* or *not at all* to *extremely*, and the aggregated results of positive and negative questions are used to place subjects on each axis of the Positive Affect - Negative Affect scale (Watson et al., 1988). [2]

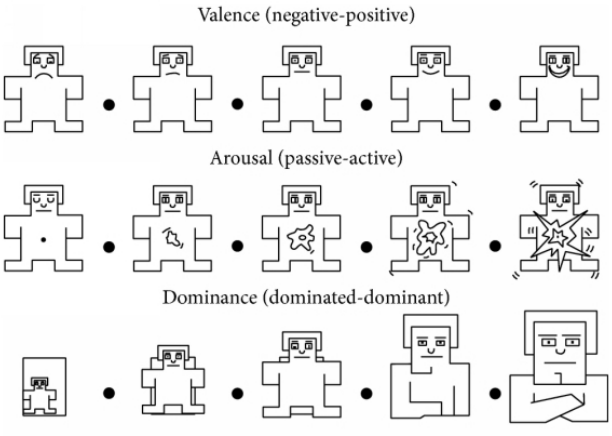

Figure 2.2: Self-Assessment Manikin scale (Source: Bradley and Lang, 1994)

Other questionnaire based self-assessments, like the State-Trait Anxiety Inventory (STAI), aim to evaluate a subject's anxiousness via personality traits and their ability to handle stressful situations (Bieling et al., 1998). Subjects rate statements like "I feel jittery" and "I am worried" from *very slightly* or *not at all* to *extremely*, and the aggregate score of the examination is used to describe the magnitude of their anxiety. To gather meaningful data regarding a patient's state and trait level anxiety, the STAI is often administered after the Trier Social Stress Test (TSST), an interview style examination where subjects are probed with unseen questions and challenging mental math problems (Allen et al., 2017).

2.2 Physiological Signals as Predictors of Emotion

The ability to systematically classify emotions from physiological signals is reliant on the assumption that emotions manifest themselves through interactions between human cognition and the autonomic nervous

[1]The circumplex model of emotion serves as a foundation for dimensional conceptualizations of emotion. The vector and PANAS models introduce slight variations to the circumplex model, such as additional dimensions to the circumplex plane.

[2]The Likert scale is commonly used for measuring strength of agreement.

system (ANS) - an idea thoroughly supported by the Schachter-Singer two-factor theory of emotion. This theory posits that external factors directly influence heart rate, respiration rate, and perspiration, which the brain then associates with feelings of fear, happiness, etc. (Schachter and Singer, 1962). Further, Schachter and Singer assert that our physiological reactions to stimuli and what we associate these reactions with (e.g., danger, reward) are the foundations of discrete emotional states. Moreover, the ANS plays a critical role in generating physiological responses aligned with unique emotional experiences (Waxenbaum et al., 2021), creating a direct link between environmental factors and physiological indicators of emotion.

Recent studies have attempted to identify which physiological modalities are most associated with changes in moods, such as stress and calmness. In general, a wide array of research suggests that heart rate variance is closely associated with emotion regulation (Kim et al., 2018). HRV measures the variation of time between heartbeats - a behavior directly influenced by the vagus nerve and ANS (Christodoulou et al., 2020). While a high HRV is often considered to be a sign of overall physical and mental health, and

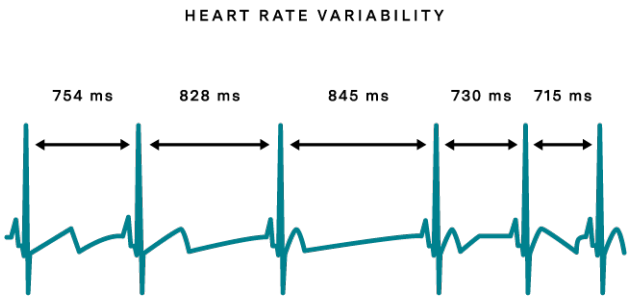

Figure 2.3: A visual representation of HRV (Source: Georgallides, 2021)

a low HRV is considered to be an indicator of stress and emotional turbulence, further research affirms that HRV is a dynamic measure that cannot purely be interpreted via magnitude (Shaffer and Ginsberg, 2017). For example, a high HRV can sometimes indicate cardiac abnormalities. Nevertheless, individuals who engage in mindfulness based intervention (MBI) (a meditation centered mental health treatment) tend to record HRV measurements more closely associated with emotional stability and effective ANS regulation (Christodoulou et al., 2020). Aside from electrocardiogram (ECG) related measures like HRV, electrodermal activity (EDA), blood volume pulse (BVP), and skin temperature and moisture are all areas of focus when it comes to drawing correlations between the mind and body (Karthikeyan et al., 2011).

2.3 Machine Learning for Time Series Analysis

Simply put, time series analysis covers statistical methodologies that aims to extract valuable insights from sequential data. There exist many machine learning methods that model time series data with the goal of forecasting future data points (e.g., stock market prediction and weather forecasting), or classifying windows of time into discrete categories (e.g., activity recognition and anomaly detection). Time series

classification (TSC) approaches can be broadly categorized into two types: similarity based and learning based. The former involves projecting sequences into a vector space and using supervised or unsupervised classification to identify similarities and differences among them. On the other hand, the latter utilizes more complex techniques, such as machine learning and optimization based algorithms. One possible similarity based TSC approach is to classify time windows by their k-nearest neighbors according to Dynamic Time Warping (DTM)[3] (Nurwanto et al., 2016).

Additionally, Traditional linear models (e.g., logistic regression) can be used to categorize individual time sequences from hand-crafted extracted features (mean, minimum, maximum, etc.). Significant improvements to the performance of linear models are generally observed when using decision trees and random forests, but gradient boosted trees often prove to be more effective than most classical machine learning architectures on a variety of time-series based tasks (Gertz et al., 2020). Still, these models have little capability to extract meaningful relationships from complex sequences as they lack any sort of recurrence mechanism.

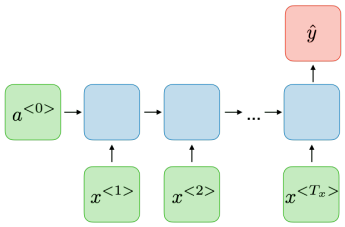

Figure 2.4: RNN for time series classification (Source: Amidi and Amidi, 2019)

In recent years, recurrent neural networks (RNN) have achieved state-of-the-art performance across many tasks in several domains for time series analysis. By allowing previous outputs to influence the network's current state, RNNs maintain an "internal memory" that is very powerful when it comes to understanding sequential data. For example, in a many-to-one RNN (see Figure 2.4), n time steps are used to generate a single output value; this is a common setup for classification and regression tasks. Alternatively, many-to-many RNNs can be used for image captioning and machine translation, though transformers are generally preferred for these types of problems (Lakew et al., 2018).

[3]DTM measures the similarity between time series of varying lengths and translation as the Euclidean distance between their optimal alignment (Müller, 2007)

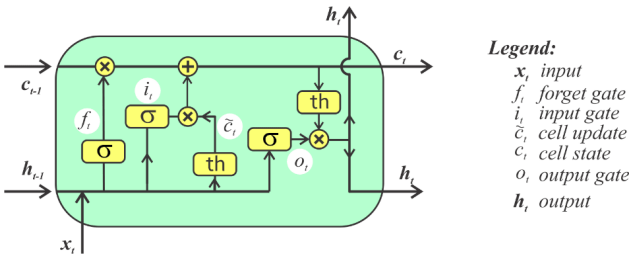

Figure 2.5: LSTM cell (Source: Hrnjica, 2019). The forget gate discards information from the cell state, the input gate adds information to the cell state, and the output gate regulates the flow of information from cell memory.

A popular and more sophisticated adaptation to the vanilla RNN is the Long Short-Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997). To combat vanishing and exploding gradients and the vanilla RNN architecture's inability to handle lengthy input sequences, LSTM networks incorporate *gates* that regulate the flow of information throughout training. In short, gates allow LSTM networks to discard less important time sequences from their internal memory, and update them with more influential values as they are observed.

With deep learning architectures, recurrence is not always needed to produce meaningful classification results. One such non-recurrent architecture, InceptionTime (Fawaz et al., 2019), has the ability to extract temporal features over multiple time frames with an ensemble of one-dimensional convolutional neural networks (CNN). InceptionTime is less computationally expensive to train and more performant than state-of-the-art similarity based time series classifiers (e.g., HIVE-COTE). Some TSC architectures further boost their performance by combining both LSTM and convolutional layers with one another.

There also exists a wide variety of self-attention based models (e.g., TapNet) that leverage transformers for TSC (Yoon et al., 2019). While transformers have demonstrated their strengths in natural language processing and sequence-to-sequence translation, their use for time series analysis is a blossoming area of research, and it is unclear if they can reliably surpass the performance of RNNs and CNNs. Currently, data requirements and computational bottlenecks (e.g., quadratic time complexity of self-attention) typically make transformers a poor choice for simpler time series tasks (Lara-Benítez et al., 2021).

2.4 Affective State Recognition

A number of works combine the ideas detailed in Sections 1.1 - 1.3 to classify emotions from multimodal physiological data. One such study conducted by Guo et al. (2016) involved showing subjects movie clips with the goal of eliciting five different emotions: sadness, anger, fear, happiness, and calmness. HRV features extracted from ECGs recorded during each emotion-specific viewing session were used for principal component analysis (PCA) and support-vector machine classification. With this approach, an accuracy of 56.90% was achieved across all five emotions. Similar research conducted by Shu et al. (2020) leveraged emotional movie clips to gather HRV data corresponding to a subject's affective state. After extracting

amplitude and frequency information from 25 subject's HRV measurements, gradient-boosted trees were used to achieve an accuracy of approximately 84% when classifying between neutral, happy, and sad.

