
Using Machine Learning to Develop Algorithms to Perform Mood Classification in Real-time

Henry Liu

A thesis submitted in fulfilment of the requirements for the degree of
Masters of Engineering in Computer Systems,
The University of Auckland, 2022.

Abstract

Globally, depression is a leading cause of disability and impaired quality of life. The COVID-19 global pandemic has seen an increase of mental health disorders particularly depression, which often goes untreated. Several key barriers contribute to this, including social stigma, difficulty accessing mental health services as they are overwhelmed, and lack of timely objective assessment approaches. A system that helps monitor mood changes on a 24/7 basis will help identify when someone may be starting to develop depression and therefore aid with earlier diagnosis and relapse prediction.

The current literature has shown traditional machine learning techniques such as regression and ensemble learning to be effective in tackling the problem of mood classification and prediction. In this work, we use deep learning and neural network approaches to tackle this problem with the aim of improving we can improve the accuracy and lowering the variance of existing methods.

We acquired a dataset from an existing study (of $n = 14$ participants) using regression and ensemble learning techniques and developed our own Neural Network model using multilayer perceptron models to tackle the classification task. We tried a neural network approach across multiple subsets of their dataset with varying success. Our best model performed on par with the existing Shah et al model with 17 out of 32 total measurements improving on that of Shah et al and providing consistently provided lower variance than their model.

Additionally, where we collected a similar dataset using technology and sensors from smartphones and smartwatches for one month. A total of 15 healthy individuals participated in the study. This data was then used in traditional monolithic and compositional neural network models to perform mood classification using smartphone-enhanced ecological momentary assessments and physiological data collected from a Fitbit smartwatch.

The compositional and monolithic neural network approaches developed using Keras provided promising results with accuracies ranging from roughly 60 -

90%. However, there were limitations to our collected dataset as the distribution of labels was limited and may have resulted in over-fitting. This was expected as all participants were healthy controls, as it was challenging to recruit a depressed group due to the COVID-19 pandemic. Future work can be done by adding extra features and samples to improve the neural network models but overall this work showed that the deep learning approach has the potential for accurate mood prediction.

Acknowledgements

If you were to ask me where I thought I would be five years ago, I definitely wouldn't have been able to guess that I would be completing a Master's degree. Before I continue with the body of work, I must thank a wide range of people who have provided me with unconditional support throughout my life and academic journey; without them, this thesis would not be possible. Sir Issac Newton once said in a letter, "If I have seen further it is by standing on the shoulders of Giants" the people mentioned below are my giants.

Firstly, I must thank my supervisors, Dr Partha S. Roop and Dr Frederick Sundram. I must first start by apologising to you both, and I know I may not have been the easiest student to supervise at times. Thank you, Partha, for the guidance you have provided me throughout this journey, from back during part IV to now. You were one of the main reasons I decided to continue my studies. You constantly pushed me to provide high-quality work and provided me with a collaborative and supportive environment. Thank you, Fred, for everything you have done for me during this journey. The feedback and guidance you have provided me have been invaluable. I would also like to thank you for all the hours we worked together to set up the study and screen participants. Without your help, I wouldn't have collected the data necessary to write parts of this thesis.

Secondly, I would like to thank the MoodAI and PRETgroup research teams for their guidance and feedback during my journey. In particular, I would like to mention and thank both Aron Jeremiah, and Sobhan Chatterjee from the Department of Engineering and Dr Amy Chan from the School of Pharmacy. Aron, it's been a pleasure working with you for the last three years. It's been a joy to see to grow and develop your skills as an engineer. I would also like to thank you, Aron, for your help and contribution to the MoodAI project and my thesis. I wish you all the best in your pursuit of your Master's degree and beyond. I know you will succeed in anything you decide to do. I want to thank Sobhan for the invaluable help and guidance he provided me relating to the machine learning aspects of this

thesis. Without your help and advice, those parts would not have been possible. Finally, to Amy, thank you for all the invaluable feedback and guidance you provided me during this journey. I would also like to congratulate you on becoming a mother. I wish you and your family all the best in whatever the future has planned.

Finally, I must thank my family and friends for their unconditional support and company throughout my academic journey. I want to thank my mother Mei and father Yan for everything you have done to provide for me and the unconditional love and support you have provided me; thank you both for making all of this possible. To my sister Ashley, thanks for your love and support. It's been a pleasure watching you grow into the magnificent you are today. I wish you all the best, and I know you will succeed in whatever you decide to study at University. The last couple of years has felt somewhat isolated due to the social distancing rules to combat the global pandemic. I found myself working and studying from home, physically separated from friends and colleagues. In particular, I would like to mention Joshua Holroyd, Joseph Matthews, Kevin Yi, Hamish Douglass, Byron Lam, Martin Huang, and Tom Stevenson. The companionship and support from you all was the most important to me during my academic journey; without it, I couldn't have said I would have made it through.

Thanks again to everyone who supported me throughout this journey,

Henry Liu

Contents

LIST OF FIGURES	xii
LIST OF TABLES	xii
LIST OF ACRONYMS	xiv
1 INTRODUCTION	1
1.1 Background	2
1.2 Digital Bio-markers	2
1.2.1 Heart Rate and Heart Rate Variability (HRV)	2
1.2.2 Physical Activity	3
1.2.3 Sleep	4
1.3 Machine Learning (ML)	4
1.4 Regression	5
1.5 Ensemble Learning	5
1.6 Deep Learning (DL)	6
1.7 Artificial Neural Networks (ANN)	6
1.8 Multilayer Perceptron (MLPs)	7
1.9 Recurrent Neural Networks (RNNs)	7
1.10 Long Short-Term Memory (LSTMs)	7
1.11 Compositional Neural Networks (C _p NNs)	9
1.12 Cross-Validation	10
1.13 Hyper-Parameter Optimisation	10
1.14 Thesis Outline	10
2 LITERATURE REVIEW AND EXISTING WORK	13
2.1 Existing Work	13
2.2 Shah et al	14

	2.2.1	<i>Introduction</i>	14
	2.2.2	<i>Study Design</i>	14
	2.2.3	<i>Study Recruitment</i>	15
	2.2.4	<i>Study Procedures</i>	15
	2.2.5	<i>Study Methodology</i>	16
	2.2.6	<i>Study Results</i>	17
	2.2.7	<i>Conclusions</i>	17
2.3		Bai et al (Mood Mirror)	18
	2.3.1	<i>Introduction</i>	18
	2.3.2	<i>Study Design</i>	18
	2.3.3	<i>Study Recruitment</i>	18
	2.3.4	<i>Study Procedures</i>	19
	2.3.5	<i>Study Methodology</i>	19
	2.3.6	<i>Study Results</i>	20
	2.3.7	<i>Conclusions</i>	20
2.4		Rykov et al	21
	2.4.1	<i>Introduction</i>	21
	2.4.2	<i>Study Recruitment</i>	21
	2.4.3	<i>Study Procedures</i>	21
	2.4.4	<i>Study Methodology</i>	21
	2.4.5	<i>Study Results</i>	22
	2.4.6	<i>Conclusions</i>	22
2.5		Current Gaps	22
2.6		Problem Statement	23
2.7		Research Contributions	23
3		MOODAI STUDY	25
	3.1	Ethics Committee Review and Approval	25
	3.2	Registration with the Australian New Zealand Clinical Trials Registry	27
	3.3	Study Funding	27
	3.4	Study Design	27
	3.5	Study Recruitment	28
	3.6	Study Inclusion/Exclusion Criteria	28
	3.7	Pre-screening Process	29
	3.8	Study Screening	31
	3.9	Choice of devices	33
	3.10	Study Procedure	34

3.11	Recruitment Results	35
3.12	COVID-19 Disruptions	35
3.13	MoodAI Website Design	36
	3.13.1 Purpose	36
	3.13.2 Framework	36
	3.13.3 Participant Access	36
3.14	Data collected	38
	3.14.1 Ecological Momentary Assessments (EMAs)	38
	3.14.2 End of Day Journal	38
	3.14.3 Physiological Data	41
	3.14.4 Audio diaries	43
	3.14.5 MoodAI Architecture Summary	43
3.15	Participant Feedback and Acceptability	44
4	SHAH ET AL DATASET	49
4.1	Introduction	49
4.2	Data preparation	49
4.3	Study Methodology Comparison	51
4.4	Machine Learning Models	52
4.5	Model Results Comparison	54
	4.5.1 Entire Data Set Results	54
	4.5.2 Subset Results	56
4.6	Shah et al Conclusions	57
5	MOODAI DATA SET AND MACHINE LEARNING PIPELINE	59
5.1	Data Retrieval	59
5.2	Feature Extraction	60
	5.2.1 Activity Cluster	60
	5.2.2 Sleep Cluster	60
	5.2.3 Heart Cluster	60
5.3	Data Preparation	60
5.4	Exploratory data analysis	64
5.5	Machine learning models	65
	5.5.1 Multi-feature Monolithic Model	66
	5.5.2 Multi-feature Compositional Model	67
6	CONCLUSIONS AND FUTURE WORK	71

6.1	Limitations	71
6.2	Strengths	72
6.3	Future Work	73
	6.3.1 <i>Mood Index</i>	73
	6.3.2 <i>Audio and Speech Analysis</i>	73
	6.3.3 <i>Platform Agnostic</i>	73
	6.3.4 <i>Embedded Implementation</i>	74
	6.3.5 <i>MoodAI Platform Availability</i>	74
6.4	Summary of Thesis	74
A	SHAH ET AL SUPPLEMENTARY INFORMATION	77
B	PARTICIPANT INFORMATION SHEET (PIS)	79
C	PATIENT HEALTH QUESTIONNAIRE 9 (PHQ-9)	89
D	ALCOHOL USE DISORDERS IDENTIFICATION TEST (AUDIT)	91
E	COLUMBIA-SUICIDE SEVERITY RATING SCALE (C-SSRS)	93
F	MINI-INTERNATIONAL NEUROPSYCHIATRIC INTERVIEW (M.I.N.I)	95
G	MONTGOMERY-ÅSBERG DEPRESSION RATING SCALE (MADRS)	97
H	TECHNOLOGY ACCEPTANCE MODEL QUESTIONNAIRE (TAM-Q)	99
I	SHAH ET AL MODEL COMPARISONS	103
	BIBLIOGRAPHY	115

List of Figures

1.1	A Basic Multilayer Perceptron Neural Network	8
1.2	A Simple Recurrent Neural Network	9
3.1	Summary of entire study process	26
3.2	Study provided OnePlus Nord Smartphone	34
3.3	Study provided Fitbit Sense Smartwatch	34
3.4	MoodAI website login page	37
3.5	MoodAI Neutral Ecological Momentary Assessment	39
3.6	MoodAI Negative Ecological Momentary Assessment	40
3.7	MoodAI Positive Ecological Momentary Assessment	41
3.8	MoodAI Daily End of Day Journal	42
3.9	MoodAI Daily Audio Diary Question 1	44
3.10	MoodAI Daily Audio Diary Question 2	45
3.11	MoodAI Daily Audio Diary Question 3	45
3.12	Overview of the MoodAI System Architecture	46
5.1	Distribution of EMAs across all participants	64
5.2	Proposed MoodAI Compositional Neural Network Model	69
I.1	Comparison of Mean MAPE over different folds using subset of Shah et al data	104
I.2	Comparison of Std MAPE over different folds using subset of Shah et al data	105
I.3	Comparison of Mean MAE over different folds using subset of Shah et al data	106
I.4	Comparison of Std MAE over different folds using subset of Shah et al data	107

List of Tables

2.1	Results of the Shah et al’s best machine learning model for each participant	17
2.2	Best Mood Mirror Results using all collected features	20
3.1	Participant feedback about MoodAI Study	47
4.1	Number of samples for each participant after data preparation	51
4.2	Table of selected subset of features and description	53
4.3	Comparison of best Shah et al regression based machine learning models using 3 fold cross validation MoodAI deep learning model using Shah et al data set	55
4.4	Comparison of best Shah et al regression based machine learning models using 4 fold cross validation MoodAI deep learning model using Shah et al data set	56
4.5	Comparison of best Shah et al regression based machine learning models using 4 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set	58
5.1	Table of MoodAI Activity Cluster features and descriptions	61
5.2	Table of MoodAI Sleep Cluster features and descriptions	62
5.3	Table of MoodAI Heart Cluster features and descriptions	62
5.4	Number of samples for each participant in the MoodAI dataset after data preparation	63
5.5	Distribution of EMA scores across all participants	65
5.6	Results from Tuned MoodAI Monolithic LSTM Model	67
5.7	Results from Tuned MoodAI Activity Cluster LSTM Model	68
5.8	Results from Tuned MoodAI Sleep Cluster LSTM Model	68
5.9	Results from Tuned MoodAI Heart Cluster LSTM Model	70
5.10	Results from Tuned MoodAI Compositional Decision MLP Model	70

I.1	Comparison of best Shah et al regression based machine learning models using 5 fold cross validation MoodAI deep learning model using Shah et al data set	108
I.2	Comparison of best Shah et al regression based machine learning models using 8 fold cross validation MoodAI deep learning model using Shah et al data set	109
I.3	Comparison of best Shah et al regression based machine learning models using 10 fold cross validation MoodAI deep learning model using Shah et al data set	110
I.4	Comparison of best Shah et al regression based machine learning models using 3 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set	111
I.5	Comparison of best Shah et al regression based machine learning models using 5 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set	112
I.6	Comparison of best Shah et al regression based machine learning models using 8 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set	113
I.7	Comparison of best Shah et al regression based machine learning models using 10 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set	114

List of Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUDIT	Alcohol Use Disorders Identification Test
BPM	Beats Per Minute
CNN	Convolutional Neural Network
C _p NN	Compositional Neural Network
C-SSRS	Columbia-Suicide Severity Rating Scale
CSV	Comma Separated Variables
DL	Deep Learning
DSM	Diagnostic and Statistical Manual of Mental Disorders
EEG	Electroencephalography
EMA	Ecological Momentary Assessment
GPU	Graphics processing unit
HAM-D	Hamilton Depression Rating Scale
HRV	Heart Rate Variability
IBI	Inter-beat Interval
KNN	K-nearest neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short-Term Memory
MADRS	Montgomery–Åsberg Depression Rating Scale
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDD	Major Depressive Disorder
M.I.N.I	Mini-International Neuropsychiatric Interview
ML	Machine Learning
MLP	Multilayer Perceptron
NaN	Not a Number
NLP	Natural Language Processing
PHQ-9	Patient Health Questionnaire 9
PIS	Participant Information Sheet
PPG	Photoplethysmography
ReLU	Rectified Linear Unit
REM	Rapid Eye Movement
RMSSD	Root Mean Squared of Successive Differences
RNN	Recurrent Neural Network
Std	Standard Deviation
SVM	Support Vector Machine
TAM-Q	Technology Acceptance Model Questionnaire
UoA	University of Auckland
XGBoost	Extreme Gradient Boosting

Chapter 1

Introduction

Globally, depression is highly prevalent and a leading cause of disability and impaired quality of life [1] [2]. Depression is the most prevalent mental health disorder in New Zealand [3] and is the second leading cause of disability globally [4]. Depression not only has a burden on people but also on the economy, with an estimated cost of 210 billion USD per year [5].

Despite depression's broad reach and availability of effective treatments, it often goes untreated. Several key barriers contribute to this, including social stigma, difficulty accessing mental health services, and lack of timely objective assessment approaches. This has been further exacerbated by an increase in psychological distress due to the COVID-19 pandemic [6] [7].

A system that helps monitor mood changes on a 24/7 basis will help identify when someone may be starting to develop mental health disorders, and then immediately provide support and resources that could prove valuable for early diagnosis and detection of mental health disorders. Capable smartphones are being more readily available, an average of 73.45% [8] of adults in the top 20 countries owned a smartphone. Wearable technology has become more readily available and with advancements in sensor technology wearables could aid in reshaping the healthcare industry [9]. Smartwatch technology allows sensor data to be used for the detection of different health illnesses, both physical and mental [10].

Studies have shown there is a link between depression, anxiety and other mental health disorders and a variety of different biomarkers. Biomarkers such as heart rate and heart rate variability have been known to reflect sympathetic and parasympathetic activity and mental illness tends to cause an imbalance in the form of increased sympathetic activity [11]. This is reflected in decreased HRV and increased resting heart rate [12] [13] [14].

With a larger dataset machine learning can be an extremely powerful tool for data processing. ML has been used for applied to areas such as natural language processing, and time series forecasting. Recent developments in ML [15] and the widespread adoption of affordable smartphones and smartwatches will facilitate technology as a lever to support objective mental health diagnosis and monitoring on a global scale [16].

1.1 Background

There are several terms and concepts that need to be defined and are essential in understanding this thesis. These include the concepts of:

- Digital biomarkers and their relation to mental wellness
- Different machine learning models and their applications

1.2 Digital Bio-markers

Digital bio-markers are measurable health indicators that can be used collected and then analysed to infer different physical and mental health conditions. Existing studies have explored the idea of using digital bio-markers for the classification of depression and results indicate that they have been shown to be effective [17]. Many studies have used emerging smartwatch technology to capture digital biomarkers such as heart rate, heart rate variability, sleep activity, and physical activity. A study conducted by Byun et al [18] in 2019 determined with a 74.4% accuracy that participants with major depressive disorder could be differentiated from healthy control participants. Some digital biomarkers that have been shown to be predictors of depressive relapse include disrupted sleep, reduced sociability, changes in mood, prosody and cognitive function [19]. Overall, digital biomarkers can provide an scalable, unobtrusive, time-sensitive, and cost-effective method for depression detection [20].

1.2.1 Heart Rate and Heart Rate Variability (HRV)

Heart rate is a well-known term that indicates how frequently your heart beats. Heart rate is commonly measured in beats per minute (BPM) and can range from 60 - 100 BPM in normal human adults.

Given a BPM heart rate measurement it is common to think that a human heart beats at consistent intervals. This is untrue as there is in fact a variation in the beat-to-beat interval and this is classified as heart rate variability (HRV). Heart rate and heart rate variability can provide great insights into the cardiovascular health of a human [21].

There are two central nervous systems, the sympathetic nervous system commands the body's fight or flight response, and the parasympathetic nervous system which controls the body's rest and digest response. In a healthy individual, there must be a balance between these two nervous systems [22]. Individuals with depression and stress-related disorders tend to have an imbalance in the form of increased sympathetic activity [11].

Literature has indicated a clear inverse relationship between long-term mental health issues and HRV [23]. Athletes and individuals who are more physically fit may have healthy heart rates that are lower than the normal 60. Unwell individuals, for example, those with chronic insomnia, will show an increased resting heart rate [24].

A low heart rate variability indicates the heart's inability to respond quickly and increases an individual's susceptibility to sickness and disease, both physiological and psychological whereas a high heart rate variability indicates a strong ability to adapt to physiological changes and helps to resist sickness and disease.

Both HR and HRV relate closely with mental health disorders such as depression and anxiety as generally individuals with mental health illnesses also show an increase in resting HR and a decreased HRV [25]. This makes both HR and HRV effective biomarkers for the detection of depression severity [26].

1.2.2 Physical Activity

Physical activity is a general measure of the exercise and steps an individual performs each day. Physical activity is commonly broken down into step count, intensive exercise, and sometimes calorie intake and expenditure. As discussed in the heart rate and heart rate variability section, a higher level of physical activity can help to improve heart rate variability and lower resting heart rate.

Similar to heart rate and heart rate variability physical activity has also been shown to be negatively correlated with depressive symptoms [27]. Literature has shown that a high level of physical activity reduces the risk of major depressive disorder in an individual [28] [29] [30]. The connection between physical activity

levels and depression symptoms could be related to heart rate and heart rate variability but may be due to the link between the level of sedentary behaviour which is the inverse of physical activity and depressive symptoms [31]. These factors make physical activity an effective digital bio-marker for determining depression severity.

1.2.3 Sleep

Sleep is a measure of the number of hours and frequency of an individual's sleep. This can be split into different segments including, light, deep, and REM sleep. Sleep plays an important part in staying healthy as it allows the brain to carry out many important functions. Sleep is essential to every part of the body and it helps the body fight disease and develop immunity, and reduce the risk of disease.

Sleep disturbance is very common in depressed individuals [32] in particular shortening of REM latency, lengthening of the duration of the first REM period and heightening of REM density [33]. Sleep disturbance in depressed individuals tends to occur in the early hours of the morning between 2 - 4 am [34]. Many studies have already captured information about sleep duration and used them as an effective digital bio-marker for depression severity detection.

1.3 Machine Learning (ML)

Machine learning is a field of computer science that focuses on the development of computer systems that have the ability to learn to perform a task without being explicitly programmed like traditional algorithms. The machine essentially draws inferences that it learns from patterns in the data.

There are three main types of machine learning [35]. The first type is supervised learning, this is when an ML model develops relationships based on looking at a set of input-output examples. The second is unsupervised learning, where the model tries to determine relationships on an input dataset without being provided labels. The final type is reinforcement learning is when the model learns constantly and learns based on feedback from an environment. Other types such as semi-supervised learning, transduction, and learning to learn hybrid combinations of the three base types.

1.4 Regression

Regression in a mathematical context is a statistical way of measuring the relationship between a set of inputs and outputs. In a machine learning context. The most common form of regression is linear regression which tunes the simple equation

$$y = mx + c$$

to be able to as accurately as possible predict an output given a given input [36]. Other forms of regression include non-linear which tries to minimise the error of a more complex function non-linear function given an input-output set. Some common regression algorithms include support vector machines and K nearest neighbours. Support vector machines are a supervised algorithm that creates a plane in a 2D or 3D space given a set of training features and then creates a clear plane that distinctly separates the two or more categories. K nearest neighbours is a non-linear regression algorithm that looks at how new data points are related to existing categories in the dataset. The output is determined by weighting already classified neighbouring points based on the distance of existing points to the new point. These algorithms have been commonly used for classification and market prediction tasks.

1.5 Ensemble Learning

Ensemble learning is a form of machine learning where instead of having one model that is trained to solve one problem, ensemble learning combines multiple to provide better predictive performance [37]. The most common ensemble learning algorithm is the random forest [38] algorithm. The random forest algorithm creates decision trees that are generated on randomly selected data points, this is called bootstrapping. Each tree will have its own output and the classification with the majority vote is then determined to be the result, this process is called aggregation. The randomisation in the algorithm means trees will be trained on features which may be really relevant and some which may not be so relevant, this helps to reduce variance as if all trees had the same features then they would all perform similarly. Generally, the size of the feature subset and the size of the forest result in the algorithm being quite computationally intensive.

1.6 Deep Learning (DL)

Deep learning is a sub-field of machine learning that enables machines to learn from past experiences in order to learn patterns and identify abstract objects [39]. Deep learning was inspired by training a computer to learn how a human brain learns. It is often used interchangeably with Artificial Neural Networks. Deep learning neural algorithms are much more computationally intensive to train than traditional machine learning algorithms and they usually require powerful machines such as high-performing GPUs to be trained.

1.7 Artificial Neural Networks (ANN)

Artificial neural networks are a sub-field of machine learning inspired by the ability of biological systems to process data. Neurons are electrically active cells in the human brain that communicate with thousands of other neurons in the human brain. ANNs create artificial neurons and assign numerical values to recreate biological neurons. The artificial neurons are activated using a range of mathematical functions, these include a simple linear function, to more complex functions such as Tanh, Rectified Linear Unit (ReLU)

Neural networks have shown to be successful in a wide range of applications including classification, pattern recognition, data analysis, and control [40].

The goal of a neural network is to solve complex problems by training and analysing existing examples. A feedforward neural network is essentially a non-linear mathematical function that is capable of mapping a set of inputs to outputs [40] after learning off a training set. A neural network learns from features in the dataset to create a complex abstraction. Neural networks have been shown to have the capability to excel where conventional algorithms could not develop connections.

A limitation of neural networks is that they learn by observation, if there are only limited examples of such classification the neural network will not be able to provide accurate predictions [39]. Another limitation is that ANNs are considered black-box approaches as sometimes they are sufficiently complex that it is difficult to explain or interpret how the model work.

1.8 Multilayer Perceptron (MLPs)

MLPs or also commonly known as deep feed-forward neural networks are the most common type of artificial neural network. Each node in each layer is connected to every node in the following layer. The way inputs are processed at each node and then passed to the next node until the output is generated is why MLPs are a feed-forward neural network.

Training an MLP involves generating outputs with input-output training data. In the first few iterations, the model output will likely be quite different to the actual output, this is represented as some sort of error or loss. This is then fed back via the process of back-propagation to each node so that it can adjust its weight. Many iterations of this is done until the error in the output is minimal.

1.9 Recurrent Neural Networks (RNNs)

Recurrent neural networks are a form of artificial neural networks which involve using the output of one step is provided alongside the inputs for the next step [41]. A simple RNN architecture (as shown in Figure 1.2) consists of an input layer, a number of hidden layers (only one hidden layer in this example), and an output layer (only one output in this example). The architecture of an RNN (See Figure 1.2) is very similar to that of an MLP as seen in Figure 1.1 but with additional arrows in the hidden layer which indicate the outputs from one step being fed back as input to the next time step [42]. Similar to other ANNs, RNNs have been proven successful for tasks such as face detection, speech recognition, prediction problems, and many more.

1.10 Long Short-Term Memory (LSTMs)

A long short-term memory is a type of recurrent neural network with a slight difference where it maintains an internal state or memory so information can be retrieved from relevant previous time steps [43]. The LSTM neural network addresses the problem of the long-term dependency problem of RNNs, which is when a lot of prior information starts piling up, RNNs become less effective at learning new things. The current input, previous output, and the current state are used in the node's calculation [44].

