MACHINE LEARNING FOR WELL-BEING ASSESSMENT: EXPLORING DIGITAL BIOMARKERS

by

Berrenur Saylam

B.S., Computer Engineering, Galatasaray University, 2014

B.S., Industrial Engineering, Galatasaray University, 2015

M.S., Complex Systems, École Normale Supérieure de Lyon, 2016

M.Eng., Computer Engineering, Galatasaray University, 2017

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2024

To my parents

for being the roots of my strength and the wings of my dreams!

ACKNOWLEDGEMENTS

"*Do not lose hope, nor be sad. You will surely be victorious if you are true believers!*"[1]

I am profoundly grateful to my mother, whose expertise in both Western and Eastern languages has been a remarkable asset throughout my academic journey. Her ability to engage in and attend a conference conducted in French at Sorbonne University alongside me was not only a privilege but a testament to her unwavering support and intellectual curiosity.

I also wish to express my deep appreciation to my father, whose early encouragement of critical thinking and philosophical inquiry has significantly shaped my intellectual development. Our thought-provoking discussions have been instrumental in refining my analytical skills and perspective.

I extend my heartfelt thanks to all the remarkable individuals from around the world who have embraced a human-first approach. Whether through personal interactions or virtual engagements, your honesty, compassion, and support have had a profound impact on my life experience. Each act of kindness and every genuine connection have enriched my journey and reinforced our shared humanity. I look forward to continuing to encounter such inspiring individuals throughout my future endeavors.

I am hopeful for a world where empathy and kindness are increasingly prevalent, as these qualities are essential to our collective well-being. Despite any challenges, our shared humanity binds us, and I believe that spreading these values can make a meaningful difference.

I would also like to extend my deepest gratitude to my advisors, Dr. Özlem Durmaz İncel and Dr. Arzucan Özgür Türkmen. Dr. İncel's guidance has been instru-

[1] Qur'an, Âl-i İmrân 3:139

mental in shaping the direction and quality of my research, and I deeply appreciate her mentorship. I am also particularly grateful to Dr. Türkmen for stepping up and taking the lead during the periods when Dr. İncel was on leave. I am also thankful to the jury members for their valuable contributions to my thesis. Their feedback has greatly enhanced the quality of this work, and I sincerely appreciate the time and effort they have invested in being part of the committee.

Lastly, I would like to express my sincere gratitude to the Turkish Directorate of Strategy and Budget for their support under the TAM Project number 2007K12-873, to YÖK-YUDAB Overseas Research Scholarships for Research Assistants and to TÜBİTAK for BİDEB 2211-A academic reward.

İstanbul, 2024

ABSTRACT

MACHINE LEARNING FOR WELL-BEING ASSESSMENT: EXPLORING DIGITAL BIOMARKERS

This thesis investigates the potential of wearable sensors and self-reported questionnaires by exploring different factors of human well-being using digital biomarkers. The primary goal is to identify and validate these biomarkers, comparing them against traditional psychological studies. Various machine learning models are trained and validated for analyzing different well-being factors, including sleep, mental well-being (stress, anxiety, depression, positive and negative affect), and academic achievement. Two extensive datasets are utilized in these explorations: NetHealth, which is collected from more than 700 college students over 4 years, and Tesserae collected over a year of 757 office workers.

Advanced techniques, incorporating the analysis of time-lagged data to capture temporal patterns and multi-task learning, are employed to unravel complex relationships among well-being parameters. The research unfolds systematically, progressing from single well-being factor exploration to time-based and multi-task methodologies.

This thesis emphasizes the importance of incorporating temporal dimensions and multi-task learning strategies for a more comprehensive understanding of well-being and the factors influencing it. Our findings offer valuable insights into the identification of reliable biomarkers and the relationships between various well-being aspects within two different target groups of university students and office workers. It aims to provide a new perspective, moving beyond single-factor exploration to enhance the comprehensiveness and applicability of well-being studies.

ÖZET

İYİLİK HALİ DEĞERLENDİRMESİ İÇİN MAKİNE ÖĞRENİMİ: DİJİTAL BİYOBELİRTEÇLERİN KEŞFİ

Bu tez, dijital biyobelirteçleri kullanarak insan refahının farklı faktörlerini keşfetmek için giyilebilir sensörlerin ve kişisel anketlerin potansiyelini araştırmaktadır. Temel amaç, bu biyobelirteçleri tanımlamak ve doğrulamak, bunları geleneksel psikolojik çalışmalarla karşılaştırmaktır. Makine öğrenimi modelleri; uyku, zihinsel sağlık (stres, kaygı, depresyon, olumlu ve olumsuz duygulanım ve akademik başarı dahil) farklı sağlık faktörlerini analiz etmek için eğitilmiş ve doğrulanmıştır. Bu araştırmalarda iki kapsamlı veri kümesi kullanılmıştır: NetHealth; 4 yıl boyunca 700'den fazla üniversite öğrencisinden ve Tesserae bir yıl boyunca 757 ofis çalışanından toplanmıştır.

Çoklu görev öğrenmeyi içeren daha ileri makine öğrenmesi yöntemleri ile, sağlıklı yaşam parametreleri arasındaki karmaşık ilişkileri çözmeye katkı sağlar. Araştırma, sistemli bir şekilde ilerleyerek, başlangıçta tek bir faktörün keşfinden zaman temelli ve çoklu görev yöntemlerine doğru evrilmektedir.

Tez kapsamında, çoklu modal kişisel veri analizi için gelişmiş yöntemler ile sağlıklı yaşamın anlaşılmasına katkıda bulunmaktadır. Dijital biyobelirteçlerin çeşitli veri kümelerinden çıkarılması ve doğrulanmasında, zaman temelli ve çoklu modal yöntemleri tezin temelini oluşturmaktadır. Bu çalışma, tek bir faktör keşfinin ötesine geçerek sağlıklı yaşam çalışmalarının kapsamlılığını ve uygulanabilirliğini artırmayı hedefleyen yeni bir bakış açısı sunmaktadır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv
ABSTRACT vi
ÖZET vii
LIST OF FIGURES xii
LIST OF TABLES xvi
LIST OF ACRONYMS/ABBREVIATIONS xviii
1. INTRODUCTION 1
1.1. Motivation and Objectives 2
1.1.1. Research Gaps in The Literature 3
1.1.2. Research Questions 5
1.1.2.1. Group 1: Digital Biomarkers Targeting Well-Being . . 5
1.1.2.2. Group 2: Validation and Prediction 6
1.2. Contributions and Findings 6
1.2.1. Novelty 7
1.2.2. Findings 7
1.2.2.1. Academic Performance Classification 8
1.2.2.2. Sleep Efficiency Prediction 8
1.2.2.3. Stress State Prediction 8
1.2.2.4. Temporal Learning Approaches 9
1.2.2.5. Multi-Task Learning Approaches 9
1.2.3. Academic Outputs of the Thesis 10
1.3. Thesis Outline 10
2. BACKGROUND 13
2.1. Sensors Integrated into Wearables 13
2.2. Digital Biomarkers 14
2.3. From Sensors to High Level Recognition 15
2.3.1. Sleep 17
2.3.2. Mood 17

2.3.3. Stress . . . 18
2.4. Ground Truth Methods for the Analysis of Psychological Targets . . . 22
2.5. Datasets Utilized in the Thesis . . . 23
2.5.1. Tesserae Dataset . . . 24
2.5.1.1. General Overview . . . 24
2.5.1.2. Thesis Related Sub-Dataset . . . 24
2.5.2. NetHealth Dataset . . . 29
2.5.2.1. General Overview . . . 29
2.5.2.2. Thesis Related Sub-Dataset . . . 31
3. LITERATURE REVIEW ON WELL-BEING PREDICTION . . . 35
3.1. Workplace Well-Being and Job Performance . . . 37
3.2. Academic Grade . . . 42
3.3. General Well-Being . . . 43
3.3.1. Sleep . . . 43
3.3.2. Physical Activity . . . 47
3.3.3. Stress . . . 48
3.3.4. Mental Health . . . 50
3.4. Other Studies Utilizing the Datasets . . . 53
3.5. Discussion . . . 53
4. ACADEMIC GRADES AND SLEEP QUALITY AMONG STUDENTS . . . 58
4.1. Academic Grade Prediction . . . 59
4.1.1. Motivation & Related Work . . . 60
4.1.2. Data Preprocessing . . . 61
4.1.2.1. Handling Missing Values . . . 61
4.1.2.2. Correlation . . . 62
4.1.2.3. Distribution of Target Values: Grades . . . 62
4.1.3. Model Details and Performance Metrics . . . 62
4.1.4. Classification Performance Evaluation . . . 63
4.1.5. Discussion . . . 64
4.2. Sleep Quality Prediction . . . 66
4.2.1. Motivation & Related Work . . . 66

4.2.2. Data Preprocessing . . . 69
4.2.3. Model Details and Performance Metrics . . . 69
4.2.4. Important Factors . . . 70
4.2.5. Discussion and Conclusion . . . 71
5. STRESS LEVELS AMONG OFFICE WORKERS . . . 74
5.1. Motivation . . . 74
5.2. Related Studies . . . 75
5.3. Data Preprocessing . . . 77
5.3.1. Details of the Stress Classes . . . 77
5.3.2. Data Preprocessing . . . 77
5.4. Classifier and Validation . . . 79
5.5. Results . . . 79
5.5.1. Results on Without Personality Dataset . . . 79
5.5.1.1. With Class Imbalance . . . 79
5.5.1.2. Performance after Applying SMOTE . . . 82
5.5.2. With Personality Dataset Results . . . 85
5.5.2.1. With Class Imbalance . . . 86
5.5.2.2. Performance after Applying SMOTE . . . 89
5.5.3. Discussion and Comparison . . . 93
5.6. Conclusion . . . 96
6. TIME-BASED WELL-BEING PREDICTION FOR WORKERS . . . 97
6.1. Background and Motivation for a Time-Based Approach . . . 99
6.2. Related Studies . . . 101
6.3. Data Preprocessing . . . 105
6.3.1. Creation of Lagged (Time-Based) Dataset . . . 108
6.4. Prediction Algorithms . . . 109
6.4.1. Random Forest . . . 109
6.4.2. XGBoost . . . 109
6.4.3. Long Short-Term Memory . . . 110
6.5. Performance of Well-Being Factors Prediction . . . 110
6.5.1. Implementation Details . . . 110

6.5.2. Digital Biomarkers of Well-Being Factors . . . 111
6.5.3. Conventional vs. Time-Based Prediction Performances . . . 116
6.6. Conclusions . . . 121
7. MULTI-TASK LEARNING FOR MENTAL HEALTH OF STUDENTS . . . 124
7.1. Studies on Time-Based and Multi-Task Learning for Mental Health . . 127
7.2. Methodology . . . 133
7.2.1. Time-Based Modeling . . . 133
7.2.2. Multi-Task Learning . . . 133
7.2.3. Data Preprocessing . . . 134
7.2.3.1. Target Values . . . 134
7.2.3.2. Missing Data . . . 135
7.2.3.3. Data Preprocessing for Time-Based Methods . . . 136
7.2.4. Models and Evaluation Metrics . . . 137
7.3. Results . . . 139
7.3.1. Quantification of Digital Biomarkers Related to DAS . . . 139
7.3.1.1. Wearable-Only Digital Biomarkers for DAS . . . 145
7.3.2. Performance Comparison . . . 146
7.4. Conclusion . . . 151
8. CONCLUSION . . . 152
8.1. Future Work and Limitations . . . 154
REFERENCES . . . 157
APPENDIX A: TELİF HAKLARIYLA KORUNAN MATERYALİN KULLANIMI . . . 189

LIST OF FIGURES

Figure 2.1. Wearable devices and sensed phenomena on the human body. . . . 14

Figure 2.2. Wearable-based digital biomarkers for psychological targets. . . . 16

Figure 2.3. From raw sensor data to high-level reasoning. 21

Figure 2.4. Tesserae dataset questionnaire contents. 27

Figure 2.5. Sleep efficiency over all semesters. 30

Figure 2.6. NetHealth dataset utilization. 32

Figure 4.1. Target value distribution: Grade. 63

Figure 4.2. Confusion matrix. 65

Figure 4.3. Feature importance for classification. 66

Figure 4.4. Most important factors for sleep quality prediction. 71

Figure 5.1. Distribution of stress classes. 77

Figure 5.2. Feature ranking with class imbalance without personality. 81

Figure 5.3. With class imbalance. 81

Figure 5.4. After SMOTE application. 81

Figure 5.5. Wake-up time vs stress. 82

Figure 5.6. Bed time vs stress. 83

Figure 5.7. Sleep duration vs stress. 83

Figure 5.8. HR samples vs stress. 83

Figure 5.9. Confusion matrix without personality with class imbalance. 84

Figure 5.10. Feature ranking after SMOTE without personality. 85

Figure 5.11. Confusion matrix without personality balanced data. 85

Figure 5.12. Feature ranking with class imbalance with personality. 87

Figure 5.13. With class imbalance. 87

Figure 5.14. After SMOTE application. 87

Figure 5.15. Anxiety vs stress. 88

Figure 5.16. Agreeableness vs stress. 88

Figure 5.17. Conscientiousness vs stress. 88

Figure 5.18. Neuroticism vs stress. 89

Figure 5.19. Confusion matrix with personality with class imbalance. 89

Figure 5.20. Feature ranking after SMOTE with personality. 92

Figure 5.21. Confusion matrix with personality balanced data. 92

Figure 5.22. Positive causal effect. 93

Figure 5.23. Negative causal effect. 95

Figure 6.1. Stress. 112

Figure 6.2. Anxiety. 113

Figure 6.3. Positive affect. 114

Figure 6.4. Negative affect. 115

Figure 6.5. Modality rankings. 116

Figure 6.6. Stress. 117

Figure 6.7. Anxiety. 118

Figure 6.8. Positive affect. 118

Figure 6.9. Negative affect. 118

Figure 7.1. Visualization of multi-task LSTM anxiety predictions. 139

Figure 7.2. Waterfall plot with 20 values for depression. 140

Figure 7.3. Waterfall plot with 20 values for anxiety. 141

Figure 7.4. Waterfall plot with 20 values for stress. 142

Figure 7.5. Bar plot with 10 values for depression. 143

Figure 7.6. Bar plot with 10 values for anxiety. 144

Figure 7.7. Bar plot with 10 values for stress. 144

Figure 7.8. Depression modality. 145

Figure 7.9. Anxiety modality. 145

Figure 7.10. Stress modality. 146

Figure 7.11. Depression. 146

Figure 7.12. Anxiety. 147

Figure 7.13. Stress. 147

LIST OF TABLES

Table 2.1. List of input features (Part 1). 26

Table 2.2. List of input features (Part 2). 28

Table 2.3. Details of the features (wearable and course data) 33

Table 2.4. Details of the features (survey data). 34

Table 3.1. Summaries of papers on NetHealth and Tesserae datasets. 36

Table 4.1. Classification performance details. 64

Table 4.2. Sleep quality prediction results. 72

Table 4.3. Details of model parameters. 73

Table 5.1. Without personality with class imbalance. 82

Table 5.2. Without personality after SMOTE. 86

Table 5.3. With personality with class imbalance. 90

Table 5.4. With personality after SMOTE. 93

Table 6.1. A snapshot of the data for positive affect. 107

Table 6.2. MAE prediction performances (best results in bold). 123

Table 7.1. DAS model predictions: MAE with best performances in bold. . . 149

Table 7.2. Further results of the time-based predictions. 150

LIST OF ACRONYMS/ABBREVIATIONS

AASM	American Academy of Sleep Medicine
AUROC	Area Under the Receiver Operating Characteristic
AUC	Area Under the Curve
BFI	Big Five Inventory
CNN	Convolutional Neural Networks
DAG	Directed Acyclic Graph
DAS	Depression-Anxiety-Stress
ECG	Electrocardiogram
EDA	Exploratory Data Analysis
EEG	Electroencephalogram
EHR	Electronic Health Record
EMA	Ecological Momentary Assessment
EMG	Electromyogram
EOG	Electrooculogram
GMM	Gaussian Mixture Model
GPA	Grade Point Average
HLM	Hierarchical Linear Models
HR	Heart Rate
HRV	Heart Rate Variability
KNN	K-nearest Neighbors
LGCM	Latent Growth-Curve Models
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MICE	Multiple Imputation by Chained Equations
MHMS	Mental Health Monitoring Systems
MSE	Mean Squared Error
MTL	Multi-Task Learning

NDU	Notre Dame University
NMF	Non-negative Matrix Factorization
NuSVR	Nu-Support Vector Regression
OCSVM	One-Class Support Vector Machines
PANAS	Positive and Negative Affect Schedule
PB-state	Physiological and Behavioral State
PCC	Pearson Correlation Coefficient
PHQ-9	Patient Health Questionnaire
PSG	Polysomnography
PSQI	Pittsburgh Sleep Quality Index
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
RF	Random Forest
RNN	Recurrent Neural Networks
SC	Skin Conductance
SHAP	SHapley Additive exPlanations
SMAPE	Symmetric Mean Absolute Percentage Error
SMOTE	Synthetic Minority Over-sampling Technique
STL	Single-Task Learning
SVM	Support Vector Machine
SWB	Subjective Well-being
WHO	The World Health Organization
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

The advent of wearable technology, exemplified by devices such as smartwatches, fitness trackers, and biosensors, has revolutionized the landscape of personal health monitoring. These devices offer a remarkable ability to continuously capture various physiological signals, providing real-time insights into human activities, health parameters, and behavioral patterns. The advantage of wearables lies in their ability to monitor individual metrics and their capacity to decipher temporal patterns, presenting a dynamic source of data that opens avenues for personalized, context-aware applications.

Integrating inertial sensors, such as accelerometers and gyroscopes, with physiological sensors, such as heart rate monitors and skin conductance sensors, has enabled the classification of diverse activities, ranging from basic locomotion to intricate daily functions. From monitoring stress levels and sleep quality to analyzing patterns of daily activities and academic performance, wearables have emerged as versatile tools with applications spanning healthcare, fitness, and workplace productivity enhancement [1–4].

Human well-being monitoring and prediction is a particularly significant application area that has received considerable attention. Well-being is defined as the satisfaction and excitement about life [5], e.g., being comfortable, healthy, happy. How one feels, both physically and mentally, is the essence of living [6]. A low level of well-being corresponds to overall dissatisfaction. Well-being is crucial for public health, which integrates mental and physical health. It is associated with self-perceived health, longevity, health behaviors, mental and physical illnesses, social connectedness, and productivity. In addition, by measuring and tracking well-being, it is possible to promote overall health and prevent diseases [5], [7].

WHO describes a well-being index (WHO-5) predicated on factors such as cheerfulness, calmness, and vigor. Wearables have emerged as valuable tools in the quest to monitor and predict the overall states of physical, mental, and social health, the combination of factors that collectively shape an individual's sense of contentment, happiness, and fulfilment in life.

The increasing awareness of mental health in modern culture and the growing amount of data collected by wearables highlight the importance of this application field. Using information from wearables and surveys, traditional machine-learning algorithms from logistic regression to decision trees and ensemble methods have been crucial in categorizing or forecasting well-being.

Digital biomarkers refer to objective and quantifiable health measures, both physiological and behavioral, derived from data collected by wearable devices, such as smartwatches, sleep sensors, step counters, as well as smartphones [8]. We can count sleep, stress, and mood as certain types of digital biomarkers. In literature, digital biomarkers are studied separately in terms of modalities, and satisfactory results are obtained in terms of the accuracy of the system [9–13]. These heterogeneous modalities correspond to various aspects of an individual, encompassing both physiological and psychological factors. Few studies are focusing on several modalities simultaneously [4], [14]. To be able to measure several modalities at the same time and to obtain a holistic view of a person's current status, cross-level integrative analyses are essential [15]. The integrative approach makes it easy to understand intra-modality relations as well.

