Personalization of Affect Recognition Models Using Biosignals and Personality

Anna de Wolff

Utrecht University
MSc Human-Computer Interaction

Utrecht University

a.c.j.dewolff@students.uu.nl
Student ID: 1289935

First supervisor: Dr. Hanna J. Hauptmann
Second supervisor: Dr. Chris Janssen
30th June 2025

Abstract

Person-specific data for affect recognition models is scarce, and generalized models often struggle with accuracy due to individual differences. Cluster-based personalization has been proposed as a compromise between generalized and fully personalized affect recognition from biosignals. This study investigated whether grouping participants by Big-Five personality profiles can improve user-independent prediction of continuous valence and arousal ratings. Cardiac and electrodermal features were extracted from two public datasets, AMIGOS and PhyMER. The datasets consisted of both short recordings (50-250 seconds) and long recordings (14-24 minutes). Three variants were evaluated: Baseline, which used only biosignals; Clusters, which added personality-based cluster assignments; and Traits, which added personality scores directly. Each variant was implemented with the Random Forests (RF) and Support Vector Regression (SVR) algorithms. As anticipated, the Baseline models across datasets struggled with large individual differences, resulting in limited explained variance of valence and arousal levels. However, informing models of personality-based subgroups did not improve the generalization of the models. Neither personality representation produced notable or systematic improvements over the Baseline. Generally, models trained on short segments exhibited improved performance compared to those trained on long segments. RF performed better for arousal, while SVR performed better for valence. Furthermore, our results confirm that recognizing continuous valence and arousal levels of unseen users is a nontrivial task, and individual variability likely diffuses the relationships between biosignals and affective labels. Future work should focus on improving the understanding of the physiological and psychological processes underlying self-reported affective states, using larger datasets and more advanced modeling techniques, such as Long Short-Term Memory (LSTM).

Contents

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Hanna, for taking me on as a supervisee halfway through the project and for her invaluable support and expertise throughout. I am also thankful to Anneloes and Egon for sharing their knowledge in affective signal processing. Lastly, I would like to thank my partner, family, and friends. I could not have completed this work without their tremendous support.

A note on generative AI: in writing this thesis, ChatGPT was used as a supplementary tool for exploration, explaining material, and refining text and code. All intellectual contributions remain my responsibility, and any generated content has been verified and edited before use.

1 Introduction

Recent advances in artificial intelligence have demonstrated their growing relevance in everyday life, largely due to their ability to process complex patterns that are too intricate to be revealed otherwise.

One application area is affective computing. Affective computing focuses on recognizing, interpreting, and adapting to human affective states, such as emotions or moods. One of the objectives is to introduce human characteristics into computing, through the subtleties of interaction [1]. Such subtleties include facial expressions, vocal tone, body language, or physiological processes (e.g., Heart Rate (HR), Skin Conductance (SC)). Hence, a great interest has surged in recognizing human affect automatically, based on these cues (e.g., [2], [3]). Furthermore, automatic affect recognition serves to advance Human-Computer Interaction (HCI), where systems aim to understand and respond appropriately to their users [1], [4]. Improved affect recognition models contribute to the development of more intuitive, empathetic, and responsive systems. Affect recognition is especially relevant in human-robot interaction, health monitoring, mental health support, and recommendation systems [1].

Affect is an umbrella term used to describe various internal states, including emotions and moods, among others [5]. Emotion can be described as a brief psychophysiological response, typically triggered by an internal or external event [6]. In contrast, mood is more stable, of less intensity, and longer duration. Mood is thus less closely tied to a specific stimulus, rendering it challenging to evoke and identify [5]. Both mood and emotion are associated with changes in physiology, cognition, and behavior, and can be operationalized using the Circumplex mood model [7]. The Circumplex model consists of two axes: valence and arousal. Here, valence indicates the tone of an affective experience, such as pleasant or unpleasant, whereas arousal describes how intensely the affective experience is perceived.

A fundamental characteristic of any affective process is that they are not just cognitive, but is also instantiated within the nervous system [8]. This notion drives the primary motive of the current research: affect is embodied. Whether physical changes are the result of cognitive processes or vice versa, affect is manifested by the sympathetic nervous system [9]. Hence, biological signals can act as indicators for affective processes within individuals. Accordingly, modalities such as Electroencephalogram (EEG), Electrocardiography (ECG), and Electrodermal Activity (EDA) have received considerable attention for their potential to enable automatic recognition of affective states [3], [10]. To use biosignals for affect recognition, annotations of affective states are required alongside the recordings. Most often, this type of data is collected in controlled settings, where participants are exposed to stimuli, such as images or videos that illicit a specific affective response. During stimulus exposure, biological sensors record data. Typically after exposure, participants provide self-reported ratings of affect (e.g., valence, arousal). This way, physiological recordings are labeled with subjective experiences.

How individuals respond to external stimuli is unique to each person, both physically and cognitively [9], [11]. From a physiological perspective, individuals differ in their body's capability to enact a physical response, which is influenced by factors such as sweat gland reactivity or baseline levels in cardiac or hormonal activity [12]. From a cognitive perspective, individuals differ in their perception and processing of events. Such differences are associated with personality traits and cognitive styles [13]. Unsurprisingly, personality traits explain affect perception and mental well-being very well [13]–[15]. For example, extroversion is generally associated with positive affect and social behavior. Likewise, neuroticism correlates with higher sensitivity to negative affect and anxiety, which is in turn reflected by increased skin conductance [16], [17]. Such findings underscore the utility of incorporating both biosignals and personality traits in affect recognition.

When developing affect recognition models, predictions can be based on data derived from a single individual or on aggregated data of multiple participants. The former, also known as personalized models, learn behavioral and physiological patterns unique to a single subject. Whereas the latter, or generalized

models, make generalized predictions across subjects [2]. Given the inherent variation among individuals mentioned previously, generalized models often perform poorly, whereas affect predictions based on data from a single person are more accurate [2]. However, training personalized models requires extensive data on a single person for reliable results. Large amounts of well-labeled data of a single person are rarely available, as data collection is costly and time-consuming. Likewise, in applications such as recommender systems, the same problem arises when a new user is introduced [2], [18]. To improve the predictive accuracy of generalized models, so-called personalization techniques have been developed. One such technique is *cluster-based modeling*, which reduces variance between individuals by training a model on subsets of people with similar characteristics. Similar people can be grouped on factors such as physiological patterns, age, gender, or personality traits [2], [18]–[20]. Previous research shows that cluster-based models generally outperform generalized models in recognizing affect, although the findings are mixed [19], [21], [22].

Due to the established relationships between personality, affective processes, and physiological changes (e.g., [13], [14], [16], [17]), personalization based on personality offers an interesting yet largely unexplored method for improving affect recognition. Therefore, this thesis investigates whether clustering participants by personality can enhance the generalization of continuous valence and arousal predictions to unseen individuals. The proposed solution is benchmarked against a fully generalized baseline model. To evaluate the added value of personality-based clustering, the solution is compared to a third model that directly includes personality traits. Two classical machine learning frameworks are used, Support Vector Regression (SVR) and Random Forests (RF), which are widely accepted and commonly applied in related studies [23], [24]. Lastly, the analyses are conducted on two separate open-source datasets to ensure the findings generalize well.

This thesis aims to make multiple contributions. First, we explore clustering on personality as a personalization method. Second, we aim to achieve a higher granularity of valence and arousal by treating them as continuous variables rather than binary. Third, by incorporating personality into affect prediction, we aim to elucidate the relationships between personality, affect, and biosignals. Lastly, we investigate the differences between brief affective responses (i.e., emotions) and longer affective experiences (i.e., moods) through short and long recordings. Finally, the performance of two different machine learning algorithms (e.g., RF, SVR) is evaluated.

The datasets used in this study comprise continuous recordings of physiological metrics, including cardiac and electrodermal activity. These recordings are annotated with affective states in terms of valence and arousal levels. Physiological recordings were either of short (50-250 s) or long duration (16-24 minutes). Finally, the datasets include assessment of personality traits, such as the Big Five Inventory (BFI) [25]. The three datasets that meet these requirements are AMIGOS [15], ASCERTAIN [26], and PhyMER [27].

This thesis is organized as follows: Chapter 2 provides the theoretical background knowledge and a review of related work on affect, biological signals, personality, and model personalization. Chapter 3 outlines the methods, describing the datasets, preprocessing pipeline, feature extraction, model design, and evaluation. Chapter 4 presents the experimental results of the models. Finally, Chapter 5 concludes with a discussion of results, the limitations of this study, and suggestions for future work.

2 Theoretical background and related works

This section provides the theoretical foundation of the research. First, a brief introduction to the field of affective computing is provided. Next, theories on emotion and mood as affective constructs are discussed. Next, the practices within affective signal processing and physiological theory are outlined. Then, individual variability and personality theory are discussed. Lastly, the issue of generalized models is raised, followed by the research questions and a discussion of available datasets.

2.1 Affective computing

To have computational systems recognize and adapt to human affective states, constructs of affect, emotion, and mood ought to be distinguished first [1]. Traditionally, defining and understanding affect was a point of interest only to the psychological and philosophical field [10]. An important contribution that changed this was Picard's work on affective computing [1]. Her publication sparked an interest among computer scientists in developing computer systems that can adapt to human emotions and cognition. Picard distinguished three different objectives within affective computing applications: 1) Systems that detect the emotions of the user, 2). Systems that express what a human would perceive as emotion and 3). Systems that actually "feel" emotion. The current thesis most closely situates itself within the first objective.

In practice, affective computing has the potential to transform how we interact with computers. Some examples include healthcare, where affect recognition software can improve mental health monitoring, either for self-help purposes or automatic intervention planning [28]. Next, affect detection can enhance the functionality of social robots and avatars by improving adaptivity and empathy in interactions. For example, in elderly care facilities, companion robots have been found to have positive effects on mood, loneliness, and social connection with other elderly individuals [29]. Lastly, affect detection models could potentially enhance recommendations in entertainment platforms such as streaming services. Different application areas have varying ethical considerations. Consequently, there is skepticism towards closing the gap between the highly emotional human and the emotionally inert computer [10]. Concurrently, many authors argue that artificial intelligence must incorporate affect as a fundamental aspect to advance to its next stage of evolution [1], [30].

2.2 Defining affect, mood and emotion

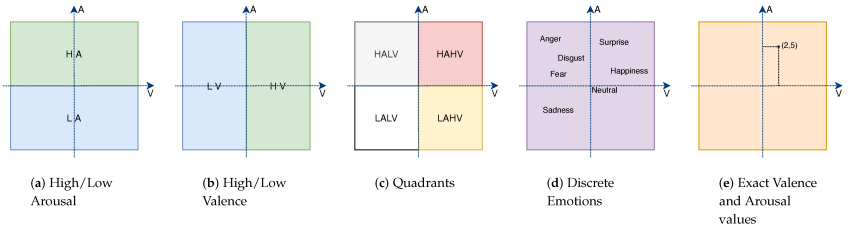
For an automated affect recognition system to be effective, the concepts of affect, mood, and emotion must first be operationalised [1]. First and foremost, the term 'affect' is broad and encompasses various processes beyond just emotion [30]. Illustrating this, Russell's work describes pure affect as a neuro-physiological state, a non-reflective feeling present in moods and emotions [31]. Yet, these terms are often used interchangeably in literature, complicating its definition, interpretation, and application [30], [32].

Emotion, for instance, can be regarded as an affective construct. Emotion is a brief, coordinated response of the body to an event or trigger, whether originating internally or externally. For example, joy or anger [6]. Alternatively, mood is described as a diffuse affective state of low intensity, but relatively long duration, and often without a direct cause. For example, contentment or irritability [33]. Compared to emotion, mood is typically experienced less intensely and for longer durations [5]. While emotional responses tend to last a few seconds, moods can span over minutes, hours, or even days. Similarly, moods are more persistent, diffuse, and not necessarily tied to a specific event that causes their onset. Instead, moods portray an individual's overall affective state for an extended period. Lastly, mood is more an internal experience, while emotion is more external and visible to others [5]. With these considerations,

eliciting and pinpointing mood is generally more challenging in a laboratory setting.

There are different approaches to operationalizing affective constructs. For example, emotions are sometimes discretely organized, for example in Paul Ekman's theory of basic emotions [34]. Ekman initially differentiated a set of six basic emotions that are universally recognized across cultures: happiness, sadness, fear, anger, surprise, and disgust [34]. Arguably, using discrete categories to describe any affective construct neglects their intensity and dynamic, time-varying nature [4]. With this objective, dynamic conceptualizations of affect have been established. For example, the Circumplex model or Positive and Negative Affect Schedule (PANAS) [28], [35]. The Circumplex Mood model approaches affective states using two bipolar dimensions: pleasure-displeasure and arousal-sleepiness [7]. The pleasure-displeasure axis describes the *valence* of an experience, while the arousal-sleepiness axis represents the level of activation, or *arousal*. Affective concepts are distributed around the circumference of this circle, such as excitement (high valence and high arousal) and depression (low valence and low arousal) [7]. However, the Circumplex model ignores the possibility of simultaneous conflicting emotions. Alternatively, the PANAS model addresses this disadvantage by including separate measures for positive and negative affect [36].

The figure below illustrates different approaches to conceptualizing affective states in terms of valence and arousal. In the current research, the targets are continuous valence and arousal values, as shown in subfigure (e).



(a) High/Low Arousal (b) High/Low Valence (c) Quadrants (d) Discrete Emotions (e) Exact Valence and Arousal values

Figure 1: Different approaches to affect recognition in literature using valence-arousal dimensions. Five ways of representing the two-dimensional valence-arousal space are shown, where the y axis represents arousal, and the x axis represents valence. Subfigures (a) and (b) display a binary classification of high/low arousal and high/low valence, which is most commonly seen in affective literature. Subfigure (c) displays the four quadrants (high-arousal/low-valence, high-arousal/low-valence, low-arousal/low-valence, low-arousal/high-valence). Subfigure (d) depicts discrete emotions, positioned on the V/A space. Subfigure (d) shows exact valence/arousal values, which are the conceptualization of focus in this thesis. Figure adapted from [37]

As a consequence of the challenges associated with modeling mood, such as its longer duration, lower intensity, and the difficulty of eliciting it, mood has received limited interest within the field of affect recognition [5]. However, one particular contribution of Katsimerou and Redi is notable. Their study proposed a computational model that defined mood as a series of recognized emotions on a valence/arousal space, which memorized previous states yet exponentially discounted their importance in the overall mood prediction [5]. Their model not only took duration as a characteristic of mood into account, but also the layered, unfolding nature of mood.

2.3 Affective signal processing

An important reason why the Circumplex model is so commonly used for affect recognition research is that it approaches affective experiences as psychophysiological [5]. This is critical to understanding mood, emotion, and other affective processes; they are not just cognitive states, but are also instantiated within the body [8]. Affective constructs are therefore tangible to a certain extent, as they are reflected by the nervous system. In particular, the activation of the Autonomic Nervous System (ANS) regulates bodily changes such as facial expressions, respiration, voice tremors, body posture, cardiac activity, or

sweat excretion [8].

As such, the field distinguishes research towards the areas of speech, vision, and biosignals [30]. Affective Signal Processing (ASP) is concerned mainly with the latter, and is correspondingly defined as processing biosignals related to affect. ASP relies on the same notion that emotions are manifestations by the sympathetic nervous system [9]. Consequently, the relationship between physiology, cognition, and emotion has been the subject of considerable academic interest [8], [9], [30].

Two advantages of using biosignals for affective computing can be noted, as opposed to speech or vision. Firstly, biological processes can be easily measured using non-invasive, commercial sensors, which are now commonly embedded in most smart devices. In addition, biosensors are unaffected by varying light and auditory conditions, rendering them reliable and convenient in practice [38]. Secondly, biosignals are free from social masking, meaning they cannot be easily manipulated by the participant [39]. Similarly, the processing of biosignals can reveal more subtleties, which may not be detected in body language or speech [30]. Furthermore, different biological processes can serve as indicators of affect, and the ones used depend on the application area [30]. The most popular non-invasive biosignals are reviewed in the following paragraphs, including cardiovascular and sweat gland activity.

First, cardiovascular activity is an effective psychophysiological indicator of arousal and stress [40]. This relationship is established through the bidirectional communication between the heart and the brain, which influences perception, emotion, and intuition among others [41]. More specifically, the ANS, controlled by the medulla in the brainstem, regulates heart rhythm. The ANS controls its activity through two axes, the sympathetic and parasympathetic nervous systems. During stress or excitement, the sympathetic division of the ANS accelerates heart rate through the release of the neurotransmitter norepinephrine. This process is part of the "fight or flight" response and primes the body for action [42]. In contrast, in calm or relaxed states, the parasympathetic division inhibits the heart by releasing acetylcholine, entering the "rest and digest" state [42].

One measurement for cardiac activity is Electrocardiography (ECG). The ECG records electrical potential generated by the heart muscles as it goes through cardiac cycles, or heartbeats [8]. A normal ECG cycle consists of three waves. The first wave, the P wave, represents atrial depolarization, which occurs just before the upper heart chambers contract. The second wave, the QRS complex, is the most prominent and appears as the lower chambers of the heart contract, generating a peak in the signal. Finally, the T wave reflects the repolarization, or the recovery of the chambers [41]. From this signal, Inter-Beat-Interval (IBI), Heart Rate (HR), and Heart Rate Variability (HRV) can be extracted. IBI is the interval between successive R peaks within the QRS complex. Based on this, HR can be extracted by taking the reciprocal of the IBI and represents the number of beats per minute (bpm). Lastly, HRV quantifies the variability in the IBI over time [41]. ECG is typically measured by electrodes placed on the chest or limbs [8]. Pairs of electrodes measure the potential difference between two locations, forming a lead. The number of electrodes required depends on the application area. In psychophysiological applications, three electrodes are often sufficient. In such a configuration, two electrodes are typically placed at each arm crook and one reference electrode on the foot [15]. ECG in particular is a favorable modality in affect recognition, as it can be measured from any part of the body and contains relatively higher amplitude than other biosensors [41].

In a similar vein, cardiac activity can also be measured using Photoplethysmography (PPG). PPG is a non-invasive, optical measure for Blood Volume Pulse (BVP) [43]. As the heart pumps blood through the body with each cardiac cycle, vessels expand [8]. This expansion, or increase in blood volume, can be detected by directing an LED light onto the skin. In turn, variations in light absorption provide information about pulsatile blood flow [8]. Similarly to ECG, the PPG signal represents each cardiac cycle and appears as a waveform, the shape of which can differ from subject to subject [8]. The signal contains an AC and a DC component [43]. The DC component represents the baseline of blood flow. The

AC component is related to the time-varying cardiac activity. The amplitude of the AC component is directly proportional to the pulse pressure, and can be used to derive IBI, HR, and HRV [43]. Lastly, PPG is commonly integrated into wearable technologies, such as smartwatches, simplifying its measurement. However, due to the indirect nature of PPG, it is more sensitive to artifacts and is often unsuited for identifying waveform components [8].

Another powerful index for affective processes is sweat gland activity, otherwise known as Electrodermal Activity (EDA) or Galvanic Skin Response (GSR). In particular, changes in EDA reflect activity of the ANS, in particular the sympathetic branch [8]. As the Sympathetic Nervous System (SNS) is activated by emotional stimuli, norepinephrine is released and binds to sweat gland receptors, inducing sweat secretion [44]. Increased sweat on the skin increases electrical conductivity [45]. EDA signals do not distinguish between positively and negatively experienced arousal and are therefore an unreliable measure for valence [44]. EDA is measured through electrodes that detect the change in conductance between two points over time. The raw signal consists of two components: tonic and phasic activity. Phasic activity, or Skin Conductance Response (SCR), reflects momentary changes in the SNS, [8]. The phasic activity (i.e., a peak) typically occurs 3 seconds after a stimulus and flattens out afterward. In contrast, tonic activity reflects the more gradual changes in skin conductance, independent of direct stimuli [8]. Furthermore, there are considerations when using EDA as a biosignal. First, its relatively slow response makes it less suitable for detecting rapid changes. Secondly, like other biosignals, individual differences exist in EDA, especially in the electrodermal lability or stability of the response [8]. This may be attributed to the differences in how people process information [8]. EDA recording can also vary depending on the placement location of the electrodes [8].

Other biosignals commonly used but excluded from the present scope include Electroencephalogram (EEG) for brain activity, Electrooculography (EOG) for eye movement, Electromyography (EMG) for muscle movement, and skin temperature (TEMP).

2.4 Individual differences and personality

How individuals respond to external stimuli is unique to each person, both physically and cognitively [9], [11]. From a physiological perspective, individuals differ in terms of their physiological makeup, such as electrodermal liability, baseline levels in cardiac or hormonal activity [8]. For instance, a higher number and density of sweat glands is associated with increased anxiety, as increased sweating can lead to increased association with feelings of stress or anxiety [46]. Similarly, variations in HRV between individuals are related to stress and regulation of affect. For example, high HRV is associated with anxiety, depression, and irritability [12]. In other words, variance in physiological characteristics at least partially accounts for the variance in affective responses.

Put differently, one's physiological makeup can be treated as the environment influencing the cognitive appraisal of an event [11]. This notion follows the theory of constructed emotion, which states that affect is not just the result of a stimulus itself, but is shaped through its environment; the interactions between physiological states, prior experiences, and contextual interpretation [11]. Just as physiological differences can be considered an environment in which one appraises situations, personality traits also create a contextual lens through which individuals interpret events [11]. Accordingly, previous research widely established that personality explains affect perception and mental well-being very well [13]–[15]. Given its role in shaping affect perception, the nature of personality and its structure are clarified first.

Personality can be defined as the individual differences in patterns of affect, cognition, and behavior [47]. One's personality is characterized by a set of traits that are consistent across situations and over time. In this context, a trait is a pattern of correlated 'reaction habits' and describes a set of behavioral tendencies that correspond to one's personality [47]. Viewing traits as stable dispositions, lexical studies

used factor analysis to detect consistent patterns of personality-related adjectives (e.g., [48]). This, in particular, marked a shift from theory-driven to data-driven psychology. Later, McCrae and Costa [49] refined these factors into the Five-Factor Model (FFM), which is known today. The FFM includes openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism. For an overview, see Table 1.

Trait	**Characteristics**
Openness	Imagination, curiosity, variety of experience versus consistency, pragmatism, and analytical thinking
Conscientiousness	Organization, self-discipline, structure versus adaptability, spontaneity, and disorganization
Extroversion	Sociability, energy, approachability versus independence from social environment, detachment, and reservation
Agreeableness	Kindness, empathy, harmony versus self-centeredness, harshness, and argumentativeness
Neuroticism	Negative emotionality, anxiety, instability versus emotional stability, calmness, and composure

Table 1: Personality traits and characteristics of both ends of the spectrum [49], [50] Summary of the Five-Factor Model (FFM) of personality, which includes openness, conscientiousness, extroversion, agreeableness, and neuroticism. Traits are associated with distinct patterns in affect, cognition, and behavior. These characteristics help explain variations in affective responses and are commonly measured using validated self-report questionnaires such as the BFI or NEO-PI-R.

Several scales have been developed to measure different dimensions of personality. Most often, personality is measured through self-report questionnaires. Perhaps the most well-known is the Big Five Inventory (BFI) [25], which consists of 44 items. Next, the NEO Personality Inventory is another validated scale, containing either 60 or 240 items, depending on the version [51]. Alternatively, the Newcastle Personality Assessor (NPA) is commonly used, comprising 12 items [52]. These measures treat personality as a continuum, as individuals can exhibit both poles of a trait, for example, extroversion and introversion [25]. Consequently, the results of personality scales such as the BFI yield a score for each of the dimensions measured, providing a summary of one's personality profile. Example items from the BFI include "I feel comfortable around people" for extroversion, or "I pay attention to details" for conscientiousness.

In light of the descriptions provided in Table 1, it is unsurprising that personality is associated with affective experiences and physiology [13]–[15]. Personality traits reflect inherent variations in how people perceive or react to positive and negative emotional stimuli, leading to differences in general levels of positive or negative affect [13]. More specifically, Gross et al. [53] proposed the affect-reactivity model, which states that extroverts are more reactive to positive stimuli than introverts. Similar affective responses are expected when presented with neutral or negative stimuli [53]. For example, pleasant affect is strongly related to extroversion. Similarly, neuroticism correlates with a greater sensitivity to negative emotions and anxiety, reflected by increased skin conductance and decreased HRV [8], [17], [53]. Therefore, incorporating both personality traits and physiological markers may help explain a greater proportion of variance in mood.

In summary, physiological makeup and personality are factors that at least partially account for the variability in affective experience among people [12]–[14], [46]. Therefore, including both biological signals and personality is expected to enhance the accuracy of mood prediction models.

2.5 Generalized versus personalized models

In addition to personality and biological characteristics, factors such as gender, age, eating and sleeping patterns, acute circumstances, and cultural differences can help explain how the same stimuli can elicit different physiological and affective experiences among individuals [9], [11]. The notion that individuals respond differently to the same events, due to inherent variation among people, explains why user-specific predictions in affect recognition models tend to be more accurate than user-independent predictions [2].

User-specific, or personalized prediction models, contrast with generalized models, which use aggregated data from multiple individuals indiscriminately to learn and predict patterns. Generalized models are also referred to as user-independent models [2]. However, a significant limitation of personalized models is that they require extensive data on a single user. Not only are large quantities of well-labeled data difficult to acquire, but user-independent models also aim to learn patterns across people, enabling inference on unseen users [2], [18]. In application, this translates to reduced calibration time for new users, also regarded as the cold-start problem.

While personalized models require extensive data on individual users, generalized models often fail to predict affect accurately. Therefore, personalization approaches exist as a middle-ground solution between fully personalized and fully generalized models. There are several approaches to personalization, as defined by Han et al., [2]. Firstly, *hybrid modeling* utilizes the unseen user's data only partially, integrating it with the rest of the training data from N-1 participants. The training and testing are then conducted iteratively for each participant, once as the unseen user. Secondly, *fine-tuning* entails pretraining a generalized model, which is then further adapted using a smaller amount of data specific to the target user. Alternatively, group-based personalization approaches exist. For instance, *cluster-based* entails training models separately for subgroups of users with similar characteristics, so that predictions are stratified across these similar subgroups. This way, personalization occurs as the variance between participants is reduced. Similarly, *multi-task learning* involves training a single deep neural network on multiple tasks simultaneously, thereby tailoring the learning process to each similar group while sharing information among the groups [2], [18]. Both cluster-based and multi-task personalization can be tailored to be completely user-independent, meaning the model will not use any part of an unseen user's data [2].

Apart from a clear advantage in predictive accuracy of fully personalized models, there is no clear performance hierarchy among different personalization approaches [2]. However, some methods are more suited to a specific context than others. In this thesis, two constraints are faced. First, the objective is to predict unseen users, which rules out hybrid modeling and fine-tuning [2]. Second, this thesis uses only classical machine learning methods rather than layered architectures that underpin multi-task learning [2], [18]. The motivation for using classical machine learning as opposed to more advanced architectures was data scarcity. Deep learning algorithms such as Neural Networks or Transformers require massive datasets to avoid overfitting due to the high number of parameters. Additionally, classical machine learning models provide greater interpretability and explainability. The complex nature of advanced machine learning models renders them difficult to interpret, and they fail to provide insight into the learned patterns influencing the predictions [54].

The findings on the effectiveness of clustering as a personalization approach tend to be mixed [18], [19], [22]. For example, Can et al. [18] compared a clustering approach based on perceived stress scores to both a personalized and generalized approach for ambulatory stress detection using HR and EDA data [18]. They found that user-specific models achieved the best accuracy results, followed by the cluster-specific and generalized approaches, respectively. Similarly, Tervonen et al. [19] compared user-specific, cluster-specific, and generalized models for detecting stress in participants. Clustering was based on individual distributions of the extracted features. They confirmed that user-specific models performed better than generalized approaches, yet cluster-specific models showed similar performance to generalized models [19].

Adler et al. [22] examined longitudinal data and assessed the efficacy of clustering based on digital trace data. Although personalization generally increased the alignment of training and testing data, clustering did not guarantee improved performance in every instance. For example, increasing cluster size generally increased prediction errors [22]. Another systematic review by Han et al. of various personalization approaches [2] found that cluster-specific modeling slightly outperformed generalized models, though this was not consistent across datasets or algorithms. Therefore, the efficacy of clustering as a personalization approach seems to depend on the quality of the clusters and datasets, which may explain the varied results.

Research towards clustering based on personality is scarce. One particular publication by Yusef et al. [21] employs k-means clustering, based on the Big Five personality traits, for collaborative filtering. Collaborative filtering is a standard algorithm in recommender systems, which often faces the cold-start problem when limited data is available on new users. The authors reported that their proposed method achieved better results in reducing the mean absolute error and increasing precision compared to other methods [21]. A similar study confirms these findings [55], supporting the potential of using personality traits to personalize predictions when limited data is available.

The fact that this area remains underexplored motivates the approach presented in the current work. Clustering based on personality data has several advantages. First, obtaining personality data is relatively straightforward, for example, by the BFI. Similarly, personality scores are relatively stable and unlikely to change over time [56]. Additionally, clustering on other physiological features may be problematic. For example, clustering based on physiological signals could unintentionally reflect sensor placement, demographic factors, or day-to-day variations rather than stable differences among individuals. In addition, clustering based on physiological features from the same trials used for testing may inadvertently lead to information leakage between the input features and the target labels.

2.6 Affective databases

A wide range of databases exists specifically for affect recognition. Some of the first studies in affect recognition focused solely on the visual modality, such as facial expressions or body posture [57], [58]. Later, studies emerged that included other modalities as well, such as biosignals [4], [15], [59]. Affective datasets are most commonly gathered in controlled settings, where affective experiences are elicited using stimuli. Stimuli can be affect-inducing videos or pictures, but they can also entail affect-inducing tasks, such as group discussion or cognitive tasks. For example, a study by Schmidt et al. [59] elicited stress in participants through public speaking and an arithmetic task [59]. However, not all affective databases are collected in a laboratory setting. Attention is increasing towards long-term studies that follow cohorts of people instead. The participants in such studies typically wear biosensing smartwatches throughout the study and complete daily surveys or Experience-based Sampling (ESM) questionnaires for multiple consecutive days [23], [60]. In such studies, affect is not elicited, but is expected to occur naturally throughout everyday life.