The architecture of an LSTM is similar to that of an RNN (see Figure 1.2 but with the addition of an internal state. There are three gates that control what sort

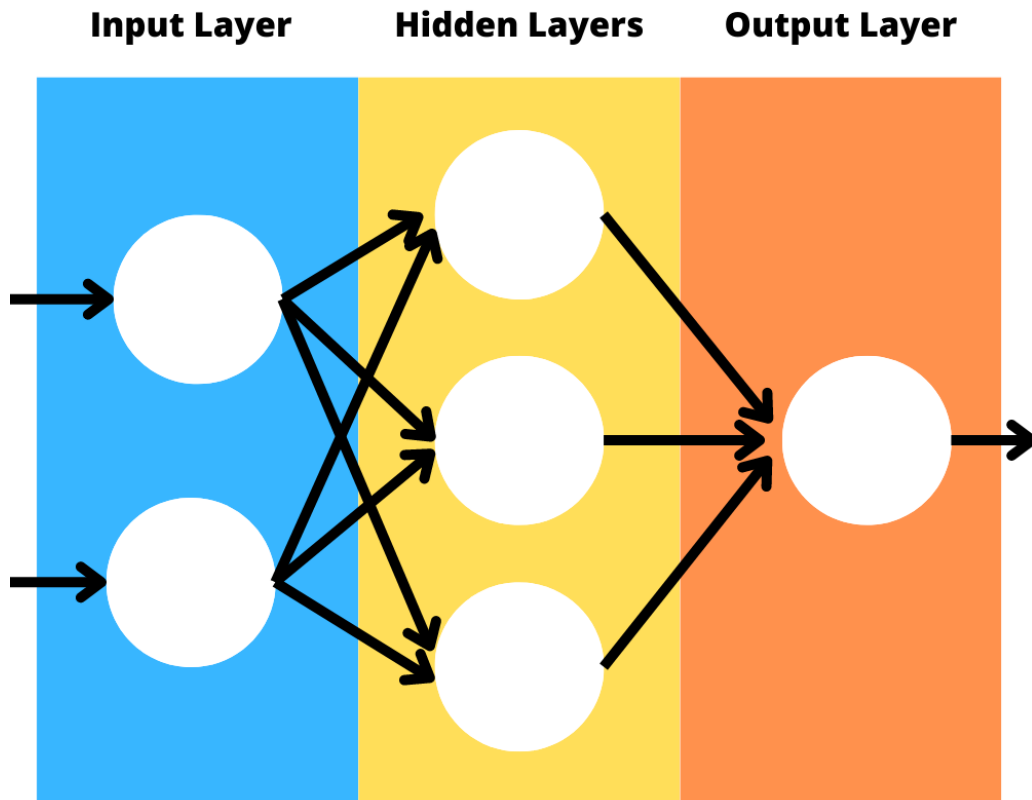


Figure 1.1: A Basic Multilayer Perceptron Neural Network

of information is maintained within the state and these are the forget gate, input gate, and output gate. The forget gate determines what state information stored in the internal state is no longer relevant. The input gate determines which incoming information should be stored in the internal state. The output gate determines what part of the internal state will be used as part of the output. These units allow a neural network to remember the stuff it needs to retain the context of the problem, while also forgetting things that are no longer applicable. LSTMs have proven to be particularly effective in looking at time-series data, natural language processing of text, handwriting and speech and also time series forecasting.

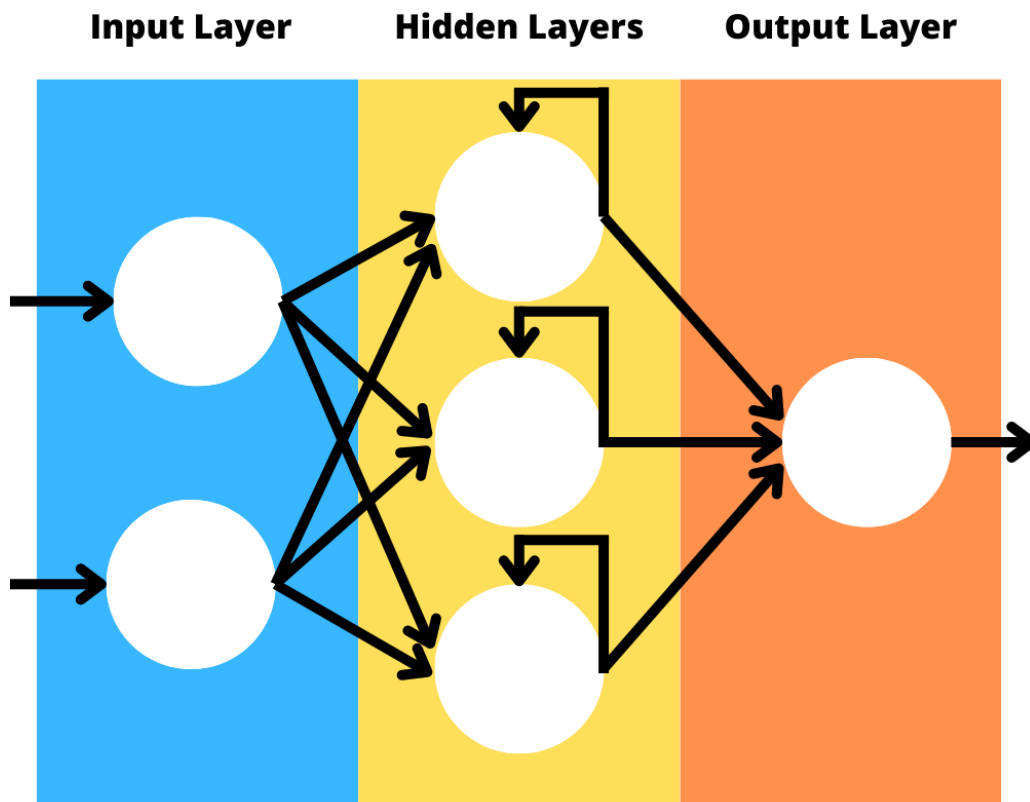


Figure 1.2: A Simple Recurrent Neural Network

1.11 Compositional Neural Networks (C_pNNs)

Traditionally neural networks are trained as a singular large monolithic model to perform a certain task. A limitation of monolithic models is as the number of features grows, the neural network model becomes exponentially hard to train, understand, and explain. The premise of compositional neural networks is to break down large compositional models into smaller models [45]. These smaller models are trained separately and have their own individual outputs, then they are put into a merge block to determine the final output.

1.12 Cross-Validation

Cross-validation is a method of validating different learning algorithms by dividing the dataset into a chosen number of folds. Typically this is divided into k segments and this method is called k -fold cross-validation. The learning algorithm is trained using all combinations of $K - 1$ train folds and 1 test fold. The purpose of performing cross-validation is to gauge the generalisability of the algorithm [46] and to compare different algorithms to the same dataset. Cross-validation allows for the entire dataset to be used in model creation and is an effective method, especially when working with smaller datasets.

1.13 Hyper-Parameter Optimisation

There is a wide range of hyper-parameter optimisation algorithms but the two most simple and commonly used are grid search and random search [47]. Grid search is a brute force method where hyper-parameters are assigned a range in which the grid should search. The hyper-parameters are then adjusted between this range with the desired step size. A model is then created for every combination of every hyper-parameter wanting to be tuned. Grid search is an exhaustive method and is extremely computationally intensive as it scales factorially with step size and number of hyper-parameters to tune. Random search is a hyper-parameter optimisation method that randomly picks parameters within a given space and it does this a predefined amount of times. The idea behind random search is that we don't need to be all points in a similar space as they will provide similar results. This makes the random search much less computationally intensive and has shown to be sufficiently effective for neural network training.

1.14 Thesis Outline

The remainder of this thesis is as follows. Chapter 2 presents a review of existing literature and studies which have applied forms of ML in the mental health space. This is followed by analysing three studies in particular which are some of the most recent and relevant work in the space. Finally, we discuss a summary of the existing work and the needs gaps that this thesis attempts to fill.

Chapter 3 explains in detail the MoodAI study design procedure and methodology undertaken in the development of the MoodAI study. This ranges from study

recruitment, ethics approval, study goals, participant screening and the development of all the tools required to reproduce the study.

Chapter 4 presents a comparison between the methodology and results of the Shah et al [48] work and a MoodAI DL and NN model. We explore the idea of whether or not NNs are able to provide better performance over traditional regression and ensemble learning ML models.

Chapter 5 presents the data collection and ML pipeline performed using data collected from the MoodAI study. We train compositional and monolithic ML models with the goal of being able to perform accurate mood classification.

Finally, chapter 6 concludes the work by presenting, strengths and limitations of the work, reflections on the results of the work as well any potential future directions for related research relating to the work presented in this thesis.

Chapter 2

Literature Review and Existing Work

2.1 Existing Work

Many existing studies from all around the world have tried to perform some form of depression prediction with ML models using a variety of digital biomarkers such as sleep, step activity and heart rate along with clinical instruments. Three papers by Jacobson et al [20] [49] [50] utilised digital biomarkers to monitor the severity of MDD and anxiety. Jacobson et al used regression and ensemble learning-based ML models such as extreme gradient boosting, support vector machines ridge regression, random forest, and KNN to predict severity with an 80% accuracy.

Ghandeharioun et al [51] conducted a similar study using features collected from smartphone sensors and wristbands to predict the participant's HAM-D. To do this lasso, ridge, and elastic net regression models and random forest ensemble learning processes. A 10-fold cross-validated model achieves an RMSE of 7 points with the max score on the HAM-D being 52 points on the 17-item version [52].

Tazawa et al [53] is a paper published in 2020 used ML algorithms to screen for depression and the severity of depression symptoms. Tazawa et al used wearable data from smartwatches such as step activity, calorie expenditure, sleep time, heart, and skin temperature. The study recruited a total of 45 depressed participants and 41 healthy controls. Participants were screened by a clinician using the 17-item HAM-D instrument. Regression and ensemble learning models such as SVM, random forest, and XGBoost were considered, but XGBoost appeared to be

the most effective model. 10-fold cross-validation was used to validate the model and the model was able to identify symptomatic patients with a 76% accuracy.

Sano et al [54] conducted a study to identify high stress and poor mental health in participants using wearable sensors and smartphones. Physiological, phone and mobility data proved to be effective predictors of stress and mental health. The ML models used were similar to those in Ghandeharioun et al [51] being LASSO, and SVM. The best models achieved results of 78.3% accuracy for classifying high and low-stress groups, 87% accuracy for classifying high and low mental health groups, 73.5% accuracy for stress classification, and 79% accuracy for mental health classification.

All the aforementioned studies utilised regression and ensemble learning-based ML models. DL and NNs are some of the most recent developments in AI technology and recent work using DL and NNs in healthcare suggest that there is potential to apply DL and NNs in mental health diagnosis and treatment [15].

The following studies will be discussed in detail as they are the most recent and closely related to the work discussed in this thesis:

2.2 Shah et al

2.2.1 Introduction

The Shah et al work was a study conducted by a team from the University of California, San Diego, USA. The study is titled "Personalized machine learning of depressed mood using wearables" [48] published in June 2021.

2.2.2 Study Design

The Shah et al study was designed around determining the predictors of depressed mood in depressed adult humans by leveraging smartphone-based EMAs together with smartwatches and neurocognitive assessments combined with EEG measurements. They then applied ML models to the collected data to try to predict depressed mood ratings and also determine the most impactful features to guide personal intervention.

2.2.3 Study Recruitment

The study recruited adult human subjects from the University of California San Diego College Mental Health Program. No structured clinical interview was conducted as part of recruitment for this study.

All participants needed to meet the following inclusion/exclusion criteria:

- Experiencing moderate depression symptoms determined by a PHQ-9 assessment scoring between 10 and 17 inclusively.
- Absence of any suicidal behaviours determined by the use of the C-SSRS.
- Be stable on any current psychotropic medications.

A total of 14 participants took part in the study and the mean age of the recruited participants was 21.6 ± 2.8 years including a majority of 72% (10) of those participants being female.

2.2.4 Study Procedures

Participants were monitored and assessed for a period of 1 month. All data collected by the Shah et al study were in the years prior to the emergence of the COVID-19 global pandemic.

Participants in the Shah et al study were tasked to complete were the following:

- Neurocognitive Assessments
Neurocognitive assessments are assessments that are designed to test certain brain functions. A total of six assessments were used in the study; these included inhibitory control, interference processing, working memory, emotion bias, internal attention, and reward processing. These assessments were done a total of three times during the study duration, at the beginning, middle and end of the one-month duration.
- EEGs
The neurocognitive assessments were done simultaneously with the EEGs. The EEGs were used to record brain activity during the neurocognitive assessments.

- EMAs
EMAs were done using an app called BrainE [55]. This app would prompt participants at 8 am, 12 pm, 4 pm, and 8 pm to complete depression and anxiety ratings on a 7-point Likert scale.
- Stress Assessment
Stress assessments in the form of 30-second breathing assessments were done at the same time and frequency as the EMAs.
- Diet Reporting
Fats, sugars, and caffeine were recorded on a 0-6 scale at the same time and frequency as the EMAs.
- Smartwatch data
Lifestyle data such as sleep, physical activity, and stress were captured using a Samsung Galaxy smartwatch. The smartwatch was expected to be worn for the entire duration of the study except when charging.

2.2.5 Study Methodology

Shah et al extracted a total of 43 features from the collected multidimensional data to be used in their ML models. Supplementary information provided by Shah et al including a complete list and description of all 43 features can be found in Appendix A. The ML models were trained on the 43 different features using the depressed mood rating as the label. A total of seven regression and ensemble learning-based ML models were used, these included the following:

- Elastic Net
- Random Forest
- Gradient Boosted Trees
- Ada Boosted Trees
- Poisson Regressor
- Support Vector Regressor
- Voting Regressor

A nested cross-validation technique and hyperparameter tuning was performed to improve the performance of each of the ML models.

2.2.6 Study Results

Results from all the different ML models showed that there was no one model which performed the best among all participants. Table 2.1 shows the best results achieved from the best machine learning model for every participant. Results using the SHapley Additive exPlanations (SHAP), a game theory-based algorithm indicated that the top five most relevant features to the ML models [56] from most important to least important were anxiety, diet, physical activity, breathing & stress, neurocognition, and sleep in that order. Overall the Shah et al work resulted in an average MAPE of $27.9 \pm 10.3\%$ and an MAE of $0.77 \pm 0.27\%$ across all participants.

Subject ID	MAPE		MAE	
	Mean	Std	Mean	Std
1	7.55%	5.55%	0.358	0.291
10	25.45%	10.13%	0.900	0.248
12	26.27%	14.44%	0.650	0.330
14	40.88%	11.87%	1.007	0.335
15	10.24%	2.53%	0.378	0.088
18	24.05%	11.80%	0.882	0.356
19	29.11%	6.24%	0.651	0.202
20	31.55%	6.22%	1.055	0.485
21	33.28%	11.59%	0.824	0.372
23	35.12%	15.30%	0.812	0.167
24	6.40%	6.91%	0.208	0.267
26	36.41%	9.63%	1.152	0.217
28	21.23%	7.56%	0.657	0.131
29	63.14%	26.13%	1.274	0.322

Table 2.1: Results of the Shah et al’s best machine learning model for each participant

2.2.7 Conclusions

Overall the work of Shah et al showed promising results using regression and ensemble learning-based ML models using data collected with smartphones and smartwatches. Certain participants such as subjects id 1 & 24 show models with single-digit MAPE and correspondingly low MAE. Although the average MAPE and MAE can still be improved given the limitations in the quantity of collected dataset it is not always possible to generalise the prediction models. This thesis furthers the work of Shah et al by applying DL and NN-based ML techniques

using the Shah et al dataset to try to improve on the results from Shah et al (see Chapter 4).

2.3 Bai et al (Mood Mirror)

2.3.1 Introduction

Mood Mirror [57] was a study conducted by Beijing's Capital Medical University. This study was published under the title "Tracking and Monitoring Mood Stability of Patients With Major Depressive Disorder by Machine Learning Models Using Passive Digital Data: Prospective Naturalistic Multicenter Study" published in March 2021.

2.3.2 Study Design

The Mood Mirror study was conducted as a multisite, noninterventional prospective study at 4 different psychiatric hospitals in Beijing, China. The study utilised an in-house designed app called Mood Mirror installed on Android smartphones along with a Mi Band 2 which recorded sleep, heart rate, and step count data. The Mood Mirror app also facilitated the capture of daily EMAs and Biweekly PHQ-9 assessments. The study was designed to determine the correlation between all these features for patients with depression.

2.3.3 Study Recruitment

Participants were recruited from the 4 clinics across Beijing. The recruitment period was from February 2019 to April 2020.

Participants of the Mood Mirror study needed to meet the following inclusion/exclusion criteria to be admitted into the study:

- Be within the ages of 18 to 60 years
- Diagnosed with MDD determined by a clinician using the DSM IV [58]
- No current diagnosis of substance abuse
- Own an Android phone capable of installing the Mood Mirror app

2.3.4 Study Procedures

Participants were observed for a total of 12 weeks. Participants were reimbursed a total of ¥500 for completing the entire 12-week observation period.

Participants of the Mood Mirror study were required to complete the following tasks over their 12-week observation period:

- Bi-weekly physician-administered HAM-D assessments [52]
The HAM-D is one of the most common clinician-administrated depression rating scales.
- Bi-weekly PHQ-9 assessments
The PHQ-9 is a self-administered instrument for diagnosing and measuring the severity of depression in an individual.
- Daily mood ratings
Essentially an EMA using a visual 7-point Likert scale completed via the Mood Mirror app.
- Passive smartphone usage data capture
Consent to providing phone interaction and usage data such as call and text logs, GPS location, and screen status.
- Wristband physiological data capture
Wear the provided wristband capturing data about a participant's sleep, step and heart rate.

2.3.5 Study Methodology

Data collected from the Mood Mirror app and wristband were separated into feature groups. These groups were sleep, step count, heart rate, and phone data. Different regression and ensemble learning-based ML approaches were applied to each of the different features individually as well as all the features as a whole to predict between two different groups of participants. The following ML models were used:

- Support Vector Machines
- K-nearest Neighbours

- Decision Trees
- Naive Bayes
- Random Forest
- Logistic Regression

2.3.6 Study Results

Accuracies of the best model for all binary classification task showed promising results with greater than 70%. The best classification group was between steady-remission and swing-moderate achieving an accuracy of above 80%. Similar to Shah et al, there was no one ML model which was best across the different classification categories. The most effective features used in the model were sleep, step count and heart rate data. See Table 2.2 for a more detailed breakdown for the Mood Mirror results.

Different class of mood states	Best Model	Average Percentage Accuracy	Average Std Percentage
Steady and Swing	KNN	76.67%	8.47%
Steady-remission and Swing-drastric	Naïve Bayes	74.29%	9.27%
Steady-remission and Swing-moderate	KNN	80.56%	15.28%
Steady-depressed and Swing-drastric	Logistic Regression	75.91%	13.18%
Steady-depressed and Swing-moderate	SVM	74.73%	8.44%

Table 2.2: Best Mood Mirror Results using all collected features

2.3.7 Conclusions

Results of Mood Mirror models appear to be promising, with most models showing accuracies greater than 70% using data collected passively from smartphones and smartwatches. The Mood Mirror study’s main strength was the high participant count of 334, largely due to the collaboration with 4 different psychiatric hospitals. While the participant count was high compared to other studies this may explain why accuracies were lower than in some of the other studies discussed. Similar to

that of Shah et al [48] the models being used are also regression and ensemble learning-based ML models. This once again poses the question of whether DL and NN approaches are able to provide improvements in accuracy and variance over traditional ML models.

2.4 Rykov et al

2.4.1 Introduction

Rykov et al is a study published under the title "Digital Biomarkers for Depression Screening With Wearable Devices: Cross-sectional Study With Machine Learning Modeling" [34]. The study was designed to be a cross-sectional study of $n = 290$ healthy working adults. The goal of the study was to utilise digital biomarkers and sensor data from a Fitbit smartwatch to determine the correlation between depression symptoms and digital biomarkers in a general population.

2.4.2 Study Recruitment

The 290 healthy working adults over the age of 21 years old were recruited from Nanyang Technological University in Singapore. The average age of participants was 33 years old with ages ranging from 21 to 64 years. There was a slight bias of female participants with $n=170$ or 63.7%.

2.4.3 Study Procedures

Participants were involved with the study for two weeks and required to complete PHQ-9 assessments as a self-reported depression screening at the start and end of the study period. During the study period, they were provided and required to wear a Fitbit Charge 2 smartwatch. The smartwatch captured different biomarkers such as physical activity, sleep patterns, and circadian rhythms.

2.4.4 Study Methodology

The averaged scores from PHQ-9 assessments were used to determine the classify the participants into different groups based on depression symptom severity. Pre-processing was done on the lifestyle data collected from the Fitbit Charge 2 smartwatch. This involved excluding any data from participants that were not actively wearing the smartwatch for more than 20 hours a day and also if there were not at least 10 days worth of complete data. There were participants ($n=24$) that

scored 0 for their PHQ-9 assessment and they were also excluded. Supervised ML models were used to predict symptom severity to determine the predictive ability of the captured digital biomarkers. 4 fold cross-validation was used to validate the machine learning model.

2.4.5 Study Results

Regression analysis showed three digital biomarkers that were highly correlated with depression symptom severity. These markers were step count, variation in nighttime heart rate, and interday stability. Their best model which was detecting participants with a high risk of depression had an accuracy of 80%, a sensitivity of 82%, and a specificity of 78%. Unfortunately, the types of machine learning models used were not explicitly disclosed.

2.4.6 Conclusions

Overall the work by Rykov et al indicates that physiological data from the latest consumer wearables can help aid in the depression screening, but existing models are unable to achieve accuracies greater than 80%. The major limitation of the Rykov et al work was the low sampling frequency, participants were only monitored for 2 weeks using a singular label for the entire set of biomarkers. Additionally, there was a heavy bias in the number of individuals with PHQ-9 scores between 0 and 4 as the recruited participants were healthy adults.

Digital biomarkers from consumer-grade wearable sensors appear to be viable for use early detection and diagnosis of depression, but future work can be done by using smartphone enriched data and applying different ML models could provide improved results.

2.5 Current Gaps

There has been a large body of work done using traditional statistical ML methods such as regression and ensemble learning [34] [20] [49] [50] [51] [48] [57]. These studies have proven the effectiveness of ML for diagnosis and prediction of mental health disorders. Review of existing studies indicated that the regression and ensemble learning models were unable to consistently achieve accuracies above 80%.

DL and NNs have had proved successful in a wide range of applications from computer vision to applications in healthcare [59] [15]. The viability of DL has not

yet been fully explored in area of digital biomarkers and mood state prediction but recent applications in related fields such the use of audio and visual data for depression prediction and severity recognition [60] [61] [62].

2.6 Problem Statement

Depression is one of the leading causes of disability and impaired quality of life. Smartphone and smartwatch technology is now at a stage where we are able to use the data collected from their sensors to make judgments about mental health. There have been existing work using regression and ensemble learning-based approaches with limited success. The concepts of DL and NN have still yet to be applied in the context of mood classification. This thesis will explore whether or not it is possible to use digital biomarkers and DL and NN methods to classify mood state more accurately than existing regression and ensemble learning ML models.

2.7 Research Contributions

The main contribution of this thesis is:

- Designing a longitudinal observational study to capture digital biomarkers and EMAs from healthy controls using smartphones and smartwatches.
- Applying DL and NN approaches to the existing Shah et al dataset and comparing the effectiveness of regression and ensemble learning models with DL models.
- Applying monolithic and compositional approaches to the collected (MoodAI) dataset to determine ability of models to predict mood state using collected EMAs and digital biomarkers.

Chapter 3

MoodAI Study

The MoodAI study was conducted as a collaboration between the University of Auckland Faculty of Engineering Department of Electrical, Computer, and Software and the University of Auckland Faculty of Medical and Health Sciences, Department of Psychological Medicine. The MoodAI study was separated into two distinct stages. Firstly, the development of the protocols and infrastructure such as the development of the MoodAI web app and backend services. Secondly, was the study involvement stage where we recruited and monitored participants each for a period of one month. The study utilises a Fitbit Sense smartwatch - one of Fitbit's high-end consumer-grade fitness trackers. The accuracy and validity of Fitbit smartwatches have been investigated by many studies and have been determined to provide a suitable accuracy for the measurement of physical activity [63] [64] [65] [66] [67], heart rate [68] [69] [70], sleep [71] [69], HRV, and Breathing rate [72]. As a result, Fitbit smartwatches have been utilised in many observational studies.

The entire study process is described in detail in this Chapter. A summary of the entire process can be seen in Figure 3.1.

