DETECTING AFFECT STATE LEVELS USING SMARTPHONE, SMARTWATCH & CONTEXTUAL DATA

FINNEGAN SCHONKNECHT

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2120964

COMMITTEE

dr. Hendrickson
Thomas Quadt

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 29th, 2024

WORD COUNT

8774

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. Drew Hendrickson, for his continued guidance and support throughout the process of writing this thesis. I also want to thank Kang et al. (2023) for providing me with the data to use in this thesis. Finally, I would like to thank my Mam, Isabelle, Carol, Shelley, Ethan and Sully for their unconditional support throughout this year.

DETECTING AFFECT STATE LEVELS USING SMARTPHONE, SMARTWATCH & CONTEXTUAL DATA

FINNEGAN SCHONKNECHT

Abstract

Although prior research has been done in affect state detection, no studies have utilised the dataset and methodology used in this paper for affect state detection. This paper aims to investigate the feasibility of detecting affect states, stress, valence, and arousal using sensor data from smartphones and smartwatches. A comparison of traditional machine learning models (Random Forest, XGBoost, and Support Vector Machine) with recent time-series models (LSTM and GRUs) in predicting affect states will be analysed. The dataset used in this paper is provided by Kang et al. (2023), comprising smartphone, smartwatch and self-reported affect state levels from 80 participants tracked over different 7-day periods. This paper addresses two key trends: (1) the increasing societal focus on mental health and (2) advancements in physiological and smartphone-based tracking technologies. Additionally, it addresses the current research gap in applying traditional and time-series machine learning models on this dataset for affect state detection. The primary findings of this thesis improve on the results achieved by Kang et al. (2023), with macro-F1 scores of 0.567, 0.559 and 0.584 for stress, valence and arousal, respectively. The best-performing model is Random Forest across all affect states. Results from this thesis indicate that the use of time-series models such as LSTM and GRUs do not perform any better than traditional models for this prediction task, which contradicts existing literature, with the best-performing Random Forest model showing notable error patterns. Although the results in this paper do not compare to the state-of-the-art in the broader context of affect state detection, it provides a promising starting point for future research utilising this dataset and unobtrusive tracking methods.

CONTENTS

1 INTRODUCTION

1.1 *Problem Statement*

Human emotions are inherently difficult to predict due to numerous environmental factors, individual variability in emotional responses, and their subjective nature. Affective computing aims to detect a user's affect states (emotions) using external observations (Schmidt et al., 2019). Russell (1980) proposed that all affect states (emotions) can be condensed into two dimensions: valence and arousal. Studies have utilised this along with various tracking techniques in order to detect these states; analysing EEG activity, typing patterns, physiological data, smartphone usage, and contextual data have all been shown at various levels of accuracy as predictors of valence and arousal levels in humans (Aalbers et al., 2023; Osmani et al., 2015; Taylor et al., 2020).

In recent years, the intersection of machine learning and behavioural psychology has led to significant advancements in understanding and predicting individuals' affect states. While traditional machine learning models have demonstrated success in predicting self-reported affect state levels, recent time-series model advancements have opened up new areas of research which warrant exploring. Investigating the efficacy of these models on new datasets aims to provide useful insights into methods and models from previous studies while also enhancing the understanding and predictive abilities of models for affect state detection.

1.2 *Social and Scientific Relevance*

Mental health is a critical issue in today's tech-driven era, with smartphones being the most widely used devices (Sidoti et al., 2024). Studies have shown that smartphone usage, along with physiological and contextual data, can detect stress and anxiety levels (Bogomolov et al., 2014; Kyriakou et al., 2019; Vahedi & Saiphoo, 2018). Understanding the relationship between affect states and these data modalities can aid in predicting mental health conditions. Furthermore, long-term exposure to stress hormones (glucocorticoids) has been shown to be associated with depressive disorders (Lupien et al., 2009), further motivating the societal relevance of this research. In 2022, 30.9% of deaths among young people (under 30) were due to suicide, a 17.7% increase since 2000 ("Suicide deaths in 2022", 2023). Hence, identifying accurate proxies for behaviours or patterns leading to undesirable affect states could help in identifying and providing support for people suffering from poor mental health. Given that 97% of people under the age of 60 own a smartphone (Sidoti et al., 2024), coupled with the advancement of smartwatch tracking technology, real-time physiological and behavioural patterns can be tracked unobtrusively. Smartphone app developers or even smartphone manufacturers could integrate affect state detection systems into their devices, allowing for unobtrusive and mobile affect state detection. Aalbers et al. (2023) notes the importance of early detection and treatment of stress, as untreated high levels of stress can lead to more severe mental health problems, such as depression.

Early detection of these undesirable affect states could see drops in suicides, self-harming and people's reliance on mood-altering medications.

From a scientific perspective, this paper evaluates the extent to which both static and time-series machine learning models perform in affect state detection. Applying feature extraction methods that existing literature has found successful, further analysis is done to understand how the models perform overall and on certain groups within the data. Furthermore, if the results in this paper hold up to existing studies, it can provide insights into the effectiveness of both static and time-series models and the included data modalities for affect state detection.

1.3 *Research Questions*

MQ: *To what extent is it possible to detect users' perceived stress levels based on their smartphone usage, smartwatch and contextual data? Additionally, what errors and biases are present in the best performing model(s)?*

The overarching goal of this paper is to evaluate the feasibility and to what extent the models can detect affect states. Various feature extraction techniques and models selected from existing literature have been used to formulate this prediction task. The subsequent sub-questions will outline the research workflow for this paper.

SQ1 - To what extent will the use of a time-series model, namely LSTM and GRU networks, improve macro F1 score for stress, valence & arousal?

To answer this question, LSTM and GRUs will be utilised to evaluate whether they can outperform traditional machine learning models such as Random Forest, SVM and XGBoost in affect state prediction. This research question stems from the current gap in the literature on applying these time-series models to this dataset.

SQ2 - How does the prediction performance of the best performing model differ per class (low/high) level for stress, valence and arousal?

The aim of these models is to identify undesirable levels of stress, valence and arousal. The models should be able to identify high levels of stress, arousal (when coupled with negative valence) and low valence, as these are the affect states that are harmful to the participants. Confusion matrices and further metrics will be evaluated across the best-performing models in order to identify and comprehend what errors the models are making.

SQ3 - How much does the F1-score of the best-performing model differ for disparate groups, namely personality traits?

This research question is inspired by the findings of Bogomolov et al. (2014) and Spathis et al. (2019); results from these papers indicate that model accuracy differs depending on individual memberships to the Big-5 personality groups;

Neuroticism, Conscientiousness, Openness and Agreeableness, Extraversion. To answer this question, we evaluate the best performing model per state on a separate test set, macro-F1 score of the models are reported on classifying each personality trait group, evaluating whether the models show any bias towards any of the personality groups.

2 LITERATURE REVIEW

2.1 *Affect States*

Posner et al. (2005) suggested that all emotions can be classified within a two-dimensional space defined by valence and arousal. As can be seen in Figure 1, arousal spans from high to low, while valence ranges from positive to negative. Emotions are characterised by high arousal and positive valence, for instance, excitement or happiness, whereas those with low arousal and negative valence may indicate feelings of depression or boredom. Building on this, Giannakakis et al. (2019) suggest that although the definition of stress is somewhat subjective, it is generally associated with negative emotional states and can be mapped onto the arousal and valence space, as feelings of high arousal and negative valence (upper left quadrant). These affect states can be further contextualised within external social patterns (Berkman et al., 2014) and physiological biomarkers (Osotsi et al., 2020). However, it's essential to recognise that the relationship between valence and arousal is deeply personalised. Some individuals may associate positive feelings with excitement and other high-arousal states, while others may correlate them with feelings of calmness and contentment (Kuppens et al., 2013). Multiple studies have supported this with their individual-specific trained models outperforming their group models (Aalbers et al., 2023; Osmani et al., 2015; Osotsi et al., 2020).

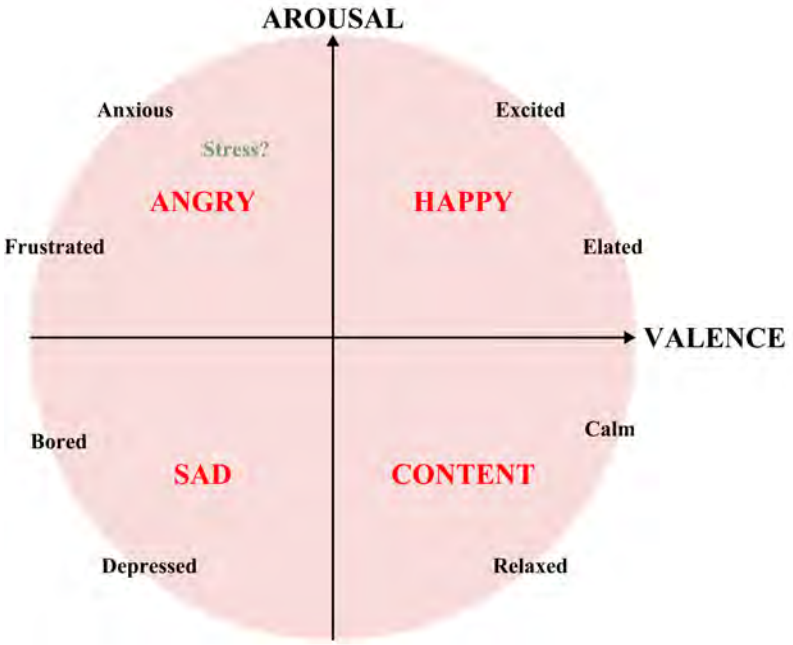

Figure 1: Valence - Arousal Space. Source: the authors illustration based on Kuppens et al. (2013) and Posner et al. (2005)

2.2 *Affect States and Physiological Data*

There have been several studies that utilise physiological features to detect levels of valence and arousal. Bulagang et al. (2021) conducted a study in which participants viewed videos, with each video representing different quadrants of the valence-arousal space. By monitoring heart rate, electrodermal activity (EDA) and skin temperature, their intra-participant models achieved 100% accuracy in detecting levels of valence and arousal. Siirtola et al. (2023) employed an RNN variant to predict self-reported affect states using various self-reporting affect and emotional state surveys, achieving R^2 scores of 0.71 and 0.81 for valence and arousal, respectively. It is worth noting that both of these studies used data that was collected in a controlled laboratory setting, which has been shown to significantly improve accuracy when compared to physiological data that is collected outside of a laboratory setting. Subsequent studies have attempted to utilise real-world physiological data in order to predict affect states. Umematsu et al. (2019) used a wristband tracker in order to predict stress, health & happiness, with physiological data being one of the top predictors across all models. Due to the similarity of the data and prediction task, model performance can be compared to Umematsu et al. (2019), with their state-of-the-art (SOTA) LSTM and Logistic Regression models achieving accuracies of 89.6% 73.4%, respectively. Recent advancements in smartwatches and wearable devices have allowed researchers to shift from obtrusive, highly controlled laboratory experiments to real-world applications in which users go about their daily routines without being hindered by these tracking devices. This has opened lots of doors for researchers to deploy these models into real-world settings and

provide real-world solutions for users. Kang et al. (2023), the original researchers of the dataset used in this study, unlike previous literature (Taylor et al., 2020; Umematsu et al., 2019), did not extract physiological specific features from their raw data. For example, Umematsu et al. (2019) extracted EDA (Electrodermal Activity) artefacts using methods developed by Taylor et al. (2015). They extract the following across all physiological features tracked: mean, standard deviation, skewness, kurtosis, absolute differences, binned entropy, median and time-series complexity. As this paper uses the physiological data provided by Kang et al. (2023), we follow similar data treatment methods.