With WESAD, the data set used later in this paper, Schmidt et al. (2018) used sliding window statistical feature extraction with various tree based models to classify 60-second physiological signal snapshots by baseline, stress, and amusement data collection sessions. With chest based inputs, they report a best accuracy of 80.34% with AdaBoost on their three-class problem. Garg et al. (2021) take a similar approach to that of Schmidt et al., but they only report a best three-class accuracy of 65.73% with 10-second signal windows.

With respect to deep-learning based approaches to ASR on the WESAD data set, Huyn et al. (2021) proposed a CNN optimized with neural architecture search, achieving a state-of-the-art classification accuracy of 83.43%. Alternatively, Rovinska and Khan (2022) employ a support vector machine to classify latent vectors from a CNN autoencoder, reporting a best accuracy and F_1-score of 85.66% and 82.82 respectively, but these results come from an altered preprocessing setup that yields fewer samples of only one second in length.

Chapter 3

WESAD: A Multimodal Dataset for Wearable Stress and Affect Detection

3.1 Introduction

As mentioned in Chapter 1, advancements in sensing capabilities have created opportunities for the efficient measurement and collection of physiological signals for emotion recognition. While brain wave and facial-expression based multimodal data sets exist for this very task (Koelstra et al., 2012), practical emotion recognition systems must be low-profile and reliant on consumer-friendly wearables. After all, electroencephalogram (EEG) sensors for brain wave measurement and constant facial expression monitoring are quite invasive and simply do not translate to non-laboratory environments.

With WESAD, Schmidt et al. (2018) present a multimodal data set comprised of various physiological signals collected with non-intrusive, chest and wrist-worn wearable devices. These signals are labeled by discrete emotional categories (neutral, stressed, amused, calm) from the sessions in which they were recorded, and the results of mental health self-assessments described in Section 2.1. Further, the inclusion of periodically collected mental health self-assessments in WESAD allows for novel emotion recognition approaches that are detailed in future chapters of this work.

3.2 Data Collection

3.2.1 Subjects

15 graduate students participated in the WESAD study. 13 of the 15 subjects were male, and their average age was 27.5 ± 2.4 years. Participants were rejected from the study due to pre-existing medical conditions, pregnancy, and heavy smoking habits to prevent the collection of abnormally influenced physiological signals.

3.2.2 Sensors and Signal Modalities

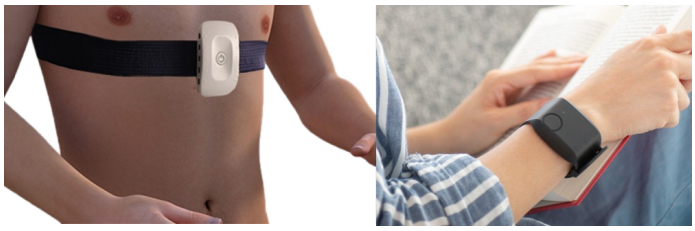

Figure 3.1: WESAD chest and wrist-worn sensors (Source: "E4 wristband | Real-time physiological signals | Wearable PPG, EDA, Temperature, Motion sensors", 2020)

Table 3.1: Physiological signal modalities in WESAD and their sensing sources

Signal	Source	Unit of measurement
Acceleration (ACC)	Chest, wrist	g
Electrocardiogram (ECG)	Chest	mV
Temperature (TEMP)	Chest, wrist	$°C$
Electrodermal activity (EDA)	Chest, wrist	μS
Blood volume pulse (BVP)	Wrist	mV
Respiration rate (RESP)	Chest	bpm
Electromyography (EMG)	Chest	mV

All signal modalities in WESAD were collected using the *RespiBAN Professional* and the *Empatica E4* (see Figure 3.1 left and right). Recorded signal modalities and their sensing sources are detailed in Table 3.1. All chest signals were sampled at 700 Hz, while BVP, EDA, TEMP, and ACC signals captured on the Empatica E4 were sampled at 64 Hz, 4 Hz, 4 Hz, and 32 Hz respectively.

Here, ACC represents the gravitational force (g) applied in each spatial dimension (x, y, and z) and is limited to $[-2g, +2g]$. ECG measures electrical activity of the heart over time in millivolts (mV), TEMP measures skin temperature in degrees Celsius, and EDA measures electrical conductance of the skin from sweat gland activity in microsiemens (μS). BVP refers to the rhythmic expansion and contraction of blood vessels with each heartbeat, and is measured in millivolts. RESP is measured in breaths per minute (bpm), and EMG measures muscular activity in millivolts.

3.2.3 Methodology

WESAD data collection is comprised of six steps: baseline measurement, amusement measurement, two meditation sessions, rest, and stress measurement. After being outfitted with chest and wrist wearables, in order to obtain valid baseline measurements, subjects were first instructed to read for 20 minutes in a sitting or standing position. Next, to elicit an amused response, subjects were shown 11 funny movie clips over a span of 392 seconds. To simulate stress, participants took part in an adaptation of the Trier

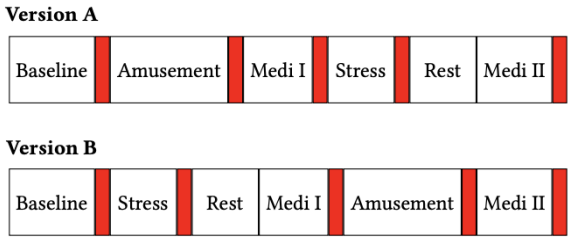

Figure 3.2: WESAD data collection protocol (Source: Schmidt et al., 2018)

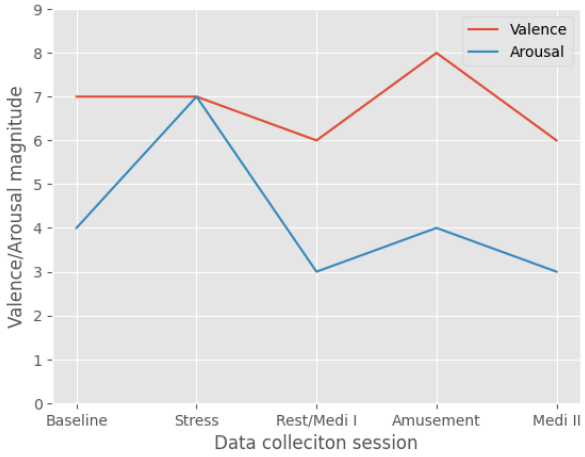

Figure 3.3: Subject 2's valence and arousal across data collection Version B

Social Stress Test (TSST). Schmidt et al.'s version of the TSST required participants to deliver a three-minute speech about their strengths and weaknesses to a panel of three graduate faculty members. After the speech, subjects were told to count down from 2023 to zero in increments of 17 - any mistake would restart the countdown. To bring the subjects back down to a neutral affective state, participants took part in a seven-minute guided audio meditation session after both the stress and amusement portions of the study. Additionally, subjects were given a rest period after the stress session that was separate from the guided meditation. In total, data collection lasted around two hours per subject. Slight variations were made to the data collection schedule in an effort to introduce variance into stressed and amused samples. The two main WESAD data collection schedules are detailed in Figure 3.2.

The red sections pictured in Figure 3.2 represent times when subjects were instructed to complete periodic mental health self-assessments. These assessments include: PANAS (see Figure 3.5), STAI (see Figure 3.4), and SAM. The results of these assessments were then used to verify the goal of each data collection session (see Table 3.2). Schmidt et al. (2018) note that the average PANAS, STAI, and SAM scores across subjects indicates higher levels of engagement and distress after the stress session, and lower feelings of anxiety after the amusement session. For example, as seen in Figure 3.3, Subject 2 experiences heightened levels of arousal during the stress session that subside after meditation.

Table 3.2: Questionnaire evaluation (Source: Schmidt et al., 2018)

	PANAS		SAM		STAI
Session	Positive	Negative	Valence	Arousal	
Baseline	25.5 ± 6.0	12.3 ± 2.0	6.7 ± 0.9	2.5 ± 0.9	10.8 ± 1.9
Stress	31.3 ± 4.7	22.0 ± 6.4	4.5 ± 1.6	6.8 ± 1.8	18.5 ± 2.0
Amusement	25.8 ± 5.1	11.4 ± 2.1	7.5 ± 0.6	3.0 ± 1.6	9.3 ± 2.0

These findings suggest that the WESAD data collection methodology was *mostly* successful in eliciting the emotional states that Schmidt et al. intended to capture with each session. Still, there are only minor differences in the results of each self-assessment between baseline and amusement sessions. Moreover, the amusement session was successful in reducing stress levels, but it is unclear whether subjects truly felt amused with Schmidt et al.'s emotion elicitation approach.