Both naturalistic and laboratory studies rely on ground truth labeling. These labels bind affective states to the physiological recordings, which is necessary for the training and testing of recognition models [10]. The affective construct targeted varies by the dataset; some affective datasets and research treat emotions as static, discrete categories, while others approach affect as a dynamic construct, for example, by measuring affect on a valence-arousal space. In a laboratory setting, the stimulus itself can be labeled or the affective state that the stimulus elicits in the participant. Labeling of the stimulus is often performed by external researchers before the experiment [2], [27]. In this case, the stimuli are typically assigned to one for example discrete emotions (e.g., 'fear' or 'anger'), a quadrant in the valence-arousal space (e.g., 'high arousal/low valence'), or a specific point within that space [15], [26].

Alternatively, the participant's affective response to the stimulus is labeled. This can be either internal or external [4], [15]. In external annotation, outside evaluators label, for example, video recordings of the participants as they engage with the stimuli. A study by Miranda-Correa et al. [15] verified this approach and found high inter-rater correlations between external evaluators, as well as high correlation between internal and external annotations [15]. However, internal annotation is more commonly used. Internal innotation requires the participants to annotate their affective experiences through self-report. Oftentimes, before and after the presentation of stimuli, users complete a Likert-based questionnaire aimed at assessing their retrospective affective state [4], [35], [61]. See section 2.2 for an in-depth discussion of different measures. Moving on, while retrospective annotation is more straightforward, stimuli can also be annotated for the entire stimulus duration [4]. For example, the CASE dataset by Sharma et al. [4] developed a joystick-based annotation interface that spans the valence-arousal space of the Circumplex model throughout the entire stimulus duration. In contrast, discrete dynamic annotation requires participants to annotate their affective state in certain time intervals, usually before and after stimuli [15], [26]. The latter is more commonly found, as it simplifies both the annotation and analysis processes.

Most studies, although they have obtained continuous annotations (e.g., Circumplex model) convert these dynamic scales to categories. In turn, binary prediction models are used to predict high valence/low valence or high arousal/low arousal [23], [37]. Reducing the scale to a binary range can be problematic, because everyone within a class is treated as if they are equal. For instance, individuals whose ratings are similar but are on either side of the threshold are categorized differently, and vice versa. Alternatively, continuous prediction models achieve higher granularity by predicting values along the range of valence or arousal levels, rather than dichotomizing them [10], [37]. To predict continuous variables, regression-based frameworks are needed [37]. For example, a study by Galvao et al. [37] utilized EEG signals to successfully predict exact valence and arousal levels in a subject-independent scenario [37]. Below, a Table of all public affective datasets that include personality measures are shown 2.

#	Dataset	Signal	Ground truth	N	Duration	Personality
[15]	AMIGOS (2021)	EEG, ECG, EDA, body posture & facial expressions	1-9 SAM for Arousal, Valence, Engagement, Liking, Familiarity, Ekman's basic emotions, per stimuli annotation	40	<250 s (short) 14-22 min (long)	BFI, PANAS
[26]	ASCERTAIN (2018)	EEG, ECG, EDA, facial expressions	1-7 SAM for Arousal, Valence, Engagement, Liking, Familiarity, per stimuli annotation	58	51-127s	BFI
[27]	PhyMER (2023)	EEG, PPG, EDA, HR, TEMP	1-9 SAM for valence/arousal and seven basic emotions, per stimuli annotation	30	61 to 181s	NPA
[62]	BIRAFFE2 (2022)	ECG, EDA, facial expressions	Arousal, Valence, per stimuli annotation	102	6s	NEO-FFI
[23]	K-EMOPHONE (2023)	GSR, PPG, HST, ACC, step count	Arousal, valence, stress, attention, task disturbance, 16 times a day	77	7 days	BFI
[60]	DAPPER (2021)	HR, GSR, ACC	TIPI-CI, PANAS, Arousal, Valence, 12 times a day	88	5 days	BFI

Legend: EEG: Electroencephalography; ECG: Electrocardiography; ACC: acceleration; TEMP: Skin Temperature; BVP: Blood Volume Pulse; HST: Heart Sound Technique; SAM: Self-Assessment Manikins; PANAS: Positive And Negative Affect Scale; NPA: Newcastle Personality Assessor; BFI: Big-Five Inventory; TIPI-CI: Ten-Item Personality Inventory - Chinese version

Table 2: Detailed overview of candidate affective databases. Summary of publicly available affective databases that incorporate physiological recordings for emotion or affect recognition. The datasets vary in signal modalities, ground truth labeling strategies, participant sample sizes, and study durations. Affective annotations differ in terms of temporal granularity and whether they target continuous dimensions (e.g., valence, arousal) or discrete emotion categories. Only datasets that also include personality trait measures were considered in this overview.

2.7 Research gap and questions

In summary, generalized models struggle to predict affect in users accurately due to individual variability [2]. While personalized models perform better by accounting for individual differences, they require extensive user-specific data. Research shows a promising alternative: cluster-based modeling. Cluster-based modeling improves affect recognition by grouping people with similar characteristics, thereby reducing variance within the training data [2]. Furthermore, existing research has primarily focused on the binary classification of valence and arousal, which neglects the broad spectrum of affective states [23], [37]. Therefore, in this study, valence and arousal are treated as continuous variables. Similarly, most previous studies have focused on short-term emotional responses only, with limited research on longer-lasting

affective processes such as mood [5].

To the best of our knowledge, no prior research exists that addresses these gaps and uses personality-based clustering in the context of affect recognition. Therefore, the current thesis proposes cluster-based personalization for predicting continuous valence and arousal levels in unseen participants. Additionally, we evaluate input features extracted from recordings of both short (50-250 seconds) and long (14-24 minutes) duration. To achieve this, two supervised machine learning algorithms (Random Forests (RF) and Support Vector Regression (SVR)) that utilize biosignals as input features are employed, and the results are validated using two different datasets.

> ***RQ1:*** *Does including personality-based cluster information improve the generalization of exact valence and arousal predictions to unseen participants, using biosignals recorded from both short and long recordings?*
>
> ***RQ2:*** *Which machine learning algorithm, Random Forests or Support Vector Machines, performs best in the context of recognizing continuous valence and arousal levels from physiological features?*

To answer these questions, two open-sourced affective datasets containing biosignals and personality records are used.

3 Methods

This section describes the methodological approach for answering the research questions. First, the dataset selection process is described. Next, the data preparation pipeline is discussed. This preparation included preprocessing of the annotation data, biosignals, and personality scores. Afterwards, time-domain features were extracted from the biosignals. Next, the clustering based on personality is described. For the experiment, the creation of different affect recognition model variants is discussed, along with the training and testing procedures. The variants include the baseline model, our proposed cluster-informed model, and a model with personality traits for comparison. Lastly, the used algorithms, Support Vector Machines (SVM) and Random Forests (RF), along with their evaluation metrics, are discussed.

3.1 Datasets

A list of criteria was created to guide the selection of the datasets. The criteria were ordered in importance and partitioned into 'basic' and 'additional' categories. See Table 3 for an overview of all criteria.

Category	Label	Description of Criteria
Basic Criteria	C1	The dataset must be open-source.
	C2	The dataset must contain continuous recordings of physiological markers such as ECG or EDA during affect elicitation.
	C3	The dataset must include dynamic annotations for affect, for example, using the Circumplex model or PANAS.
	C4	The dataset must provide a validated personality test, such as the Big-Five Inventory (BFI).
Additional Criteria	C5	The dataset includes recordings spanning at least 60 seconds.
	C6	The dataset demonstrates good construct validity and data quality.

Table 3: Ordered list of dataset selection criteria Divided into two categories, the basic criteria represent the minimal criteria for conducting our proposed research. From a large array of datasets, after a evaluation on the basic criteria, only six remained (BIRAFFE2 [62], K-EMOPHONE [23] and DAPPER [60], AMIGOS [15], ASCERTAIN [26] and PhyMER [27]). After evaluating the additional criteria, which ensured the methodological soundness of the datasets, two datasets were selected: AMIGOS [15] and PhyMER [27].

The basic criteria were fundamental to this research; for instance, a dataset that is not open-source cannot be used. Selection on the basic criteria (C1-C4) resulted in six datasets: BIRAFFE2 (2022), [62], K-EMOPHONE (2023) [23] and DAPPER (2021) [60], AMIGOS (2021) [15], ASCERTAIN (2018) [26] and PhyMER (2023) [27]. The additional criteria ensured the quality and relevance of the dataset.

Criteria C5 required recordings to span at least 60 seconds. This was relevant as the reliability of certain features decreases when recordings are considerably less than 60 seconds, for instance, metrics like Heart Rate (HR) and Heart Rate Variability (HRV) [63]. In addition, longer periods of stimulus presentation allow for the unfolding of more complex and stable affective states [10]. To that end, BIRAFFE2 [62] was unsuitable for our aims, as the biological recordings were 6 seconds for each trial.

Criterion C6 regarded the construct validity and data quality of datasets. Construct validity is the degree to which a test or measure accurately measures the theoretical concept it is intended to assess [64]. Poor construct validity can lead to inconsistent and biased data, which compromises the reliability of the results. The self-reported affect questionnaire in the K-EMOPHONE dataset [23] raised concerns regarding face validity. In each of the 16 daily questionnaires, participants were prompted to report their emotion since the last questionnaire, which ranged from 2 hours to 30 minutes. As emotions are transient

and typically do not persist for extended periods, this suggests that the questionnaire captures a different construct than the one formulated. In the DAPPER [60] database, similar concerns emerged. Then, both K-EMOPHONE [23] and DAPPER (2021) [60] are naturalistic studies. As a consequence, both datasets likely contain large proportions of unbalanced and noisy data. Similarly, self-reports throughout the day are difficult to align with biological recordings, which limits their validity as ground truth annotations. Overall, these points highlight methodological shortcomings that compromise the quality of the data. Therefore, both K-EMOPHONE [23] and DAPPER [60] were excluded from the present scope.

This selection process resulted in three suitable datasets: AMIGOS [15], ASCERTAIN [26], and PhyMER [27]. To the best of our knowledge, there are no other open-source datasets that meet the criteria and fit our purpose. However, the ASCERTAIN database was ultimately not made available. Therefore, we proceeded with the AMIGOS and PhyMER datasets only. AMIGOS, or "A dataset for Multimodal research of affect, personality traits and mood on Individuals and GrOupS", is a lab-based study investigating affect, personality traits, and mood through physiological signals in both individual and group settings. This dataset was particularly relevant, as it included both long (14-24 minutes) and short (50-250 seconds) stimulus presentations. Next, PhyMER, short for "Physiological Dataset for Multimodal Emotion Recognition", is a Korean laboratory-based study that collected physiological signals and personality traits. PhyMER's study included only brief stimulus presentations. For a detailed overview of the experimental setups in both AMIGOS and PhyMER, see Table 4.

	AMIGOS	**PhyMer**
Participants	N = 40; (13 female), aged 21–40 (mean 28.3). Ethnicity not disclosed.	N = 30; (15 female), aged 20–30 years. Korean students.
Experimental design	Participants watched 16 short clips; 37 watched an additional 4 long videos. In long videos, 20 watched stimuli alone, while 17 in groups of three). Retrospective annotation.	Participants watched 23 video clips in three 15-minute blocks. Each trial: stimulus presentation, annotation, neutral stimulus (1 min) with color bars and calm music.
Stimuli	16 short (<250 s) and 4 long videos (14–24 min) categorized into each quadrant of the V/A space.	23 clips (61–181 s) adapted to Korean audiences to evoke specific affective responses, rated on V/A and basic emotions.
Biosignals and sensors	EEG, ECG, EDA; frontal/full-body videos. ECG and EDA measured with Shimmer 2R. ECG electrodes at arm crooks and ankle reference (256 Hz). EDA electrodes on index and middle finger (128 Hz).	EEG, EDA, PPG, and TEMP using the Empatica E4 wristband. PPG (64 Hz); EDA (4 Hz) via two dry electrodes on wrist.
Personality	44-item BFI; 20-item PANAS (post-tests).	12-item Korean-adapted NPA (pre-test).
Ground truth	9-pt SAM-based scale for valence, arousal, control, familiarity, liking, basic emotions (internal assessment). External assessments by three examiners.	9-pt SAM-based scale for valence, arousal, basic emotions.
Total trials	788	690

Legend: EEG: Electroencephalography; ECG: Electrocardiography; EDA: Electrodermal Activity; PPG: Photoplethysmography; TEMP: Skin Temperature; PANAS: Positive and Negative Affect Schedule; BFI: Big Five Inventory; NPA: Newcastle Personality Assessor; SAM: Self-Assessment Manikin.

Table 4: Detailed overview of AMIGOS and PhyMer datasets Summary of the selected dataset's characteristics, such as participant information, experimental design, stimuli used for affect elicitation, and which biosignals were measured with which sensor equipment. In addition, the personality inventory used and ground truth assessment. Finally, the total number of trials (participants * trials) is reported.

3.2 Missing data

Before analysis, several participants were missing from the datasets. In AMIGOS, Participants 8, 24, and 28 did not complete the long-video protocol and therefore contributed only sixteen trials each. Additionally, participant 32 did not have labels for the long video experiment, and the short video experiment

data for participant 9 were missing. Additionally, participants 18 and 28 were missing personality assessments, and to ensure a fair comparison between the baseline and our proposed model, both were omitted from the models. During EDA feature extraction, three participants' trials were unsuitable for reliable EDA peak detection due to data corruption. For these trials, median imputation was applied to replace approximately eight missing values.

In the PhyMER dataset, eight samples of subject 10 (recordings 9-16) were missing due to device malfunction during the experiment [27].

3.3 Annotation processing

When training any supervised learning model, the ground truth of the target variable is necessary to learn the relationship between input features and the target. Ground truth annotation, therefore, refers to the process of labeling data with accurate information used as a reference. In the datasets used in this study, biosignals were annotated with self-reported valence and arousal levels. In regression, the continuous values were used directly.

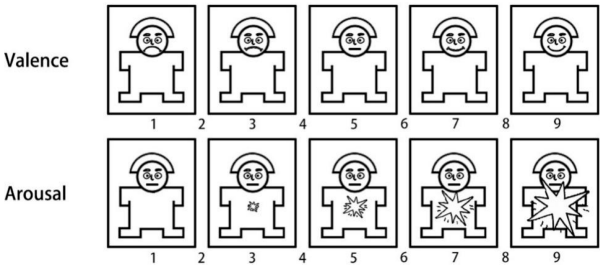


Figure 2: 1-9 point Self-Assessment Manikins (SAM)-based annotation for valence and arousal. Depiction of the valence and arousal rating system for participants in both datasets. In valence, a low score (1) represents unpleasant/negative affect, while a high score (9) indicates pleasant/positive affect. For arousal, low (1) represents no activation, while high (9) represents high activation. Figure adopted from [65]

3.3.1 Binarisation

Our primary objective was to predict continuous valence and arousal levels using regression. However, we also train binary classification versions of our algorithms. This was to enable a direct comparison with the authors of AMIGOS, as they reported only binary classification results ([15]. For the classification experiment, binary targets were derived from the continuous labels in the AMIGOS data. Matching the approach in [15], the labels were divided into low and high bins by using the median value of each affective dimension as the threshold. See equation 1 for an illustration of the decision boundary.

$$\text{Target bin} = \begin{cases} 1 & \text{if value} > \text{median}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The resulting binary columns `Arousal_bin` and `Valence_bin` were appended to the table with labels and used only for the classification experiments; all regression models continued to use the raw ratings.

3.4 Biosignal processing

The following section describes the pipeline for transforming raw biological data into interpretable input for the models. The preprocessing pipeline included winsorization, filtering, and scaling. Next, features

were extracted from the preprocessed data. For both preprocessing and feature extraction, the `NeuroKit2` package in Python 3.9 was used [66]. See Figure 3 for an overview of the pipeline.

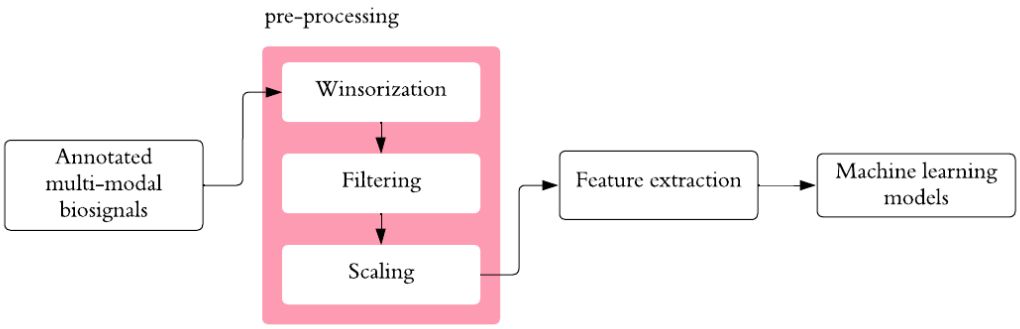


Figure 3: High-level overview of processing pipeline The processing pipeline for the transformation of raw physiological signals into input for the models. It includes four main steps: (1) winsorization for outlier attenuation, (2) signal-specific Butterworth filtering for noise reduction and preparing signals for feature extraction, (3) within-subject Z-score normalization for baseline differences between individuals, and (4) feature extraction.

3.4.1 Winsorization

The first preprocessing step was noise removal. In the signal data of both datasets, a common source of noise was sudden, heavy distortions of the signal, almost certainly attributable to sensor malfunctions, as they exceed biologically plausible ranges and/or patterns. Most likely, such artifacts were caused by sudden movements or interrupted skin contact during the data collection process. This resulted in data points that significantly deviated from normal values. Outliers, especially those that are non-random, introduce noise to the features. Outliers were defined as the values falling in the outer 1% of each participant's trial distribution. Winsorization was applied to reduce the impact of those outliers on the signal (i.e., a within-subject, within-trial trimming). This approach was adapted from [2]. The threshold of 0.01 was relatively mild and chosen to minimize the impact on the raw signal. In the Photoplethysmography (PPG) data of PhyMER, a slightly stricter winsorization setting was applied (0.03) to benefit feature extraction, as these particular recordings were more affected by distortions.

3.4.2 Filtering

Besides outliers, there are more complex forms of noise. A signal can be consistently interfered with by outside factors, such as muscle activity due to respiration or powerline interference [8]. Various filtering techniques exist that can smooth signals and remove such noise. There are three main types of filtering: low-pass, high-pass, and band-pass. Low-pass filtering eliminates high-frequency noise by only allowing low frequencies to pass. In contrast, high-pass filters remove low-frequency noise, often due to baseline wandering. Lastly, band-pass filtering removes both low- and high-frequency components, leaving only the relevant signal intact. In affect recognition research, the Butterworth filter is commonly used due to its uniform sensitivity to the targeted frequencies while reducing unwanted frequencies [67]. Therefore, this filter is applied with varying configurations depending on the signal.

To revisit, ECG and PPG are both measures of cardiac activity that are indicative of affective processes. ECG records the electrical impulses generated by the heart as it undergoes depolarization and repolarization during each cardiac cycle, measured in millivolts (mV)[68]. The peak-to-peak voltage typically ranges from 0.1 mV to 5 mV, depending on the individual and electrode placement [68]. The main frequency components of the ECG signal lie between 0.05 Hz and 150 Hz [41]. The frequencies in ECG

signals that should be preserved for feature extraction are 0.67–5 Hz for detecting the HR and P wave. The QRS complex can be detected within 10 to 50 Hz, and the T wave at 1–7 Hz [41]. Therefore, a high-pass filter at 0.5 Hz plus powerline filtering was applied for ECG signals. This configuration is the default in the `NeuroKit2` library, and is similarly implemented in comparable papers [15], [69].

In contrast, PPG is an optical technique used to measure blood volume changes in the microvascular bed of tissue [43]. PPG is typically measured on the fingertip or wrist, and outputs a waveform [43]. PPG signals are inherently more noisy than ECG signals, and are often contaminated by high-frequency noise [8]. The components of interest typically range from 0.5 Hz to 5 Hz [70]. Due to the low sampling rate of the PPG data (4 Hz), the signal was smoothed using a second-order Butterworth low-pass filter with a configuration of 0.45 Hz.

EDA measures sweat gland activity and reflects sympathetic nervous system activity [44]. The EDA signal is typically encoded in microsiemens (µS), as it represents electrical conductance. The EDA amplitude ranges from 0.1 µS to 50 µS , depending on individual differences, electrode placement, and environmental conditions [8]. Furthermore, the EDA signal is partitioned into tonic (Skin Conductance Level (SCL)) and phasic (Skin Conductance Response (SCR)) activity [8]. Tonic activity reflects the baseline arousal state over time, while phasic activity represents fluctuations related to stimuli. SCR is only a small fraction of SCL [8]. SCRs, or peaks, are significant transient fluctuations in phasic electrodermal activity [8]. SCRs are characterized by its onsets and offsets of fluctuations [8]. Generally, onset thresholds are defined at around 0.01–0.05 µ/S, although methods for dynamic threshold detection exist [66]. SCR can be the result of a stimulus, but can also occur spontaneously. This is referred to as Non-Specific Skin Conductance Response (NS-SCR) [8]. All EDA signals were filtered with a 4th-order Butterworth low-pass filter at 3 Hz using zero-phase filtering, which is the default configuration in the `NeuroKit2` library.

3.4.3 Scaling

As previously argued, individual differences in physiology can hinder the generalization of predictions. Differences in baselines between people can be partially mitigated through within-subject scaling of the raw signal [27]. Scaling was expected to improve the model's ability to generalize by partially accounting for individual variability. Therefore, each individual's sensor data was transformed to be relative to their baseline [71]. This is done before feature extraction by subtracting one's mean μ from each value, and dividing this by the Standard Deviation (SD) σ. Rescaling does not change any underlying patterns, but only shifts the data [71]. Z-score normalization is defined in the following equation:

$$Z_i = \frac{X_i - \mu_i}{\sigma_i} \tag{2}$$

Where X_i is the original value i. μ_i is the mean value of individual i. σ_i is the SD of individual i. This outputs a z-score, Z_i, which is the standardized value of the individual i.

3.4.4 Feature extraction

In this section, the process of feature extraction is discussed. Theoretical concepts will be briefly revisited to provide context, for example, calculation methods and ranges within which the features are expected to fall. In classical machine learning, feature extraction is a crucial step with the primary goal of reducing dimensionality and making the processed data interpretable for the models. It can both help distill meaningful information and reveal new, meaningful information. For AMIGOS, features for ECG and EDA were extracted. For PhyMER, PPG and EDA features were extracted. In turn, parameters were derived from those features.

When constructing features from signal data, a distinction can be made between temporal and spectral features. Temporal features describe variations in the signal over time, usually based on direct measurements of signal amplitude, duration, or inter-event intervals [8]. Consequently, temporal features are often more interpretable and have lower computational complexity. On the other hand, spectral features describe the signal in the frequency domain, reflecting information carried by different frequency components [8]. In the current thesis, only time domain features were extracted. Due to the complexity of frequency analysis, it was necessary to limit the scope of the research to fit a realistic time frame.

In similar research, features are sometimes computed from shorter segments of the recordings to increase the amount of available data or to capture local fluctuations in the signal [2], [37], [72]. However, the current research examines long-term in addition to short-term affective processes. We decided not to segment, based on the premise that preserving the context (i.e., considering the entire recording) offers a more ecologically valid representation of affective states.

ECG and PPG HR can be detected by calculating the average number of R-peaks per minute. In the `NeuroKit2` package, R-peaks are detected by finding local maxima in the QRS complexes based on the steepness of the absolute gradient of the ECG signal [66]. From HR, simple parameters such as mean HR and mean Inter-Beat-Interval (IBI) can be calculated over time fragments [73]. HR at resting state typically ranges from 60-100 bpm, and IBI ranges from 785–1.160 ms [63].

Based on HR, HRV can be calculated, which is the variability between successive R-R peaks within a time window [8]. Low variability indicates activation of the fight/flight response, while high variability indicates a rest/digest response [8]. HRV can be calculated in multiple ways, and conceptualizing HRV in multiple ways may provide a more comprehensive view on HRV. Calculation methods for HRV include the Standard Deviation of NN intervals (SDNN) [73]. SDNN reflects all the cyclic components responsible for variability in the period of recording [8]. NN intervals are the time differences between successive normal heartbeats, e.g., RR intervals after removing ectopic beats [8]. Specifically, the SDNN computes the deviation of each NN interval from the mean over a given period. This is squared to correct for negative values, as only the size of the difference matters, not the direction [73]. In the current study, SDNN is calculated over each recording and represents the total variability in heartbeats during this time. SDNN at resting state typically ranges between 32 to 93 ms, which can be higher in athletes [74].

Next, Root Mean Square of Successive Differences (RMSSD) is another way of measuring HRV. RMSSD takes the root of the average squared differences between two peaks (NN interval). Similar to SDNN, the square is taken to correct for negative values. Next, the average of the differences is taken to enable analysis across NN intervals. Lastly, the root is taken to convert back to milliseconds, making it directly comparable to NN intervals [73]. RMSSD values typically range between 19–75 ms, which can be higher in athletes, but rarely exceed the range of 10 to 300 ms [74].

Next, pNN50 represents the percentage of successive NN intervals that differ by more than 50 ms [73]. Typical values range from 1% to 80%, depending on factors such as age, activity level, and recording conditions. A high pNN50 value reflects greater heart rate variability (HRV), often associated with a relaxed or resting state. In contrast, a low pNN50 indicates reduced HRV, suggesting increased physiological arousal or stress [73]. The interpretation of HRV ranges (e.g., SDNN, RMSSD, and pNN50) depends on whether the recording period was short-term (5 min), ultra-short term (60-240 s), or over 24 hours [63].

Apart from HR and HRV parameters, the temporal characteristics of specific wave components can be derived. Recall that the different wave components represent distinct phases in the cardiac cycle. First, the duration of the QRS complex, or the time from Q wave onset to S wave end, can be calculated. This represents the efficiency of electrical conduction in the ventricles, and shortened QRS duration tends to be related to sympathetic activation [8]. The average QRS complex lasts 0.08 to 0.12 seconds. Similarly, the PR interval can be calculated, which is the time from P wave onset to QRS onset [8]. Similarly, a shorter

PR interval can indicate sympathetic activation. The average PR interval lasts 0.12 to 0.20 seconds. Lastly, the QT interval duration is the time from QRS onset to T wave end [8]. When this interval is longer, the ventricles take longer to repolarize, which is generally linked to sympathetic activation [8]. The average PR interval lasts 0.35 to 0.43 seconds. To extract these features, the signal is delineated into the components representative each cardiac cycle using the `Neurokit2` library [66]. As the heart goes through this cycle with each heartbeat, the average duration of QRS, PR, and QT phases per minute is taken.

HRV features extracted from ECG can also be derived from PPG. However, as PPG is an indirect measure of cardiac activity, the signal of pulsative flow is significantly less stable and defined compared to the direct electrical currents generated by the heart [8]. Therefore, distinct wave components are hardly distinguishable from PPG data, which is why these features were excluded from the analysis of PPG data. Similarly, PPG is more sensitive to motion artifacts [8]. As HRV metrics are more sensitive to outliers and noise, and extended measurement periods typically yield more reliable results [63]. Therefore, HRV metrics extracted from PPG may not be as informative as when extracted from ECG.

EDA Two features were derived from EDA; SCL and SCR. In total, eight parameters were extracted from those two features. Again, only time domain features were taken into account.

In the context of longer, ongoing stimuli, the most useful EDA measures are SCL and frequency of peaks, or NS-SCRs [8]. As chronic stimuli span over longer periods, they can be viewed as modulating increases and decreases in tonic arousal [8]. By contrast, analysis of phasic activity provides information on the reactivity of the ANS [8]. Parameters from both SCL and SCR were constructed.

Quantifying certain SCR components, such as amplitude, is not a straightforward task. In particular, the response interference effect complicates analysis [8]. Response interference occurs when the peak to be analyzed is interfered with a new peak due to the relatively slow reactive nature of EDA, which results in an accumulation of responses [8]. Therefore, it should be noted that there is no perfect solution to this problem, and the calculation of certain SCR parameters depends on experiment design and which analysis tools are used for calculation [8]. For example, the `NeuroKit2` package in Python uses adaptive thresholding to detect rapid increases (onsets) in conductance, or peaks [66]. It filters out small fluctuations using amplitude and rise time criteria to detect only significant SCRs [66]. With the considerations in mind, parameters like mean SCR amplitude, SD of the mean amplitude, NS-SCR frequency, mean tRise, mean tRecovery, and mean tHalfRecovery are typically used in analysis of the phasic activity [73], [75].

Mean SCR amplitude is the average value of the phasic component over a given time window [73]. Amplitude was calculated by subtracting the highest point from the SCR from the local minimum just before the peak. When taking the mean amplitude over a certain period, a lower value indicates less arousal and vice versa [45]. Similarly, the SD of the mean SCR amplitude can be taken to assess the variability over a certain period [73]. A higher SD indicates more fluctuations in EDA, which suggests frequent sympathetic responses [8]. Typical amplitudes of peaks range from 0.2 to 1.0 S [73]

Next, the number of SCRs in the absence of identifiable stimuli, or NS-SCRs was taken[73]. NS-SCR frequency, can be counted over a specific time window, and indicates how evoking a given moment was for a participant [8]. On average, 1 to 3 SCRs occur per minute in a resting state, but this can increase to around 25 to 30 in states of high arousal [73]. To facilitate comparison between stimuli of varying lengths, the NS-SCR frequency was corrected for recording length.