1.1. Motivation and Objectives

This thesis focuses on exploring well-being parameters through digital biomarkers obtained from wearables and questionnaires. The main goal is to establish the reliability of these digital biomarkers, derived from both questionnaires and devices like smartwatches and biosensors, in indicating various aspects of well-being such as sleep quality, stress levels, anxiety, depression, and positive-negative affect.

To achieve this objective, the thesis adopts a structured approach. It begins by examining a single factor of well-being, such as sleep or stress. The initial investigation focuses on a single well-being factor, such as sleep or stress, analyzed across various datasets. Subsequently, it progresses towards examining time-based and multi-task methodologies, which handle multiple target variables like stress and anxiety simultaneously.

To facilitate this analysis, two large-scale datasets are utilized. Focusing specifically on workplace well-being prediction, the Tesserae dataset [16] serves as a pivotal resource collected over an extended period from office workers. Within the context of this dataset, we focused on predicting stress, anxiety, and positive and negative affect levels. This selection aligns with the well-being index derived from the WHO-5 questionnaire [17], encompassing these factors and providing a comprehensive framework for the thesis. Furthermore, the NetHealth dataset [18] is a valuable resource for investigating student well-being due to its extended longitudinal coverage (spanning four years). We utilized the NetHealth dataset for the prediction of student grades, sleep patterns, and mental health factors.

1.1.1. Research Gaps in The Literature

A review of the existing literature on well-being prediction revealed several key knowledge gaps. This thesis aims to address these gaps.

Firstly, many studies rely on *small-scale datasets* with limited duration and participant diversity. These datasets often restrict the exploration of temporal aspects of well-being, such as long-term trends and fluctuations over time. In contrast, this thesis emphasizes the use of two extensive datasets (details are given in Section 2) comprising hundreds of participants over prolonged periods, from months to years. This approach allows for a more comprehensive analysis of how well-being evolves over time (addressed in Section 6 and Section 7), and the factors that influence its trajectory.

Secondly, a major challenge is the *interpretability of ML models* used for well-being prediction. Many models are considered 'black box', making their internal workings opaque. This limits the ability to explain why certain features or digital biomarkers predict changes in well-being. Moreover, there is a gap in understanding how these predictive features relate to established well-being digital biomarkers in psychological literature. This thesis addresses these gaps by analyzing features and digital biomarkers of well-being in detail (in Section 5, Section 6 and Section 7), aiming for deeper insights into well-being prediction mechanisms.

Another prominent gap in the literature is the tendency to focus on predicting specific aspects of well-being, such as stress or anxiety, rather than considering the broader, *multifaceted nature of well-being*. Well-being encompasses various dimensions, including sleep, physical activity, stress, depression, positive and negative affect, which interact and influence each other. This thesis aims to bridge this gap by exploring multiple dimensions of well-being simultaneously (all sections, particularly Section 6 and Section 7), offering a more holistic understanding than traditional single-measure approaches.

Furthermore, *temporal dynamics* are crucial in understanding well-being, yet the temporal dimension of data is often overlooked, particularly regarding how previous assessments influence future levels of well-being. The fluctuation of well-being over time and the impact of past states on future outcomes are pivotal factors frequently disregarded in current research. Wearables, capable of collecting longitudinal data, offer a unique opportunity to explore stress, anxiety, and emotional states across time. Integrating time-based prediction techniques with wearable data introduces an innovative approach to capturing temporal patterns. This comprehensive strategy not only aids in the early detection of well-being fluctuations but also enables timely interventions to promote mental well-being. Through the utilization of longitudinal studies (Section 6 and Section 7), this thesis aims to uncover contributing factors to these fluctuations and to comprehensively understand the temporal dynamics of well-being.

While some studies have employed advanced methodologies such as time-series analysis [19–21] and multi-task learning [22–25] for predicting mood using wearable sensor data or self-reported mood data, these approaches are notably absent when it comes to applying them to large-scale datasets for predicting overall well-being comprehensively.

Furthermore, the integration of deep-learning architectures with data from wearable devices has led to the development of *hybrid models* that exploit both sensor data and contextual information. These models utilize diverse inputs, including physiological signals, accelerometer data, and contextual factors like time of day, location, social interactions, and self-reported mood states. The effectiveness of these models can be significantly enhanced by combining multiple data sources that impact an individual's well-being levels. This thesis aims to bridge this gap by employing these advanced techniques on extensive datasets, aiming to enhance the accuracy and practical utility of models for predicting well-being.

To sum up, this thesis identifies and addresses critical gaps in the existing literature on well-being prediction. By doing so, it aims to contribute to the field, offering insights that can inform interventions and strategies aimed at improving individual and societal well-being outcomes.

1.1.2. Research Questions

The research questions addressed by the sub-studies conducted within the scope of this thesis can be grouped as follows

1.1.2.1. Group 1: Digital Biomarkers Targeting Well-Being.

- *RQ1:* What are the highly related digital biomarkers targeting student's academic grade in the NetHealth dataset?
- *RQ2:* What are the highly related digital biomarkers targeting student's sleep in

the NetHealth dataset?

- *RQ3:* What are the highly related digital biomarkers targeting worker's stress in the Tesserae dataset?
- *RQ4:* What are the underlying digital biomarkers of the targeted well-being factors in the NetHealth and Tesserae datasets?

1.1.2.2. Group 2: Validation and Prediction.

- *RQ5:* How do the machine learning algorithms (e.g., RF, CNN) perform for classifying/predicting the target values of well-being?
- *RQ6:* Are the resulting revealed digital biomarkers compatible with conventional psychological studies?
- *RQ7:* Does the historical levels of well-being impact future levels?
- *RQ8:* Is it possible to improve time-based prediction performances by employing multi-task learning since the targeted well-being parameters have commonalities?

1.2. Contributions and Findings

This section outlines how each research question is addressed within specific sections of the thesis and provides a summary of findings for exploration in those sections.

The detailed findings from these studies are provided in the respective sections: RQ1 and RQ2 are addressed in Section 4, where student's academic grade and sleep related digital biomarkers are analyzed using the NetHealth dataset. RQ3 is addressed in Section 5, where stress-related digital biomarkers are analyzed using the Tesserae dataset. RQ4 is covered in Section 6 and Section 7 with a focus of several well-being digital biomarkers such as stress, anxiety, depression, positive and negative affect. RQ5 and RQ6 are covered in all sections to show the model performances and compare the obtained digital biomarkers validity in psychological studies. RQ7 is explored in Section 6 and Section 7, with a focus on incorporating time-based analysis techniques. RQ8 is addressed in Section 7 by including multi-task learning along with time-based analysis.

1.2.1. Novelty

The main contribution of the thesis is the system design and development of multi-modal personal data analysis methodologies and extracting related digital biomarkers. To the best of our knowledge, even though several studies exist in the literature regarding the quantification of digital biomarkers on one target variable, this thesis differs in terms of the datasets, employing the time-aspect and multi-modality, further advancing by dealing with their personalized versions in an incremental manner. Developing such systems and algorithms may facilitate further analysis of real-world applicability.

Given research questions delve into validating the identified digital biomarkers against established psychological frameworks and exploring predictive modeling techniques for assessing temporal trends in well-being. Our findings underscore the integration of machine learning approaches, such as multi-task learning, to enhance predictive accuracy across varying temporal resolutions. This approach addresses gaps in current literature by demonstrating the applicability of advanced computational methods in well-being research.

In summary, our thesis contributes to the field by not only identifying key digital biomarkers associated with diverse well-being factors but also by validating their utility in psychological contexts and enhancing prediction capabilities through innovative machine learning methodologies.

1.2.2. Findings

In this thesis, we investigate the recognition of both simple and complex well-being factors using various machine learning algorithms. Through this exploration, we provide insights into the underlying digital biomarkers of targeted well-being factors, as detailed in our studies [26–30]. Additionally, we assess the compatibility of these digital biomarkers with conventional psychological studies, shedding light on their reliability and validity in mental health assessment.

1.2.2.1. Academic Performance Classification. In the context of academic grade classification (Section 4.1), we employed the RF machine learning technique to classify and predict students' grades based on a combination of survey and wearable data. Feature importance analysis identified key factors influencing model performance, highlighting the significance of survey parameters related to sleep, the Big Five personality traits, health, mental health, personal information, and origin. Notably, none of the parameters collected from wearables appeared on the highly important parameter list, suggesting that personality features and subjective survey responses dominate objectively measured parameters.

1.2.2.2. Sleep Efficiency Prediction. For predicting sleep efficiency values obtained from Fitbit devices (Section 4.2), we explored prominent CNN architectures alongside the conventional RF model to compare outcomes comprehensively. Despite promising prediction results from CNNs, the RF model exhibited superior performance, albeit with a marginal difference in MAE. This subtle variation may be attributed to the inherent characteristics of our dataset, aligning with findings where RF consistently outperformed various deep learning techniques on the same dataset. Feature reduction experiments demonstrated that deep learning without explicit feature extraction outperforms models incorporating manually selected features. However, the feature ranking derived from this process provides valuable insights for interpretation. The top influential factors revealed that sleep efficiency is significantly influenced by wearable activity measurements and survey-related factors such as exercise, depression, and personality traits, supporting existing literature on the interconnectedness of sleep, stress, physical activity, and mood.

1.2.2.3. Stress State Prediction. We examined digital biomarkers related to stress using the Tesserae dataset collected from office workers (Section 5). Using RF, we classified stress states based on multi-modal data and employed SMOTE to address class imbalance. The analysis revealed that phone activity and sleep-related markers are most relevant to stress, with the highest classification accuracy of 80% achieved when

including personality-related parameters. These findings confirm literature that examines the relationship between sleep, well-being, stress, positive affect, mood, and physical activity.

1.2.2.4. Temporal Learning Approaches. We delve into the potential enhancement of prediction performances by incorporating time-based (lagged) versions of conventional algorithms (Section 6 and Section 7). This approach reveals that lagged versions of algorithms can improve prediction accuracy by considering historical data and temporal dependencies. Comparing our findings with literature utilizing time-based techniques on other well-being datasets, we observe similar performance improvements from daily to weekly resolutions. This method shows promise in capturing the interplay between various well-being indicators over time.

1.2.2.5. Multi-Task Learning Approaches. Section 7 uniquely employs different well-being factors by employing MTL to further improve prediction accuracy by leveraging the commonalities among targeted well-being parameters (Section 7). Incorporating MTL techniques on student well-being prediction revealed the potential to capture temporal trends by introducing lagged versions of both baseline and MTL algorithms that account for diverse time windows. This approach extends the application of MTL to the NetHealth dataset, addressing a gap in current research.

Incorporating a temporal perspective, this thesis scrutinizes the influence of past measurements on mental health outcomes. The multi-task learning model is tailored to discern intricate relationships among depression, anxiety, and stress, acknowledging both shared patterns and individual nuances within the dataset along with consideration of time relations. By combining time-based predictive models and multi-task learning, this thesis aims to provide insights into the various aspects that influence mental health within the university setting.

1.2.3. Academic Outputs of the Thesis

The studies conducted within the scope of this thesis are as follows

- Saylam, B., E.Y. Ekmekci, E. Altunoğlu, and Ö.D. İncel, "Academic Performance Relation with Behavioral Trends and Personal Characteristics: Wearable Device Perspective", *Pervasive Health and Smart Sensing Conference*, pp. 35-39, Ljubljana, Slovenija, 2022. (Section 4.1)
- Kilic, O., B. Saylam, and Ö. D. İncel, "Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning", *International Conference on Machine Learning Technologies*, pp. 116-120, Stockholm, Sweden, 2023. (Section 4.2)
- Saylam, B., and Ö. D. İncel, "Extracting Digital Biomarkers for Unobtrusive Stress State Screening from Multimodal Wearable Data", *International Conference on Intelligent Edge Processing in the IoT era*, pp. 139-160, Online, 2022. (Section 5)
- Saylam, B., and Ö. D. İncel, "Quantifying Digital Biomarkers for Well-Being: Stress, Anxiety, Positive and Negative Affect via Wearable Devices and Their Time-Based Predictions", *Sensors*, Vol. 23, No. 21, pp. 8987-9005, 2023. (Section 6)
- Saylam, B., and Ö. D. İncel, "Multitask Learning for Mental Health: Depression, Anxiety, Stress (DAS) using Wearables", *Diagnostics*, Vol. 14, No. 5, pp. 501-518, 2024. (Section 7)

1.3. Thesis Outline

In this section 1, we set the foundation for our research by outlining the motivation and specific objectives. We discuss the distinctive contributions our work brings to the existing body of knowledge, emphasizing their significance. Additionally, we provide an overview of the thesis structure to guide readers through the document.

In Section 2, we provide essential background information and introduces the Tesserae and NetHealth datasets. We delve into the role of wearables in health monitoring, define digital biomarkers, and discuss the relevance of ground truth data and behavioral questionnaires.

Section 3 reviews current research on mental health and well-being monitoring using the Tesserae and NetHealth datasets. It examines studies predicting well-being factors, noting whether they utilize the entire datasets or specific portions relevant to our research. Key papers are highlighted, and distinctions between these studies and our thesis are drawn. Secondary studies using other datasets are also acknowledged in the related works sections.

In Section 4, our first set of results focuses on the NetHealth dataset, specifically exploring the prediction of academic grades and sleep quality among college students. We present methodologies and findings, shedding light on the factors influencing academic performance and the relationship between wearables and sleep patterns.

In Section 5, we turn our attention to stress prediction among workers using the Tesserae dataset. Our objective is to extract novel digital biomarkers associated with stress, employing diverse parameters from mobile phones, wearables and questionnaires. We utilize the RF algorithm to assess the impact of essential parameters on stress levels.

Section 6 delves into the exploration of time-based patterns in well-being, specifically stress, anxiety, positive, and negative affect. We apply baseline methods like RF, XGBoost, and LSTM, enhancing their performances by incorporating time-lagged versions of the dataset.

Examining the interconnected nature of mental well-being, Section 7 focuses on MTL for various mental health factors. We explore related digital biomarkers, apply baseline methods, and assess the impact of MTL on model performance.

In the final section (Section 8), we conclude the thesis by summarizing key findings and emphasizing the overall significance of our research. This section serves as a synthesis of the entire document, providing insights for future research directions.

2. BACKGROUND

In this section, we delve into a comprehensive exploration of wearables and digital biomarkers used within the context of this thesis. Our focus extends to the landscape of sleep quality assessment, and the evaluation of mood and stress. By navigating through the background of these components, we aim to pave the way for a detailed understanding of their role in predicting well-being factors. In addition, we also provide some of the ground truth methods used in the context of wearable studies to validate the machine/deep learning algorithms. Furthermore, we explain the datasets and their parts used in the scope of the thesis.

2.1. Sensors Integrated into Wearables

Because of their low cost, small size and computational power, wearable devices have become part of our daily lives, and they provide continuous measurements with embedded sensors. The main aim of sensor embeddings into devices was to extract information about the device's context. Nowadays, it is also used to understand the health and behavior of a carrying person. Some of the health-related phenomena and corresponding wearables are shown in Figure 2.1. For instance, with wearable devices, it is possible to track movement, location, sound, and light as physical conditions [31–33]; communication, social interaction, sleep, mood, stress as behavioral conditions [3], [11], [13]; and depression, bipolar disorder, schizophrenia as clinical conditions [34–36]. These behaviors, thoughts, feelings, traits, or identification of states are called as *behavioral markers* in the literature [37].

The literature identifies various digital biomarkers [8], including *pharmacodynamic response biomarkers* to measure changes in response to pharmacological agents [38], *diagnostic biomarkers* for disease detection [39], and *monitoring biomarkers* for tracking medical conditions [40]. This thesis focuses on monitoring digital biomarkers and collects data in a user-friendly manner during daily activities.

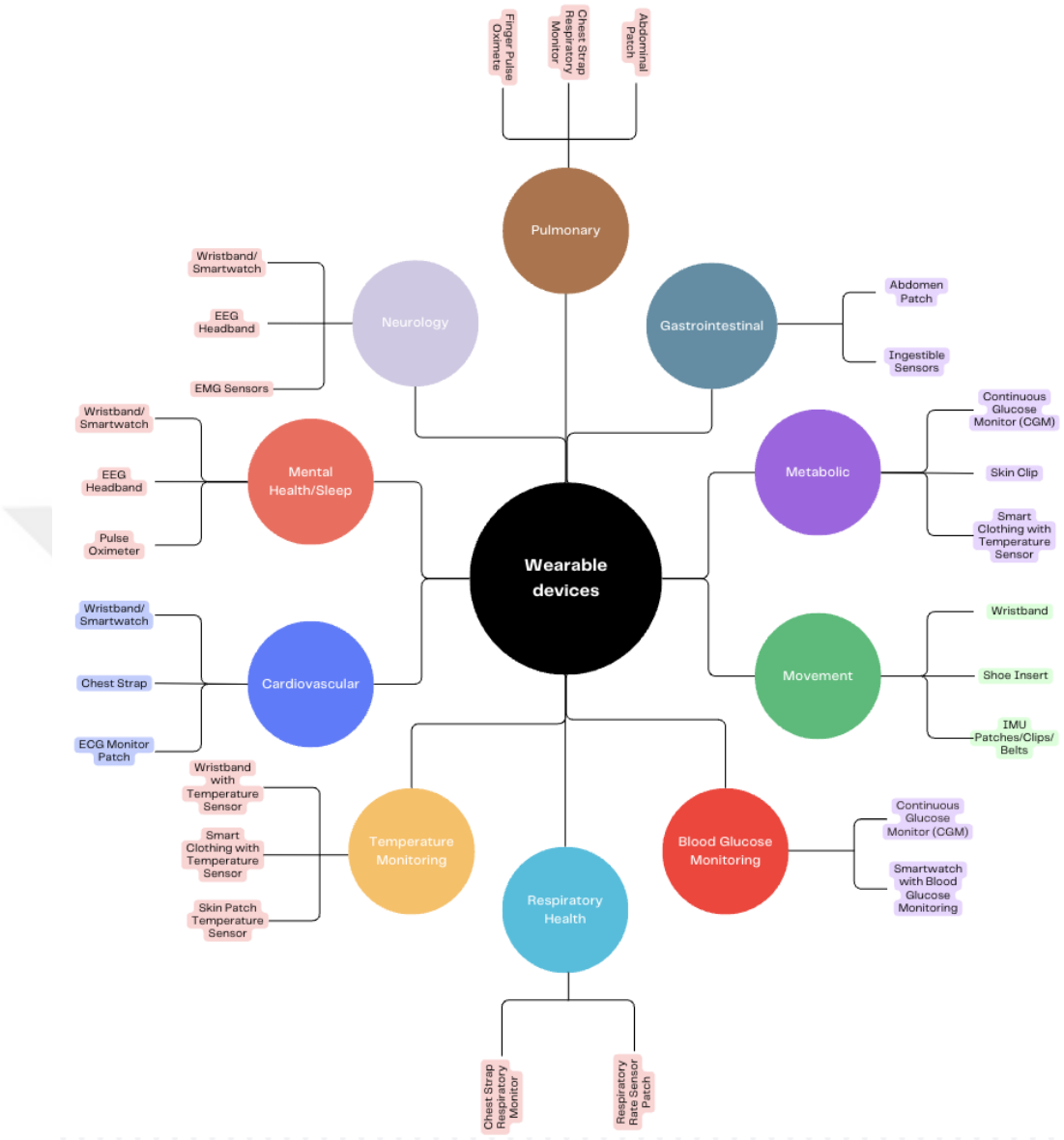

Figure 2.1. Wearable devices and sensed phenomena on the human body.