2.3 *Affect States and Smartphone Usage / Contextual Data*

Smartphone usage and the effects it has on individuals have been extensively studied in the fields of machine learning, human-computer interaction and psychology. Due to the sophistication of smartphones today, researchers do not have to rely on self-reported smartphone usage; rather, software that can track users' interactions with their smartphones has become the preferred method. Likamwa et al. (2013) introduced a novel app named *MoodScope*, which tracked participants' communication history and usage of their smartphones, achieving an initial accuracy of 66% and 93% with a further two months of training, showcasing the predictive power of smartphone usage. Their results found calls and text messages as some of the most influential features, Aalbers et al. (2023) noted similar results, with time spent on messenger apps, day of week and hour of day contributing to moments of subjective stress. Although some studies have indicated that smartphone usage can aid in mental well-being (Stawarz et al., 2019), multiple studies have shown that excessive smartphone usage can affect both physical and mental well-being (Daniyal et al., 2022). X. Zhang et al. (2018) showcased this with their app *MoodExplorer*, utilising microphone, accelerometer, GPS and phone usage data such as SMS call frequency to contacts, screen on-off ratio and usage amount. Results indicate a high correlation between smartphone usage patterns and their self-reported emotional states, with their factor graph model achieving an accuracy of 76%. The smartphone usage feature extraction methods employed in this paper were inspired by X. Zhang et al. (2018). Conversely, Bogomolov et al. (2014) suggested that smartphone usage data alone does not provide good predictive properties in predicting self-reported daily stress levels. Only when they were combined with personality and weather data did they yield informative results. Additionally, personality traits all showed high predictive power, with the five personality traits being Neuroticism, Conscientiousness, Openness and Agreeableness, and Extraversion.

2.4 *Time-series Models*

Recurrent Neural Networks (RNNs) have shown their applicability in affect state prediction, Suhara et al. (2017) explored this by forecasting depressed moods with the use of self-reported moods, activity levels & sleeping. Their method

of applying embedding layers in order to learn semantic relationships between categorical variables was adapted for this paper. Variations of RNNs have also seen promising results, with LSTM and GRUs being two prominent models that have shown promise in affect state prediction (Acikmese & Alptekin, 2019; Suhara et al., 2017). Umematsu et al. (2019) proposed LSTM model utilised physiological, smartphone and behavioural survey data in order to predict tomorrow's well-being (stress, health and happiness), achieving an average accuracy of 84.2%. The intra-correlation of valence & arousal and their effect on stress levels have allowed deep learning models to employ multi-task-learning (MTL) in order to parallelise the learning emotional states (Tran et al., 2021). The shared learning can aid in model accuracy and help improve the generalisation of the model (Kandemir et al., 2014; Y. Zhang & Yang, 2022), Y. Zhang and Yang (2022) defined this as 'Task Relation Learning'. Two of the most prominent studies utilising these techniques in affect state prediction are Jaques et al. (2015) and Taylor et al. (2020). Recently, more complex RNNs have been explored, Spathis et al. (2019), their study utilised an encoder-decoder Multi-task LSTM for valence and arousal prediction. Their models achieved MSE levels of 0.14 and 0.16 for valence and arousal, respectively. However, they note that their models performed better on emotionally stable participants, potentially limiting the application of the models to those without mental disorders. Due to the similarity of the data and the models used, results in this paper for time-series models can be compared to those achieved by Spathis et al. (2019) and Umematsu et al. (2019)

The aim of this paper is to apply techniques similar to the studies discussed above to build on the work done by Kang et al. (2023). Using their results as a baseline to be improved upon. Due to the recent publication date of this dataset, other than the original researchers, only one paper has applied machine learning models to this dataset (Alikhanov et al., 2024). No papers to data have applied the feature extraction techniques that have proven successful in previous studies on this dataset, alongside the application of static and time-series models.

3 METHODOLOGY

3.1 *Baseline Models*

Two baseline models were utilised for the purpose of this study: the Majority Class model and the Naive Bayes model. The aim of these models is to establish a lower bound for this prediction task and provide benchmarks for the static and time-series models to compare to. The Majority Class Classifier predicts the most frequent class in the data and is commonly used as a baseline in existing literature (Bogomolov et al., 2014; Kang et al., 2023).

The Naive Bayes model, based on Bayes theorem of conditional probabilities (Joyce, 2021), is known for its simplicity yet widespread applicability across machine learning domains, including affect state prediction (Jaques et al., 2015; Kang et al., 2023; Romeo et al., 2019), for this reason it was included in this paper.

3.2 *Static Models*

In this section, we will discuss the three static models used in this paper: Random Forest (RF), eXtreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM).

The first model is the RF, which is a tree-based algorithm consisting of multiple decision trees. RF has been featured in many emotion prediction studies (Aalbers et al., 2022; Bogomolov et al., 2014; Gjoreski et al., 2017; Kang et al., 2023; Schmidt et al., 2019), this can be put down to two reasons. Firstly, the RF is one of the most popular machine learning algorithms, with its simplistic nature and interpretable results. Secondly, it deals with high-dimensional data quite well, which is often the case in these studies.

The next model that was included in this paper was the XGBoost Classifier, first introduced in a 2016 paper from Chen and Guestrin (2016); this algorithm has found widespread applications in affect state prediction, with both its regression and classification variants (Asare et al., 2021; Kang et al., 2023).

The final model that was chosen for this paper was the SVM, as seen with the two previous models. This was chosen as a result of its popularity and performance on similar prediction tasks (Bogomolov et al., 2014; Gjoreski et al., 2017; Jaques et al., 2015; Osmani et al., 2015).

3.3 *Time-series Models*

Long Short-Term Memory Networks (LSTMs) were introduced in order to overcome the shortcomings of Recurrent Neural Networks (RNN) (Hochreiter & Schmidhuber, 1997). LSTMs use three gates to regulate information retention and forgetting: the forget gate discards unnecessary information, the input gate updates the cell state with new information, and the output gate outputs relevant information to the next hidden state. These gates help control gradients and preserve information over long sequences. LSTMs are computationally complex, so Gated Recurrent Units (GRUs) were proposed as an alternative with a similar structure but only two gates. There is no consensus on which model is better for capturing long-term dependencies, but LSTMs generally achieve higher accuracy with sufficient data and resources (Murray et al., 2022). Given the popularity of LSTMs for similar prediction tasks (Acikmese & Alptekin, 2019; Siirtola et al., 2023; Spathis et al., 2019; Umematsu et al., 2019), they were included in this thesis for comparison with existing literature. Finally, due to the GRUs' less complex structure and performance on smaller datasets (Oğuz & Ertuğrul, 2023), they were also included in this paper.

3.3.1 *Time-series Model Architecture*

The architecture for the LSTM and GRU used in this thesis was consistent with existing literature, mainly following the architecture employed by Acikmese and Alptekin (2019) and Umematsu et al. (2019). Both models used embedding layers

in order to handle the categorical variables, following Suhara et al. (2017). Once these categorical variables have been fed into their embedding layers, they are concatenated with the normalised numeric sequences. The models then consist of 3 layers, with one hidden LSTM / GRU layer. Following this, a dropout layer was added before the dense layer in order to help with overfitting (Umematsu et al., 2019). Then, a dense output layer with a sigmoid activation function was used to produce the binary classification output. Batch size, the number of units and dropout rate were tuned during the training of the models, which will discussed in more detail in Section 3.4.2. Binary cross-entropy loss and the RMSprop optimiser were used for loss function and optimisation respectively (Umematsu et al., 2019).

3.4 *Hyperparameter Tuning*

3.4.1 *Static Models*

First, the *Participant ID* was used to split the train data into 5 folds using sklearn's GroupKFold split (Pedregosa et al., 2011), each fold contained twelve random participants' full data, maintaining the temporal nature of the data with each participant's data. Four of the folds were used for training the model on a set of hyperparameters, the remaining fold was used to evaluate these hyperparameters. RandomSearchCV (Pedregosa et al., 2011) was used to search the hyperparameter space due to the limited computational resources. Appendix B (page 36) shows the range of parameters that were searched for, and Appendix C (page 37) shows the optimal hyperparameters chosen for each model, which were then used to evaluate the models on a separate test set.

In relation to the SVM model, the Radial Basis Function (RBF) kernel was chosen for its ability to handle non-linear decision boundaries (Pedregosa et al., 2011), which is consistent with similar studies (Jaques et al., 2015; Umematsu et al., 2019).

3.4.2 *Time-series Models*

Greff et al. (2017) showed that tuning hyperparameters for LSTM/GRU models can be done independently. So, using nested loops, the batch size, number of LSTM/GRU units and dropout rate were tuned (Siirtola et al., 2023; Spathis et al., 2019; Umematsu et al., 2019). The participants in this study had variable sequence lengths (ESM responses), and to ensure the time-series models were fed sequential user data, we padded the user sequences with 24 time-step sequences to create uniform sequence lengths. Values of -1999 were used for all features in these padded sequences. All participants were padded to have a length of 128 sequences, as the length had to be divisible by the batch size. To ensure these new padded sequences did not contribute to the loss function or macro-F1 score, a masking layer was added to ignore values of -1999, effectively ignoring all padded sequences and respective targets. The range for both batch size and number of units was 32, 64 and 128. Sklearn's TimeSeriesSplit (Pedregosa et al., 2011) was used to split into five folds during training. The models were then trained on the four 4 folds

of the data, and the hyperparameters were evaluated on the one validation fold. Essentially, using one participant's sequences as both the train and validation sets for each batch. Early stopping with a patience of 5 was added into the optimisation loop, this stopped that iteration of model evaluation if the validation score did not improve within 5 epochs. The best hyperparameters were then used for evaluation on the test set. Appendix C (page 37) shows the optimal hyperparameters found for both models.

4 EXPERIMENTAL SETUP

This section will give a detailed overview of the experimental process and pipeline of this paper. Figure 2 illustrates this paper's general workflow, with more detailed figures throughout this section, to provide clarity on certain processes.

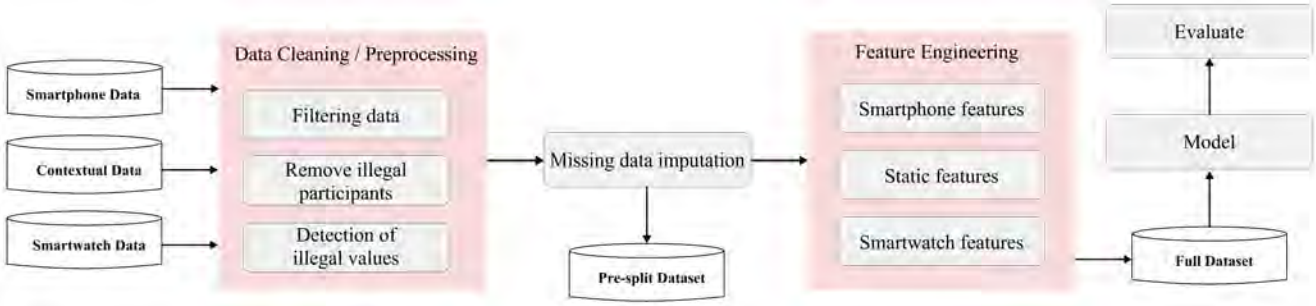

Figure 2: General workflow of this paper, indicating data cleaning and feature engineering processes. Source: The author's illustration.

4.1 *Dataset description*

The dataset obtained for this study was provided by Kang et al. (2023), in which the researchers collected affect and cognitive state labels while measuring smartphone usage/sensors and physiological data collected from smartwatches. The dataset consists of 80 participants (24 females) with a mean age of 21.8 (range = 17–38). The participants were tracked from April 30 to May 8 2019.