Peak rise time (tRise) is the time interval between the onset of the SCR and its peak amplitude, typically taking between 1 to 3 seconds [73]. This indicates how quickly the skin conductance response builds up to form a peak [8]. Longer rise times may indicate more gradual tension increase, while shorter

rise times indicate a sudden increase [8]. When extracting this parameter from a longer time window, multiple peaks may occur, which leads us to calculate the mean tRise over each timeframe. On average, the rise time of an SCR is between 1 to 3 seconds [8]

Similarly, the average time it takes a peak to return to baseline was calculated. Known as Peak Recovery Time (tRecovery), which measures the nervous system's ability to recover [8]. Longer recovery times may indicate prolonged autonomic activation. However, considering the response interference effect, SCR half recovery time is a more robust parameter to capture recovery time. Specifically, this is the temporal interval between SCR peak and point of 50% recovery of SCR amplitude, typically lasting in 2 to 10 seconds [8]. To account for trial length, the average of the tHalfRecovery time is taken.

Finally, the slope and the mean of the tonic component were taken. The SCL Slope is the average change in SCL within a time fragment. It can be calculated by by fitting a linear model to the SCL signal. The `scipy.stats` library in Python includes a function that computes the slope of any line using ordinary least squares [76]. Mean tonic levels fall within 2 to 20 S, while the gradual change of SCL (slope) is typically 1 to 3 S [8]. Table 5 provides an overview of each biosignal, their extracted features, parameters, and calculation methods.

Signal	**Feature**	**Parameter**	**Calculation Method**
ECG(8)	Heart Rate (HR)	Mean HR	–
		Heart Rate Variability (HRV)	SDNN
			RMSSD
			PNN50
	Inter-Beat Interval (IBI)	Mean IBI	–
	QRS Complex	Duration	–
	PR Interval	Duration	–
	QT Interval	Duration	–
PPG(5)	Heart Rate (HR)	Mean HR	–
		Heart Rate Variability (HRV)	SDNN
			RMSSD
			PNN50
	Inter-Beat Interval (IBI)	Mean IBI	–
EDA(7)	Skin Conductance Response (SCR)	Mean SCR Amplitude	–
		SD of SCR Amplitude	–
		NS-SCR Frequency	–
		Mean Rise Time	–
		Mean Recovery Time 50%	–
	Skin Conductance Level (SCL)	Mean SCL Amplitude	–
		SCL Slope (Δ)	–

Legend: HRV: Heart Rate Variability, SDNN: Standard Deviation of NN intervals, RMSSD: Root Mean Square of Successive Differences, PNN50: Percentage of NN50, QRS complex: Ventricular depolarization complex in ECG, PPG: Photoplethysmogram, EDA: Electrodermal Activity, tHalfRecovery: Time to half recovery after peak response.

Table 5: Overview of Extracted Features for Each Modality Overview of signal types, their corresponding time-domain features, parameters and different ways of calculating that parameter. Note that from PPG, distinct wave components cannot be reliably extracted. Features were corrected for recording length to facilitate comparison between recordings of varying lengths.

In addition to the physiological features, the AMIGOS dataset included a contextual binary variable indicating whether the participant viewed the long video alone or in a group. As the original AMIGOS authors reported effects of social context on affective responses, this variable was included as a binary variable in the long-video experiments [15].

Outlier removal Earlier removal of extreme values in the raw signal data did not remove all anomalies. Therefore, after feature extraction, the distributions were inspected for outliers that fell outside physiologically plausible ranges. These ranges were based on the literature and are defined earlier in Section 3.4.4. Any feature that was exhibiting suspicious behavior was flagged for investigation. In the AMIGOS short video experiment, all features fell in expected ranges, and no extreme outliers were detected. This was not the case in the long video experiment of AMIGOS. In particular, the HRV measures (SDNN and RMSSD) contained extreme values, which by far exceeded the expected ranges. Longer trials generally contained more artifacts, and SDNN and RMSSD are inherently sensitive to measurement errors. For example, the presence of just two artifacts within a 5-minute segment can significantly distort its value [63]. To address this, we excluded any trial in which RMSSD exceeded 300 ms, a threshold beyond the range of physiologically plausible values (10-300 ms) [74]. This threshold also captured the same extreme values in SDNN. Five trials were discarded.

When inspecting the PhyMER features, most fell in expected ranges. However, SDNN and RMSSD exhibited unrealistic behavior in the outer ends of the distribution, similarly seen in the AMIGOS long video data. Since PPG is an indirect measure of cardiac activity and inherently noisy, it was unsurprising that it yielded unreliable HRV metrics. Four trials in which RMSSD exceeded >300 ms were discarded.

Between-person variability As large inter-individual variability can hinder cross-participant generalization, the proportion of total variance attributable to between-participant differences was assessed using the one-way random effects Intraclass Correlation Coefficient (ICC). The ICC is typically used for determining the reliability of a metric, indicating how much of the total variance in a measure can be attributed to systematic differences between participants, rather than measurement error [77]. In the application in this thesis, this 'measurement error' represents within-subject variability. Hence, ICC is applied to quantify the between-person variance of the different features, as the data consists of multiple measurements. Variance components were estimated from weighted mean squares to account for the unequal number of trials per participant, as some were discarded earlier due to data corruption. The formula is defined as follows:

$$ICC = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}} \tag{3}$$

Here, $\sigma^2_{\text{between}}$ denotes the variance between subjects, and σ^2_{within} represents the variance within subjects. If the ICC for a feature is less than 0.5, most variability occurs within participants (across trials), and between-person differences are small. In contrast, when the ICC approaches 1, variability is dominated by differences between individuals [77]. Although a high proportion of between-person variance indicates that a feature is a reliable individual characteristic, this stability across individuals can hinder generalization to unseen participants. Formula adopted from [77].

3.5 Personality data processing

In both datasets, the personality data consisted of precomputed trait scores (i.e., one score per person for each trait). The five traits were Extroversion, Neuroticism, Openness, Conscientiousness, and Agreeableness.

First, Neuroticism scores in AMIGOS were inverted, reflecting 'Emotional Stability' instead. This was reversed to ensure consistency across datasets. Next, trait columns were scaled to fall in equal ranges. Scaling was done using the standard scaler from the `scikit-learn` library, which is equal to a z-score normalization for between-subjects. Each individual's score for a certain trait was subtracted from the corresponding trait mean, divided by the trait's SD. This transforms each facet score to be relative to the mean across all participants on each respective trait. This step was crucial as variables with larger ranges can dominate the clustering process [78].

3.5.1 Clustering

For the proposed personalization approach in this thesis, clustering served to identify subgroups of individuals based on their personality profiles. Two clustering algorithms were considered, K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). K-means clustering is one of the most commonly used clustering methods, particularly employed by studies focused on cluster-based personalization (e.g., [2], [19], [21]). However, the personality traits in PhyMER were slightly non-normally distributed, and centroid-based algorithms, such as k-means, are sensitive to asymmetric or heavy-tailed distributions [79]. In this case, density-based methods such as DBSCAN may perform better as they do not assume specific cluster shapes, sizes, or distributions [80]. Therefore, these two approaches were compared to determine which suited the data best. Below, both algorithms and two evaluation metrics are briefly discussed. Afterwards, the selection process is described.

First, DBSCAN is an unsupervised, non-parametric clustering algorithm. It is particularly suitable for arbitrarily shaped clusters and can handle noise and outliers relatively well [78]. Clusters are formed based on the density of the datapoints [78]. Two parameters are selected before clustering: the neighborhood radius ϵ and the minimum number of points required to form a dense region. Next, the algorithm randomly selects an unvisited data point and determines whether it is a core point. A data point is a core point when it has at least the number of $minPts$ neighbors within ϵ. Clusters are formed around the core points, and all reachable points are assigned to this cluster. If a data point is not a core point, it is labeled as noise. These steps are repeated until all points are visited [80]. DBSCAN automatically determines the optimal number of clusters.

In contrast, K-means groups data into k distinct clusters based on Euclidian distance [79]. Before clustering, the number of k is determined. The algorithm then randomly selects k initial centroids from the data and assigns each point to the cluster whose centroid is closest, typically based on Euclidean distance. The centroids are recomputed as the mean of all points in their respective clusters. The last two steps are repeated until the clusters are stable and the centroids no longer change significantly [79]. The optimal number of k can be found by comparing the resulting silhouette scores for different values of k.

The silhouette score indicates how well any data point fits within its assigned cluster, and how distinct that cluster is from others [81]. A coefficient closer to 1 reflects a good match, and a coefficient closer to -1 reflects a poor match. If many data points have a low silhouette score, the number of clusters may be too low [81]. Silhouette score is therefore a common evaluation metric for cluster quality. Alternatively, Davies-Boulding Index (DBI) measures the average similarity of each cluster with its most similar cluster, where similarity is defined by the intra-cluster (compactness) and inter-cluster (separation) distances [82]. Here, lower values are preferred, and indicate the clusters are well-separated and/or compact. Hence, the optimal clustering method achieves the highest silhouette score and lowest DBI index.

Although DBSCAN initially appeared promising due to its ability to detect clusters of arbitrary shape and handle outliers, its use in this context proved problematic. As DBSCAN labels a subset of samples as noise, a large portion of participants were excluded from any cluster, regardless of parameter

configuration. For instance, even with a permissive maximum distance for points to be assigned to a cluster (eps = 4.5) and a low number of minimal required samples to form a dense region (min samples = 2), DBSCAN still labeled approximately 15% of data points as noise (silhouette score = 0.34). This behavior was unsuitable for the use of clustering in this study, which relied on assigning every participant to a cluster, unless the subjects labeled as noise are excluded. Therefore, K-Means clustering was adopted instead, which ensured complete cluster assignment across all participants, albeit at the slight expense of cluster cohesion.

3.6 Models

To build the models, two different machine learning algorithms were employed: RF and SVM. These models were chosen because they are among the most widely used in the context of affect recognition research (e.g., [26], [27], [37], [41]). Additionally, both algorithms can be applied to classification as well as regression tasks.

Firstly, RF have been reported by prior research to be particularly effective in exact valence and arousal estimation [37]. RF is a tree-based ensemble method that aggregates multiple individual decision trees, of which the predictions of a class or a continuous variable are averaged [83]. Each tree is trained on a bootstrapped sample (sampling with replacement) of the training samples. At each split, a random subset of the features is considered. Due to the averaging of multiple trees, RF tends to reduce the risk of overfitting compared to a regular decision tree. In turn, RF is relatively robust to error and outliers [83]. The most critical hyperparameters are the number of trees (n_estimators), the maximum depth of each tree (max_depth), and the number of features considered for splitting at each node (max_features). Increasing the number of estimators generally improves performance, but it also increases computation time. Max depth controls the complexity of each tree, with deeper trees being more prone to overfitting. Lastly, max features balances diversity among trees and model accuracy by limiting the feature subset used at each split [83].

Secondly, SVM are one of the most frequently adopted algorithms in affect recognition research, according to a meta-analysis by Hasnul et al. [41]. For predicting continuous variables, Support Vector Regression (SVR) is used. SVR uses supervised learning to identify an optimal decision boundary that minimizes error within a defined margin [84]. In Support Vector Classification (SVC), the boundary maximizes the margin between classes instead. SVM in general are particularly suited in high-dimensional feature spaces and are capable of capturing nonlinear relationships through their different kernel functions (e.g., radial basis function (RBF) or polynomial kernels) [84]. Besides kernel options, the most critical hyperparameters are gamma and C. First, the regularization parameter C balances the margin size against prediction errors. A high C value allows for fewer errors and fits the boundary closer to the training data. In turn, gamma dictates the influence of each data point. With a low gamma, the influence is small, and the boundary fits the training data more loosely.

Moving on, the training and testing procedure is now discussed. A random 80/20 train/test split was made on the data to prevent the model from seeing any of the testing data beforehand. Here, all trials of the test participants were excluded, adhering to a subject-independent approach. On the training partition of the data, the input features were scaled to fall in comparable ranges using a standard scaler, which standardizes features by removing the mean and scaling to unit variance on each feature independently using the samples in the training set.

To optimize model performance, a grid search was employed. Grid search exhaustively evaluates all specified parameter values by training and validating models for each combination of hyperparameters. The grid search was executed via Leave-One-Out Cross-Validation (LOOCV), which is suitable for small datasets. LOOCV splits the data into a training and test set, with only one participant occupying the

test set [78]. One such configuration is considered a fold. Then, within a set range of different hyperparameters, the mean best-performing hyperparameters are selected for the respective fold. This process is repeated until each participant has appeared as the test set once. This yields the hyperparameter set that results in the best performance across folds according to a chosen metric. In the regression tasks, Root Mean Squared Error (RMSE) was used for optimization, while classification optimized on Area Under the Curve (AUC). The best hyperparameter combination was then used to refit the model on the entire training set. Ultimately, the model was evaluated on the held-out 20% of the data. All models were subject to an identical training and evaluation pipeline.

To evaluate the effectiveness of our proposed solution, we built three variants of the affect recognition models. First, a baseline model was constructed. The baseline was a fully generalized model that received input only from the physiological features, excluding any personality data.

For the proposed solution, personality-based cluster assignments were added to the model. This approach differed from the typical cluster-based modeling in the literature. Cluster-based modeling trains a separate model on each identified cluster [2]. However, this was not feasible with the current size of our datasets. Training models separately for each cluster would increase the risk of overfitting, as the resulting number of trials per trained model would be too small after partitioning. For instance, three clusters of equal size would result in around 200 trials per trained model. Therefore, we approximated cluster-based modeling, by featuring the cluster assignment of each participant in the model instead. This would make our approach 'cluster-informed' rather than truly stratified.

Including the absolute cluster number (e.g., 0, 1, 2) would inadvertly bias the model towards higher values, which is problematic as the labels carry no numerical meaning. Therefore, the Principal Component (PC) projection from each cluster centroid was taken instead. In K-means, the cluster centroid represents the average of the data points within that cluster [78]. Therefore, taking the PC would represent a compressed representation of the average profile of each cluster. In other words, this value identifies cluster membership while retaining *some* information about the relative position of that cluster in the personality space. As the sole purpose of the PC was to identify cluster membership, only one PC was necessary.

To assess the added value of the clustering, a model was implemented that included the five Big Five trait scores directly as features. This included normalized scores for each participant on Extroversion, Openness, Conscientiousness, Neuroticism, and Agreeableness. Incorporating a Traits model serves two purposes. First, it helped determine whether any effects observed in the Clusters model are attributable to clustering itself rather than to the traits. Second, it tested whether personality traits alone provide useful information for affect recognition. If the cluster-informed model outperforms both the baseline and the Traits model, this suggests that clustering captures meaningful structure beyond what the trait scores convey directly.

3.7 Evaluation metrics

Evaluation metrics are essential to understanding model performance. A combination of metrics is needed for a complete understanding. Regression models require different metrics than classification tasks, so both will be briefly reviewed in this section. For the continuous prediction models, we used Mean Absolute Error (MAE), RMSE, and Coefficient of Determination (R^2). For the classification models, we reported accuracy, balanced accuracy, F1 score, and AUC. These metrics were chosen to enable comparison with related works.

3.7.1 MAE, RMSE and R2

Firstly, MAE calculates the average absolute difference between the predicted values ($\hat{y}$) and the actual values (y). MAE is therefore a direct measure of prediction accuracy, and lower MAE generally indicates improved predictive accuracy [85]. MAE is expressed in units of the target variable.

Next, RMSE is the square root of the average squared difference between predicted and actual values [85]. RMSE penalizes larger deviations more than smaller ones, which can lead to biased evaluations in the presence of extreme values. RMSE is also expressed in units of the target variable, making it directly interpretable [85]. Similar to MAE, lower RMSE indicates more accurate predictions [85].

Lastly, R^2 quantifies the proportion of variance in the target variable that is explained by the model [86]. It is essentially one minus the proportion of residual sums of squares (SSE) over the total sums of squares (SST). Hence, an R^2 value close to 1 indicates that the model captures most of the variability in the data. In contrast, a value near or below 0 suggests that the model does not improve upon a naive baseline, which always predicts the target mean [86]. While R^2 typically ranges from 0 to 1, it can take negative values. Negative R^2 values occur when the SSE is larger than the SST; in this case, the model would perform better by simply predicting the mean.

To summarize, the best-performing regression model has the lowest MAE, Mean Squared Error (MSE), and RMSE, indicating less prediction error. Meanwhile, the best-fitted model has a higher R^2 score, which explains more of the variance in the target variable.

3.7.2 Accuracy, F1 and AUC

While metrics for regression-based tasks quantify the magnitude and distribution of errors, classification-based tasks evaluate the correctness of the classifications. While there are various metrics for classification, we will report Accuracy, Balanced Accuracy, F1, and AUC to enable comparison with related works.

First, accuracy measures the proportion of correctly classified cases over the total number of predictions. Although intuitive and easy to interpret, accuracy can be misleading in imbalanced datasets, and does not differentiate between false positives and false negatives [78]. Therefore, more representative metrics are needed for fair comparison between studies.

An accurate metric under imbalance is the F1 Score. The F1 score is the harmonic mean of the true positives out of all positive classifications (precision) and out of all true positives, how many have been classified as a true positive. While precision reflects the model's ability to avoid false positives, recall demonstrates the model's ability to avoid false negatives [78]. However, the F1 score focuses only on performance with respect to the positive class; it does not consider how well negative cases are classified or the proportion of negative cases in the dataset. Since we are interested in accurately identifying both high (positive) and low (negative) valence and arousal, we additionally report Balanced Accuracy to provide a more comprehensive evaluation. Balanced accuracy weighs both classes the same, regardless of their frequency within the dataset.

Lastly, the AUC summarizes the model's ability to discriminate between the two classes across all possible classification thresholds, in contrast to previous metrics that consider a single threshold (e.g., 0.5) [87]. The AUC corresponds to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. An AUC of 0.5 indicates chance-level performance, while a value of 1.0 reflects perfect discrimination [87].

4 Results

This section details the results of the conducted methodology. First, the physiological features were explored. This includes reviewing the descriptive statistics, collinearity among features, their correlations with affective labels, lastly their variability between participants and trials. Then, the personality data clustering process is outlined, which is followed by a brief exploration of the personality features. Next, the optimized hyperparameters are summarized. Finally, the empirical results of the affect recognition models for both datasets are presented.

4.1 Data exploration

4.1.1 Descriptives

The first step in analyzing the physiological data involved computing and examining descriptive statistics for the extracted features. This step helped verify their behavior and improve understanding of the data. Recall that a small number of features were crafted for input to the models. In AMIGOS, nine features from the Electrocardiography (ECG) signal and eight from Electrodermal Activity (EDA) were extracted, totaling 16 different features. Additionally, in AMIGOS, a binary feature was added to indicate whether the participant had watched the stimuli alone or in groups. In PhyMER, six different features were extracted from Photoplethysmography (PPG) and eight features from EDA.

See table 6 for the descriptive statistics of the features derived from the short video experiment in AMIGOS.

Feature	**Mean**	**SD**	**Min**	**Max**
HR_mean (bpm)	76.74	10.19	52.92	112.80
SDNN (ms)	53.98	19.09	12.79	123.60
RMSSD (ms)	34.14	14.93	10.34	132.22
pNN50 (%)	11.59	10.70	0.00	52.86
Mean_IBI (s)	0.80	0.11	0.53	1.13
QRS_mean (s)	0.07	0.02	0.04	0.16
PR_mean (s)	0.17	0.02	0.14	0.25
QT_mean (s)	0.33	0.03	0.21	0.40
MeanSCRAmplitude (μS)	0.47	0.35	0.00	2.69
SDSCRAmplitude (μS)	0.33	0.25	0.00	1.90
NSSCRFrequency	5.89	3.26	0.00	27.81
MeanRiseTime (s)	2.70	1.62	0.68	13.68
MeanRecoveryTime_50 (s)	1.82	1.02	0.14	9.14
MeanSCL (μS)	0.05	0.87	-2.35	3.28
SCLSlope (μS)	0.00	0.01	-0.04	0.07

Table 6: Descriptive statistics of physiological features extracted from short-video segments in AMIGOS (total trials: 623) Summary of the distribution (mean, SD, min–max) of each extracted feature across all trials. Before feature extraction, winsorization, filtering, and normalization of the raw signal have been performed. Most features behaved normally, and most values fell in physiologically plausible ranges.

Feature	Mean	SD	Min	Max
HR_Mean (bpm)	72.67	7.71	52.73	89.21
SDNN (ms)	78.01	29.11	28.27	273.74
RMSSD (ms)	51.31	24.96	17.50	203.93
pNN50 (%)	16.20	11.51	1.18	52.13
MeanIBI (s)	0.84	0.09	0.67	1.14
QRS_Mean (s)	0.07	0.02	0.05	0.12
PR_Mean (s)	0.18	0.02	0.14	0.22
QT_ Mean(s)	0.34	0.03	0.28	0.42
MeanSCRAmplitude (μS)	0.44	0.30	0.01	1.89
SDSCRAmplitude (μS)	0.32	0.20	0.00	1.26
NSSCRFrequency	3.18	2.12	0.04	8.39
MeanRiseTime (s)	2.39	1.33	1.51	12.00
MeanRecoveryTime_50 (s)	2.13	0.61	0.89	4.14
MeanSCL (μS)	0.00	0.70	−1.41	1.51
SCLSlope (μS)	0.00	0.00	−0.00	0.00

Table 7: Descriptive statistics of physiological features extracted from long-video segments in AMIGOS (total trials: 141). Summary of the distribution (mean, SD, min–max) of each extracted feature across all trials. Before feature extraction, winsorization, filtering, and normalization of the raw signal have been performed. Most features behaved normally, and most values fell in physiologically plausible ranges. Large mean RMSSD and SDNN may indicate more noise in the data.

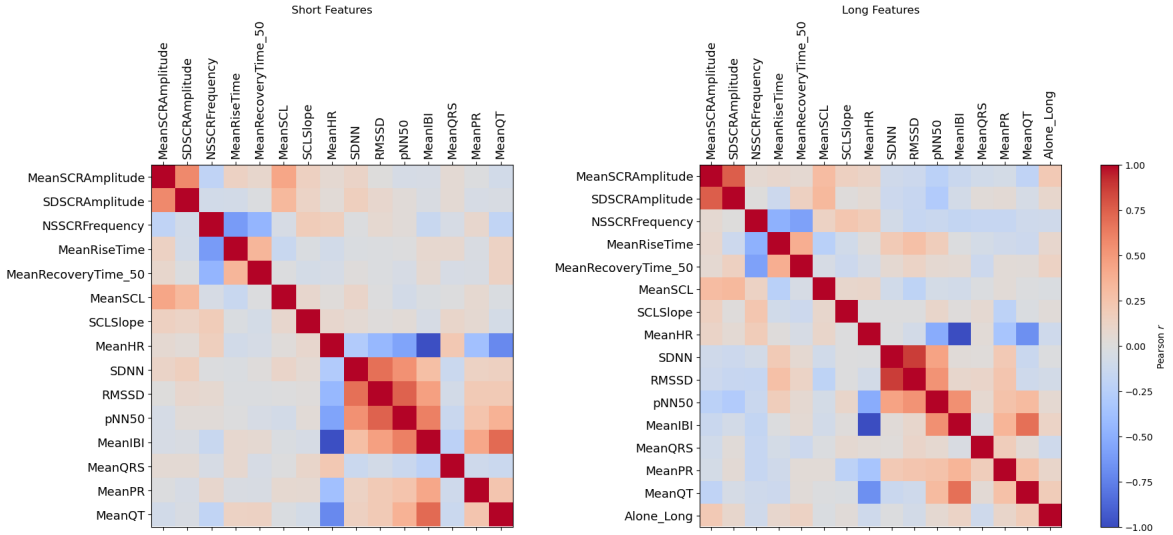
Lastly, most features extracted from the PhyMER dataset fell in expected ranges, see table 8

Feature	Mean	SD	Min	Max
Mean_HR (bpm)	73.42	9.88	51.54	99.51
SDNN (ms)	87.89	40.76	23.66	260.33
RMSSD (ms)	111.79	69.10	21.27	296.34
pNN50 (%)	41.59	23.74	0.55	92.50
Mean_IBI (s)	0.83	0.11	0.60	1.16
MeanSCRAmplitude (μS)	0.11	0.15	0.00	1.27
SDSCRAmplitude (μS)	0.08	0.14	0.00	1.69
NSSCRFrequency (count)	7.80	5.62	0.00	26.44
MeanRiseTime (s)	2.76	2.26	0.99	16.00
MeanRecoveryTime$_{50}$ (s)	1.71	0.98	0.50	6.5
MeanSCL (μS)	0.02	1.03	−3.02	6.06
SCLSlope (μS)	0.00	0.01	−0.05	0.06

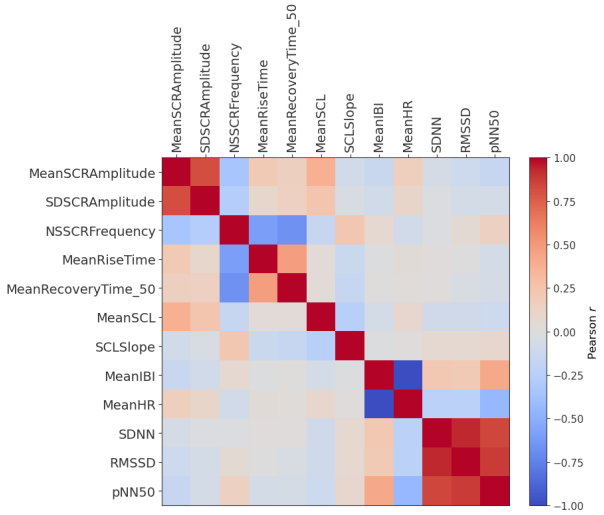
Table 8: Descriptive statistics for physiological features of dataset PhyMER (total trials: 676) Summary of the distribution (mean, SD, min–max) of each extracted feature across all trials. Before feature extraction, winsorization, filtering, and normalization of the raw signal have been performed. Most features behaved normally, and most values fell in physiologically plausible ranges. However, similar to the long video experiment in AMIGOS, the HRV measures exhibited some unrealistic values at the upper end of the distribution.

4.1.2 Feature collinearity

To continue, all features in both AMIGOS and the PhyMER datasets were screened for collinearity by visualizing the pairwise Pearson correlation coefficients in the form of heatmaps. This allowed for further examination and confirmation of feature behavior. See Figure 4 for an overview.



(a) AMIGOS Video Correlations



(b) PhyMER Correlation Heatmap

Figure 4: Pairwise Pearson Correlation Heatmaps for Physiological and Personality Cluster Features in AMIGOS and PhyMER Datasets. These figures present the pairwise Pearson correlations for all extracted physiological features in the AMIGOS and PhyMER datasets. The heatmaps in subfigure (a) display correlations for the short video experiment (left) and the long video experiment (right) in the AMIGOS dataset. Subfigure (b) shows the feature correlations within the PhyMER dataset. Color intensity indicates the strength of the correlation, with blue representing negative associations and red representing positive associations. Features have been corrected for recording length.

In AMIGOS, Heart Rate Variability (HRV) features were strongly intercorrelated with one another, as expected, since they are different measures of the same phenomenon. Standard Deviation of NN intervals (SDNN), Root Mean Square of Successive Differences (RMSSD), and pNN50 exhibited correlations of $|r| =$ 0.44–0.81 in short and $|r| =$ 0.68-1 in long videos. Similarly, Mean Heart Rate (HR) correlated negatively with HRV features (e.g., SDNN, RMSSD, pNN50), as HRV metrics are mathematically and physiologically linked to HR [88]. Similarly, Mean HR showed a strong negative correlation with Mean IBI ($r =$ -0.8),

consistent with the inverse relationship between HR and Inter-Beat-Interval (IBI) [8]. In AMIGOS, the phasic EDA features were moderately correlated, indicating internal coherence; Mean Skin Conductance Response (SCR) Amplitude and Standard Deviation (SD) of the mean SCR Amplitude ($r = 0.7$), as both are derived from the same distribution of detected SCR. Mean Recovery Time 50 correlated negatively with NS-SCR Frequency; when the Autonomic Nervous System (ANS) cycles through activation and recovery more quickly (i.e., more peaks), there is less time between successive responses (i.e., shorter recovery times) [8]. In the long videos, features related to HRV(SDNN, RMSSD, and pNN50) correlated more strongly with one another than in the short videos, but more weakly with other features. This may be explained by the fact that the longer trials contained more motion artifacts, and that shared noise among the features may have increased correlations spuriously.

In PhyMER (subfigure (b), similar patterns emerged. For instance, strong intercorrelations among HRV were again observed, but to a more extreme extent compared to AMIGOS ($|r| = 0.8$-0.95). Similarly, electrodermal features were moderately to strongly correlated with each other ($|r| = 0.20$-0.84). Note that the different wave components in PhyMER are not displayed as they could not be extracted from PPG.

4.1.3 Feature target correlations

To assess the direct associations between features and the labels, we calculated Pearson correlations between each feature and the self-reported arousal and valence ratings. The univariate feature-label associations of both the long and short experiments in AMIGOS are visualized in Figure 5.