Read each statement and circle the number that best indicates how you feel right now	**Very slightly or not at all**	**Somewhat**	**Moderately**	**Very much**
I feel at ease	1	2	3	4
I feel nervous	1	2	3	4
I am jittery	1	2	3	4
I am relaxed	1	2	3	4
I am worried	1	2	3	4
I feel pleasant	1	2	3	4

Figure 3.4: Sample STAI self-assessment

Indicate the extent you have felt this way over the past week.		Very slightly or not at all	A little	Moderately	Quite a bit	Extremely
PANAS 1	Interested	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 2	Distressed	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 3	Excited	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 4	Upset	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 5	Strong	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 6	Guilty	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 7	Scared	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 8	Hostile	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 9	Enthusiastic	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 10	Proud	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 11	Irritable	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 12	Alert	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 13	Ashamed	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 14	Inspired	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 15	Nervous	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 16	Determined	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 17	Attentive	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 18	Jittery	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 19	Active	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
PANAS 20	Afraid	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

Scoring:

***Positive Affect Score*:** Add the scores on items 1, 3, 5, 9, 10, 12, 14, 16, 17, and 19. Scores can range from 10 – 50, with higher scores representing higher levels of positive affect.
Mean Scores: 33.3 (SD±7.2)

***Negative Affect Score*:** Add the scores on items 2, 4, 6, 7, 8, 11, 13, 15, 18, and 20. Scores can range from 10 – 50, with lower scores representing lower levels of negative affect.
Mean Score: 17.4 (SD ± 6.2)

Your scores on the PANAS: Positive: ____ Negative: ____

Figure 3.5: Sample PANAS self-assessment (Source: Watson et al., 1988)

Chapter 4

Methodology

4.1 Tasks

Given the fact that WESAD labels its contents by both data collection sessions (baseline, stress, amusement, meditation) and mental health self-assessments, it is worth exploring the relationship between the available signals and a number of target variables and their representations. If we formally define a classification model as

$$y = f(x|\theta);\ x \in R^n \tag{4.1}$$

where f is a user defined function that estimates a mapping between an input x and discrete variable y given some number of learned parameters θ, we can define a WESAD classification task as

$$T_{task} = \{0, 1, ..., n\} \tag{4.2}$$

where T_{task} represents the set of possible classification outputs that f can map its input feature set x to.

4.1.1 $T_{Session}$

Consider three WESAD data collection session types: baseline, stress, and amusement. If we map each session condition to 0, 1, and 2 respectively, we can define $T_{Session}$ as a multi-class classification task where

$$T_{Session} = \{0, 1, 2\} \tag{4.3}$$

In an effort to preserve the size of our data set, baseline and meditation sessions are combined to make up class 0.

4.1.2 T_{PANAS}

Consider the PANAS self-assessment pictured in Figure 3.5, where exam results are denoted by two measures, PA (positive affect) and NA (negative affect), ranging from $10 - 50$. We define T_{PANAS} as a

multi-class multi-output classification task, where PA and NA are each discretized into four equal width bins and used as subtasks T_{PA} and T_{NA}. Moreover, we define T_{PANAS} as

$$T_{PANAS} = \{T_{PA} \times T_{NA}\} \tag{4.4}$$

where bin 0 maps to $[10, 19]$, bin 1 maps to $[20, 29]$, bin 2 maps to $[30, 39]$, and bin 3 maps to $[40, 50]$ for both T_{PA} and T_{NA}.

4.1.3 T_{STAI}

Similar to PANAS, the STAI assessment requires subjects to rate statements about their emotional state on a 4 point Likert scale. The values attributed to each statement are aggregated and used to represent the overall magnitude of a subject's stress level. We discretize this aggregated score into three bins that are used as labels in T_{STAI}. Formally, T_{STAI} is defined as a multi-class classification task

$$T_{STAI} = \{0, 1, 2\} \tag{4.5}$$

where classes 0, 1, and 2 represent the ranges $[1, 8]$, $[9, 14]$, and $[15, 24]$ respectively. Notice that bin 1 is narrower than bins 0 and 2; due to the skew of STAI responses in WESAD, we chose bin widths that do not result in empty bins so as to keep T_{STAI} as a three-class task.

4.1.4 T_{SAM}

In the context of WESAD, the SAM was used to place subjects on the valence-arousal scale (see Figures 2.1 and 2.2). Valence and arousal each range from $1 - 9$, and for T_{SAM}, these axes are each divided into equal width bins, where bin 0 maps to $[1, 3]$, bin 1 maps to $[4, 6]$, and bin 2 maps to $[7, 9]$. We define T_{SAM} as a multi-class multi-output classification task comprised of subtasks $T_{Valence}$ and $T_{Arousal}$, where the output of T_{SAM} is denoted by

$$T_{SAM} = \{T_{Valence} \times T_{Arousal}\} \tag{4.6}$$

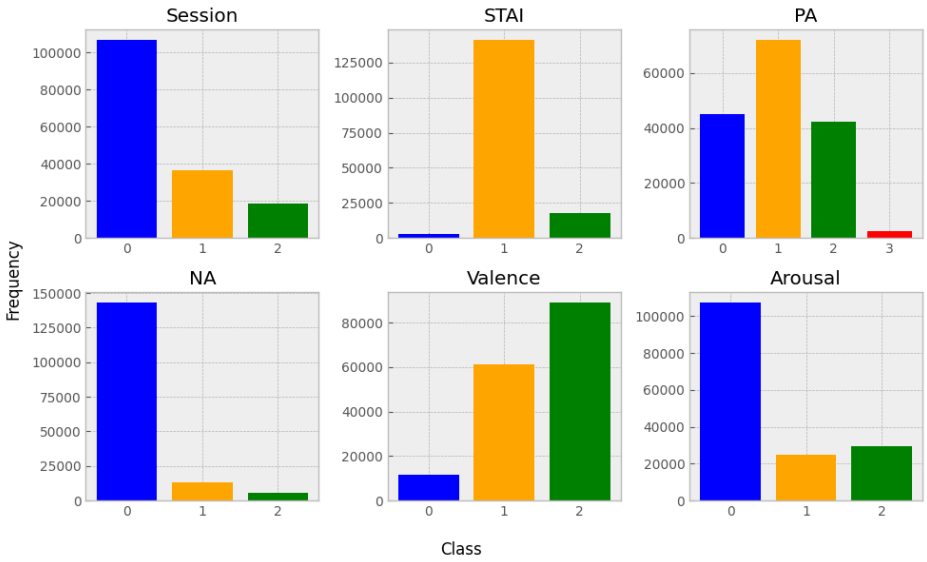

Figure 4.1: Class distribution for each WESAD task. Significant class imbalance is present across all tasks.

4.2 Features and Signals

When selecting input features from WESAD for ASR-related tasks, it important that we consider individual signal modalities, their sampling rates, and how easily they can be measured in an everyday setting. Further, the goal of this work is to introduce a practical framework for evaluating mental health from the physiological signals collected exclusively via wrist-worn wearables; we focus our feature set on a mixture of signals collected with both the *RespiBAN Professional* chest sensor and *Empatica E4* wrist sensor to maximize the number of modalities available, while remaining faithful to wrist measurements whenever possible. Specifically, we solely rely on the *RespiBAN Professional* for ECG, EMG, and RESP measurements, while all remaining signal modalities come from the wrist sensor (i.e., ACC, TEMP, EDA, and BVP). It is true that chest sensors can, and normally do, produce more accurate measurements than wrist-worn devices, but recent studies suggest that signals measured across the chest and wrist have comparable predictive power over affective states (Pinge et al., 2022). Moreover, the inclusion of chest-measured signals in this study can be attributed to data availability constraints. A full list of features and their descriptions is detailed in Table 4.1.

4.2.1 Signal Preprocessing

To effectively use signals collected from both the chest and wrist, we employ a Fourier based resampling method to match the sampling frequency of each signal modality to 70 Hz.[1] This sampling rate was chosen

[1]Fourier-transform based resampling was accomplished with SciPy, an open source Python library for general signal processing. See https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.resample.html for more details.

to preserve the interpretability of our high-fidelity chest measurements while also keeping signal length within reason for further processing (Mahdiani et al., 2015).

4.2.2 Feature Extraction

In the context of our time series analysis, feature extraction involves deriving insightful measures from individual time segments. If we consider ACC to be comprised of three separate signals: ACC_X, ACC_Y, and ACC_Z, the following statistical features are extracted from ECG, RESP, ACC, TEMP, EDA, BVP, and EMG signals:

- minimum
- maximum
- mean
- standard deviation

Additionally, the peak detection algorithm proposed by Christov (2004) is used to extract individual heartbeats from the ECG measurements of each subject;[2] segmenting signal peaks and beat-to-beat intervals from an ECG is a crucial initial step towards calculating HRV and a number of HRV-related metrics. When interpreting an ECG as influxes in voltage during the cardiac cycle over time where periodic voltage spikes are identified as instantaneous heartbeats, we can use a peak detection algorithm d to identify the times in which heartbeats occur over the duration of an ECG signal e. Consider

$$t = d(e) \tag{4.7}$$

to represent the set of times in which peaks are detected in e; the set of all beat-to-beat intervals in e can be represented by

$$Intervals = \{t_{i+1} - t_i | 1 \leq i < |t| - 1\} \tag{4.8}$$

We continue by deriving the following signal specific features from $Intervals$:

$$SDNN = \sqrt{\frac{1}{|Intervals| - 1} \sum_{n=1}^{|Intervals|} (Intervals_n - Intervals_\mu)^2} \tag{4.9}$$

$$RMSSD = \sqrt{\frac{1}{|Intervals| - 1} \sum_{n=1}^{|Intervals|} Intervals_n^2} \tag{4.10}$$

[2] ECG peak detection and HRV metric calculation is performed using pyHRV, an open source Python library for physiological signal processing. Source code is available at https://github.com/PGomes92/pyhrv.