The k-emophone dataset consists of four distinct sub-datasets. The first is the pre-survey data, in which participants completed the Korean-translated BFI-15, measuring openness, conscientiousness, neuroticism, extroversion, and agreeableness, along with their age and gender.

The second is the daily Experience Sampling Method (ESM) surveys. The participants' smartphones were randomly prompted 16 times a day with a survey, and with each prompt, the users responded to questions about emotion, stress, attention, task disturbance, changes in emotions, and the duration of the current emotion. Each of the questions was measured on a 7-point Likert scale. Appendix D (page 39 displays the ESM survey questions with a 7-point Likert scale ranging from -3 to +3.

Thirdly is the real-world sensor data, tracked from smartphones and smartwatches. Smartphone data consists of network traffic, social communication and application usage. Smartwatch data contained sensor readings pertaining to physiological data such as heart rate, skin temperature, and steps (see Figure 3).

Lastly is post-survey data, which consisted of surveys in which participants reported stress, depression, and psychiatric disorders using the Perceived Stress Scale (PSS), the Patient Health Questionnaire (PHQ-9), and an altered General Health Questionnaire (GHQ-12).

The data was organised into folders for each participant, containing 23 CSV files. ESM responses were stored in a separate CSV file, with columns named *ResponseTime* and *Pcode* to track responses. Figure 3 illustrates the tracked variables and data structure, with sampling methods including periodic, adaptive, and event-based sampling.

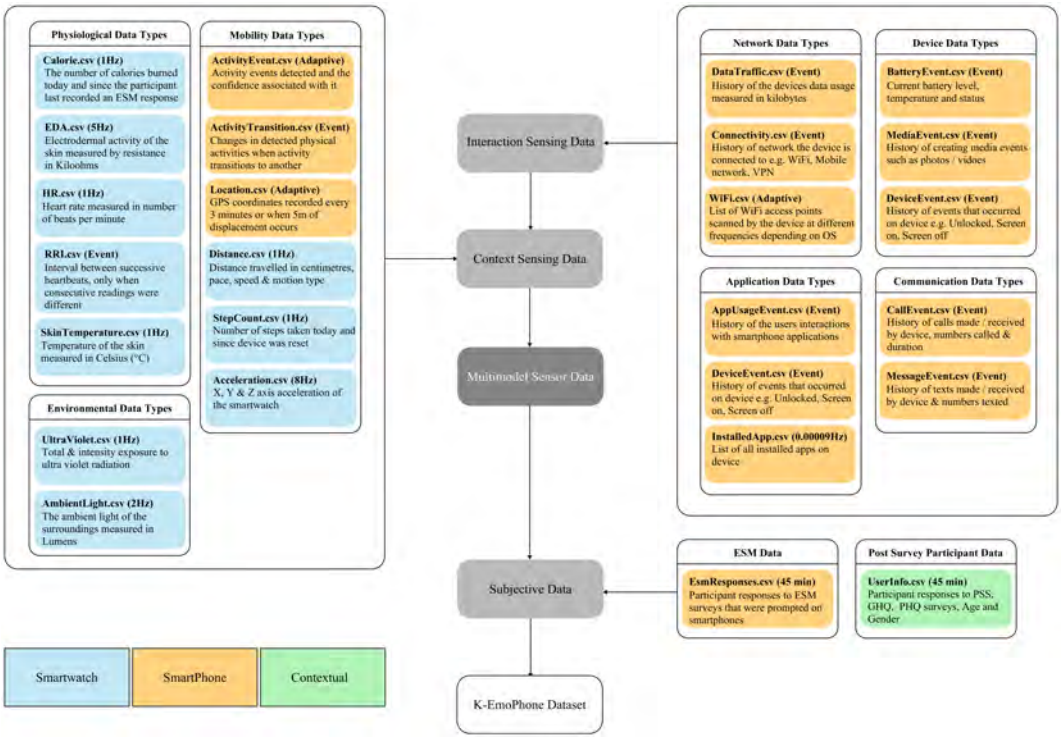

Figure 3: K-emophone data. Source: The authors illustration based on Kang et al. (2023)

4.2 *Dataset Cleaning and Preprocessing*

Firstly, participants P27, P59 and P65 were excluded from the study. P27's physiological data tracked by the MS Band 2 smartwatch generated significantly larger amounts of data than other participants, P59 had no tracked smartphone usage, and P65 did not report any ESM responses. Consequently, we removed them from this paper. Similar to Kang et al. (2023), we excluded participants with fewer than 35 ESM responses, less than 50% of the prompted responses, resulting in the exclusion of 2 participants. Additionally, we discarded ESM responses where 10

minutes had elapsed, and all values for stress, valence, and arousal were 0, as they were non-informative. This resulted in 5,459 valid ESM responses.

4.2.1 *Data Filtering*

The MS Band 2 smartwatch collected physiological data throughout the study. However, due to its nature as a smartwatch rather than a SOTA physiological tracker, some data types were found to be excessively noisy. Specifically, Electrodermal Activity (EDA), which measures changes in skin electrical properties, was excluded for all participants. EDA values typically range from 1.000 Ω to 100.000 Ω (Fish & Geddes, 2009). However, our data averaged 350.000 Ω across all participants. For this reason, we excluded EDA data.

The skin temperature and heart rate features were filtered by excluding values outside the established industry standards. For skin temperature, this was values outside 31° and 38° (C. M. Lee et al., 2019); for heart rate, this was values outside 30 and 200.

4.2.2 *Missing Data Imputation*

Missing data was present across all data types in the dataset. Due to the similarity of the data and the prediction task, the method employed by Spathis et al. (2019) of filling missing values with 0 was used.

4.2.3 *Target Construction*

Similar studies have discarded the middle 20% of responses to address class imbalances (Umematsu et al., 2019). As shown in Figure 4, there is an imbalance of target variables across participants and affect states. Following Kang et al. (2023), we set the threshold for low/high affect states as the mean value for each participant's affect state. This approach, seen in Figure 5, helps create a more evenly distributed target in an attempt to reduce model bias.

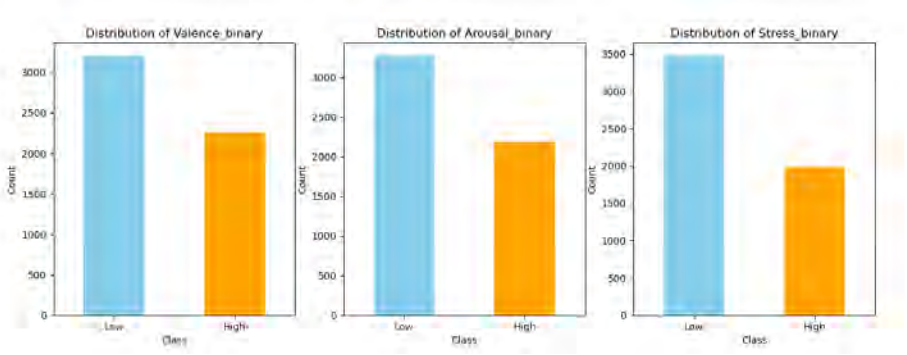

Figure 4: Non-personalised target variables

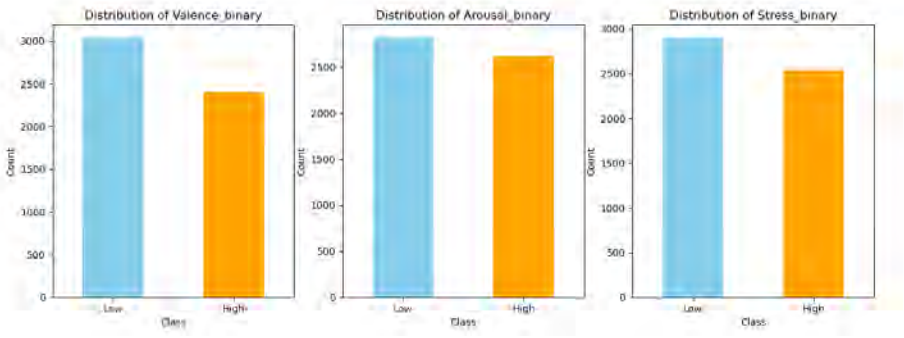

Figure 5: Personalised target variables

4.3 *Common Feature Engineering for Static and Time-series Models*

4.3.1 *Static Features*

Twelve static features were extracted from the data, these features were not extracted using the pre-defined windows. They are as follows;

Personality traits, Age and Gender
Using the pre-study surveys of the participants, the five personality traits were included as features for each participant: Openness, Agreeableness, Conscientiousness, Extraversion and Neuroticism. This was motivated by the findings of Bogomolov et al. (2014) in which personality traits were found to be important features in stress recognition. The age and gender of participants were also included as features, with gender being dummy-coded.

PSS, PHQ and GHQ Levels
Using the post-study survey in which participants indicated their perceived stress levels, degree of depression and severity of psychiatric disorders, these were all included as numeric features ranging from 3 - 15, as they have shown useful in similar prediction tasks

ESM Features
Additionally, the day of the week and hour the ESM response was taken at was extracted as they have also both proven to be influential factors when predicting affect states (Aalbers et al., 2023; H. Lee et al., 2012). The hour feature ranged from 0 - 24, and the day feature was one-hot encoded with 1 indicating the day of the ESM response.

Sleep Proxy
Finally, a sleep proxy was included. Sleep time of the participants was not directly tracked in the original study, however, it was included given previous studies emphasis on the importance of sleep time and quality on emotion levels (Aalbers et al., 2023; Giannakakis et al., 2019; Jaques et al., 2015). This was done with the use of the *DeviceEvent.csv* feature, which tracked whenever a user unlocked/locked

their phone and got a notification. Excluding the hours of 11 am - 7 pm (the assumption was made that users were awake), we took the longest time between the screen-off event and the next unlock event. This resulted in a sleep proxy for every day of the study for each participant, which was included as a static feature for every ESM response on that day.

4.3.2 *Windowed Features*

Call / SMS Events
Social support and human interactions have also been seen as an important factor when predicting participants well-being and emotion levels (Likamwa et al., 2013; Taylor et al., 2020; Umematsu et al., 2019; C. Zhang et al., 2018). Consequently, similar to Bogomolov et al. (2014) and X. Zhang et al. (2018) the following was extracted; total calls/messages, unique numbers texted/called, unique numbers that called the participant, and time spent calling for each window.

Phone Usage
Similar to Jaques et al. (2015), and in an attempt to constrain the number of features, the apps were aggregated into their respective categories and the following features were extracted for each window: top 5 categories interacted with categories, number of interactions per category, entropy of categories used, frequency of smartphone unlocks, and proportion of time spent on the smartphone. Appendix E (page 40) shows the full list of app categories.
Location
The location data included in this dataset was encoded in order to protect the privacy of the participants who were involved in the original study. As a result of this, the actual location of the participants could not be derived, but the spatial relationship between coordinates remained. Hence, the treatment of the data is the same as Kang et al. (2023). First, each pair of longitude and latitude coordinates were encoded using a 7-bit geohash in which the same coordinates were represented by the same 7-character sequence. Then, the entropy of these new geohash clusters was calculated.

Figure 6 illustrates the train/test split and separate processes needed for the static and time-series models.