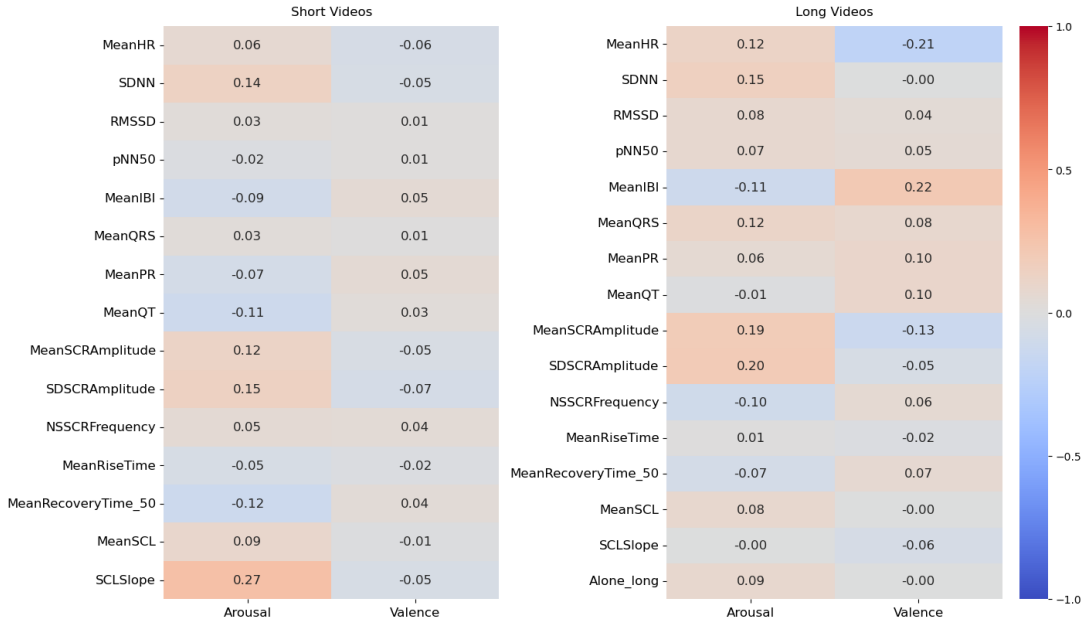


Figure 5: Univariate Feature Correlations with Arousal and Valence for AMIGOS. Pearson correlation coefficients were computed between each physiological feature and self-reported affective ratings in the AMIGOS dataset. The left and right heatmaps correspond to the short and long video conditions, respectively. In general, features were more strongly related to arousal than to valence. The strongest correlation is observed for SCL Slope in the short videos with arousal. In the long segments, correlations between label and cardiac features increased compared to the short segments.

In the short video experiment of AMIGOS, arousal exhibited the strongest correlation with Skin Conductance Level (SCL) Slope ($r = 0.27$), suggesting that steeper increases in SCL were moderately associated with heightened arousal. This was then followed by SD SD SCR Amplitude ($r = 0.15$) and SDNN ($r = 0.14$). No feature showed a substantial correlation with valence for short videos ($|r| < 0.07$).

In the long video experiment of AMIGOS, the feature with the relatively strongest correlation to arousal was SD SCR Amplitude ($r = 0.20$). This was followed by Mean SCR Amplitude ($r = 0.19$) and SDNN ($r = 0.15$). For valence, the features correlating most strongly with the labels in the long videos were Mean IBI ($r = 0.22$) and its reciprocal, Mean HR ($r = 0.21$). Finally, watching the long videos in group setting (Alone_long = 1) was weakly associated with higher arousal ratings, but not with valence ($r = 0.09$ and $r = 0.0$, respectively). Notably, the correlation between SCL Slope and arousal in the long videos shrank considerably, compared to the short videos ($r = 0.06$). Likely, computing slopes over longer windows averaged out local fluctuations. In contrast, the cardiac features, particularly Mean HR, Mean IBI, and the three wave components (Mean QRS, Mean PR, Mean QT), showed increased correlation with both valence and arousal compared to the short videos. Possibly, the measurement of cardiac metrics requires more extended periods to exhibit clear, consistent changes in response to affective stimuli [8]. Alternatively, the presence of more noise in longer trials inflated spurious correlations.

For PhyMER, the linear feature-label associations were similarly visualized and evaluated by computing Pearson correlations for each pair (see Figure 6).

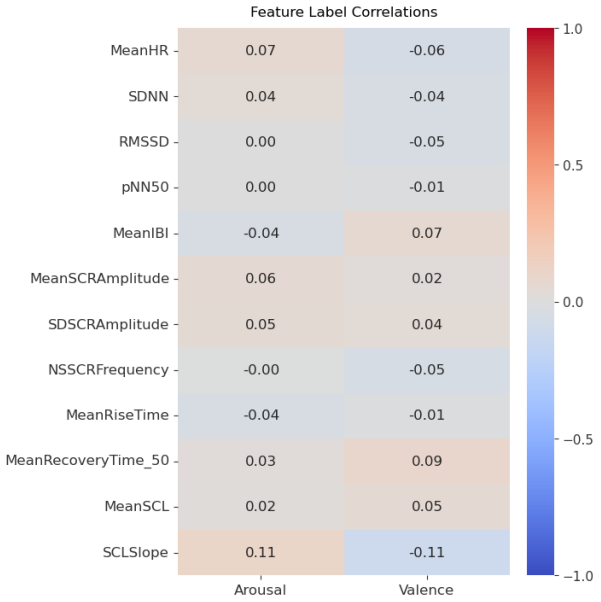


Figure 6: Univariate Feature Correlations with Arousal and Valence in PhyMER. Heatmap of Pearson correlations between physiological features and affective labels in the PhyMER dataset. Although most associations were near zero, and features showed patterns similar to the AMIGOS short-video label-feature correlations. The overall weak linear relationships suggest limited predictive power from individual features.

The overall magnitude of correlations between the physiological features and labels in PhyMER was mostly close to zero, with all correlation coefficients falling between -0.11 and 0.11. Although to a lesser extent, the features extracted from PhyMER shared patterns similar to those in the short dataset in AMIGOS. For arousal, the largest correlations were SCL Slope ($r = 0.11$), Mean HR ($r = 0.07$) and Mean SCR Amplitude ($r = 0.06$). For valence, this was SCL Slope (r = -0.11), Mean Recovery Time 50% ($r = 0.09$) and Mean IBI ($r = 0.07$), with steeper SCL slopes, slower recovery times and high IBI related to more negative affect. For valence, the strongest feature-label correlations were observed for SCL Slope ($r = 0.11$), Mean Recovery Time 50% ($r = 0.09$), and Mean IBI ($r = 0.07$).

The label-feature correlations within PhyMER were relatively low compared to those in AMIGOS data. This may be attributed to lower sensor quality (commercial instead of medical-grade), sampling rate, or environmental factors in PhyMER. However, low correlations were observed across datasets, suggesting that affective responses may be more effectively captured through nonlinear models, feature

interactions, or subject-specific modeling approaches.

4.1.4 Between-person variability

Lastly, the between-participant differences of physiological features were assessed. Specifically, we calculated the proportion of total variance attributable to between-participant differences, using a one-way random-effects Intraclass Correlation Coefficient (ICC). The results provide insight into inter-individual variability, which can potentially hinder cross-participant generalization.

Figure 7 displays the ICCs for both long and short experiments.

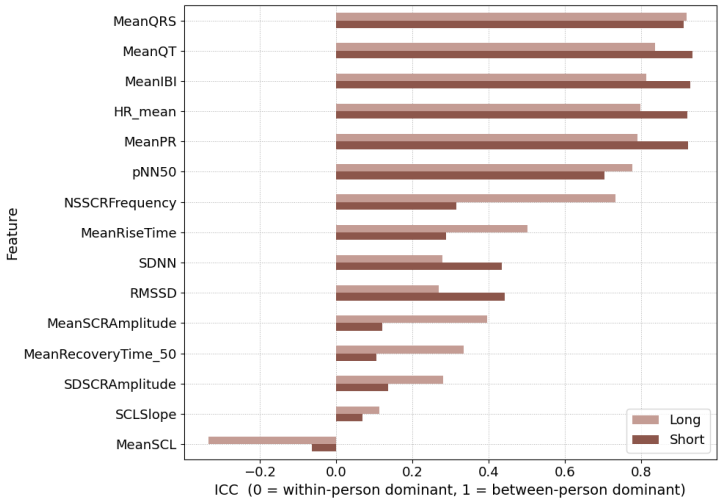


Figure 7: Weighted Intraclass Correlation Coefficients (ICC) for Physiological Feature Variability in AMIGOS. Weighted ICC for physiological features in the AMIGOS short and long video experiments. ICC values are displayed the x-axis with the features on the y-axis. Values close to 0 indicate variability occurs predominantly within participants, while values closer to 1 indicate variability predominantly occurs between participants. Cardiac features exhibit high between-participant variability (ICC = 0.4–0.9), while electrodermal features show greater within-person variability. This holds implications for model generalization. Notably, Mean SCL yielded negative ICC values, likely due to normalization effects.

In AMIGOS, cardiac features exhibited prominent variability among individuals, with many features exceeding the 0.5 threshold, which impacts generalization abilities. The negative ICCs observed for Mean SCL (-0.1 in short videos, -0.4 in long videos) indicate that its variance is predominantly within-participants. This may explain why SCL Slope was one of the most strongly related features to valence and arousal, as observed earlier.

Similarly, Figure 4.1.4 shows the ICC values for the physiological features in PhyMER.

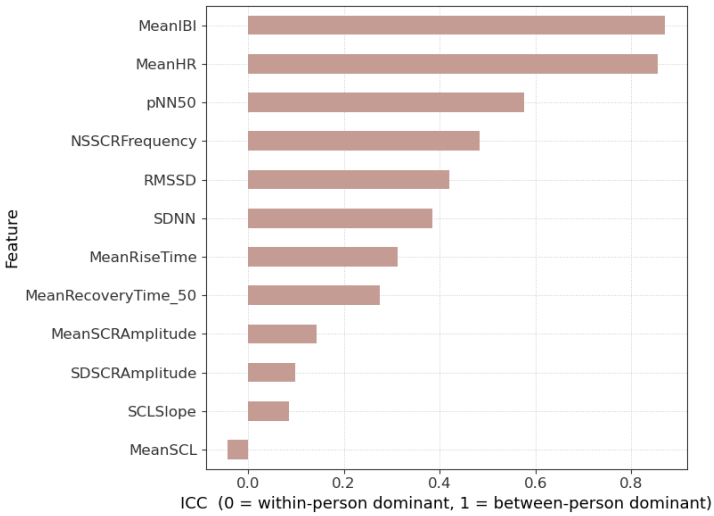


Figure 8: Weighted Intraclass Correlation Coefficients (ICCs) for Physiological Feature Variability in PhyMER. Weighted ICC for physiological features in PhyMER. ICC values are displayed the x-axis with the features on the y-axis. Values close to 0 indicate variability occurs predominantly within participants, while values closer to 1 indicate variability predominantly occurs between participants. Electrodermal features exhibit low or negative ICC, indicating that within-participant variability predominates. In contrast, cardiac metrics (e.g., pNN50, Mean HR, Mean IBI) exhibit high between-participant variance (ICC 0.54–0.87).

The PhyMER feature variance structures display patterns similar to AMIGOS. Specifically, the variance within electrodermal and HRV features was primarily due to within-person differences (ICC = -0.04-0.40), while cardiac metrics, particularly those related to heart rate (pNN50, Mean HR, Mean IBI), were primarily driven by between-person differences (ICC = 0.54 – 0.87).

In both datasets, electrodermal features were consistently stable within participants, aside from NS-SCR frequency, which likely varied a lot across trials. In both datasets, cardiac features related to heart rate (e.g., different wave components, Mean HR, Mean IBI) showed most subject-to-subject variability. Overall, inter-personal variability in both AMIGOS and PhyMER was relatively high, although to a lesser extent in PhyMER.

4.1.5 Label Distributions

Having reviewed the input parameters, the following analysis outlines the characteristics of the target variables (i.e., valence and arousal). Understanding the distributions aids in the interpretation of any model outcomes. In both datasets, valence and arousal were rated on a scale from 1 to 9.

Figure 9 illustrates the distribution of arousal and valence ratings for the AMIGOS short video dataset. Arousal ratings exhibited a bimodal distribution (SD: 1.78), with peaks observed around the moderate and high ranges. Similarly, valence ratings displayed a bimodal pattern as well (SD= 2.28), indicating concentrations in both negative and positive ratings.

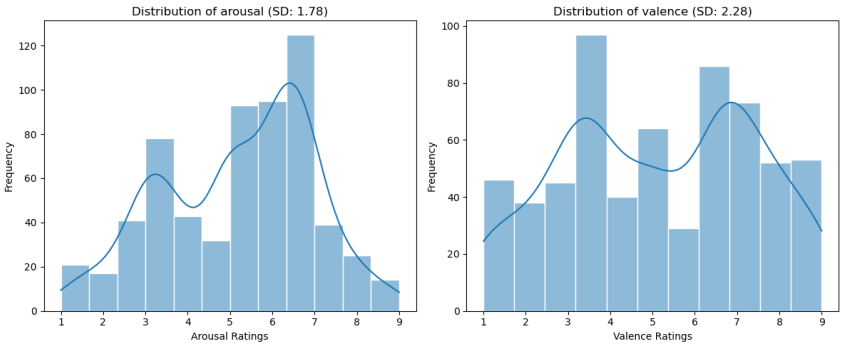


Figure 9: Label distributions for short video dataset AMIGOS The figure shows both valence and arousal distributions, with on the x axis the range (1-9) and frequency on the y axis (0-120). Both arousal (left) and valence (right) exhibit a slight bimodal pattern, though valence appeared more uniformly distributed than arousal.

For the AMIGOS long video dataset, shown in Figure 10, the distributions of both arousal (SD: 1.78) and valence (SD: 2.28) ratings were similar to those observed in the short video dataset, although arousal displays a roughly unimodal shape with a clear skew to the right (SD= 1.78). In turn, valence continued to show a more spread distribution (SD= 2.28), with a roughly unimodal pattern skewed towards the upper range.

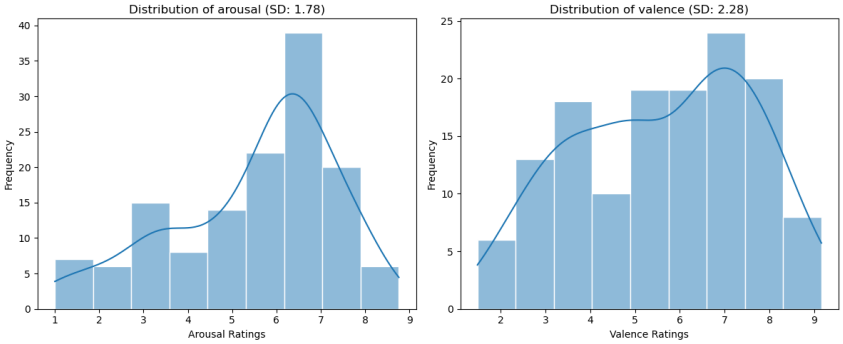


Figure 10: Label distributions for long video dataset AMIGOS The figure shows both valence and arousal distributions, with on the x axis the range (1-9) and frequency on the y axis (0-40). Both valence and arousal exhibit a right-skewed distribution, with most ratings concentrated at the higher end of both scales.

The PhyMER dataset displayed different patterns, as shown in Figure 11. Arousal ratings (SD: 2.24) presented a slight bimodal distribution, with decreased occupancy in the middle range. Valence ratings (SD: 1.92) in PhyMER also exhibited a unimodal distribution, skewed to the left, with predominantly lower valence scores (i.e., negative affect) and a lower prevalence of highly positive ratings. The overall distributions in PhyMER indicate a stronger tendency towards higher arousal and more negative valence compared to AMIGOS.

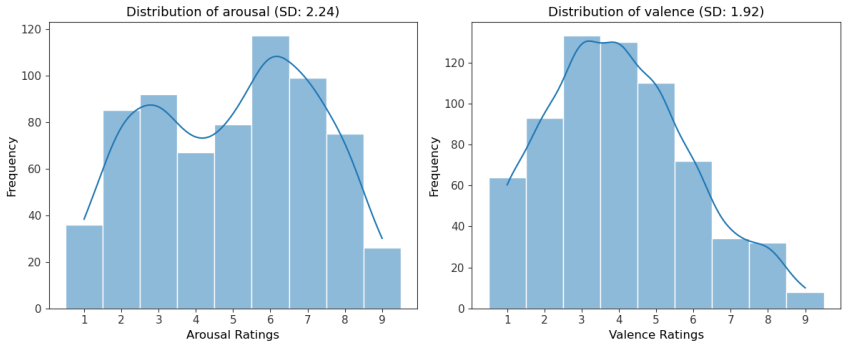


Figure 11: Label distributions for PhyMER dataset The figure shows both valence and arousal distributions, with on the x axis the range (1-9) and frequency on the y axis (0-120). Arousal (left) presented a slight bimodal distribution, with fewer ratings in the middle. Valence (right) exhibited a unimodal pattern with a slight left skew.

4.2 Personality-based Clustering

The next step is to identify the subgroups based on personality scores. This was necessary for the cluster-informed personalization explored in this thesis. Hence, the purpose of clustering in this thesis was to identify subgroups of similar personality profiles within the datasets. In this section, we present the results of K-means clustering on both datasets.

Figure 12 shows the clusters identified in the AMIGOS dataset, visualized using the first two principal components from a Principal Component Analysis (PCA) projection.

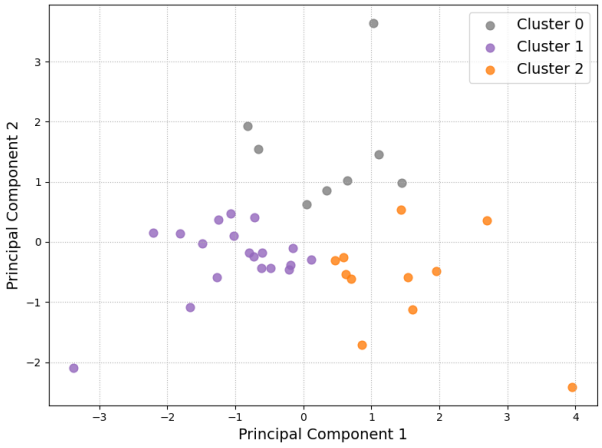


Figure 12: K-means Clustering Results of AMIGOS Scatterplot of k-means clustering applied to (standardized) Big Five trait scores in the AMIGOS dataset (N = 38). The first two principal components from a PCA are plotted on the axes for visualization purposes. Each dot represents a participant, colored according to their assigned cluster. Cohesion and separation were poor: (silhouette score = 0.210, DN = 1.49). PCs of each cluster centroid were used as input to the models.

The K-means clustering stabilized around n_init = 100 with a fixed random seed. The highest silhouette score in the AMIGOS data was observed at k = 2 (silhouette = 0.228), while k = 3 obtained a more favorable Davies-Bouldin score (Davies-Boulding Index (DBI)). In favor of equal comparison grounds with PhyMER, k = 3 (silhouette = 0.210) was selected, which was traded off with a minor decrease in silhouette score. See Appendix A for the plotted silhouette scores.

Figure 13 depicts the clusters identified in the PhyMER dataset, visualized using the first two principal components from a PCA projection.

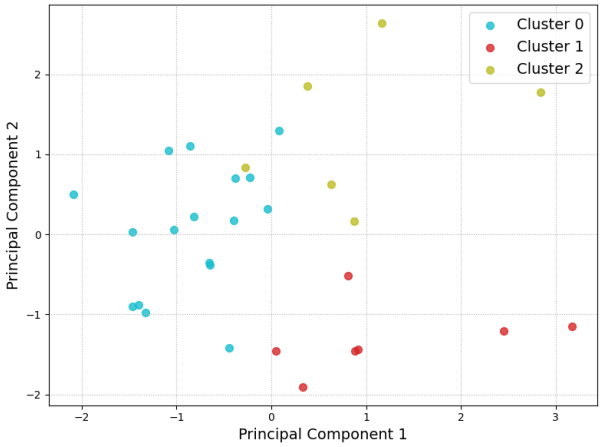
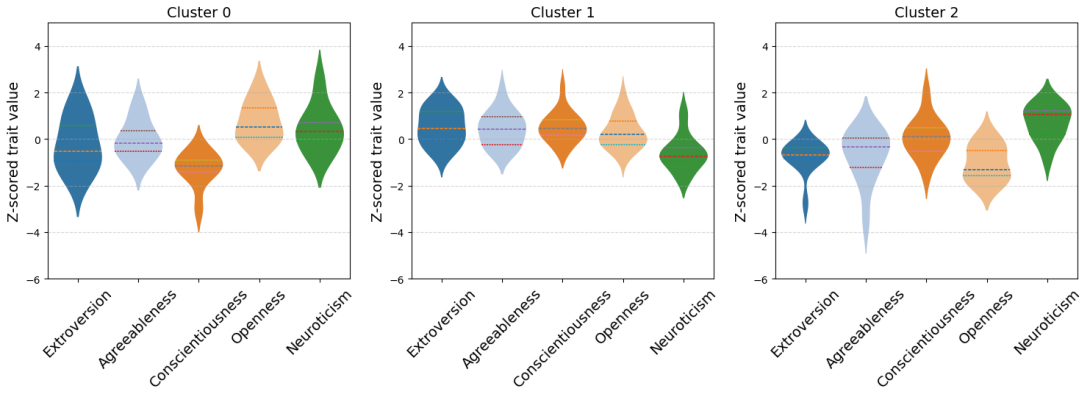


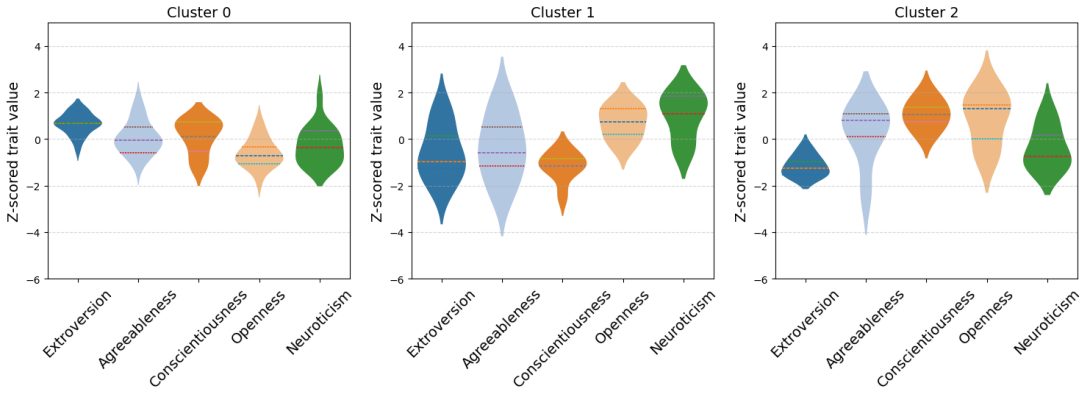
Figure 13: K-means Clustering Results of PhyMER Scatterplot of k-means clustering applied to (standardized) Big Five trait scores in the PhyMER dataset (N = 30). The first two principal components from a PCA are plotted on the axes for visualization purposes. Each dot represents a participant, colored according to their assigned cluster. Cluster quality was moderate: groups are somewhat cohesive yet still showed overlap (silhouette score = 0.26, DB = 1.38). The first PCs of each cluster centroid were used as input to the models.

In PhyMER, the optimal silhouette score was found at k = 3 (silhouette score = 0.26). See Appendix A for the plotted silhouette scores. The average similarity between and within each cluster and its closest neighbors was fair (DBI = 1.38), which suggests the clusters were reasonably separate, but still contained moderate internal spread. Although not visually apparent, clusters were slightly more separated and coherent than in AMIGOS. However, the clusters in both datasets were not clearly defined and had relatively high overlap. In both datasets, a few outliers could be identified, which likely influenced cluster cohesion.

After identifying the clusters, we may inspect the distributions of each cluster and investigate the characteristics of each cluster. Figure 14 depicts violin plots of the z-scored Big Five personality traits across the three clusters derived from AMIGOS (a) and PhyMER (b) datasets.



(a) AMIGOS Cluster Distribution



(b) PhyMER Cluster Distribution

Figure 14: K-means Cluster Distributions in AMIGOS and PhyMER. Distributions of clusters on standardized personality traits, where each subplot represents an identified cluster. The traits are displayed the x-axis with the z-scores on the y-axis. Each bar represents the number of participants assigned to each cluster. Most overlap in trait profiles was observed between AMIGOS Cluster 1 and PhyMER Cluster 2, and between AMIGOS Cluster 2 and PhyMER Cluster 1. Cluster 0 in both datasets displays distinct profiles. While there appeared to be some underlying groups across the datasets, interpretations remain descriptive rather than definitive.

To interpret the composition of the identified clusters, the Big Five personality traits were analyzed within each cluster. The descriptions below summarize the average trait profiles for each cluster, enabling the interpretation of the behavioral tendencies that characterize each cluster. Values are reported as standardized means (z-scores), with higher scores indicating stronger endorsement of the trait.

AMIGOS In AMIGOS, cluster 0 contained 19 participants (50%). On average, participants in this cluster scored slightly below average in extroversion with large variance (Mean = -0.23, SD= 1.16). Agreeableness was centered around the mean with a large spread (Mean = -0.01, SD= 0.84). With a notably below average conscientiousness of more than one SD (Mean = -1.28, SD = 0.77). Moderately high openness (Mean = 0.74, SD = 0.86) and slightly above average on neuroticism with a wide spread (Mean = 0.46, SD = 1.03). Overall, this cluster seemed to capture individuals who are slightly introverted, but relatively open and less organized.

Cluster 1 consisted of 11 participants (28.9%). Average Extroversion and Agreeableness scores were above the mean with a wide spread (Mean = 0.54, SD = 0.78) and (Mean = 0.43, SD = 0.76) respectively. Similar values were found for Conscientiousness (Mean = 0.5, SD = 0.67). Openness was slightly above average with moderate spread (Mean = 0.29, SD = 0.71). Lastly, Neuroticism was slightly below average (Mean = -0.6, SD = 0.77). Based on this, cluster 1 seemed to reflect relatively extroverted, agreeable, organized, and emotionally stable individuals.

Cluster 2 consisted of 8 participants (21.1%). In this cluster, extroversion appeared below average (Mean = -0.77, SD = 0.71). Agreeableness also seemed to be below average, albeit with a broad distribution (Mean = -0.731, SD = 0.156). Conscientiousness scores were centered around the mean (Mean = .07, SD = 0.88), and Openness scores were relatively low at (Mean = -1.03, SD= 0.78). Lastly, Neuroticism was notably above average (Mean = 0.7, SD = 0.76). Overall, this cluster corresponded to introverted, relatively neurotic individuals who were less open to experience.

PhyMER In PhyMER, 17 (56.6%) participants belonged to cluster 0. On average, more than half a SD above average and relatively narrowly distributed in Extraversion (Mean = 0.64, SD = 0.45). Agreeableness was centered near zero with moderate spread (Mean = 0.06, SD = 0.72). Conscientiousness was slightly above average (Mean = 0.11, SD = 0.75), and Openness relatively low (Mean = -0.60, SD = 0.63). Neuroticism was slightly below average, but with a large standard deviation (Mean = -0.28, SD = 0.82). Overall, this cluster likely represents individuals who are extroverted, emotionally stable, but less open to experience.

Secondly, 7 participants belonged to cluster 1 (23%). On average, participants exhibited low extroversion, with a wide spread, indicating a large disparity among them (Mean = -0.58, SD = 1.17). Agreeableness was centered slightly below zero and similarly wide spread, reflecting no clear tendency among participants (Mean = -0.36, SD = 1.4). Conscientiousness appeared clearly below average (Mean = -1.17, SD = 0.64). With moderately high Openness (Mean = 0.7, SD = 0.71) and Neuroticism, although very spread (Mean = 1, SD = 0.99). Overall, this cluster represented relatively introverted individuals who were slightly disagreeable and scored lower on the conscientiousness dimension. The cluster portrays individuals who are somewhat introverted, with neurotic tendencies, yet have a high openness to experience.

Lastly, 6 participants belonged to cluster 2 (20%). Extraversion scores appeared mostly below the mean with limited variance (Mean = -1.15, SD = 0.45). Agreeableness centered around the mean with large variance (Mean = 0.24, SD = 1.31). Conscientiousness (Mean = 1.06, SD = 0.67) and Openness (Mean = 0.88, SD = 1.15) both were predominantly above the mean. Lastly, Neuroticism scores were below average (Mean = -0.37, SD = 0.93). Overall, cluster 2 was characterized by introverted and organized individuals who were also open and emotionally stable.

If both datasets are representative of the broader population, similar personality-based clusters would be expected to emerge across them. Based on visual inspection of the descriptions, some similarities in cluster tendencies across datasets can be observed. The most consistent overlap was observed between AMIGOS Cluster 1 and PhyMER Cluster 2. Both clusters included participants who scored above average on conscientiousness and openness, and below average on neuroticism. Secondly, similarities were found between AMIGOS Cluster 2 and PhyMER Cluster 1. Both clusters included participants with low extroversion, low conscientiousness, and high neuroticism. Lastly, AMIGOS Cluster 0 and PhyMER Cluster 0 showed the least overlap. The former featured individuals with below-average conscientiousness and slightly above-average neuroticism, while the latter included participants who were more extroverted, emotionally stable, and less open to experience. This suggests that Cluster 0 captured different profiles in each dataset.

The observed patterns suggest that the clustering identified some similar latent subgroups across the datasets. However, interpretations should be treated as descriptive rather than definitive, as cluster cohesion in both datasets was low, indicated by the silhouette and scores, and sample sizes were relatively small.

4.3 Personality Features

With the personality clusters defined, Pearson's r was used to examine the direct univariate associations between self-reported arousal and valence and the personality-based features. These included the individual Big Five traits and the Principal Component (PC)s representing the participant's assigned cluster centroid. This analysis provided initial insights into the relationship between personality traits and affective self-report ratings.

In Figure 15, correlations between the Big Five traits and cluster PCs can be observed for both the short and long AMIGOS datasets.

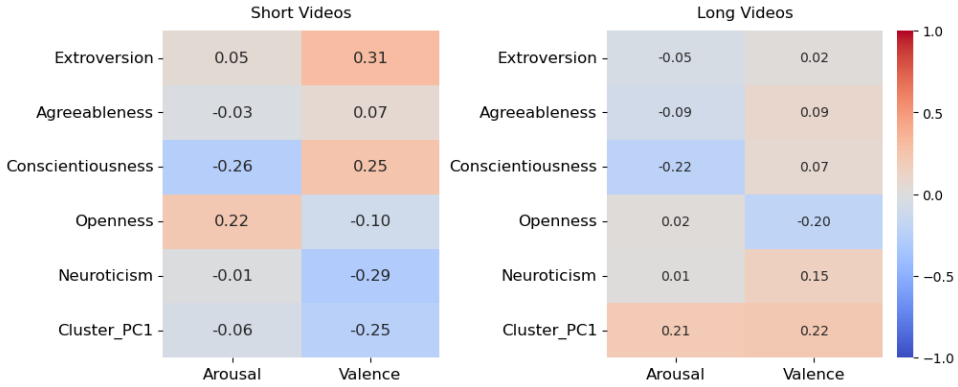


Figure 15: Correlations Between Personality Features and Affective Ratings in AMIGOS Pearson's r correlations between Big Five traits, the first PC of the assigned cluster centroid and self-reported arousal and valence are shown for both short (left) and long (right) video conditions in the AMIGOS dataset. While most associations were weak, notable patterns include negative correlations between conscientiousness and arousal, as well as a moderate link between extroversion and valence in short videos. PC1 showed modest correlations, particularly in the long condition.