$SDNN$ represents the standard deviation of all beat-to-beat intervals and $RMSSD$ represents the root mean square of all beat-to-beat intervals.

It is worth mentioning that the statistical and signal specific features described in Table 4.1 are extracted in a sliding window fashion, where the window size is 60 seconds and the window shift is 0.25 seconds (Kreibig, 2010; Schmidt et al., 2018). With a sampling rate of 70 Hz, our preprocessing schema generates 161886 individual windows of 4200 time steps in length.

Table 4.1: Signals and features used after feature selection and preprocessing

Signal	Source	Feature	Description
ACC	Wrist	$ACC_{X,Y,Z\mu}$	Mean acceleration in the x, y, or z axis
		$ACC_{X,Y,Z\sigma}$	Standard deviation of acceleration in the x, y, or z axis
		$ACC_{X,Y,Zmin}$	Minimum acceleration in the x, y, or z axis
		$ACC_{X,Y,Zmax}$	Maximum acceleration in the x, y, or z axis
ECG	Chest	ECG_{μ}	Mean electrocardiogram voltage
		ECG_{σ}	Standard deviation of electrocardiogram voltage
		ECG_{min}	Minimum electrocardiogram voltage
		ECG_{max}	Maximum electrocardiogram voltage
		$SDNN$	Standard deviation of $Intervals$
		$RMSSD$	Root mean square of $Intervals$
TEMP	Wrist	$TEMP_{\mu}$	Mean temperature
		$TEMP_{\sigma}$	Standard deviation of temperature
		$TEMP_{min}$	Minimum temperature
		$TEMP_{max}$	Maximum temperature
EDA	Wrist	EDA_{μ}	Mean electrodermal voltage
		EDA_{σ}	Standard deviation of electrodermal voltage
		EDA_{min}	Minimum electrodermal voltage
		EDA_{max}	Maximum electrodermal voltage
BVP	Wrist	BVP_{μ}	Mean blood volume pulse
		BVP_{σ}	Standard deviation of blood volume pulse
		BVP_{min}	Minimum blood volume pulse
		BVP_{max}	Maximum blood volume pulse
EMG	Chest	EMG_{μ}	Mean electromyography voltage
		EMG_{σ}	Standard deviation of electromyography voltage
		EMG_{min}	Minimum electromyography voltage
		EMG_{max}	Maximum electromyography voltage
RESP	Chest	$RESP_{\mu}$	Mean respiration rate
		$RESP_{\sigma}$	Standard deviation of respiration rate
		$RESP_{min}$	Minimum respiration rate
		$RESP_{max}$	Maximum respiration rate

Additionally, prior to being used as inputs to any statistical models, these windows undergo z-score normalization.

$$x_{normalized} = \frac{x - \mu}{\sigma} \tag{4.11}$$

4.3 Baseline Classification Algorithms

In accordance with the approaches of Schmidt et al. (2018) and Garg et al. (2021), we train the following machine learning classifiers to learn a mapping between the manually extracted features in Table 4.1 and the labels in $T_{Session}$ for comparison with deep learning models:

1. k - nearest neighbors
2. Decision tree
3. Random forest
4. AdaBoost
5. Linear discriminant analysis
6. Gradient boosting

Relevant hyperparameter and training details specific to each model are covered in the next chapter.

4.4 Our Approach

While using manual extracted features for supervised TSC can certainly yield favorable results, performance on downstream tasks is often reliant on one's domain expertise and the quality of the input features in question. Deep learning architectures, on the other hand, circumvent this issue by learning to extract meaningful features from raw input signals, instead of these features being specified beforehand. We propose an end-to-end network that takes raw, 60-second ACC, ECG, TEMP, EDA, BVP, EMG, and RESP windows as inputs.

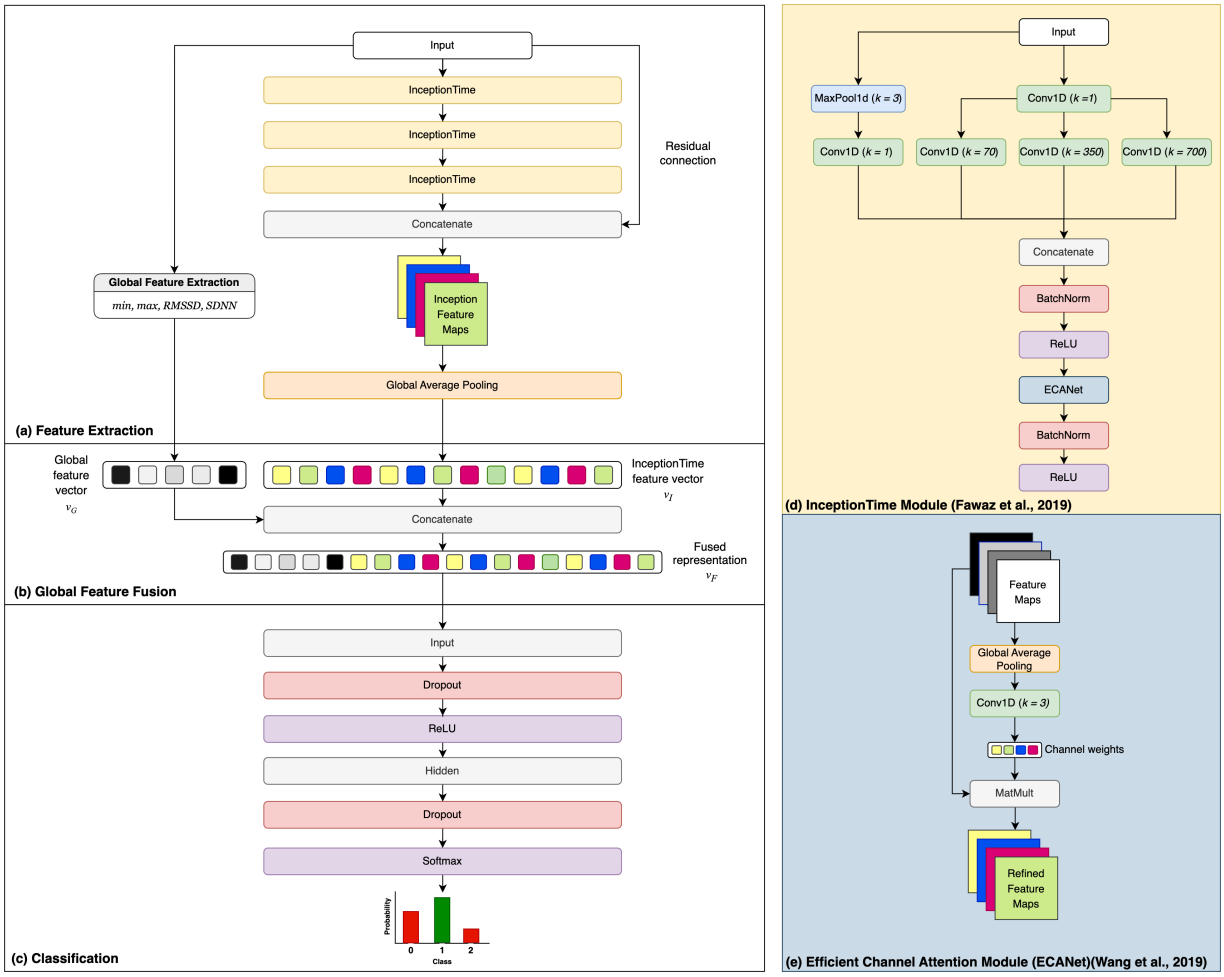

Figure 4.2: Overview of TAASR, our proposed classification architecture. It consists of a modified InceptionTime module (d) with efficient channel attention (e) and a global feature fusion branch (b).

Each signal modality is normalized separately before being fed into a feature extraction network (Figure 4.2.a) comprised of three InceptionTime modules (Figure 4.2.d) (Fawaz et al., 2019) with Efficient-Channel-Attention based feature refinement (Figure 4.2.e) (Wang et al., 2019), joined by a residual connection. When combined, these layers enable TAASR to learn which features from which time frames are the most important at increasingly complex levels of abstraction. Next, global average pooling is used to produce a translation invariant vector v_I from InceptionTime feature maps encoding physiological information extracted over 1, 5, and 10 second windows. v_I is then fused via concatenation with v_G, a vector of globally extracted features from Table 4.1, to form v_F, a multi-level, temporally-aware, signal representation vector. v_G is comprised of $SDNN$, $RMSSD$, and the min and max values of each input modality. Mean and standard deviation are excluded from v_G since their values remain constant across normalized windows ($\mu = 0, \sigma = 1$). v_F is passed through an MLP classification head (Figure 4.2.c) with one hidden layer before the softmax activation function is used to generate output class probabilities. In total, TAASR is comprised of 489726 trainable parameters. We train this architecture to specifically

perform the $T_{Session}$ three-class classification task, but we also explore augmenting the $T_{Session}$ training process via multi-task learning with mental health self-assessments. Experiments and results are detailed thoroughly in the following chapter.

4.4.1 InceptionTime

Classifying 60-second windows of raw physiological signals poses an interesting challenge - with a length of 4200 time steps per input sequence, any model applied to our data set must be able to efficiently capture extraordinarily long-term dependencies. While LSTM networks have been historically championed as long time lag task solvers (Hochreiter and Schmidhuber, 1997), recent empirical evaluations of sequence modeling approaches suggest that CNNs are more adept at handling lengthy input sequences (Bai et al., 2018). We follow this intuition by building upon a vanilla InceptionTime network for the feature extraction backbone of TAASR.