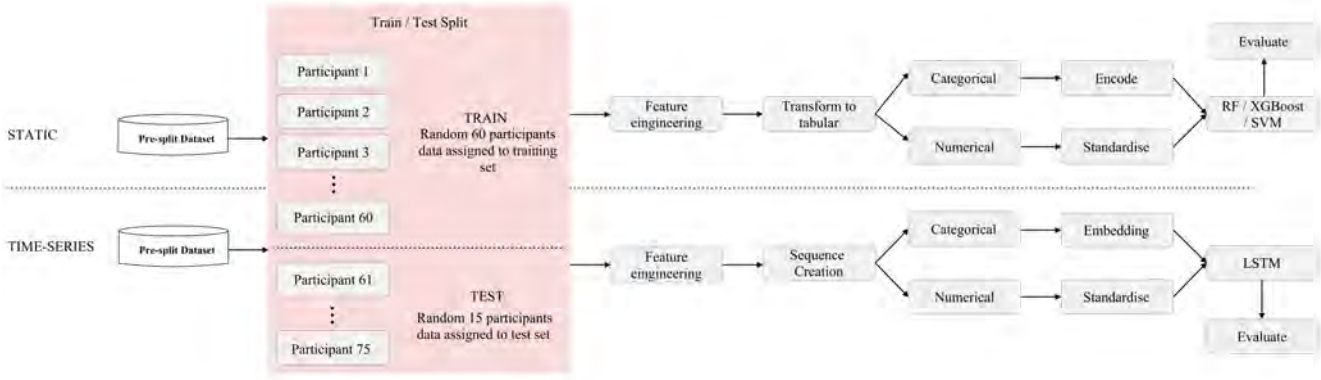

Figure 6: Workflow of train/test split and separate processes for static vs time-series models. Source: The author's illustration

4.4 *Static Model Feature Engineering*

To make the time-series data compatible with our static models, we transformed it into tabular format, following Kang et al. (2023) and Umematsu et al. (2019). With only 5438 target variables, using each timestamp would result in the p > n (big-p, little-n) problem. To address this, we aggregated the data into eight windows: 30 seconds, 1 minute, 5 minutes, 10 minutes, 30 minutes, 1 hour, 3 hours, 6 hours, and 12 hours. There is no standard for window sizes in the literature, so we used those from Kang et al. (2023) and added a 12-hour window for a fair comparison between static and time-series models. For each window and feature (not including the features described in Section 4.3), we extracted mean, median, standard deviation, and entropy. This resulted in 1211 total features., figure 7 shows the tabular data structure for the static models.

Response_Time	P_Code	30_Sec_Feature_1_Entropy	1_Min_Feature_1_mean	...	12_Hour_Feature_1206_mean	Stress_Binary
12:45:26	17	0.569	35	...	1.501	1
13:30:55	17	0.005	12	...	0.257	0
09:05:22	65	1.00	76	...	42.56	1
...	...	...	...	...	...	...

Figure 7: Tabular data structure for static models

4.4.1 *Categorical Variable Treatment*

Due to the high cardinality of the categorical data, methods employed to mitigate this were used in relation to both static and time-series models. This further motivated the feature extraction techniques of app usage and location data. If all of the app interactions and geohash location values were one-hot encoded, it would lead to a feature size that is not compatible with the number of outcome variables available in this dataset. Hence, the remaining low dimensional categorical

variables were one-hot encoded: gender, day of the week, and the top 5 categories of app interactions (for each window).

4.5 *Time-series Model Feature Engineering*

Feeding the raw data would result in excessively long sequences, harming model performance (Spathis et al., 2019). Hence, resampling of the data is required. While the literature on applying LSTM and GRUs to this type of data exists, detailed methodologies for resampling and feeding data into models are sparse. We followed Acikmese and Alptekin (2019) for sequence length and windowing, resulting in sequences of length 24. If 12 hours of data were unavailable, we inserted -1999 time-steps into sequences and utilised the masking layer described in Section 3.4.2 to ignore these time-steps. Figure 8 illustrates 128 the multi-time-step sequences created for each participant. Each time step had 113 numerical features extracted from the data.

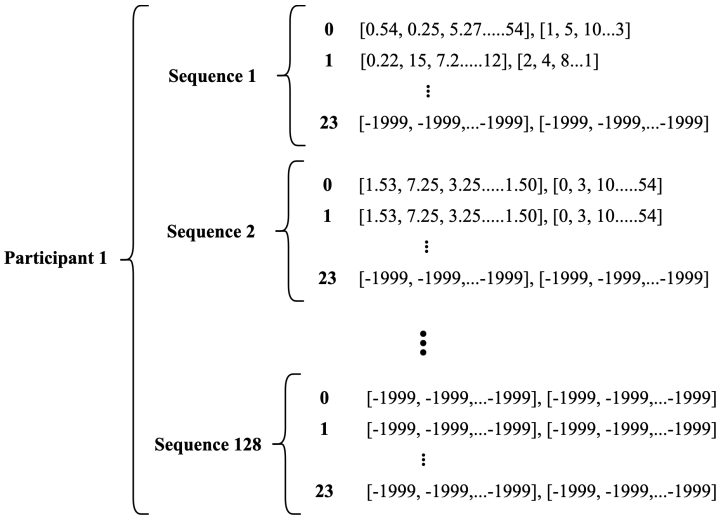

Figure 8: Sequence structure for LSTM and GRU models, with -1999 values for padded timesteps and sequences

4.5.1 *Categorical Variable Treatment*

Inspiration was taken from Suhara et al. (2017), who utilised embedding layers for their categorical variables. The embedding layers represent the categorical values in a continuous vector space in which categories with similar meanings or values are grouped closer together than those that are dissimilar. These relationships are learned during the training process with the use of the Keras Embedding Layer (Chollet et al., 2015). The dimension in which to reduce these variables is freely chosen. However, literature indicating the embedding size is sparse, Suhara et al. (2017) do not specify the rule they follow. Hence, a rule proposed by Howard and Gugger (2020) was used:

$$\textit{Embedding Size} = \begin{cases} \textit{Dimensionality} \ / \ 2, & \text{for } \textit{Dimensionality} < 50 \\ 50, & \text{for } \textit{Dimensionality} \geq 50 \end{cases}$$

The categorical variables that used the embedding layers were as follows: Gender, the day of the week and the top 5 app categories for the 30-minute window, totalling 7 embedding layers. Each category was first represented as integer indices, then fed into their respective layers

4.6 *Robustness Check*

Robustness checks were present for the training of all the models present in this paper. As can be seen in Table 2, the standard deviation of the model macro-F1 score was reported. For the baseline and static models, this was calculated across the 5 folds in which the models were trained and the hyperparameters were tuned. For the time-series models, the standard deviation of the macro-F1 score was calculated across epochs. The training score variability can be seen in Table 2, the figures in brackets indicating the standard deviation of the models' performance during training.

4.7 *Out-of-sample Model Evaluation*

The dataset was split into a training set (80% of participants) and a test set (20% of participants), ensuring no overlap between the two sets, following Spathis et al. (2019). This method was chosen to assess the model's ability to generalise to unseen participant data, reflecting real-world scenarios where models encounter new participants. Although another common method is to do a time-series split of the participants data (Aalbers et al., 2022; Kang et al., 2023). However, as previously stated, reserving 100% of a participant's data for testing can aid in more realistic model performance for real-world settings, which is why this method was chosen.

4.8 *Evaluation Metrics*

Macro-F1 was deemed the most appropriate metric to report due to the class imbalance across all three affect states, and it is consistent with the original paper (Kang et al., 2023).

4.9 *Algorithms and Software*

All of the coding and analysis for this paper was in either a Jupyter or Google Colab notebook, using version 3.11.4 of Python. Google Colab was used for some of the model training as access to better computing resources was necessary. Table 1 shows all of the software packages that were used throughout the entire coding process.

Package	Version	Source
Pandas	2.2.1	(McKinney et al., 2010)
Numpy	1.26.0	(Harris et al., 2020)
Seaborn	0.12	(Waskom et al., 2017)
Scikit-learn	1.3	(Pedregosa et al., 2011)
Matplotlib	3.8.0	(Hunter, 2007)
Keras	3.1	(Chollet et al., 2015)
TensorFlow	2.16.1	(Abadi et al., 2016)

Table 1: Packages to be used

5 RESULTS

This section analyses the results of all models trained to detect stress, valence, and arousal separately, following the structure of the research questions defined in Section 1.3. First, we compare and analyse static and time-series models. Then, we select the best model(s) for each affect state and analyse feature importance, providing insights into model performance and the effectiveness of feature engineering techniques. Finally, we conduct error analysis and disparate group analysis on the best-performing model(s).

5.1 *Baseline, Static and Time-series Model Performance*

Starting with stress detection, the RF model performs best with a test macro F1 score of 0.567, surpassing the next-best model (GRU) by 4.6 percentage points. All static and time-series models outperform the majority classifier, indicating they can predict the most common class and capture some level of underlying data patterns. However, the LSTM performs worse than the Naive Bayes model, with evidence suggesting overfitting is present in the LSTM, as there is a drop of 9.3 percentage points from training to testing. The best-performing time-series model is the GRU, achieving a score of 0.521. The results for stress detection would not suggest that the time-series models are able to capture long-range temporal dependencies more efficiently than the static models.

For valence detection, the best performing model was the RF model, achieving a test macro-F1 score of 0.559, outperforming the next best model by 0.6 percentage points. Similarly to what is seen in stress detection, all of the static models outperform the baseline models. Unlike stress detection, we see both time-series models outperform the baseline models. Furthermore, the GRU otperforms the LSTM by 3.8 percentage points. All of the static and time-series models showed an increase in test macro-F1 scores when compared to stress detection, alternatively both the baseline models performed worse. The second best performing model for valence detection is the GRU, which has comparable results to the RF model. This

may indicate as the detection task becomes more complex, the shortcomings and simplicity of the baseline models become more prevalent, and the static time-series models are able to capture the trends in the data more efficiently.

Finally, for arousal detection, the RF model again performs best with a macro-F1 score of 0.584, outperforming the next-best model by 0.8 percentage points. The time-series models struggle in arousal detection, performing worse than all static models and showing signs of overfitting, with decreases of 6.3 and 6.7 percentage points from training to testing for LSTM and GRU, respectively.

Overall, static models generally perform better across all affect states, for stress and valence detection the GRU showed comparative results, being the second best performing model across static and time-series models. Additionally, the GRU model outperforms the LSTM across all three states by an average of 2.6 percentage points in test macro-f1 scores. However, the time-series models seem to suffer from signs of overfitting, with notable drops from train to test in macro-F1 scores. Finally, the RF model performs best across all three affect states, hence any further analysis in this paper will be conducted on the RF model.

		Macro F1-Score	
		Train (Std)	Test
Stress	*Majority Classifier*	*0.346 (0.010)*	*0.353*
	Naive Bayes	*0.502 (0.021)*	*0.489*
	Random Forest	0.513(0.032)	**0.567**
	XGBoost	0.460(0.027)	0.409
	Support Vector Machine	0.494 (0.021)	0.514
	LSTM	0.576 (0.101)	0.483
	GRU	**0.600** (0.104)	0.521
Valence	*Majority Classifier*	*0.359 (0.018)*	*0.352*
	Naive Bayes	*0.486 (0.036)*	*0.417*
	Random Forest	0.561 (0.020)	**0.559**
	XGBoost	0.537 (0.014)	0.518
	Support Vector Machine	0.543 (0.023)	0.545
	LSTM	0.563 (0.036)	0.532
	GRU	**0.585** (0.089)	0.553
Arousal	*Majority Classifier*	*0.338 (0.007)*	*0.353*
	Naive Bayes	*0.524 (0.038)*	*0.431*
	Random Forest	**0.586** (0.034)	**0.584**
	XGBoost	0.556 (0.035)	0.561
	Support Vector Machine	0.562 (0.015)	0.576
	LSTM	0.539 (0.068)	0.476
	GRU	0.561 (0.060)	0.494

Table 2: Overview of results achieved by baseline, static and time-series models, the best score for each class is marked in bold

5.2 *Feature Importance*

Analysing feature importance for the affect states aims to enhance understanding of the models and provide insights into which features contribute to model predictions. This evaluation assesses how the three types of data used (smartphone, smartwatch, and contextual) contribute to model performance. Similar to Aalbers et al. (2023), SHAP (SHapley Additive exPlanations) values are utilised to evaluate feature importance across all affect states and their impact on each class (low/high). These values were calculated on the test set, with colours representing feature values (Y-axis), where higher entropy levels for features like *"1_Hour_Steps_Entropy"* show as red and vice versa. Along the X-axis, SHAP values indicate how strongly each feature influences the model's prediction of low or high levels of each affect state.