In the short video experiment of AMIGOS, extroversion correlated negligibly with arousal (r = 0.05) but moderately with valence (r = 0.31), indicating that more extroverted participants tended to evaluate their affective state as more pleasant. Conscientiousness was inversely related to arousal (r = -0.26) and positively to valence (r = 0.25); higher-scoring individuals reported calmer yet more pleasant states. Openness showed a small positive link with arousal (r = 0.22) and a weak negative link with valence (r = -0.10). Neuroticism displayed a small negative correlation with valence (r = -0.29) but not with arousal (r = -0.01). The observation that higher neuroticism is related to negative affect aligns with theoretical expectations [16]. The cluster component PC1 was weakly negatively associated with both both arousal (r = -0.06) and valence (r = -0.25).

When examining the long recordings, it was observed that the effects decreased as the stimulus duration increased. Extroversion was no longer predictive of either label ($|r|$ 0.05). Openness is no longer related to arousal (r = 0.02), but it has a slightly increased negative association with valence (r = -0.20) compared to the short videos. Conscientiousness kept a weak inverse relation with arousal (r = -0.22) but its link with valence became negligible (r = 0.07). Interestingly, neuroticism reversed direction, correlating weakly and positively with valence (r = 0.15). The PC now correlated positively, although weakly, with arousal (r = 0.21) and valence (r = 0.22).

Similarly, Figure 16 summarizes the correlations between personality features, both individual traits and cluster-derived components, and arousal and valence ratings in the PhyMER dataset.

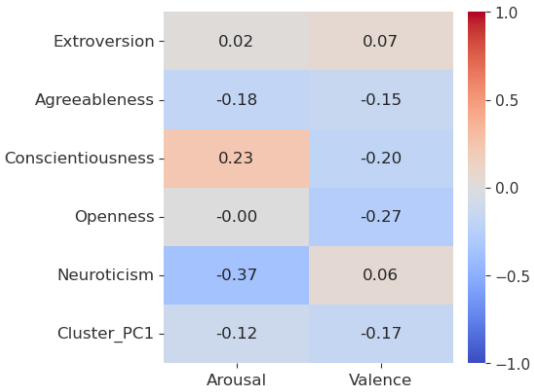


Figure 16: Correlations Between Personality Features and Affective Ratings in PhyMER Pearson's r correlations between Big Five traits, the first PC of the assigned cluster centroid, and self-reported arousal and valence in the PhyMER dataset. Most associations are weak to moderate. Neuroticism shows the strongest (negative) correlation with arousal.

In the PhyMER dataset, extroversion was close to zero for both arousal (r = 0.02) and valence (r = 0.07). Agreeableness showed small negative associations with both arousal (r = -0.18) and valence (r = -0.15). Conscientiousness showed a positive relation with arousal; higher scores predicted higher arousal (r = 0.23) but lower valence (r = -0.20), though the correlation was weak. Openness was unrelated to arousal (r = 0) but moderately negative for valence (r = -0.27). Neuroticism had the strongest relationship with arousal, showing a moderate negative correlation (r = -0.37), while being unrelated to valence (r = 0.06). The correlation between the PC and the labels showed similarities with the AMIGOS short-video data; the PC was negative for both labels.

Across datasets, the correlations between personality and affective dimensions were generally small to moderate. Importantly, the relationships appeared dataset-specific, as the directions of some relationships were reversed across datasets. For instance, neuroticism was negatively related with valence in the AMIGOS short dataset (r = -0.25). This this effect reversed in the long dataset (r = +0.15), and disappeared in the PhyMER dataset (r = -0.06). This suggests that the results should be interpreted with caution and suggesting potential systematic differences between datasets, such as in participants, stimuli, or the quality of personality assessments.

4.4 Hyperparameters outcomes

This section summarizes the outcomes of hyperparameter optimization, which helps improve understanding of the models' behavior and the characteristics of the data. The hyperparameter tuning was performed using a grid search within a nested Leave-One-Out Cross-Validation (LOOCV) setup, in which the parameters that yielded the lowest Root Mean Squared Error (RMSE) on average were selected. The best performing hyperparameter configurations for each model are provided in Appendix D.

AMIGOS For the short dataset, the Random Forests (RF) models generally favored shallower trees (max depth = 5), with sqrt or 0.5 as the max features setting and small min samples leaf values (3–5). These settings indicate that the model was reducing the risk of overfitting by favoring simpler trees. Support Vector Regression (SVR) models performed best with a radial basis function (RBF) kernel for arousal prediction (C = 10.0, gamma = 1.0). This suggests that the relationship between features and arousal labels was nonlinear and required relatively flexible, localized decision surfaces (higher gamma). The model benefited from allowing more flexibility (higher C) without overfitting. However, valence prediction required more regularization (C = 0.1) and a smaller gamma (0.0001). The small gamma implies that patterns related to valence were less clearly separable, which required the model to generalize

more broadly. The lower C further enforced regularization, a method by which the model prevents overfitting by penalizing overly complex models. This indicates that the features contained less consistent signal for predicting valence compared to arousal

In the long dataset, Random Forests favored deeper trees (max depth = 10) and consistently used a strict parameter for feature selection (max features = sqrt). This suggests that in the long videos, the relationship between input features and labels was more obscured compared with the short videos. SVR models used performance with RBF kernels across all settings, which suggests the data was nonlinear. The high regularization (C = 10.0) and moderate error tolerance (= 0.2) indicate an emphasis on closely fitting the data while allowing for minor deviations.

Overall, the settings suggest that the relationship between input features and labels was not clearly distinguishable, due to redundancy or weakly informative features in the dataset, which explains the need for regularization in all models.

PhyMER For the RF models, the baselines preferred deeper trees (max depth = 10) while the models with traits or clusters preferred shallow trees (max depth = 5). Similarly, either half of the features (max features = 0.5) or the square root (max features = sqrt) was used, indicating the model preferred the use of fewer features at each split. These settings suggest a preference for balancing model complexity with regularization, which prevented the model from creating highly specific trees and thereby improving generalization.

For SVR, the optimal parameters varied substantially between models. In general, most models performed best with the RBF kernel (nonlinear), suggesting that the relationships between the physiological features and the affective dimensions were complex and nonlinear. The best performing SVR model (Traits on valence) used a relatively high regularization parameter (`C` = 10), and a small epsilon-insensitive margin (`epsilon` = 0.05) with an RBF kernel. This configuration indicates the best model on the training data was sensitive to training errors (small epsilon) and heavily penalized those errors (large C). This particular configuration may have been an attempt to capture high individual variability in valence responses, potentially leading to overfitting on the training data and therefore poor generalization to unseen users.

4.5 Model performance

This section presents the empirical results of the regression models RF and SVR for predicting Arousal and Valence using both the AMIGOS and PhyMER datasets. For AMIGOS only, the results from the binary classification are presented to verify the pipeline with the authors. Performance was evaluated across different models (Baseline, Traits, Clusters). Recall that the Baseline model consisted of only the biosignal features. Traits included the participant's Big Five scores, as well as biosignal features. Lastly, the Clusters model included the PC of the personality-based cluster assignment in addition to the biosignal features. This PC was intended to introduce a group-based identifier to the model, which would enable the model to identify group-specific patterns. This was therefore a proxy to true cluster-based modeling, as this was not feasible with the data at our disposal. As evaluation metrics, Coefficient of Determination (R^2), Mean Absolute Error (MAE) and RMSE were used.

4.5.1 AMIGOS

Table 9 summarizes the predictive performance of the models on short and long videos, separately for valence and arousal.

			Short Video Segments			Long Video Segments		
Variant	**Algorithm**	**Measure**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**
No Personality (Baseline)	RF	Arousal	0.038	1.436	1.776	-0.176	1.718	1.928
		Valence	-0.010	1.814	2.066	0.002	1.819	2.055
	SVR	Arousal	-0.006	1.521	1.816	-0.020	1.549	1.796
		Valence	-0.005	1.790	2.061	-0.020	1.809	2.077
Traits	RF	Arousal	0.014	1.462	1.798	-0.096	1.613	1.861
		Valence	-0.008	1.795	2.064	-0.042	1.861	2.099
	SVR	Arousal	-0.004	1.513	1.814	-0.024	1.550	1.799
		Valence	-0.005	1.790	2.060	-0.021	1.810	2.077
Clusters	RF	Arousal	-0.046	1.504	1.852	-0.188	1.721	1.938
		Valence	-0.014	1.813	2.070	-0.025	1.844	2.082
	SVR	Arousal	-0.007	1.518	1.817	-0.022	1.550	1.798
		Valence	-0.005	1.790	2.061	-0.021	1.810	2.078

Table 9: Model Performance for AMIGOS Performance of regression models (RF and SVR) across Baseline, trait-informed, and cluster-informed variants for predicting arousal and valence in the long and short datasets. Generally, no model outperformed the Baseline; all R^2 values remained close to or below zero, indicating poor generalization to unseen participants. Longer segments consistently yielded negative results, indicating poorer performance of the models trained on data from longer segments.

Summary The models generally did not capture any of the variance in self-reported affective dimensions for unseen participants; most R^2 values were negative. Arousal was better predicted than valence in the short videos. For instance, arousal predictions spanned from -0.046 to 0.038, while valence was exclusively negative, ranging from -0.005 to -0.14. For long videos, this was the opposite; valence was consistently less negative (-0.024 to 0.002) than arousal (-0.188 to -0.020). The best-performing model, the Baseline RF model for arousal, explained only 4% of the total variance in arousal ratings.

Generally, short video segments were easier to model than long segments. Averaged over algorithms and variants, short-arousal and short-valence showed mean R^2 values of –0.002 (arousal) and –0.008 (valence), whereas the R^2 in the long segments were on average –0.088 (arousal) and –0.021 (valence). Short segments yielded less-negative R^2 values for arousal than for valence, while this ordering was reversed for long segments. Negative R^2 values indicate that a model that continuously predicts the mean would perform better than the fitted models.

The MAE and RMSE supported this observation. The lowest errors were obtained by the baseline model using RF for the short-arousal task, which had an MAE of 1.436 (18.1%), and RMSE of 1.776 (22.2%). The percentages in parentheses are normalized by the scale width (9-1 = 8). All other models were within ± 2% of these minima on the normalized scale, which implies that differences in prediction error were small.

Interestingly, SVR appeared to be a better algorithm for valence, while RF performed better for arousal in the short videos. RF consistently outperformed SVR for arousal tasks, whereas the opposite occured for valence prediction. However, when examining the long-video results, this is again reversed; now, SVR explained more variance in arousal, while RF explained more variance in valence. In the long videos, this pattern was not observed in the error metrics.

Figure 17 depicts ($\hat{y}$) versus actual (y) values of the RF baseline model. See Appendix B.1 for the complete set of figures.

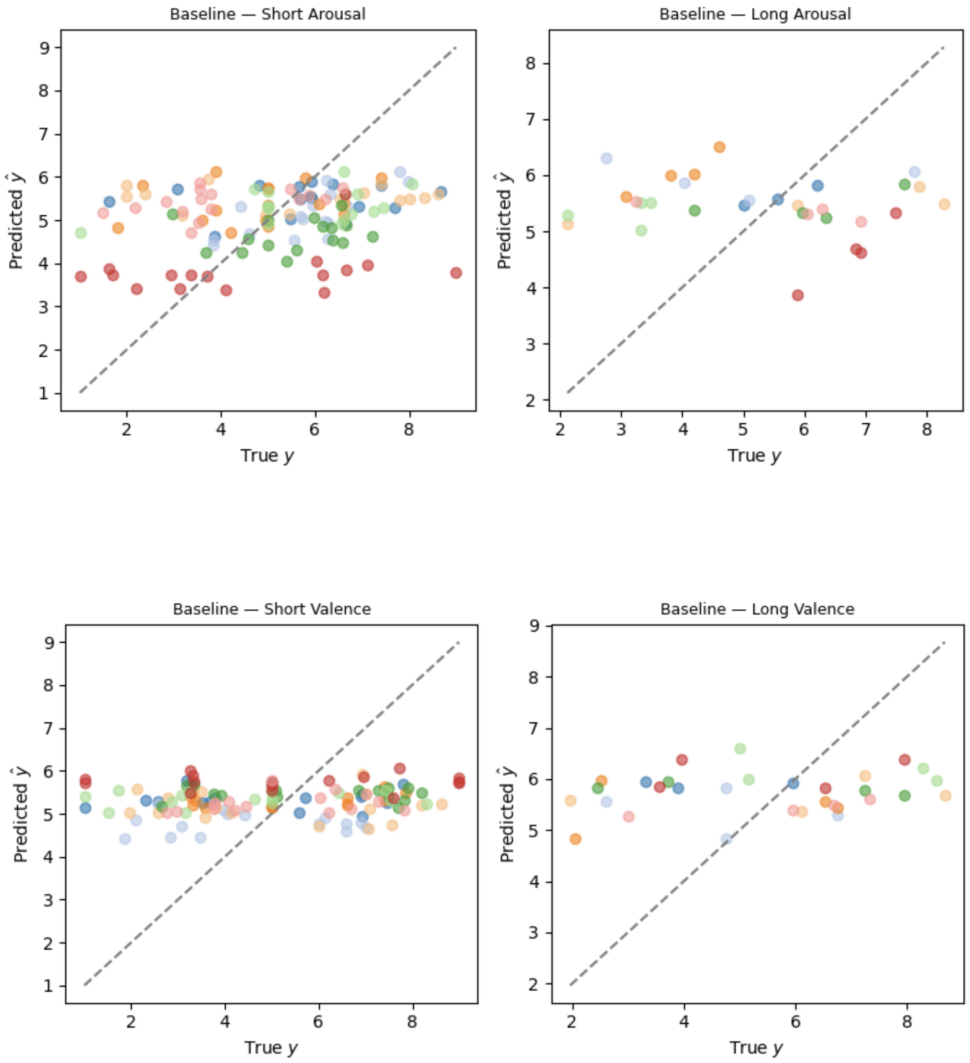


Figure 17: True y vs. Predicted $\hat{y}$ Arousal and Valence Ratings (Baseline, RF for AMIGOS. Scatterplots showing predicted versus actual self-reported arousal (top) and valence (bottom) ratings for both short and long video segments in the AMIGOS dataset. The predictions clustered near the mean, illustrating the model's tendency to regress toward the average label value. This indicates the model failed to learn meaningful patterns from the features, especially for valence. Predictions deviated only slightly from the global mean, explaining the poor generalization and low R^2 scores. Color indicates individual participants, and some inter-participant variability in prediction accuracy can be observed.

When observing the predicted ($\hat{y}$) versus actual (y) values of the RF baseline model, there is little to no trend observed. The model continuously predicted values around the mean, especially for valence (see Appendix B.1). This further illustrates the model has not learned any meaningful pattern between the input features and its labels; the model minimized MAE and RMSE by staying near the densest region of the label space. As a consequence, it failed to correctly estimate both high and low ratings in either affective dimension.

The diagnostic plots further showed this regression to the mean. The residual-versus-fitted plots (See Appendix C.1) showed the fitted values are tightly clustered, particularly for the SVR models. In some instances, the predicted $\hat{y}$ values were nearly constant, explaining low R^2 scores. The RF models were slightly more flexible, corresponding to marginally improved performance. Some group-level bias was also present (points are color-coded by Participant); the models appeared to perform better for some participants than for others.

Personality as personalization Finally, adding personality information did not improve predictive accuracy; most models yielded virtually identical results. In the short video experiment with RF on arousal, a decrease in explained variance was observed when adding personality information to the model. For example, compared to the Baseline, Traits decreased by -0.024 explained variance, while the Clusters model decreased by -0.084 compared to the Baseline. This pattern was not observed in the long video experiment. Instead, the long videos appeared to benefit slightly from additional context. For instance, Traits RF for arousal increased explained variance by +0.08 compared to the Baseline, which decreased again with -0.092 in Clusters. This observation only stood for arousal, as explained variance for valence decreased in the Traits and Cluster models. Furthermore, the SVR models showed little to no difference in performance across all variants for both short and long video segments, underscoring the limited impact of personality features on SVR performance.

Ultimately, the observed differences across variants were marginal and hold no practical significance. This suggests that the current methods of incorporating personality information did not yield any benefit for predicting affect in unseen participants.

4.5.2 PhyMER

Next, the regression output from the PhyMER dataset is presented. All results are summarized in Table 10

Setting	Algorithm	Task	R^2	MAE	RMSE
No personality (Baseline)	RF	Arousal	0.061	1.874	2.152
		Valence	-0.049	1.605	1.977
	SVR	Arousal	0.007	1.880	2.213
		Valence	-0.009	1.577	1.938
Traits	RF	Arousal	0.043	1.856	2.172
		Valence	-0.021	1.602	1.950
	SVR	Arousal	-0.022	1.903	2.245
		Valence	0.003	1.559	1.928
Clusters	RF	Arousal	0.069	1.842	2.143
		Valence	-0.037	1.610	1.966
	SVR	Arousal	-0.073	1.946	2.301
		Valence	-0.015	1.577	1.944

Table 10: Model performance with PhyMER. Performance of Random Forests (RF) and Support Vector Regression (SVR) across model variants; Baseline (physiological only), Traits (Big Five traits added), and Clusters (cluster-based components added) for predicting continuous arousal and valence labels. Metrics included R^2, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). All models exhibited limited explanatory power, with R^2 values that were either close to or below zero. RF models generally outperformed SVR, particularly in terms of arousal. The lowest errors and highest R^2 were obtained using the Clusters variant with R^2.

Summary With PhyMER, slightly improved results were observed compared to AMIGOS; however, none of the models still explained much of the variance in valence or arousal in unseen participants. R^2 values were similarly close to zero and in some instances negative. Arousal yielded marginally better results. For example, the explained variance in arousal ranged from -0.073 to 0.069. For valence, R^2 values were almost all negative, ranging from -0.037 to 0.003. In general, RF models performed better

for arousal than for valence, while SVR performed better for valence. The model that explained the most variance was Clusters, RF for arousal ($R^2 = 0.069$). This was followed by the Baseline, RF for arousal ($R^2 = 0.061$) and lastly Traits, RF for arousal ($R^2 = 0.043$). Nevertheless, these poor fits suggested that the models learned almost no meaningful patterns to predict arousal and valence across participants.

In terms of error, the MAE ranged from 1.56 to 1.95, corresponding to relative errors between 19.49% and 24.33%, while RMSE ranged from 1.93 to 2.30 (24.10% to 28.76%, normalized). In other words, the models predict on average within two Likert-scale points of the actual ratings. As RMSE penalizes large errors more heavily, the gap between MAE and RMSE suggests that the model prediction accuracy may vary between participants or trials. Overall, RF models tended to outperform SVR models in terms of both MAE and RMSE for most variants. The best-performing configuration was Clusters, RF for arousal, resulting in an MAE of 1.84 (23.03%) and an RMSE of 2.14 (26.79%). For valence, the lowest RMSE was obtained for Traits, SVR (1.93, 24.10%).

The residual-versus-fitted plots for all model variants (see Appendix C.3) further showed both RF and SVR models produced predictions clustered around the mean. The broader fitted range in RF models corresponded to marginally better performance, while the SVR models showed minimal variance in fitted values, supporting the low R^2 scores.

In Figure 18, an example visualization is provided for one of the models (Baseline, RF), which depicts the predicted $\hat{y}$ values plotted against actual y, color coded by participantID. Evident from the figure, the model failed to capture the actual variance in y labels, as no cohesion was observed between the predicted and actual labels, and the model predicted values around the mean. The additional scatter plots (see Appendix B.3) demonstrated comparable performance patterns across conditions.

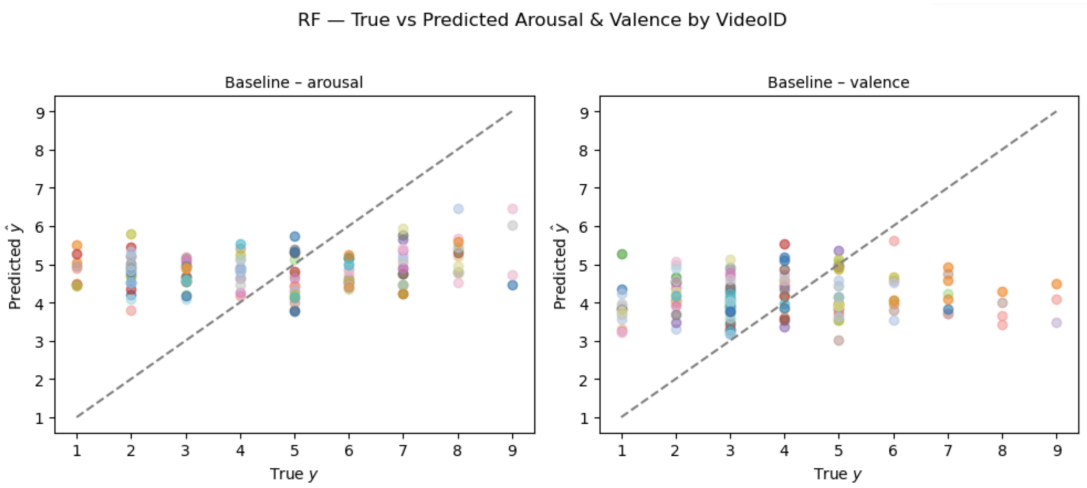


Figure 18: True y vs predicted $\hat{y}$ labels for baseline RF model in PhyMER Scatter plots show predicted $\hat{y}$ values against true y values for arousal (left) and valence (right), color-coded by participant. The model consistently predicts values near the center of the scale, failing to capture the true range of the target variables and reflecting its limited generalization ability. Note that the original label data in PhyMER were integers and contained no floats.

Personality as personalization method When comparing the different settings, the incorporation of personality information, whether through direct trait inclusion or cluster assignment, did not lead to a meaningful improvement in model performance for the PhyMER dataset. The highest explained variance and lowest errors were found in the Cluster, RF on arousal, and the Traits, SVR on valence.

Adding the cluster assignments using RF reduced the error the most for arousal (-0.032 MAE and -0.009 RMSE compared with the Baseline) and increased explained variance by 0.008. For valence, using raw trait scores with SVR gave the lowest errors, but the gain over Baseline was minimal (+0.012 R^2,

-0.018 MAE, and -0.010 RMSE).

In conclusion, injecting personality information, whether as raw trait scores or cluster assignments, offered a marginal benefit, although this benefit was neither notable nor consistent. The maximal improvements over the no-personality Baseline were a decrease in MAE of approximately 0.02-0.03 (0.3% of the 8-point scale) and RMSE of 0.01 (0.1%). Therefore, adding personality did not improve generalization within the PhyMER dataset.

4.5.3 Binary classification

Finally, to verify our pipeline and ensure an equal ground for comparison with the AMIGOS paper [15], which only reported binary classification results, we additionally conducted a binary classification analysis. For this task, again RF and Support Vector Classification (SVC) were used. Classes in the short dataset were effectively balanced (arousal = 318:305; valence = 289:334). In the test set, this was (arousal = 66:62; valence = 68:60). This was not the case for the long video dataset, and as this analysis serves the main purpose of validation of the pipeline, we only include the short-video classification results. The results of the short dataset are summarized in Table 11.

Setting	Algorithm	Task	Accuracy	Balanced Acc.	AUC	F1
No Personality (Baseline)	RF	Arousal	0.531	0.529	0.538	0.565
	SVC	Arousal	0.563	0.563	0.572	0.569
	RF	Valence	0.492	0.503	0.505	0.496
	SVC	Valence	0.516	0.528	0.484	0.523
Traits	RF	Arousal	0.586	0.577	0.518	0.662
	SVC	Arousal	0.531	0.526	0.561	0.589
	RF	Valence	0.523	0.510	0.525	0.430
	SVC	Valence	0.539	0.531	0.434	0.468
Clusters	RF	Arousal	0.555	0.553	0.559	0.584
	SVC	Arousal	0.563	0.567	0.507	0.533
	RF	Valence	0.477	0.487	0.512	0.481
	SVC	Valence	0.578	0.572	0.491	0.518

Table 11: Binary classification performance for short video segments in AMIGOS Performance of Random Forests (RF) and Support Vector Classification (SVC) across model variants; Baseline (physiological only), Traits (Big Five traits added), and Clusters (cluster-based components added) for predicting binary arousal and valence labels. Overall, performance gains from including personality features were minimal or inconsistent. Arousal was easier to predict than valence, and RF generally outperformed SVC across metrics.

The Traits obtained the best F1 score, RF model on arousal (F1 = 0.662), while the highest AUC was reached by the Baseline SVC model on arousal (Area Under the Curve (AUC) = 0.572). No single configuration excelled simultaneously on both metrics. Furthermore, these peak scores only marginally surpassed the threshold of random chance, signaling limited practical utility of the classifiers.

Across algorithms and tasks, adding personality information provided little to no effect and was mildly harmful to discrimination power. Relative to the Baseline, the Traits variant produced a 2.9% reduction in AUC and 0.2% in F1 score. Similarly, the Clusters variant reduced AUC by 1.4% and F1 score by 1.7%. Models struggled more with predicting valence than arousal; across all settings and algorithms, arousal classification achieved a 10.3% higher AUC and 20.1% higher F1 score. Lastly, RF outperformed SVC by an average of 3.5% in AUC and 0.6% in F1 score.

5 Discussion

We now discuss the implications of our findings and how they situate in the body of affective literature. In this study, we investigated whether supplementing affect recognition models with personality-based cluster assignments improved the generalization to unseen participants. Affect was represented as two continuous dimensions, valence and arousal, based on participants' self-reported ratings. We employed two classical machine learning algorithms for this task: Random Forests (RF) and Support Vector Regression (SVR). To validate our pipeline with the authors AMIGOS and PhyMER, we additionally performed binary classification for the AMIGOS dataset. To our knowledge, no prior work has explored personality-based clustering as a personalization strategy, and few studies have addressed affect prediction using continuous variables. Because data restrictions prevented true cluster-based modeling, we approximated it by including the first Principal Component (PC) of each cluster centroid as a group-level identifier. To answer our research questions, we trained a baseline model only on features extracted from biosignals. Next, our proposed model included two additional features, which contained the first PC of the personality-based cluster assignment as the group identifier. To assess the added value of clustering, we included a model that incorporated the five personality traits directly alongside the physiological features. Furthermore, we investigated the implications of training the models on data obtained from shorter stimuli (50-250 s) and longer stimuli (14-24 min). Finally, to determine the generalization of our findings across datasets, two separate datasets were analyzed: AMIGOS and PhyMER.

First, the Baseline model performed poorly across datasets and algorithms in predicting absolute valence and arousal levels for unseen participants. For example, the baseline RF model trained on the short segment data in the AMIGOS dataset explained only 4% of the variance in arousal and yielded a negative Coefficient of Determination (R^2) for valence (-1%). However, including personality information, either directly or via cluster assignment, did not substantially or consistently improve the generalization capabilities of the models. For example, the explained variance of arousal in RF trained on the short segment data in AMIGOS decreased to 1.4% when adding trait information. This decreased to -4.6% when using cluster assignments. In PhyMER, the Clusters model showed a slight improvement in explained variance (+0.08 R^2). Next, the models struggled to accurately predict valence and arousal more in the long segments than in the short segments; prediction errors were systematically larger, and the explained variance was almost exclusively negative in the long segments. Furthermore, RF tended to outperform SVR, although this improvement was minor and inconsistent across tasks. Generally, arousal was more accurately estimated than valence. In PhyMER, arousal consistently preceded valence in terms of prediction accuracy. In AMIGOS, neither valence nor arousal consistently outperformed the other. However, AMIGOS' optimized hyperparameters suggest the valence-related patterns were less defined, as the model favored more diffuse decision boundaries than when predicting arousal. Lastly, the descriptive analyses revealed weak correlations of the physiological features with valence and arousal labels. Lastly, the Intraclass Correlation Coefficient (ICC) analysis revealed great interpersonal differences across features.

5.1 Interpretation of findings

The performance of our baseline models was comparable to those in the original AMIGOS [15] and PhyMER [27] papers. In AMIGOS, a binary Support Vector Classification (SVC) classification model yielded an F1 score of 0.570 for short-valence and 0.585 for short-arousal, trained on features from Electroencephalogram (EEG), Electrodermal Activity (EDA), and Electrocardiography (ECG) modalities. Our binary SVC based on features from just EDA and ECG yielded similar results, with F1 scores of 0.523 for short-valence and 0.569 for short-arousal. Furthermore, the authors of PhyMER conducted subject-independent regression using Extreme Gradient Boosting (XGBoost) trained on EEG, EDA, and

Photoplethysmography (PPG) features. The authors reported an Mean Absolute Error (MAE) of 1.941 for arousal and a MAE of 1.587 for valence. Likewise, our baseline SVR model, trained on features extracted from EDA and PPG, yielded MAE = 1.880 for arousal and MAE 1.577 for valence. The close agreement between our baseline results and those reported in the AMIGOS and PhyMER corroborates our implementation and validates our methodological choices. Furthermore, the low R^2 values for user-independent valence and arousal prediction were confirmed by Siirtola et al. [89] who reported R^2 values ranging from -1.14 to 0.02 across models for both valence and arousal using a RF model. Interestingly, Siirtola et al. [89] found that the explained variance of valence and arousal significantly improved to range from 0.64 to 0.80 when using more advanced frameworks, such as Long Short-Term Memory (LSTM). Notably, another study by Galvao et al. [37] accurately predicted exact arousal and valence ratings using classical machine learning. For example, their RF achieved MAE of 0.115 for arousal, and MAE of 0.158 for valence. Galvao et al. achieved these results using spectral EEG analysis, which may partially explain the discrepancy between the outcomes.