2.2. Digital Biomarkers

Recognizing the pivotal role of lifestyle modifications in enhancing health and preventing chronic diseases is a widely acknowledged fact [7]. This realization prompts adopting changes that promote a healthier life sustained over time. Wearables and mobile sensors emerge as instrumental tools in this pursuit, tracking daily behavioral patterns and furnishing the overall state for targeted lifestyle adjustments.

The capacity of wearables, including smartwatches, sleep sensors, and step counters, to monitor health parameters from daily activity to sleep quality and stress during various tasks is transformative. This digitization of healthcare measurements has given rise to the concept of *digital biomarkers*. They comprise a range of metrics or attributes calculated via digital health technologies. These digital biomarkers are quantitative, objective measurements obtained from wearables and phones that are behavioral and physiological in nature [8].

Digital biomarkers hold particular significance in continuously monitoring and early detection of critical conditions, such as depression. For instance, researchers [41, 42] employ location entropy, measuring the change of time spent at various locations, as an indicator of depression. Spending equal time in diverse locations is notably associated with lower levels of depression. Behavioral methods also play a crucial role in extracting non-verbal cues related to emotions and stress through attributes like facial expressions, ear movements, and overall body motion [43]. The terms 'activity' and 'behavioral recognition' are also sometimes interchangeably used in studies [44,45].

2.3. From Sensors to High Level Recognition: Sleep, Mood, Stress Digital Biomarkers

The growing concerns in modern society regarding stress and anxiety have become prominent, significantly impacting people's overall well-being, productivity, and quality of life. Recognizing the critical need to detect and effectively manage these conditions, there is a rising interest in leveraging the sensing capabilities of wearable devices. As wearable technology advances to incorporate an expanding range of sensing capabilities including skin conductance, heart rate, and more researchers are focusing on extracting valuable insights from individuals' daily behaviours. while uncovering the various factors influencing a person's activity, behavior, social context, mood, and stress levels.

A typical methodology cycle can be seen in Figure 2.2, which starts with collecting data actively and passively using wearable devices and smartphones; extracting digital biomarkers related to well-being states such as depression, stress, anxiety, and overall well-being; and analyzing these digital biomarkers using predictive models or analytical techniques. This process enables continuous monitoring and assessment of psychological states in real-world settings, leveraging the extensive data captured by wearables to gain insights into mental health and well-being. Furthermore, the objective can extend beyond mere monitoring; it aims to address potential issues before they escalate to more severe states. This proactive approach can further involve providing timely feedback and personalized recommendations for daily living to prevent the worsening of stress-related conditions [13].

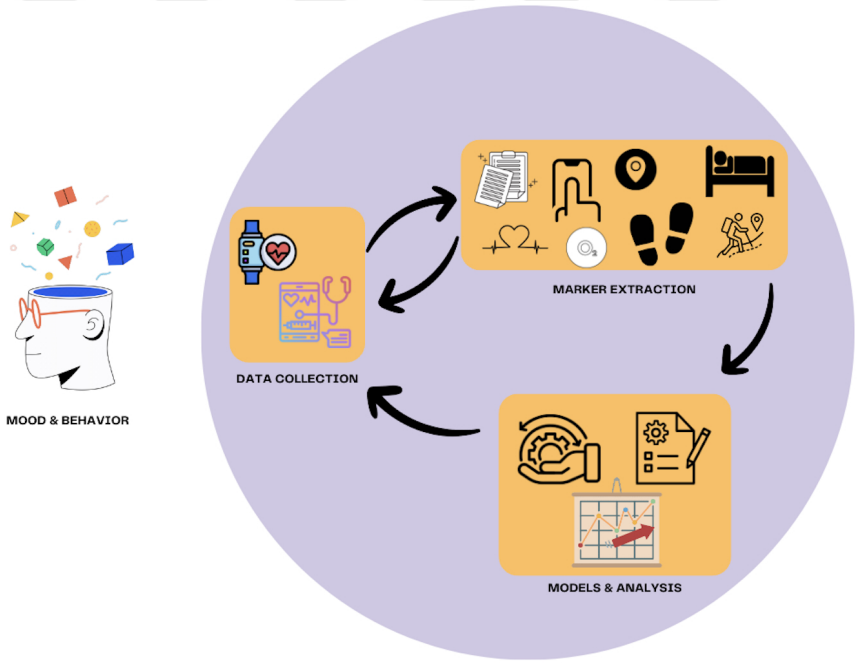

Figure 2.2. Wearable-based digital biomarkers for psychological targets.

In the following, we define and provide background on the well-being conditions that are analyzed throughout the thesis.

2.3.1. Sleep

Research in sleep monitoring holds significant importance due to its profound impact on daily performance and mood. The duration, timing, regularity, and quality of sleep are critical factors influencing both physical and psychological health. Irregular or poor sleep patterns can trigger conditions such as diabetes, obesity, depression, and immune system failures, making this research area paramount [12], [46].

PSG stands as the gold standard for sleep measurement, employing EEG, ECG, EMG, EOG, and other factors [47]. While PSG provides a detailed picture of sleep patterns, its obtrusiveness and high cost pose challenges for long-term monitoring, limiting its applicability primarily to in-clinic research.

Actigraphy [48] emerges as a practical alternative to PSG, utilizing body movement to infer sleep and wake states. With higher prediction quality (85% - 96%), actigraphy enables more convenient long-term sleep monitoring. Recognized by the AASM for detecting specific sleep disorders, it still faces challenges. Some individuals find it obtrusive, and its wrist-watch design, relying on accelerometers, may be unreliable during periods of inactivity.

Subjective methods, such as self-rated questionnaires (e.g., Pittsburgh Sleep Quality Index - PSQI) and sleep diaries, offer additional insights on sleep quality. Ongoing research explores integrating smartphones and wearables to address actigraphy's limitations, combining various parameters for improved sleep/wake detection [12]. Other studies delve into sleep/wake classification through smartphone usage patterns, contributing to a comprehensive understanding of sleep patterns and monitoring [49].

2.3.2. Mood

Positive affect, which encompasses emotions like joy, happiness, enthusiasm, and contentment, plays a vital role in well-being and mental health assessment. It repre-

sents the spectrum of pleasant feelings contributing to an individual's daily emotional experiences. Understanding and measuring positive affect are integral to assessing overall happiness and fostering psychological resilience.

Conversely, *negative affect* embodies emotions like sadness, anger, fear, and disgust, constituting distressing feelings that can impact mood and well-being. The comprehension and measurement of negative affect are crucial for evaluating emotional health and identifying psychological distress. While negative affect can be a natural response to adverse events, persistent or severe negative affect may signal underlying emotional or mental health issues.

Anxiety, as a persistent state, entails excessive worry, fear, or apprehension concerning future events or situations. It is characterized by heightened alertness and an underlying sense of unease. While anxiety can be a natural response to stressors, its impact becomes concerning when it reaches overwhelming levels, becomes intrusive, or impedes daily functioning. In such cases, it may indicate the presence of an anxiety disorder. These anxiety disorders, including generalized anxiety disorder, social anxiety disorder, and panic disorder, present unique features and symptoms, adding complexity to anxiety-related challenges. Recognizing the nuances of anxiety and its diverse forms is essential for accurate identification, diagnosis, and tailored intervention.

Depression is a pervasive mental health condition marked by persistent and deep-seated feelings of sadness, hopelessness, and diminished interest or pleasure in once-enjoyable activities, leading to substantial impairment in daily functioning and overall quality of life [50, 51].

2.3.3. Stress

Stress, whether psychological or physiological, arises as a reaction to challenging situations or perceived threats, eliciting complex emotions such as tension, pressure, and unease. This complex response can manifest in both mental and physical

symptoms, acting as a natural reaction to life events. However, when stress becomes excessive or chronic, it negatively impacts one's mental and physical health.

This multifaceted response is influenced by a myriad of factors spanning environmental (work, living conditions), psycho-social (life events, everyday challenges, social relationships), biological (genetics, physical health, neurobiology), psychological (personality traits, cognitive factors, resilience), lifestyle (diet, nutrition, physical activity, sleep patterns, substance use), and cultural-societal (beliefs-expectations, discrimination, social inequality) realms.

Originally defined as a "non-specific response to any demand given by the body" by an endocrinologist in 1936 [52], stress, as conceptualized by behavioral scientists, represents "the response of the body ending with anxiety or discomfort in the face of any threat" [53]. Associated with conditions like Acute Stress Disorders and Post Traumatic Stress Disorder [52], [54,55], heightened stress levels correlate with increased occurrences of heart diseases and depressive symptoms [56, 57].

Understanding and addressing stress requires continuous advancements in monitoring and prediction methodologies to navigate its intricate nature effectively. There is a lot of research on the prediction of stress, and recent advancements [14], [58] in research methodologies have shifted towards predictive models, moving away from binary mood classifications [59, 60]. Wearable devices have become indispensable in monitoring and managing stress levels, aiming to maintain a balance within healthier boundaries [11], [61–63]. Commonly utilized sensors in wearables [11], [62], [64] include accelerometers, blood pressure monitors, SC sensors, HRV monitors, and HR sensors, often integrated with ground truth data collection through surveys and questionnaires.

To sum up this section, we can see the higher-level information extraction hierarchy in Figure 2.3 from raw sensor modalities to behavioral markers, which will then help to infer clinical states.

Figure 2.3 illustrates the progression from raw sensor data to high-level clinical reasoning through a structured, multi-step process. The initial stage involves data collection using a variety of sensors, including biometric sensors, activity trackers, environmental sensors, and social interaction sensors. Each sensor type captures specific primary data features: biometric data (such as heart rate or body temperature), activity patterns (like steps taken or sleep duration), environmental factors (such as ambient temperature or light levels), and social interaction features (such as frequency and duration of social contacts).

These primary data features are then analyzed to extract advanced behavioral markers. For example, biometric data can inform movement analysis, speech and language data can be derived from social interaction features, visual analysis can be conducted using environmental factors, and pattern recognition can be applied across various datasets to identify significant trends and anomalies.

The final stage involves clinical evaluation, where these advanced behavioral markers are used for diagnostic assessment, treatment monitoring, risk assessment, and overall health and well-being evaluation. Each subtype of data and its corresponding behavioral marker contribute to a comprehensive understanding of an individual's health, enabling precise and personalized clinical interventions. For instance, movement patterns can aid in diagnosing neurological conditions, speech and language markers can monitor cognitive health, visual analysis can detect environmental risk factors, and pattern recognition can help in predicting health risks. This integrative approach demonstrates the value of combining sensor data with advanced analytics to enhance clinical decision-making and improve patient outcomes.

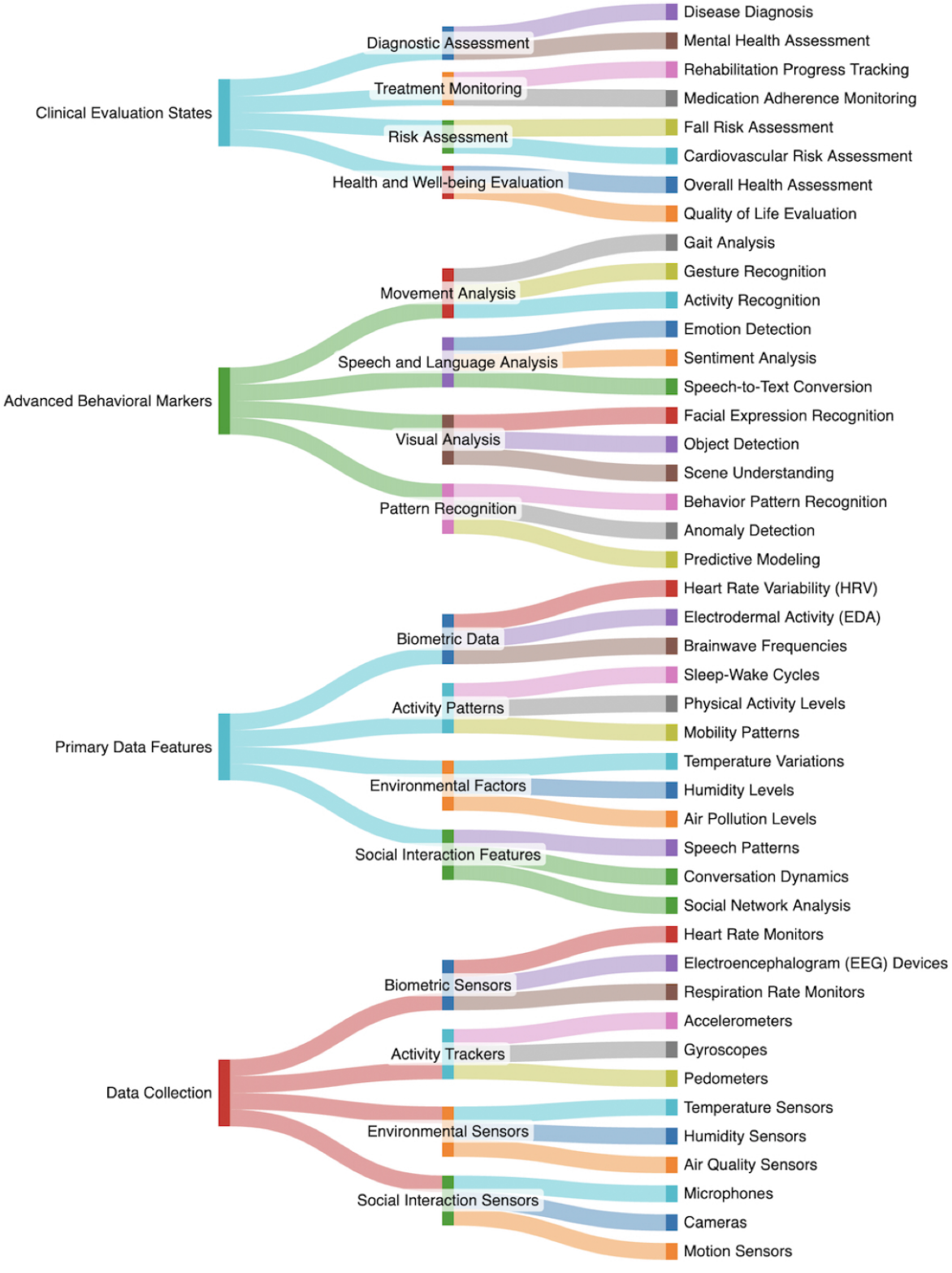

Figure 2.3. From raw sensor data to high-level reasoning.

2.4. Ground Truth Methods for the Analysis of Psychological Targets

Defined as "the association between data and the outcome" [65], ground truth data assumes qualitative or quantitative forms depending on the studies. These values are indispensable for testing and enhancing the performance of supervised machine learning techniques. Despite the aims of personal sensing research to develop unobtrusive and ambulatory systems, corresponding labels are required to train the system.

Ground truth serves as a reliable benchmark for evaluating and validating the predictions in data analysis. In the context of this thesis, questionnaires are a common method for gathering subjective responses and ground truth data. However, it is important to acknowledge that alternative methods exist beyond questionnaires. Other approaches, such as direct observations, clinical assessments, or physiological measurements, may also be employed in different studies, emphasizing the diverse range of techniques available for establishing ground truth in research.

The following provides an example list of commonly used questionnaires for behavioral markers, offering insight into specific tools employed in the research process. It is crucial to recognize that the following methods discussed are illustrative. The choice of questionnaires depends on the specific requirements and objectives. The surveys and questionnaires used throughout the thesis in each section are elaborated in Sections 2.5.1.2 and 2.5.2.2.

- *EMA:* It captures thoughts, emotions, behavior, physiological states of individuals, and context in their natural environment using wearables. It offers advantages over traditional assessments by soliciting real-time states rather than reflections on the current situation, reducing bias. EMA includes active (self-report) and passive (observational via wearables) types and is commonly used in mood disorders, depression, eating disorders, mental health, and sleep research. Typical questions: "Rate your overall stress level at this moment? (1-5 rating)", "Rate your overall mood at this moment? (1-5)", "What are you doing now? (e.g.,

watching TV, doing chores, chatting, going shopping)", "Who are you with right now? (free filling)", "Is there something occupying your mind that you cannot get out of your mind? (yes/no)", "How much difficulty are you having in getting it out of your mind? (1-5)", "How well did you sleep last night? (1-5)" [66].

- *PHQ-9:* This depression questionnaire, widely used in studies, comprises 9 questions and takes about 3 minutes to complete. It addresses mood, anxiety, alcohol use, eating habits, and somatoform disorders. The score spans 5 levels: none, mild, moderate, moderately severe, and severe depression [67].
- *Big Five Personality Information:* This questionnaire provides insights into people's behavior by analyzing their personality characteristics. Typically administered at the beginning of an experiment, it aims to capture stable traits that are not expected to change during the wearable data collection process. The questionnaire comprises 5 components, measuring extraversion, agreeableness, emotional stability, conscientiousness, and openness.
- *PSQI:* This index is focused on assessing sleep quality across various dimensions, including latency, duration, efficiency, medication use, disturbances, and dysfunction. Each question's response is weighted between 0-3, and with 7 components, the overall score ranges from 0-21. A lower score indicates better sleep quality.

2.5. Datasets Utilized in the Thesis

It is important to note that, in the context of the thesis, instead of collecting new data, the decision was made to utilize available two datasets due to their extensive nature, encompassing a large number of participants over a considerable period. Additionally, these datasets were collected in real-world settings rather than controlled laboratory environments. This choice was made to avoid potential limitations associated with collecting new data, such as constraints on sample size and environmental control.

2.5.1. Tesserae Dataset [16]

2.5.1.1. General Overview. The Tesserae Dataset [2] [16], [68], collected to measure workplace performance, psychological traits, and physical characteristics over a year from 757 subjects, utilized a phone and wearable device (Garmin) as instruments. Data collected from the phone included screen lock/unlocks, data usage, charging, location, phone states, beacon sightings, and wearable data streaming. The Garmin watch, worn during the study period, recorded heart rate, HRV, step count, sleep duration, sleep staging, and stair count values, boasting 5-7 days of battery life. Bluetooth beacons were utilized for extracting location information, such as home or office. In addition to non-verbal measurements from devices, verbal data from social media platforms like Facebook and LinkedIn were gathered, though inaccessible. The data collection campaign was approved by the University of Notre Dame as a research project, with participants providing consent.

For ground truth data collection, various questionnaires covering job performance, intelligence, mood, anxiety, health measures, exercise, sleep, and stress were distributed. Participants filled out all questions at the study's onset (approximately one-hour duration) and daily surveys, summarizing each measurement type, requiring only a few minutes. The overall score of each questionnaire corresponds to one column in the daily survey scores file. Exit section questionnaires and follow-up surveys were also collected but remained inaccessible. Therefore, the questionnaires are categorized into four main parts: Initial Ground Truth, Daily Surveys, Exit, and Follow-up Survey. The details of the collected ground truth questionnaires can be found in Figure 2.4.

2.5.1.2. Thesis Related Sub-Dataset. The shared data file did not provide access to all measurements, so our analysis was conducted on a subset of the original dataset. While data was collected from 757 participants, we only had access to data from 727 participants. Despite detailed sampling measurements from the wearable device occurring from 3 minutes to 15 minutes, depending on data modalities, our analysis was

[2] https://tesserae.nd.edu/

conducted on a daily basis since the daily surveys could be utilized as the ground truth data for the targeted psychological aspects. Despite the year-long data collection, only 56 days of daily survey data were utilized as target variables throughout the thesis[3] .

To construct our dataset, we merged HR, Activity, Daily, Stress, Sleep, Weather, Phone Activity, HR WorkDesk Home and Daily Scores files in the dataset based on ParticipantID and Timestamp. This allowed us to obtain wearable and ground truth data for each participant on a daily basis. During merging, it was noted that for some participants, the amount of data from wearables exceeded the ground truth survey answers, indicating extended data collection periods with the devices. However, as ground truth information was essential for our analysis, we only considered participants' data with ground truth responses, i.e., responses to the daily surveys, which are BFI, PANAS, Omnibus Anxiety Questionnaire, MITRE Omnibus Stress Questionnaire.