3.1 Ethics Committee Review and Approval

The MoodAI study protocol was approved by the Auckland Health Research Ethics Committee under the project title: "Real-time assessment of mood changes and machine learning" (Ref AH22436). Participants were required to provide informed consent in order to participate in the study. All participant data was de-identified and a randomly generated 28-digit Universal Unique Identifier (UUID)

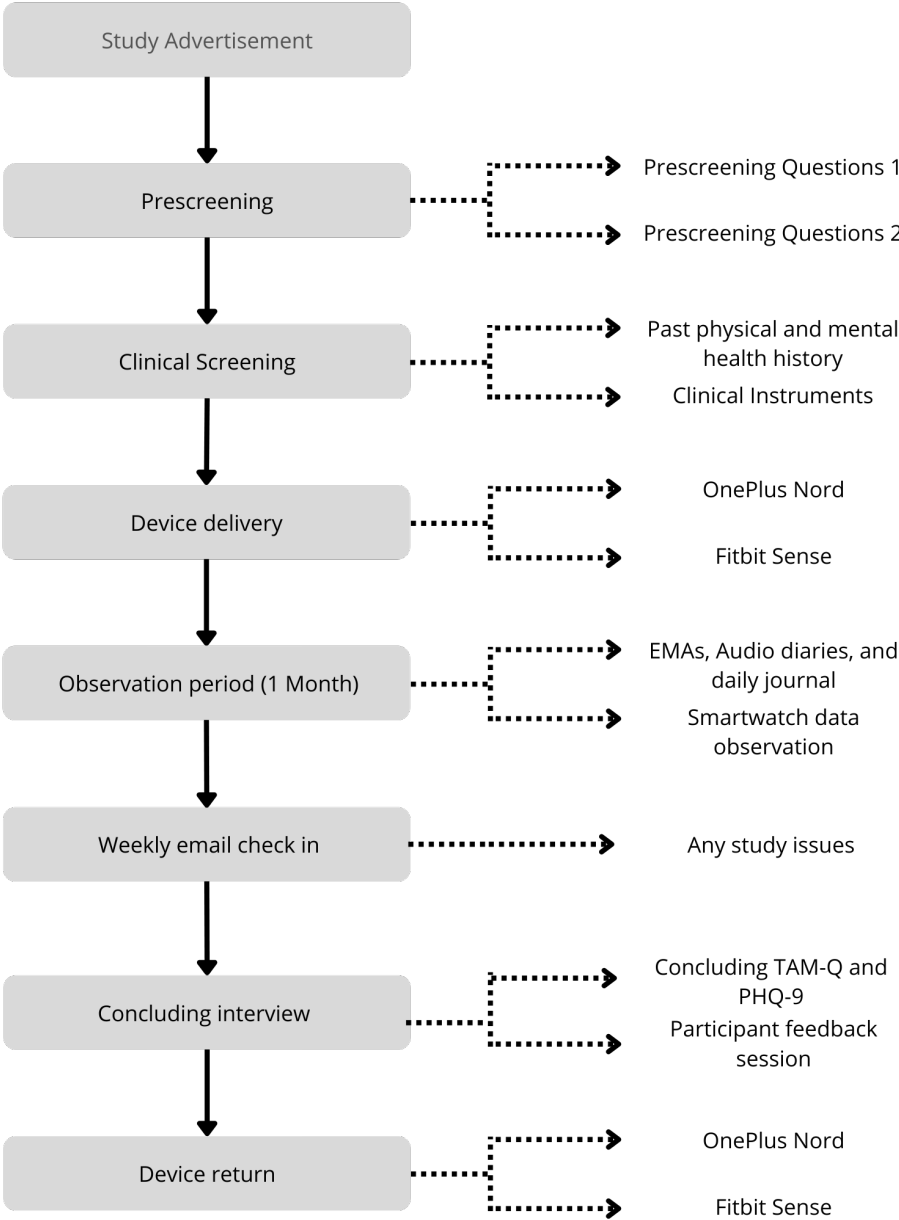


Figure 3.1: Summary of entire study process

was used to represent them in the collected dataset. Only the primary investigator and other researchers approved by the primary investigator had access to the identifying table. Collected data was stored securely on a study-managed web server securely hosted on the Google Cloud Platform. Participants were able to withdraw from the study at any point during the one-month observation period as mentioned in the provided participant information sheet (see Appendix B), but data collected up until then was still able to be included for analysis.

3.2 Registration with the Australian New Zealand Clinical Trials Registry

MoodAI was registered with the Australian New Zealand Clinical Trials Registry (ANZCTR) under the title: "Real-time assessment comparing mood changes and machine learning in adults with mild-moderate depression and a group of healthy volunteers" (ACTRN12621000803897) [73]. The ANZCTR is an online registry of clinical trials managed by the National Health and Medical Research Council (NHMRC) Clinical Trials Centre, University of Sydney. The registry provides research transparency, facilitates trial participation, promotes research collaboration, and overall enhanced the quality of the study.

3.3 Study Funding

The MoodAI work discussed in this thesis is linked to a Health Research Council of New Zealand research project titled "Real-time assessment of mood changes and machine learning" (project number: 3722256) [74]. The Health Research Council did not have a further role in the study design process, collection of data, analysis, development of machine learning models, or writing of any reports and literature relating to the study.

3.4 Study Design

The study was developed to be a case-control longitudinal study to monitor both mood and physiological features over a one-month period. To achieve this, smartphones and smartwatches were utilised to digitally capture ecological momentary assessments (EMAs) [75] and physiological biomarkers such as heart rate, sleep, and physical activity. The study design methodology was to keep required tasks

as simple and accessible as possible. Many studies in this field have previously consisted of invasive and time-demanding tasks to complete. These demanding tasks reduce participant engagement over time and result in missing data. Additionally, the study also captures daily audio diaries which will be used for future work in the area of speech and vocal features in classifying mood. These features have been shown to have links to mood and mood disorders [76]. The observation period for participants was one-month and participants were provided a total of \$100 worth of supermarket vouchers as compensation for their time.

3.5 Study Recruitment

The study consisted of two populations; mild to moderate depression population and a healthy control population. Study recruitment was conducted in slightly different ways depending on the population. For participants with mild-moderate depression, recruitment was intended to be via primary healthcare providers around the Auckland region, one provider, in particular, was Tamaki Health. We needed to ensure our observation methods were approved by their General Practitioner (GP) to minimise if not avoid any unexpected risks to their health. Due to COVID-19 disruptions, we were unable to recruit any participants in the mild-moderate depression group, this is further discussed in Chapter 3.12.

Healthy controls did not require such approval from their GP and were recruited via either advertisements sent through the University of Auckland Department of Electrical, Computer and Software Engineering mailing lists, in-person lecture advertisements or by word of mouth from the research team and participants.

3.6 Study Inclusion/Exclusion Criteria

In order to participate in the study, participants needed to meet a series of inclusion/exclusion criteria. The inclusion/exclusion criteria help establish an ideal pool of participants for the monitoring task. The inclusion/exclusion criteria are as follows:

Inclusion Criteria:

- Participants are male or female, aged between 18 to 60 years

- Participants are willing and able to give informed consent for participation in the study
- Participants are willing to undergo a detailed clinical screening interview by a clinician and member of the research team
- Participants are willing to use the study-provided smartwatch and smart-phone(if needed) for the duration of the study
- Have a recent depression diagnosis (Only applicable to the depressed group)

Exclusion Criteria:

- Inability to speak or read English to a level that enables informed consent and/or participation in the study
- Unable or disinterested in using the study provided technology
- Substance abuse or dependence in the last six months
- Any other unstable medical or neurological condition
- Any other condition judged by the research team as likely to impact the ability to complete the study

3.7 Pre-screening Process

After a potential participant has shown interest in participating in the study, a member of the research team would conduct the pre-screening process. This process consists of asking a series of questions about relevant past and present medical history and certain daily activities. The purpose of the pre-screening process was to ensure that participants meet the study's inclusion and exclusion criteria. The questions help to identify if there are any immediate health concerns with the participant before proceeding further with recruitment. The detailed screening process is very resource-intensive on the research team and also very detailed from the participant's point of view as many clinical measurement tools are used so the pre-screening process helps to efficiently filter out potentially unsuitable participants.

The first set of pre-screening questions are as follows:

- How did you hear about the study?
- What is your main reason for taking part in the study?
- Would you be willing to use the technology in the study?
- Are you currently attending your GP or specialist mental health services for depression?
- Are you currently on any medication for your mental health?
- Do you consume alcohol daily?
- Are you currently suicidal?
- Are you between the ages of 18 to 60?

If there were no major concerns with the participant following their answers to the first set of questions, demographic and contact details were collected to contact and arrange delivery of the device to the participant.

The second set of pre-screening questions are as follows:

- Date of birth/Age:
- Weight and Height:
- COVID-19 Vaccine Status:
- Name and Address:
- Contact phone number and email address:
- Would you prefer to use the study-provided phone or your own personal phone? If you choose to use your personal phone, please specify the model.
- Need a study-provided smartphone and SIM card?

3.8 Study Screening

If participants passed all the required pre-screening checks they would then proceed to a detailed clinical screening session before being accepted to the study. The purpose of this detailed screening session was to look into any past physical and mental health that may impact their ability to participate in the study and to ensure that they are in the correct study group.

The clinical screening session was originally planned to be in person in the University of Auckland Grafton Campus but needed to be transitioned to be done via Zoom due to lockdown restrictions. The clinical screening session with the participant was conducted with a total of two people from the research team - A clinician performing the clinical assessments and myself to help with data entry and demonstrating and answering questions about the technology utilised in the study.

Details and results from clinical assessments collected during the participant screening session were collected on REDCap (Research Electronic Data Capture) which is a browser-based surveying platform designed specifically for research use and is commonly used by studies in the medical field [34] [77].

The study screening involved a study clinician assessing any past and present physical or mental health conditions they may have had. The structured assessments listed below assessed suicidality, psychiatric comorbidities, depression screening, depression severity and alcohol use.

The instruments used in the clinical screening were as follows:

- Patient Health Questionnaire 9 (PHQ-9) See Appendix [78] [79]

The PHQ-9 is a well-recognised clinical instrument used for depression screening utilised by many studies [80] [34] [48]. It consists of a total of 9 items each ranging from 0 to 3 points, for a total of up to 27 points. It is used in the MoodAI study to determine whether or not participants are in the right population group. The guideline for participants in the mild-moderate depression group is 10 - 19 inclusive. Scores of 20 and above indicate potentially severe depression and participation may not be suitable for the participant. Healthy controls are expected to have scored less than 10 otherwise they may have an undiagnosed major depressive disorder and will be advised to seek the relevant help. The full PHQ-9 instrument used in the screening can be found in Appendix C.

- Alcohol Use Disorders Identification Test (AUDIT) See Appendix [81]

A participant’s alcohol consumption was essential to ensuring the physiological data collected from the study was a fair representation of a healthy participant. Alcohol has been known to decrease heart rate variability and also may also result in increased heart rate [82] [83] [84]. The AUDIT is a questionnaire designed by the World Health Organisation as a simple method of screening for excessive alcohol consumption. The AUDIT is an effective means of screening for the spectrum of alcohol use disorder [85]. The AUDIT consists of a total of ten questions with varying answers worth up to 4 points each, the AUDIT score ranges from 0 to 40. In order to be considered eligible for the study, the participant needed to have an AUDIT score of less than 16. AUDIT scores greater than 16 may indicate potentially harmful drinking alcohol or alcohol dependence [86]. The full AUDIT instrument used in the screening can be found in Appendix D.
- Columbia-Suicide Severity Rating Scale (C-SSRS) [87]

The C-SSRS is a detailed questionnaire that is used to determine the presence of suicidal intentions or behaviour. Unlike the AUDIT and PHQ-9, there is no score and explicit guideline for acceptance, but instead, the screening clinician needed to determine whether or not the participant had a history of suicidal ideations/intents/acts or is currently suicidal. The full C-SSRS instrument used in the screening can be found in Appendix E.
- Mini-International Neuropsychiatric Interview (M.I.N.I) See Appendix [88]

The M.I.N.I is a structured diagnostic clinical interview that determines the presence of comorbidities. Comorbidities are any other disease or condition which may be present.

Certain comorbidities such as chronic heart comorbidities or breathing-related comorbidities could limit a participant’s ability to perform exercise and thus impacted the data collected from the smartwatch. It is at the discretion of the examining clinician and research team to determine if the presence of a comorbidity could potentially impact data collection. If the comorbidity is determined to be relevant, the participant was excluded from the study. The full M.I.N.I instrument used in the screening can be found in Appendix F.
- Montgomery–Åsberg depression rating scale (MADRS) See Appendix [89] [90]

The MADRS is a clinical instrument used by clinicians to determine the presence and severity of depression. The MADRS consists of a total of 10 items; Apparent sadness, Reported sadness, Inner tension, Reduced sleep, Reduced appetite, Concentration difficulties, Lassitude, Inability to feel, Pessimistic thoughts, and Suicidal thoughts. Depending on the severity, each of the aforementioned items is assigned a value from 0 to 6. A total MADRS score is then calculated as the total of the scores for each item. This score ranges from 0 to 50 and the score requirement for the mild-moderate depression population was between 7 to 34 inclusive, and for the healthy control population, the total score should be no more than 6. The full MADRS instrument used in the screening can be found in Appendix G.

- Technology Acceptance Model Questionnaire (TAM-Q) See Appendix
The purpose of TAM-Q is to determine a participant's acceptance of the technology provided in the study. Similar to [91] the TAM-Q adapter for the MoodAI study utilised a five-point Likert scale with the following set points: (1) Strongly disagree, (2) Disagree, (3) Neutral, (4) Agree, (5) Strongly agree. The TAM-Q consisted of questions about the usefulness of wearable technology and whether the participant had any data security concerns. The full TAM-Q questionnaire used in the screening can be found in Appendix H.

3.9 Choice of devices

Many other studies limited smartphone usage to only Android smartphones [57] [92] rather than giving participants a choice. The MoodAI Study hoped to be able to include a wider variety of participants regardless of their smartphone ownership or preference. This was further reflected in the choice of the Fitbit Sense smartwatch and web app approach which is compatible with both Android and iOS operating systems. During this study observation period participants were loaned a Fitbit Sense smartwatch (see Figure 3.3) and a OnePlus Nord smartphone (if they wanted or needed a separate phone to participate in the study)(see Figure 3.2).



Figure 3.2: Study provided OnePlus Nord Smartphone



Figure 3.3: Study provided Fitbit Sense Smartwatch

3.10 Study Procedure

Participants were observed for up to a period of one month. Participants were provided with a \$50 supermarket voucher as reimbursement for the time spent participating in the pre-screening and screening process. Participants were required to complete daily tasks on their provided smartphone and also expected to wear the smartwatch for the entire monitoring period (excluding charging, showering and cleaning). The tasks included EMAs, a daily journal, and a daily audio diary. These tasks and their frequency will be discussed in more detail in Section 3.14.

A team member would check in with each participant weekly to ensure there are no issues with data collection or the study in general. Additionally, an initial check-in would be performed after the first day of the study to ensure there are no immediate issues.

At the study conclusion, a team member would scheduled a short session to discuss device return and the completion of end-of-study TAM-Q and PHQ-9 questionnaires. Participants were required to complete an end-of-study PHQ-9 and TAM-Q to determine if there were any changes in the presence or severity of depression and their thoughts on technology acceptance respectively. Participants were also provided the opportunity to give any additional feedback relating to any part of the study. Participants were provided with a \$50 supermarket voucher as reimbursement following their conclusion of the study.

3.11 Recruitment Results

As of 31 May 2022, the study recruited and completed the observational period for a total of 15 participants. All participants were healthy controls and at that stage, no participants with mild-moderate depression had been recruited, for reasons discussed further in the limitations section of this thesis. Of those 15 participants, 6 were female (40%) and the rest were male. The ages of participants in the study ranged from 19 to 52 with the mean age being 29 years old. All 15 participants that were recruited were recruited via the channels mentioned in the Study Recruitment section. There was a diverse spread of cultures of participants from Indian, NZ European, Chinese, Māori, and other European backgrounds. Participants from a Pacific Island background were under-represented in this study despite efforts to recruit.

3.12 COVID-19 Disruptions

We planned on observing both a depressed group and a healthy control group. We did not manage to recruit any participants for the depressed group despite our best efforts. Our primary healthcare referrer Tamaki Health was overwhelmed with providing resources during the COVID-19 pandemic and was facilitating the vaccine roll-out.

Recruitment began in August 2021 and during that period, there were multiple lockdowns put in place by the New Zealand government restricting our ability to interact with participants. Study screening was unable to continue in person and thus the study transitioned to online screening sessions via Zoom and devices were couriered to suitable participants.

Since we were unable to recruit any participants in the depressed group we Instead reached out to a research group from the University of San Diego [48]

and utilised their published dataset as a substitute and means of comparing the effectiveness of the different methods, this is discussed further in Chapter 4.

3.13 MoodAI Website Design

3.13.1 Purpose

The purpose of the MoodAI website was to create "a one place find all" solution for participants to be able to access anywhere and on any device. The use of a web platform meant that it was accessible on both Android and iOS smartphones as well as PCs and laptops if desired. Many studies utilised multiple third-party app[93] [48], this adds an extra burden on participants to complete required tasks. A study-controlled website also provides extra confidence to participants the data is managed directly by the research team. This is reflected in the participant feedback and acceptability section 3.15.

3.13.2 Framework

The MoodAI web app was developed and hosted using the BaaS provided by Google titled Firebase. Firebase provides resources for users to create mobile and web applications quickly as they provide useful features that would normally require a lot of development time to build. Firebase's features include security, authentication, No-SQL database, cloud storage, cloud functions and Hosting. Other clinical studies have utilised a multitude of features such as the No-SQL database [94]. These features allow for creating a highly effective and scalable web app with minimal cost. Utilising the Firebase JavaScript SDK, the web app was developed using standard HTML, CSS and JavaScript.

3.13.3 Participant Access

As discussed in section 3.1 all participant data were de-identified and they were provided unique randomly generated details to use to log into the website. Figure 3.4 shows the MoodAI logo and design of the login page. Firebase utilises industry-leading security authentication provided by Google. Security rules are put in place are handled by the server and put in place so participants are not able to access any data not belonging to them [95].

MOODAI

Home

About Us

Help

Sign in with email

Email

NEXT

By continuing, you are indicating that you accept our [Terms of Service](#) and [Privacy Policy](#).

Figure 3.4: MoodAI website login page

3.14 Data collected

3.14.1 Ecological Momentary Assessments (EMAs)

EMAs are user-generated self-ratings about their mood or stress captured at multiple periods in time [75]. EMAs can provide clinicians a good insight into how a patient is progressing throughout the day. The EMA collection process is able to be enhanced through the use of smartphones [96]. Smartphones allow for EMAs to be captured multiple times a day and remotely at the convenience of the participant [30] [97] [98] [48]. The captured EMAs could then be used to determine the presence of many mood disorders such as [99] bipolar disorder [100] [92], and depression [101].

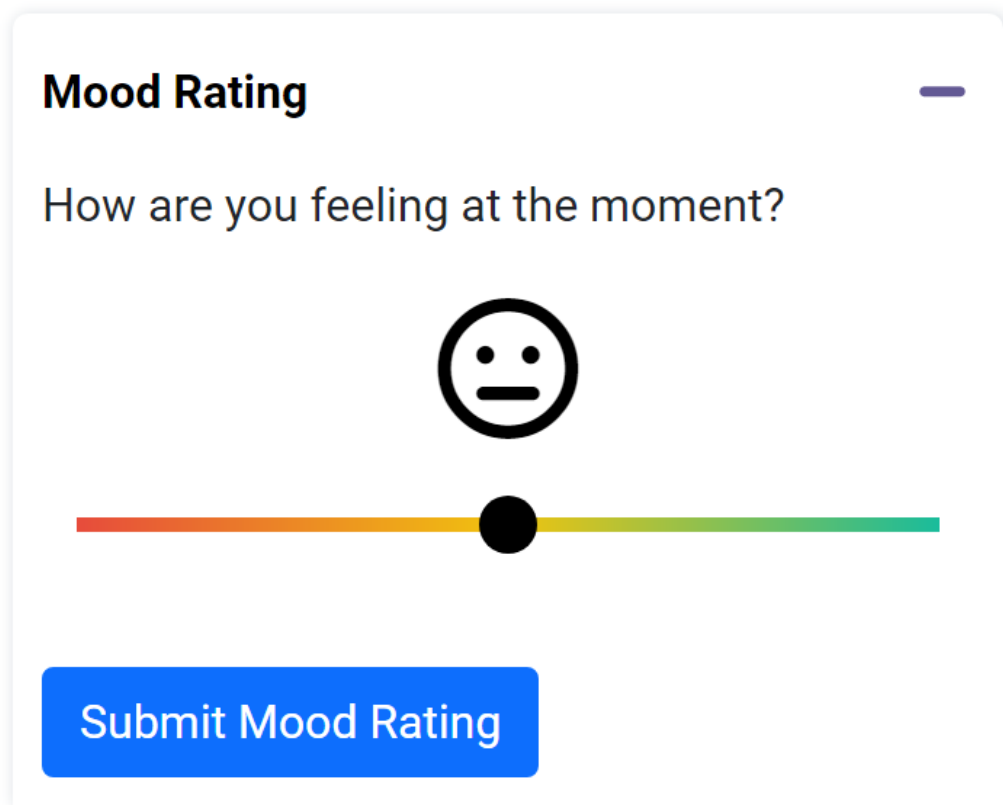
Researchers are increasingly utilising EMAs in depression studies as they provide valuable insights in hard-to-monitor behaviours and mood changes [101] [102] [48]. There have been variations in Likert scales used in study administered EMAs. The most commonly used were 5-point [92] [103], 7-point [57] [30] [48] [104], and 10-point [105] Likert scales.

The final decision for the MoodAI study was to utilise a 5-point Likert scale. The main reason behind this decision was to make it as simple as possible for participants. A 5-point Likert scale will reduce the trouble of deciding between too many options, while also allowing for enough granularity to distinguish between different mood states.

Participants were expected to submit EMAs a total of 5 times a day. Participants received SMS reminders on their study phones at 9 am, 12 pm, 3 pm, 6 pm, and 9 pm to instruct them to complete the EMA as soon as possible. The assessment was then done on the MoodAI website accessible via their smartphone. Figures 3.5, 3.6, 3.7 show the final designs of the EMAs included in the MoodAI web app. Although the EMA utilised a 5-point Likert scale in the background, a visual slider and emojis were created to disguise the 5-point Likert scale [92]. Many studies have utilised a visual scale [57] as it is more engaging and enjoyable from the point of view of the participant, thus increasing the likelihood of the task being completed and therefore an improved completion percentage.

3.14.2 End of Day Journal

The end-of-day journal is a short questionnaire to be completed by participants daily following the daily audio diary. The purpose of this journal is for the par-



Mood Rating

How are you feeling at the moment?

☹️

Submit Mood Rating

Figure 3.5: MoodAI Neutral Ecological Momentary Assessment

participant to inform the research team of any changes in mental or physical health that might have arisen during the course of the monitoring period.

As discussed in Chapter 3.8 Study Screening, the presence of a comorbidity could potentially impact data collection. This journal would be checked prior to performing weekly check-ins with participants.

The end-of-day journal as it appears on the MoodAI web app can be seen in Figure 3.8. It consisted of three questions and an optional field for any other health issues that might have arisen impacting the data collected or their ability to continue participating in the study.

The purpose of each of the questions is as follows:

- Have you consumed any Alcohol and/or Drugs within the last 24 hours?
As stated in Chapter 3.8 Alcohol consumption is known to decrease HRV and potentially increase heart rate [82] [83] [84]. Similar to alcohol drug use

Mood Rating

How are you feeling at the moment?

☹️

Submit Mood Rating

Figure 3.6: MoodAI Negative Ecological Momentary Assessment

has also been shown to decrease HRV [106] [107]. This question allows us to get a timestamp which we can refer back to if there were any significant dips in any given participant's physiological measures.

- Have you had a large meal within the last 3 hours?
High intakes of fats and high carbohydrates have been found to reduce HRV [108]. Large meals can potentially suppress HRV in the short term and have long-term baseline impacts if increased intakes persist over a period of time. Therefore, this question allows us to determine if a decrease in HRV may have been due to a large meal temporarily suppressing HRV. Participants had final judgement but were instructed a large meal was a meal that was greater than the quantity of what they would normally consume for that meal.
- Have you done any intensive exercise within the last 3 hours?
Exercise and physical activity has shown to have a direct link to HRV [109].

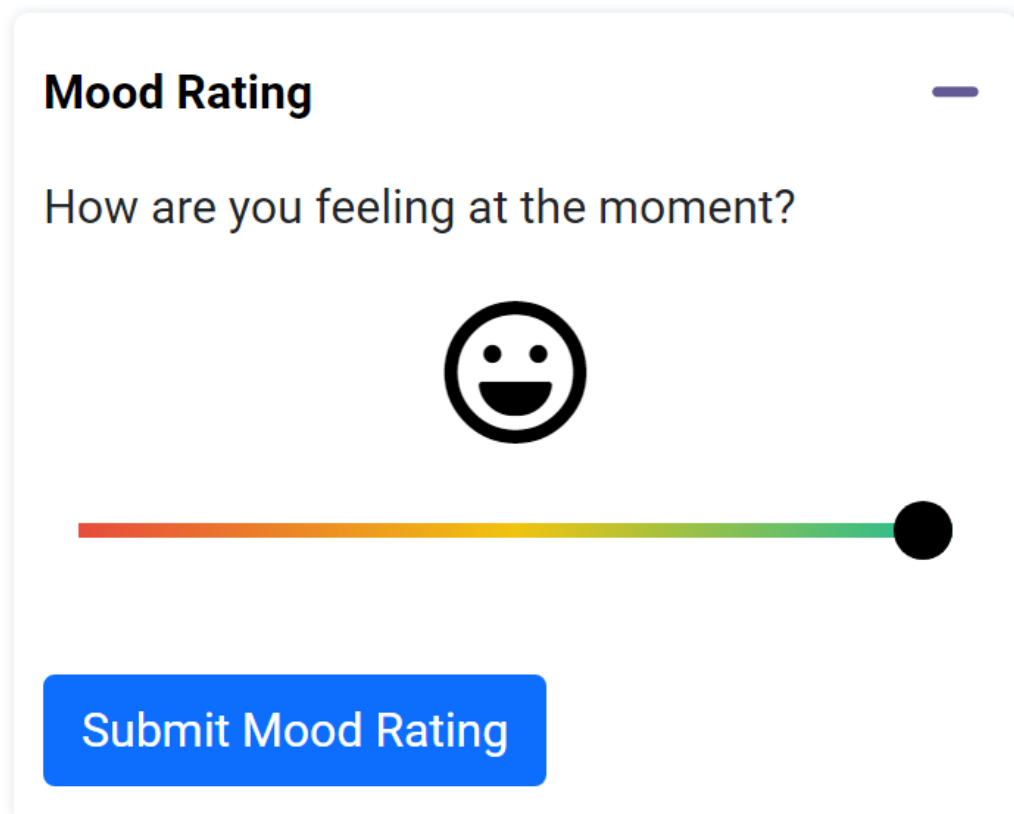
The image shows a mobile application interface for a mood rating. At the top, the title "Mood Rating" is displayed in bold black text. Below the title is the question "How are you feeling at the moment?". In the center, there is a large black and white smiley face icon. Below the icon is a horizontal color gradient bar that transitions from red on the left to green on the right, with a black circular slider positioned at the far right end. At the bottom of the interface is a blue button with the text "Submit Mood Rating" in white.