Whether clustering is a viable method for improving generalization remains inconclusive. Any improvements in our results were inconsistent and of limited practical relevance. Related works that compared various personalization approaches, including clustering, reported inconsistent findings. For example, Han et al. [78] found minor effects of clustering on generalization and concluded that the most effective strategy was dataset- and algorithm-dependent, which aligns with our observations. For example, using AMIGOS data, Han et al. [78] found an AUC of 0.504 for arousal, which increased minimally to 0.514 for their cluster-based model. Similarly, Tervonen et al. [19] found that cluster-based models yielded comparable to generalized models for stress detection. However, other studies have reported positive results for cluster-based modeling. For instance, Can et al. (2020) [18] clustered individuals based on their baseline stress levels and found that clustering improved performance than a generalized model, although increasing cluster size negatively affected predictions. Similarly, Adler et al. [22] clustered participants based digital trace data distributions. They found that personalization generally aligned the training and testing data, though this effect was inconsistent, and performance dropped as the number of included near neighbors increased.

The weak quality of the clusters themselves may explain the lack of consistent performance gain in our study. The silhouette coefficients in both datasets were poor to acceptable (silhouette = 0.22, 0.25). This indicates that while algorithm identified some latent structure, the resulting groups were overlapping and had large internal spread. As a consequence, the diffuse boundaries may have obscured any patterns specific to that cluster. Possibly, the clustering of the traits led to information loss, which obscured individual variability that was better preserved in the raw scores. Second, as the sample size of both datasets was somewhat limited, we approximated cluster-based modeling with *cluster-informed* modeling. Due to data scarcity, we provided the model with the first PC of each participant's cluster assignment, rather than training a separate model on each subset. Our proxy is thus a gentler but less potent solution than training a model exclusively on users who share similar characteristics. Lastly, while clustering based on personality did not prove to be effective, clustering based on a variable more directly related to affect may yield improved results, for example, by grouping people based on physiological characteristics, such as their resting state baselines.

Although the Traits model slightly outperformed the Clusters model in some instances, the addition of traits generally did not substantially improve performance. This finding was partially confirmed by other research; for example, Pant et al. found that personality has a weak to low impact on emotion recognition when using classical machine learning models on the PhyMER dataset [27]. Likewise, Martinez et al. [90] found that neither age, sex, nor personality traits significantly correlated with the classification of arousal and valence using the AMIGOS dataset. However, other studies did find adding personality led to model improvement. For example, Zhao et al. [91] found that incorporating personality improved the prediction of valence and arousal levels by 10%. These findings suggest that the impact of personality

on affect recognition is not definitive, and depends on the modeling approach, dataset characteristics, or the way personality information is integrated.

The limited impact of personality traits in our models may result from the coarse granularity of personality data, which does not align well with the fluctuating physiological features used to recognize valence and arousal levels. For example, a body of literature suggests that certain traits, such as neuroticism or extroversion, correspond to baseline differences in physiology and affective styles across individuals (e.g., [8], [53]). However, our results suggest that momentary changes do not reflect personality differences enough. Instead, personality may be more strongly associated with stable affective dispositions, such as mood, rather than emotional fluctuations. As a result, the additional trait features introduced noise without providing much explanatory power.

Next, our analyses indicated that longer-duration affective states (i.e., mood) were more challenging to model than brief, short-term affective responses (i.e., emotions). Specifically, predictive models trained on short-video features achieved higher accuracy for valence and arousal than those trained on long-video features. This aligns with findings from the AMIGOS article, which reported that short videos elicited more distinct and measurable physiological and affective responses. In contrast, affective reactions to long videos were more subtle and variable [15]. Katsimerou & Redi [5] attempted to address this problem by proposing a model that considers mood as a series of brief affective states, where more distant states received less weight in the final prediction.

Several factors can explain the performance gap between long and short segments. First, during preprocessing, we found that longer video segments contained more sensor noise, which is unsurprising, as the risk of artifacts increases with time. Similarly, longer recordings are subject to more natural fluctuation. As participants assign their affective ratings retrospectively, those labels may not accurately reflect every fluctuation within the session. These fluctuations can negatively affect label-feature correlations. As we did not take these temporal dynamics into account, information on the underlying processes that constitute mood was essentially lost. Lastly, the long video data had a significantly smaller sample size (141) than the short video data (676). The consequence is that outliers, whether random or non-random, have had an increased impact on the predictions and further harmed generalization. Taken together, these findings suggest that longer-duration affective states introduce additional modeling challenges, such as higher temporal variability compared to short-term emotional responses [5], [15].

Lastly, a finding that was consistent with the literature was that valence was more complex to predict than arousal. Affective computing studies generally agree on this finding (e.g., [37], [89]–[91]). Both the model performance results and the univariate correlation analysis in our study indicated that the relationship between biosignals and valence is weaker or more complex compared to arousal. This supports the notion that biological signals are not as effective predictors of valence as they are of arousal. For example, research on electrodermal activity suggests EDA cannot reliably distinguish different levels of valence [8], [44]. Alternatively, some articles have suggested that the discrepancy may result from extreme valence labels being less represented than extreme arousal levels [89], [91].

Furthermore, predictions for valence tended to have no explained variance but consistently lower error metrics (MAE, Root Mean Squared Error (RMSE) than arousal. This may be explained by the fact that the models predicted more closely to the mean for valence. Seemingly acceptable errors can still be achieved by consistently predicting values close to the mean, even though no variance is explained.

Lastly, our analyses revealed that different algorithms performed differently for valence and arousal. In both datasets, RF outperformed SVR in arousal prediction. In contrast, SVR almost always performed better on valence tasks, both in explaining variance and minimizing MAE and RMSE. Similar works did not support this finding. For instance, [37] found that RF consistently outperformed SVR in terms of MAE and RMSE, for both valence and arousal. However, their study was based on spectral EEG features, which appear to be more strongly related to valence/arousal ratings than EDA or ECG features. Two

interacting factors may explain this observation.

One explanation for this may be due to the extreme regularization settings of SVR. The hyperparameters for SVR indeed revealed that the models struggled more with identifying clear patterns for valence than for arousal. As a response, SVR minimized error by only predicting the mean, which was particularly the case in the AMIGOS data. In this case, a risk-averse model was more effective than a flexible model.

5.2 Implications

Moreover, our research confirms that generalizing exact valence and arousal levels based on biosignals is a nontrivial task, especially with limited data. The inconsistent findings in related studies suggest that affect recognition performance is dataset and model-dependent, further complicating comparisons and generalizations across studies. Our results indicate that in AMIGOS and PhyMER datasets, the relationship between physiological features and affecive ratings was diffuse. As demonstrated by our ICC analysis, we showed that there was, in fact, greater variability between participants than within participants for most features. While this variability between participants reflects the robustness of these features, it also highlights the inherent generalization problem that we sought to address in this article: the intrinsic variability of people hinders the generalization capabilities of affect recognition models. Such individual differences are manifested in different and complex ways. First, baseline differences and physiological reactivity dictate how people respond differently to the same stimuli. In turn, the cognitive appraisal of these physiological changes differs between individuals (i.e., theory of constructed emotion), and again, how these appraisals are translated into self-reported ratings. Taken together, these individual differences introduce noise within the relationship between features and labels. This noise confounds recognition models unless person-level effects are explicitly accounted for. In other words, the generalized model may instead learn to discern participants rather than affective states.

Our study underlines that estimating exact valence and arousal values is more challenging than treating them as categories. Previous literature has predominantly treated affect recognition as a classification problem; however, some implications should be considered when reducing affective constructs, such as valence and arousal, to 'high' and 'low' categories. For example, our best model using AMIGOS data achieved an F1 score of 0.662 when predicting arousal, whereas our regression model in the same instance explained only 1.4% of the variance. Essentially, in binary classification, granularity is sacrificed for seemingly acceptable performance. Our results suggest that reducing affective states into binary categories is not actually informative. Thus, the results of binary classification of affective states should be considered with nuance, and further work is needed to improve the understanding of which physiological or psychological processes explain self-reported ratings in valence and affect.

Finally, our research demonstrates the importance of transparency in reported metrics for model performance. For example, the relative error (MAE) displayed by our models ranged between 18% and 24.33% across the datasets. These errors themselves do not appear problematic and may even imply the model is relatively precise. However, the models' predictions against the true y values reveal that the model has not learned any meaningful pattern, and has regressed to the mean. In our case, most self-reported ratings fell within the central portion of the entire range. Hence, the relative errors may appear small, but relative to the actual distributions, these errors are roughly equal to the naturally occurring fluctuations. This highlights a well-known problem: metrics should always be interpreted in combination and with supporting visualization. Therefore, other studies that fail to report crucial information necessary to see the whole picture should be treated with caution.

5.3 Limitations and future work

In this section, the limitations and considerations of our research are discussed. In addition, possible directions for future work are suggested.

Firstly, using cluster-informed modeling as a proxy for cluster-based modeling poses a limitation. Including cluster assignment information as features does not force the model to learn only from similar participants, but instead provides the model the *option* to learn group differences, which it might not, especially if other features dominate. Future studies should investigate the effects of true stratified analyses based on personality clusters, particularly in when enough data is available.

The next consideration is the how we operationalized affective constructs. Questions arise on whether we accurately modeled mood in the longer segments. For instance, moods are diffuse, complex underlying affective states that set the overall tone of an individual's experience, which can last over hours, days, or weeks [5], [28]. Whether a 14 to 24-minute segment of affect-inducing stimuli can provide enough context to elicit a certain mood in someone is possible, but not guaranteed. In turn, whether this is effectively captured through the Circumplex model (valence-arousal) may also be overly simplistic. Aside from the limitations of the Circumplex model, retrospectively assigned labels may not uniformly correspond to the whole recording. For example, longer segments contain more physiological and affective fluctuations, which can lead to misalignment between features and labels. One solution internal fluctuation is to segment recordings to improve feature-label correspondence. However, segmenting reduces the context in which the affective process is situated. Not only is this problematic when aiming to infer affective states in real-life applications, but it is particularly so for concepts like mood, where context is indispensable. This poses a complex challenge; while modeling mood effectively requires increased context, increased contexts imply more measurement errors, fluctuation, and feature-label misalignment. Future research should thus focus on modeling long-term affective states, with more advance models or more elaborate ground truth assessments. For example, the Positive and Negative Affect Schedule (PANAS), which is a measurement oriented towards more stable affective states.

This aligns closely with the following limitation: this study did not consider temporal dependencies. Temporal dependencies are particularly informative when dealing with time-series data; the physiological and affective states of one moment are a result of the states that preceded it. Models such as LSTM are developed explicitly for capturing such temporal dependencies by treating the recording data as sequences. LSTM models are capable of learning long-term dependencies within the data, which is especially beneficial when modeling affective experiences over increased duration [5]. Therefore, future research should focus on employing models that account for temporal dynamics, especially when modeling more complex, affective processes of longer duration.

In general, LSTM models are more promising for affect recognition as they are part of the deep learning family. This poses another limitation of the current study; only classical machine learning algorithms were considered. Deep learning models, such as neural networks, can automatically extract complex features and learn hierarchical representations directly from raw signals. Unlike algorithms such as RF and SVR, which rely on manual feature engineering based on domain knowledge, deep learning can capture complex patterns that are far more intricate than those that can be manually extracted. The potency of deep learning is compelling; papers consistently report prediction accuracies from 0.70 to 0.90 when employing complex deep learning models (e.g., [37], [91]).

However, advanced machine learning models require larger datasets to perform reliably. This highlights a key limitation of the present study: this study relied on limited samples of laboratory-collected data, with both datasets containing few participants. This scarcity is due to the inherent limitations of experimental settings, which are time-consuming and costly. Under limited data constraints, the model lacks sufficient representation of specific patterns present in the data. As a result, the model is fitted

too closely to the data at hand, resulting in poor generalization. Future work could experiment with combining datasets that share similar characteristics or collecting increased sample sizes.

Relying on data collected in a laboratory setting limits the generalizability and ecological validity of the findings. Although experimental setups often result in improved data quality compared to naturalistic settings, stimuli may not elicit genuine responses in participants. People typically experience affect as more authentic, layered, and intense in real life [60]. Hence, laboratory studies consistently underrepresent some affective responses, such as the outer ends of the valence-arousal space. Thus, for future research, it would be beneficial to investigate how well affect recognition models extend to data collected in natural settings. For example, by collecting data on affective responses in participants' natural environments through Experience-based Sampling (ESM) or other in-situ data collection methods, which offer a more representative and ecologically valid understanding of how affect unfolds. Powerful models like LSTM can help mitigate the limitations of naturally collected data, such as label-feature misalignment.

Finally, our models likely would have benefited from a feature reduction step, as the low feature-label correlations indicated the presence of feature redundancy. However, due to time constraints, we did not perform this additional step. Although models like RF internally select features by splitting only on features that maximize utility, some form of feature selection could still be beneficial, especially for the SVR models [83]. Furthermore, some studies find EDA to be the most critical biosignal driving performance [90], [92]. For example, a study by Kamasak et al. [92] demonstrated that using only EDA features was sufficient to achieve 70% accuracy in emotion recognition using RF. Future research could benefit from identifying the most influential features for affect recognition models, in particular EDA.

5.4 Conclusion

While affect recognition models based on biosignals face challenges due to individual variability, personalization techniques for user-independent models can help improve generalization to unseen users. In this study, we investigated the potential of using personality-informed clustering as a personalization technique for recognizing continuous valence and arousal states in unseen participants. Our analyses confirmed that fully generalized models indeed perform poorly across participants; however, neither supplementing the models with personality-based cluster assignments nor trait information consistently improved generalization. We found that brief affective experiences were better modeled than longer-duration affective experiences, and predicting arousal generally yielded better results than valence. The algorithms performed differently for each task; RF tended to perform better for arousal, while SVR performed better for valence. Future studies should prioritize using larger datasets, either by combining existing resources or by collecting data in naturalistic settings over extended periods. In addition, research should address the challenges of modeling long-term affective states, such as mood. Advanced models, such as LSTM, are particularly promising because they can capture the temporal dynamics essential for representing longer-duration affective processes. Our study confirms that generalizing exact valence and arousal levels based on biosignals is a nontrivial task. Individual variability diffuses the relationships between physiology and affective ratings, and more research is needed to understanding the psychophysiological processes that underlie these self-reported ratings. Finally, we urge similar research to interpret the predictive performance of classification models with caution, and maintain transparency by reporting all metrics relevant to understanding the performance of affect recognition models.

Code availability

The code for this thesis is made available at my GitHub repository: https://github.com/vegaknak/affect-recognition

Glossary

R^2 Coefficient of Determination. 28, 29, 44–48, 50, 51

ANS Autonomic Nervous System. 8–10, 33

ASP Affective Signal Processing. 9

AUC Area Under the Curve. 28, 29, 49

BFI Big Five Inventory. 6, 11, 13

BVP Blood Volume Pulse. 9

DBI Davies-Boulding Index. 26, 38, 39

DBSCAN Density-Based Spatial Clustering of Applications with Noise. 26, 27

ECG Electrocardiography. 5, 9, 17, 20–24, 30, 50, 52

EDA Electrodermal Activity. 5, 10, 12, 17, 19, 21, 23, 24, 30, 33, 50–52, 55

EEG Electroencephalogram. 5, 50–52

ESM Experience-based Sampling. 13, 55

GSR Galvanic Skin Response. 10

HCI Human-Computer Interaction. 5

HR Heart Rate. 5, 9, 10, 12, 17, 22, 32–34, 36

HRV Heart Rate Variability. 9–11, 17, 22, 23, 25, 32, 33, 36

IBI Inter-Beat-Interval. 9, 10, 22, 33, 34, 36

ICC Intraclass Correlation Coefficient. 25, 35, 36, 50

LOOCV Leave-One-Out Cross-Validation. 27, 43

LSTM Long Short-Term Memory. 1, 51, 54, 55

MAE Mean Absolute Error. 28, 29, 44–46, 48, 49, 51–53

MSE Mean Squared Error. 29

NPA Newcastle Personality Assessor. 11

NS-SCR Non-Specific Skin Conductance Response. 21, 23

PANAS Positive and Negative Affect Schedule. 8, 54

PC Principal Component. 28, 42–44, 50, 51

PCA Principal Component Analysis. 38

PPG Photoplethysmography. 9, 10, 20–25, 30, 33, 51

RF Random Forests. 1, 6, 16, 17, 27, 43–52, 54, 55, 70, 71, 74

RMSE Root Mean Squared Error. 28, 29, 43–46, 48, 49, 52

RMSSD Root Mean Square of Successive Differences. 22, 25, 32, 33

SAM Self-Assessment Manikins. 19

SC Skin Conductance. 5

SCL Skin Conductance Level. 21, 23, 24, 33–35

SCR Skin Conductance Response. 10, 21, 23, 24, 33, 34, 36

SD Standard Deviation. 21, 23, 26, 33, 34, 36, 37, 40, 41

SDNN Standard Deviation of NN intervals. 22, 25, 32–34

SNS Sympathetic Nervous System. 10

SVC Support Vector Classification. 27, 49, 50

SVM Support Vector Machines. 17, 27

SVR Support Vector Regression. 1, 6, 16, 27, 43–48, 50–55, 72, 73

XGBoost Extreme Gradient Boosting. 50

References

[1] R. W. Picard, 'Affective computing for hci,' in *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I - Volume I*, L. Erlbaum Associates Inc., 1999, pp. 829–833, ISBN: 0805833919.

[2] Y. Han, P. Zhang, M. Park and U. Lee, 'Systematic evaluation of personalized deep learning models for affect recognition,' *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–35, 2024. DOI: 10.1145/3699724.

[3] E. L. Van Den Broek, V. Lisỳ, J. H. Janssen, J. H. Westerink, M. H. Schut and K. Tuinenbreijer, 'Affective man-machine interface: Unveiling human emotions through biosignals,' in *Biomedical Engineering Systems and Technologies: International Joint Conference, BIOSTEC 2009 Porto, Portugal, January 14-17, 2009, Revised Selected Papers 2*, Springer, 2010, pp. 21–47.

[4] K. Sharma, C. Castellini, E. L. Van Den Broek, A. Albu-Schaeffer and F. Schwenker, 'A dataset of continuous affect annotations and physiological signals for emotion analysis,' *Scientific data*, vol. 6, no. 1, p. 196, 2019. DOI: 10.1038/s41597-019-0209-7.

[5] C. Katsimerou, J. A. Redi and I. Heynderickx, 'A computational model for mood recognition,' in *User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings 22*, Springer, 2014, pp. 122–133. DOI: 10.1007/978-3-319-08786-3_11.

[6] L. F. Barrett, M. Lewis and J. M. Haviland-Jones, *Handbook of emotions*. Guilford Publications, 2016, ISBN: 978-1-4625-2534-8.

[7] J. A. Russell, 'A circumplex model of affect.,' *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[8] J. T. Cacioppo, L. G. Tassinary and G. Berntson, *Handbook of psychophysiology*. Cambridge University Press, 2007, ISBN: 9781107415782.

[9] W. James, 'The physical basis of emotion,' *Psychological Review*, vol. 101, no. 2, pp. 205–210, 1994. DOI: 10.1037/0033-295X.101.2.205. [Online]. Available: https://doi.org/10.1037/0033-295X.101.2.205.

[10] R. A. Calvo and S. D'Mello, 'Affect detection: An interdisciplinary review of models, methods, and their applications,' *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010. DOI: 10.1109/T-AFFC.2010.1.

[11] L. F. Barrett, 'The theory of constructed emotion: An active inference account of interoception and categorization,' *Social cognitive and affective neuroscience*, vol. 12, no. 1, pp. 1–23, 2017.

[12] J. W. Hughes and C. M. Stoney, 'Depressed mood is related to high-frequency heart rate variability during stressors,' *Psychosomatic medicine*, vol. 62, no. 6, pp. 796–803, 2000.

[13] R. E. Lucas and B. M. Baird, 'Extraversion and emotional reactivity.,' *Journal of personality and social psychology*, vol. 86, no. 3, p. 473, 2004.

[14] D. Marengo, K. L. Davis, G. Ö. Gradwohl and C. Montag, 'A meta-analysis on individual differences in primary emotional systems and big five personality traits,' *Scientific reports*, vol. 11, no. 1, p. 7453, 2021.

[15] J. A. Miranda-Correa, M. K. Abadi, N. Sebe and I. Patras, 'Amigos: A dataset for affect, personality and mood research on individuals and groups,' *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 479–493, 2018. DOI: 10.1109/TAFFC.2018.2884461.

[16] H. J. Eysenck, 'Dimensions of personality: The biosocial approach to personality,' in *Explorations in temperament: International perspectives on theory and measurement*, Springer, 1991, pp. 87–103.

[17] M. A. Klados, P. Konstantinidi, R. Dacosta-Aguayo, V.-D. Kostaridou, A. Vinciarelli and M. Zervakis, 'Automatic recognition of personality profiles using eeg functional connectivity during emotional processing,' *Brain sciences*, vol. 10, no. 5, p. 278, 2020. DOI: 10.3390/brainsci10050278.

[18] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva and C. Ersoy, 'Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches,' *IEEE Access*, vol. 8, pp. 38 146–38 163, 2020.

[19] J. Tervonen, S. Puttonen, M. J. Sillanpää *et al.*, 'Personalized mental stress detection with self-organizing map: From laboratory to the field,' *Computers in Biology and Medicine*, vol. 124, p. 103 935, 2020.

[20] S. Taylor, N. Jaques, E. Nosakhare, A. Sano and R. Picard, 'Personalized multitask learning for predicting tomorrow's mood, stress, and health,' *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 200–213, 2017. DOI: 10.1109/TAFFC.2017.2784832.

[21] Z. Yusefi Hafshejani, M. Kaedi and A. Fatemi, 'Improving sparsity and new user problems in collaborative filtering by clustering the personality factors,' *Electronic Commerce Research*, vol. 18, pp. 813–836, 2018.

[22] D. A. Adler, F. Wang, D. C. Mohr and T. Choudhury, 'Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies,' *Plos one*, vol. 17, no. 4, e0266516, 2022.

[23] S. Kang, W. Choi, C. Y. Park *et al.*, 'K-emophone: A mobile and wearable dataset with in-situ emotion, stress, and attention labels,' *Scientific data*, vol. 10, no. 1, p. 351, 2023. DOI: 10.1038/s41597-023-02102-2.

[24] L. Berkemeier, W. Kamphuis, A.-M. Brouwer *et al.*, 'Measuring affective state: Subject-dependent and -independent prediction based on longitudinal multimodal sensing,' *IEEE Transactions on Affective Computing*, vol. 16, no. 2, pp. 827–843, 2025. DOI: 10.1109/TAFFC.2024.3474098.

[25] O. P. John, E. M. Donahue and R. L. Kentle, *Big five inventory (bfi)*, APA PsycTests, 1991. DOI: 10.1037/t07550-000.

[26] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler and N. Sebe, 'Ascertain: Emotion and personality recognition using commercial sensors,' *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016. DOI: 10.1109/TAFFC.2016.2625250.

[27] S. Pant, H.-J. Yang, E. Lim, S.-H. Kim and S.-B. Yoo, 'Phymer: Physiological dataset for multimodal emotion recognition with personality as a context,' *IEEE Access*, vol. 11, pp. 107 638–107 656, 2023. DOI: 10.1109/ACCESS.2023.3320053.

[28] R. LiKamWa, Y. Liu, N. D. Lane and L. Zhong, 'Moodscope: Building a mood sensor from smartphone usage patterns,' in *Proceedings of the 11th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys '13)*, Taipei, Taiwan: Association for Computing Machinery, 2013, pp. 389–402, ISBN: 978-1-4503-1672-9. DOI: 10.1145/2462456.2464449. [Online]. Available: https://doi.org/10.1145/2462456.2464449.

[29] S. Gongora Alonso, S. Hamrioui, I. de la Torre Diez, E. Motta Cruz, M. Lopez-Coronado and M. Franco, 'Social robots for people with aging and dementia: A systematic review of literature,' *Telemedicine and e-Health*, vol. 25, no. 7, pp. 533–540, 2019.

[30] E. L. van den Broek, 'Affective signal processing (asp): Unraveling the mystery of emotions,' Ph.D. dissertation, University of Twente, Enschede, The Netherlands, 2011, ISBN: 978-90-365-3243-3. DOI: 10.3990/1.9789036532433.

[31] J. A. Russell, 'Core affect and the psychological construction of emotion.,' *Psychological review*, vol. 110, no. 1, p. 145, 2003. DOI: 10.1037/0022-3514.39.6.1161.

[32] M. Power and T. Dalgleish, *Cognition and emotion: From order to disorder*. Psychology press, 2015, ISBN: 978-1-84872-067-5.

[33] J. C. Borod, *The neuropsychology of emotion*. Oxford University Press, 2000, ISBN: 0-19-511464-7.

[34] P. Ekman, 'Facial expressions of emotion: An old controversy and new findings,' *Philosophical transactions of the royal society of London. Series B: Biological Sciences*, vol. 335, no. 1273, pp. 63–69, 1992.

[35] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras and N. Sebe, 'Decaf: Meg-based multimodal database for decoding affective physiological responses,' *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.

[36] D. Watson, L. A. Clark and A. Tellegen, 'Development and validation of brief measures of positive and negative affect: The panas scales.,' *Journal of personality and social psychology*, vol. 54, no. 6, pp. 1063–1070, 1988. DOI: 10.1037/0022-3514.54.6.1063.

[37] F. Galvão, S. M. Alarcão and M. J. Fonseca, 'Predicting exact valence and arousal values from eeg,' *Sensors*, vol. 21, no. 10, p. 3414, 2021.

[38] M. Soleymani, F. Villaro-Dixon, T. Pun and G. Chanel, 'Toolbox for emotional feature extraction from physiological signals (teap),' *Frontiers in ICT*, vol. 4, p. 1, 2017.

[39] L. G. Tassinary and J. T. Cacioppo, *Unobservable facial actions and emotion*, 1992.

[40] P. Rainville, A. Bechara, N. Naqvi and A. R. Damasio, 'Basic emotions are associated with distinct patterns of cardiorespiratory activity,' *International journal of psychophysiology*, vol. 61, no. 1, pp. 5–18, 2006.

[41] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana and A. A. Aziz, 'Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review,' *Sensors*, vol. 21, no. 15, p. 5015, 2021.

[42] P. J. Bota, C. Wang, A. L. Fred and H. P. Da Silva, 'A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals,' *IEEE access*, vol. 7, pp. 140 990–141 020, 2019.

[43] J. Allen, 'Photoplethysmography and its application in clinical physiological measurement,' *Physiological measurement*, vol. 28, no. 3, R1, 2007.

[44] H. D. Critchley, 'Electrodermal responses: What happens in the brain,' *The Neuroscientist*, vol. 8, no. 2, pp. 132–142, 2002.

[45] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012, ISBN: 978-1-4614-1125-3.

[46] M. Harker, 'Psychological sweating: A systematic review focused on aetiology and cutaneous response,' *Skin pharmacology and physiology*, vol. 26, no. 2, pp. 92–100, 2013.

[47] M. Zuckerman, *Psychobiology of personality*. Cambridge University Press, 1991, vol. 10, ISBN: 0-521-35095-6.

[48] E. C. Tupes and R. E. Christal, 'Recurrent personality factors based on trait ratings,' *Journal of personality*, vol. 60, no. 2, pp. 225–251, 1992.

[49] R. R. McCrae and P. T. Costa Jr, 'Personality trait structure as a human universal.,' *American psychologist*, vol. 52, no. 5, p. 509, 1997.

[50] R. R. McCrae and O. P. John, 'An introduction to the five-factor model and its applications,' *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[51] P. T. Costa Jr and R. R. McCrae, *Neo Personality Inventory*. American Psychological Association, 2000.

[52] D. Nettle, *Personality: What makes you the way you are*. Oxford University Press, 2007, ISBN: 978-0-19-921142-5.

[53] J. J. Gross, S. K. Sutton and T. Ketelaar, 'Relations between affect and personality: Support for the affect-level and affective-reactivity views,' *Personality and social psychology bulletin*, vol. 24, no. 3, pp. 279–288, 1998.

[54] A. Kargarandehkordi, M. Kaisti and P. Washington, 'Personalization of affective models using classical machine learning: A feasibility study,' *Applied Sciences*, vol. 14, no. 4, p. 1337, 2024.

[55] R. Hu and P. Pu, 'Using personality information in collaborative filtering for new users,' *Recommender Systems and the Social Web*, vol. 17, pp. 60–70, 2010.

[56] W. Bleidorn, T. Schwaba, A. Zheng *et al.*, 'Personality stability and change: A meta-analysis of longitudinal studies.,' *Psychological bulletin*, vol. 148, no. 7-8, p. 588, 2022.

[57] G. McKeown, M. Valstar, R. Cowie, M. Pantic and M. Schroder, 'The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent,' *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2011. DOI: 10.1109/T-AFFC.2011.20.

[58] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn and R. Picard, 'Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected,' in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 881–888. DOI: 10.1109/CVPRW.2013.130.

[59] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger and K. Van Laerhoven, 'Introducing wesad, a multimodal dataset for wearable stress and affect detection,' in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408. DOI: 10.1145/3242969.3242985.

[60] X. Shui, M. Zhang, Z. Li, X. Hu, F. Wang and D. Zhang, 'A dataset of daily ambulatory psychological and physiological recording for emotion research,' *Scientific data*, vol. 8, no. 1, p. 161, 2021. DOI: 10.1038/s41597-021-00902-1.

[61] S. Koelstra, C. Muhl, M. Soleymani *et al.*, 'Deap: A database for emotion analysis; using physiological signals,' *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[62] K. Kutt, D. Drążyk, L. Żuchowska, M. Szelążek, S. Bobek and G. J. Nalepa, 'Biraffe2, a multimodal dataset for emotion-based personalization in rich affective game environments,' *Scientific data*, vol. 9, no. 1, p. 274, 2022. DOI: 10.1038/s41597-022-01363-1.