5.2.1 *Stress*

Figure 9 illustrates that nine out of the top ten most influential features for stress prediction are physiological. The only non-physiological feature shown is the median confidence level of the smartphone in detecting walking activity (5_Min_Walking_Device_Confidence_Median). Among the physiological features, those related to step counts over one-hour and three-hour intervals are prominent, indicating their significance in stress detection. Additionally, the calories burned feature also shows influence, with similar informative patterns observed in one-hour and three-hour windows for both steps and calories burned features.

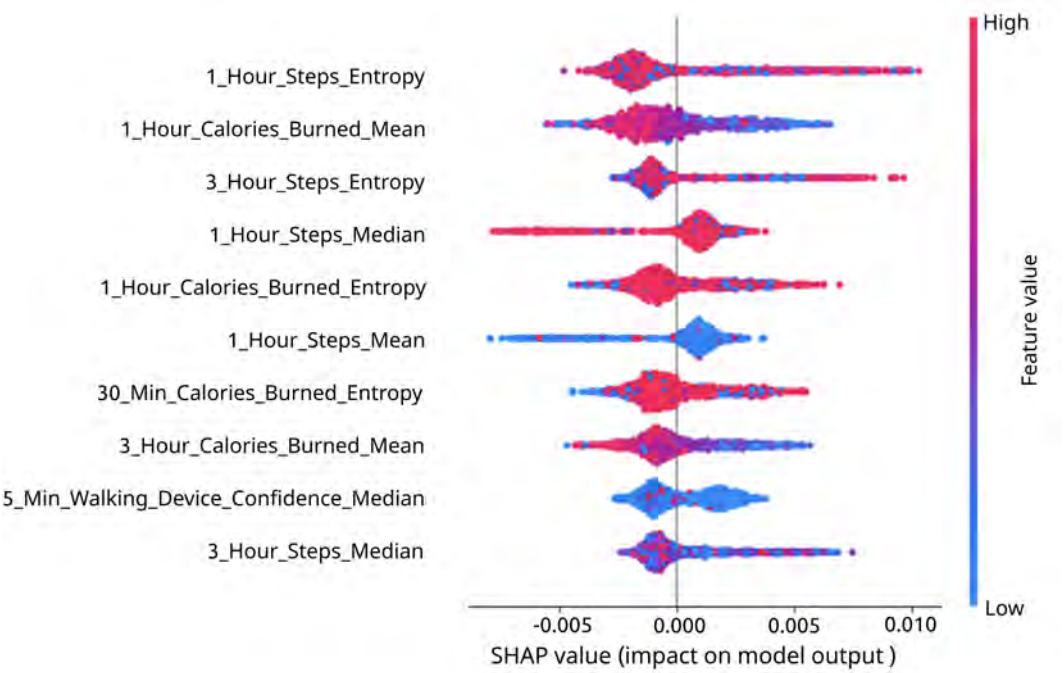

Figure 9: SHAP values for RF model detecting low stress (class 0)

5.2.2 *Valence*

Figure 10 displays SHAP values obtained from the test set for predicting low valence levels using the RF model. Unlike stress predictions, low valence levels are associated with negative emotions or affect states (Figure 1). The four most important features identified include entropy and mean values of calories burned by participants, where higher values steer predictions away from low valence. Additionally, entropy and mean values of distance travelled are influential features. Skin temperature entropy over a three-hour period also proves influential for valence detection, with high entropy values influencing the model to predict high valence levels.

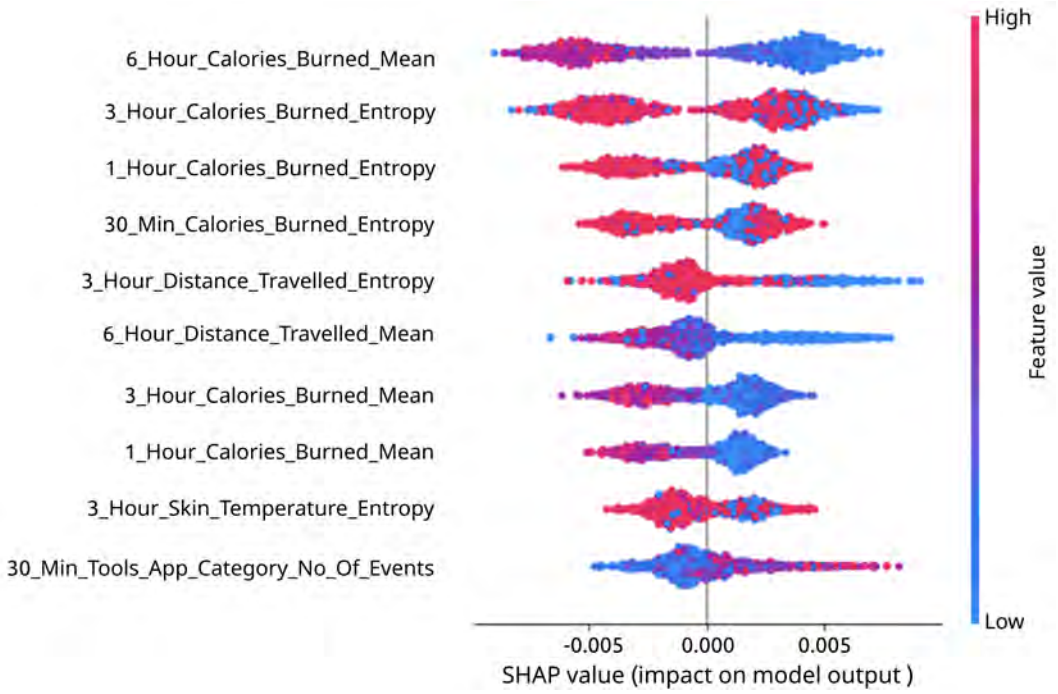

Figure 10: SHAP values for RF model detecting low valence (class 0

5.2.3 *Arousal*

Figure 11 graphs the SHAP values obtained for predicting low states of arousal. Similar to stress and valence predictions, physiological data stands out as the most influential category for arousal detection. Features such as RRI intervals (time between successive heart rates) and accelerometer variables dominate the top ten features. In contrast to stress and valence prediction models, the number of user interactions with communication apps also proves influential for arousal prediction, where fewer interactions push the model towards predicting high arousal levels.

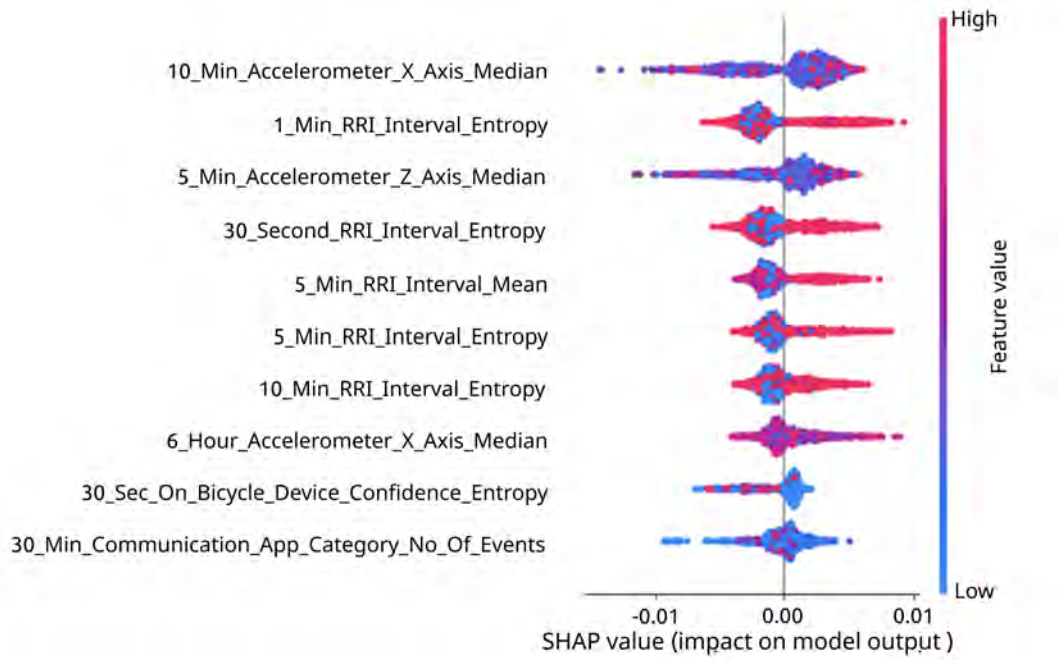

Figure 11: SHAP values for RF model detecting low arousal (class 0

All of the models rely heavily upon physiological data in order to make their predictions. It must be noted that this does not imply a correlation between these features and low / high affect state levels but rather an insight into what features the model relies upon for predictions.

5.3 *Error Analysis*

Error analysis was conducted on the best model across all three affect states. The RF model was the best-performing model across all three affect states, with all further analysis being done on this model.

5.3.1 *Stress*

As can be seen in Figure 12, which is the confusion matrix for stress detection. It shows the most prominent error made by the model was mislabeling high-stress states as low stress, with 276 instances of high-stress levels being mislabeled by the model. This would be considered a Type II error and harmful for the model to make. For this classification task, a Type II error is much more harmful than a Type 1, Type II errors could lead to undiagnosed high stress levels, which if gone untreated could lead to chronic stress or even more severe mental health issues (Aalbers et al., 2022).

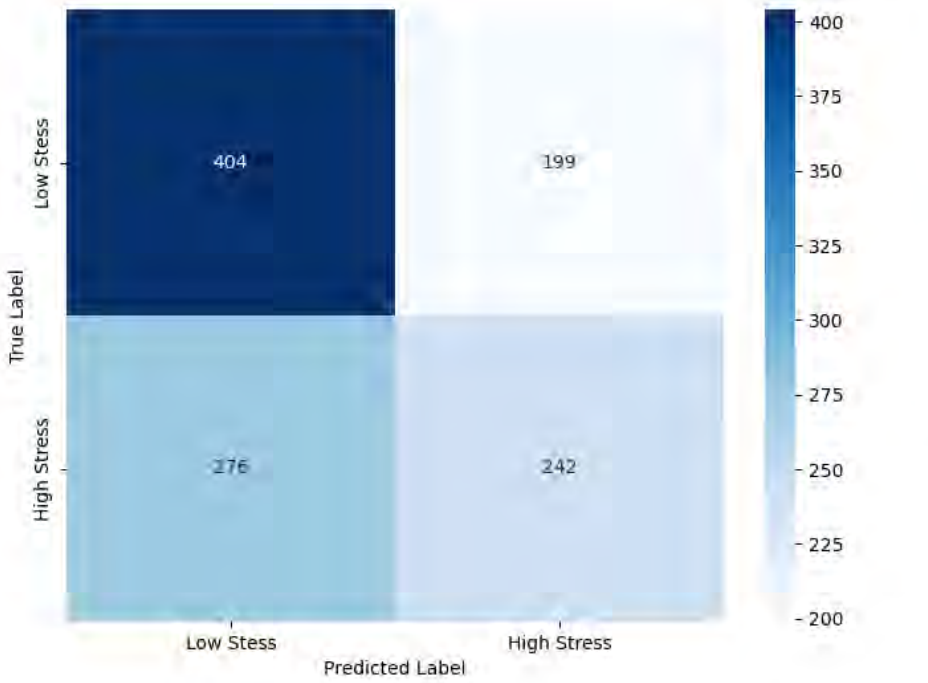

Figure 12: Stress Confusion Matrix

Additionally, precision, recall and macro-F1 scores for each class were constructed. Referring to Table 3, it can be seen that across all three metrics, the detection of high-stress levels performs significantly worse than low stress. For instance, the high-stress class reports a recall of 0.47, again showing the inability of this model to detect high-stress levels accurately.