Fawaz et al.'s proposed InceptionTime architecture consists of three InceptionTime modules joined by a residual connection. Each InceptionTime module is comprised of:

1. a bottleneck layer that reduces the dimensionality of the input time series from M to $m < M$. This reduces model complexity and should encourage generalization. Max pooling is also used to feed a down sampled version of the inputs through another bottleneck layer, making the model invariant to small perturbations (Fawaz et al., 2019). We tested several different bottleneck sizes and found that a dimensionality of 16 yields the most favorable results. It may seem counter intuitive that our bottleneck layer increases model complexity, rather than decreasing it, but any attempts to set its size < 16 resulted in training instability and reduced performance.

2. a CNN ensemble layer. Three one-dimensional CNNs with varying kernel lengths are used for hierarchical feature extraction. Fawaz et al. (2019) set their kernel sizes to 10, 20, and 40 with 32 filters each. We set our InceptionTime module kernel sizes to 70, 350, and 700 with 16 filters each - these expanded kernels help TAASR capture dependencies over lengthy input sequences. With fewer output classes compared to the tasks covered in the original InceptionTime paper, fewer features are needed to learn $T_{Session}$, hence our decision to reduce the number of filters to 16.

3. hierarchical feature map concatenation. The outputs of the bottleneck and CNN ensemble layers are concatenated depth-wise to form a final feature map with $4 \times$ number of filters dimensions.

Stacking InceptionTime modules atop one another enables the network to gradually extract high and low-level features from the input time series, and the residual connection permits easy flow of information throughout the architecture as a whole.

4.4.2 Channel Attention

The concept of *attention*, or the brain's ability to "focus on what's important", has recently gained popularity among deep learning researchers. In practice, a multitude of neural attention mechanisms have

been used to improve the performance of natural language processing (Vaswani et al., 2017) and computer vision (Woo et al., 2018) systems.

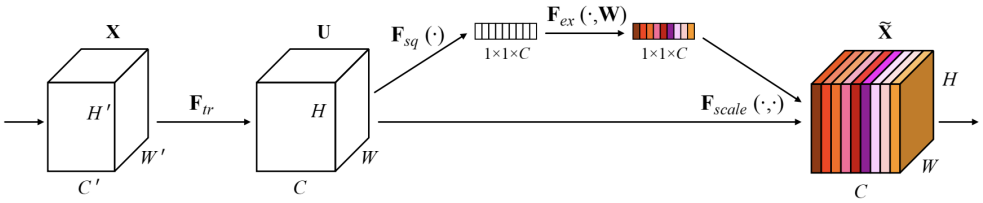

Figure 4.3: Squeeze-and-Excitation Block (Source: Hu et al., 2017)

One such attention mechanism, Squeeze-and-Excitation (SENet)(Figure 4.3), serves as a CNN feature refinement module that uses global channel interdependencies to accentuate important features and suppress less informative ones (Hu et al., 2017). The first component of SENet, *squeeze*, uses global average pooling to spatially compress a CNN's output U of size $C \times W \times H$ into a global information embedding vector S of size $1 \times 1 \times C$. Next, in the *excitation* phase, S is passed through a multi-layer perceptron with ReLU activations to capture non-linear dependencies between channels before the sigmoid function is used to generate channel weights, forming a recalibrated global information vector E. Finally, channel-wise multiplication between E and U results in a recalibrated set of CNN feature maps $\tilde{X}$.

While Hu et al. (2017) demonstrate that SENet achieves state-of-the-art performance across many tasks, with Efficient Channel Attention (ECANet) (Figure 4.2.e), Wang et al. (2019) suggest that capturing cross-channel dependencies with a multi-layer perceptron adds unnecessary complexity to the overall channel attention architecture. Moreover, ECANet improves upon the computational efficiency of SENet by modeling cross-channel interactions in S with $1D$ convolutions of size k, where k represents the number of neighboring channels that can impact a given channel's predicted attention weight (Wang et al., 2019). This convolution is depth preserving, eliminating the need for dimensionality reduction to produce E as performed in SENet's bottlenecked MLP (*excitation*). The kernel size k of ECANet is adaptively determined by

$$k = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{4.12}$$

where C is the number of channels and γ and b are manually determined hyperparameters. The result of this calculation should be rounded to the nearest odd number. Wang et al. set γ and b to 2 and 1 respectively. In a similar fashion to SENet, ECANet produces a pseudo-global information embedding vector of size $1 \times 1 \times C$ that is passed through the sigmoid activation before being used to recalibrate its input feature maps. With 64 output channels from TAASR's InceptionTime backbone, we use the aforementioned hyperparameters with Equation 4.12 to get a k of three for TAASR's channel attention module. As depicted in Figure 4.2.d, ECANet is the terminal architectural component of each of TAASR's InceptionTime modules, outputting refined feature maps that are batch normalized and fed through the ReLU activation function before being propagated through successive InceptionTime layers.

4.4.3 MLP Classification Head

The final component of TAASR, an MLP classification head, uses v_F, a multi-level, temporally-aware representation vector generated via global average pooling on InceptionTime feature maps, to output class probabilities for any given training task. As depicted in Figure 4.2.c, our classification head is comprised of an input layer, one hidden layer, and an output layer. v_F is compressed from 84 features to 32 features in the hidden layer before being further reduced to the number of classes in the output layer. ReLU activation is used in the hidden layer to help the classification head learn non-linear relationships between affective states, and dropout ($p = 0.5$) between layers helps mitigate overfitting during training (Srivastava et al., 2014).

4.5 Evaluation

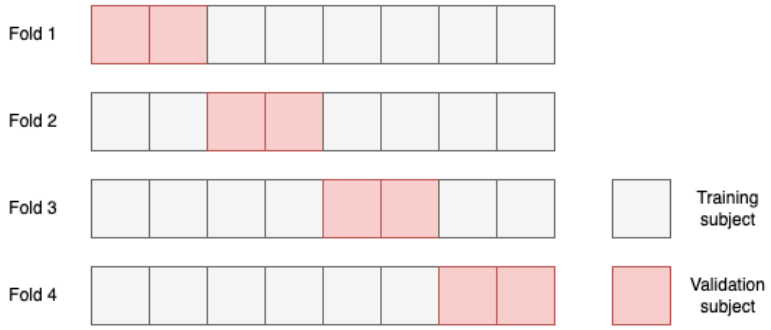

Figure 4.4: Example of leave-two-subjects-out cross-validation with 8 subjects

Due to the imbalanced nature of our classification tasks, both accuracy and F_1-score are used to evaluate each model's performance. In the multi-class setting, F_1-score is calculated for each individual class and averaged to obtain a final measurement. Evaluation is performed via leave-two-subjects-out cross-validation to ensure generalization to unseen subjects (see Figure 4.4). Prior works (Schmidt et al., 2018; Rashid et al., 2021; Huynh et al., 2021) utilize LOSO (leave-one-subject-out) cross-validation, but due to computational restraints we reserve two subjects for each cross-validation fold, resulting in seven training splits. Since there are 15 subjects in WESAD, the last fold includes three subjects.

Chapter 5

Experiments and Results

5.1 Experimental Setup

5.1.1 Software and Hardware

All experiments described in this section were conducted with Python on an NVIDIA V100 GPU provided by Google Cloud Computing Services. PyTorch was used for the development of TAASR and Scikit-learn was used for all baseline classification algorithms.

5.1.2 Baseline Hyperparameters

We employ a grid-search based hyperparameter optimization scheme to tune each baseline classification algorithm in a brute-force fashion. For kNN, we find that classification accuracy and F_1-score are fairly resistant to any changes in k, the number of neighbors to be considered for any given prediction. A grid search over $[1, 10]$ results in an optimal k of 5. Conversely, for our data set, AdaBoost, random forest, and gradient boosting are significantly more sensitive to hyperparameter tuning. We train and evaluate each of the aforementioned learning algorithms with ensemble sizes $n\_estimators$ between 10 and 100 in increments of 10, resulting in tuned ensembles sizes of 30, 90, and 100 for AdaBoost, random forest, and gradient boosting respectively (see Figure 5.1). For each tree based algorithm, the minimum number of samples required to split a node is set to 20 (Schmidt et al., 2018).