After constructing the data file, we obtained 36,294 rows of data from all participants. Additionally, files were merged by considering the common columns in each file. Consequently, common columns such as ParticipantID, Timestamp, and Date were not included in the final construction, as a one-time indication sufficed for the entire dataset.

In the final dataset, we have 269 columns, with 15 of them representing ground truth information. These 15 columns include survey name, stress, anxiety, sleep, positive affect, negative affect, extraversion, agreeableness, conscientiousness, neuroticism, openness, total phone activity duration, survey sent time, survey start time, and survey finish time. Furthermore, we excluded other well-being factor columns while targeting a specific one. For example, in the case of stress prediction, we removed *anxiety, pos affect, neg affect*. The list of input features are in Table 2.1 and 2.2.

The constructed Tesserae sub-dataset is utilized in Section 5 for stress classification and in Section 6 for the prediction of stress, anxiety, positive and negative affect

[3]https://osf.io/yvw2f/wiki/EMAs/

levels using a time-based approach. This section solely covers the data construction process for the Tesserae dataset, with further preprocessing for analysis explained in Section 5.3.2 for stress classification and in Section 6.3 for time-based analysis.

Table 2.1. List of input features (Part 1).

Category	**Features**
General	ParticipantID, Timestamp
Activity	active secs, highly active secs, total activity secs
Sleep	adjusted bed time, adjusted sleep duration, adjusted wakeup time
Daily Activities	active kilocalories, active time seconds, activity stress seconds, average hr, average stress level, distance in meters, duration in seconds, floors climbed, high stress duration seconds, low stress duration seconds, medium stress duration seconds, moderate intensity duration seconds, resting hr, stress duration seconds, steps, vigorous intensity duration seconds
HR	ave hr, max hr, median hr, min hr, mode hr, range hr
Stress	ave stress, max stress, median stress, min stress, mode stress, range stress
Daily Scores (Ground Truth)	agreeableness, anxiety, conscientiousness, date, extraversion, negative affect, neuroticism, openness, positive affect, stress, total pad

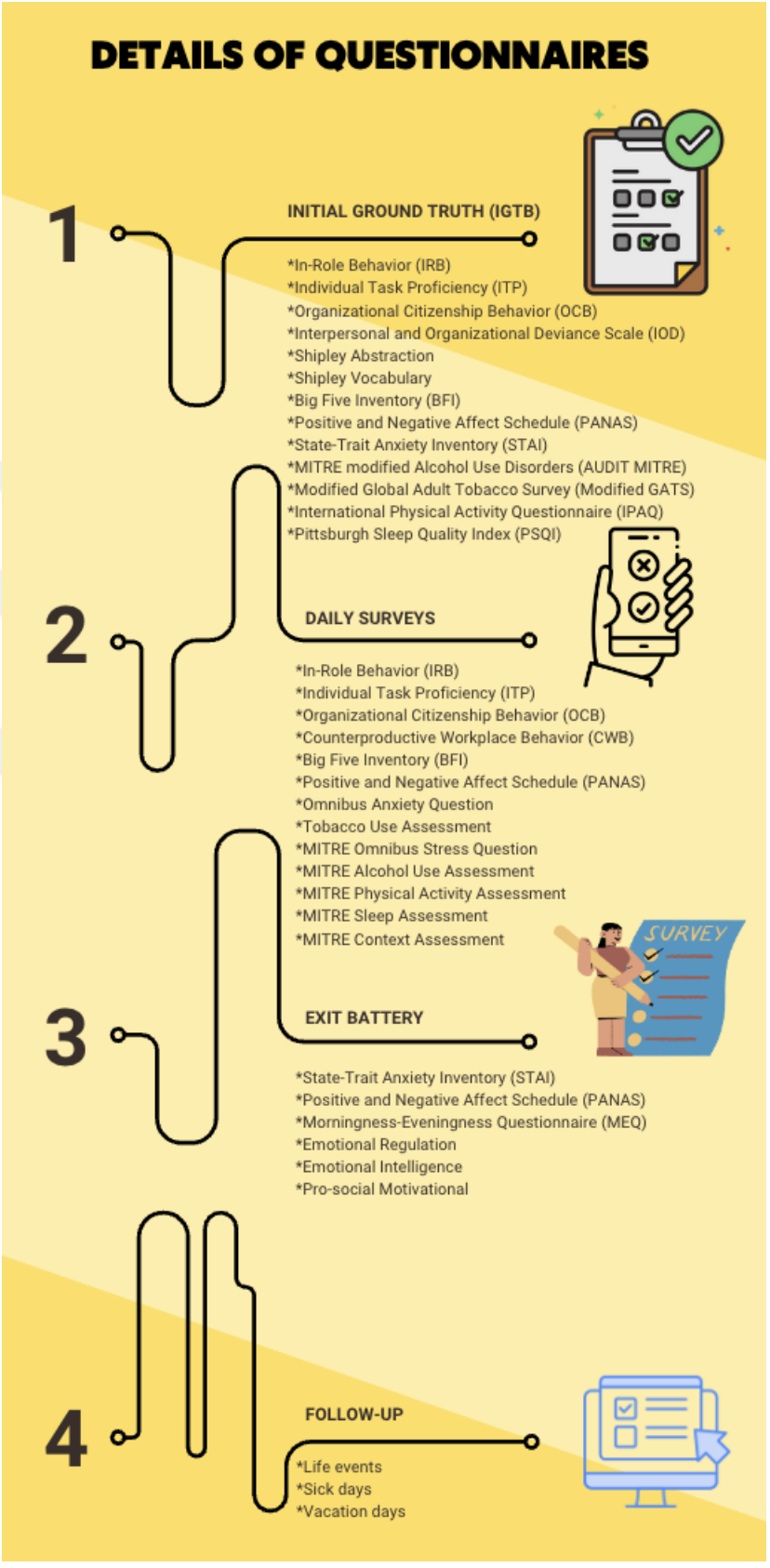

Figure 2.4. Tesserae dataset questionnaire contents.

Table 2.2. List of input features (Part 2).

Category	**Features**
Mixed Physical Activity	act in vehicle ep0, act in vehicle ep1, act in vehicle ep2, act in vehicle ep3, act in vehicle ep4, act on bike ep0, act on bike ep1, act on bike ep2, act on bike ep3, act on bike ep4, act running ep0, act running ep1, act running ep2, act running ep3, act running ep4, act unknown ep0, act unknown ep1, act unknown ep2, act unknown ep3, act unknown ep4, act walking ep0, act walking ep1, act walking ep2, act walking ep3, act walking ep4, bbidp median, calories active hrs, caloriesdp median, garmin calories max, garmin calories, mean, garmin calories median, garmin calories min, garmin calories std, garmin steps max, garmin steps mean, garmin steps median, garmin steps min, garmin steps std, gimbal active hrs, gimbal ep0, gimbal ep1, gimbal ep2, gimbal ep3, gimbal ep4, gimbaldp median, quality activity, quality bbi, quality calories, quality gimbal, quality hr, quality steps, steps active hrs, stepsdp median
Weather	ave cloudcover, ave daytime cloudcover, ave daytime feelslikef, ave daytime heatidxf, ave daytime humidity, ave daytime pressure, ave daytime tempf, ave daytime visibility, ave daytime windchillf, ave daytime windgustmph, ave daytime windspeedmph, ave feelslikef, ave heatidxf, ave humidity, ave pressure, ave tempf, ave visibility, ave windchillf, ave windgustmph, ave windspeedmph, max cloudcover, max daytime cloudcover, max daytime feelslikef, max daytime tempf, max daytime visibility, max daytime heatidxf, max daytime humidity, max daytime pressure, max daytime windchillf, max daytime windgustmph, max daytime windspeedmph, max feelslikef, max heatidxf, max humidity, max pressure, max tempf, max visibility, max windchillf, max windgustmph, max windspeedmph, min cloudcover, min daytime cloudcover, min daytime feelslikef, min daytime heatidxf, min daytime humidity, min daytime pressure, min daytime tempf, min daytime visibility, min daytime windchillf, min daytime windgustmph, min daytime windspeedmph, min feelslikef, min heatidxf, min humidity, min pressure, min tempf, min visibility, min windchillf, min windgustmph, min windspeedmph, mode daytime weathercode, mode daytime weatherdesc, mode weathercode, mode weatherdesc, mildest daytime weathercode, mildest daytime weatherdesc, mildest weathercode, mildest weatherdesc, severest daytime weathercode, severest daytime weatherdesc, severest weathercode, severest weatherdesc, total daytime precipmm, total precipmm, total snowcm

2.5.2. NetHealth Dataset [18]

2.5.2.1. General Overview. The NetHealth dataset[4] , documented in various sources such as [18] and [26], provides a comprehensive longitudinal perspective on university students' behaviors from Fall 2015 to Spring 2019. During the 2015-2017 period, approximately 700 students participated, with 300 continuing into the 2017-2019 period. Data collection involved the use of a smartphone, a Fitbit for health-related data, and surveys for ground truths. This approach was inspired by previous studies such as the NetSense study [69] at the University of Notre Dame and the Friends and Family study [70] at MIT, which also utilized wearable devices and phone agents for data collection. Online surveys were conducted once a semester, with the collection procedure approved through IRB protocols, and each participant providing consent. However, not all data collected is publicly shared due to privacy concerns, despite the dataset offering a comprehensive, multi-year longitudinal view.

Some survey questions were only asked in specific semesters (waves), leading to missing information that restricts the use of ground-truth data from every semester in training algorithms. Data collection involved multiple modalities descirbed as following

- *Communication:* Smartphone-based logs include calls, SMS, WhatsApp, and subcategories like Voice, Facetime, etc., aimed at constructing social networks.
- *Wearable:* Fitbit measurements capture activity and sleep details such as steps, active minutes, heart rate, sleep duration, and waking time. An example of data visualization showing the participant's overall pattern for all semesters can be found in Figure 2.5.
- *Courses and Grades:* Administrative records from ND Registrar's Office reveal course and grade information. These data were collected at the end of the regular data collection procedure in Spring 2019, with consent obtained from only 220 students. The information comprises the academic period, corresponding course number, and final grade.

[4] `http://sites.nd.edu/NetHealth/`

- *Calendar:* A weekly academic calendar featuring class start dates, breaks, holidays, etc., covering 194 weeks.
- *Survey:* There are two main survey types: basic survey and network survey, collected over 8 periods corresponding to summer 2015, winter 2016, summer 2016, fall 2016, spring 2017, fall 2017, spring 2018, and spring 2019. These self-reported questionnaires encompass physical and mental health, social-psychological states, tastes, self-reported behaviors, demographics, and background traits. Specific survey details utilized in each section will be explained within the corresponding sections throughout the thesis document.
- *Network Survey:* Interaction network details include relationship types, closeness, trust, frequency of interaction, relationship duration, and person's information such as sex, age, ethnicity, and educational attainment.

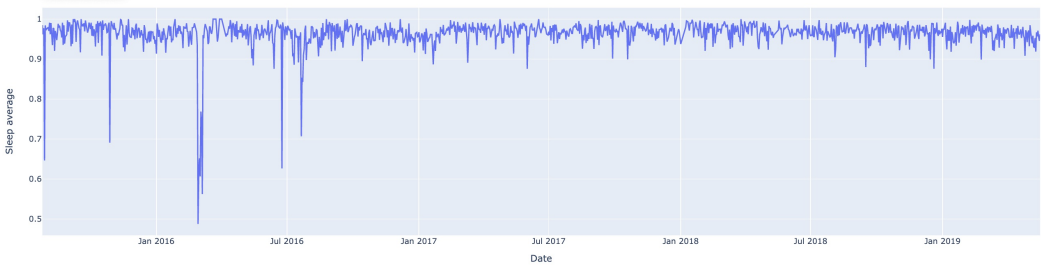

Figure 2.5. Sleep efficiency over all semesters.

The NetHealth study has highlighted unique challenges associated with commercial wearables and data synchronization. Fitbit syncing issues were common, often resulting in data loss and impacting compliance rates. Limited storage capacity in commercial wearables poses additional challenges, necessitating the development of scripts to manage API call limits, data extraction, and synchronization issues. Evaluating the quality of datasets collected from college cohorts is essential for understanding the reliability and validity of the findings. Despite compliance challenges, their dataset collection provides valuable insights into sleep and physical activity patterns among

college students. Evaluating dataset quality involves assessing student compliance, identifying missing data, and analyzing trends and correlations within the dataset.

2.5.2.2. Thesis Related Sub-Dataset. To facilitate our analysis, we first constructed a sub-dataset from NetHealth, aligning with our research objectives. Given the diverse nature of the dataset, encompassing various data types and parameters, we meticulously selected relevant features before starting our thesis.

The NetHealth dataset comprises eight semesters, each with distinct survey questions and responses. Notably, certain semesters lack specific questions; for example, semesters 1, 2, 3, and 7 have no queries related to stress, while semesters 4, 5, 6, and 8 include them. Similarly, sleep ground truth is absent in semesters 5 and 7. To ensure a robust dataset and as an initial analysis, we opted to focus on semester (wave) 1, characterized by relatively higher response rates compared to other semesters in Section 4. On the other hand, for more complex analysis, we handled missing values to utilize all semesters' information in Section 7 for multi-task learning of depression, anxiety and stress.

Throughout the thesis, we used the following sub-datasets of NetHealth: wearables (activity and sleep), surveys (bad habits, big-five personality inventory, education, exercise, health, mental health, origin, personal information, sex, and sleep), courses and grades. However, depending on the target value that we are focusing on in separate studies, we removed the irrelevant ones and the ones in the target from the parameter space. To get the overall picture of the dataset and make the transitions clearer between the sections, please refer to Figure 2.6. Light blue colours are referring different target values in different sections, which are academic grade (Section 4.1), sleep (Section 4.2), and mental health (Section 7) separately. Please note that the academic grade is only utilized in the grade classification section (Section 4.1) For Section 4.2, since we are focusing on sleep parameters from wearables as a target variable, sleep related survey parameters are excluded from the study's parameter space Likewise, in Section 7, since we are focusing on mental health parameters, they are excluded from

the input parameter space Details of the corresponding measured values can be examined in Table 2.3 and 2.4. Additionally, since we are focusing on only one semester's data in Section 4, in the results, some parameters have sub-index (e.g. $_1$) indication, denoting measurement from semester 1. Otherwise, for Section 7, since we are using all semesters' data, there is no such indication.

All experiments in the thesis are performed on macOS, a 2.7GHz quad-core Intel Core i7 processor with 16 GB of 2133 MHz LPDDR3 SDRAM using the Google Colab platform and Jupyter Notebooks with Python 3.10 scripts.

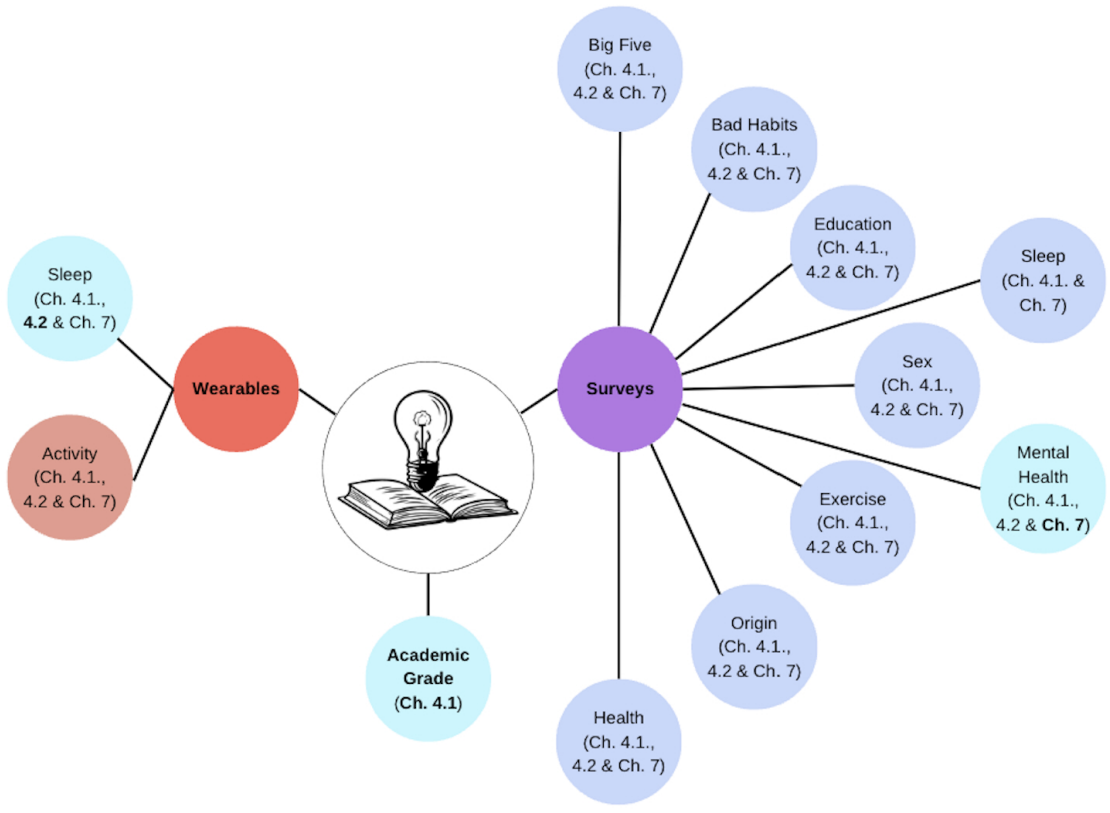

Figure 2.6. Utilization of the NetHealth dataset in various sections (bold ones indicate usage as a target value).

Table 2.3. Details of the features (wearable and course data)

Details of the features (wearable and course data)	
Dataset	**Measured Values**
Wearable data (Activity)	complypercent (percent minutes using Fitbit), meanrate (mean heart rate), sdrate (st. dev. heart rate), steps, floors, sedentaryminutes, lightlyactiveminutes, fairlyactiveminutes, veryactiveminutes, lowrangemins (low range minutes), fatburnmins, cardiomins, peakmins, lowrangecal, fatburncal, cardiocal, peakcal
Wearable data (Sleep)	timetobed (time went to bed), timeoutofbed (time out of bed), bedtimedur (minutes in bed), minstofallasleep (minutes to fall asleep), minsafterwakeup (minutes in bed after waking), minsasleep (minutes asleep), minsawake (minutes awake during sleep period), Efficiency (minsasleep/(minsasleep + minsawake))
Courses and grades	AcademicPeriod, CourseReferenceNumber, FinalGrade

Table 2.4. Details of the features (survey data).