Figure 3.7: MoodAI Positive Ecological Momentary Assessment

Studies have found that increased physical activity results in an increased HRV compared to a similar individual who engages in less physical activity. Physical activity has also been shown to be linked to depression [28] [29] and those who exercise frequently are less likely to have a major depressive disorder [27] [30]. This question was subject to the participant's judgement, but Intensive exercise was defined to participants as any exercise which would elevate your heart rate over normal levels. Examples included but were not limited to running or swimming.

Data collected from the end-of-day journal was stored securely on the MoodAI Firestore database located on the Firebase BaaS.

3.14.3 Physiological Data

Fitbit wearables are becoming more accessible and an important tool to assess physiological features such as heart rate, sleep and activity. The links between

End of day Journal —

Have you consumed any Alcohol and/or Drugs within the last 24 hours?

Yes No

Have you had a large meal within the last 3 hours?

Yes No

Have you done any intensive exercise within the last 3 hours?

Yes No

Other Entries

Any other things that may be affecting your health..

Figure 3.8: MoodAI Daily End of Day Journal

these physical activity, sleep, and the heart with mental health is clear [28] [29] [28] [110]. Many studies have been adding smartwatches [48] such as Fitbit smartwatches as a means of enhancing their study [51] [34] [69].

The data collected from the provided Fitbit Sense is passively done. The Fitbit was pre-paired to the participant's phone and syncs data to the Fitbit cloud on a regular basis. Data is able to be stored for up to 30 days to ensure no data loss when out of internet connectivity for a prolonged period of time. A lot of different data is collected on the Fitbit, in particular, we focused on activity, sleep, and heart-based data. Data is then extracted via a Python Fitbit API and Fitbit REST APIs to be pre-processed for the machine learning models discussed in chapter 5.

3.14.4 Audio diaries

Speech characteristics of those with mental health issues differ from those of healthy individuals [76]. Features extracted from voice diaries provide valuable insight into the presence and severity of mental health disorders [111]. Existing studies have already tried utilising smartphones to record audio logs and extract prosodic, spectral and, cepstral features to determine the presence of mental health illnesses [19] [112] [113].

For the MoodAI study, we took inspiration from Dickerson et al [113] and Place et al [114] to developed a daily audio diary consisting of three questions in total. The user interface of the audio diary questions as seen on the MoodAI web app as seen by participants can be seen in the figures 3.9, 3.10, and 3.11 respectively. The first question was a stock sentence that provided a common ground for a trend to be developed from vocal features. The second question asked about a positive part of the participant's day by getting them to describe things they found enjoyable. The third and final question was the opposite of that and asked about any negative or things they disliked about their day. The audio diary was to be completed once a day together with the End of Day Journal. The total length of the audio diary was expected to be between 2 to 5 minutes in length.

3.14.5 MoodAI Architecture Summary

The figure 3.12 provides a summary of the entire MoodAI architecture from the type of data captured from the smartphone and smartwatch to how the data is stored and managed.

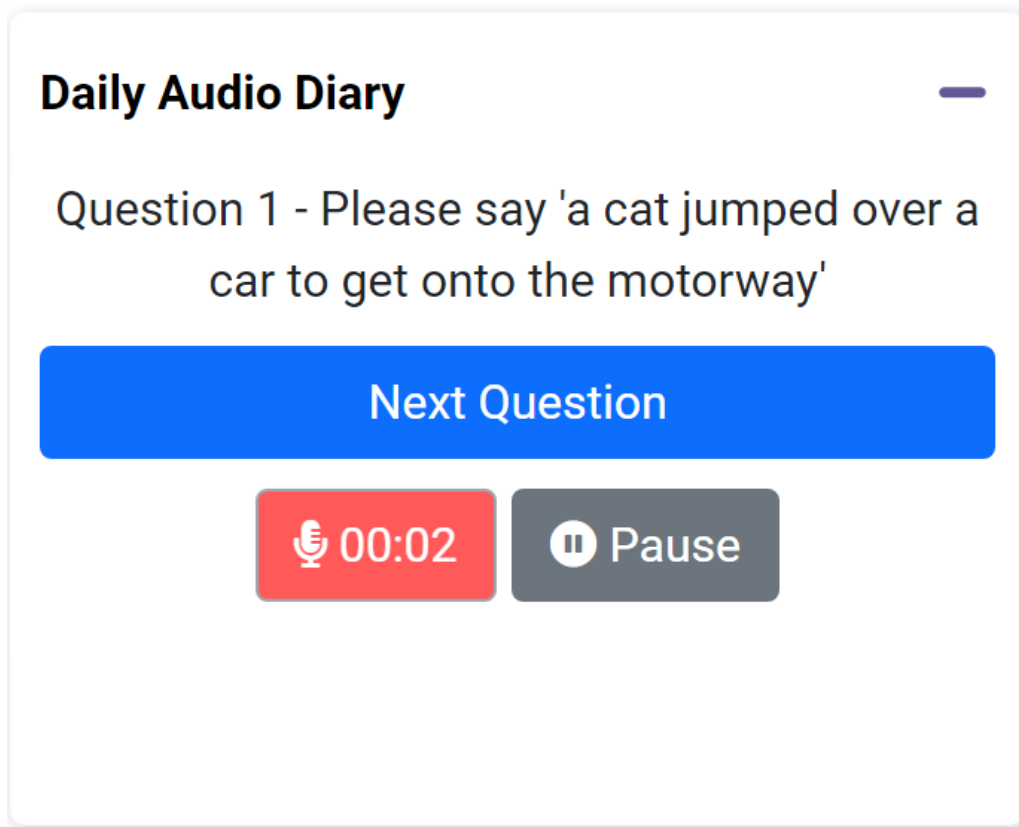


Figure 3.9: MoodAI Daily Audio Diary Question 1

3.15 Participant Feedback and Acceptability

At the end of the study, participants were required to complete an end-of-study TAM-Q and PHQ-9. Notable trends in the TAM-Q were that participants were not worried about the data security of the MoodAI platform or that of Fitbit. All 15 participants either answered strongly agree or agree to the study system being easy to use. 12 out of the 15 recruited participants thought the platform would provide benefit to those with depression, with the remaining 3 answering neutral for that question. There was a greater split in the acceptance of the smartwatch where 10 participants answered that wearables would use useful in their everyday life, 4 participants were neutral and, 1 participant was concerned about wearing the smartwatch overnight.

Between the start of study PHQ-9 and the end of study PHQ-9, there were no significant differences in PHQ-9 score, all differences were within ± 2 points and did not cross any major thresholds.

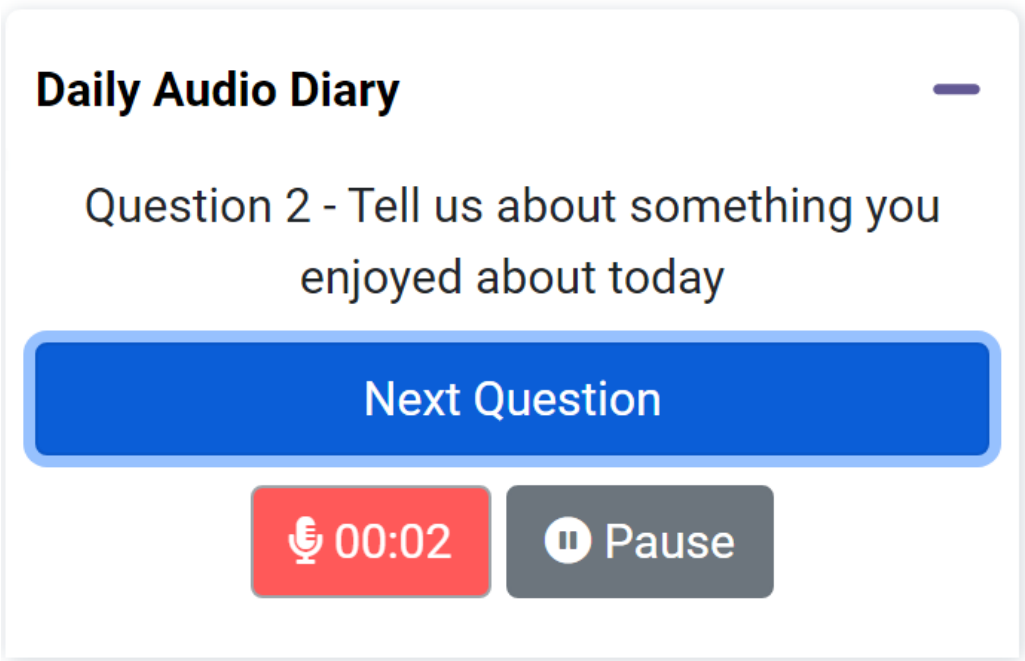


Figure 3.10: MoodAI Daily Audio Diary Question 2

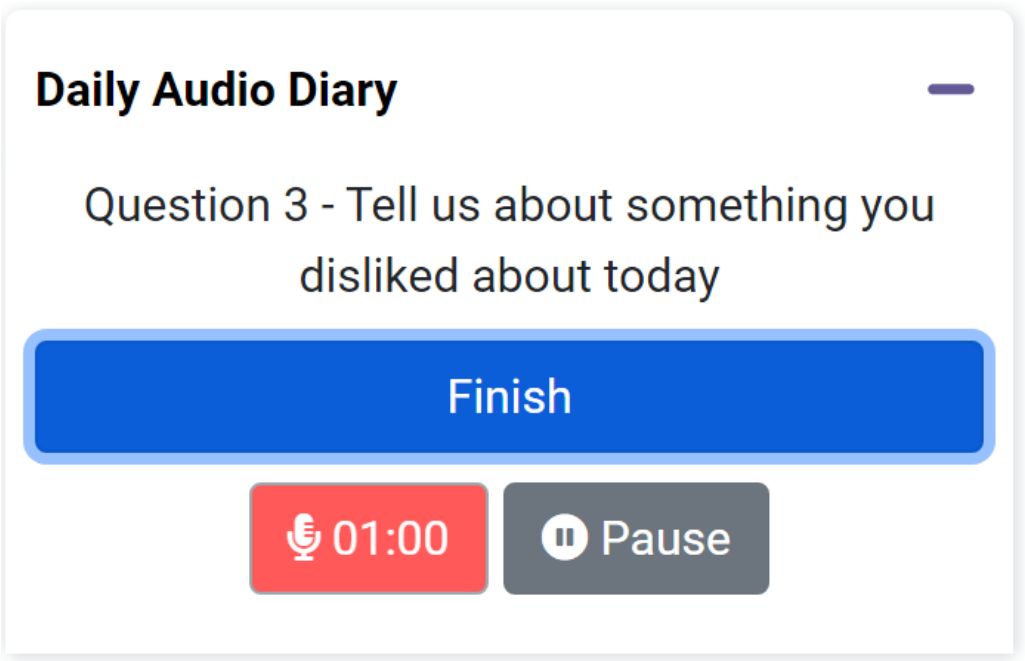


Figure 3.11: MoodAI Daily Audio Diary Question 3

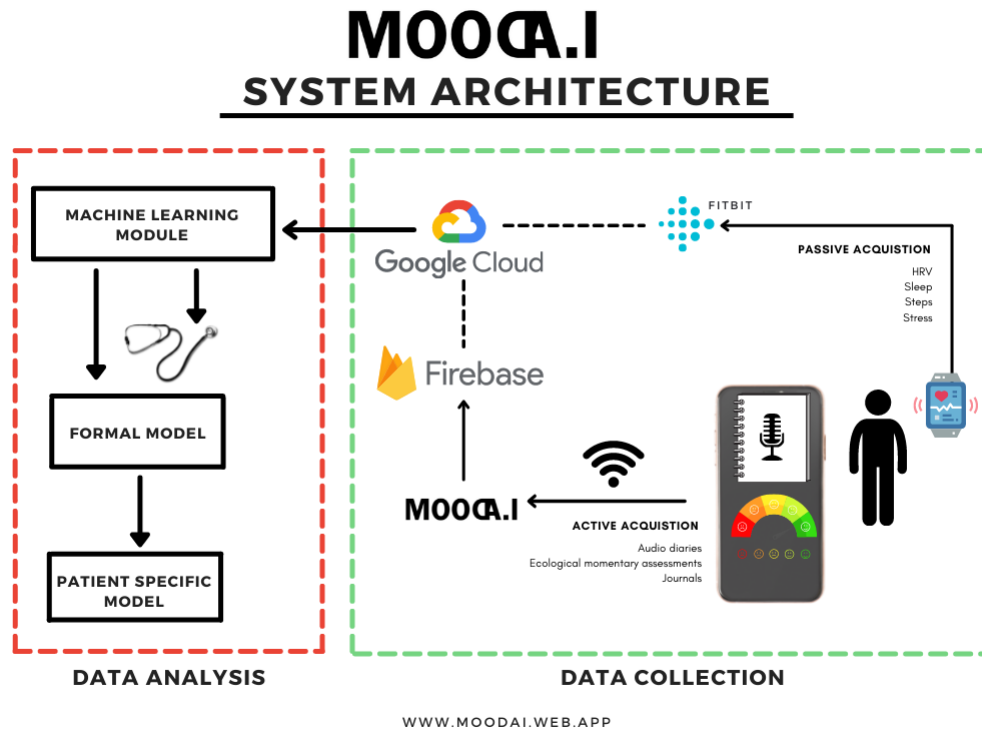


Figure 3.12: Overview of the MoodAI System Architecture

Additionally, participants were encouraged to provide any additional feedback about any aspect of the study. Table 3.1 indicates some notable feedback points.

Participant Id	Feedback
0	Mood ratings provided a good reflection opportunity
1	Found it uncomfortable to sleep at night with smartwatch
3	Definitely recommend this study to anyone willing to monitor their physical and mental health
4	Weekly stats about how you are doing would be useful
5	Would be useful to show you how many you've done during the day
7	Not too many things happened daily to be discussed in the audio diary
8	Would be useful to show you how many you've done during the day.
9	Effects on own mental health were positive, the audio diaries allowed the participant to reflect on what to be grateful for. There were far more things to be thankful for than negative. Found the negative things were quite minor.
10	Completing the mood ratings at the exact time was difficult at times. Good, there was flexibility.
11	Thought the study was a good experience, the first real study that I've been a part of. Struggled to meet the 5 times a day requirement for EMAs because it didn't fit my schedule.
12	MoodAI system was simple and easy to use. Fitbit system provided too much detail, and this could overwhelm depressed individuals

Table 3.1: Participant feedback about MoodAI Study

Chapter 4

Shah et al dataset

4.1 Introduction

Data collected and analysed as part of the paper "Personalised machine learning of depressed mood using wearables" [48] written by Shah et al was provided under a co-authorship agreement for any future publications resulting from work done with their dataset. The dataset was collected from 14 participants with moderate depressive symptoms assessed using the PHQ-9. As stated in Section 2.2 this data was collected in the years prior to the emergence of the COVID-19 global pandemic.

I would like to thank the team of Shah et al, for providing the data used in this analysis. This data collected from depressed participants allowed us to test deep learning models on depressed data despite not having recruited any depressed participants in the MoodAI study due to the COVID-19 pandemic.

4.2 Data preparation

Shah et al provided their raw and featurised datasets in zip files. After examining the files supplied by Shah et al, the featurised dataset was the ideal choice for use in the ML models as it included all the features pre-processed by Shah et al into the 43 different features as shown in the paper's supplementary information (see Appendix A). The featurised dataset consisted of CSVs files containing data collected from 14 participants throughout a one-month study participation period.

Before the data from the provided files can be used in the ML models pre-processing steps were needed to remove any issues with the dataset. Python was used to analyse and pre-process the data. We used Pandas [115] [116] one of the

most popular software libraries in Python for data manipulation and operations. These CSV data is read in and stored as a DataFrame data structure. After conducting preliminary exploratory data analysis the following issues were found in the dataset.

Firstly, for some participants, the HRV feature titled "ppg_std" contained what appeared to be a missing or invalid value. The HRV feature indicates the standard deviation of PPG values captured from the Samsung Galaxy smart-watch's sensor within a ± 15 -minute window around each depressed mood EMA. The value shown was consistently 999, which we deemed to be the default value recorded if an error or missing value scenario occurred. It would not have been possible to train using varying amount of features for each sample. As a result, all data from five participants (ids 10, 15, 18, 21, and 23) were removed due to the aforementioned issues with the HRV feature.

Secondly, we removed rows within the DataFrame of each participant which contained values that were considered NaN. These values would cause errors if used to train the models and thus needed to be removed.

Thirdly, we removed participant id 24 from the dataset as there were issues with the values recorded from the 2 physical activity and sleep-based features titled "exercise_calorie", "exercise_duration", and "prev_night_sleep" which all recorded a value of 999 for every entry. Similar to the first issue relating to HRV the value of 999 seems to indicate a missing or invalid value. For some participants, only part of the data was missing for the sleep feature and we did not want to remove any more data as there was already a limited quantity. Instead of removing the affected rows, we calculated a mean sleep duration for that participant using all non-zero values and assigned that as the missing or invalid sleep value. The most common method is last value carried forward [117], but in the context of sleep we felt it was more appropriate to calculate a mean as it would be more representative of a typical night.

Finally, the dataset was not arranged chronologically, so we sorted the DataFrames using the "datestamp" column as the sorting key.

After pre-processing the dataset, we were left with 8 participants with varying sample lengths as shown in Table 4.1. Assuming the duration of one month and an EMA frequency of 4 times a day, the maximum amount of samples per participant is 120. Notable participants are ids 14, 20, and 29, who have workable samples of less than 50% of the expected amount. It is expected that the models for these participants may not perform as well as the models for other participants due to the lack of training data available.

Participant ID	Number of samples
1	82
12	98
14	33
19	99
20	55
26	105
28	105
29	59

Table 4.1: Number of samples for each participant after data preparation

4.3 Study Methodology Comparison

As discussed separately in Chapter 2 and 3, there are several key differences between the studies. Firstly, there was no structured clinical interview conducted as part of the Shah et al study and depression symptom diagnoses were done solely using the PHQ-9. In the MoodAI study, we performed detailed participant recruitment and a clinician administered several clinical instruments (See Chapter 3). The Shah et al study captured neurocognitive assessment combined with EEGs which were performed in a lab environment, which the MoodAI study did not capture. MoodAI developed a custom web app that collected the necessary data whereas, Shah et al utilised a third-party app called MindLog. HRV was also measured differently, Shah et al calculated the standard deviation of values from the smartwatch’s PPG sensor around a ± 15 -minute window, whereas MoodAI calculated the RMSSD around a ± 15 minute window, a more robust HRV measurement [118].

The idea of combining the provided Shah et al dataset and the MoodAI dataset was something that was explored. Upon further investigation, there were significant differences in sensor data and methodology. Firstly, the sampling frequency for EMA was 4 times a day for Shah et al and 5 times a day for MoodAI. The format of the EMAs was also different as Shah et al split it into anxiety and depression and used a 7-point Likert scale, whereas MoodAI only assessed general mood and used a 5-point Likert scale. Shah et al use an older Samsung Galaxy smartwatch whereas, MoodAI utilises a newer Fitbit Sense smartwatch. The sensors and algorithms used by these two companies are different and are not disclosed/proprietary, it would have been difficult if not impossible to account for these differences.

Given these differences, we came to the conclusion that given the difficulty in adapting the datasets and the fact that it may not be fair for either dataset to combine as there were differences in sensors, we decided to analyse them separately in this Chapter and in Chapter 5.

4.4 Machine Learning Models

The goal of the DL models was to replicate the steps taken by Shah et al but utilise a different analytical approach so we can create a head-to-head comparison between the two different approaches.

The ML Models discussed in this Chapter and in Chapter 5 were developed using the Python DL API Keras [119] which utilises the open source machine learning software library developed by Google TensorFlow [120].

We developed a MLP model architecture with an input layer, 2 hidden layers, and an output layer. A set seed was used when training models so that any random number generation (for example during the random generation of model weights) can be reproduced. Due to the low quantity of sampled data, we utilised an approach called stratified k-fold for cross-validation [121]. Similar to the regular k-fold, data is split into k segments where k-1 segments are used for training and 1 segment is used for testing the model. The difference with the stratified k-fold is that data is split such that the folds preserve the percentage of samples for each class. This method is particularly useful when the dataset is small as it allows samples which are uncommon to both be present in the train and learn set, resulting in better model performance [122]. Shah et al used a nested k-fold cross validation method to improve the performance of a small dataset, whereas our models used stratified k-fold to accomplish the same result.

Hyper-parameter tuning was also done to the model optimise model [123] by tuning parameters such as batch size, epochs, and number of neurons. This was done using a grid search approach using the Python itertools library by training models on a set of possible combinations of hyper-parameters.

The DL model was trained to predict depressed mood ratings by using two different sets of data, one using the entire 43 feature data set and another using a subset of 14 features which were easily attainable using the currently MoodAI system architecture as described in Figure 3.12 and Chapter 5. These included features were features that were easily completed and acquirable remotely such as EMAs and lifestyle data captured from smartphone and smartwatch sensors.

The excluded features were features acquired from the neurocognitive assessment and EEG measurements which were conducted by Shah et al. Still, since these features needed to be during a scheduled laboratory visit, we deemed this was too inaccessible for the average participant.

A brief description of the chosen 14 feature subset can be seen in Table 4.2, a more detailed explanation of all 43 features can be found in Appendix A.

Feature Number	Feature Name	Feature Description
1	Anxious	How anxious the participant was at the time of EMA on a scale of 1-7
2	Distracted	How distracted the participant was at the time of EMA on a scale of 1-7
3	Past day fats	Total fatty items consumed in last 24 hours prior to EMA rating
4	Past day sugars	Total sugary items consumed in the last 24 hours prior to EMA rating
5	Past day caffeine	Total cups of caffeine consumed in the last 24 hours prior to each EMA rating
6	Mean heart rate \pm 30 mins	Mean heart rate from smartwatch within 15 minute window of EMA rating
7	PPG standard deviation \pm 15 mins	Standard deviation of PPG data from smartwatch within 15 minute window of EMA rating
8	Cumulative step calorie	Total steps taken in the last 12 hours prior to EMA rating
9	Cumulative step speed	Average velocity of all walking in the last 12 hours prior to EMA rating
10	Cumulative step distance	Total step distance travelled in last 12 hours prior to EMA rating
11	Exercise calorie	Total calories burnt from exercise in the last 24 hours prior to EMA rating
12	Exercise duration	Total exercise duration performed in the last 24 hours prior to EMA rating
13	Previous night sleep	Hours slept in the night previous to EMA rating
14	Time of day	Hour when the EMA was taken

Table 4.2: Table of selected subset of features and description

4.5 Model Results Comparison

Shah et al created 7 different regression and ensemble learning-based ML models with varying accuracies. The best ML method for each participant was compared to the DL models developed in this thesis. The models were compared using 4 different metrics being mean MAPE, std MAPE, mean MAE, and std MAE. MAPE (formula shown in Equation 4.1) indicates the percentage difference between the model’s predicted output and the actual output. MAE indicates the difference between the model’s predicted output and the actual output but in absolute terms. We recreated these metrics using the same methods as Shah et al. The MAPE and MAE is calculated for each fold of the k-fold cross validation and an average and standard deviation is computed using the list of MAPE and MAE values. Results from the MLP were outputted from 0-6 and needed to be adjusted to match the 7-point Likert scale used by Shah et al otherwise there would be inconsistencies in the MAPE equation (see Equation 4.1).

$$MAPE = \frac{1}{n} \sum_{k=1}^n \left| \frac{P_k - A_k}{A_k} \right| \quad (4.1)$$

Where A_k is the actual value and P_k is the predicted value

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - x_i| \quad (4.2)$$

We compared our DL approach to the best model for each of the participants in the Shah et al dataset. Shah et al used a nested k-fold cross validation method with 4 outer folds and 10 inner folds. In our models, we explored a variety of different folds for each participant to determine what number of folds would provide the best model results. There were 4 metrics across 8 different participants for a total of 32 total comparisons between the DL model and the best Shah models. The following sections discuss in detail the results of the two different datasets.