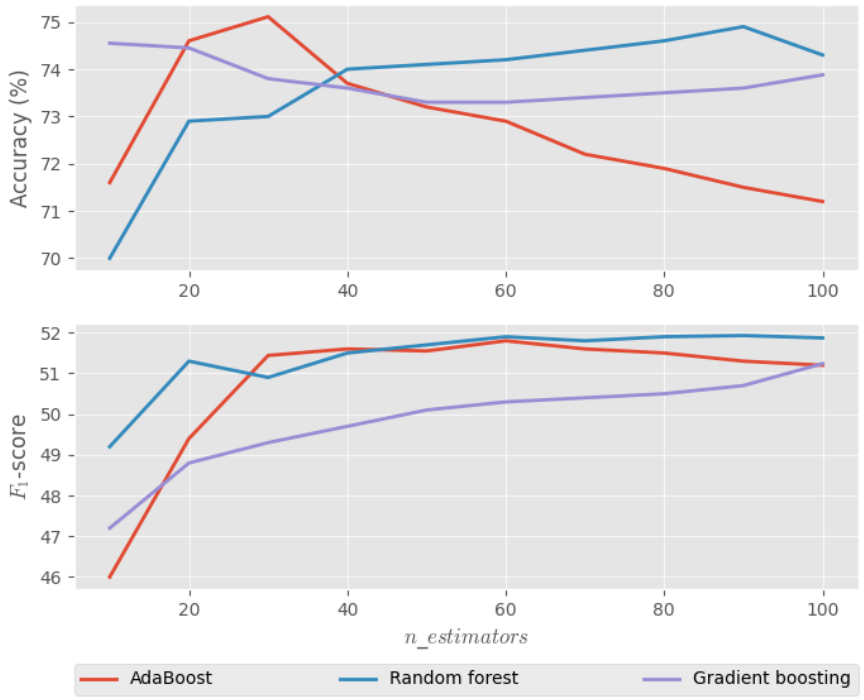

Figure 5.1: Ensemble classification algorithm performance when varying $n\_estimators$. Optimal performance for AdaBoost is reached at 30 estimators, while larger ensemble sizes (i.e., 90 and 100) are needed to reach optimal performance for random forest and gradient boosting.

5.1.3 Training TAASR

In order to make consistent comparisons and highlight any architectural improvements between TAASR and InceptionTime, we borrow numerous hyperparameter values vital to training from the original InceptionTime paper (Fawaz et al., 2019). This includes optimizer (Adam), learning rate (0.0001), and batch size (64). TAASR is trained with categorical cross-entropy loss for 10 epochs to satisfy the multi-class nature of $T_{Session}$. Early stopping is used to terminate training after three consecutive epochs of increasing loss. To remedy class imbalance in $T_{Session}$, each term of categorical cross-entropy is scaled by a class weight given by

$$w_c = \frac{num\_samples}{num\_classes \times num\_samples_c} \tag{5.1}$$

Furthermore, the loss of some sample o is calculated by a weighted sum of cross-entropy for each class c, where classes with fewer samples are increasingly penalized by w_c, and only the ground truth class term ($y_c = 1$) contributes to the overall total. By scrutinizing predictions for minority classes we prevent TAASR from overfitting to overrepresented training examples.

$$L(o) = -\sum_{c=1}^{C} w_c y_c \log(f(o)_c) \tag{5.2}$$

5.2 Incrementally Modifying InceptionTime for $T_{Session}$

Table 5.1: Accuracy and F_1-score for each InceptionTime variant trained to highlight the benefits of our contributions. k_{small} denotes $k \in \{10, 20, 40\}$, k_{large} denotes $k \in \{70, 350, 700\}$, and d denotes the depth of the network.

InceptionTime variant	$T_{Session}$	
	Accuracy (%)	F_1-score
k_{small}, $d = 1$	64.91 ± 1.82	48.18 ± 2.19
k_{large}, $d = 1$	66.80 ± 1.34	53.77 ± 0.93
k_{small}, $d = 3$	71.47 ± 0.84	57.32 ± 0.99
k_{large}, $d = 3$	72.99 ± 1.24	59.06 ± 1.20
k_{large}, $d = 3$, w/ fusion	75.05 ± 1.08	62.09 ± 0.87
$\mathbf{k_{large}}$, $\mathbf{d = 3}$, w/ fusion and attention	$\mathbf{81.12 \pm 1.40}$	$\mathbf{68.95 \pm 0.85}$

To demonstrate the effectiveness of TAASR as an end-to-end learning architecture for $T_{Session}$, we begin by experimenting with a vanilla InceptionTime network with 1 layer and train models of continually increasing complexity until the full TAASR setup is reached. The results of these experiments are listed in Table 5.1

A single InceptionTime module with basic kernel sizes $k \in \{10, 20, 40\}$ is less accurate than sophisticated guessing (64.91% vs. 66.17%), but this is expected, since: (1) with a sufficiently large data set, deeper networks are better at generalizing (LeCun et al., 2015) and (2) the kernel sizes proposed by Fawaz et al. (2019) have little capacity to capture meaningful patterns given our input sampling frequency of 70 Hz. By simply increasing the kernel sizes to $k \in \{70, 350, 700\}$ to cover 1, 5, and 10 second windows, we observe a $\approx 2\%$ increase in accuracy and a $\approx 5\%$ increase in F_1-score. Still, with just one layer, this basic architecture learns feature maps at only one level of abstraction. To remedy this shortcoming, we chain three InceptionTime modules together and perform two separate training runs - one with small kernels and one with large kernels - to demonstrate the importance of network depth and proper receptive field sizes. This results in significant increases in both accuracy and F_1-score for each model - we report an accuracy of 71.47% and an F_1-score of 57.32 for InceptionTime with three layers and basic kernels and an accuracy of 72.99% and an F_1-score of 59.06 for InceptionTime with three layers and large kernels.

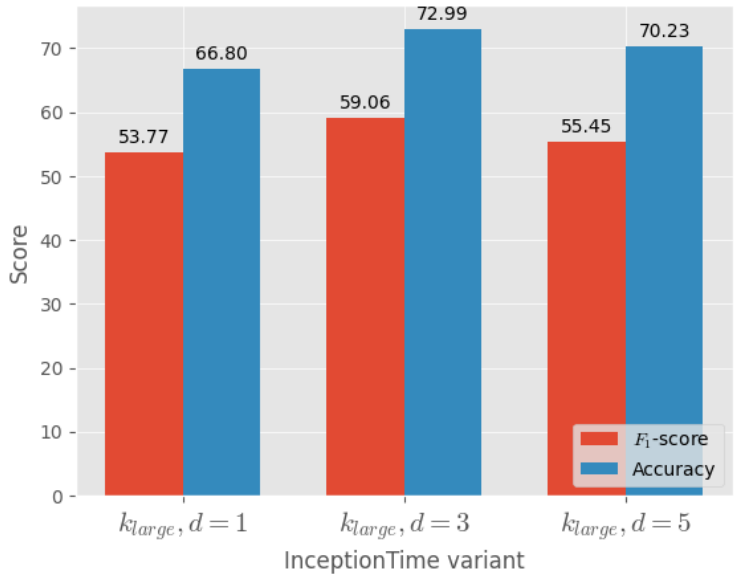

Figure 5.2: InceptionTime (k_{large}) performance for $T_{Session}$ when varying network depth

It is worth noting that any attempts to train InceptionTime variants with more than three layers resulted in noticeable drops in both accuracy and F_1-score, highlighting a distinct trade-off between model complexity and performance when optimizing for $T_{Session}$ (see Figure 5.2).

In our next experiment, we build upon a three layer InceptionTime network with big kernels by fusing its output representation with manually extracted global statistics. Each InceptionTime module learns to extract meaningful features over three individual time scales, and while we can always incorporate additional CNNs into the InceptionTime ensemble to expand the architecture's degree of multi-level awareness, this could lead to over-parameterization and longer training times. For simplicity, we leverage a number of the features in Table 4.1 to give the model understanding over a 60-second time frame. Moreover, when combining learned features with global window summary statistics via vector concatenation, and providing the network with explicit HRV information, we observe a $\approx 2\%$ increase in accuracy and a $\approx\ 3\%$ increase in F_1-score over the previous best InceptionTime variant with minimal increases to network complexity.

Finally, we explore the effect of incorporating an attention module into each of the three InceptionTime modules in TAASR's feature extraction backbone. By enabling InceptionTime to recalibrate extracted features at several stages of abstraction, we anticipate that a more nuanced understanding of $T_{Session}$ will be propagated through the network as a whole, resulting in compounded performance benefits from TAASR's individual components. The full TAASR architecture sees increases of $\approx 6\%$ to both accuracy and F_1-score when augmented with ECANet. Interestingly, training converged after six epochs, close to double the average training length of our previous experiments, demonstrating that TAASR's superior classification performance comes from an improved ability to recognize affective states which emerges further along in training.

5.3 Multi-Task Learning

Multi-task learning (MTL) describes a paradigm for training machine learning models where individual but related tasks are learned simultaneously so as to improve task-specific performance (Caruana, 1998). Here, to make use of the MTL framework, we make the assumption that data collection sessions in $T_{Session}$ are inherently related to the outcomes of mental health self-assessments (i.e., T_{STAI}, T_{PA}, T_{NA}) collected in WESAD. For example, subjects participating in a stress session may systematically report higher levels of stress in each self-assessment, whereas average levels of valence and arousal may be exclusively reported during baseline sessions. Further, when considering these cross-task relationships, it is certainly worth exploring modifying a single TAASR model to handle multiple classification tasks to augment the performance of our main focus, $T_{Session}$.

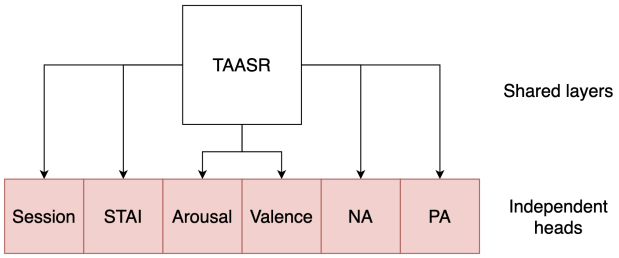

Figure 5.3: High-level overview of TAASR-MT

TAASR-MT (see Figure 5.3), our proposed multi-task version of TAASR, consists of a shared trunk with task-specific MLP classification heads - a setup commonly referred to as *hard parameter sharing* (Ruder, 2017). This architecture enables feature extraction to be learned jointly, producing a task-agnostic representation vector v_F that can be tuned for further use in any number of independent classification branches.