True label	**Precision**	**Recall**	**Macro F1-score**
Low Stress (0)	0.59	0.67	0.63
High Stress (1)	0.55	0.47	0.50

Table 3: Precision, Recall and Macro-F1 scores of Random Forest model for stress detection

5.3.2 *Valence*

Referring to Figure 13, the RF models most prominent error is a Type II error. Occurs when low valence levels are misclassified for high valence levels. As was seen for stress detection, this is a much more harmful error for the model to make, with low valence levels being associated with emotions such as depression and anxiety (Figure 1). The model seems to be performing better on the majority class present in the data, which is high levels of valence (Figure 5).

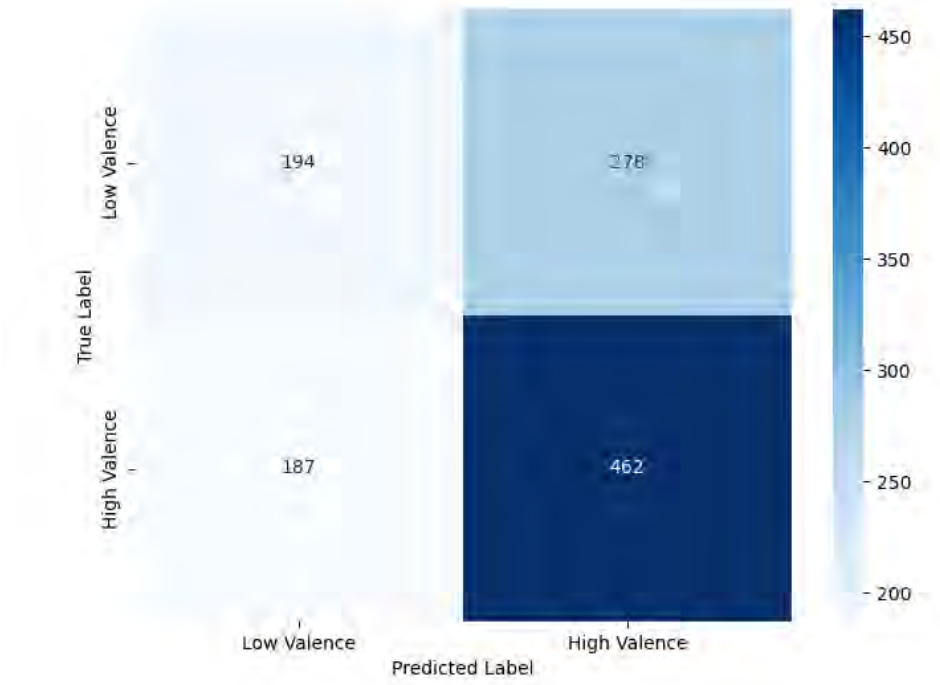

Figure 13: Valence Confusion Matrix

Furthermore, Table 4 displays the per-class metrics for valence detection. As can be seen for low valence levels (0 class), the model performs considerably worse with a macro-F1 score of 0.45, with a 21 percentage point difference in macro-F1 score between the classes.

True label	Precision	Recall	Macro F1-score
Low Valence (0)	0.49	0.41	0.45
High Valence (1)	0.62	0.71	0.66

Table 4: Precision, Recall and Macro-F1 scores of Random Forest model for valence detection

5.3.3 *Arousal*

Finally, we analyse the RF models performance for arousal detection. Figure 14 shows the arousal confusion matrix. Unlike stress and valence levels, low and high states of arousal are both associated with negative and positive emotions. Hence, we cannot say with certainty what errors are more harmful. It shows that similar to valence states, low levels of arousal are being misclassified as high arousal (Type-II errors). Referring back to Figure 5, low levels of arousal are the majority class. Interestingly, the models most prominent error is misclassifying the majority class, which could suggest the model is overfitting the majority class during training (or underfitting the minority).

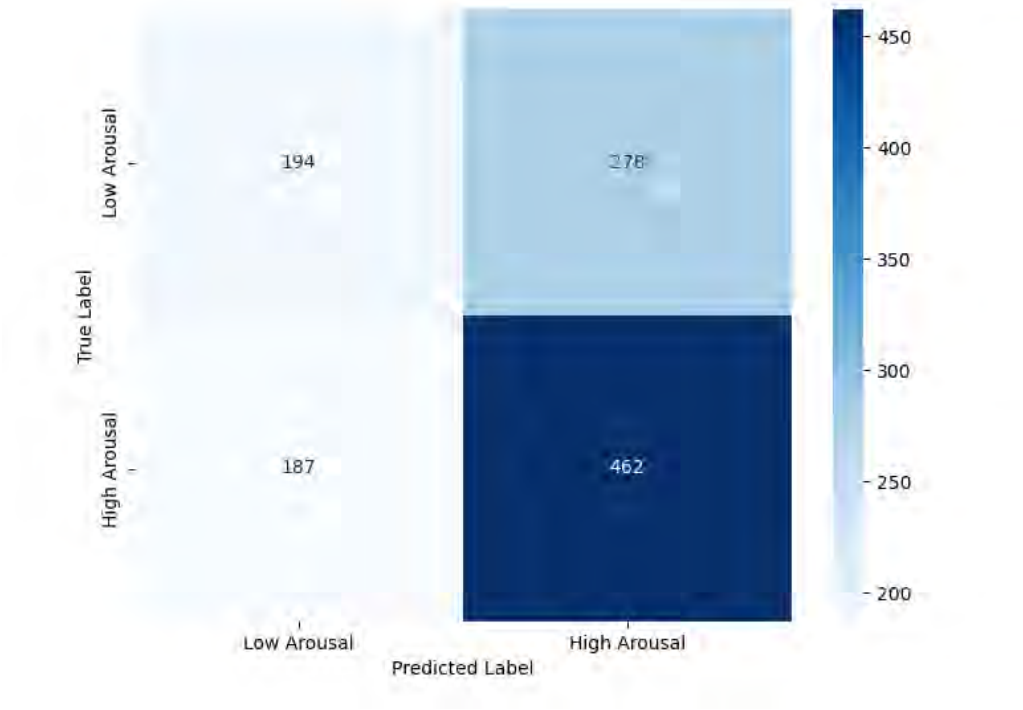

Figure 14: Arousal Confusion Matrix

Additionally, Table 5 shows the class-wise breakdown for arousal detection. The model performs similarly for both classes, achieving matching macro-F1 scores. Low arousal outperforms high arousal by 4 percentage points in the precision metric, indicating the model is more confident in its positive predictions for low arousal. This is because a higher precision means that a higher proportion of the positive predictions are true positives.

True label	Precision	Recall	Macro F1-score
Low Arousal (0)	0.61	0.55	0.58
High Arousal (1)	0.57	0.58	0.58

Table 5: Precision, Recall and Macro-F1 scores of Random Forest model for arousal detection

Overall the macro-F1 scores across the models indicate positive results in comparison with the literature. However, upon further analysis, it must be noted that the RF model makes significant error patterns when classifying the more consequential and harmful affect states. This can clearly be seen in the breakdown of stress and valence detection. If this model were to be used in a real-world scenario, these kinds of errors could cause serious harm and lead to unhealthy, undiagnosed and undesirable affect states for users.

5.4 *Disparate Group Analysis*

For disparate group analysis, the predictive performance of the best-performing model across all three affect states was analysed for participants' membership personality groups in order to inform us of the generalisation capabilities of the model. Table 6 shows the test set macro-F1 scores of each personality trait across all affect states. Participants with high levels of Conscientiousness achieve the highest results across all three affect states. This is closely followed by high levels of Agreeableness with macro-F1 scores of 0.58, 0.63 and 0.58 for stress, valence and arousal, respectively. Conscientiousness, Agreeableness and Openness all achieve relatively similar results, suggesting negligible differences in prediction performance. However, there is a notable difference in macro-F1 scores for high levels of Neuroticism and Extraversion, with both traits being the lowest-performing traits.

	Macro-F1		
Personality Traits	**Stress**	**Valence**	**Arousal**
Neuroticism	0.45	0.49	0.50
Conscientiousness	0.59	0.67	0.60
Openness	0.55	0.59	0.56
Agreeableness	0.58	0.63	0.58
Extraversion	0.46	0.55	0.52

Table 6: Macro-F1 Scores for Personality Traits Across Different Affect States

6 DISCUSSION

6.1 *Time-series vs Static Models*

One of the research questions for this thesis, was to evaluate whether SOTA time-series models could outperform more traditional machine learning models for affect state detection. This was evaluated in Section 5.1, and clearly shows that they performed notably worse on this prediction task and dataset. Studies that have tackled similar prediction tasks while utilising time-series models have seen an increase in model performance, which contradicts the results obtained in this thesis. Spathis et al. (2019) saw their encoder-decode LSTM outperform all of the static and baseline models for valence and arousal prediction. Similarly, Umematsu et al. (2019) found that the LSTM model using daytime physiological data was able to predict tomorrow's stress with an accuracy of 80%. It is worth noting that the dataset used in this thesis is considerably smaller than those used in similar studies. Spathis et al. (2019) and Umematsu et al. (2019) used datasets tracking their participants from November 12, 2010, to May 21, 2011, and February 2013, until October 2016, respectively. The inferior results obtained in this thesis could be attributed to the smaller size of the dataset, but this cannot be stated with certainty.

However, this does open up promising areas for future research. The extensive tracking of smartphone usage and physiological features is very promising, and given the unobtrusive nature of the tracking, expanding this tracking for longer periods is achievable.

6.2 *Results Compared to the Literature*

The results of the best models in this paper are notably higher (see Section 4.9) than the results obtained by Kang et al. (2023), which is the only other paper to use this dataset for affect state detection. This paper's best-performing model for stress detection was Random Forest, achieving a test macro-F1 score of 0.567. This improved results obtained by Kang et al. (2023) by 2.3 percentage points and shows comparable results to Bogomolov et al. (2014), who achieved a macro-F1 score of 0.57. The same is found for valence and arousal detection, with this paper's model outperforming Kang et al. (2023) by 11.2 and 5 percentage points, respectively. The results achieved in this study are also comparable to other similar studies, Schmidt et al. (2019) obtained a macro-F1 score of 0.47 for stress prediction (binary classification), with the use of physiological data. Acikmese and Alptekin (2019) LSTM model outperformed this paper's best stress detection results by 6 percentage points. However, it outperformed this paper's best time-series result by 11. When compared to SOTA studies that utilise similar models and feature extraction techniques, the results of this paper differ considerably. Umematsu et al. (2019) achieved an accuracy of 83.5% with their LSTM model outperforming the LSTM and GRU models in this paper considerably. The feature extraction techniques employed in this paper were taken from a number of different studies. However, feature importance analysis using SHAP values (Section 5.2) showed that none of the manually extracted features (Section 4.3) appeared in the top 10 most important features across all three affect states. Furthermore, physiological features appeared as the most prominent category of important features, aligning with results obtained by Umematsu et al. (2019). Although the manually extracted features taken from existing literature did not appear in the most important or influential features. The results obtained from this paper outperform those of Kang et al. (2023), potentially indicating that despite none of the features individually contributing significantly to model predictions, their aggregated influence increases model performance. However, further analysis would need to be done in order to state this with certainty.