4.5.1 Entire Data Set Results

The best performing folds were the 3 and 4 fold implementation, Tables 4.3 4.4 show a side by side comparison of the MAPE and MAE metrics from a these folds against the best models from the Shah et al dataset. Results from other folds can be found in Appendix A Tables I.1 I.2 I.3.

The difference row for each participant represents the difference between the result from the Shah et al model and DL models. A positive difference value

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	6.56%	1.44%	0.317	0.045
	Diff	0.99%	4.11%	0.041	0.246
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	31.62%	5.27%	0.682	0.116
	Diff	-5.35%	9.17%	-0.032	0.214
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	74.62%	50.68%	1.667	0.758
	Diff	-33.74%	-38.81%	-0.660	-0.423
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	31.45%	5.99%	0.808	0.151
	Diff	-2.34%	0.25%	-0.157	0.051
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	31.63%	10.08%	1.006	0.265
	Diff	-25.23%	-3.17%	-0.798	0.002
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	56.02%	3.07%	1.644	0.041
	Diff	-19.61%	6.56%	-0.492	0.176
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	29.20%	11.27%	0.895	0.340
	Diff	-7.97%	-3.71%	-0.238	-0.209
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	65.12%	10.16%	1.175	0.174
	Diff	-1.98%	15.97%	0.099	0.148

Table 4.3: Comparison of best Shah et al regression based machine learning models using 3 fold cross validation MoodAI deep learning model using Shah et al data set

indicates that the DL model performed better than that of the best Shah et al model in that metric whereas a negative value represents the opposite.

Results from tuned 3 and 4-fold models appeared to perform the best among all tested folds. Both these folds performed better than the Shah et al model in 14 out of 32 total metrics. Something notable about our DL models was that the standard deviation of MAE and MAPE for 7 out of 8 participants was lower than that of the Shah et al model. Where the DL model failed was achieving a consistently lower mean MAPE.

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	8.28%	1.50%	0.402	0.047
	Diff	-0.73%	4.05%	-0.044	0.244
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	32.45%	4.50%	0.672	0.115
	Diff	-6.18%	9.94%	-0.022	0.215
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	70.34%	4.56%	1.788	0.223
	Diff	-29.46%	7.31%	-0.781	0.112
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	37.56%	11.21%	0.862	0.245
	Diff	-8.45%	-4.97%	-0.211	-0.043
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	25.76%	12.63%	0.760	0.306
	Diff	-19.36%	-5.72%	-0.552	-0.039
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	38.87%	5.29%	1.288	0.140
	Diff	-2.46%	4.34%	-0.136	0.077
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	23.33%	4.35%	0.725	0.094
	Diff	-2.10%	3.21%	-0.068	0.037
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	59.81%	14.04%	1.210	0.230
	Diff	3.33%	12.09%	0.064	0.092

Table 4.4: Comparison of best Shah et al regression based machine learning models using 4 fold cross validation MoodAI deep learning model using Shah et al data set

4.5.2 Subset Results

Multiple different folds evaluated to determine the best performing model. The four fold method appeared to be the best among the folds. Table 4.5 provides a head-to-head comparison of MAPE and MAE metrics of the 4 fold implementation against the best models from the Shah et al paper. Tables from other folds can be found in Appendix I Tables I.4 I.5 I.6 I.7. Similar to the resultant tables from the entire dataset, the difference row for each participant represents the difference between the result from the Shah et al model and DL models. A positive difference indicates that the DL model performed better than that of the best Shah et al model in that metric.

Results from the tuned 4 fold appeared to be the best among all the different folds. The 4-fold model proved better than the best Shah et al models in 17 out of 32 metrics. Except for participants 20 and 14, in all other instances where the 4-fold model performed worse, the difference was within a small margin of error (see Table 4.5). We notice a trend in the standard deviation metrics where the DL model performed better in 14 out of 16 metrics. This is the opposite with mean MAE and MAPE metrics where Shah et al perform better in 13 out of 16 metrics. One consideration that needs to be made with the results from participants 14 and 20 is that their sample sizes are considerably lower than that of the other samples with only 33 and 55 samples respectively. The lack of samples may have negatively impacted the ability of the DL model to draw any conclusive relationships between the features and depressed EMA rating.

Graphs showing a direct comparison between different folds can be seen in Appendix I Figures I.1 I.2 I.3 I.4. The graphs reinforce the conclusions drawn from the tables about the standard deviation of MAE and MAPE being consistently lower than that of the best Shah et al results and the mean MAE and MAPE being consistently higher than that of the best Shah et al results.

4.6 Shah et al Conclusions

In conclusion, the deep learning approach we developed appeared to perform on par with that of the best models from the Shah et al work. Results from the reduced feature set indicated that it performed better than the full feature set. It is possible that in the full feature set there may only be a handful of features with high importance, other less relevant features may be hindering the model and making it harder for the model to make the correct inferences. The deep learning models showed consistent reductions in model variance but were overall worse in terms of mean MAPE and MAE. Although the accuracy of our models perform slightly worse than that of Shah et al, the lower variance means that the consistency of the DL model is greater than that of the Shah et al mode. The range of error values from the model remains in a smaller range given different combinations of training data. Since deep learning models require a large amount of data to properly train, the limited quality of samples in the Shah et al dataset may be an explanation for these results. Given more data, results indicate that deep learning models could potentially improve on traditional regression and ensemble learning methods.

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	7.47%	2.05%	0.364	0.081
	Diff	0.08%	3.50%	-0.006	0.210
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	31.96%	4.98%	0.721	0.174
	Diff	-5.69%	9.46%	-0.071	0.156
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	67.68%	4.20%	1.535	0.302
	Diff	-26.80%	7.67%	-0.528	0.033
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	36.07%	5.03%	0.850	0.095
	Diff	-6.96%	1.21%	-0.199	0.107
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	26.06%	12.03%	0.777	0.337
	Diff	-19.66%	-5.12%	-0.569	-0.070
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	38.80%	9.55%	1.240	0.179
	Diff	-2.39%	0.08%	-0.088	0.038
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	25.91%	5.78%	0.821	0.120
	Diff	-4.68%	1.78%	-0.164	0.011
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	56.07%	15.33%	1.138	0.174
	Diff	7.07%	10.80%	0.136	0.148

Table 4.5: Comparison of best Shah et al regression based machine learning models using 4 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set

Chapter 5

MoodAI Data set and Machine Learning Pipeline

5.1 Data Retrieval

As discussed in Chapter 3 participants underwent a one-month study observation period. During the one month, data was collected from their study provided One-Plus Nord smartphone (as seen in Figure 3.2) and Fitbit Sense smartwatch (as seen in Figure 3.3).

The MoodAI architecture (as seen in Figure 3.12) collected data from various sources. Firstly, the MoodAI web app facilitated the capture and storage of EMAs, audio dairies, and end-of-day journals. EMAs and end-of-day journals were stored as documents in the Firebase Firestore database. We created two separate collections for EMAs and end-of-day journals, with documents for each separate entry. Audio dairies were stored in Firebase’s Cloud storage. Data was pulled from these sources using custom Python scripts created using the Firebase Admin Python SDK [124]. Participant EMAs were stored in CSV files whereas audio dairies were stored as WAV files.

Secondly, participants were given a study Fitbit Sense smartwatch data to wear over the one-month observation period. Data from the smartwatch was regularly synced with the study-provided smartphone via Bluetooth and sent directly to Fitbit servers for storage. Fitbit servers’ data were retrieved using the Python Fitbit API [125] and Fitbit Web API [126]. Data was extracted using RESTful requests based on the submission times for each EMA.

5.2 Feature Extraction

Data from Fitbit included a variety of lifestyle features, we grouped closely related features into clusters. For the MoodAI dataset, we split the lifestyle data into activity, sleep, and heart clusters. We were unable to understand the implementation of all features fully. This was due to Fitbit keeping many formulas and algorithms confidential. The subset of features selected in the cluster was selected from the list of features available on the Fitbit Web API [126], excluding those that required hardware not present on the Fitbit Sense smartwatch.

5.2.1 Activity Cluster

The activity cluster groups together various features related to a participant’s level of physical activity. Table 5.1 indicates the selected activity cluster features and descriptions of what they represent. All features in the activity cluster relate to any activity undertaken on the day of the EMA rating. A study conducted by Lu et al [69] using a Fitbit Charge HR includes many of the same activity features.

5.2.2 Sleep Cluster

The sleep cluster includes features relating to each participant’s quality and quantity of sleep. These features include the breakdown of sleep stages, sleep duration and time spent in bed. Table 5.2 shows the complete list and description of the features included in the sleep cluster. All features in the sleep cluster relate to the sleep session on the day before each EMA rating. A study conducted by Lu et al [69] selected similar sleep features to features included in this cluster.

5.2.3 Heart Cluster

The heart cluster includes features relating to measurements taken about the heart for each participant. These features include heart rate and heart rate variability. Features from this cluster are calculated within a window around each EMA rating.

5.3 Data Preparation

After successfully collecting the raw data, several steps needed to be undertaken such that the dataset could be used to train the ML models.

Feature Number	Feature Name	Feature Description
1	Time of Day	Hour of the day when the EMA rating was taken (0 - 23)
2	Daily Steps	Number of steps reported by Fitbit Sense smartwatch for the day of the EMA rating
3	Very Active Minutes	Total minutes the participant where the Fitbit Sense smartwatch reported the participant was very active on the day of the EMA rating
4	Fairly Active Minutes	Total minutes the participant where the Fitbit Sense smartwatch reported the participant was fairly active on the day of the EMA rating
5	Lightly Active Minutes	Total minutes the participant where the Fitbit Sense smartwatch reported the participant was lightly active on the day of the EMA rating
6	Sedentary Minutes	Total minutes the participant where the Fitbit Sense smartwatch reported the participant was sedentary on the day of the EMA rating

Table 5.1: Table of MoodAI Activity Cluster features and descriptions

Firstly, it was expected that there would be errors and missing values in the dataset. This may have been due to participants missing submissions or unexpected errors with the Fitbit smartwatch. Therefore, rows containing NaN values were removed from the dataset.

Secondly, the time of day feature was generated as part of the data preparation process, this was done by extracting the hour component from each EMA submission date time stamp.

HRV was only provided from Fitbit during hours when the participant was asleep, instead, we calculated an extrapolation of HRV using an RMSSD formula (See Equation 5.1). We calculated RMSSD in 1-minute segments and then calculate the average HRV using all 1-minute seconds in a ± 15 -minute window around each EMA rating.

The RMSSD formula utilises Inter-beat Interval (IBI), but since Fitbit does not provide access to individual IBI, we extrapolate it using HR data that has up to 1-second granularity (see Equations 5.3 and 5.2). We also performed a weighted

Feature Number	Feature Name	Feature Description
1	Time of Day	Hour of the day when the EMA rating was taken (0 - 23)
2	Light Sleep Duration	The length of time in seconds the participant was in the light sleep stage
3	Deep Sleep Duration	The length of time in seconds the participant was in the deep sleep stage
4	REM Sleep Duration	The length of time in seconds the participant was in REM sleep stage
5	Wake Duration	The total number of minutes the participant was awake
6	Minutes After Wake up	The total number of minutes it took for a participant to get back to sleep after waking up during a sleep session
7	Minutes to Fall Asleep	The number of minutes the participant took to fall asleep
8	Time in Bed	The total number of minutes the participant was in bed
9	Minutes Asleep	The total number of minutes the participant was asleep

Table 5.2: Table of MoodAI Sleep Cluster features and descriptions

Feature Number	Feature Name	Feature Description
1	Time of day	Hour of the day when the EMA rating was taken (0 - 23)
2	Average Heart Rate	Average Heart rate taken from all recorded heart rate values within ± 30 -minute window around EMA
3	Average HRV	Average 1 min RMSSD HRV values from all recorded heart rate values within ± 15 minute window around EMA
4	Resting Heart Rate	Resting heart rate from the day of EMA rating

Table 5.3: Table of MoodAI Heart Cluster features and descriptions

Participant No.	Number of samples
0	155
1	79
3	155
4	35
5	35
6	131
7	157
8	133
9	154
10	116
11	141
12	153
14	131
15	118

Table 5.4: Number of samples for each participant in the MoodAI dataset after data preparation

average by multiplying each IBI by the time difference till the next heart rate measurement and then dividing by 60 seconds to fill in any missing data.

After data preparation, one participant (Participant No. 13) was removed due to errors with heart rate and sleep data from their smartwatch, leaving us with data from 14 participants. Table 5.4 shows the number of samples per participant.

$$RMSSD = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (|IBI_{i+1} - IBI_i|)^2} \quad (5.1)$$

$$HeartRate(bpm) = \frac{60}{IBI} \quad (5.2)$$

$$IBI = \frac{HR}{60} \quad (5.3)$$

K-fold cross-validation techniques shuffle the data removing any sort of temporal component from the dataset. To retain a temporal component to the dataset, instead of training the model on 1D rows of features, we created a 2D row of features using a sliding window technique, this gives the model a small window of historical values to look back on. The Matrix shown below indicates how each 2D input matrix was constructed. x represents different features numbered 0 to n and t represents the timestamp of the feature from i to the total number of samples.

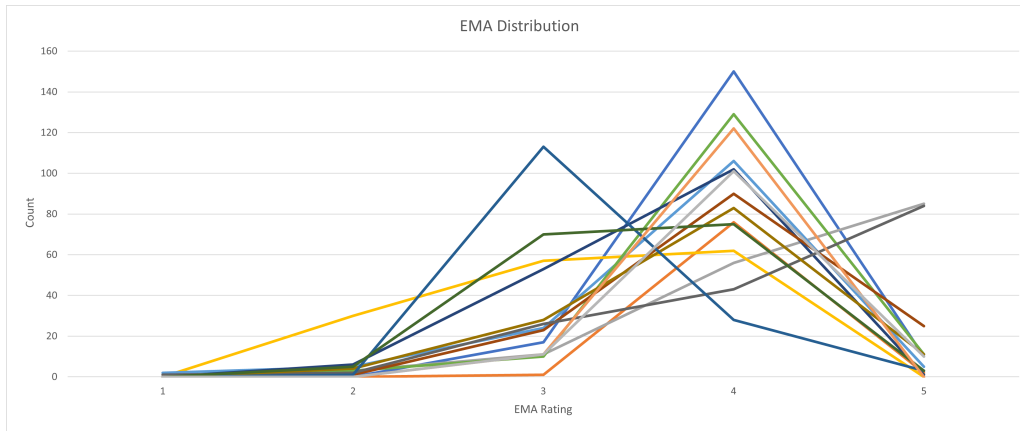


Figure 5.1: Distribution of EMAs across all participants

$$\begin{bmatrix} x_{t,0} & \dots & \dots & x_{t,n} \\ x_{t-1,0} & \dots & \dots & x_{t-1,n} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{t-i,0} & \dots & \dots & x_{t-i,n} \end{bmatrix}$$

Where n = number of features and i = size of sliding window

5.4 Exploratory data analysis

We performed an exploratory data analysis on the labels to determine the distribution of samples. As shown in Figure 5.1 we found that there was a heavy bias towards the EMA score of 4 for most participants which one participant showing a bias towards the EMA score of 3. Despite wanting a more bell-shaped distribution, we recognise that all participants were healthy controls and thus it was expected that mood states would tend towards higher scores. Table 5.5 shows the percentage distribution of submissions across all EMA scores. We can see that 10 out of 14 participants have a heavy bias toward the 4 EMA score, with over 50% of all their ratings being 4s. Participant 11 prefers the EMA score of 3 with a 78% frequency for that score. Whereas participants 3 and 9 had a bias towards the EMA score of 5 with over 50% of all their ratings being 5s.

Participant No.	1	2	3	4	5
0	0.00%	0.00%	9.6%	84.75%	5.65%
1	0.00%	0.00%	1.27%	96.20%	2.53%
3	0.00%	1.94%	7.10%	36.13%	54.94%
4	0.00%	20.13%	38.26%	41.61%	0.00%
5	1.41%	3.52%	16.90%	74.65%	3.52%
6	0.00%	1.96%	6.54%	84.31%	7.19%
7	0.00%	3.70%	32.72%	62.96%	0.62%
8	0.00%	0.72%	16.55%	64.75%	17.99%
9	0.64%	1.28%	16.67%	27.56%	53.85%
10	0.00%	3.17%	22.22%	65.87%	8.73%
11	0.00%	0.69%	77.93%	19.31%	2.07%
12	0.00%	3.27%	45.75%	49.02%	1.96%
14	0.00%	0.00%	8.27%	91.73%	0.00%
15	0.00%	0.00%	9.02%	82.79%	8.20%

Table 5.5: Distribution of EMA scores across all participants

5.5 Machine learning models

Similar to the models discussed in Chapter 4 The ML Models discussed in this Chapter were developed using the Python DL API Keras [119] which utilises the open source machine learning software library developed by Google TensorFlow [120]. As discussed below, we explored multiple different model architectures, but one common point among all models trained was the same EMAs as the labels.

For the following models, we used similar k-fold cross validation and hyper-parameter tuning methods. Hyper-parameter tuning was also done to optimise the model [123] by adjusting model parameters [123]. In the MoodAI models, we adjusted parameters such as neurons, epochs, batch size, and sliding window size. Similar to the Shah et al models discussed in Chapter 4 this was done using a grid search approach on the Python itertools library by training models on a set of possible combinations of hyper-parameters. Grid search is more computationally intensive but provides a more comprehensive search space when compared to random search.

Nested within each iteration of hyper-parameter tuning was stratified k-fold cross-validation. Stratified k-fold cross validation splits the dataset into k segments where k-1 segments are used for training and 1 segment is used for testing the model [121]. Stratified k-fold allows us to preserve the percentage of samples for each class among the folds. This is particularly useful for the MoodAI dataset as there is a bias towards the EMA score of 4 and samples of other classes are

infrequent. Similar to what we used for the Shah et al dataset, we used 4-fold stratified cross-validation for all models discussed below.

We created two models using the collected dataset, a monolithic model and a compositional model. Separate models were created for each participant using only their data, making each model participant-specific. The monolithic model consists of far more features than each cluster model, making it inherently more complicated to understand and more computationally intensive to train. The compositional model splits features into smaller models [45], trains and tunes each model separately, and then combined using an additional MLP. Although the compositional model has more individual components, each component is much easier to understand and explain. The goal of this work was to determine which model architecture performed better.

5.5.1 Multi-feature Monolithic Model

5.5.1.1 Model Architecture

We used an LSTM model with an input layer, 1 hidden layer, and an output layer for the multi-feature monolithic model. We used set seeds for any random number generation in model training so that results would be reproducible. The monolithic model includes all features from each cluster (except only one instance of the time of day feature) for a total of 17 features.

5.5.1.2 Results

Test accuracies from the tuned multi-feature monolithic model appear to be positive. Table 5.6 shows the test accuracies of the model ranging from 65% to 100%, averaging 87% across all participants in the dataset. Although the accuracies are high there are a few notable observations. Firstly, participants 9 only achieved an accuracy of 65.71%. If we refer to Table 5.5 we can see that participant 9 was one of the two participants to show a bias towards the EMA score of 5. Secondly, participants 1, 5, and 6 achieved extraordinarily high accuracies of 100%. Accuracies of 100% are generally not possible in a real-world environment that the test set is intended to simulate. These participants have shown to have the highest percentage of 4 EMA ratings; thus, our models may have also developed such bias.

Participant No.	Model Test Accuracy
0	94.44%
1	100.00%
3	80.00%
4	83.33%
5	100.00%
6	100.00%
7	82.05%
8	80.65%
9	65.71%
10	88.89%
11	85.71%
12	77.14%
14	96.77%
15	89.66%

Table 5.6: Results from Tuned MoodAI Monolithic LSTM Model

5.5.2 Multi-feature Compositional Model

The compositional model consisted of 3 separate clusters; activity, sleep, and heart clusters. Three separate LSTMs one for each cluster trained. Each LSTM was then optimised using stratified k-fold and hyper-parameter tuning to optimise each model. The best model parameters were then used to train a final decision layer. Figure 5.2 shows the implemented compositional neural network architecture.

Looking at the results from each cluster’s LSTM as shown in Tables 5.7, 5.2, and 5.9 we see similar results to the multi-feature monolithic model. Model accuracies range from 60% to 100% with the average model accuracy across all three clusters being approximately 86%. Similar to the multi-feature monolithic model, there still exists the problem of the large variations in model accuracy among participants. Participants 1, 5, and 6 showed 100% model accuracy in some of the LSTMs, this may be due to their high concentration of EMA scores of 4 resulting in the model overfitting. The lack of samples limited the ability of stratified k-fold to generate sufficiently different test samples while preserving class sample percentage. Participant number 5 has shown consistently high accuracies among all the models, one potential reason may be the lack of samples compared to other participants. However, participant 6 also achieves similarly high accuracies while having approximately 4x the samples of participant 5.

Participant No.	Model Accuracy
0	96.97%
1	100.00%
3	76.67%
4	68.97%
5	88.89%
6	100.00%
7	90.32%
8	80.77%
9	67.74%
10	91.67%
11	96.30%
12	79.31%
14	93.10%
15	88.89%

Table 5.7: Results from Tuned MoodAI Activity Cluster LSTM Model

Participant No.	Model Accuracy
0	96.97%
1	100.00%
3	76.67%
4	62.96%
5	88.89%
6	96.55%
7	90.32%
8	84.00%
9	66.67%
10	91.67%
11	96.30%
12	75.86%
14	89.66%
15	92.59%

Table 5.8: Results from Tuned MoodAI Sleep Cluster LSTM Model

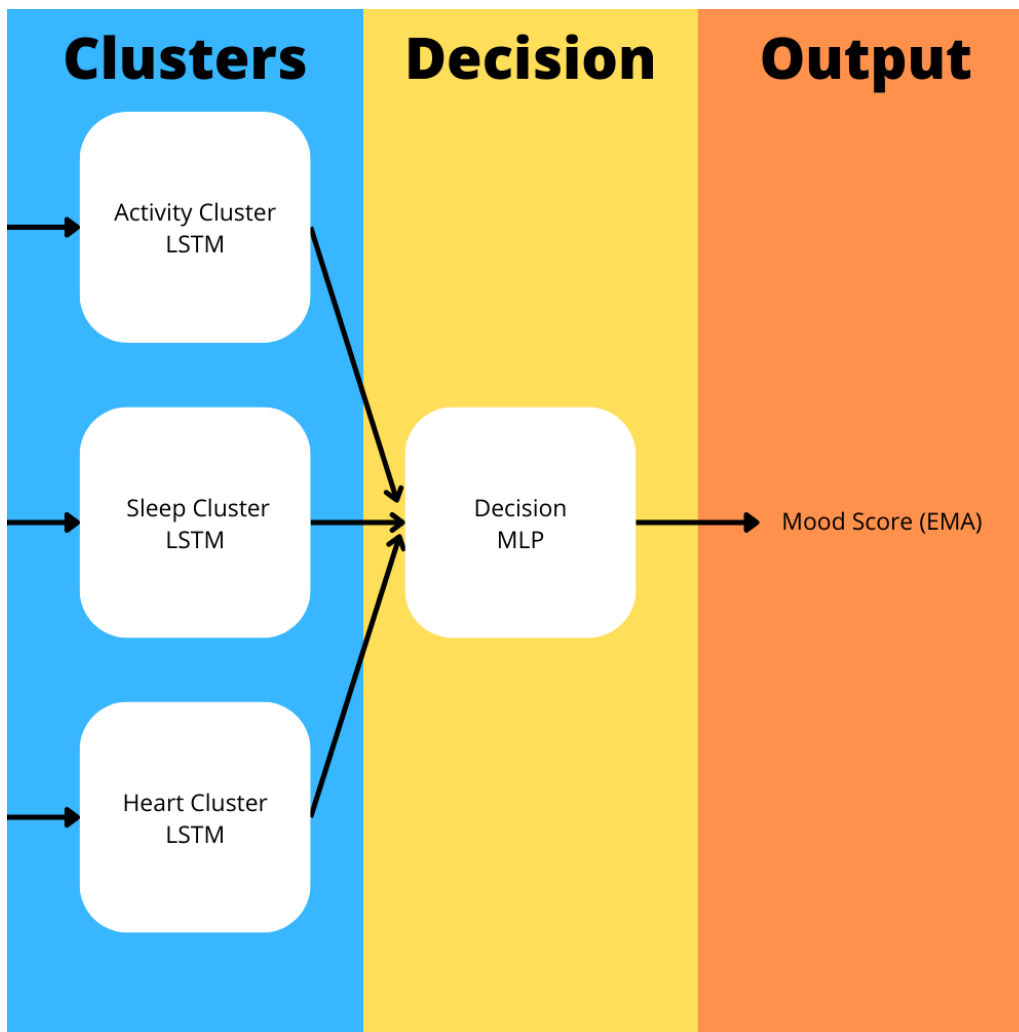


Figure 5.2: Proposed MoodAI Compositional Neural Network Model

5.5.2.1 Model Architecture

The final decision layer was a simple MLP with an input layer, a single hidden layer, and an output layer. The inputs to this model were the predicted EMAs from the models of each cluster (3 in total). We then trained the model using the actual participant EMAs as the labels. Table 5.10 indicates the model accuracies achieved from the decision layer MLP. Results appear to show that the compositional model performs better than the monolithic model. However, there were signs of overfitting for each cluster's LSTM. The overfitting in each of the clusters has likely been magnified as the input to the compositional decision MLP were the predicted

Participant No.	Model Accuracy
0	94.44%
1	100.00%
3	72.97%
4	85.71%
5	100.00%
6	100.00%
7	76.92%
8	80.65%
9	63.16%
10	92.59%
11	84.38%
12	75.00%
14	96.88%
15	89.66%

Table 5.9: Results from Tuned MoodAI Heart Cluster LSTM Model

Participant No.	Model Accuracy
0	96.23%
1	100.00%
3	91.23%
4	95.01%
5	100.00%
6	100.00%
7	71.88%
8	82.23%
9	64.69%
10	80.00%
11	72.41%
12	82.49%
14	98.21%
15	91.98%

Table 5.10: Results from Tuned MoodAI Compositional Decision MLP Model

EMA values from each of the separate models. This is due to limitations in the spread of samples contained in the dataset.