Dataset	**Measured Values**
BigFive/Personal inventory	Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness
Bad habits	usetobacco (used tobacco), usebeer (drank beer), usewine (drank wine or liquor), usedrugs (used rec drugs like marij. or cocaine), usedrugs_prescr (used presc. drugs not prescribed), usecaffine (drank caffenated drinks)
Education	hs (high school type), hssex (high school sex composition), hsgrade (high school average grade), apexams (# of hs ap exams), degreeintent (highest intended degree), hrswork (paid hours senior year), ndfirst (Notre Dame first choice of applied colleges?)
Exercise	hsclubrc (club activities), exercise (excersise), clubsports (play club, intramural or rec sports) , varsitysports (play varsity sports), swimming (swim), Dieting (special type of diet), PhysicalDisability (physical disability)
Health	SelfEsteem (on the whole, I am satisfied with myself), Trust (most people can be trusted), SRQE_Ext (external self-regulation (exercise)), SRQE_Introj (introjective self-regulation (exercise)), SRQE_Ident (identified self-regulation (exercise), SelfEff_exercise_scale (when i am feeling tired), SelfEff_diet_scale (self_efficacy score (diet items)), selfreg_scale (i have trouble making plans to help me reach my goals)
Origin	momdec (is your mother deceased?), momusa (was mother born outside usa?), daddec (is your dad deceased?), dadusa (was your dad born outside usa?), parentstatus (parents living together or divorced/living apart), dadage (father's age), momage (mom's age), numsib (number of siblings), birthorder (which # in birth order are you?), parentincome (parent's total income last year), parenteduc (combined parent education), momrace (mother's race), dadrace (father's race), momrelig (mother's religious preference), dadrelig (father's religious preference), yourelig (your religious prefence)
Personal info	selsa_rom (romantic loneliness), selsa_fam (family loneliness), selsa_soc (social loneliness)
Sex	gender (gender)
Mental health	STAITraitTotal (state_trait anxiety score), CESDOverall (CES depression score), BAIsum (beck anxiety score), STAITraitGroup (state_trait anxiety 2 category), CESDGroup (CES depression - 2 categories), BAIgroup (beck anxiety (3 category)), majorevent (life changes)
Sleep	PSQI_duration (computed time in bed), PSQIGlobal (PSQI total score), PSQIGroup (PSQI two categories), MEQTotal (MEQ (chronotype) score - high score morning person), MEQGroup (MEQ (chronotype) groups - 5 categories))

3. LITERATURE REVIEW ON WELL-BEING PREDICTION

This section explores current studies on mental health and well-being monitoring, focusing on the Tesserae and NetHealth datasets. These datasets are crucial for researchers studying well-being. The main goal is to review studies using these datasets to predict well-being factors, noting whether they use the entire dataset or specific parts. We also identify and compare papers that predict well-being factors using these datasets with this thesis. Secondary studies using other datasets but focusing on mental health and well-being are acknowledged in related work sections of other sections.

Our literature search targeted two main groups: studies from the Tesserae dataset [16] and the NetHealth dataset [18]. We used Google Scholar to find papers linked to these datasets and extended the search using Semantic Scholar [5] and Connected Papers [6] for similar research. This search yielded over 170 papers.

The thesis focuses on three research clusters: workplace well-being, academic performance, and general well-being (including sleep, physical activity, stress, and mental health). Relevant studies are summarized in Table 3.1, with detailed discussion in Section 3.5.

The thesis focuses on three main clusters of research: well-being at workplace, academic grade, and general well-being, which covers sleep, physical activity, stress, and mental health. Relevant studies in literature focusing on these clusters are summarized in Table 3.1, and each cluster are discussed in Section 3.5.

[5] https://www.semanticscholar.org/
[6] https://www.connectedpapers.com/

Table 3.1. Summaries of papers relevant to this thesis using NetHealth and Tesserae datasets.

Topic Cluster	Paper	Dataset	Context - target factor	Method	Objective/output
Workplace & Job Performances	[16]	Tesserae	Longitudinal sensing of workers	Sensor data analysis	Conducted longitudinal sensing
	[71]	Tesserae	Workplace sensor survey	Survey analysis	Examined factors influencing acceptance
	[72]	Tesserae	Workplace behavior analysis	Sensor data analysis	Analyzed behavioral patterns
	[73]	Tesserae	Job performance prediction	Multimodal sensor data analysis	Developed models to predict performance
	[74]	Tesserae	Prediction of human traits	Sensor data analysis	Jointly predicted human traits
	[75]	Tesserae	Employee well-being assessment	Sensor data analysis	Assessed employees' well-being
	[76]	Tesserae	Job performance differentiation	Machine learning	Differentiated between job performers
	[77]	Tesserae	Job promotion detection	Mobile sensing data analysis	Detected job promotions
	[78]	Tesserae	Person-organization fit assessment	Activity routine analysis	Studied person-organization fit
Sleep	[18]	Both	Sleep and physical activity patterns	Fitbit devices, API, data collection scripts	Compliance, sleep/activity patterns
	[79]	NetHealth	Effects of pre-sleep phone use	Fitbit, smartphone app, surveys	Sleep quality
	[80]	NetHealth	Relationship between nightly sleep duration and GPA	Fitbit, questionnaires	Sleep duration, GPA
	[81]	NetHealth	Impact of deviations from normal bedtimes on resting heart rate	Fitbit, statistical models	Resting heart rate
	[82]	Tesserae	Effects of seasons/weather on sleep patterns	Multimodal sensing, linear-effect models	Sleep duration/timing influenced by seasons/weather
	[83]	Tesserae	Effects of snoozing behavior on sleep physiology/personality traits	Surveys, sleep/heart rate trackers	Snooze prevalence, associations
	[84]	Both	Improved sleep detection using wearable/phone data	Data fusion, statistical analysis	Improved sleep detection
	[20]	Tesserae	Sleep patterns/alignment in remote teams during COVID-19	Wearables, surveys	Sleep patterns, teamwork
Physical Activity	[85]	NetHealth	Subjective Well-Being	Classification Algorithms	Prediction accuracy
	[86]	NetHealth	Monitoring Heart Rate Evolution Patterns	Growth-Curve Models	Heart rate changes
Stress	[87]	Tesserae	Alignment between work routines and organizational objectives	Longitudinal sensor data analysis	Identified alignment between work routines and goals
	[88]	Tesserae	Response to Life Events	Fitness Tracker Data	Analyzed physical health changes
	[19]	Tesserae	Stress Prediction	Chaos Theory & Deep Learning	Improved stress prediction
Mental Health	[89]	Tesserae	Mental Well-being	Activity Analysis	Identified gap in mental health
	[90]	Tesserae	Self-esteem	Phone Data Analysis	Predicted self-esteem
	[91]	NetHealth	Subjective Well-being	Fitbit Step Counts	Investigated relationship
This thesis	[26–30]	Both	Student's academic grade, sleep and well-being	Sensor data and surveys	Biomarker identification and validation, ML performances (conventional, time-based, multi-task), Interpretation

3.1. Workplace Well-Being and Job Performance

In today's fast-paced work environments, maintaining employee well-being while optimizing performance has become a crucial focus for organizations worldwide. Nepal et al. [71] surveyed passive sensing technologies, emphasizing their potential to monitor employee behavior and well-being in real-time, despite challenges such as privacy concerns and data biases. This section focuses on workplace well-being and technology studies, highlighting how advancements in sensing technologies, data analytics, and machine learning reshape our understanding of employee health and performance. Since the Tesserae dataset was collected from workers while the NetHealth dataset was from students, this section focused on only studies using the Tesserae dataset.

The Tesserae project, led by Mattingly et al. [16], conducted a comprehensive year-long study on 757 information workers across the U.S., starting in January 2018. Data collection involved Bluetooth beacons tracking movements in both office and home environments, smartwatches monitoring heart rate, physical activity, and sleep, phone agents gathering additional contextual information, and social media analyzing social context and interactions, as explained in Section 2. Furthermore, psychometric surveys were employed to construct ground truth variables, assessing job performance, personality traits, cognitive abilities, affect, and well-being from various perspectives. Notably, the study involves participants from diverse organizational backgrounds, including supervisors, non-supervisors, and various industry sectors, providing a comprehensive understanding of performance differentiators across roles and company types. The analysis commenced with EDA to identify data patterns, trends, and anomalies.

The project faced challenges, including enrollment and retention, managed by offering detailed incentives and compliance-based payments. Privacy concerns were addressed with strict data anonymization and consent protocols. Technological issues, such as remote enrollment logistics and sensor malfunctions, were also resolved.

The Tesserae project demonstrated the feasibility of using multimodal sensors to gain a comprehensive understanding of workplace dynamics. It highlighted the importance of integrating various data sources to effectively assess employee well-being and performance. The insights gained underscored the potential of advanced sensing technologies in organizational research, providing valuable perspectives on how workplace environments impact stress and job performance.

In the pursuit of enhancing job performance predictions, a recent study by Lin et al. [72], has delved into leveraging machine learning algorithms to analyze sensor-collected data and demographic information to predict personality traits. The data from 403 participants of the Tesserae dataset was used due to funding constraints and missing data. Machine learning algorithms, including linear regression, support vector regression, decision trees, random forests, and gradient-boosting regressors, were employed to predict personality traits such as neuroticism, conscientiousness, extraversion, agreeableness, and openness. Performance metrics such as MAPE and MAE were utilized to evaluate model effectiveness. The research yielded insights into the feasibility of predicting personality traits and their impact on job performance using sensor-collected data. Moreover, the comprehensive analysis of health behaviors, psychological constructs, and their influence on job performance underscores the potential for tailored interventions to improve productivity and well-being in organizational settings.

In another study, Das et al. [73] delve into the impact of daily activities and personality traits on workplace performance using the Tesserae dataset from 603 participants to explore the intersection of personality traits and job performance, introducing the concept of organizational personas. The methodology was centered on a person-focused approach, employing clustering algorithms to identify dominant patterns of personality traits and daily activities. The study found that organizational personas could explain workplace performance beyond traditional personality traits, validating the independent association of personality and activities with job performance. However, no interaction effect was found between the two persona patterns, indicating that

while both dimensions are important, they do not interact in ways that significantly alter their impact on performance. Despite the insights gained, the study faced limitations, such as data skewed towards information workers and the passive sensing technology's limitations in capturing the full scope of workers' actual situations. The personas identified were also limited to measurable activity contexts, which may not fully represent comprehensive job performance indicators.

In the realm of occupational well-being, predictive analytics play a pivotal role in assessing job performance and psychological traits. Robles et al. [74] explore the integration of predictive analysis techniques to forecast individuals' job performance, well-being, and psychological traits. By leveraging diverse data sources from the Tesserae dataset, including psychometric surveys, sensors, social media, and daily data streams, this study offers a comprehensive framework for understanding and optimizing workplace dynamics. The research methodology encompasses the development of predictive analysis benchmarks for job performance, well-being, and psychological traits, with a focus on automating the discovery of variable relations and data separability. Techniques such as linear regression, Bayesian ridge regression, support vector regression, decision trees, and random forest regression are employed to model complex relationships among variables.

The research findings highlight the efficacy of predictive analysis in forecasting job performance, psychological constructs, and physical variables. While physical variables are reliably sensed and estimated, predicting psychological constructs such as anxiety, tobacco consumption, interpersonal deviance, and agreeableness poses greater challenges. These insights have profound implications for organizational decision-making, facilitating targeted interventions to enhance workplace productivity, employee satisfaction, and overall well-being. This study underscores the transformative potential of predictive analytics in optimizing workplace performance and well-being.

Expanding on this, Nepal et al. [75] explore how commuting stress influences job performance, productivity, and organizational behavior through mobile sensing

technology. The study used mobile phones and wearable sensors to collect data on commuting patterns and their effects on workplace performance. It involved 275 information workers commuting between home and work locations across the United States over a one-year period, again in the scope of Tesserae project. Data were gathered using GPS logs from participants' smartphones every 2-10 minutes, capturing location-related context, physiology, and behavioral patterns during commutes. The feature extraction process focused on identifying patterns in commuting behavior that could impact job performance. Machine learning models such as ExtraTrees, NuSVR, Adaboost, and a Stacked Model were employed to analyze the data. The results, evaluated using metrics like MAE and PCC, demonstrated the potential of mobile sensing to predict workplace performance based on commuting behavior.

Despite the promising findings, the study faced several limitations. It lacked an assessment of environmental and job-related factors and did not consider participants' subjective feelings about their commutes. The sample size and diversity were limited, and self-reported data could be subjectively biased. The study also focused on between-person differences rather than within-person fluctuations.

In another Tesserae dataset-related study, Mirjafari et al. [76] explore how mobile sensing data can effectively distinguish between higher and lower job performers over a cohort of 554 information workers. By analyzing features such as sleep patterns and heart rate captured through mobile sensors, this study unveils the relationship between behavior and workplace performance, offering revelations for organizational optimization. Machine learning techniques, such as a gradient boosting classifier, are employed to classify workers as higher or lower performers. The classifier achieves an AUROC of 0.83, indicative of robust model performance. The findings underscore the efficacy of mobile sensing data in differentiating between higher and lower job performers, transcending traditional performance evaluation metrics. Furthermore, the objective assessment facilitated by mobile sensing allows for tailored interventions aimed at enhancing workplace productivity and employee well-being.

Furthermore, in literature, Nepal et al. [77] shed light on the objective measurement of behavioural and physiological changes post-promotion through mobile sensing technology. By analyzing data collected from phones, wearables, and Bluetooth beacons, this study shows the impact of job promotion events on information workers' well-being and performance. The research methodology adopts a mobile sensing approach to track the physiological and behavioral patterns of information workers who have been promoted, spanning a one-year period from 141 information workers in the Tesserae dataset. A random convolutions-based approach is employed to extract patterns from multivariate time series signals, facilitating the identification of key features such as activity, heart rate, stress levels, and workplace behavior. Various classification models are evaluated to distinguish between promoted and non-promoted periods, achieving an AUC of 0.72.

The research findings underscore the efficacy of mobile sensing in objectively detecting job promotion events and analyzing subsequent changes in information workers' behavior and physiology. While the study highlights the potential of mobile sensing technology in providing valuable insights into workplace dynamics, limitations such as small sample size, short-term observation period, and reliance on self-reports for job performance metrics are acknowledged. Nevertheless, the development of predictive classification models represents a significant advancement in understanding and addressing the implications of job promotions on employee well-being and performance.

Finally, Das et al. [78] elucidate a compelling connection between routine fit and job performance, underscoring the pivotal role of congruence between individual routines and organizational norms in fostering workplace effectiveness. Through meticulous analysis of empirical data, the researchers discern a robust correlation between routine fit and various indicators of job performance. Firstly, they posit that individuals whose routines align closely with the rhythm and tempo of the organization are better equipped to navigate their professional responsibilities with efficiency and efficacy. By synchronizing their activities with the broader organizational context, employees are poised to capitalize on synergies and minimize friction, thereby optimiz-

ing their contributions to the organizational goals. Moreover, the study reveals that routine fit is associated with higher levels of task performance among employees. Individuals whose routines exhibit a high degree of congruence with organizational norms demonstrate greater proficiency in executing their assigned tasks, leading to tangible improvements in productivity and output. Furthermore, authors observe a negative correlation between routine fit and workplace deviance, suggesting that employees who experience a strong alignment between their routines and organizational expectations are less likely to engage in counterproductive behaviors that undermine organizational cohesion and effectiveness. Importantly, the researchers highlight the multidimensional nature of routine fit, acknowledging that it encompasses not only objective measures of task performance but also subjective indicators such as perceived stress and arousal levels. By incorporating both objective and subjective assessments, they offer a holistic understanding of how routine fit influences job performance outcomes. In summary, researcher's findings underscore the significance of routine fit as a predictor of job performance within organizational settings. By illuminating the intricate interplay between individual routines and workplace effectiveness, the study informs organizational practices aimed at enhancing employee productivity, fostering a culture of engagement, and maximizing organizational performance.

3.2. Academic Grade

Concerning student's academic grade studies and their affecting factors, using NetHealth dataset, there has been only one specific study in literature which focus on sleep and academical grade's relation. Creswell et al. [80] investigate the relationship between nightly sleep duration and college students' grade point averages, emphasizing the importance of adequate sleep for academic success. Their study, drawing from data collected from first-year college students across multiple universities, namely Carnegie Mellon University, University of Washington, Notre Dame University (NetHealth dataset), employs wrist actigraphy to track sleep patterns and their impact on academic performance.

Using actigraphy data from students at three universities, the study found that insufficient sleep predicts lower end-of-term GPA. Each additional hour of nightly sleep was associated with a 0.07 improvement in GPA. The study utilized linear regression models to analyze the relationship between various sleep features and academic performance, emphasizing that consistent, sufficient sleep is crucial for academic success. Despite considering other sleep variables such as bedtime variability and sleep window midpoint, total nightly sleep emerges as the most robust predictor of academic success. By establishing the critical role of nightly sleep duration in academic performance, the study underscores the necessity for students to prioritize sufficient sleep to optimize their learning potential and cognitive functions during their first year of college.

3.3. General Well-Being

3.3.1. Sleep

Sleep plays a vital role in human physiology, influencing many aspects of health and well-being. Recent technological advancements have allowed researchers to delve deeper into understanding sleep patterns and their effects on individuals. This section examines several studies, ranging from the impact of bedtime variations on heart rate to predicting sleep quality and its link to subjective well-being through the use of wearable devices, particularly using the NetHealth and Tesserae datasets which are also utilized in this thesis.

Several studies have employed similar methodologies to measure sleep and physical activity patterns among college cohorts. Purta et al. [18] distributed Fitbit Charge HR devices to 500 college freshmen and utilized phone agents to collect longitudinal data within the NetHealth data collection campaign. The details of the dataset were explained in Section 2.

It is important to consider how pre-sleep habits, particularly phone use, can affect sleep quality. Vhaduri et al. [79] examine the influence of pre-sleep phone usage on

sleep quality among college students, underscoring the importance of limiting phone use before sleep. Their research, based on the NetHealth mobile crowd sensing dataset, included sensor data from smartphones and Fitbit devices, such as phone screen unlock sessions and audio port usage. It aimed to capture pre-sleep phone behaviors and their effects on sleep quality, focusing on patterns among over 400 college students. The methodology involved rigorous statistical analyses, including Two-Sample t-Tests, to examine variations in sleep quality due to different pre-sleep phone use patterns. By analyzing measures like awake count, awake duration, restless count, and restless duration, the researchers revealed the negative impact of bedtime phone usage on sleep quality.

Their findings highlight the detrimental effect of phone use during bedtime on sleep quality compared to pre-bedtime phone use, emphasizing the need for individuals to limit phone usage before sleep for better sleep outcomes. The researchers discovered how using phones before bed hurts sleep quality. This is a common problem in today's world, but one that hasn't been studied much before. Their findings remind us that creating a good sleep environment and routine is important for our health and thinking, especially now that we use electronic devices all the time.

Faust et al. [81] further explore the effects of changes in bedtime regularity on resting heart rate among college students, highlighting the importance of consistent sleep schedules for heart health. Their study, utilizing data from 557 college students collected through Fitbit Charge HRs within the NetHealth dataset, analyzes the impact of bedtime deviations on resting heart rate. They found that even minor deviations in bedtime, such as going to bed 30 minutes later, were associated with significantly higher resting heart rates throughout sleep, persisting into the following day. Additionally, bedtimes of at least 1 hour earlier were linked to higher resting heart rates during sleep, although they normalized by the end of the sleep session and did not extend into the following day. These findings underscore the importance of maintaining proper sleep habits, particularly bedtime regularity, as high variability in bedtimes may have adverse effects on cardiovascular health.

Transitioning from students' perspective to office workers, [82] investigates how seasonal and weather changes affect sleep patterns using the Tesserae dataset. The study delves into the intricate relationship between environmental factors and sleep behaviors by examining 216 individuals across four seasons. The research meticulously controlled for demographic variables and trait-like constructs known to influence sleep behavior. Findings highlight statistically significant effects of seasons and weather on sleep patterns, with the most noticeable effects observed during the spring season. Wake times tended to be earlier, and sleep duration decreased. Additionally, bedtimes and wake times shifted slightly later with increasing outdoor temperatures. This study contributes to determinants of sleep, elucidating the subtle yet impactful influence of seasons and weather on our nightly rest.

As we consider the external environmental factors affecting sleep, it is essential to explore how integrating technology can enhance our understanding of sleep patterns. Transitioning to more specific aspects of sleep behavior, Mattingly et al. [83] delve into the examination of snoozing, focusing on its prevalence, physiological effects, and influencing factors. They define snoozing as the act of waking to multiple alarms and aim to establish it as a construct in need of further exploration. Data from 450 participants were collected using a combination of surveys and wearable devices, providing comprehensive insights into snoozing behavior. The findings reveal that 57% of participants engaged in snoozing, with factors such as gender, age, step count, sleep quality, and chronotype influencing the likelihood of snoozing. Snoozers exhibited lighter sleep before waking and demonstrated an elevated resting HR, suggesting potential physiological implications associated with this waking method. Despite these effects, snoozers did not report feeling sleepier or experiencing shorter sleep durations compared to non-snoozers.