Given the multi-class nature of the tasks we derive from WESAD, each classification head's task-specific understanding is given by weighted categorical cross-entropy, while the sum of losses across tasks is used to optimize the network as a whole. We formally define TAASR-MT's loss function, L_{MT}, as a multi-task loss where the multi-task cross-entropy of a sample o is given by the sum of the losses of each task t in some global task set T.

$$L_{MT}(o) = -\sum_{t=1}^{T}\sum_{c=1}^{C} w_{t,c} y_{t,c} \log(f(o)_c) \tag{5.3}$$

After training TAASR-MT according to the experimental setup described in Section 5.1.3 with a revised loss function, we observe convergence after an average of 6.85 ± 0.47 epochs, which is slightly longer than a specialized TAASR architecture's training time. TAASR-MT's multi-task learning objective yields a $\approx 1\%$ increase in $T_{Session}$ F_1-score, indicating a degree of positive transfer between the

Table 5.2: Accuracy and F_1-score for each task used to train a multi-headed TAASR architecture with multi-task loss

Task	TAASR-MT	
	Accuracy (%)	F_1-score
$T_{Session}$	76.54 ± 1.28	70.02 ± 0.74
T_{PA}	54.07 ± 2.23	51.19 ± 2.40
T_{NA}	84.67 ± 1.16	80.72 ± 0.99
T_{STAI}	33.15 ± 3.48	16.70 ± 2.67
$T_{Valence}$	90.43 ± 0.91	90.33 ± 1.88
$T_{Arousal}$	81.16 ± 1.59	74.57 ± 1.37

knowledge needed to classify WESAD data collection sessions and mental health self-assessments from physiological signal snapshots. Conversely, we see a $\approx 5\%$ dip in $T_{Session}$ accuracy attributed to negative transfer from poorly learned tasks (i.e. T_{PA} and T_{STAI}). Moreover, self-assessment results vary only subtly across WESAD data collection sessions, making class separation difficult (Schmidt et al., 2018). Our assessment discretization scheme may further contribute to assessment-specific learning challenges and negative transfer by removing any useful information provided by continuous assessment representations.

5.4 Results

Table 5.3: Accuracy and F_1-score of models trained for the $T_{Session}$ classification task (*baseline* vs. *stressed* vs. *amused*)

Classification algorithm	$T_{Session}$	
	Accuracy (%)	F_1-score
AdaBoost	75.11 ± 0.69	51.44 ± 1.15
LDA	76.56 ± 0.88	58.07 ± 0.83
kNN	57.90 ± 1.05	42.99 ± 2.24
Decision tree	66.97 ± 1.27	52.48 ± 0.98
Random forest	74.90 ± 1.12	51.93 ± 0.90
Gradient boosting	73.88 ± 0.64	51.24 ± 0.71
InceptionTime	71.47 ± 0.84	57.32 ± 0.99
TAASR	$\mathbf{81.12 \pm 1.40}$	68.95 ± 0.85
TAASR-MT	76.54 ± 1.28	$\mathbf{70.02 \pm 0.74}$
Random guess	33.33	29.42
Sophisticated guess	66.17	26.68

A full comparison between baseline classification algorithms, InceptionTime, TAASR, and TAASR-MT is detailed in Table 5.3. With an accuracy of $76.56 \pm 0.88\%$ and an F_1-score of 58.07 ± 0.83, LDA trained with hand-crafted inputs performs better than vanilla InceptionTime across each classification metric, indicating that end-to-end learning techniques for TSC are not necessarily sufficient for task-specific performance improvements. Conversely, while AdaBoost, kNN, decision tree, random forest, and gradient boosting outperform InceptionTime in terms of accuracy, their F_1-scores indicate poor classification robustness on $T_{Session}$. In this case, the benefits of end-to-end learning are immense, but only observed after incorporating channel attention, properly sized receptive fields, and multi-task learning (optionally) into InceptionTime.

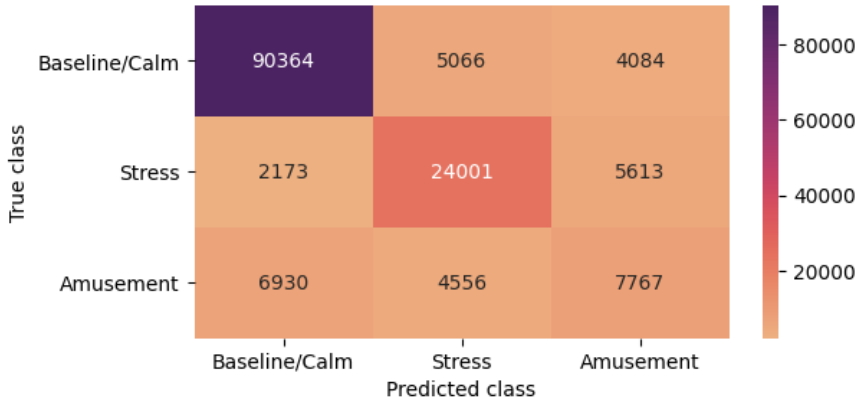

Figure 5.4: TAASR confusion matrix for $T_{Session}$

Table 5.4: TAASR's class-specific performance

Class	Accuracy (%)	F_1-score
Baseline/Calm	90.85	90.83
Amusement	44.47	42.31
Stress	71.38	73.71

Depicted in Figure 5.4 is the confusion matrix for TAASR's $T_{Session}$ predictions over each cross-validation fold. As shown in Table 5.4, 90.85% of all baseline samples and 71.38% of all stress samples were positively identified, but just 44.47% of amusement samples were classified correctly. While TAASR often confuses amusement samples for baseline ones, our amusement-specific accuracy outperforms that of Schmidt et al. (2018) by $\approx 10\%$. It is worth noting that Schmidt et al. report a superior overall F_1-score to both TAASR and TAASR-MT (72.51 ± 0.17) with AdaBoost trained exclusively on manually extracted chest based sensor inputs, and inferior performance when wrist based signals are included. We use signals collected on the wrist whenever possible to present results consistent with practically collected data (i.e., everyday affective state recognition is most practically accomplished with wrist based wearables; chest sensors can be cumbersome and uncomfortable) and report a higher classification accuracy by $\approx 1\%$ and an F_1-score within 2% of their best performing chest based model. Additionally, TAASR and TAASR-MT outperform Schmidt et al.'s best wrist based model by $\approx 5\%$ for accuracy and ≈ 4 for F_1-score, but direct comparisons should be interpreted with caution given our mixed feature set and altered cross-validation scheme.

Huynh et al. (2021) also applied deep learning methods to WESAD with StressNAS, a chest based deep CNN optimized with a neural architecture search. They report a best classification accuracy of 83.43% for $T_{Session}$, outperforming both TAASR and Schmidt et al.'s AdaBoost, but only after a 50 hour architecture ranking process. Huynh et al. also train and evaluate a fully connected MLP and a ResNet-like architecture on chest based ACC, EDA, BVP, and TEMP signals for comparison with StressNAS.

They report accuracies of 78.11% and 79.48% for each of the aforementioned models, narrowly underperforming TAASR. Nevertheless, Huynh et al. fail to include F_1-scores for any of their StressNAS experiments, so it is unclear whether their ASR approach holds any merit.

Chapter 6

Concluding Remarks

6.1 Conclusion

In this work, we investigated deep learning approaches with primarily wrist based input features for affective state recognition, and compared their performance with alternative machine learning approaches. Through training a modified InceptionTime network - TAASR - we were able to successfully capture hierarchical temporal dependencies from ECG, RESP, EMG, TEMP, ACC, BVP, and EDA signals for various emotion classification tasks.

To justify TAASR as an improvement over Fawaz et al.'s (2019) InceptionTime, we conducted a series of experiments that built upon a single InceptionTime module, highlighting the performance benefits of TAASR's individual components. Our experiments explored network depth, convolutional kernel sizes, incorporating globally extracted features, and attention for iterative feature refinement. Further, we improved TAASR's $T_{Session}$ classification robustness with TAASR-MT, a multi-task network that uses hard parameter sharing and multi-task loss to transfer knowledge across session and mental health self-assessment classification tasks.

In a three-class setting, TAASR demonstrates a strong ability to recognize baseline/calm and stress samples, whereas identifying amusement samples proves to be a more difficult undertaking. Nevertheless, results indicate that our end-to-end learning approach is more accurate than various baseline architectures that make use of similar input features - and comparable to those that make exclusive use of high fidelity chest signals. Moreover, by evaluating TAASR and TAASR-MT with a leave-two-subjects-out cross-validation scheme, we demonstrate that generalization to unseen subjects is possible with learned feature extraction.

6.2 Implications and Ethical Concerns

Portable ASR systems can help physicians and public health officials easily gather important characteristics about their patients and subjects of study. For example, TAASR-MT's ability to predict the binned outcome of certain self-assessments could enable physicians to rapidly identify and diagnose, or confirm

the diagnosis of mental health disorders like depression and anxiety. Further, should wearables detect increased levels of stress or calmness in certain populations, public health officials can use this information to develop policies and programs catered to specific communities. Hospitals could also use automatic ASR monitoring to track their patient's pain levels after a risky procedure (Campbell et al., 2019). In a commercial setting, mobile applications could be integrated with TAASR, permitting smartwatch owners to use their devices to reinforce healthy lifestyle choices.