Chapter 6

Conclusions and Future Work

6.1 Limitations

The original study intention was to recruit a total of 20 participants with mild-moderate depression and 20 healthy controls. The recruitment of participants with mild-moderate depression was intended to be done via primary healthcare providers in Auckland such as Tamaki Health. This is because, of the need for a clinical confirmation of a present or past depression diagnosis. Additionally, this depressed population is potentially at risk and therefore there was a necessity to get confirmation from their general practitioner.

Recruitment for the MoodAI study began in August 2021 and continued through to May 2022. This period was the peak of the COVID-19 global pandemic in New Zealand. With the emergence of the delta and omicron variants of COVID-19 and the urgent vaccine roll-out, primary healthcare providers such as Tamaki Health needed to focus on meeting community needs. As a result, no participants with mild-moderate depressive symptoms were recruited.

Additionally, there were multiple lockdowns and social distancing measures put in place by the New Zealand government. Study screening was unable to continue in person without breaking the procedures put in place by the New Zealand government and the University of Auckland. The study adapted accordingly to lockdown restrictions and after conducting one in-person screening session, the study transitioned to online screening sessions via zoom and devices were couriered to suitable participants.

We supplemented the lack of participants in the depressed group by retrieving a data sample from Shah et al 4 which consisted entirely of participants with self-diagnosed depression.

Machine learning models require large amounts of data to be able to better perform classification tasks. As of June 2022, a total of 15 participants were recruited and completed a one-month monitoring period. Each participant was monitored for approximately 4 weeks and entered 5 ecological momentary assessments per day. This results in a maximum of 150 samples per person.

We were hoping for a normal distribution of EMA scores closely representing a bell curve. Since we were unable to recruit any participants with mild-moderate depression, we were unable to achieve the desired distribution. Instead we observed a heavy bias towards the EMA score of 4 in our healthy controls. This result was not unexpected as the healthy controls were determined to be mentally healthy during their clinical screening sessions.

Similar to many other studies in this area, we needed to build a larger dataset to improve the performances of our models [69]. However, building such a large dataset is difficult due to the nature of the data and the amount of time and resources required.

6.2 Strengths

Although we were unable to recruit a depressed group for the MoodAI study due to COVID restrictions, we acquired a published data set from a research group from the University of San Diego, California. This data set referred to as the Shah et al data set [48] consisted of similar bio-markers captured in the MoodAI study but was made up entirely of participants with moderate depression symptoms. This data helped to supplement the inability to collect data from depressed participants as part of the MoodAI study. The work done by Shah et al used only regression and ensemble learning-based machine learning methods so we were able to explore if there were any benefits to using deep learning approaches. We trained models using both the entire feature set and also a limited subset of the features. Results from the subset were on par with the best results reported by Shah et al. Although our MAPE and MAE were generally worse than that of Shah et al, our deep learning model did manage to consistently output a lower standard deviation in both MAE and MAPE measures.

Despite the results from our models being somewhat inconclusive, this thesis was able to provide a framework as to how deep learning models can be constructed using different digital bio-markers. The compositional neural network approach where we broke down features into different clusters and trained separate smaller

neural networks before combining them in a final decision layer to determine the overall outcome is something that to the best of my knowledge has not yet been applied in the context of depression classification poses some promising future research questions.

Overall, despite all the aforementioned limitations, the work discussed in this thesis successfully explores the idea of predicting mood states using bio-physical data collected from smartphones and smartwatches.

6.3 Future Work

6.3.1 Mood Index

Due to COVID-19 disruptions, we were unable to recruit any depressed participants as part of the MoodAI dataset. The EMAs were heavily clustered towards a certain rating and thus resulted in overfitting. Our research team has developed a mood index that is calculated by taking into account a trend in consecutive EMAs. A rolling mood index could then be used as the label to train and tune the ML models. This could show improvements over the current confusion matrix which is heavily biased towards the 4 EMA rating.

6.3.2 Audio and Speech Analysis

A current body of literature has shown speech patterns recorded on smartphones can be used to predict and monitor mood states [76]. As part of the MoodAI study, audio diaries were collected on a daily basis. Common prosodic features such as frequency and speaking rate along with cepstral features (Mel-frequency cepstral coefficients) can be analysed and included in future machine learning models.

6.3.3 Platform Agnostic

The MoodAI work utilised Fitbit Sense smartwatches for bio-physical data collection. For better accessibility, expanding supported smartwatches to support offerings from Apple, Garmin, Samsung, Huawei and other Google Wear OS-based smartwatches would provide a much further reach. This way participants could volunteer using their own personal smartwatches. This extra data could be used to improve the accuracy of the machine learning model.

6.3.4 Embedded Implementation

The compositional neural network (CpNN) approach discussed in the paper by Yang, Xin, et al titled "A compositional approach using Keras for neural networks in real-time systems" [45] greatly inspired the neural network models discussed in this thesis. Future work be done to utilise the developed semantics proposed by Yang, Xin, et al [45] to convert the CpNN to C code using their CpNN2C compiler which builds on top of the Keras2C tool developed by Conlin, Rory, et al [127]. This could then be run on an embedded wearable device like a smartwatch, thus providing further features and improving the mental health capabilities of smartwatches.

6.3.5 MoodAI Platform Availability

Rates of access to mental health services are currently inequitable with numerous factors contributing including stigma, cost, lack of time and difficulties with access. Digital technologies provide advantages such as 24-hour accessibility, low cost, and familiarity for users. The MoodAI platform could be further refined and could be made widely/freely available to help monitor/track depression as part of self-management or to aid clinicians with monitoring trends. This work could also be adapted for use in other mental health or physical health conditions such as chronic pain.

6.4 Summary of Thesis

In this thesis, we developed a longitudinal observational study where we acquired ethics approval, performed clinical screenings, and collected a variety of lifestyle and ecological momentary assessment data from human participants. A total of 15 participants were recruited and monitored over one month. We developed the MoodAI app for participants to interact with and capture EMAs, audio diaries, and daily journals. The MoodAI system architecture provides a framework for others conducting similar observational studies to follow.

From the system architecture, we constructed a dataset consisting of 17 features relating to physical activity, sleep, and heart rate. From this, a machine learning pipeline was developed to predict mood states using LSTM neural networks. While on the surface results from both the monolithic and compositional models appear promising with accuracies ranging from 60 to 100%. The inability

to recruit a depressed group due to the global pandemic resulted in limitations in the distribution of dataset labels. Healthy participants proved to have consistent mood ratings. The dataset's skew and the limited samples seemed to indicate the model suffered from overfitting issues.

We attempted to supplement the lack of depressed participants by collaborating with a research team from the University of San Diego, California, USA. This team collected a depressed dataset in the years before the pandemic. This collaboration yielded fruitful results as we were able to perform a head-to-head comparison of their regression and our deep learning-based approach. Using only a subset of the dataset we showed that the deep learning models discussed in this thesis performed similarly in terms of accuracy but performed better in terms of model variance when compared to their regression-based models.

In conclusion, although this thesis did not provide conclusive results on whether or not deep learning can be successfully used to perform mood classification, we developed a scalable framework upon which future research can build on.

Appendix A

Shah et al Supplementary Information

For copyright purposes the supplementary information by Shah et al has been removed.

It can be found online here:

https://static-content.springer.com/esm/art%3A10.1038%2Fs41398-021-01445-0/MediaObjects/41398_2021_1445_MOESM1_ESM.pdf

Appendix B

Participant Information Sheet (PIS)



Department of Psychological Medicine
Level 3, Building 507,
22-30 Park Avenue,
Grafton,
Auckland 1023, New Zealand
+64 (0) 9 923 6531

Postal address:
Department of Psychological Medicine
Faculty of Medical and Health Sciences
The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

PARTICIPANT INFORMATION SHEET

Study title: Real-time assessment of mood changes and machine learning

Name of researchers: Drs Frederick Sundram, Amy Chan and Partha Roop

Contact email address for primary researcher: f.sundram@auckland.ac.nz

Voluntary Invitation to Participate

You are invited to participate in this study because you are between the ages of 18 and 60 and interested in using smartphones and smartwatches to help us understand more about depression. Your general practitioner (GP) has either informed you about this study or contacted the study investigators about your interest to participate. Alternatively, you may have found out about this study via email distribution lists, social media (e.g., Facebook), word of mouth or advertisements. This study involves monitoring physiological health using a provided smartwatch (Fitbit Sense) and a series of audio recordings using the provided smartphone (OnePlus Nord). There will be a demonstration by the team of how to use these devices. Participation will result in data from these devices being collected over a month. This study is funded by the Health Research Council of New Zealand.

Overview of this study

Aim: This project aims to gather data from smartphones and smartwatches over a one-month duration. Most of the data captured will be passive without a need for you to do anything. However, there will also be data captured that would need your interaction either with the smartphone or smartwatch. The data captured will be used to develop a future system that can detect changes in mood in people with depression. Eventually, the data will inform the development of algorithms which can be used to automatically detect low mood in real-time and generate appropriate prompts to help individuals access mental health support more quickly.

Rationale: Depression is extremely common in New Zealand and across the world but there are problems with accessing mental health support in a timely way because mental health services are often stretched. Additionally, people tend not to present early for help with their depression. A system that helps monitor changes in mood on a 24/7 basis will help identify when someone may be becoming depressed and aid with accessing mental health support earlier.

Duration: Should you be willing to participate, your involvement in the study will be for up to one month. During this timeframe, most data will be acquired without the need for your interaction with the smartwatch and smartphone. Some of this data will be acquired when you are asleep at night while other data will be acquired when you are awake.

Benefits: The smartwatch can provide you with detailed insight into your physiological wellbeing. Readings such as heart rate, breathing rate, heart rate variability and sleep quality may offer insight into your health. Also, your data will help develop a system that will help others in addressing their depression and accessing mental health supports earlier.

Study Design

In this study, we will be recruiting people with depression and healthy individuals. All the data from the smartwatches will be gathered passively i.e., without your interaction with the device and will be acquired during sleep and during awake hours. Data gathered will include physiological measures such as heart rate, breathing rate, heart rate variability, sleep quality, number of footsteps taken during the day and distance travelled. However, data from smartphones will require interaction including ecological momentary assessments (mood ratings) and daily audio diaries. To recruit the number of participants we require, the project's expected duration will be one year; however, each participant will only be involved for up to one month. As part of assessing suitability to participate, there will be clinical screenings taking place at the start of the study via interview at the Clinical Research Centre at the University of Auckland in Grafton. This would ideally be done in person, but should there be any Covid-19 related disruptions, this may need to occur remotely via Zoom.

What will participating in the study involve?

Participation in the study will require the completion of an initial screening in person to assess suitability to participate. There will be one visit to the Clinical Research Centre at Grafton, at the beginning of the study when you will be provided with the required devices and instructions after undergoing screening. It is expected that this visit would take up to three hours. If considered suitable, there will be a demonstration of how to use the devices (both smartwatch and smartphone). You will then be provided with the devices (Fitbit Sense and OnePlus Nord) and a SIM card for the study-supplied smartphone. The Fitbit Sense is to be worn during sleep and for most of the day. Exceptions are made for wearing the Fitbit Sense for example when charging of this device and when having a shower or when there is any discomfort in your wrist. Participants are also expected to record short audio diaries (2-5 mins in length) once a day via the study-provided smartphone. These audio diaries will be completed through a web application called MoodAI.

Of note, all the data captured through the Fitbit Sense is passively acquired without any interaction whereas the acquisition of audio diaries will require interaction with the smartphone. You do not need to create any user accounts for Fitbit, this will be done for you and you will be provided with details of the account. This account will sync your physiological health data with the Fitbit Health App. To allow transfer of data from your Fitbit device or smartphone, the smartphone will need to be connected to the internet. Audio diaries will be done through the MoodAI web application daily and well as ecological momentary assessments

up to four times daily – these are mood ratings on a scale from happy to sad. There will be notifications to prompt you to complete these tasks during the day.

At the end of the study, we can arrange to collect the study-provided devices from your home to minimise travel disruption.

Inclusion criteria

- Participants are male or female, aged 18 to 60 years.
- Participants are willing and able to give informed consent for participation in the study.
- Participants are willing to undergo a clinical screening interview by a member of the research team.
- Participants are willing to use the smartphone and smartwatch for the duration of the study.

Exclusion criteria

- Inability to speak or read English to a level that enables informed consent and/or participation in the study.
- Those who are unable or disinterested in using the technology.
- Any other condition judged by the research team as likely to impact on the ability to complete the study.
- Substance abuse or dependence in the last six months.
- Any other unstable medical or neurological condition.

Risks and Incidental findings

There are no immediate risks from participating in this study. We do not expect any incidental findings to arise because of this study. However, the research team will advise your general practitioner (GP) of your participation if the level of your depression is deemed safe for participation in this study. Should you experience any worsening of your mood during the study, you should speak with your GP - clinical supports are additionally available for you to access should your mood worsen during the study while you might be awaiting an appointment with your GP (see support hotline numbers in the contact details section). The research team may exclude you from participation should your depression levels worsen at the start or over the course of the study and the research team is not providing clinical care or support to you. Should you not have depression, the research team will not be contacting your GP.

Compensation and Reimbursement

You will be reimbursed with \$100 worth of supermarket vouchers and up to \$20 for transport costs. Supermarket voucher reimbursement will be split into two stages. \$50 at the beginning of the study and a following \$50 if you remain till the end of the one-month study. We will cover your transport costs at the start of the study.

Withdrawal from the study

Your participation is entirely voluntary, and you may decline the invitation to participate. If you choose not to participate, you may contact any of the investigators found in the contact details section. You are free to withdraw from the study after this initial clinical screening without giving any reasons. Of note, the data collected from the time you consented to participate in the study till the time you withdraw from the study may be used. If you withdraw from this study, you must return any devices provided to you as part of the study. These include the Fitbit Sense smartwatch and the OnePlus Nord smartphone. They must be returned timely, and in good condition, otherwise, you may be liable for the device.

Data Collection Procedures

You will be provided with the choice of up to two pieces of hardware:

- 1) a Fitbit Sense smartwatch and;
- 2) a OnePlus Nord Android smartphone

The Fitbit Sense smartwatch will be provided, but the OnePlus Nord Android smartphone will be offered to participants who do not wish to use their personal phone for the study, or for those who do not have a smartphone that is able to pair with the provided smartwatch or submit audio diaries.

You will also be provided account details which will be used to access the different services required for this study.

GPS functionality of provided devices will be always enabled. This is solely to ensure that the devices are not stolen and to aid with recovery.

At the end of the study, you must return **all** devices provided to you for the purpose of the study. A researcher will contact you to organise the most convenient way for you to return the devices. The default is for a researcher to schedule a delivery pickup.

Data Collected

Your physiological health data will be collected over a one-month period and will be sent to our data collection server without needing your input.

Fitbit data:

You are required to wear a Fitbit Sense smartwatch throughout the day and night. This device monitors physiological wellbeing. The data captured are:

- Resting Heart Rate
- Instantaneous Heart Rate
- Heart Rate Variability – During sleep
- Breathing Rate – During sleep
- Skin Temperature – During sleep
- Oxygen Saturation – During sleep
- Sleep Score – During sleep



The OnePlus Nord smartphone and Fitbit Sense smartwatch provided through this study will already be pre-paired via Bluetooth. The OnePlus Nord smartphone will have the Fitbit app preinstalled. This app will periodically transfer the physiological data captured/stored on the smartwatch to the OnePlus Nord smartphone and onwards to online Fitbit servers. This data will be stored under the participant's provided Fitbit account.

The study-provided Fitbit account will be created through Google services and the name and email address will be de-identified so you will not be identifiable. Fitbit data will be retrieved manually by the research team. This data will be exported at the end of the first week and month/end of study.

MoodAI web app:

The MoodAI web application will be hosted on the Firebase Google Cloud Platform. This web application has security/authentication built in via the Google Cloud Platform and participants can sign in using their provided Google account. This web app will be maintained by members of this study who are part of the Department of Computer Engineering at the University of Auckland.

The purpose of the web app is for participants to be able to upload daily self-recorded audio diaries and complete mood ratings (ecological momentary assessments). This web app will prompt users to complete these items daily using notifications either via text messaging/SMS or native notifications on the phone.

Audio diaries:

You are required to record a daily audio diary of 2-5 minutes in duration. This will be done through the study specific MoodAI web app. Access to this web app will be via a web link using a web browser. Authentication and access will be via the study-provided Google account (do not use your own personal Google account if you have one). Instructions on what you should record will be provided by the web app. Audio diaries must be done in English and not any other language, as a result a level of proficient English is a requirement to participate in this study. These audio diaries will be stored on a secure Google database managed by researchers of the study.

Ecological momentary assessments (EMAs):

EMAs are regular mood ratings over the course of the day. EMAs will also be completed via the MoodAI web app and should be done up to 4 times daily and at the time audio diaries are completed. It will be a simple scale indicating how you are feeling at the time of capture. This data will also be stored on a secure Google database managed by researchers of the study.

Study conclusion

The study-provided smartphone and smartwatch will be collected by the research team. There will be a final export of Fitbit data and check if EMA and audio diaries have been successfully acquired. The study-provided Google account will be discontinued, and the OnePlus Nord smartphone and Fitbit smartwatch will be reset to factory settings with deletion of your data on these devices.

Data storage, retention, destruction, and future use

During this study, the research team will record information about you and your study participation. This includes the results of any study assessments. If needed, further information from your hospital records and your GP may also be collected.

Identifiable Information

Identifiable information is any data that could identify you (e.g., your name, date of birth, or address). Only the research team will have access to your identifiable information.

De-identified (Coded) Information

To make sure your personal information is kept confidential, information that identifies you will not be included in any report generated by the research team. Instead, you will be identified by a unique code. The research team will keep a list linking your unique code with your name, so that you can be identified by your coded data if needed. The results of the study may be published or presented, but not in a form that would be expected to identify you.

Future Research Using Your Information.

If you agree, your coded information may be used for future research performed by members of this research team related to interventions for depression. If you agree, your coded information may also be used for other medical and/or scientific research that is unrelated to the current study. The primary investigator will review the study to determine ethical viability before any information is provided.

Your information may be used for 10 years after the study conclusion for future research unless you withdraw your consent. After that point, your information will be destroyed. However, it may be extremely difficult or impossible to access your information, or withdraw consent for its use, once your information has been shared for future research.

Security and Storage of Your Information.

Your identifiable information is held on a private web server named MoodAI during the study. This webserver will be hosted on the Google Cloud Platform and designed and maintained by researchers involved in this project and affiliated with the Computer Engineering Department at the University of Auckland. After the study, it is transferred to a dedicated Dropbox file storage system created for this project and maintained by the University of Auckland, and stored for at least 10 years, then destroyed. All storage will comply with local and/or international data security guidelines.

Risks.

Although every effort will be made to protect your privacy, absolute confidentiality of your information cannot be guaranteed. Even with coded and anonymised information, there is no guarantee that you cannot be identified. The risk of people accessing and misusing your information (e.g., making it harder for you to get or keep a job or health insurance) is currently very small, but may increase in the future as people find new ways of tracing information.

Although your coded information is maintained by researchers here in New Zealand, it will be hosted on the Google Cloud Platform which is located overseas.

Rights to Access Your Information.

You have the right to request access to your information held by the research team. You also have the right to request that any information you disagree with is corrected.

Please ask if you would like to access the results of your clinical screening and safety assessments during the study.

If you have any questions about the collection and use of information about you, you should contact the research team.

Rights to Withdraw Your Information.

You may withdraw your consent for the collection and use of your information at any time, by informing the research team.

If you withdraw your consent, your study participation will end, and the research team will stop collecting information from you.

Information collected up until your withdrawal from the study will continue to be used and included in the study. This is to protect the quality of the study.

Ownership Rights.

Information from this study may lead to discoveries and inventions or the development of a commercial product. The rights to these will belong to the study's Primary Investigator, Dr Frederick Sundram. You and your family will not receive any financial benefits or compensation, nor have any rights in any developments, inventions, or other discoveries that might come from this information.

Use of New Technologies (e.g., Artificial Intelligence, Health Apps).

The study will involve the use of the Fitbit Health App. This is necessary as it enables the retrieval of data from the Fitbit Sense smartwatch. Participants will be assigned Fitbit accounts that are de-identified. Anyone outside of the research team cannot use the data gathered and cannot be used to identify you. The data will be synced to Fitbit servers and is encrypted by Fitbit. You are not expected to subscribe to any paid Fitbit app features.

Results of the study

At the end of the study, you may request a copy of the physiological data recorded from the Fitbit Sense smartwatch and the audio recordings you would have recorded on the OnePlus Nord smartphone. The audio diaries as well as any physiological data captured by the Fitbit Sense can be accessed via the study-provided smartphone at any time. Guidance on how to view this data will be provided and demonstrated at the start of the study. Once produced, any journal articles outlining the study's outcome will be offered to participants who have opted into receiving study results on the consent form.



Study Funding

This study is funded by the Health Research Council (HRC) of New Zealand. The webpage containing details about this project is found here: <https://www.hrc.govt.nz/resources/research-repository/real-time-assessment-mood-changes-and-machine-learning>

Contact Details

If you have any questions, concerns, or complaints about the study at any stage, you can contact:

Dr Frederick Sundram, Primary Investigator

+64 9 923 7521

f.sundram@auckland.ac.nz

Dr Amy Chan, Co-investigator

+64 9 923 5524

a.chan@auckland.ac.nz

Dr Partha Roop, Co-investigator

+64 9 923 5583

p.roop@auckland.ac.nz

If you want to talk to someone who is not involved with the study, you can contact an independent health and disability advocate on:

Phone: 0800 555 050

Fax: 0800 2 SUPPORT (0800 2787 7678)

Email: advocacy@advocacy.org.nz

Website: <https://www.advocacy.org.nz/>

If you require Māori cultural support, talk to your whānau in the first instance. You may also contact the administrator for He Kamaka Waiora (Māori Health Team) by telephoning 09 486 8324 ext 2324, or contact the Auckland and Waitematā District Health Boards Māori Research Committee or Māori Research Advisor by phoning 09 4868920 ext 3204 to discuss any questions or complaints about the study.

For concerns of an ethical nature, you can contact the Chair of the Auckland Health Research Ethics Committee at:

Email: ahrec@auckland.ac.nz

Phone: 09 373 7599 x 83711

Address: Auckland Health Research Ethics Committee, The University of Auckland,
Private Bag 92019, Auckland 1142.

If you require free counselling services, please try some of the following helplines:



Need to talk? Free call or text 1737 any time.

Talk to a trained counsellor or call:

- the Depression helpline – 0800 111 757
- Alcohol drug helpline – 0800 787 797
- Gambling helpline – 0800 654 655
- Healthline – 0800 611 116 – to get help from a registered nurse 24/7.
- Lifeline – 0800 543 354
- Samaritans – 0800 726 666

Approved by Auckland Health Research Ethics Committee on **03/06/2021** for 3 years.
Reference Number **AH22426**.

Appendix C

Patient Health Questionnaire 9 (PHQ-9)

For copyright purposes the exact PHQ 9 Questionnaire used in the study has been removed.

It can be found online here:

<https://www.healthnavigator.org.nz/tools/p/patient-health-questionnaire-9-phq-9/>

Appendix D

Alcohol Use Disorders Identification Test (AUDIT)

For copyright purposes the exact AUDIT Questionnaire used in the study has been removed.

It can be found online here:

<https://auditscreen.org/check-your-drinking/>

Appendix E

Columbia-Suicide Severity Rating Scale (C-SSRS)

For copyright purposes the exact C-SSRS Questionnaire used in the study has been removed.

It can be found online here:

https://cssrs.columbia.edu/wp-content/uploads/C-SSRS_Pediatric-SLC_11.14.16.pdf

Appendix F

Mini-International Neuropsychiatric Interview (M.I.N.I)

For copyright purposes the exact M.I.N.I Questionnaire used in the study has been removed.

It can be found online here:

<https://harmresearch.org/mini-international-neuropsychiatric-interview-mini/>

Appendix G

Montgomery–Åsberg depression rating scale (MADRS)

For copyright purposes the exact MADRS Questionnaire used in the study has been removed.

It can be found online here:

<https://www.mdcalc.com/calc/4058/montgomery-asberg-depression-rating-scale-madrs>

Appendix H

Technology Acceptance Model Questionnaire (TAM-Q)

Technology Acceptance Questionnaire – MoodAI (Baseline)

Below are some statements that other people have made about wearables such as their FitBit and the Mood AI platform. Please tick below how much you agree with these statements. There are no right or wrong answers, we are simply interested in your experience and your views.