6.3 *Error & Disparate Group Analysis*

The final part of this paper was to evaluate and analyse the error and prediction patterns of the best-performing model(s). Although this study outperformed Kang et al. (2023) across all affect states, the RF model showed poor predictive performance in predicting the less desired and potentially harmful affect states (high stress and low valence levels). Although low-stress states were the majority

class (Figure 5), low valence was the majority class, indicating that the model was not necessarily biased towards the majority class. This limitation of the models calls into question the efficacy of these models in real-world scenarios.

The disparate group analysis in this paper is inspired by the findings of Bogomolov et al. (2014) and Spathis et al. (2019). Particularly Spathis et al. (2019), who found that in their study, participants who displayed high levels of Openness were associated with more accurate model predictions. The proposed reason for this was that participants with high levels of Openness tend to be more open to new ideas and have intellectual curiosity, hence they might use the app more honestly and subsequently be easier to predict their affect states. Furthermore, the results in this paper contracted the findings of Spathis et al. (2019), as no difference in model performance can be seen for participants with high levels of Openness, with the Openness trait achieving similar scores to high levels of Agreeableness and Conscientiousness. There does not seem to be any notable difference in macro-F1 score across most of the personality traits across all of the affect states, with the exception of Neuroticism. As is noted in Bogomolov et al. (2014), previous studies have found that people with high levels of Neuroticism respond more negatively to stressors and report higher levels of daily stress. They further stated that students with high levels of Extraversion and Conscientiousness respond more positively and are less affected by stressors. This could potentially attributed to the models difficulty in detecting affect states for participants with high levels of Neuroticism, as it was not a user-specific model, the model leans the patterns that lead to high levels of stress, valence and arousal for the other four personality traits. However, these rules and patterns may not hold for participants with high levels of Neuroticism. Although, it must be noted that without further analysis and model evaluation this cannot be stated with certainty.

6.4 *Limitations and Future Directions*

Despite careful filtering, the final dataset for this study contained only 5,348 data points, significantly fewer than similar studies (Osmani et al., 2015; Umematsu et al., 2019). This limited dataset size may hinder the model's ability to learn relevant patterns, especially for time-series models that require more data (Yamak et al., 2019). Additionally, The models were trained on group participant data rather than user-specific data. Previous studies have shown that user-specific models yield more accurate results for affect state prediction. For instance, Osmani et al. (2015) found their non-user-specific model achieved 54% accuracy, 34.1 percentage points lower than the user-specific model. Similar results were reported by Aalbers et al. (2023). Due to limited data (an average of 71 ESM responses per participant), a user-specific approach was not feasible. Future research could explore combining user-specific models with longer tracking periods to enhance affect state detection. Finally, the homogeneity of the data, with all participants being Korean and under the age of 38, the model's generalisation capabilities may suffer when presented with a heterogeneous participant, such as an older European. However, this limitation is common in existing literature. For example, the SNAPSHOT (Sano

et al., 2018) and StudentLife (Wang et al., 2014) datasets, considered SOTA in emotion detection, also involve homogeneous participant groups. Future research could involve more heterogeneous participant groups to improve generalisation.

7 CONCLUSION

The aim of this thesis was to evaluate to what extent it is possible to detect participants' perceived affect at state levels. We will start by evaluating the sub-questions, which will build to answering the main research question:

SQ1 - To what extent will the use of a time-series model, namely, LSTM and GRU Networks, improve macro F1-score for stress, valence and arousal?

As summarised in Section 5.1, the use of the LSTM and GRU models did not show any evidence to indicate they were able to improve macro-F1 scores across all three affect states. This is not consistent with the literature, but as referenced in Section 6.4, the size of the dataset is considerably smaller than similar studies.

SQ2 - How does prediction performance of the best performing model differ per class (low / high) level for stress, valence and arousal

The aim of these models was to identify undesirable levels of stress, valence and arousal, it is clear from the results that the models have a hard time distinguishing the different low/high states across all affect states. Particularly for stress and valence, the error patterns show the model struggles to detect undesirable levels of those affect states (high stress and low valence).

SQ2 - How much does the F1-score of the best-performing model differ for disparate groups, namely personality traits?

The macro-f1 score does not differ substantially across four of the five personality traits. This contrasts with results obtained by Spathis et al. (2019), which showed high levels of Openness had better model performances. However, the results in this paper showed Neuroticism performed worst out of all traits. This follows the findings of Bogomolov et al. (2014). However, overall macro-F1 scores did not differ substantially across the five personality traits.

MQ - To what extent is it possible to detect users' perceived stress levels based on their smartphone usage, smartwatch and contextual data? Additionally, what errors and biases are present in the best-performing model?

This paper demonstrates that combining smartphone, smartwatch, and contextual data effectively detects stress, valence, and arousal states, achieving macro-F1 scores of 0.567, 0.559, and 0.584, respectively. Despite notable error patterns in detecting undesirable affect states, bias assessment of the best-performing models indicated minimal differences in predictive performance across various personality traits. While this research does not yet match the results of SOTA studies in this

field, it represents the SOTA for this particular dataset and outperforms some of the existing literature (Bogomolov et al., 2014; Kang et al., 2023; Spathis et al., 2019). The study provides a framework for affect state detection, utilising feature extraction techniques and models employed by existing literature. However, the models used in this paper are not yet viable for real-world applications, showing notable error patterns. Nonetheless, given the importance of affect state detection, further research into unobtrusive tracking techniques involving smartphone and smartwatch data, coupled with the techniques applied in this study, is of societal benefit. These techniques offer the potential for enhanced mental health monitoring, personalised interventions, and increased accessibility of affect state detection technology through unobtrusive tracking. In conclusion, this research lays the groundwork for advancements in affect state detection and its practical applications, ultimately contributing to societal well-being.

8 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

Data Source: The K-EmoPhone dataset has been acquired from https://zenodo.org through an online request. The original owner of the data used in this thesis (Kang et al., 2023) retains ownership of the data during and after the completion of this thesis. The authors of this study acknowledge that they do not have any legal claim to the data. The obtained data is anonymised. Work on this thesis did not involve collecting data from human participants or animals. All the figures belong to the author; when figures were inspired by others, it is explicitly stated. Grammarly and a generative language model (ChatGPT, "OpenAI" (2024)) was used for spelling and grammar corrections. yes and the

REFERENCES

1,916 suicide deaths in 2022, 54 more than in 2021 [Last Modified: 2023-05-15T15:00:00+02:00]. (2023, May). Retrieved May 18, 2024, from https://www.cbs.nl/en-gb/news/2023/19/1-916-suicide-deaths-in-2022-54-more-than-in-2021

Aalbers, G., Hendrickson, A. T., Abeele, M. M. V., & Keijsers, L. (2023). Smartphone-Tracked Digital Markers of Momentary Subjective Stress in College Students: Idiographic Machine Learning Analysis [Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada]. *JMIR mHealth and uHealth*, *11*(1), e37469. https://doi.org/10.2196/37469

Aalbers, G., vanden Abeele, M. M. P., Hendrickson, A. T., de Marez, L., & Keijsers, L. (2022). Caught in the moment: Are there person-specific associations between momentary procrastination and passively measured smartphone use? [Publisher: SAGE Publications]. *Mobile Media & Communication*, *10*(1), 115–135. https://doi.org/10.1177/2050157921993896

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–283.

Acikmese, Y., & Alptekin, S. E. (2019). Prediction of stress levels with LSTM and passive mobile sensors. *Procedia Computer Science*, *159*, 658–667. https://doi.org/10.1016/j.procs.2019.09.221

Alikhanov, J., Zhang, P., Noh, Y., & Kim, H. (2024). Design of Contextual Filtered Features for Better Smartphone-User Receptivity Prediction [Conference Name: IEEE Internet of Things Journal]. *IEEE Internet of Things Journal*, *11*(7), 11707–11722. https://doi.org/10.1109/JIOT.2023.3331715

Asare, K. O., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., & Ferreira, D. (2021). Predicting Depression From Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study [Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada]. *JMIR mHealth and uHealth*, *9*(7), e26540. https://doi.org/10.2196/26540

Berkman, L. F., Kawachi, I., & Glymour, M. M. (2014). *Social Epidemiology* [Google-Books-ID: qHpYCwAAQBAJ]. Oxford University Press.

Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., Alex, & Pentland. (2014). Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*. https://doi.org/10.1145/2647868.2654933

Bulagang, A. F., Mountstephens, J., & Teo, J. (2021). Multiclass emotion prediction using heart rate and virtual reality stimuli. *Journal of Big Data*, *8*(1), 12. https://doi.org/10.1186/s40537-020-00401-x

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System [arXiv:1603.02754 [cs]]. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
Comment: KDD'16 changed all figures to type1.

Chollet, F., et al. (2015). *Keras*. https://github.com/fchollet/keras

Daniyal, M., Javaid, S. F., Hassan, A., & Khan, M. A. B. (2022). The Relationship between Cellphone Usage on the Physical and Mental Wellbeing of University Students: A Cross-Sectional Study [Number: 15 Publisher: Multidisciplinary Digital Publishing Institute]. *International Journal of Environmental Research and Public Health*, *19*(15), 9352. https://doi.org/10.3390/ijerph19159352

Fish, R. M., & Geddes, L. A. (2009). Conduction of Electrical Current to and Through the Human Body: A Review. *Eplasty*, *9*, e44. Retrieved June 23, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2763825/

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2019). Review on Psychological Stress Detection Using Biosignals. *IEEE Transactions on Affective Computing*, *PP*, 1–1. https://doi.org/10.1109/TAFFC.2019.2927337

Gjoreski, M., Luštrek, M., Gams, M., & Gjoreski, H. (2017). Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics*, *73*, 159–170. https://doi.org/10.1016/j.jbi.2017.08.006

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A Search Space Odyssey [arXiv:1503.04069 [cs]]. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924
Comment: 12 pages, 6 figures.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

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

Howard, J., & Gugger, S. (2020). Deep Learning for Coders with Fastai and PyTorch.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(3), 90–95.