The study employed a logistic regression model to differentiate between snoozers and non-snoozers, analyzing various demographic, psychological, and physiological traits. Results indicated that being female, younger, having fewer steps, lower conscientiousness, and experiencing more disturbed sleep were associated with an increased

likelihood of being a snoozer. Furthermore, the methodology involved the use of wearable devices to track sleep staging and heart rate activity, providing valuable insights into the physiological effects of snoozing. The study highlights the importance of further research into snoozing, emphasizing its potential implications for sleep quality, physiological health, and overall well-being.

Building on the impact of wearables to the sleep detection, Martinez et al. [84] enhance sleep detection accuracy by integrating phone usage data with wearable technology. They focus on improving sleep detection by fusing phone agent and wearable data streams to enhance the accuracy of sleep predictions. Their research paper targets achieving more accurate sleep detection by combining the movement detection capabilities of wearables with the technology usage data detected by cell phones. Data collected from the Tesserae study, involving 757 participants from cognitively demanding professions, showed that combining phone activity data with wearable data significantly improves sleep measurement accuracy.

The fusion of these data streams helps address the limitations of wearables, which can overestimate sleep duration when users are lying still but awake and using their phones. The study aimed to establish a more accurate model for predicting self-reported sleep duration by adjusting the sleep value reported by wearables based on phone activity data. To achieve this, the researchers utilized an app based on StudentLife [68] to collect usage data from both iOS and Android models. They employed a logistic regression model to fuse phone activity with wearable data and generate improved sleep estimates. The findings revealed that combining sensor streams enhances wearable measurement accuracy, particularly in scenarios where wearables may overestimate sleep due to phone usage while lying still. Adjustments to the sleep value reported by wearables were distributed across many individuals for a modest amount, rather than concentrated with large adjustments for a few participants. Additionally, the study identified that Android users, younger participants, and supervisors had more phone agent adjustments on average.

3.3.2. Physical Activity

Physical activity is a cornerstone of human health, significantly affecting numerous aspects of physical and mental well-being. Recent advancements in sensing technology have enabled researchers to gain deeper insights into exercise patterns and their effects on individuals. This section explores various aspects of physical activity, including how external conditions, personal habits, and technological advancements impact exercise frequency, intensity, and outcomes by focusing on the related studies utilizing the NetHealth dataset.

Focusing on the intersection of physical activity, sleep, and mental well-being, in a recent study, Kilic et al. [85] conducted a study aiming to predict the SWB of university students using the NetHealth dataset, more precisely 285 students after preprocessing and cleaning. They used machine learning techniques such as Naive Bayes, KNN, SVM, and Ensemble classifiers to predict SWB levels from sleep and physical activity data collected from sensors. The SVM classifier excelled in terms of accuracy, precision, and F1-Score, achieving a prediction accuracy of 59.9% and a recall of 62.1%. This research highlights the significant potential of utilizing physical activity and sleep data collected through sensors to predict SWB levels. The findings emphasize the critical role that physical activity and sleep play in mental well-being, offering a novel approach to assessing and potentially improving the overall well-being of students through technology-driven insights. The research emphasized the limitations posed by traditional questionnaires in terms of time, cost, and implementation flow, advocating for the utilization of sensor data for SWB prediction.

The exploration of physical activity levels among college students reveals significant insights into the factors that influence daily movement patterns. In a study utilizing Fitbit data from the NetHealth project, Wang et al. [86] analyzed the physical activity of 619 participants over 637 days, highlighting the impact of social factors on activity levels. This extensive dataset enabled the use of LGCMs and HLMs to investigate temporal growth trends and variations in physical activity. Despite limited

overall growth in physical activity levels, notable exceptions were found in daily steps. The study identified that social, environmental, and behavioral factors significantly impact physical activity, with social interactions and peer networks playing a critical role. Interestingly, while males showed a higher susceptibility to peer influence, females were less affected. Additionally, the size of daily social networks emerged as a predictor of increased physical activity levels, underscoring the complex interplay of factors that drive physical activity among college students.

3.3.3. Stress

Stress, a common aspect of modern life, significantly impacts individual well-being. The body's response to negative events involves both psychological and physiological dimensions. Recent research highlights how stressors affect physical health and how coping strategies can alleviate these effects. Wearable technology has quantified these responses, providing insights into stress-induced changes in physical well-being.

HRV has emerged as a potential biomarker for stress, yet its associations in real-life settings remain unclear. Through a longitudinal study involving the Tesserae dataset, researchers investigated the relationship between HRV, measured by fitness trackers, and perceived stress among information workers. Despite observing a weak association, particularly during work hours, [87] cautioned against solely relying on HRV for stress detection, emphasizing the complexities of real-world contexts. Conducted as part of the Tesserae Project, the research involved 657 participants. The methodology included using mixed-effects generalized linear models to predict EMA responses from HRV features, which were calculated over time windows ranging from 5 minutes to 24 hours. Key HRV features analyzed were the low frequency to high-frequency ratio, very low-frequency power, triangular index, and standard deviation of the averages of normal-to-normal intervals. These models were applied to various time frames, with HRV features computed during work hours (8 AM to 6 PM) showing the best performance in explaining perceived stress variance.

Despite the robust data collection and analysis, the study found that HRV alone was a weak predictor of perceived stress in naturalistic settings, explaining only about 1% of the variance in stress levels. HRV was more reflective of stress when specific, isolated stressors were present, highlighting the complexity of using HRV as a standalone measure for stress detection in everyday life. The study also identified several limitations, including sample homogeneity in terms of income and education, and accuracy concerns with self-reported stress durations and timings. A follow-up study with 327 participants was conducted a year later and aimed to refine the link between HRV and stress, focusing on the participants' most stressful time of the day over a 1-week period.

Faust et al. [88] conducted a comprehensive investigation into the effects of negative life events on physical well-being, shedding light on the intricate relationship between stressors and physiological responses. Their study underscores the importance of understanding how individuals' physical health changes in response to adverse life circumstances and the role of coping strategies in mitigating these effects. The research methodology involved techniques such as OCSVM and Bayesian gamma-hurdle regression models. OCSVMs were trained on a pre-event window to establish a physiological and behavioral state (PB-state) baseline, enabling the detection of deviations in participants' physical health and behavior post-event. This approach allowed for the quantification of physical health changes in response to negative life events and the contextualization of these changes with coping strategies. By representing daily resting heart rate, physical activity, and sleep as points in a three-dimensional space using an OCSVM with a RBF kernel, they were able to capture non-linear relationships within the data streams. This model provided a singular metric of distance from the distribution of training data, classifying each day post-event as an inlier or outlier based on the learned decision boundary. Thus, it generated a univariate sequence of values representing the negative life event response curve.

The findings of the study highlight that physical determinants of health indeed deviate in response to negative life events. Moreover, the authors demonstrate that these deviations can be mitigated through various coping strategies, emphasizing the

significance of adaptive coping mechanisms in maintaining overall well-being. However, the study acknowledges certain limitations, including a small sample size due to the lack of survey participation rate (only 330 of 757 completed the survey) that may impact generalizability and a limited observation window post-event for deviations in PB-state. Despite these constraints, the research significantly contributes to the understanding of stress-induced physiological responses and the importance of coping strategies in promoting resilience and well-being in the face of adversity.

In another study, Li et al. [19] propose a novel approach that integrates chaos theory and deep learning for stress prediction, contributing to a deeper understanding of stress dynamics and the development of personalized stress prediction models using Tesserae dataset from 478 participants. This method surpasses conventional techniques, as it outperforms both pure deep learning and chaos methods in stress prediction accuracy. Their study focuses on analyzing sequences of daily stress levels to forecast stress in advance. By leveraging chaos theory, the researchers identify predictable time ranges and unveil underlying patterns within the stress system. Deep learning techniques, enhanced with a two-layer attention mechanism, are then employed to simulate the nonlinear state of stress sequences, leading to more accurate stress predictions. The proposed approach achieves over 2% and 8% better accuracy compared to pure deep learning and chaos methods, respectively.

3.3.4. Mental Health

Understanding mental well-being is a diverse task that necessitates comprehensive analyses of various data streams and measures. In this section, we explore pioneering studies that utilize passive sensing technologies to predict mental well-being, assess self-esteem, and quantify subjective well-being. We begin by examining the Tesserae Project, which highlights the importance of integrating diverse data modalities to understand mental states. Next, we delve into the CampusLife project, where passive sensing is employed to measure self-esteem dimensions among college students. Finally, we analyze the NetHealth study, which investigates the correlation between

physical activity trends and subjective well-being, particularly focusing on weekend activity levels.

Passive sensing, which leverages data from smartphones, wearables, and social media, offers promising avenues for predicting mental well-being. A seminal study utilized the Tesserae Project dataset to analyze psychosocial and behavioral signals alongside self-reported and physiological measures of mental states [89]. By employing a triangulation framework, the study adds rigor, complexity, richness, and depth to the inquiry of predictive analysis of mental well-being using passive sensing. The researchers demonstrated that features with social signals have a smaller semantic gap with self-reported well-being, providing insights for future research in social and ubiquitous computing. This finding underscores the importance of integrating diverse data modalities for a more comprehensive understanding of mental health dynamics and addresses the challenge of predicting mental well-being through passive sensing, highlighting the complexity of mental health assessment.

Their methodology involves comparing different prediction approaches using a data triangulation framework to understand mental well-being constructs. By utilizing psychosocial signals (such as social media language features) and behavioral signals (such as physical activity features), the research investigates the semantic relationship between low-level computer representations and high-level mental well-being constructs. Furthermore, the study emphasizes the importance of evaluating ground truth measures and adopting parsimonious sensing to reduce the semantic gap in predicting mental well-being. Their findings suggest that features with psychosocial signals are better at predicting self-reported anxiety and stress, while features with behavioral signals are better at predicting stressful arousal. This understanding contributes to the ongoing discourse surrounding mental health assessment methodologies and underscores the need for multidisciplinary approaches in passive sensing research.

Self-esteem, a fundamental aspect of mental health and well-being, plays a pivotal role in shaping individuals' perceptions of themselves and their interactions with

the world around them. Passive sensing technologies, particularly through smartphones, can offer promising avenues for understanding and quantifying self-esteem. Researchers [90] present a comprehensive exploration into the feasibility of employing passive sensing techniques to measure self-esteem dimensions. Their study, conducted as part of the CampusLife project [92, 93] during the spring of 2016, comprising a cohort of 51 students. The inclusion of both actively and passively sensed data enriches the dataset to measure self-esteem indicators among college students. The research underscores the significance of passive sensing in elucidating individuals' self-perception and mental well-being.

The methodology involves harnessing smartphone sensors to capture a diverse array of passive data streams, including physical activity, communication patterns, and social interactions. Leveraging a theory-driven approach, the study constructs machine learning classifiers tailored to predict three dimensions of self-esteem: performance, social, and appearance. Key to the study's methodology is the compilation of a groundtruth dataset derived from EMAs, providing daily snapshots of participants' self-esteem across the identified dimensions. Through an analysis employing linear regression, gradient-boosted regression, and deep learning-based multilayer perceptron models, they demonstrate the efficacy of passive sensing in predicting self-esteem components. The findings reveal a high correlation coefficient of 0.60 and a low SMAPE of 7.26%, indicative of the predictive accuracy attained through passive sensing methodologies. In summary, their study contributes significantly to the expanding field of psychological assessment through passive sensing techniques [90].

Quantifying subjective well-being using trends in weekend activity has emerged as a compelling avenue for understanding the relationship between physical activity and mental well-being among college students [91]. The study delves into the intricate dynamics between physical activity trends and subjective well-being, shedding light on the pivotal role of weekend activity levels, as indicated by step counts, in elucidating individuals' mental well-being.

In their investigation, they analyze data sourced from the NetHealth study, encompassing a cohort of 125 college students over two academic years. Leveraging Fitbit Charge HR data, including metrics such as heart rate, active minutes, steps, calories burned, and sleep duration, the researchers delve into how variations in weekend activity relate to subjective well-being traits. They utilize k-means clustering to discern distinct patterns in step counts across a typical week. This approach facilitates the identification of common trends in weekend activity levels among the student cohort. The researchers reveal that higher weekend step counts correlate positively with sociability and subjective well-being traits, thereby hinting at the potential of weekend activity levels as proxies for assessing individual well-being. Despite limitations such as sample size constraints, potential selection bias, and the focus primarily on college students, the study underscores weekend activity levels as a significant potential indicator of subjective well-being. Exploring the interplay between physical activity trends and mental well-being provides insights into using quantified self-data to understand individual well-being dynamics.

3.4. Other Studies Utilizing the Datasets

Some studies use these datasets but focus on the different problems which are not within the scope of this thesis. We can cluster them as

- measuring compliance of the participants to the data collection [94–99],
- personal visualization [100, 101],
- learning based methodologies [102–104],
- imputing missing values [105–108],
- social media & network analysis [109–125].

3.5. Discussion

In this section, related papers in the literature which primarily utilized the same datasets that we use throughout this thesis have been presented. While workplace well-

being and physical activity recognition are not the primary objectives of this thesis, they are included due to their relevance to the core research areas. For instance, personality traits are found to be related to job performance [72], and in another study [73], it is even further validated that there is an independent association between personality and job performance. We internalized the importance of personality traits by including them in our parameter space during the thesis studies. We showed their importance even to other target factors such as sleep, stress, and mental health. Similarly, physical activity is recognized as a factor highly correlated with the well-being aspects investigated in this thesis. Therefore, we provide a detailed discussion of this topic within the context of our research. To maintain focus on the core thesis topics, other potentially relevant applied study areas are only mentioned. It is important to note that the data collection for these datasets spanned extended periods (see Section 2.5). Consequently, each study within this section utilized relevant data subsets collected throughout the data collection timeframe. Thus, specific details regarding the datasets were provided while explaining them.

In Section 4.2, we investigate the performance of CNNs and RF models in predicting sleep quality using data collected from wearable devices. Our findings demonstrate that RF models outperform CNNs in analyzing sleep metrics obtained through wearable technology, consistent with findings by Booth et al. [126]. The reliability of wearable technology in capturing sleep data is well-supported by existing literature, supported by studies such as Kilic et al. [85] and Phan et al. [127]. Furthermore, Sano and Picard [128] highlight the strengths of RF models in handling complex datasets and identifying significant features with minimal preprocessing requirements.Additionally, Arora et al. [129] underscore the effectiveness of CNNs in analyzing sleep quality metrics like Daily Sleep Quality, Weekly Sleep Quality, and Sleep Consistency, achieving 97.30% accuracy in weekly sleep quality classification. Moreover, studies such as Kilic et al. [27] demonstrate the potential of machine learning techniques in accurately predicting sleep quality from wearable device data. Integrating subjective surveys enhances prediction accuracy in health monitoring, as demonstrated by Sathyanarayana et al. [130] and Arora et al. [129]. Adapting models to environmental and behavioral

factors is crucial, although relying solely on wearables may overlook subjective aspects of sleep quality.

We delve into stress biomarker extraction using RF on the Tesserae dataset in Section 5, emphasizing mobile and wearable data integration. We achieve 80% accuracy in classifying five stress levels, highlighting multi-modal feature analysis including activity levels, sleep patterns, and personality traits [28]. In contrast to previous studies, our approach emphasizes the integration of diverse digital biomarkers from wearable and mobile devices. Martinez et al. [87] found a limited correlation between HRV and stress, focusing on physiological responses. Faust et al. [88] explored coping mechanisms in response to negative life events, while Li et al. [19] innovated with chaos-theory-enhanced deep learning for stress prediction. Our study contributes to the literature by leveraging RF for feature selection and stress classification across multi-modal data sources. Unlike Martinez et al.'s focus on HRV alone or Faust et al.'s study on physiological responses post-events, we integrate real-time digital biomarkers for comprehensive stress assessment. This approach not only improves classification accuracy but also enhances the understanding of real-life stress dynamics through wearable technology.

Additionally, Section 6 and Section 7 offer valuable insights by analyzing the Tesserae and NetHealth datasets, respectively. These investigations highlight the importance of incorporating temporal considerations and advanced modelling techniques for improved prediction accuracy in mental well-being factors using digital biomarkers. To the best of our knowledge, none of the well-being studies considered these multiple factors simultaneously on these two datasets. Furthermore, time-based prediction of multiple factors was not considered in previous studies using these datasets. The only study in the literature that utilized both MTL and time-series analysis for mood/well-being prediction is presented in [25]. However, it used another dataset, considered the use of only LSTM, and did not explore the extensive datasets considered in this thesis.

In Section 6, we explore the prediction of stress, anxiety, and positive and negative affect among office workers using wearable devices. Our findings align with the

digital biomarkers with conventional psychological measures, demonstrating that incorporating the time aspect, particularly through LSTM models, significantly enhances predictive performance [28]. LSTM models integrated with wearable data have shown promising results in quantifying well-being factors, highlighting their potential for personalized interventions and public health initiatives [29].

Conversely, in Section 7, using the NetHealth study, which focuses on college students, we predict depression, stress, and anxiety using similar methodologies in Section 6. We emphasize the impact of digital biomarkers on these mental health factors and validate our results against traditional psychological studies. We highlight the effectiveness of multi-tasking techniques alongside temporal modeling, particularly noting the performance improvements achieved with RF and XGBoost algorithms when considering temporal aspects. We further advance mental health prediction using wearable data, emphasizing the importance of comprehensive data-driven approaches for mental health assessment [30].

Our studies align well with previous literature emphasizing the potential of wearable devices in capturing intricate aspects of mental well-being [28], [30]. Furthermore, our findings support the integration of predictive models into practical applications, such as workplace well-being programs and public health initiatives, where early interventions based on real-time data could mitigate stress and anxiety levels [29], [30].

In reviewing related literature on these datasets (as also presented in Table 3.1), several open topics and approaches were identified that were addressed or differentiated in the thesis:

- *Integration of Digital Biomarkers:* Many studies focused on single aspects like HRV or activity levels for stress assessment. The thesis differentiates by integrating diverse digital biomarkers from wearable and mobile devices, providing a more comprehensive view of stress dynamics and mental well-being.
- *Predictive Modeling Techniques:* While some studies applied CNNs or RF individ-

ually for sleep or stress prediction, the thesis explored the comparative effectiveness of these models and extended them into advanced techniques like LSTM and XGBoost. This approach aimed at improving prediction accuracy by considering temporal aspects and diverse datasets (Tesserae, NetHealth).

- *Application in Workplace Well-being:* Existing literature often highlighted the correlation between personality traits and job performance but did not extensively connect these with broader mental well-being factors like sleep and stress. The thesis filled this gap by demonstrating the practical application of predictive models derived from wearable data in workplace well-being programs.
- *Temporal Considerations:* The thesis emphasized the importance of temporal modelling in predicting mental health factors, showing how incorporating time-series data can enhance the accuracy of predictions.
- *Practical Applications:* While some studies discussed theoretical frameworks or preliminary findings, the thesis aimed at practical applications by demonstrating how predictive models based on wearable technology data could be integrated into real-life interventions.

Overall, our thesis differentiated itself by providing a holistic approach that integrated multiple digital biomarkers, employed advanced predictive modelling techniques with temporal considerations, and focused on practical applications in workplace well-being, thereby contributing to a more comprehensive understanding and application of wearable technology in well-being assessment.