Statement	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Wearable devices are very useful to my life in general					
Wearable devices provide very useful services and information to me					
Overall, I find wearables easy to wear					
Using wearable devices is worthwhile					
I don't use a wearable device, because I am concerned about being observed					
I am afraid of my health data being tracked by wearable devices					
Using wearable devices fits into my lifestyle					
Using a wearable device makes me feel uncomfortable due to potential data security issues					
Overall, wearable devices look attractive					
I think the wearable device will be helpful to those with depression					
I think the wearable device is more relevant to physical health rather than mental health					
I am concerned about having to wear the wearable device most of the day					

Technology Acceptance Questionnaire – MoodAI (End of study)

Below are some statements that other people have made about their FitBit and the Mood AI platform. Please tick below how much you agree with these statements. There are no right or wrong answers, we are simply interested in your experience and your views.

Statement	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Wearable devices are very useful to my life in general					
Wearable devices provide very useful services and information to me					
Overall, I find wearables easy to wear					
Using wearable devices is worthwhile					
I don't use a wearable device, because I am concerned about being observed					
I am afraid of my health data being tracked by wearable devices					
Using wearable devices fits into my lifestyle					
Using a wearable device makes me feel uncomfortable due to potential data security issues					
Overall, wearable devices look attractive					
I think the wearable device will be helpful to those with depression					
I think the wearable device is more relevant to physical health rather than mental health					
I am concerned about having to wear the wearable device most of the day					

Wearable experience	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
I find it easy/convenient to wear the wearable device to sleep					
I worry about losing/misplacing the wearable device					
I recommend others to use wearable devices					

I am happy wearing my wearable devices around other people					
Using a wearable device is enjoyable to me					
I can do my daily tasks even when wearing my wearable					
System/platform experience					
The system was very useful for my mental health					
The system provided very useful services and information to me					
Overall, I found the system easy to use					
I found the system worthwhile					
I would recommend others to use this system					
Using the system was enjoyable to me					
Using this system makes me feel uncomfortable due to potential data security issues					
Using the system fits into my lifestyle					
I think this system will be helpful to those with depression					

Appendix I

Shah et al Model Comparisons

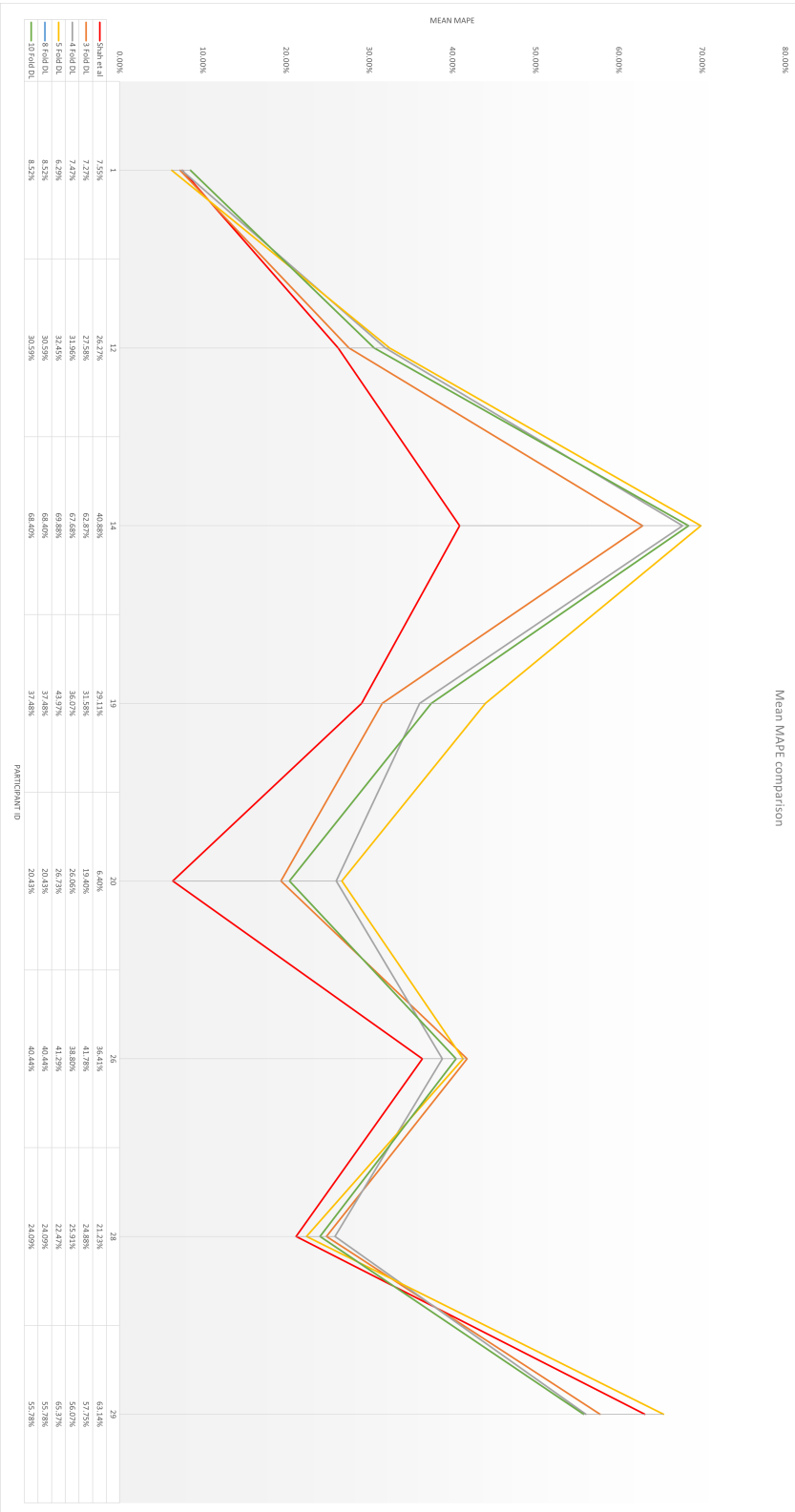


Figure I.1: Comparison of Mean MAPE over different folds using subset of Shah et al data

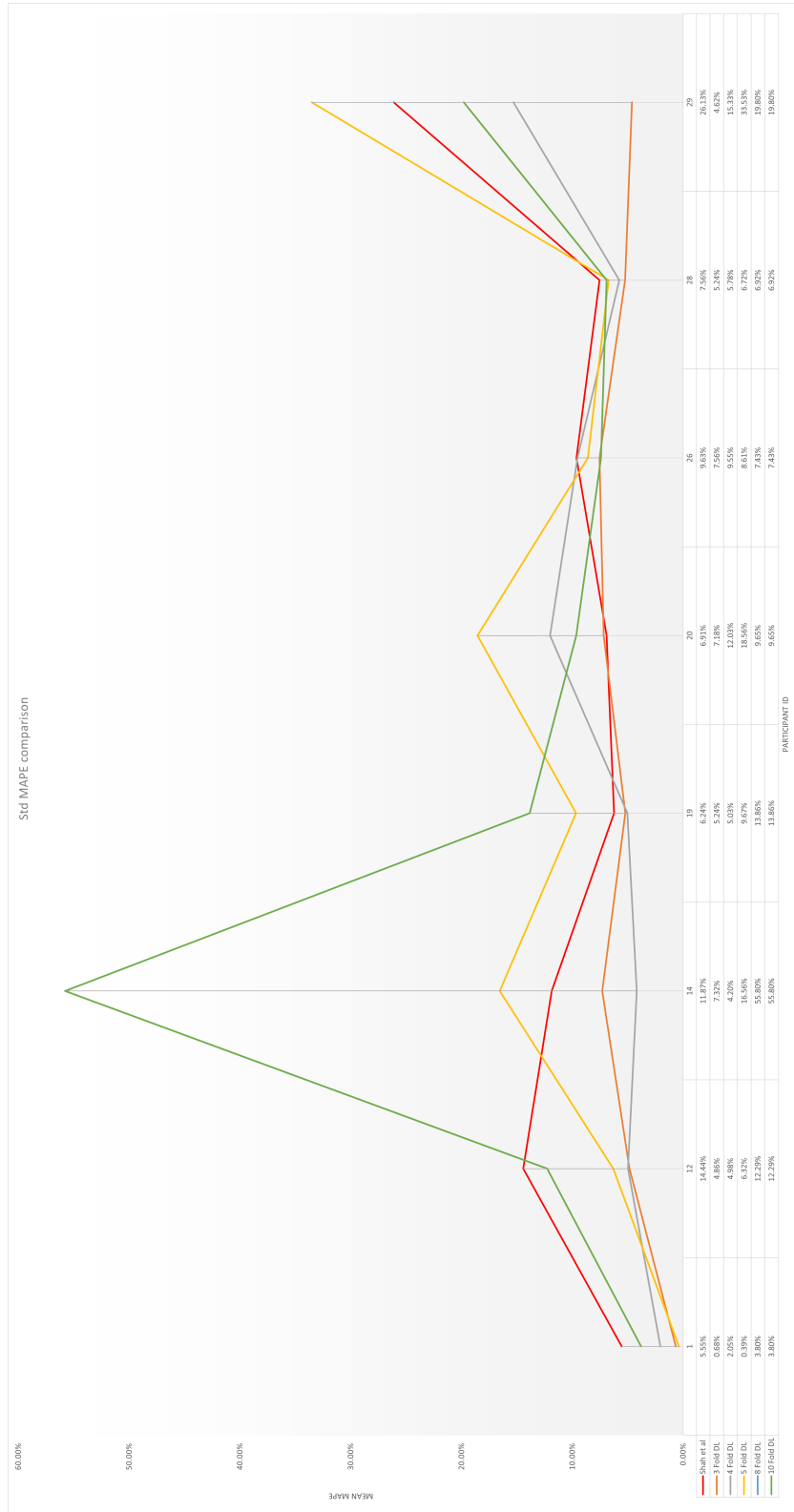


Figure I.2: Comparison of Std MAPE over different folds using subset of Shah et al data

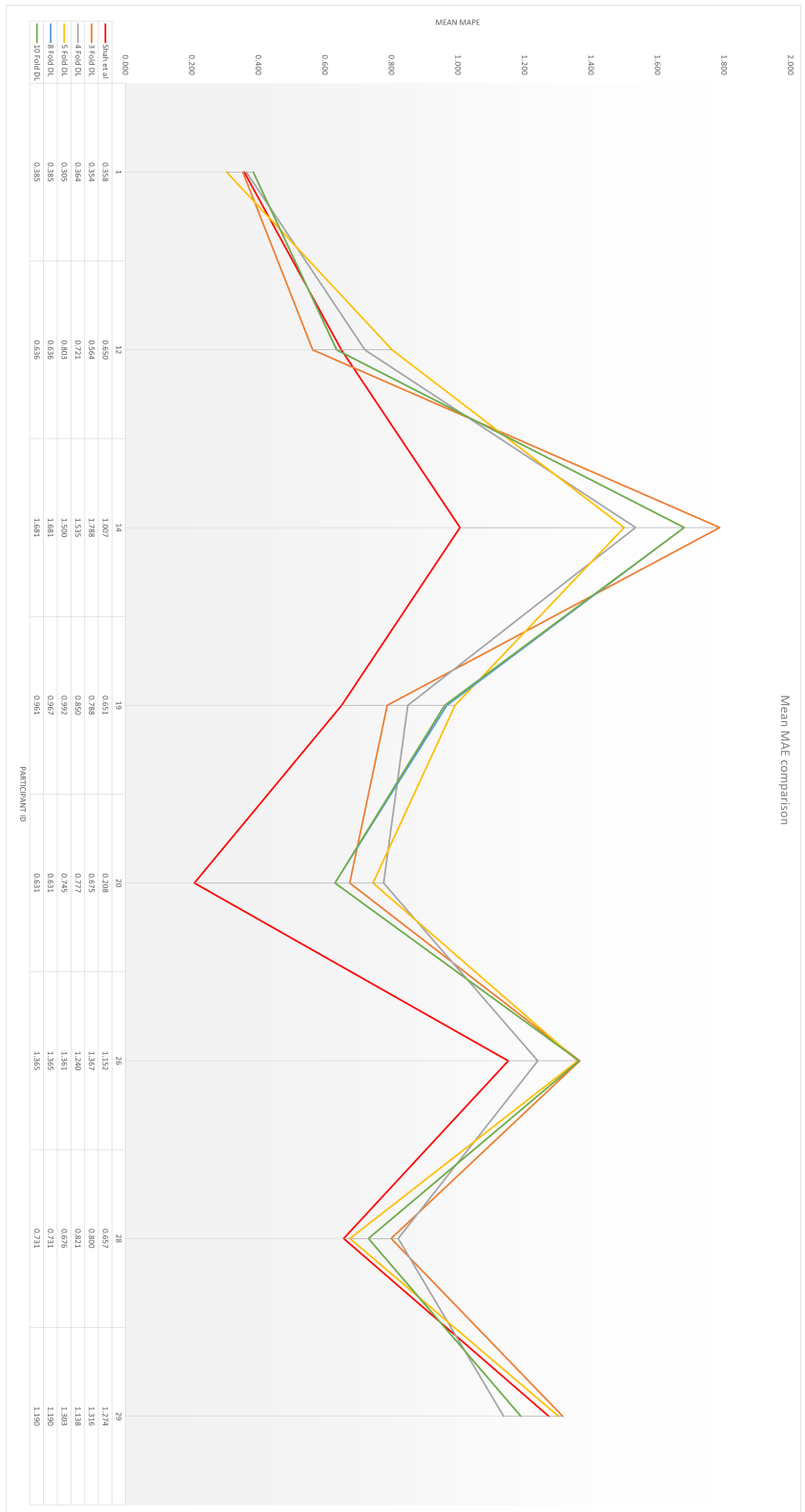


Figure 1.3: Comparison of Mean MAE over different folds using subset of Shah et al data

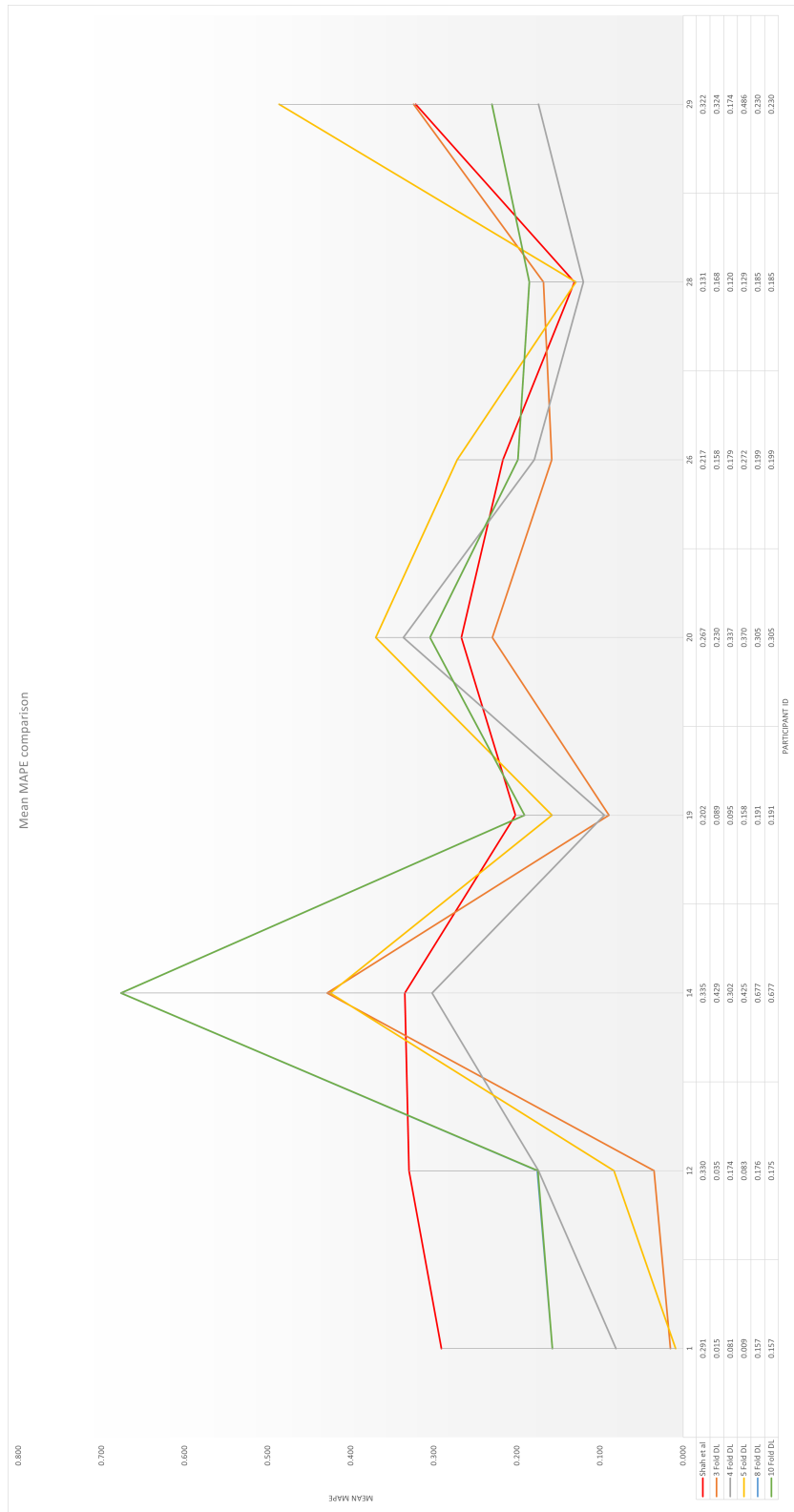


Figure I.4: Comparison of Std MAE over different folds using subset of Shah et al data

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	7.22%	0.67%	0.353	0.038
	Diff	0.33%	4.88%	0.005	0.253
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	29.12%	4.76%	0.637	0.086
	Diff	-2.85%	9.68%	0.013	0.244
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	56.84%	27.45%	1.467	0.665
	Diff	-15.96%	-15.58%	-0.460	-0.330
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	32.72%	5.55%	0.869	0.211
	Diff	-3.61%	0.69%	-0.218	-0.009
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	18.39%	5.79%	0.618	0.210
	Diff	-11.99%	1.12%	-0.410	0.057
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	45.20%	12.27%	1.428	0.248
	Diff	-8.79%	-2.64%	-0.276	-0.031
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	25.01%	9.21%	0.762	0.186
	Diff	-3.78%	-1.65%	-0.105	-0.055
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	64.37%	27.11%	1.308	0.429
	Diff	-1.23%	-0.98%	-0.034	-0.107

Table I.1: Comparison of best Shah et al regression based machine learning models using 5 fold cross validation MoodAI deep learning model using Shah et al data set

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	6.46%	3.46%	0.325	0.166
	Diff	1.09%	2.09%	0.033	0.112
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	34.91%	15.08%	0.794	0.292
	Diff	-8.64%	-0.64%	-0.144	0.038
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	51.10%	22.70%	1.300	0.631
	Diff	-10.22%	-10.83%	-0.293	-0.296
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	29.34%	9.29%	0.801	0.231
	Diff	-0.23%	-3.05%	-0.150	-0.029
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	23.58%	8.69%	0.804	0.257
	Diff	-17.18%	-1.78%	-0.596	0.010
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	39.70%	10.78%	1.365	0.243
	Diff	-3.29%	-1.15%	-0.204	-0.026
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	24.87%	8.46%	0.752	0.220
	Diff	-3.64%	-0.90%	-0.095	-0.089
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	80.27%	29.93%	1.692	0.641
	Diff	-7.14%	-3.80%	-0.418	-0.319

Table I.2: Comparison of best Shah et al regression based machine learning models using 8 fold cross validation MoodAI deep learning model using Shah et al data set

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	8.52%	4.97%	0.399	0.211
	Diff	-0.97%	0.58%	-0.041	0.080
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	30.93%	19.45%	0.648	0.331
	Diff	-4.66%	-5.01%	0.002	-0.001
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	65.21%	33.57%	1.467	0.734
	Diff	-24.33%	-21.70%	-0.460	-0.399
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	27.66%	9.01%	0.770	0.283
	Diff	1.45%	-2.77%	-0.119	-0.081
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	21.22%	11.21%	0.660	0.332
	Diff	-14.82%	-4.30%	-0.452	-0.065
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	4.05%	12.53%	1.295	0.214
	Diff	32.36%	-2.90%	-0.143	0.003
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	23.04%	7.17%	0.725	0.319
	Diff	-1.81%	0.39%	-0.068	-0.188
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	56.81%	44.83%	1.157	0.665
	Diff	6.33%	-18.70%	0.117	-0.343

Table I.3: Comparison of best Shah et al regression based machine learning models using 10 fold cross validation MoodAI deep learning model using Shah et al data set

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	7.27%	0.68%	0.354	0.015
	Diff	0.28%	4.87%	0.004	0.276
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	27.58%	4.86%	0.564	0.035
	Diff	-1.31%	9.58%	0.086	0.298
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	62.87%	7.32%	1.788	0.429
	Diff	-21.99%	4.55%	-0.781	-0.094
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	31.58%	5.24%	0.788	0.089
	Diff	-2.47%	1.00%	-0.137	0.113
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	19.40%	7.18%	0.675	0.230
	Diff	-13.00%	-0.27%	-0.467	0.037
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	41.75%	7.56%	1.367	0.158
	Diff	-5.37%	2.07%	-0.215	0.059
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	24.88%	5.24%	0.800	0.168
	Diff	-3.65%	2.32%	-0.143	-0.037
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	57.75%	4.62%	1.316	0.324
	Diff	5.39%	21.51%	-0.042	-0.002

Table I.4: Comparison of best Shah et al regression based machine learning models using 3 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	6.29%	0.39%	0.305	0.009
	Diff	1.26%	5.16%	0.053	0.282
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	32.45%	6.32%	0.803	0.083
	Diff	-6.18%	8.12%	-0.153	0.247
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	69.88%	16.56%	1.500	0.425
	Diff	-29.00%	-4.69%	-0.493	-0.090
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	43.94%	9.67%	0.992	0.158
	Diff	-14.86%	-3.43%	-0.341	0.044
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	26.73%	18.56%	0.745	0.370
	Diff	-20.33%	-11.65%	-0.537	-0.103
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	41.29%	8.61%	1.361	0.272
	Diff	-4.88%	1.02%	-0.209	-0.055
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	22.47%	6.72%	0.676	0.129
	Diff	-1.24%	0.84%	-0.019	0.002
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	65.37%	33.53%	1.303	0.486
	Diff	-2.23%	-7.40%	-0.029	-0.164

Table I.5: Comparison of best Shah et al regression based machine learning models using 5 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	8.52%	3.80%	0.385	0.157
	Diff	-0.97%	1.75%	-0.027	0.134
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	30.59%	12.29%	0.636	0.176
	Diff	-4.32%	2.15%	0.014	0.154
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	68.40%	55.80%	1.681	0.677
	Diff	-27.52%	-43.93%	-0.674	-0.342
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	37.48%	13.86%	0.967	0.191
	Diff	-8.37%	-7.62%	-0.316	0.011
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	20.43%	9.65%	0.631	0.305
	Diff	-14.03%	-2.74%	-0.423	-0.038
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	40.44%	7.43%	1.365	0.199
	Diff	-4.03%	2.20%	-0.213	0.018
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	24.09%	6.92%	0.731	0.185
	Diff	-2.86%	0.64%	-0.074	-0.054
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	55.78%	19.80%	1.190	0.230
	Diff	7.36%	6.33%	0.084	0.092

Table I.6: Comparison of best Shah et al regression based machine learning models using 8 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set

Participant ID	Model	Mean absolute % error		Mean absolute error	
		Mean	Std	Mean	Std
1	Shah et al	7.55%	5.55%	0.358	0.291
	Deep Learning	8.52%	3.80%	0.385	0.157
	Diff	-0.97%	1.75%	-0.027	0.134
12	Shah et al	26.27%	14.44%	0.650	0.330
	Deep Learning	30.59%	12.29%	0.636	0.175
	Diff	-4.32%	2.15%	0.014	0.155
14	Shah et al	40.88%	11.87%	1.007	0.335
	Deep Learning	68.40%	55.80%	1.681	0.677
	Diff	-27.52%	-43.93%	-0.674	-0.342
19	Shah et al	29.11%	6.24%	0.651	0.202
	Deep Learning	37.48%	13.86%	0.961	0.191
	Diff	-8.37%	-7.62%	-0.310	0.011
20	Shah et al	6.40%	6.91%	0.208	0.267
	Deep Learning	20.43%	9.65%	0.631	0.305
	Diff	-14.03%	-2.74%	-0.423	-0.038
26	Shah et al	36.41%	9.63%	1.152	0.217
	Deep Learning	40.44%	7.43%	1.365	0.199
	Diff	-4.03%	2.20%	-0.213	0.018
28	Shah et al	21.23%	7.56%	0.657	0.131
	Deep Learning	24.09%	6.92%	0.731	0.185
	Diff	-2.86%	0.64%	-0.074	-0.054
29	Shah et al	63.14%	26.13%	1.274	0.322
	Deep Learning	55.78%	19.80%	1.190	0.230
	Diff	7.36%	6.33%	0.084	-0.0092

Table I.7: Comparison of best Shah et al regression based machine learning models using 10 fold cross validation MoodAI deep learning model using a subset of the Shah et al data set

Bibliography

- [1] M. J. Friedrich, “Depression is the leading cause of disability around the world,” *Jama*, vol. 317, no. 15, pp. 1517–1517, 2017.
- [2] C. J. Murray, T. Vos, R. Lozano, *et al.*, “Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the global burden of disease study 2010,” *The lancet*, vol. 380, no. 9859, pp. 2197–2223, 2012.
- [3] J. E. Wells, M. A. O. Browne, K. M. Scott, *et al.*, “Prevalence, interference with life and severity of 12 month dsm-iv disorders in te rau hinengaro: The new zealand mental health survey,” *Australian and New Zealand Journal of Psychiatry*, vol. 40, no. 10, pp. 845–854, 2006.
- [4] A. J. Ferrari, F. J. Charlson, R. E. Norman, *et al.*, “Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study 2010,” *PLoS medicine*, vol. 10, no. 11, e1001547, 2013.
- [5] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, C. T. Pike, and R. C. Kessler, “The economic burden of adults with major depressive disorder in the united states (2005 and 2010),” *The Journal of clinical psychiatry*, vol. 76, no. 2, p. 5356, 2015.
- [6] B. Fatke, P. Hölzle, A. Frank, and H. Förstl, “Covid-19 crisis: Early observations on a pandemic’s psychiatric problems,” *Deutsche medizinische Wochenschrift (1946)*, vol. 145, no. 10, pp. 675–681, 2020.
- [7] C. Wang, R. Pan, X. Wan, *et al.*, “Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (covid-19) epidemic among the general population in china,” *International journal of environmental research and public health*, vol. 17, no. 5, p. 1729, 2020.