[63] F. Shaffer and J. P. Ginsberg, 'An overview of heart rate variability metrics and norms,' *Frontiers in Public Health*, vol. 5, p. 258, 2017.

[64] M. E. Strauss and G. T. Smith, 'Construct validity: Advances in theory and methodology,' *Annual Review of Clinical Psychology*, vol. 5, no. 1, pp. 1–25, 2009.

[65] F. Tian, X. Wang, W. Cheng, M. Lee and Y. Jin, 'A comparative study on the temporal effects of 2d and vr emotional arousal,' *Sensors*, vol. 22, no. 21, p. 8491, 2022.

[66] D. Makowski, T. Pham, Z. J. Lau *et al.*, 'Neurokit2: A python toolbox for neurophysiological signal processing,' *Behavior research methods*, pp. 1–8, 2021.

[67] S. K. Jagtap and M. Uplane, 'The impact of digital filtering to ecg analysis: Butterworth filter application,' in *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*, IEEE, 2012, pp. 1–6.

[68] P. J. Lang, M. K. Greenwald, M. M. Bradley and A. O. Hamm, 'Looking at pictures: Affective, facial, visceral, and behavioral reactions,' *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.

[69] S. Tivatansakul and M. Ohkura, 'Emotion recognition using ecg signals with local pattern description methods,' *International Journal of Affective Engineering*, vol. 15, no. 2, pp. 51–61, 2016.

[70] Á. Solé Morillo, J. Lambert Cause, V.-E. Baciu, B. da Silva, J. C. Garcia-Naranjo and J. Stiens, 'Ppg edukit: An adjustable photoplethysmography evaluation system for educational activities,' *Sensors*, vol. 22, no. 4, p. 1389, 2022.

[71] S. G. Patro and D.-K. K. Sahu, 'Normalization: A preprocessing stage,' *IARJSET*, Mar. 2015. DOI: 10.17148/IARJSET.2015.2305.

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

[73] R. R. Singh, S. Conjeti and R. Banerjee, 'A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals,' *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 740–754, 2013.

[74] B. S. Tegegne, T. Man, A. M. van Roon, H. Snieder and H. Riese, 'Reference values of heart rate variability from 10-second resting electrocardiograms: The lifelines cohort study,' *European journal of preventive cardiology*, vol. 27, no. 19, pp. 2191–2194, 2020.

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

[76] P. Virtanen, R. Gommers, T. E. Oliphant *et al.*, 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,' *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[77] J. P. Weir, 'Quantifying test-retest reliability using the intraclass correlation coefficient and the sem,' *The Journal of Strength & Conditioning Research*, vol. 19, no. 1, pp. 231–240, 2005.

[78] J. Han, J. Pei and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022, ISBN: 978-0-443-19768-1.

[79] K. P. Sinaga and M.-S. Yang, 'Unsupervised k-means clustering algorithm,' *IEEE access*, vol. 8, pp. 80 716–80 727, 2020.

[80] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, 'A density-based algorithm for discovering clusters in large spatial databases with noise,' in *kdd*, vol. 96, 1996, pp. 226–231.

[81] B. Li and A. Sano, 'Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress,' *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–26, 2020.

[82] D. L. Davies and D. W. Bouldin, 'A cluster separation measure,' *Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 224–227, 1979.

[83] L. Breiman, 'Random forests,' *Machine learning*, vol. 45, pp. 5–32, 2001.

[84] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua and A. Lopez, 'A comprehensive survey on support vector machine classification: Applications, challenges and trends,' *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[85] A. Botchkarev, 'A new typology design of performance metrics to measure errors in machine learning regression algorithms,' *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, pp. 045–076, 2019, ISSN: 1555-1237. DOI: 10.28945/4184.

[86] L. Plonsky and H. Ghanbar, 'Multiple regression in l2 research: A methodological synthesis and guide to interpreting r2 values,' *The Modern Language Journal*, vol. 102, no. 4, pp. 713–731, 2018.

[87] T. Fawcett, 'An introduction to roc analysis,' *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[88] J. Sacha, 'Interaction between heart rate and heart rate variability,' *Annals of Noninvasive Electrocardiology*, vol. 19, no. 3, pp. 207–216, 2014. DOI: 10.1111/anec.12148.

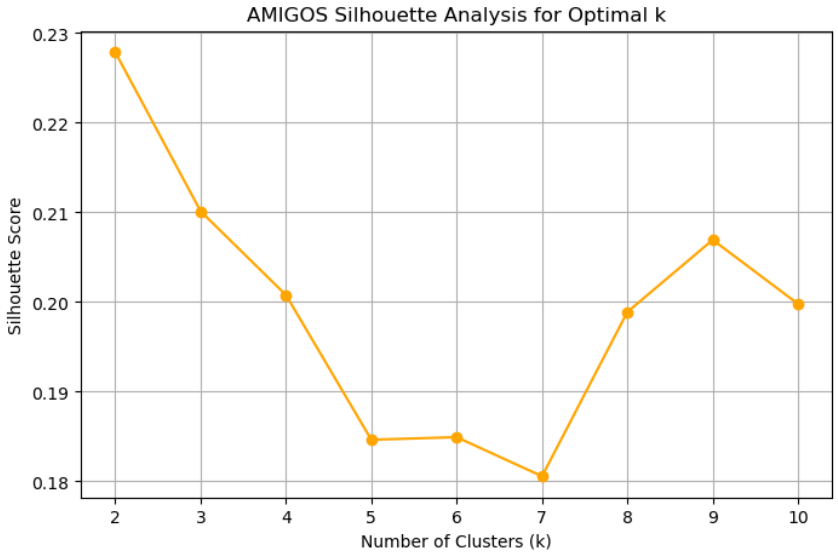
[89] P. Siirtola, S. Tamminen, G. Chandra, A. Ihalapathirana and J. Röning, 'Predicting emotion with biosignals: A comparison of classification and regression models for estimating valence and arousal level using wearable sensors,' *Sensors*, vol. 23, no. 3, p. 1598, 2023.

[90] L. A. Martínez-Tejada, Y. Maruyama, N. Yoshimura and Y. Koike, 'Analysis of personality and eeg features in emotion recognition using machine learning techniques to classify arousal and valence labels,' *Machine Learning and Knowledge Extraction*, vol. 2, no. 2, p. 7, 2020.

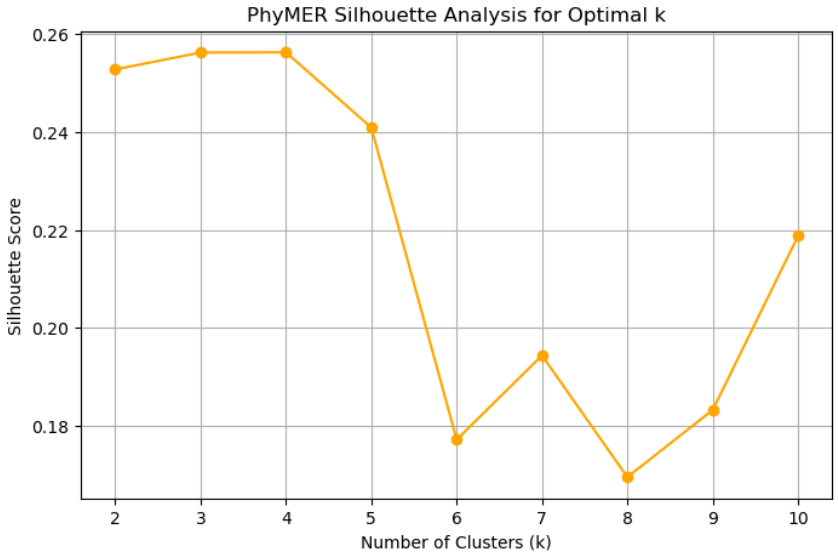
[91] S. Zhao, G. Ding, J. Han and Y. Gao, 'Personality-aware personalized emotion recognition from physiological signals.,' in *IJCAI*, 2018, pp. 1660–1667.

[92] D. Ayata, Y. Yaslan and M. Kamaşak, 'Emotion recognition via random forest and galvanic skin response: Comparison of time-based feature sets, window sizes, and wavelet approaches,' in *2016 Medical Technologies National Congress*, IEEE, 2016, pp. 1–4.

A Personality based clustering



(a) Silhouette score plot for AMIGOS K-means clustering.



(b) Silhouette score plot for PhyMER K-means clustering.

Figure 19: Silhouette analysis for optimal cluster number (k) in K-means clustering of scaled personality data in AMIGOS and PhyMER datasets. The silhouette score is plotted for different values of k to determine the optimal number of clusters. Higher scores indicate more cohesive and well-separated groupings. AMIGOS (a) shows a peak at $k = 2$, while PhyMER (b) peaks at $k = 3$. Across datasets, maximum silhouette scores remain under 0.26, suggesting weak cluster quality.

B Predicted vs Actual y values

B.1 AMIGOS Random Forests

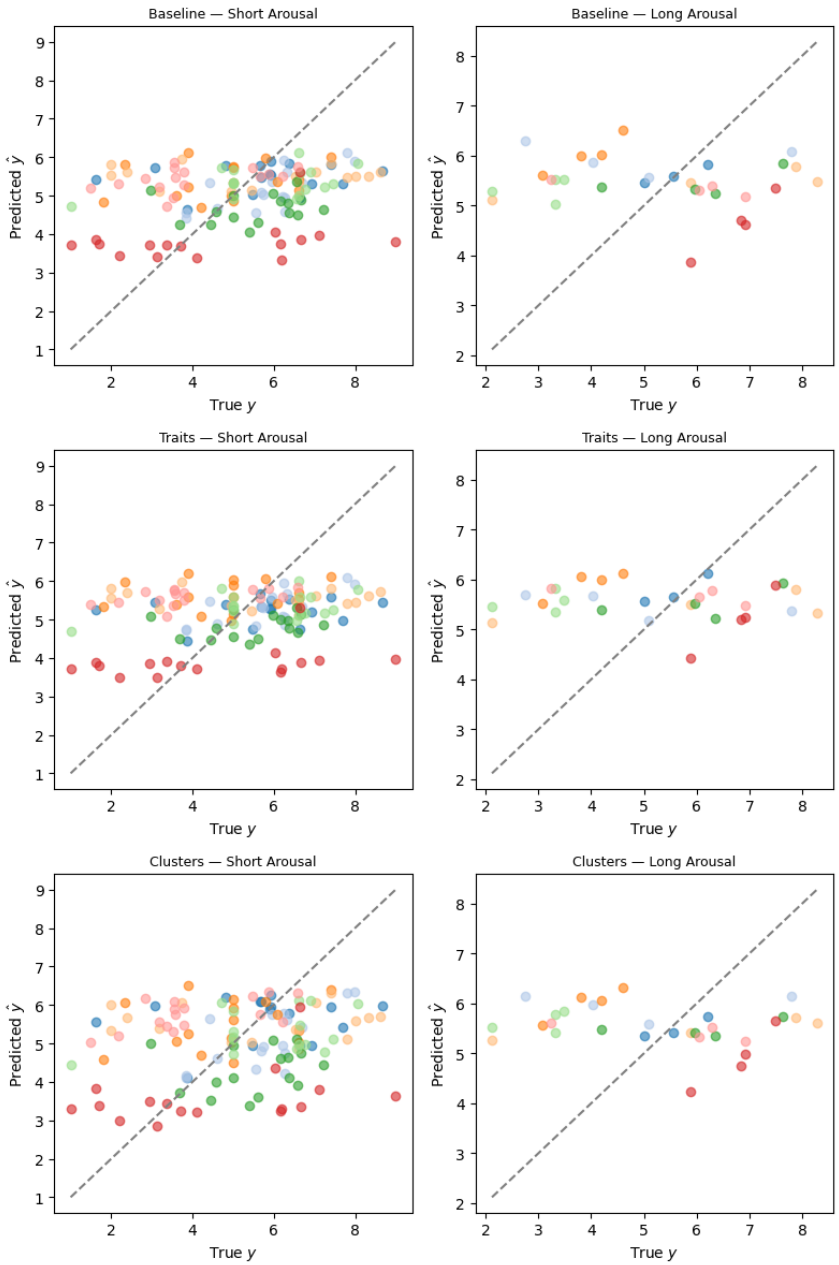


Figure 20: Random Forest (RF) predicted against actual ratings for arousal for AMIGOS Depiction of the predicted y values against actual y arousal ratings (1-9) for short (left) and long (right) videos of the AMIGOS data, color-coded by Participant ID. Some participant effects are observed; the model appears to adjust better to certain participants than to others. Minimal differences can be observed between the variants (Baseline, Traits, Clusters). The dotted diagonal represents perfect predictions. No discernible trend is observed, indicating that the model has learned a minimal pattern in the arousal scores of unseen participants.

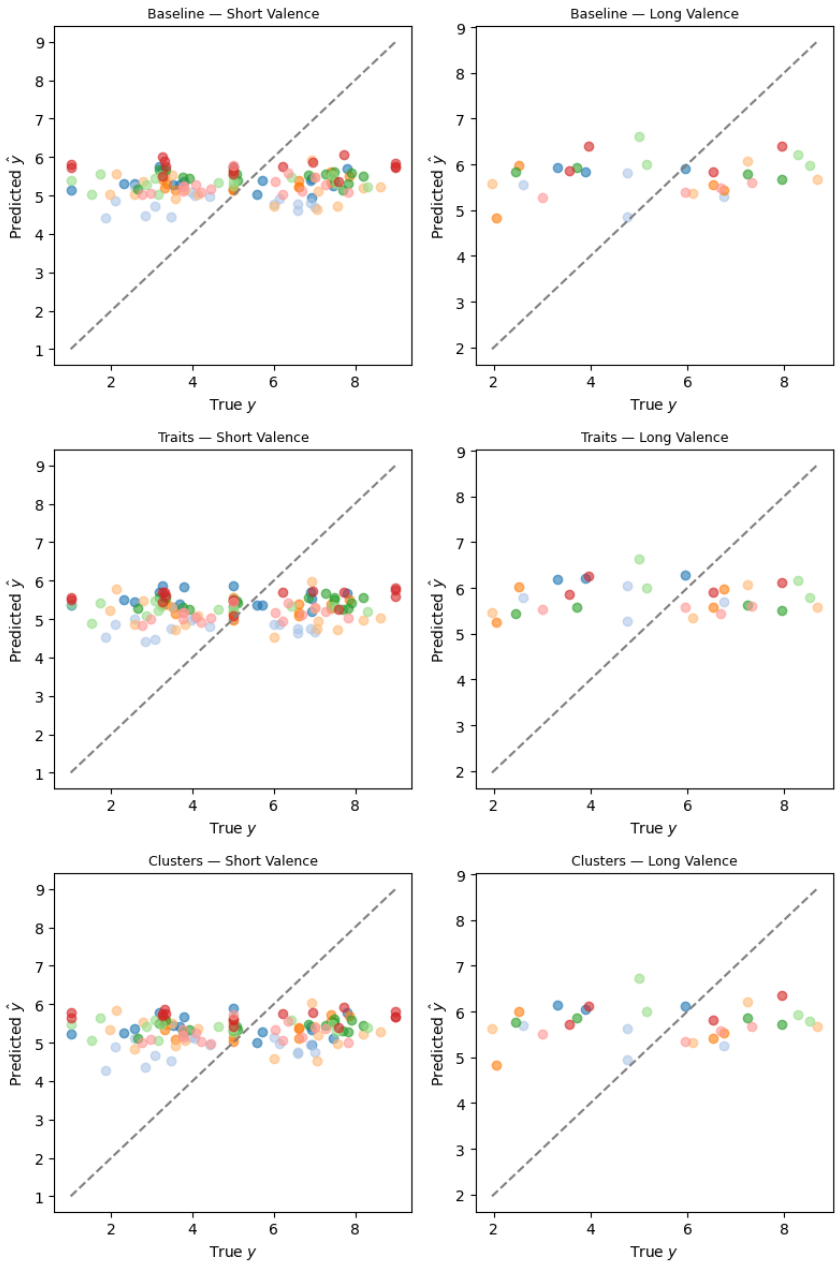


Figure 21: Random Forest (RF) predicted against actual ratings for valence for AMIGOS Depiction of the predicted y values against actual y valence ratings (1-9) for short (left) and long (right) videos of the AMIGOS data, color-coded by Participant ID. Minimal differences are observed between the variants (Baseline, Traits, Clusters). The dotted diagonal represents perfect predictions. No discernible trend is observed, indicating that the model has learned a minimal pattern in the valence scores of unseen participants. Compared to arousal, predictions are even more compressed in the middle range of the scale, indicating the model struggled more with explaining variance in valence than arousal. Slight participant effects are observed; model appears to better adjust to certain participants more than others.

B.2 AMIGOS Support Vector Regression

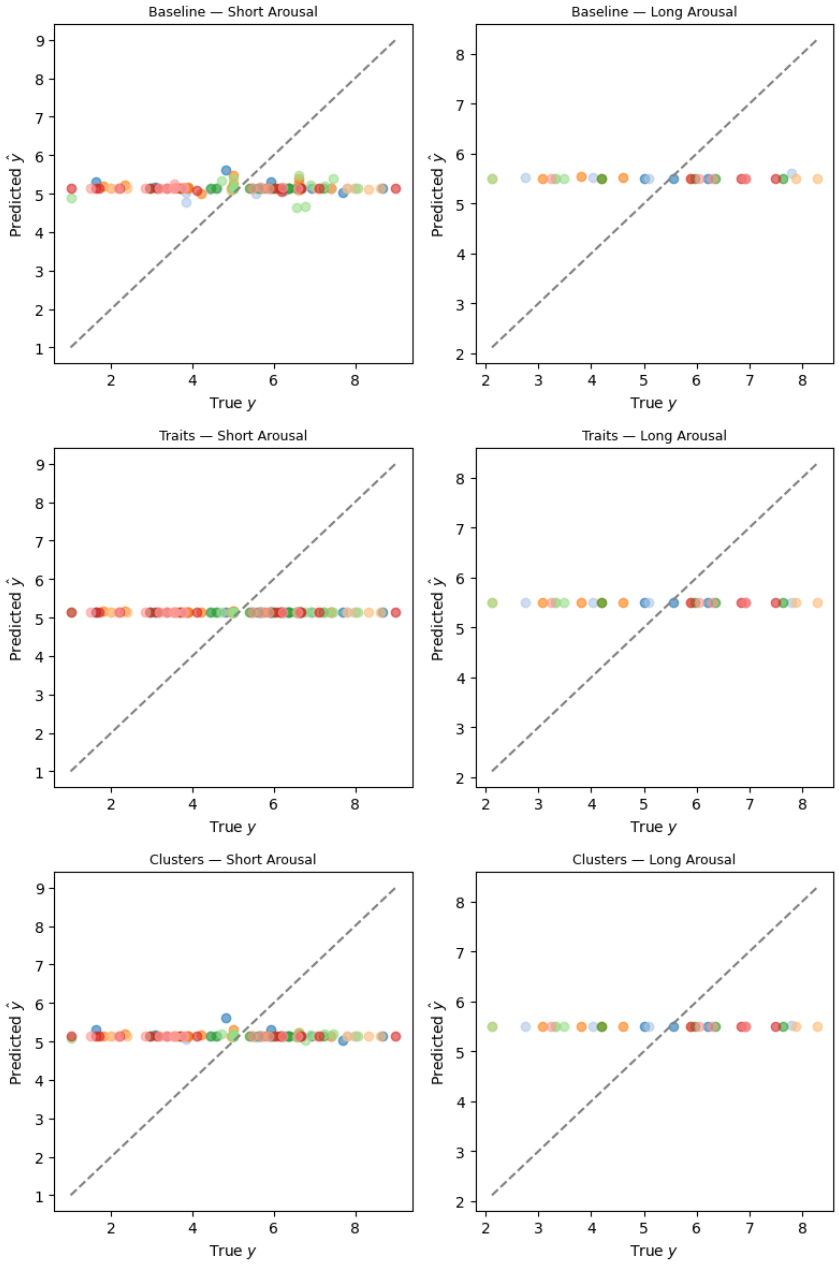


Figure 22: Support Vector Regression (SVR) predicted against actual ratings for arousal in AMIGOS Depiction of the predicted y values against actual y valence ratings (1-9) for short (left) and long (right) videos of the AMIGOS data, color-coded by Participant ID. Minimal differences are observed between the variants (Baseline, Traits, Clusters). The dotted diagonal represents perfect predictions. The model almost exclusively predicted the mean rating, indicating that a minimal pattern was learned in the valence scores of unseen participants. In comparison to Random Forests (RF) for arousal (see figure **??**), it becomes clear that the Support Vector Regression (SVR) algorithm struggled more with explaining variance than RF, as it solely predicted the mean.

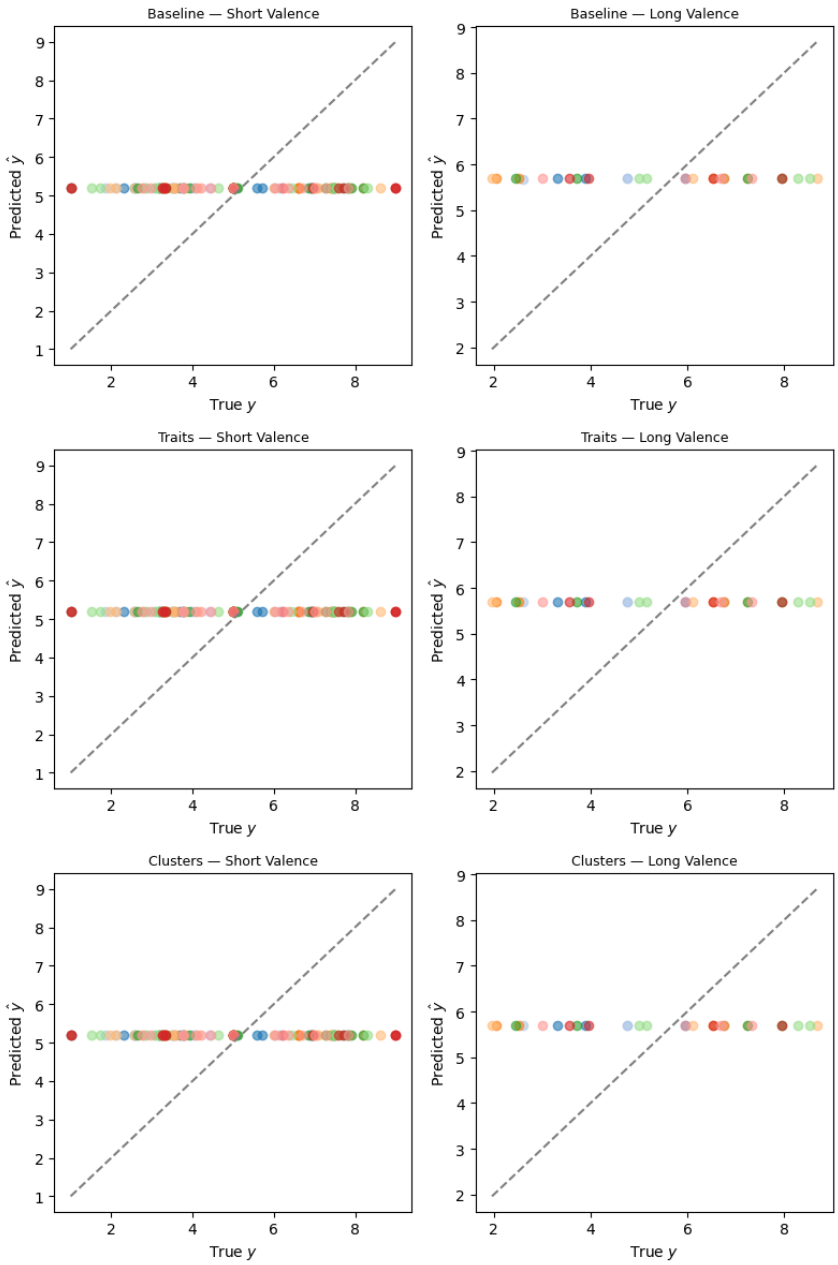


Figure 23: Support Vector Regression predicted against actual ratings for valence in AMIGOS Depiction of the predicted y values against actual y valence ratings (1-9) for short (left) and long (right) videos of the AMIGOS data, color-coded by Participant ID. Minimal differences are observed between the variants (Baseline, Traits, Clusters). The dotted diagonal represents perfect predictions. The model almost exclusively predicted the mean rating, indicating that a minimal pattern was learned in the valence scores of unseen participants. In comparison to Random Forests (RF) for valence (see figure 21), it becomes clear that the Support Vector Regression (SVR) algorithm struggled more with explaining variance than RF, as it solely predicted the mean.

B.3 PhyMER Random forests

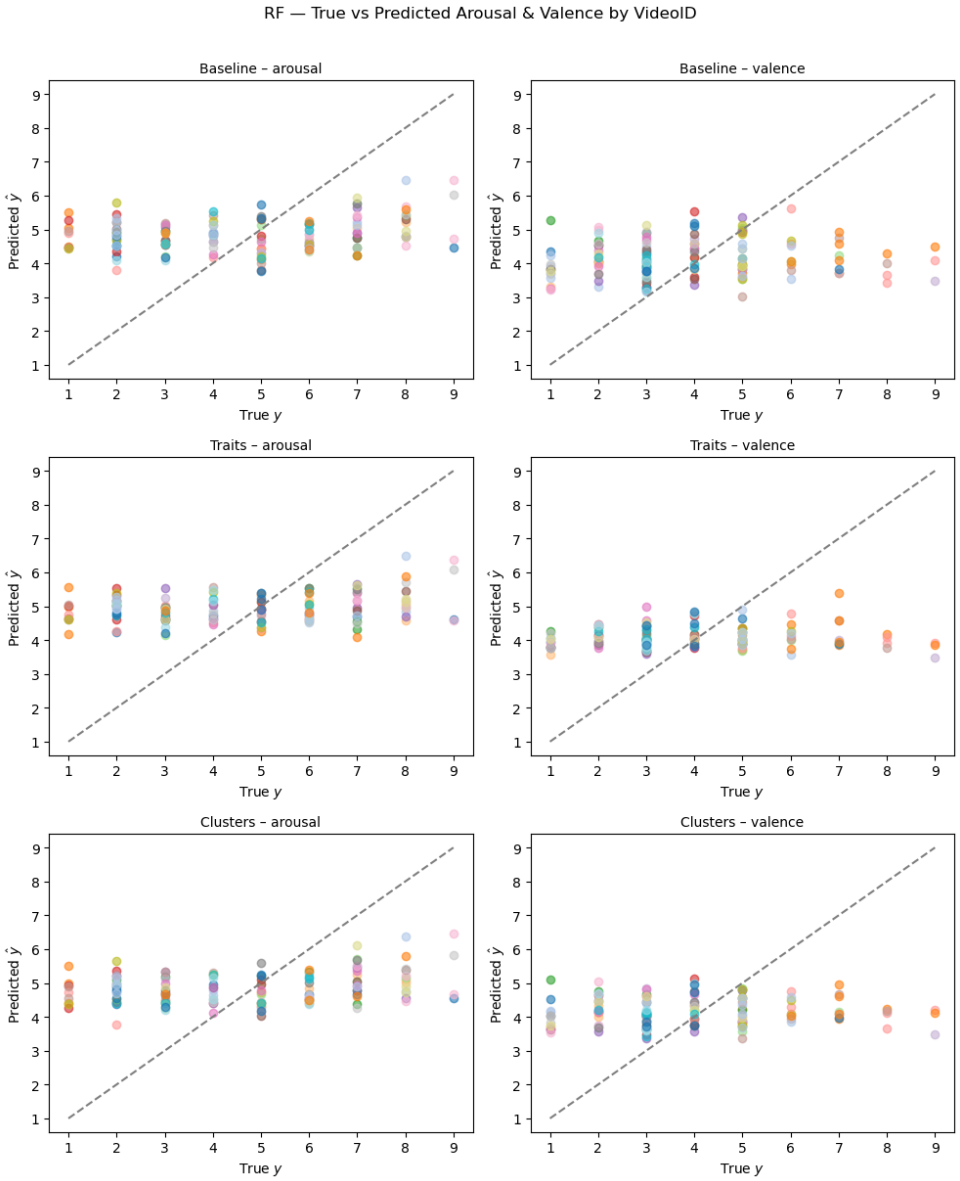


Figure 24: Random Forest (RF) predicted against actual ratings for arousal and valence in PhyMER Depiction of the predicted y values against actual y valence ratings (1-9) for arousal (left) and valence (right) videos of the PhyMER data, color-coded by Participant ID. Minimal differences are observed between the variants (Baseline, Traits, Clusters). The dotted diagonal represents perfect predictions. The model stayed close to predicting the mean. No clear positive slope is observed, indicating that a minimal pattern was learned in the valence scores of unseen participants. Note that the valence and arousal ratings contained no floats and were integers.