Despite these use cases, many see mental health as something that should remain private. In an age where society is hyperfocused on data privacy and digital rights, the idea that wearables have the ability (to some degree) to understand how you feel would likely be unsettling to most. Any physiological data used for such an application should be treated with the utmost care to ensure privacy and prevent misuse, and those with access to sensitive records should be diligently screened. Further, concerns over misuse could hinder widespread public adoption of wearables, making it difficult to improve ASR systems with larger and more informative data sets.

6.3 Limitations and Future Work

Although TAASR demonstrates reasonably good performance on $T_{Session}$, the limited number of subjects ($n = 15$) contained in WESAD highlights a need for further ASR data collection studies with wearable devices. Additionally, given the fact that the baseline, amusement, and stress sessions were conducted in accordance with specific activities, TAASR may have difficulty generalizing to new environments, even if they are associated with predictable affective states. For example, when participating in physically demanding activities, a high heart rate and excessive skin moisture should not necessarily be equated with stress. This shortcoming could be addressed by incorporating both location data and activity recognition systems (Yazdansepas et al., 2016) into the complete ASR pipeline.

Future work should be concerned with several topics. First, differentially private federated learning could be used to train TAASR in a safe and responsible manner, thus addressing several ethical concerns. Next, on top of pseudo-global channel attention provided by ECANet (Wang et al., 2019), spatial attention could be incorporated with TAASR's InceptionTime modules to give the network localized focus of individual CNN feature maps. Additionally, learned weighting of TAASR-MT's multi-task loss (Kendall et al., 2017) could remedy negative transfer so that poorly learned tasks contribute less to the model's learning objective. Investigating new self-assessment representation schemes could also be an easy way to further optimize TAASR-MT training. Given the relatively fuzzy differentiation between amusement and baseline samples in WESAD, pre-training TAASR with supervised contrastive loss (Khosla et al., 2020) could result in better class separability without the need for further data collection sessions.

Bibliography

Allen, A. P., Kennedy, P. J., Dockray, S., Cryan, J. F., Dinan, T. G., & Clarke, G. (2017). The trier social stress test: Principles and practice. https://doi.org/https://doi.org/10.1016/j.ynstr.2016.11.001

Amidi, A., & Amidi, S. (2019). Cs 230 - recurrent neural networks cheatsheet.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, *abs/1803.01271*. http://arxiv.org/abs/1803.01271

Bestelmeyer, P., Kotz, S., & Belin, P. (2017). Effects of emotional valence and arousal on the voice perception network - pmc. https://doi.org/10.1093/scan/nsx059

Bieling, P. J., Antony, M. M., & Swinson, R. P. (1998). The state–trait anxiety inventory, trait version: Structure and content re-examined. https://doi.org/https://doi.org/10.1016/S0005-7967(98)00023-0

Bradley, M., & Lang, P. (1994). Measuring emotion: The self-assessment manikin and the semantic differential.

Campbell, E., Phinyomark, A., & Scheme, E. (2019). Feature extraction and selection for pain recognition using peripheral physiological signals. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00437

Caruana, R. (1998). *Multitask learning*. Springer.

Christodoulou, G., Salami, N., & Black, D. S. (2020). 12671_2019_1296_article 554..570.

E4 wristband | real-time physiological signals | wearable ppg, eda, temperature, motion sensors. (2020).

Fawaz, H. I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P., & Petitjean, F. (2019). Inceptiontime: Finding alexnet for time series classification. *CoRR*, *abs/1909.04939*. http://arxiv.org/abs/1909.04939

Georgallides, G. (2021). What is hrv? | basis blog.

Gertz, M., Große-Butenuth, K., Junge, W., Maassen-Francke, B., Renner, C., Sparenberg, H., & Krieter, J. (2020). Using the xgboost algorithm to classify neck and leg activity sensor data using on-farm health recordings for locomotor-associated diseases. *Computers and Electronics in Agriculture*, *173*, 105404. https://doi.org/https://doi.org/10.1016/j.compag.2020.105404

Hayward. (2022). Covid-19's toll on mental health.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hrnjica, B. (2019). Lstm – bahrudin hrnjica blog — hrnjica.net.

Hu, J., Shen, L., & Sun, G. (2017). Squeeze-and-excitation networks. *CoRR*, *abs/1709.01507*. http://arxiv.org/abs/1709.01507

Huynh, L., Nguyen, T., Nguyen, T., Pirttikangas, S., & Siirtola, P. (2021). Stressnas: Affect state and stress detection using neural architecture search. *CoRR*, *abs/2108.12502*. https://arxiv.org/abs/2108.12502

Karthikeyan, P., Murugappan, M., & Yaacob, S. (2011). A review on stress inducement stimuli for assessing human stress using physiological signals. https://doi.org/10.1109/CSPA.2011.5759914

Kendall, A., Gal, Y., & Cipolla, R. (2017). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. https://doi.org/10.48550/ARXIV.1705.07115

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. https://doi.org/10.48550/ARXIV.2004.11362

Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y. H., & Koo, B.-H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry investigation*, *15*(3), 235.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

Kreibig, S. (2010). Autonomic nervous system activity in emotion: A review. *Biological psychology*, *84*, 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Lakew, S. M., Cettolo, M., & Federico, M. (2018). A comparison of transformer and recurrent neural networks on multilingual neural machine translation. *CoRR*, *abs/1806.06957*. http://arxiv.org/abs/1806.06957

Lara-Benıtez, P., Gallego-Ledesma, L., Carranza-Garcıa, M., & Luna-Romera, J. M. (2021). Evaluation of the transformer architecture for univariate time series forecasting. *Advances in Artificial Intelligence: 19th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2020/2021, Málaga, Spain, September 22–24, 2021, Proceedings 19*, 106–115.

Laricchia, F. (2022). Global connected wearable devices 2016-2022. https://www.statista.com/statistics/487291/global-connected-wearable-devices/

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lyons, E. J., Lewis, Z. H., Mayrsohn, B. G., & Rowland, J. L. (2014). Behavior change techniques implemented in electronic lifestyle activity monitors: A systematic content analysis. *J Med Internet Res*, *16*(8), e192. https://doi.org/10.2196/jmir.3469

Mahdiani, S., Jeyhani, V., Peltokangas, M., & Vehkaoja, A. (2015). Is 50 hz high enough ecg sampling frequency for accurate hrv analysis? *2015*. https://doi.org/10.1109/EMBC.2015.7319746

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, 69–84.

Nurwanto, F., Ardiyanto, I., & Wibirama, S. (2016). Light sport exercise detection based on smartwatch and smartphone using k-nearest neighbor and dynamic time warping algorithm. https://doi.org/10.1109/ICITEED.2016.7863299

Pinge, A., Bandyopadhyay, S., Ghosh, S., & Sen, S. (2022). A comparative study between ecg-based and ppg-based heart rate monitors for stress detection. *2022 14th International Conference on COMmu-*

nication Systems NETworkS (COMSNETS), 84–89. https://doi.org/10.1109/COMSNETS53615.2022.9668342

Rashid, N., Chen, L., Dautta, M., Jimenez, A., Tseng, P., & Al Faruque, M. A. (2021). Feature augmented hybrid cnn for stress recognition using wrist-based photoplethysmography sensor. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2374–2377.

Rubin, D., & Talarico, J. (2009). A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words - pmc.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *CoRR*, *abs/1706.05098*. http://arxiv.org/abs/1706.05098

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological review*, *69*(5), 379.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018). Introducing wesad, a multimodal dataset for wearable stress and affect detection. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 400–408. https://doi.org/10.1145/3242969.3242985

Shaffer, F., & Ginsberg, J. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, *5*. https://doi.org/10.3389/fpubh.2017.00258

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Staff, M. C. (2020). Mindfulness exercises - mayo clinic. https://www.mayoclinic.org/healthy-lifestyle/consumer-health/in-depth/mindfulness-exercises/art-20046356#:~:text=Mindfulness

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2019). Eca-net: Efficient channel attention for deep convolutional neural networks. *CoRR*, *abs/1910.03151*. http://arxiv.org/abs/1910.03151

Watson, D., Clark, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. https://www.semanticscholar.org/paper/Development-and-validation-of-brief-measures-of-and-Watson-Clark/f82f152244b1cb861db0f290d55302011aee28dc

Waxenbaum, J. A., Reddy, V., & Varacallo, M. (2021). Anatomy, autonomic nervous system. https://www.ncbi.nlm.nih.gov/books/NBK539845/

Woo, S., Park, J., Lee, J., & Kweon, I. S. (2018). CBAM: convolutional block attention module. *CoRR*, *abs/1807.06521*. http://arxiv.org/abs/1807.06521

Yazdansepas, D., Niazi, A. H., Gay, J. L., Maier, F. W., Ramaswamy, L., Rasheed, K., & Buman, M. P. (2016). A multi-featured approach for wearable sensor-based human activity recognition. *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 423–431. https://doi.org/10.1109/ICHI.2016.81

Yoon, S. W., Seo, J., & Moon, J. (2019). Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. https://doi.org/10.48550/ARXIV.1905.06549

Zhus, J., Ji, L., & Liu, C. (2019). Heart rate variability monitoring for emotion and disorders of emotion. https://doi.org/10.1088/1361-6579/ab1887