4. PREDICTION OF STUDENTS' ACADEMIC GRADES AND SLEEP QUALITY ON THE NETHEALTH DATASET

Gaining insights into the key factors influencing students' academic performance is crucial for developing a more precise methodology to enhance their academic lives. Existing studies often investigate the correlation between students' daily lives and academic achievements, relying on statistical techniques and subjective responses gathered through questionnaires. Over the past decade, the research community has seen a surge in the popularity of wearable devices, such as smartwatches and smartphones, due to their ability to objectively measure users' activity, sleep patterns, and mood states using integrated sensors. This opens avenues for extracting markers linked to individuals' physiological and psychological states.

Furthermore, sleep is another crucial factor in influencing daily performance, overall well-being, and life quality. With the advent of wearable devices, it has become feasible to measure sleep in an unobtrusive manner, moving beyond traditional methods such as camera recordings and image-based state extraction. Wrist-worn devices equipped with accelerometers, heart rate, and heart rate variability sensors enable direct measurements of various sleep-related features, including time to bed, time out of bed, bedtime duration, minutes to fall asleep, and minutes after wake-up.

In this section, we focus on classifying students' academic grades and predicting sleep quality using the NetHealth dataset [18] (Section 2). We address research questions RQ1, RQ5, and RQ6 in Section 4.1 and RQ2, RQ5, and RQ6 in Section 4.2, as outlined in Section 1. Although the dataset spans eight semesters, we use only the first semester's data due to minimal missing values.

4.1. Academic Grade Prediction

Understanding the intricate factors influencing academic performance is crucial for students seeking continuous improvement in their educational journey. Numerous studies have delved into these influences, encompassing diverse aspects such as family history, psychological well-being, and physical activity [18], [131, 132]. While some investigations focused on family history [131] and others on the inclusion of physical activity in the curriculum [132], certain studies considered sleep based on self-reported measures [18]. However, these studies typically adopted a unimodal approach, concentrating on isolated factors without establishing connections between various modalities. A comprehensive understanding necessitates a multi-modal approach to unveil a more holistic view.

This section concentrates on the analysis of multi-modal data obtained from objectively measured wearable device sensors and surveys encompassing subjects' origin, sex, education level, and habits, alongside state-of-the-art assessments of sleep, big five personality traits, and mental health inventories (refer to Table 2.3 and Table 2.4 for details). The primary objective is to explore the factors influencing students' academic performances. Leveraging the NetHealth open-source dataset [131], this section aims to uncover relationships between the aforementioned aspects and academic performance.

The dataset spans various academic periods (semesters) and encompasses diverse survey data. Despite occasional gaps in survey responses, the section focused on a specific period with minimal missing information. Preprocessing involved data imputation techniques to handle missing values, ensuring the dataset's readiness for the final analysis. Machine learning technique, specifically the RF algorithm, was employed for both factor selection and classification. The section identifies essential parameters related to academic performance, such as sleep, big five personalities, health, mental health, personal information, and origin data. These parameters are believed to be instrumental in comprehending the factors affecting students' academic lives and guiding improvements.

A notable contribution of this work is its endeavor to integrate diverse factors, aiming to identify the most influential predictors for students' academic performance by integrating areas like sleep, mental health, and activities within a singular section.

4.1.1. Motivation & Related Work

Examining studies utilizing NetHealth data reveals diverse analyses on topics such as biometric-based authentication [133], physical activity and sleep patterns [131], network analysis [117], [121], and physical activity prediction [134]. This section stands out as a unique exploration that combines multiple focus areas, including sleep, mental health, and activities, in the context of academic performance.

In the area of related works, various studies explore academic performance from educational, psychological, and smartphone sensing perspectives [68], [128], [135, 136]. Notably, the introduction of objectively measured signals sensed from wearables into the research domain related to students' mental health and academic performance began with the StudentLife project [135].

Wang et al. [68] examine the day-to-day and week-by-week impact of workload on various aspects, including stress, sleep, activity, mood, sociability, mental well-being, and academic performance, utilizing smartphone sensors. While revealing strong correlations, this study falls short of considering behavioral differences.

Sano et al. [128] explore the relationship between wearable device sensors, surveys, and students' grades. Using SVM with different kernel setups, the study uncovers intriguing insights, such as the impact of social features and accelerometer sensor data on discriminating higher and lower performers. While akin to our experiment in terms of multi-modal data from wearable sensors and surveys, this study goes further to explore class balancing scenarios. Another study [136] investigates factors related to academic grades, modeling behavior change points to capture individual behaviors. The study identifies positive correlations between section duration and final grades.

Daniels et al. [137] investigate the associations between physical fitness levels, academic performance, and job satisfaction, emphasizing a holistic approach to health and its implications for academic and professional success. Expanding this focus, their research examines the relationships between physical fitness, health behaviors, and occupational outcomes among university affiliates. Conducted with 166 participants between October 2018 and March 2020, the study involved fitness measurements, the use of physical activity devices, and surveys assessing happiness and job satisfaction. The findings revealed a concerning statistic: 45% of participants had poor cardiorespiratory fitness and unhealthy body composition. Furthermore, significant correlations were identified between physical fitness metrics, such as muscular endurance, and important outcomes like GPA and job satisfaction. These results underscore the necessity for universities to enhance physical activity and sleep among their affiliates to promote better health and improve occupational success. The study's recommendations call for comprehensive fitness assessments and targeted interventions to address the identified deficiencies in physical fitness.

4.1.2. Data Preprocessing

We initiated data preprocessing, starting with the elimination of highly correlated features (as discussed in Section 4.1.2.2). The final step involved the application of the Random Forest algorithm for subsequent analyses.

4.1.2.1. Handling Missing Values. Upon completing the dataset preparation for analysis, we identified missing values across several columns. To ensure comprehensive analysis while accommodating these partially missing columns, we opted for imputation. For categorical columns, we utilized the most frequent imputation, while numerical columns employed mean imputation. Notably, activity-related wearable data exhibited sufficient correlation, warranting the use of the KNN imputation technique. Consequently, this method was applied to address missing values in the activity-related wearable data. Importantly, the sleep data from wearables presented no missing values.

4.1.2.2. Correlation. In order to decrease high-dimensionality of the dataset and improve the performance of our research, we looked at the relationship between the parameters. Removal of features with correlations higher than 80% was taken into consideration. *cardiomins, fatburnmins, lowrangemins, minsasleep, minsawake, peakmins* were some of these. Considering that correlations above the threshold suggested redundant insights, the reasoning for this choice was to remove unnecessary data. Following several trials, a cutoff point of 80% was found to be reasonable, leading to the elimination of these six strongly linked characteristics. The original dataset included 93 parameters, but after post-correlation elimination, it was down to 87.

4.1.2.3. Distribution of Target Values: Grades. The distribution, illustrated in Figure 4.1, highlighted the predominance of *A* grades, with notably fewer instances from the *B-, C+, C, C-* classes. Specifically, class distribution included 41856 instances of *A*, 19321 instances of *A-*, 10048 instances of *B+*, 7265 instances of *B*, 2526 instances of *B-*, 1617 instances of *C+*, 1258 instances of *C*, 354 instances of *C-*, and 4346 instances of *S (satisfactory)*. To address potential class imbalance, the SMOTE technique was applied, resulting in a uniform distribution of 41856 instances across all classes. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class instances. This helps to balance the class distribution and improve the performance of classifiers, particularly in scenarios where one class is significantly underrepresented. Without any reformulation of imbalanced data, the model may be biased towards the majority class, leading to poor performance in predicting the minority class.

4.1.3. Model Details and Performance Metrics

In employing a classification methodology, the RF algorithm was chosen since it is an ensemble method, demonstrating superior performance compared to other methods documented in the relevant literature within this domain [64]. The RF algorithm was configured with specific parameters; 1000 estimators, Gini criterion, and maximum

features set to the square root, leveraging the scikit-learn toolkit[7] . For the training and testing phases, a split of 75% and 25% split was performed on the dataset, respectively.

4.1.4. Classification Performance Evaluation

The classification task was conducted on the preprocessed dataset without further alterations, given that the target variable was already categorical. The classification performance is detailed in Table 4.1, providing F1-score specifics for each class and the global average, alongside the accuracy metric. An average accuracy of 76% was achieved, with notable performances for classes *B-*, *C+*, *C-*, *S*. Prior to SMOTE application, the average accuracy was 65%, and lower F1-scores for certain classes were observed. Instances of confusion are illustrated in Figure 4.2, revealing, for instance, that class *A* is most frequently confused with *A+*, which is anticipated due to their proximity. Class *S* is predominantly confused with other classes, aligning with expectations as satisfactory results correspond to passing a course.

The impact of SMOTE extends beyond its function of generating instances based on similarity; it also involves under-sampling. Downsampling to equalize instances across classes resulted in 354 instances for each class, leading to a decreased average accuracy of 47%. This decline is expected due to the removal of a significant portion of data, hindering effective learning.

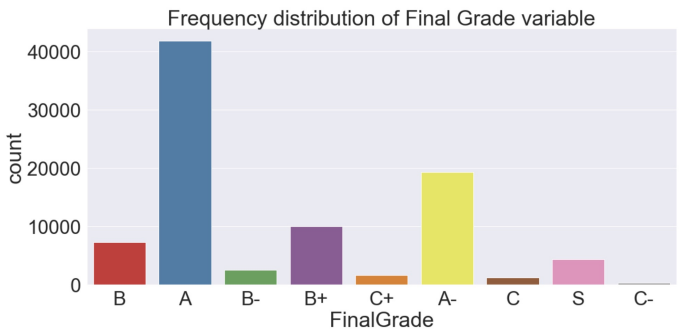

Figure 4.1. Target value distribution: Grade.

[7] https://scikit-learn.org/stable/

Table 4.1. Classification performance details.

	precision	**recall**	**F1-score**	**support**
A	0.53	0.56	0.55	10434
A-	0.52	0.44	0.47	10410
B	0.66	0.71	0.68	10449
B+	0.67	0.60	0.63	10507
B-	0.85	0.87	0.86	10507
C	0.89	0.92	0.91	10494
C+	0.88	0.89	0.88	10418
C-	0.98	0.98	0.98	10502
S	0.78	0.83	0.80	10455
accuracy			0.76	94176
macro avg	0.75	0.75	0.75	94176
weighted avg	0.75	0.76	0.75	94176

Figure 4.3 highlights the most influential factors in achieving the classification performance. The top 20 parameters, determined through RF feature selection, are presented in order of importance. Noteworthy factors include *MEQTotal* (sleep), *Trust* (health), *Extraversion* (big five), *selsa_soc* (personal info), *selsa_rom* (personal info), among others. These parameters, spanning surveys related to sleep, big five, health, mental health, personal information, and origin, contribute significantly to the classification outcome. For detailed explanations of these parameters, refer to Table 2.3 and Table 2.4.

4.1.5. Discussion

In this section, we employed the RF machine learning technique to assess the accuracy of classifying and predicting students' grades based on a combination of survey and wearable data. Additionally, we conducted feature importance analysis to identify the key factors influencing the model's performance. The results highlight the significance of survey parameters related to *sleep, big five, health, mental health, personal information, and origin*, indicating their substantial impact on the classifi-

cation outcomes. Please note that, none of the parameters collected with wearables did not appear in the highly important parameter list which may require further interpretations. The results can be interpreted as personality features and subjective survey responses dominate objectively measured parameters. Notably, our approach distinguishes itself from existing research [68], [128], [136] by incorporating the use of SMOTE.

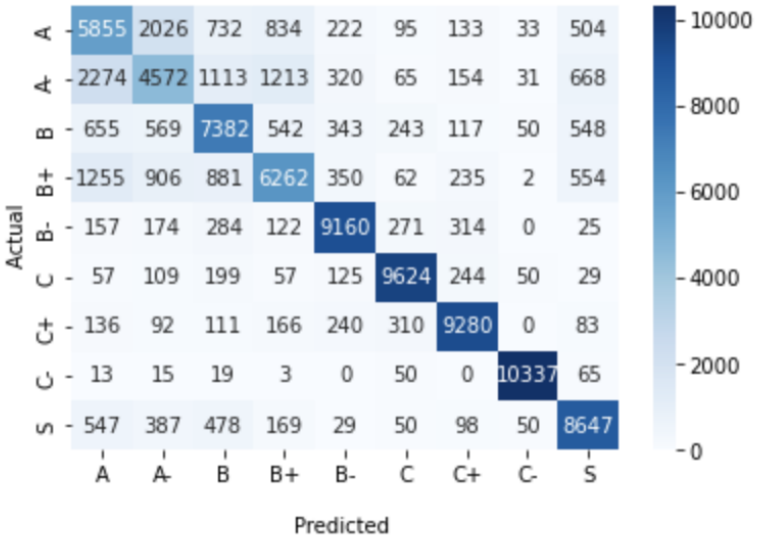

Figure 4.2. Confusion matrix.

The current study utilizes data from a prestigious university, potentially introducing a selection bias towards students with higher academic performance (e.g., prevalence of A and A+ grades). Expanding the sample to include students with a wider range of academic performance could provide more generalizable insights into the factors impacting academic outcomes.

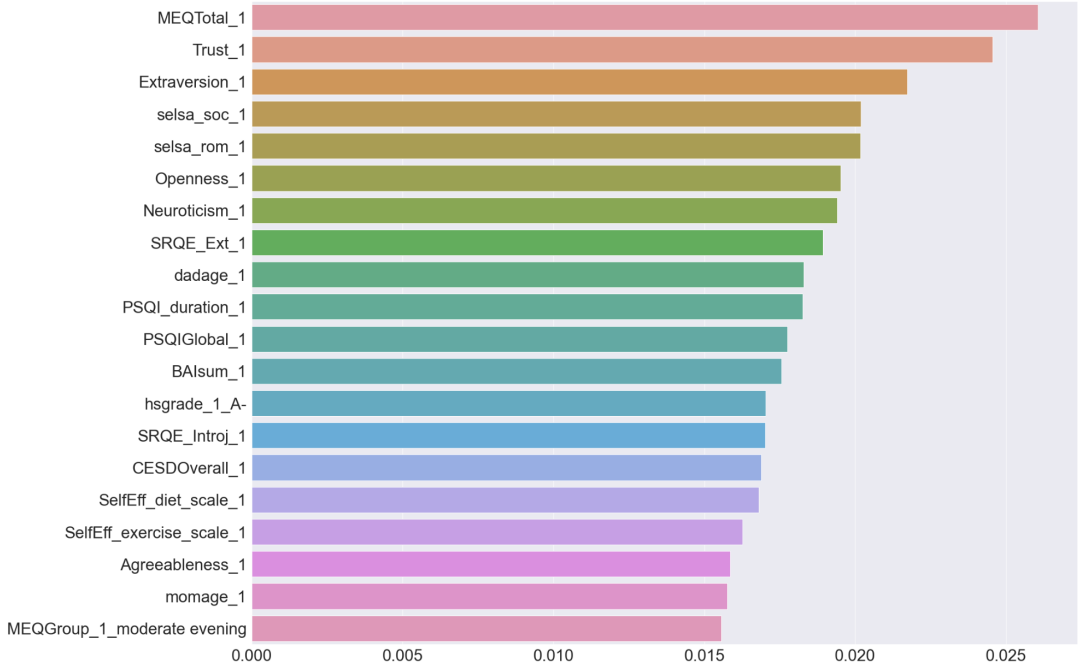

Figure 4.3. Feature importance for classification.

4.2. Sleep Quality Prediction

4.2.1. Motivation & Related Work

Understanding the underlying factors influencing sleep quality is crucial for improving individuals' living standards. With the aid of big data and precise deep-learning models, numerous factors affecting sleep quality metrics can be unveiled. This section endeavours to establish meaningful relationships between sleep measurements and comprehensive multi-modal data, incorporating personal surveys and data from everyday wearable devices. The datasets utilized in this section are derived from the NetHealth open-source dataset [18] as explained in Section 2.

In our exploration, we employ various CNN architectures renowned for their efficacy in human activity recognition [138]. Given that the dataset's measurements are collected daily from wearables, we focus on architectures tailored for daily activity recognition tasks. The intricate details of these architectures and their distinctions are

thoroughly discussed in forthcoming sections. Additionally, we deploy the RF ensemble learning method, not only for performance comparison with deep learning architectures but also for extracting essential features crucial for predicting sleep quality. Despite the generally superior performance of deep learning algorithms, especially CNNs, our findings unveil RF's enhanced performance in this specific context. This observation is consistent with the results in [126], where various deep learning techniques were compared with conventional machine learning methods, establishing RF's superiority. Our findings align with the existing literature exploring sleep quality prediction through wearable devices and questionnaires.

Notable studies such as [18], [85], [127–130], [139], contribute valuable insights to this field. While personal surveys are not universal in these studies, wearable devices emerge as a common denominator. For instance, [85] delves into using sleep and activity metrics to estimate university students' well-being, applying Naive Bayes, KNN, SVM, and Ensemble classifiers to predict "Subjective Well-being" scores based on wearable device data. Similarly, [127] focuses on predicting sleep quality classes, good and poor, using RF, LSTM, GRU, and CNN models, with GRU reported as the best-performing model despite CNN enhancing sleep quality prediction from activities. [128] investigates various target values related to personality traits, wearable device data, and mobile phone data, revealing a correlation between poor sleep quality and high stress levels. Another study [130] classifies sleep quality with deep learning models trained on wearable device data, with CNN achieving the highest accuracy (94.49%) and outperforming traditional logistic regression (64.63%).

In [140], authors proposed a novel approach based on NMF to quantify activity and sleep rhythms from mobile phone data. Through screen-on events, the study identified four distinct temporal components representing morning, night, evening, and noon activities. Despite the lack of clear categorization based on activity weights, the methodology offers valuable insights into individual behavioral patterns. In [141], researchers explored the relationship between chronotypes and social network structures using mobile phone data. Their findings revealed that individuals with "owl" chrono-

types tend to maintain larger personal networks and occupy central positions within social networks. This suggests a strong link between chronotypes and social behavior, emphasizing the need to consider sleep patterns in social network analysis. However, the study doesn't account for confounding factors like demographics or personality traits that could influence the findings.

Other studies have also contributed to sleep research;

- In [142], researchers conducted a genome-wide association study to explore the impact of eveningness chronotype on sleep duration among night shift workers, offering insights into genetic and environmental factors influencing sleep health.
- The study [143] investigated sleep disturbance in elective general surgery patients and evaluated the impact of a sleep quality improvement intervention.
- In [144], authors examined the effects of insufficient sleep and early class start times on cadet health and performance, highlighting the complex interactions between sleep, academic performance, and military training.
- In [145], authors presented SWAPP, a system for automatic sleep monitoring in children, providing tools for diagnosing and treating pediatric sleep disorders.
- The study [146] investigated the effectiveness of the PostOp Pack intervention in improving sleep quality among general surgery patients.
- In [147], authors examined gender disparities in sleep disruption during COVID-19 lockdowns, finding that women experienced more sleep disruption than men, especially among key workers.
- In [148], researchers proposed a framework for active sleep monitoring using wearables, achieving high accuracy in sleep state classification, though acknowledging the need for further validation against clinical sleep monitoring methods.

These studies, collectively, underscore the multifaceted nature of sleep research and the pivotal role of wearable devices and big data in advancing our understanding of sleep quality and its determinants.

4.2.2. Data Preprocessing

The section focuses on predicting sleep efficiency using the Fitbit sleep efficiency score obtained from the wearable device. The Fitbit device employs sensors to discern whether an individual is awake or asleep, providing additional information such as total time in bed, awake time in bed, and sleeping time. The sleep efficiency score is then computed as the ratio of minutes asleep to the sum of minutes asleep and minutes awake, represented by the formula $minsasleep/(minsasleep+minsawake)$. This score, ranging from 0 to 1, denotes the proportion of time spent asleep in bed relative to the total time in bed.