Jaques, N., Taylor, S., Azaria, A., Ghandeharioun, A., Sano, A., & Picard, R. (2015). Predicting students' happiness from physiology, phone, mobility, and behavioral data. *International Conference on Affective Computing and Intelligent Interaction and workshops : [proceedings]. ACII (Conference)*, *2015*, 222–228. https://doi.org/10.1109/ACII.2015.7344575

Joyce, J. (2021). Bayes' Theorem. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. Retrieved May 14, 2024, from https://plato.stanford.edu/archives/fall2021/entries/bayes-theorem/

Kandemir, M., Vetek, A., Gönen, M., Klami, A., & Kaski, S. (2014). Multi-task and multi-view learning of user state. *Neurocomputing*, *139*, 97–106. https://doi.org/10.1016/j.neucom.2014.02.057

Kang, S., Choi, W., Park, C. Y., Cha, N., Kim, A., Khandoker, A. H., Hadjileontiadis, L., Kim, H., Jeong, Y., & Lee, U. (2023). K-EmoPhone: A Mobile and Wearable Dataset with In-Situ Emotion, Stress, and Attention Labels. *Scientific Data*, *10*(1), 351. https://doi.org/10.1038/s41597-023-02248-2

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, *139*(4), 917–940. https://doi.org/10.1037/a0030811

Kyriakou, K., Resch, B., Sagl, G., Petutschnig, A., Werner, C., Niederseer, D., Liedlgruber, M., Wilhelm, F. H., Osborne, T., & Pykett, J. (2019). Detecting Moments of Stress from Measurements of Wearable Physiological Sensors [Number: 17 Publisher: Multidisciplinary Digital Publishing Institute]. *Sensors*, *19*(17), 3805. https://doi.org/10.3390/s19173805

Lee, C. M., Jin, S.-P., Doh, E. J., Lee, D. H., & Chung, J. H. (2019). Regional Variation of Human Skin Surface Temperature [Publisher: Korean Dermatological Association]. *Annals of Dermatology*, *31*(3), 349. https://doi.org/10.5021/ad.2019.31.3.349

Lee, H., Young Sang Choi, Sunjae Lee, & Park, I. P. (2012). Towards unobtrusive emotion recognition for affective social communication. *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, 260–264. https://doi.org/10.1109/CCNC.2012.6181098

Likamwa, R., Liu, Y., Lane, N., & Zhong, L. (2013, June). *MoodScope: Building a Mood Sensor from Smartphone Usage Patterns* [Journal Abbreviation: MobiSys 2013 - Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services Publication Title: MobiSys 2013 - Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services]. https://doi.org/10.1145/2462456.2464449

Lupien, S. J., McEwen, B. S., Gunnar, M. R., & Heim, C. (2009). Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nature Reviews Neuroscience*, *10*(6), 434–445. https://doi.org/10.1038/nrn2639

McKinney, W., et al. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Murray, C., Chaurasia, P., Hollywood, L., & Coyle, D. (2022). A Comparative Analysis of State-of-the-Art Time Series Forecasting Algorithms. *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, 89–95. https://doi.org/10.1109/CSCI58124.2022.00021

Oğuz, A., & Ertuğrul, Ö. F. (2023, January). Chapter 1 - Introduction to deep learning and diagnosis in medicine. In K. Polat & S. Öztürk (Eds.), *Diagnostic Biomedical Signal and Image Processing Applications with Deep Learning Methods* (pp. 1–40). Academic Press. https://doi.org/10.1016/B978-0-323-96129-5.00003-2

OpenAI. (2024). Retrieved June 29, 2024, from https://openai.com/

Osmani, V., Ferdous, R., & Mayora, O. (2015). Smartphone app usage as a predictor of perceived stress levels at workplace. *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*. https://doi.org/10.4108/icst.pervasivehealth.2015.260192

Osotsi, A., Oravecz, Z., Li, Q., Smyth, J., & Brick, T. R. (2020). Individualized Modeling to Distinguish Between High and Low Arousal States Using Physiological Data. *Journal of Healthcare Informatics Research*, *4*(1), 91–109. https://doi.org/10.1007/s41666-019-00064-1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, *17*(3), 715–734. https://doi.org/10.1017/S0954579405050340

Romeo, L., Cavallo, A., Pepa, L., Berthouze, N., & Pontil, M. (2019). Multiple Instance Learning for Emotion Recognition Using Physiological Signals. *IEEE Transactions on Affective Computing*, *PP*, 1–1. https://doi.org/10.1109/TAFFC.2019.2954118

Russell, J. A. (1980). A circumplex model of affect [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Sano, A., Taylor, S., McHill, A. W., Phillips, A. J., Barger, L. K., Klerman, E., & Picard, R. (2018). Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study. *Journal of Medical Internet Research*, *20*(6), e210. https://doi.org/10.2196/jmir.9410

Schmidt, P., Dürichen, R., Reiss, A., Van Laerhoven, K., & Plötz, T. (2019). Multi-target affect detection in the wild: An exploratory study. *Proceedings of the 23rd International Symposium on Wearable Computers*, 211–219. https://doi.org/10.1145/3341163.3347741

Sidoti, O., Gelles-Watnick, R., Faverio, M., Atske, S., Radde, K., & Park, E. (2024). Mobile Fact Sheet. Retrieved March 1, 2024, from https://www.pewresearch.org/internet/fact-sheet/mobile/

Siirtola, P., Tamminen, S., Chandra, G., Ihalapathirana, A., & Röning, J. (2023). Predicting Emotion with Biosignals: A Comparison of Classification and Regression Models for Estimating Valence and Arousal Level Using Wearable Sensors [Number: 3 Publisher: Multidisciplinary Digital Publishing Institute]. *Sensors*, *23*(3), 1598. https://doi.org/10.3390/s23031598

Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., & Rentfrow, J. (2019). Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2886–2894. https://doi.org/10.1145/3292500.3330730

Stawarz, K., Preist, C., & Coyle, D. (2019). Use of Smartphone Apps, Social Media, and Web-Based Resources to Support Mental Health and Well-Being: Online Survey. *JMIR Mental Health*, *6*(7), e12546. https://doi.org/10.2196/12546

Suhara, Y., Xu, Y., & Pentland, A. (2017, April). *DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks* [Pages: 724]. https://doi.org/10.1145/3038912.3052676

Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). Automatic Identification of Artifacts in Electrodermal Activity Data. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2015*, 1934–1937. https://doi.org/10.1109/EMBC.2015.7318762

Taylor, S., Jaques, N., Nosakhare, E., Sano, A., & Picard, R. (2020). Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE transactions on affective computing*, *11*(2), 200–213. https://doi.org/10.1109/TAFFC.2017.2784832

Tran, T.-D., Kim, J., Ho, N.-H., Yang, H.-J., Pant, S., Kim, S.-H., & Lee, G.-S. (2021). Stress Analysis with Dimensions of Valence and Arousal in the Wild [Number: 11 Publisher: Multidisciplinary Digital Publishing Institute]. *Applied Sciences*, *11*(11), 5194. https://doi.org/10.3390/app11115194

Umematsu, T., Sano, A., & Picard, R. W. (2019). Daytime Data and LSTM can Forecast Tomorrow's Stress, Health, and Happiness [ISSN: 1558-4615]. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2186–2190. https://doi.org/10.1109/EMBC.2019.8856862

Vahedi, Z., & Saiphoo, A. (2018). The association between smartphone use, stress, and anxiety: A meta-analytic review [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/s *Stress and Health*, *34*(3), 347–358. https://doi.org/10.1002/smi.2805

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 3–14. https://doi.org/10.1145/2632048.2632054

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., … Qalieh, A. (2017, September). *Mwaskom/seaborn: Vo.8.1 (september 2017)* (Version vo.8.1). Zenodo. https://doi.org/10.5281/zenodo.883859

Yamak, P. T., Yujian, L., & Gadosey, P. K. (2019). A Comparison between ARIMA, LSTM, and GRU for Time Series Forecasting. *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, 49–55. https://doi.org/10.1145/3377713.3377722

Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., & Atkinson, P. M. (2018). A hybrid MLP-CNN classifier for very fine resolution remotely

sensed image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *140*, 133–144. https://doi.org/10.1016/j.isprsjprs.2017.07.014

Zhang, X., Li, W., Chen, X., & Lu, S. (2018). MoodExplorer: Towards Compound Emotion Detection via Smartphone Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(4), 176:1–176:30. https://doi.org/10.1145/3161414

Zhang, Y., & Yang, Q. (2022). A Survey on Multi-Task Learning [Conference Name: IEEE Transactions on Knowledge and Data Engineering]. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203

APPENDICES

9 APPENDIX A

Abbreviation	Meaning
RF	Random Forest
XGboost	eXtreme Gradient Boosting
SVM	Support Vector Machine
RNN	Recurrent Neural Network
LSTM	Long-short term Memory
GRU	Gated Recurrent Unit
ESM	Experience Sampling Method
SOTA	State-of-the-art
SHAP	SHapley Additive exPlanations

Table 7: Abbreviations used throughout this paper and their respective meanings

10 APPENDIX B

Hyperparameter	Values
n_estimates	[100, 200, 300, 500, 1000, 1500]
max_depth	[5, 10, 15, 30, 50]
min_samples_split	[2, 5, 10, 20]
min_samples_leaf	[1, 2, 4, 10]
bootstrap	[True, False]

Table 8: Hyperparameter Space for Random Forest

Hyperparameter	Values
n_estimates	[100, 200, 300, 500, 1000]
max_depth	[2, 4, 8, 10, 15, 20]
learning_rate	[0.001, 0.01, 0.1, 1]
subsample	[0.8, 0.5, 0.2]

Table 9: Hyperparameter Space for XGBoost

Hyperparameter	Values
C	[0.1, 1, 10, 100]
gamma	[scale, auto]

Table 10: Hyperparameter Space for Support Vector Machine

11 APPENDIX C

Affect State	RF Best Hyperparameters
Stress	n_estimators: 1500 max_depth: 50 min_samples_split: 2 min_samples_leaf: 1 bootstrap: True
Valence	n_estimators: 1000 max_depth: 50 min_samples_split: 5 min_samples_leaf: 2 bootstrap: True
Arousal	n_estimators: 1000 max_depth: 30 min_samples_split: 5 min_samples_leaf: 4 bootstrap: True

Table 11: Best hyperparameters for Random Forest model

Affect State	XGBoost Best Hyperparameters
Stress	n_estimators: 1000 max_depth: 20 learning_rate: 0.1 subsample: 0.8
Valence	n_estimators: 1000 max_depth: 15 learning_rate: 0.1 subsample: 0.5
Arousal	n_estimators: 1000 max_depth: 20 learning_rate: 0.01 subsample: 0.8

Table 12: Best hyperparameters for XGBoost model

Affect State	SVM Best Hyperparameters
Stress	gamma: auto C: 1
Valence	gamma: scale C: 1
Arousal	gamma: scale C: 1

Table 13: Best hyperparameters for SVM model

Affect State	LSTM Best Hyperparameters
Stress	LSTM units: 128 Batch size: 64 Dropout rate: 0.2
Valence	LSTM units: 256 Batch size: 64 Dropout rate: 0.2
Arousal	LSTM units: 128 Batch size: 128 Dropout rate: 0.1

Table 14: Best hyperparameters found for the LSTM

Affect State	GRU Best Hyperparameters
Stress	GRU units: 128 Batch size: 64 Dropout rate: 0.1
Valence	GRU units: 256 Batch size: 128 Dropout rate: 0.4
Arousal	GRU units: 256 Batch size: 128 Dropout rate: 0.2

Table 15: Best hyperparameters found for the GRU

12 APPENDIX D

Table 16: Survey Questions and Responses

My emotion right before doing this survey was		
Q1. very negative (−3)	∼	very positive (+3)
My emotion right before doing this survey was		
Q2. very calm (−3)	∼	very excited (+3)
My attention level to my ongoing task right before doing this survey could be rated as		
Q3. very bored (−3)	∼	very engaged (+3)
My stress level right before doing this survey was		
Q4. not stressed at all (−3)	∼	very stressed (+3)
My emotion that I answered above has not changed for recent __ minutes.		
[5, 10, 15, 20, 30, 60 min/I am not sure]		
Answering this survey disturbed my ongoing task		
Q6. not disturbed at all (−3)	∼	very disturbed (+3)
How did your emotions change while you are answering the survey now?		
Q7. I felt more negative (−3)	∼	I felt more positive (+3)

Table 17: 7 point Likert scale questions participants answered for every ESM prompt. Source: author's illustration based on Kang et al. (2023)

13 APPENDIX E

Categories	
PERSONALIZATION	COMMUNICATION
	PHOTOGRAPHY
SYSTEM	TOOLS
FINANCE	PRODUCTIVITY
HEALTH AND FITNESS	VIDEO PLAYERS
MISC	TRAVEL AND LOCAL
MAPS AND NAVIGATION	MUSIC AND AUDIO
LIFESTYLE	HOUSE AND HOME
SOCIAL	GAME
ART AND DESIGN	SHOPPING
WEATHER	EDUCATION
FOOD AND DRINK	NEWS AND MAGAZINES
ENTERTAINMENT	BOOKS AND REFERENCE
SPORTS	BUSINESS
COMICS	BEAUTY
LIBRARIES AND DEMO	AUTO AND VEHICLES

Table 18: Full lit of app categories tracked