- [8] L. Silver, "Smartphone ownership is growing rapidly around the world, but not always equally," 2019.
- [9] N. Sultan, "Reflective thoughts on the potential and challenges of wearable technology for healthcare provision and medical education," *International Journal of Information Management*, vol. 35, no. 5, pp. 521–526, 2015.
- [10] E. Smets, E. Rios Velazquez, G. Schiavone, *et al.*, "Large-scale wearable data reveal digital phenotypes for daily-life stress detection," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–10, 2018.
- [11] M. W. Agelink, C. Boz, H. Ullrich, and J. Andrich, "Relationship between major depression and heart rate variability.: Clinical consequences and implications for antidepressive treatment," *Psychiatry research*, vol. 113, no. 1-2, pp. 139–149, 2002.
- [12] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: A review," *Medical and biological engineering and computing*, vol. 44, no. 12, pp. 1031–1051, 2006.
- [13] R. M. Carney, J. A. Blumenthal, P. K. Stein, *et al.*, "Depression, heart rate variability, and acute myocardial infarction," *Circulation*, vol. 104, no. 17, pp. 2024–2028, 2001.
- [14] A. L. Wheat and K. T. Larkin, "Biofeedback of heart rate variability and related physiology: A critical review," *Applied psychophysiology and biofeedback*, vol. 35, no. 3, pp. 229–242, 2010.
- [15] C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcome research: A scoping review," *Translational Psychiatry*, vol. 10, no. 1, pp. 1–26, 2020.
- [16] G. S. Malhi, A. Hamilton, G. Morris, Z. Mannie, P. Das, and T. Outhred, "The promise of digital mood tracking technologies: Are we heading on the right track?" *Evidence-based mental health*, vol. 20, no. 4, pp. 102–107, 2017.
- [17] A. Abbas, K. Schultebrucks, and I. R. Galatzer-Levy, "Digital measurement of mental health: Challenges, promises, and future directions," *Psychiatric Annals*, vol. 51, no. 1, pp. 14–20, 2021.

- [18] S. Byun, A. Y. Kim, E. H. Jang, *et al.*, “Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol,” *Computers in biology and medicine*, vol. 112, p. 103381, 2019.
- [19] F. Matcham, C. Barattieri di San Pietro, V. Bulgari, *et al.*, “Remote assessment of disease and relapse in major depressive disorder (radar-mdd): A multi-centre prospective cohort study protocol,” *BMC psychiatry*, vol. 19, no. 1, pp. 1–11, 2019.
- [20] N. C. Jacobson, H. Weingarden, and S. Wilhelm, “Digital biomarkers of mood disorders and symptom change,” *NPJ digital medicine*, vol. 2, no. 1, pp. 1–3, 2019.
- [21] K. Fox, J. S. Borer, A. J. Camm, *et al.*, “Resting heart rate in cardiovascular disease,” *Journal of the American College of Cardiology*, vol. 50, no. 9, pp. 823–830, 2007.
- [22] F. Shaffer, R. McCraty, and C. L. Zerr, “A healthy heart is not a metronome: An integrative review of the heart’s anatomy and heart rate variability,” *Frontiers in psychology*, vol. 5, p. 1040, 2014.
- [23] I. Liu, S. Ni, and K. Peng, “Happiness at your fingertips: Assessing mental health with smartphone photoplethysmogram-based heart rate variability analysis,” *Telemedicine and e-Health*, vol. 26, no. 12, pp. 1483–1491, 2020.
- [24] B. Farina, S. Dittoni, S. Colicchio, *et al.*, “Heart rate and heart rate variability modification in chronic insomnia patients,” *Behavioral sleep medicine*, vol. 12, no. 4, pp. 290–306, 2014.
- [25] J. M. Gorman and R. P. Sloan, “Heart rate variability in depressive and anxiety disorders,” *American heart journal*, vol. 140, no. 4, S77–S83, 2000.
- [26] R. Hartmann, F. M. Schmidt, C. Sander, and U. Hegerl, “Heart rate variability as indicator of clinical state in depression,” *Frontiers in psychiatry*, vol. 9, p. 735, 2019.
- [27] V. De Angel, S. Lewis, K. White, *et al.*, “Digital health tools for the passive monitoring of depression: A systematic review of methods,” *NPJ digital medicine*, vol. 5, no. 1, pp. 1–14, 2022.

- [28] J. K. Vallance, E. A. Winkler, P. A. Gardiner, G. N. Healy, B. M. Lynch, and N. Owen, "Associations of objectively-assessed physical activity and sedentary time with depression: Nhanes (2005–2006)," *Preventive medicine*, vol. 53, no. 4-5, pp. 284–288, 2011.
- [29] J. Kim, T. Nakamura, H. Kikuchi, K. Yoshiuchi, T. Sasaki, and Y. Yamamoto, "Covariation of depressive mood and spontaneous physical activity in major depressive disorder: Toward continuous monitoring of depressive mood," *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, pp. 1347–1355, 2015.
- [30] C. Brogly, J. K. Shoemaker, D. J. Lizotte, J. K. Kueper, M. Bauer, *et al.*, "A mobile app to identify lifestyle indicators related to undergraduate mental health (smart healthy campus): Observational app-based ecological momentary assessment," *JMIR formative research*, vol. 5, no. 10, e29160, 2021.
- [31] L. Zhai, Y. Zhang, and D. Zhang, "Sedentary behaviour and the risk of depression: A meta-analysis," *British journal of sports medicine*, vol. 49, no. 11, pp. 705–709, 2015.
- [32] D. Riemann, M. Berger, and U. Voderholzer, "Sleep and depression—results from psychobiological studies: An overview," *Biological psychology*, vol. 57, no. 1-3, pp. 67–103, 2001.
- [33] M. BERGER and D. RIEMANN, "Rem sleep in depression—an overview," *Journal of sleep research*, vol. 2, no. 4, pp. 211–223, 1993.
- [34] Y. Rykov, T.-Q. Thach, I. Bojic, G. Christopoulos, J. Car, *et al.*, "Digital biomarkers for depression screening with wearable devices: Cross-sectional study with machine learning modeling," *JMIR mHealth and uHealth*, vol. 9, no. 10, e24872, 2021.
- [35] T. O. Ayodele, "Types of machine learning algorithms," *New advances in machine learning*, vol. 3, pp. 19–48, 2010.
- [36] B. Yildiz, J. I. Bilbao, and A. B. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 1104–1122, 2017.
- [37] Z.-H. Zhou, "Ensemble learning," in *Machine learning*, Springer, 2021, pp. 181–210.

- [38] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [40] C. M. Bishop, “Neural networks and their applications,” *Review of scientific instruments*, vol. 65, no. 6, pp. 1803–1832, 1994.
- [41] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, *et al.*, *Gradient flow in recurrent nets: The difficulty of learning long-term dependencies*, 2001.
- [42] J. F. Kolen and S. C. Kremer, *A field guide to dynamical recurrent networks*. John Wiley & Sons, 2001.
- [43] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [44] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with lstm,” *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
- [45] X. Yang, P. Roop, H. Pearce, and J. W. Ro, “A compositional approach using keras for neural networks in real-time systems,” in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2020, pp. 1109–1114.
- [46] P. Refaailzadeh, L. Tang, and H. Liu, “Cross-validation,” in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 532–538, ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_565. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_565.
- [47] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyperparameter optimization,” *Advances in neural information processing systems*, vol. 24, 2011.
- [48] R. V. Shah, G. Grennan, M. Zafar-Khan, *et al.*, “Personalized machine learning of depressed mood using wearables,” *Translational Psychiatry*, vol. 11, no. 1, pp. 1–18, 2021.
- [49] N. C. Jacobson, H. Weingarden, and S. Wilhelm, “Using digital phenotyping to accurately detect depression severity,” *The Journal of nervous and mental disease*, vol. 207, no. 10, pp. 893–896, 2019.

- [50] N. C. Jacobson and M. D. Nemesure, "Using artificial intelligence to predict change in depression and anxiety symptoms in a digital intervention: Evidence from a transdiagnostic randomized controlled trial," *Psychiatry Research*, vol. 295, p. 113 618, 2021.
- [51] A. Ghandeharioun, S. Fedor, L. Sangermano, *et al.*, "Objective assessment of depressive symptoms with machine learning and wearable sensors data," in *2017 seventh international conference on affective computing and intelligent interaction (ACII)*, IEEE, 2017, pp. 325–332.
- [52] M. Hamilton, "The hamilton rating scale for depression," in *Assessment of depression*, Springer, 1986, pp. 143–152.
- [53] Y. Tazawa, K.-c. Liang, M. Yoshimura, *et al.*, "Evaluating depression with multimodal wristband-type wearable device: Screening and assessing patient severity utilizing machine-learning," *Heliyon*, vol. 6, no. 2, e03274, 2020.
- [54] A. Sano, S. Taylor, A. W. McHill, *et al.*, "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*, vol. 20, no. 6, e9410, 2018.
- [55] A. Misra, A. Ojeda, and J. Mishra, "Braine: A digital platform for evaluating, engaging and enhancing brain function," *Regents of the University of California Copyright SD2018-816*, 2018.
- [56] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning—a brief history, state-of-the-art and challenges," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2020, pp. 417–431.
- [57] R. Bai, L. Xiao, Y. Guo, *et al.*, "Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: Prospective naturalistic multicenter study," *JMIR mHealth and uHealth*, vol. 9, no. 3, e24365, 2021.
- [58] T. A. Widiger, A. J. Frances, H. A. E. Pincus, R. E. Ross, *et al.*, *DSM-IV sourcebook, Vol. 3*. American Psychiatric Publishing, Inc., 1997.
- [59] C. Park, C. C. Took, and J.-K. Seong, "Machine learning in biomedical engineering," *Biomedical Engineering Letters*, vol. 8, no. 1, pp. 1–3, 2018.

- [60] L. Yang, D. Jiang, W. Han, and H. Sahli, "Dcnn and dnn based multi-modal depression recognition," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2017, pp. 484–489.
- [61] R. Gupta, S. Sahu, C. Y. Espy-Wilson, and S. S. Narayanan, "An affect prediction approach through depression severity parameter incorporation in neural networks.," in *INTERSPEECH*, 2017, pp. 3122–3126.
- [62] S. Harati, A. Crowell, H. Mayberg, and S. Nemati, "Depression severity classification from speech emotion," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018, pp. 5763–5766.
- [63] M. A. Tully, C. McBride, L. Heron, and R. F. Hunter, "The validation of fitbit zip™ physical activity monitor as a measure of free-living physical activity," *BMC research notes*, vol. 7, no. 1, pp. 1–5, 2014.
- [64] M. T. Imboden, M. B. Nelson, L. A. Kaminsky, and A. H. Montoye, "Comparison of four fitbit and jawbone activity monitors with a research-grade actigraph accelerometer for estimating physical activity and energy expenditure," *British Journal of Sports Medicine*, vol. 52, no. 13, pp. 844–850, 2018.
- [65] L. M. Feehan, J. Geldman, E. C. Sayre, *et al.*, "Accuracy of fitbit devices: Systematic review and narrative syntheses of quantitative data," *JMIR mHealth and uHealth*, vol. 6, no. 8, e10527, 2018.
- [66] N. Redenius, Y. Kim, and W. Byun, "Concurrent validity of the fitbit for assessing sedentary behavior and moderate-to-vigorous physical activity," *BMC medical research methodology*, vol. 19, no. 1, pp. 1–9, 2019.
- [67] A. K. Battenberg, S. Donohoe, N. Robertson, and T. P. Schmalzried, "The accuracy of personal activity monitoring devices," in *Seminars in Arthroplasty*, Elsevier, vol. 28, 2017, pp. 71–75.
- [68] B. W. Nelson and N. B. Allen, "Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: Intraindividual validation study," *JMIR mHealth and uHealth*, vol. 7, no. 3, e10828, 2019.
- [69] J. Lu, C. Shang, C. Yue, *et al.*, "Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–21, 2018.

- [70] A. Shcherbina, C. M. Mattsson, D. Waggott, *et al.*, “Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort,” *Journal of personalized medicine*, vol. 7, no. 2, p. 3, 2017.
- [71] S. Haghayegh, S. Khoshnevis, M. H. Smolensky, K. R. Diller, R. J. Castriotta, *et al.*, “Accuracy of wristband fitbit models in assessing sleep: Systematic review and meta-analysis,” *Journal of medical Internet research*, vol. 21, no. 11, e16273, 2019.
- [72] A. Natarajan, H.-W. Su, C. Heneghan, L. Blunt, C. O’Connor, and L. Niehaus, “Measurement of respiratory rate using wearable devices and applications to covid-19 detection,” *NPJ digital medicine*, vol. 4, no. 1, pp. 1–10, 2021.
- [73] ANZCTR, *Australian new zealand clinical trials registry (anzctr)*. [Online]. Available: <https://www.anzctr.org.au/Faq.aspx>.
- [74] H. R. C. N. Zealand, *Real-time assessment of mood changes and machine learning*. [Online]. Available: <https://www.hrc.govt.nz/resources/research-repository/real-time-assessment-mood-changes-and-machine-learning>.
- [75] T. R. Kirchner and S. Shiffman, “Ecological momentary assessment. in the wileyblackwell handbook of addiction psychopharmacology,” pp. 541–565, 2013.
- [76] O. Flanagan, A. Chan, P. Roop, F. Sundram, *et al.*, “Using acoustic speech patterns from smartphones to investigate mood disorders: Scoping review,” *JMIR mHealth and uHealth*, vol. 9, no. 9, e24352, 2021.
- [77] A. C. Lahti, D. Wang, H. Pei, S. Baker, and V. A. Narayan, “Clinical utility of wearable sensors and patient-reported surveys in patients with schizophrenia: Noninterventional, observational study,” *JMIR mental health*, vol. 8, no. 8, e26234, 2021.
- [78] K. Kroenke, R. L. Spitzer, and J. B. Williams, “The phq-9: Validity of a brief depression severity measure,” *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.
- [79] K. Kroenke and R. L. Spitzer, *The phq-9: A new depression diagnostic and severity measure*, 2002.

- [80] J. Asselbergs, J. Ruwaard, M. Ejdys, N. Schrader, M. Sijbrandij, H. Riper, *et al.*, “Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: An explorative study,” *Journal of medical Internet research*, vol. 18, no. 3, e5505, 2016.
- [81] T. Babor, J. Higgins-Biddle, J. Saunders, and M. Monteiro, “The alcohol use disorders identification test world health organization,” *Geneva, Switzerland.[Google Scholar]*, 2001.
- [82] J. Ryan and L. Howes, “Relations between alcohol consumption, heart rate, and heart rate variability in men,” *Heart*, vol. 88, no. 6, pp. 641–642, 2002.
- [83] D. B. Newlin, E. A. Byrne, and S. W. Porges, “Vagal mediation of the effect of alcohol on heart rate,” *Alcoholism: Clinical and Experimental Research*, vol. 14, no. 3, pp. 421–424, 1990.
- [84] P. J. Conrod, J. B. Peterson, and R. O. Pihl, “Reliability and validity of alcohol-induced heart rate increase as a measure of sensitivity to the stimulant properties of alcohol,” *Psychopharmacology*, vol. 157, no. 1, pp. 20–30, 2001.
- [85] D. F. Reinert and J. P. Allen, “The alcohol use disorders identification test: An update of research findings,” *Alcoholism: Clinical and Experimental Research*, vol. 31, no. 2, pp. 185–199, 2007.
- [86] C. Young and T. Mayson, “The alcohol use disorders identification scale (audit) normative scores for a multiracial sample of rhodes university residence students,” *Journal of Child and Adolescent Mental Health*, vol. 22, no. 1, pp. 15–23, 2010.
- [87] K. Posner, D. Brent, C. Lucas, *et al.*, “Columbia-suicide severity rating scale (c-ssrs),” *New York, NY: Columbia University Medical Center*, vol. 10, 2008.
- [88] Y. Lecrubier, D. V. Sheehan, E. Weiller, *et al.*, “The mini international neuropsychiatric interview (mini). a short diagnostic structured interview: Reliability and validity according to the cidi,” *European psychiatry*, vol. 12, no. 5, pp. 224–231, 1997.
- [89] S. A. Montgomery and M. Åsberg, “A new depression scale designed to be sensitive to change,” *The British journal of psychiatry*, vol. 134, no. 4, pp. 382–389, 1979.

- [90] J. B. Williams and K. A. Kobak, "Development and reliability of a structured interview guide for the montgomery-Åsberg depression rating scale (sigma)," *The British Journal of Psychiatry*, vol. 192, no. 1, pp. 52–58, 2008.
- [91] H. M. Abu-Dalbouh, "A questionnaire approach based on the technology acceptance model for mobile tracking on patient progress applications," *J. Comput. Sci.*, vol. 9, no. 6, pp. 763–770, 2013.
- [92] J. Vega, B. Bell, C. Taylor, *et al.*, "Detecting mental health behaviours using mobile interactions (demmi): An exploratory study focusing on binge eating," *JMIR Mental Health*, 2022.
- [93] S. Nickels, M. D. Edwards, S. F. Poole, *et al.*, "Toward a mobile platform for real-world digital measurement of depression: User-centered design, data quality, and behavioral and clinical modeling," *JMIR mental health*, vol. 8, no. 8, e27589, 2021.
- [94] W. Morrison, L. Guerdan, J. Kanugo, T. Trull, and Y. Shang, "Tigeraware: An innovative mobile survey and sensor data collection and analytics system," in *2018 IEEE third international conference on Data Science in Cyberspace (DSC)*, IEEE, 2018, pp. 115–122.
- [95] Firebase. "Firebase authentication." (), [Online]. Available: <https://firebase.google.com/docs/auth>.
- [96] L. M. Khue, E. L. Ouh, and S. Jarzabek, "Mood self-assessment on smartphones," in *Proceedings of the conference on Wireless Health*, 2015, pp. 1–8.
- [97] K. P. Dao, K. De Cocker, H. L. Tong, A. B. Kocaballi, C. Chow, and L. Laranjo, "Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: Systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies," *JMIR mHealth and uHealth*, vol. 9, no. 11, e22890, 2021.
- [98] C. Rauschenberg, B. Boecking, I. Paetzold, *et al.*, "A compassion-focused ecological momentary intervention for enhancing resilience in help-seeking youth: Uncontrolled pilot study," *JMIR mental health*, vol. 8, no. 8, e25650, 2021.

- [99] H. Þórarinsdóttir, L. V. Kessing, M. Faurholt-Jepsen, *et al.*, “Smartphone-based self-assessment of stress in healthy adult individuals: A systematic review,” *Journal of medical Internet research*, vol. 19, no. 2, e6397, 2017.
- [100] S. Schwartz, S. Schultz, A. Reider, and E. F. Saunders, “Daily mood monitoring of symptoms using smartphones in bipolar disorder: A pilot study assessing the feasibility of ecological momentary assessment,” *Journal of Affective Disorders*, vol. 191, pp. 88–93, 2016.
- [101] M. F. Armeij, H. T. Schatten, N. Haradhvala, and I. W. Miller, “Ecological momentary assessment (ema) of depression-related phenomena,” *Current opinion in psychology*, vol. 4, pp. 21–25, 2015.
- [102] S. J. Yim, L. M. Lui, Y. Lee, *et al.*, “The utility of smartphone-based, ecological momentary assessment for depressive symptoms,” *Journal of Affective Disorders*, vol. 274, pp. 602–609, 2020.
- [103] Y. S. Yang, G. W. Ryu, I. Han, S. Oh, and M. Choi, “Ecological momentary assessment using smartphone-based mobile application for affect and stress assessment,” *Healthcare informatics research*, vol. 24, no. 4, pp. 381–386, 2018.
- [104] D. Colombo, J. Fernández-Álvarez, A. Patané, *et al.*, “Current state and future directions of technology-based ecological momentary assessment and intervention for major depressive disorder: A systematic review,” *Journal of clinical medicine*, vol. 8, no. 4, p. 465, 2019.
- [105] H. Kim, S. Lee, S. Lee, S. Hong, H. Kang, N. Kim, *et al.*, “Depression prediction by using ecological momentary assessment, actiwatch data, and machine learning: Observational study on older adults living alone,” *JMIR mHealth and uHealth*, vol. 7, no. 10, e14149, 2019.
- [106] J.-L. Elghozi, A. Girard, and D. Laude, “Effects of drugs on the autonomic control of short-term heart rate variability,” *Autonomic Neuroscience*, vol. 90, no. 1-2, pp. 116–121, 2001.
- [107] C. M. van Ravenswaaij-Arts, L. A. Kollee, J. C. Hopman, G. B. Stoeltinga, and H. P. van Geijn, “Heart rate variability,” *Annals of internal medicine*, vol. 118, no. 6, pp. 436–447, 1993.
- [108] H. A. Young and D. Benton, “Heart-rate variability: A biomarker to study the influence of nutrition on physiological and psychological health?” *Behavioural pharmacology*, vol. 29, no. 2-, p. 140, 2018.

- [109] L. Bernardi, F. Valle, M. Coco, A. Calciati, and P. Sleight, "Physical activity influences heart rate variability and very-low-frequency components in holter electrocardiograms," *Cardiovascular research*, vol. 32, no. 2, pp. 234–237, 1996.
- [110] C. Baglioni, S. Nanovska, W. Regen, *et al.*, "Sleep and mental disorders: A meta-analysis of polysomnographic research," *Psychological bulletin*, vol. 142, no. 9, p. 969, 2016.
- [111] F. Or, J. Torous, and J.-P. Onnela, "High potential but limited evidence: Using voice data from smartphones to monitor and diagnose mood disorders," *Psychiatric rehabilitation journal*, vol. 40, no. 3, p. 320, 2017.
- [112] A. Guidi, S. Salvi, M. Ottaviano, *et al.*, "Smartphone application for the analysis of prosodic features in running speech with a focus on bipolar disorders: System performance evaluation and case study," *Sensors*, vol. 15, no. 11, pp. 28 070–28 087, 2015.
- [113] R. F. Dickerson, E. I. Gorlin, and J. A. Stankovic, "Empath: A continuous remote emotional health monitoring system for depressive illness," in *Proceedings of the 2nd Conference on Wireless Health*, 2011, pp. 1–10.
- [114] S. Place, D. Blanch-Hartigan, C. Rubin, *et al.*, "Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders," *Journal of medical Internet research*, vol. 19, no. 3, e6678, 2017.
- [115] T. pandas development team, *Pandas-dev/pandas: Pandas*, version latest, Feb. 2020. DOI: 10.5281/zenodo.3509134. [Online]. Available: <https://doi.org/10.5281/zenodo.3509134>.
- [116] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.
- [117] M. K. Olsen, K. M. Stechuchak, J. D. Edinger, C. S. Ulmer, and R. F. Woolson, "Move over locf: Principled methods for handling missing data in sleep disorder trials," *Sleep medicine*, vol. 13, no. 2, pp. 123–132, 2012.
- [118] H. J. Baek and J. Shin, "Effect of missing inter-beat interval data on heart rate variability analysis using wrist-worn wearables," *Journal of Medical Systems*, vol. 41, no. 10, pp. 1–9, 2017.
- [119] F. Chollet *et al.*, *Keras*, 2015.

- [120] M. Abadi, A. Agarwal, P. Barham, *et al.*, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint arXiv:1603.04467*, 2016.
- [121] D. Berrar, *Cross-validation*. 2019.
- [122] R. Kohavi *et al.*, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Ijcai*, Montreal, Canada, vol. 14, 1995, pp. 1137–1145.
- [123] M. Feurer and F. Hutter, “Hyperparameter optimization,” in *Automated machine learning*, Springer, Cham, 2019, pp. 3–33.
- [124] Firebase. “Firebase admin python sdk.” (), [Online]. Available: <https://firebase.google.com/docs/reference/admin/python>.
- [125] sphinx-quickstart. “Python-fitbit.” (), [Online]. Available: <https://python-fitbit.readthedocs.io/en/latest/>.
- [126] Fitbit. “Fitbit web api.” (), [Online]. Available: <https://dev.fitbit.com/build/reference/web-api/>.
- [127] R. Conlin, K. Erickson, J. Abbate, and E. Kolemen, “Keras2c: A library for converting keras neural networks to real-time compatible c,” *Engineering Applications of Artificial Intelligence*, vol. 100, pp. 104–182, 2021.