B.4 PhyMER: Support Vector Regression

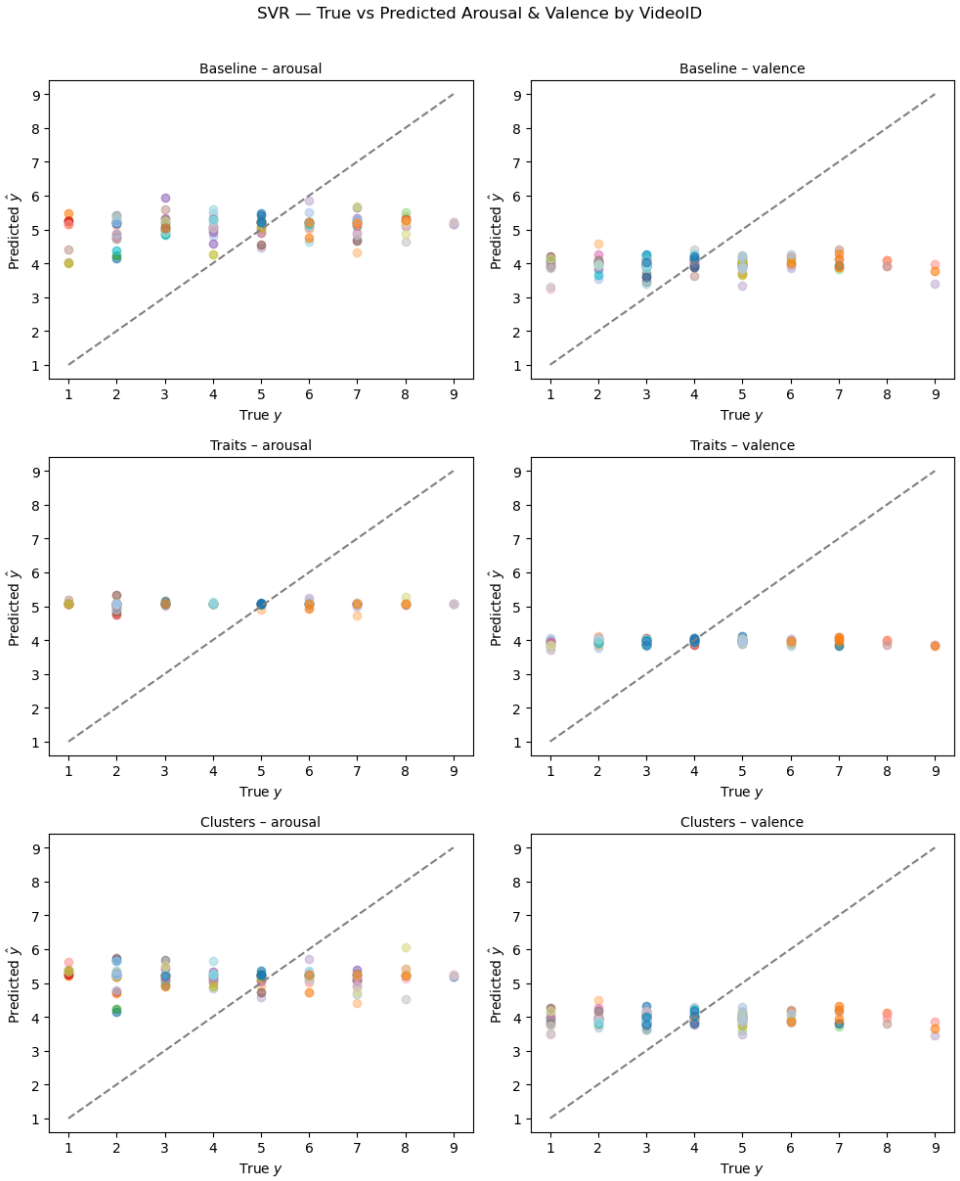


Figure 25: Support Vector Regression (SVR) predicted against actual ratings for arousal and valence in PhyMER Depiction of the predicted y values against actual y valence ratings (1-9) for arousal (left) and valence (right) of the PhyMER data, color-coded by Participant ID. Minimal differences are observed between the variants (Baseline, Traits, Clusters). The dotted diagonal represents perfect predictions. Compared to the Random Forests (RF) algorithm, predictions are further compressed around the mean, indicating that Support Vector Regression tended to explain less variance in valence and arousal than RF. Note that the valence and arousal ratings contained no floats and were integers.

C Residuals-vs-fitted plots

C.1 AMIGOS Random Forests

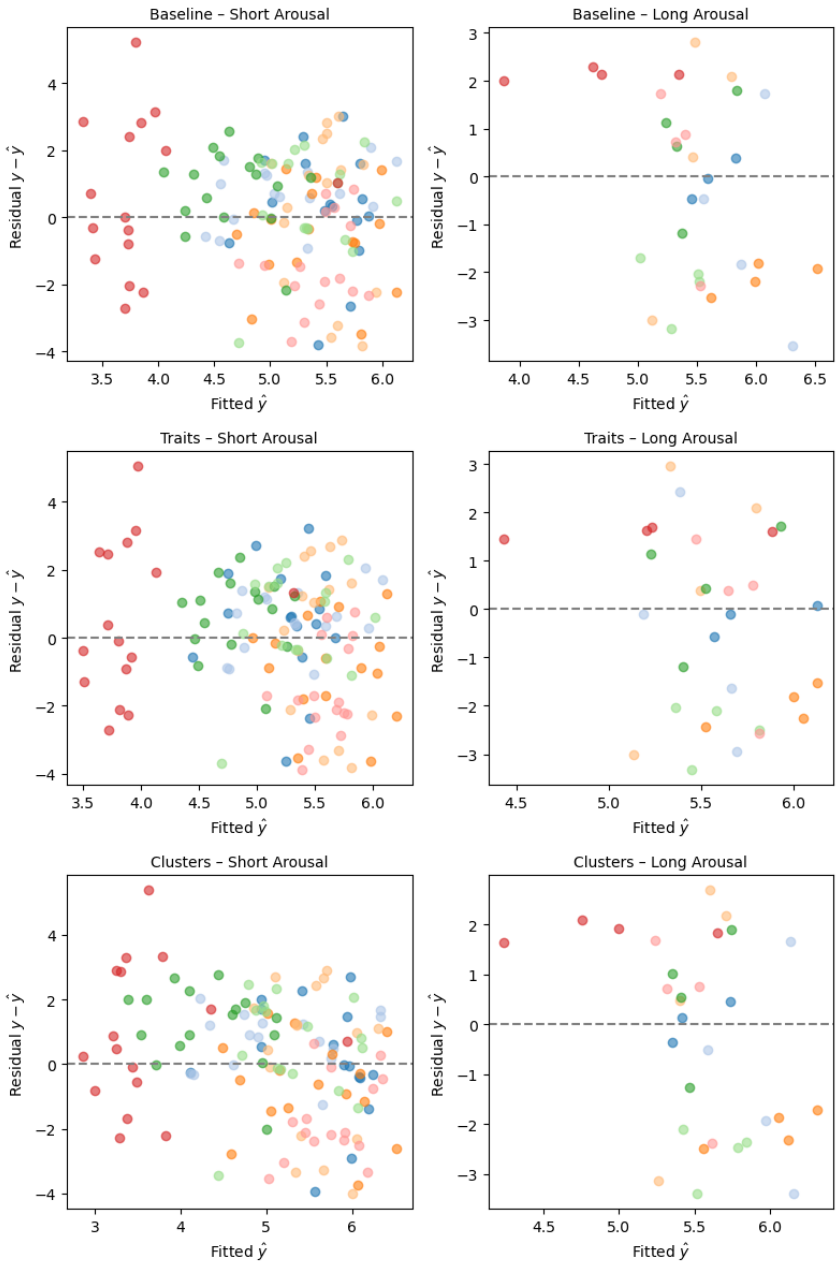


Figure 26: Residuals versus Fitted values, Arousal, RF in AMIGOS Plots of the residuals against fitted values for the RF model for arousal prediction from AMIGOS data for both short (left) and long (right) recordings, color coded by participant ID. Different plots represent different variants of the models: Baseline (top), Traits (middle), Clusters (bottom). Residuals appear roughly normally distributed in the short recordings. The long recordings exhibit a subtle inverted s-pattern, indicating a systematic underestimation of low and high arousal ratings, while overestimating moderate ones. Person-specific effects become apparent.

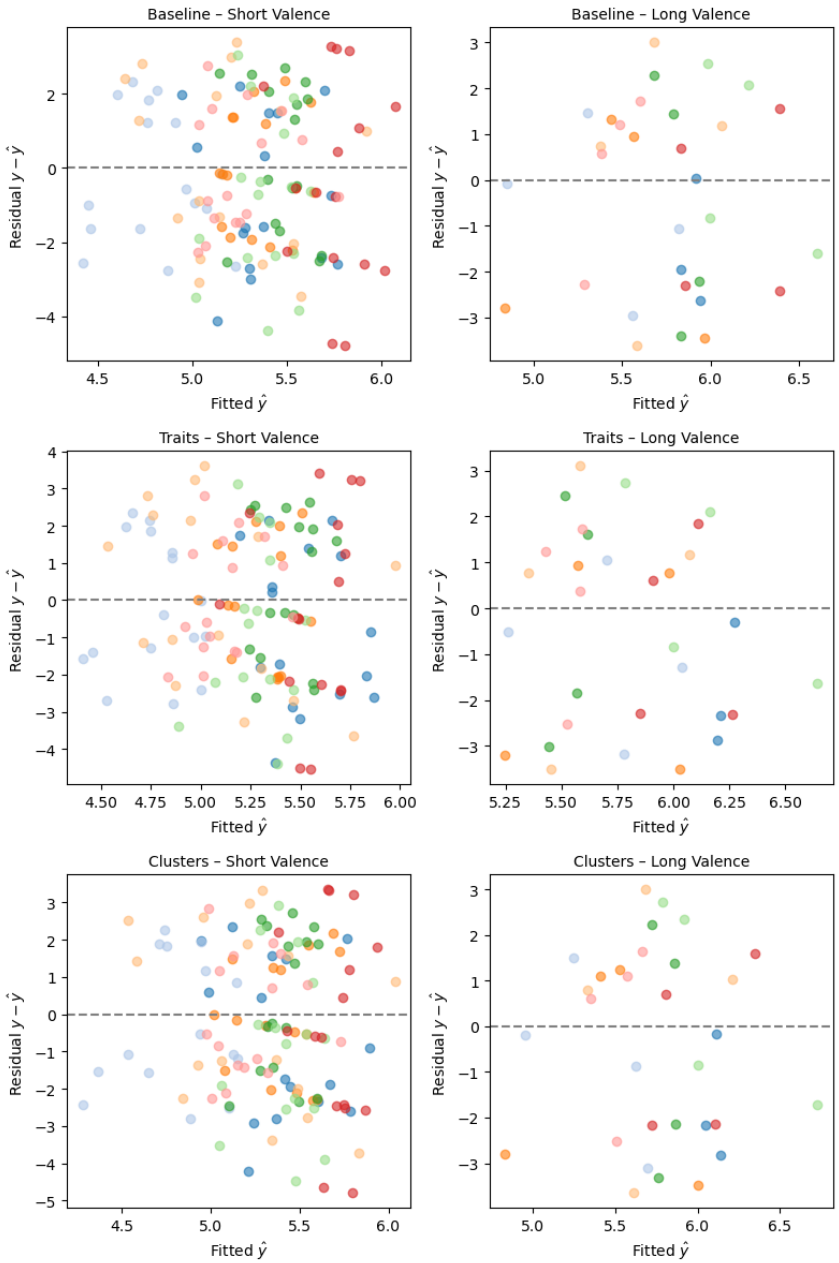


Figure 27: Residuals versus Fitted values, Valence, RF in AMIGOS Plots of the residuals against fitted values for the RF model for valence prediction from AMIGOS data for both short and long recordings, color coded by participant ID. Different plots represent different variants of the models: Baseline (top), Traits (middle), Clusters (bottom). Residuals appear roughly normally distributed, though they show compression around the mean. No clear differences in model variant are apparent.

C.2 AMIGOS Support Vector Regression

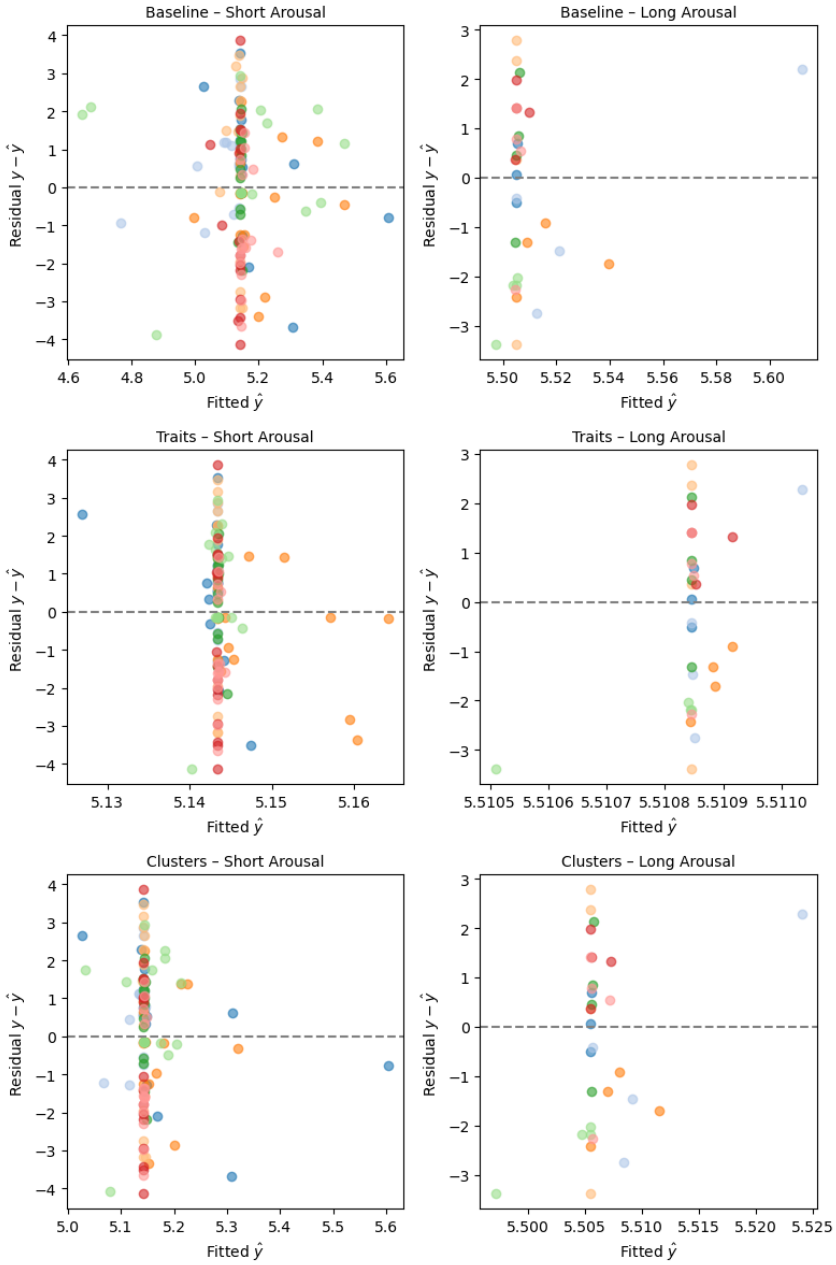


Figure 28: Residuals versus Fitted values, Arousal, SVR in AMIGOS Plots of the residuals against fitted values for the SVR model for arousal prediction from AMIGOS data for both short (left) and long (right) recordings, color coded by participant ID. Different plots represent different variants of the models: Baseline (top), Traits (middle), Clusters (bottom). Although the residuals show a great deal of spread (y-axis), the fitted y values were extremely compressed. This shows the model did not learn any pattern at all.

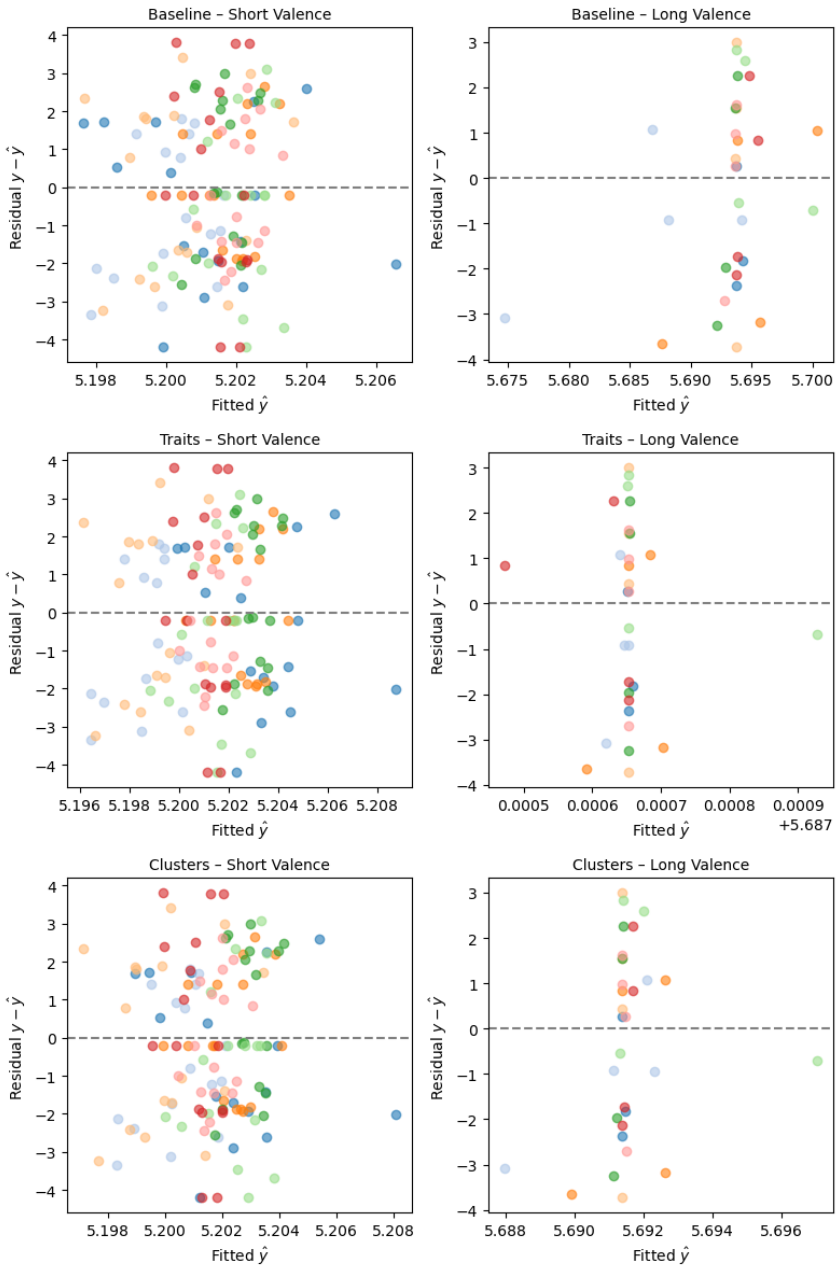


Figure 29: Residuals versus Fitted values, Valence, SVR in AMIGOS Plots of the residuals against fitted values for the SVR model for arousal prediction from AMIGOS data for both short (left) and long (right) recordings, color coded by participant ID. Different plots represent different variants of the models: Baseline (top), Traits (middle), Clusters (bottom). The fitted y-values were extremely compressed in every subplot, indicating that the model did not learn any pattern at all. The compression was more extreme in the long valence videos.

C.3 PhyMER Random Forests

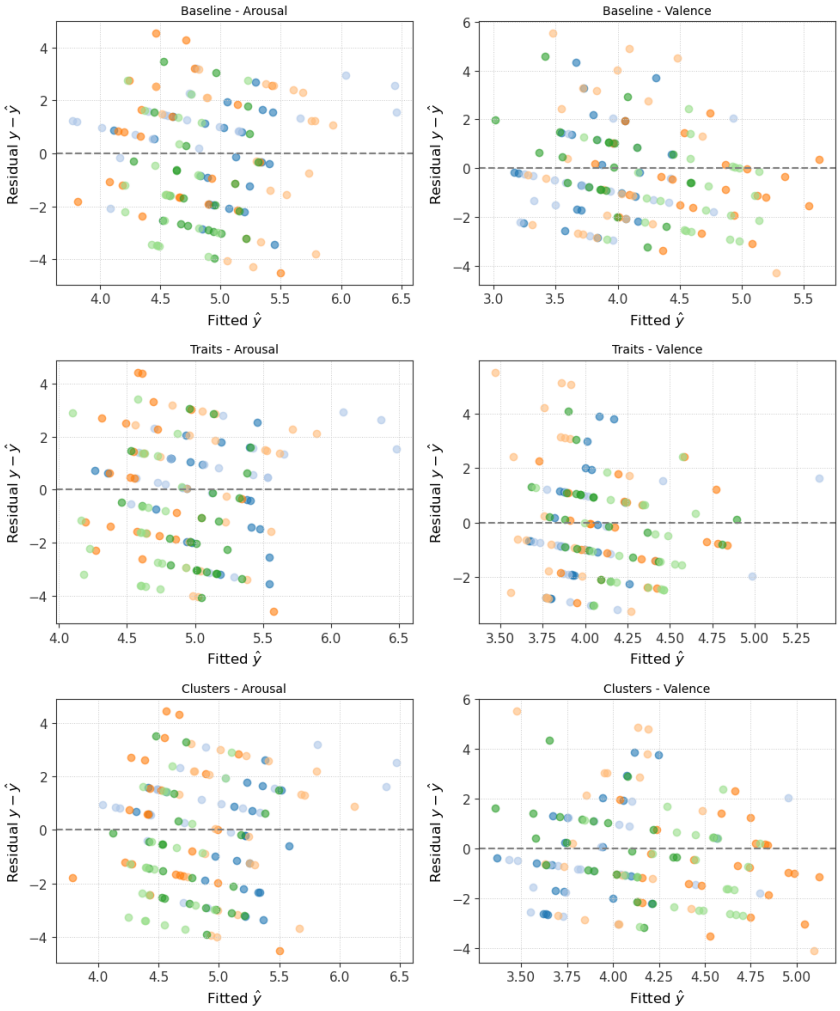


Figure 30: Residuals versus Fitted values RF

Figure 31: Residuals versus Fitted values RF in PhyMER Residuals versus fitted values for RF models across Baseline (top), Traits (middle), and Clusters (bottom) variants for arousal (left) and valence (right) prediction. Each dot represents an individual trial, colored by participant. Residuals are symmetrically distributed. Again, the range is but exhibits wide dispersion, indicating limited predictive accuracy and potential group-specific biases.

C.4 PhyMER Support Vector Regression

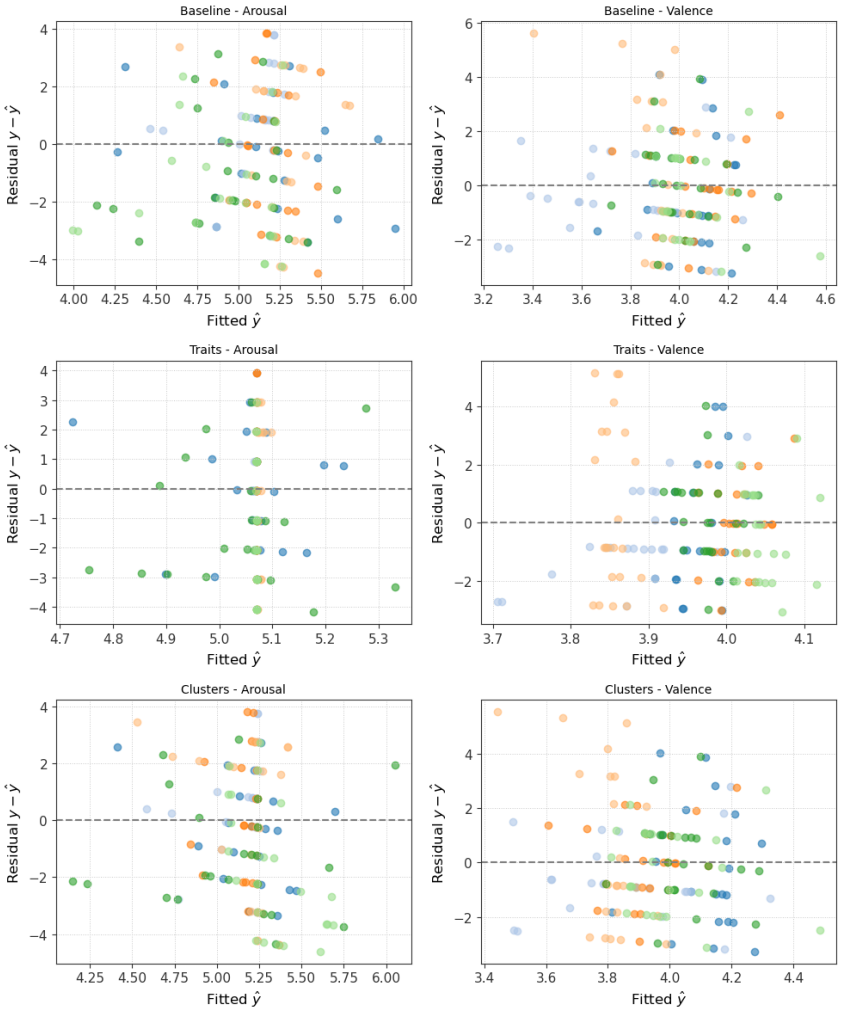


Figure 32: Residuals versus Fitted values for SVR in PhyMER, SVR Residuals versus fitted values for Support Vector Regression (SVR) models across Baseline (top), Traits (middle), and Clusters (bottom) variants. The fitted values are highly concentrated, particularly around the mean, reflecting a lack of flexibility in the SVR predictions. Residual patterns confirm the model's limited capacity to capture variation in affective responses across participants.

D Hyperparameters

D.1 AMIGOS Best parameters

Setting	Task	max depth	max features	max samples	min samples leaf	min samples split	n estimators
Baseline	arousal	5	sqrt	None	3	10	200
Baseline	valence	5	sqrt	None	5	2	150
Traits	arousal	5	sqrt	0.8	3	10	200
Traits	valence	5	sqrt	None	3	10	100
Clusters	arousal	7	0.5	None	3	2	100
Clusters	valence	5	sqrt	0.8	5	5	100

Table 12: Best hyperparameters of Random Forests (RF), Short recordings, AMIGOS Summary with best performing hyperparameters in the RF model of short-segment AMIGOS data, optimized on RMSE. The RF model generally favored shallower trees (max depth = 5), With sqrt or 0.5 as the maximum features setting and small minimum sample leaf values (3–5). This implies the model favored simpler trees, potentially to reduce the risk of overfitting to the training data.

Setting	Task	C	epsilon	gamma	kernel	coef0
Baseline	arousal	10.0	0.1	1.0	rbf	n/a
Baseline	valence	0.1	0.2	0.0001	rbf	n/a
Traits	arousal	10.0	0.1	1.0	rbf	n/a
Traits	valence	0.1	0.2	0.0001	rbf	n/a
Clusters	arousal	10.0	0.1	1.0	rbf	n/a
Clusters	valence	0.1	0.2	0.0001	rbf	n/a

Table 13: Best hyperparameters of Support Vector Regression (SVR), Short recordings, AMIGOS Summary with best performing hyperparameters in the SVR model of short-segment AMIGOS data, optimized on RMSE. The RBF kernel was repeatedly chosen as the preferred kernel option. For arousal, it required higher regularization (C), capturing smooth nonlinear relationships. In contrast, valence prediction needed lower C and a very small gamma, indicating less separable patterns and greater sensitivity to overfitting.

Setting	Task	max depth	max features	max samples	min samples leaf	min samples split	n estimators
Baseline	arousal	10	sqrt	None	5	2	150
Baseline	valence	5	sqrt	0.8	3	10	150
Traits	arousal	10	sqrt	None	3	10	200
Traits	valence	10	sqrt	0.8	3	2	100
Clusters	arousal	5	sqrt	0.8	5	10	150
Clusters	valence	10	sqrt	0.8	3	10	150

Table 14: Best hyperparameters of Random Forests (RF), long recordings, AMIGOS Summary with best performing hyperparameters in the RF model of long-segment AMIGOS data, optimized on RMSE. RF favored deeper trees than in the short segments, and consistently preferred a strict feature selection to be considered for a split (max features)

Setting	Task	C	epsilon	gamma	kernel	coef0
Baseline	arousal	10.0	0.2	1.0	rbf	n/a
Baseline	valence	10.0	0.2	1.0	rbf	n/a
Traits	arousal	10.0	0.2	1.0	rbf	n/a
Traits	valence	10.0	0.2	1.0	rbf	n/a
Clusters	arousal	10.0	0.2	1.0	rbf	n/a
Clusters	valence	10.0	0.2	1.0	rbf	n/a

Table 15: Best hyperparameters of Support Vector Regression (SVR), long recordings, AMIGOS Summary with best performing hyperparameters in the SVR model of long-segment AMIGOS data, optimized on RMSE. The model consistently sets a very high regularization (C) value in combination with a small gamma, particularly when predicting valence. This explains the observed regression to the mean in the model's predictions and suggests that the relationships between ratings and the long video segments were obscure and nonlinear.

D.2 PhyMER Best parameters

Setting	Task	max depth	max features	max samples	min samples leaf	min samples split	n estimators
Baseline	arousal	10	sqrt	0.8	3	10	200
Baseline	valence	10	0.5	None	5	2	100
Traits	arousal	10	0.5	None	10	10	150
Traits	valence	5	0.5	0.8	5	2	100
Clusters	arousal	5	0.5	None	3	10	100
Clusters	valence	7	sqrt	None	10	2	150

Table 16: Best hyperparameters of Random Forests (RF), PhyMER Summary with best performing hyperparameters in the RF model of PhyMER, optimized on RMSE. Baselines tended to prefer deeper trees (max depth = 10), while Traits or Clusters tended to prefer shallower trees (max trees = 5). Model preferred a subset of the features at each split (max features was either '0.5' or 'sqrt'). Overall, this indicates a preference for balancing model complexity (tree depth) with regularization (max features).

Setting	Task	C	epsilon	gamma	kernel	coef0
Baseline	arousal	1	0.2	1	rbf	n/a
Baseline	valence	0.1	0.05	0.1	sigmoid	0.0
Traits	arousal	10	0.2	1	rbf	n/a
Traits	valence	10	0.05	0.0001	rbf	n/a
Clusters	arousal	1	0.2	1	rbf	n/a
Clusters	valence	0.1	0.05	0.1	sigmoid	0.0

Table 17: Best hyperparameters of Support Vector Regression (SVR), PhyMER Summary with best performing hyperparameters in the SVR model of PhyMER, optimized on RMSE. The optimized parameters varied substantially between combinations. Most models preferred RBF kernels, indicating the relationships within the data were nonlinear. The best-performing model, Traits on valence, exhibited high regularization (C) and a small gamma, indicating that it closely fitted the training data.