To prepare the data for model evaluation, all individual responses were concatenated based on timestamps and participants. The survey questions were converted into numeric models for analysis. The section deals with missing values in some survey responses across different semesters, specifically concentrating on the first and second semesters' data, excluding participants who have not responded to the questionnaires. The resulting dataset comprises 200,615 rows and 92 columns.

In Figure 2.5, the variation in the target variable, i.e., sleep efficiency, is illustrated over the first and second semesters. This visualization encapsulates the collective patterns of all participants during these two semesters. The observed range of the sleep efficiency score lies between 0.93 and 0.95. Notably, two peaks are discernible, one at the study's commencement in August 2015 and another around September 2015. However, drawing conclusions about the reasons for these peaks is challenging as the specific events surrounding those days are not known.

4.2.3. Model Details and Performance Metrics

In this section, our choice of methodology centers on the application of deep learning techniques, a decision supported by their documented superior performance in the existing literature [138]. Specifically, we narrow our focus to CNN architectures

renowned for their effectiveness in the domain of human activity recognition [149–153]. Given that our dataset's measurements are collected daily from wearables, we opt to tailor our investigation to architectures designed for daily activity recognition tasks.

The selected CNN architectures (Table 4.3) are denoted by C, P, and FC, representing convolution layers, max-pooling layers, and fully connected layers, respectively. Notably, the models exhibit variations primarily in their complexities. Additionally, we deliberately opt for diverse kernel sizes, $3, 5, 12, 20$, which have previously demonstrated efficacy. We maintain consistency across architectures by employing ReLU activation for convolution layers and linear activation for fully connected layers. Filter and pool size parameters remain constant for all architectures at 32 and 2, respectively.

Applying these selected architectures to both scenarios, our objective centers on prediction. To evaluate model performance, we consider MSE and MAE as potential metrics. MAE, calculated by assessing the absolute average distance, is particularly favored in our context for its interpretability. Conversely, MSE involves averaging the squared difference between estimated and actual values. The use of MAE aligns with our aim of enhancing the interpretative aspects of our predictive modelling approach.

4.2.4. Important Factors

In our quest to identify the most influential factors affecting sleep efficiency, we leveraged RF feature importance analysis. The top 20 factors, as depicted in Figure 4.4, predominantly revolve around Fitbit activity measurements. Noteworthy variables include SRQE_introj (introjective self-regulation for exercise), CESDOverall (CES depression score), and Neuroticism (a personality trait) obtained from survey responses. Intriguingly, we also discovered that maternal age has some impact on sleep efficiency.

Subsequently, we conducted a re-evaluation of the aforementioned architectures, this time utilizing only the selected most important factors. Surprisingly, the outcomes revealed higher MAE values compared to our initial approach that incorporated all

available factors. This observation leads us to a significant conclusion: even though certain features may contribute marginally to the final performance, their exclusion results in an overall decrease in predictive performance. Thus, each feature, regardless of its individual contribution, plays a crucial role in maintaining the overall effectiveness of the predictive models in this context.

4.2.5. Discussion and Conclusion

In our investigation, the primary objective was to predict sleep efficiency values obtained from Fitbit devices. We explored and applied prominent CNN architectures, well-established in the human activity recognition domain, alongside the conventional RF model to comprehensively compare the outcomes. Additionally, we employed the feature importance technique to identify the most influential factors in our prediction model (Figure 4.4).

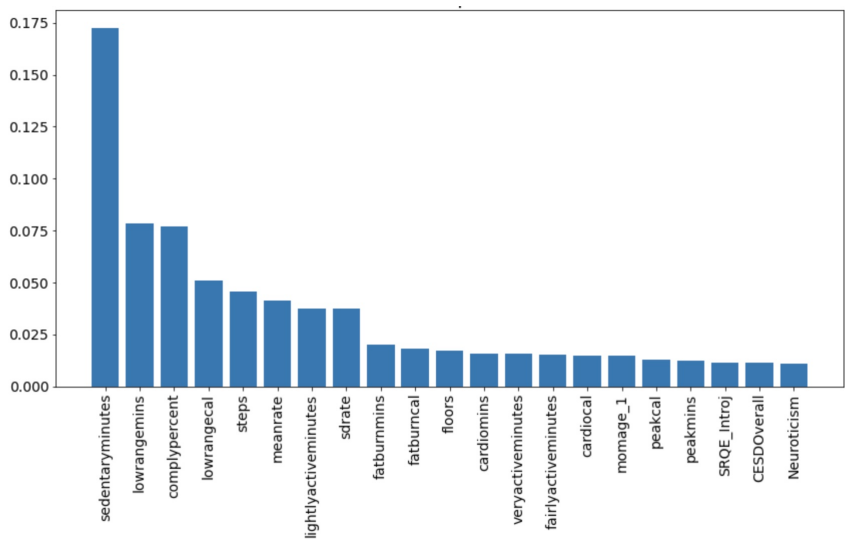

Figure 4.4. Most important factors for sleep quality prediction.

Despite the promising predictions by CNNs, with the C-P-C-P-FC architecture emerging as the most effective, RF exhibited superior performance in our context, albeit

with a marginal difference in terms of MAE (Table 4.2). This subtle performance variation may be attributed to the inherent characteristics of our dataset (a target variable with a relatively limited range between 0.93 and 0.95), aligning with findings in [126], where RF consistently outperformed various deep learning techniques.

Exploring the impact of feature reduction, our experiments demonstrated that deep learning without explicit feature extraction outperforms models incorporating manually selected features. However, the feature ranking derived from this process provides valuable insights for interpretation. Among the top 20 influential factors, sleep efficiency appears to be significantly influenced by wearable activity measurements, complemented by survey-related factors such as exercise, depression, and personality traits. These findings align with existing literature highlighting the interconnectedness of sleep, stress, physical activity, and mood [154–157]. This section uniquely contributes by incorporating these modalities in a multi-modal approach, utilizing deep learning methods on both wearable and survey data collected unobtrusively.

This section presented initial work in exploring the NetHealth dataset. In the following two sections, more comprehensive analyses are conducted using a better-labeled dataset on the Tesserae dataset. In Section 7, a more detailed analysis of well-being will be performed by revisiting the NetHealth dataset.

Table 4.2. Sleep quality prediction results.

Architecture	**MAE**
C-P-FC-FC	0.0876
C-P-C-P-FC	**0.0357**
C-P-C-P-FC	0.0429
C-C-P-C-C-P-FC	0.0844
C-P-C-P-FC-FC	0.0679
C-P-C-P-C-P	0.9361
Random Forest	**0.0282**

Table 4.3. Details of model parameters.

Architecture	Layer	Activation	Kernel Size	Pool Size	Filters	Input Size	Output Size
C-P-FC-FC	1D Conv.	ReLU	20	-	32	(93, 1)	(74, 32)
	Pooling	-	-	2	-	-	(37, 32)
	Dense	Linear	-	-	-	(93, 1)	(37, 16)
	Dense	Linear	-	-	-	-	(37, 16)
	Flatten	-	-	-	-	-	592
C-P-C-P-FC	1D Conv.	ReLU	3	-	32	(93, 1)	(91, 32)
	Pooling	-	-	2	-	-	(45, 32)
	1D Conv.	ReLU	3	-	32	-	(43, 32)
	Pooling	-	-	2	-	-	(21, 32)
	Dense	Linear	-	-	-	(50, 1)	(21, 16)
	Flatten	-	-	-	-	-	336
C-P-C-P-FC	1D Conv.	ReLU	5	-	32	(93, 1)	(89, 32)
	Pooling	-	-	2	-	-	(44, 32)
	1D Conv.	ReLU	5	-	32	-	(40, 32)
	Pooling	-	-	2	-	-	(20, 32)
	Dense	Linear	-	-	-	(50, 1)	(20, 16)
	Flatten	-	-	-	-	-	320
C-C-P-C-C-P-FC	1D Conv.	ReLU	5	-	32	(93, 1)	(89, 32)
	1D Conv.	ReLU	5	-	32	-	(85, 32)
	Pooling	-	-	2	-	-	(42, 32)
	1D Conv.	ReLU	5	-	32	-	(38, 32)
	1D Conv.	ReLU	5	-	32	-	(34, 32)
	Pooling	-	-	2	-	-	(17, 32)
	Dense	Linear	-	-	-	(93, 1)	(17, 16)
	Flatten	-	-	-	-	-	272
C-P-C-P-FC-FC	1D Conv.	ReLU	5	-	32	(93, 1)	(89, 32)
	Pooling	-	-	2	-	-	(44, 32)
	1D Conv.	ReLU	5	-	32	-	(40, 32)
	Pooling	-	-	2	-	-	(20, 32)
	Dense	Linear	-	-	-	(93, 1)	(20, 16)
	Dense	Linear	-	-	-	-	(20, 16)
	Flatten	-	-	-	-	-	320
C-P-C-P-C-P	1D Conv.	ReLU	12	-	32	(93, 1)	(82, 32)
	Pooling	-	-	2	-	-	(41, 32)
	1D Conv.	ReLU	12	-	32	-	(30, 32)
	Pooling	-	-	2	-	-	(15, 32)
	1D Conv.	ReLU	12	-	32	-	(4, 32)
	Pooling	-	-	2	-	-	(2, 32)
	Flatten	-	-	-	-	-	64

5. CLASSIFICATION OF STRESS LEVELS AMONG OFFICE WORKERS

5.1. Motivation

Clinical digital biomarkers are objectively measured and evaluated indicators of biological processes [158, 159]. These digital biomarkers, along with clinical outcomes, aim to assist in understanding the factors affecting human health.

With the advent of wearable devices such as smartwatches, sleep sensors, and step counters, it has become possible to measure health parameters like daily activity, sleep quality, and stress levels in an ambulatory and continuous manner. This digitalization of healthcare measurements has led to the term *digital biomarkers*, which are objective, quantifiable physiological and behavioral measures collected from wearables and smartphones [8]. Indicators or characteristics for a digital biomarker can be computed from one or more digital health technologies.

There are different categories of digital biomarkers in the literature [8], such as *diagnostic digital biomarkers* for detecting diseases or conditions [39], *pharmacodynamic response digital biomarkers* for measuring changes in response to pharmacological agents [38], and *monitoring digital biomarkers* for tracking specific medical conditions [40]. These categories correspond to three main types of physiological outcomes, which can be measured via sensors. Additionally, outcomes for *pharmacodynamic response digital biomarkers* can be assessed using questionnaires. This thesis falls into the monitoring digital biomarkers category since data is collected during participants' daily lives.

This section investigates the extraction of novel digital biomarkers related to stress using data from mobile phones, wearable devices and surveys. The aim is to predict participants' stress levels and identify the most important features contributing

to these predictions. The research ranks features such as sleep and activity levels, demonstrating their significance in classifying stress levels. Using daily stress responses as labels, the study classifies stress into five levels with the RF algorithm and addresses class imbalance using SMOTE. The results show that the inclusion of personality-related parameters improved overall classification accuracy. Sleep and phone activity emerged as the most crucial features for stress prediction. This study is the first to extract stress-related digital biomarkers from the Tesserae dataset and includes a causality analysis for the best-performing case to compare the biomarker rankings.

In this section, we examine Tesserae dataset [68] (Section 2), and seek to address the following research questions; RQ3, RQ5, RQ6 as stated in Section 1.

5.2. Related Studies

Many related works exist in the literature about digital biomarkers for various physiological and psychological states [67, 160–163]. Various machine learning techniques are also employed in these studies. In [163], wearable device measurements were analyzed to determine self-reported stress and mental health status using SVM models with different configurations. In addition to digital biomarkers, there are many studies on stress detection and monitoring [64, 68, 164–166].

Akbar et al. [164] utilized wearable sensors to measure physician stress during EHR inbox work. Their findings identify factors such as EHR window switching and inbox duration associated with stress, offering insights into the impact of workplace technologies on well-being. Stress patterns, including peaks in the first hour of work and evenings, were observed, leading to recommendations for reducing EHR window switching and managing work hours. However, the study's reliance on wearable sensors may introduce biases due to participant awareness of being monitored, and its limited sample size may restrict generalizability.

In [165], features from multi-modal smartphone sensors were examined to classify three stress levels (low, mid, high) using combinations of machine learning techniques such as semi-supervised learning, ensemble methods, and transfer learning.

In [166], researchers combined personal stress data with surrounding data to improve stress detection accuracy. They achieved the highest performance by combining personal and surrounding data, 79.16% on their data and 81.79% on the StudentLife dataset [68].

Stress-related studies also vary by data collection environment, such as controlled laboratory settings and daily (unobtrusive) environments. In [64], authors reviewed state-of-the-art methods using different signal types in various environments, emphasizing the use of sensors like accelerometers, body temperature, EDA, HR, HRV, and speech, along with high-level features from mobile phone usage and physical activity.

Our work presents several novel contributions that distinguish it from existing literature. Unlike previous studies, we focus on extracting new digital biomarkers related to stress using various parameters and modalities, including mobile phones and wearable devices. We aim to predict participants' stress levels and identify the most significant digital biomarkers by ranking features to determine which modalities, such as sleep and activity levels, are most crucial in classifying stress levels. Utilizing the RF algorithm, we classified five stress levels from low to high and compared the results with the usage of all parameters. Additionally, our study uniquely investigates the inclusion of personality-related parameters, specifically the Big Five personality traits, in the classification model. To address class imbalance, we applied SMOTE, achieving the best classification result of 80% overall class accuracy by incorporating personality-related parameters and resolving the class imbalance issue. Notably, this is the first study to extract digital biomarkers related to stress from the Tesserae dataset, further setting our work apart from prior research.

5.3. Data Preprocessing

5.3.1. Details of the Stress Classes

We selected stress as our dependent variable in the scope of this section, and we examined its distribution in detail. In Figure 5.1, one can see that there is a class imbalance. We have a few instances from some classes, especially from $4th$ and $5th$ classes. The model may fail to recognize them due to insufficient examples. To solve this issue, we applied SMOTE technique. After the implementation, we got 57150 rows for both our experiments which are explained in detail in Section 5.3.2, and column numbers remained the same. Here, classes 1 and 5 correspond to very low stress and very high-stress levels, respectively. To show the effect on the model performance, we will provide model results with class imbalance along with performance after applying SMOTE in Section 5.5.

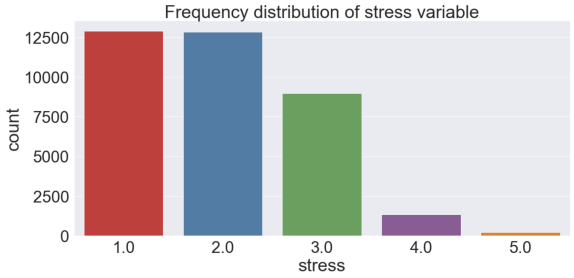

Figure 5.1. Distribution of stress classes.

5.3.2. Data Preprocessing

We deleted irrelevant attributes for our analysis, such as *survey name*, *survey sent time*, *survey start time*, and *survey end time*. We selected the stress variable among ground truth variables as a target variable. Nine missing values were in the stress column over all participants' data, and we deleted rows corresponding to them rather than applying an imputation method for simplicity.

We also examined missing values among the independent variables. Some attributes were never collected from some participants. Thus, the imputation methodology could not be applied for these attributes. Among them, if there is a lack of many participants' data, their removal by row leads to a high decrease in our dataset. Instead, we deleted those columns. These are *act (activity) still*, *light mean*, *garmin hr (heart rate) min*, *garmin hr max*, *garmin hr median*, *garmin hr mean*, *garmin hr std*, *ave (average) hr at work*, *ave hr at desk*, *ave hr at desk*, *ave hr not at work*, *call in num (number)*, *call in duration*, *call out num*, *call out duration*, *call miss num*, and their derivatives according to time episodes. However, when we have data columns with missing values at the person level, we imputed them by applying the mean operation for each participant. After this preprocessing step, we got 31772 rows and 191 columns.

In order to see the impact of personality attributes in our prediction analysis, we designed two experiments. In the first experiment, the ground truth attributes other than the stress level; extraversion, agreeableness, conscientiousness, neuroticism, openness, anxiety, sleep, positive affect, and negative affect were deleted from the constructed data file. Since we did not include personality-related columns from big five personality survey, we named this dataset as *without personality dataset*. Thus, we have 31772 rows and 185 columns. In the second experiment, we included these parameters in our feature space for the analysis rather than removing them as in the first experiment. Since we have personality information in the constructed data file, it is named as *with personality dataset*. Please note that personality related-attributes in this section differs from our published conference paper [28] where we included sleep, anxiety, positive and negative affect attributes. In the conference paper, we showed these parameters contribution as well. However, to be consistent in the scope of this thesis, we provide results when we only employ Big-five parameters as personality attributes. This reduces the column number from 195 to 191 due to the removal of the indicated four columns compared to the conference paper [28]. In the end, we have 31772 rows and 191 columns where the last column is the stress level, our target variable.

5.4. Classifier and Validation

As a classification method, we used RF algorithm in our analyses because it is an ensemble method and perform better among the other used methods in literature in this domain [64]. We applied hyper-parameter optimization with quite a large space and found that in scikit-learn n_estimators 1000, criterion *gini*, min_sample_split 2, min_samples_leaf 1, max_features *sqrt* combination reveal the best performance. Thus, we continued with that setup. In addition, we ran RF with different train/test split sizes. As we increased train size, we got higher accuracy scores. However, to overcome over-fitting, 80% - 20% train and test size are chosen, respectively.

5.5. Results

5.5.1. Results on Without Personality Dataset

In this section, we work on data with 185 columns where there are no personality information-related columns and other attributes such as anxiety, sleep, and positive and negative affect. First, we present the results of the imbalanced stress class. We provide the most important feature ranking, i.e., digital biomarkers and corresponding classification results. Then, we list important features and classification results when we solve imbalance via SMOTE application.

5.5.1.1. With Class Imbalance.

- *Most important digital biomarkers*

  We started with 184 independent digital biomarkers. We want to select only the important parameters within the parameter space that greatly impact the target, i.e., stress. Therefore, we applied RF feature selection to extract the most important digital biomarkers. The most important 20 features are shown in Figure 5.2. To see the effect of modalities, we rank these digital biomarkers according to their corresponding modalities by assigning an importance value,

e.g., 20 for the most important one and subtracting by one for the up-comings. Then, we sum and obtain modality ranking in Figure 5.3. There, we see that the most important features come from phone activity, sleep, HR, activity, and daily modalities calculated according to Figure 5.3.

In Figure 5.5, 5.6, 5.7, 5.8, we see significant changes between stress classes for each type of marker. The amount of the effects are different and can be seen in Figure 5.2. For instance, the most important marker is wake-up time according to the feature ranking. When we examine its relation with stress classes in Figure 5.5, we notice that the wake-up time interval becomes narrow concerning the increase in the stress level. Thus, it is found that the feature ranking affects the stress level. Please note that *stress* feature, from Garmin based on heart rate variability measures (the amount of time between consecutive heartbeats), has an important effect on the target stress class. It can be interpreted that the stress measurements from the Garmin confirm the subjective stress score collected as ground truth.

- *Classification*

  In Table 5.1, we present the classification performance score details for *without personality dataset* in original form. To better explain the results, we shared the first 20 important ones in Figure 5.2.

  We applied RF classification. The detailed classification results are in Table 5.1. As we have a class imbalance, we obtained very low F1-scores for the $4th$ and $5th$ classes, which present the most highly stressful classes. However, for these classes, we have high precision and low recall. It can be interpreted as correctly predicted labels; however, actual instances belonging to these classes are misclassified. We observed lower F1-score even in classes with high instances ($1st$, $2nd$); 0.68 and 0.60 respectively. In addition, we provide a confusion matrix in Figure 5.9. Class 1 is mostly confused with the $2nd$ class. In the confusion matrix, the percentages are distributed over each pair. Thus, the sum of all percentages corresponds to 100%.

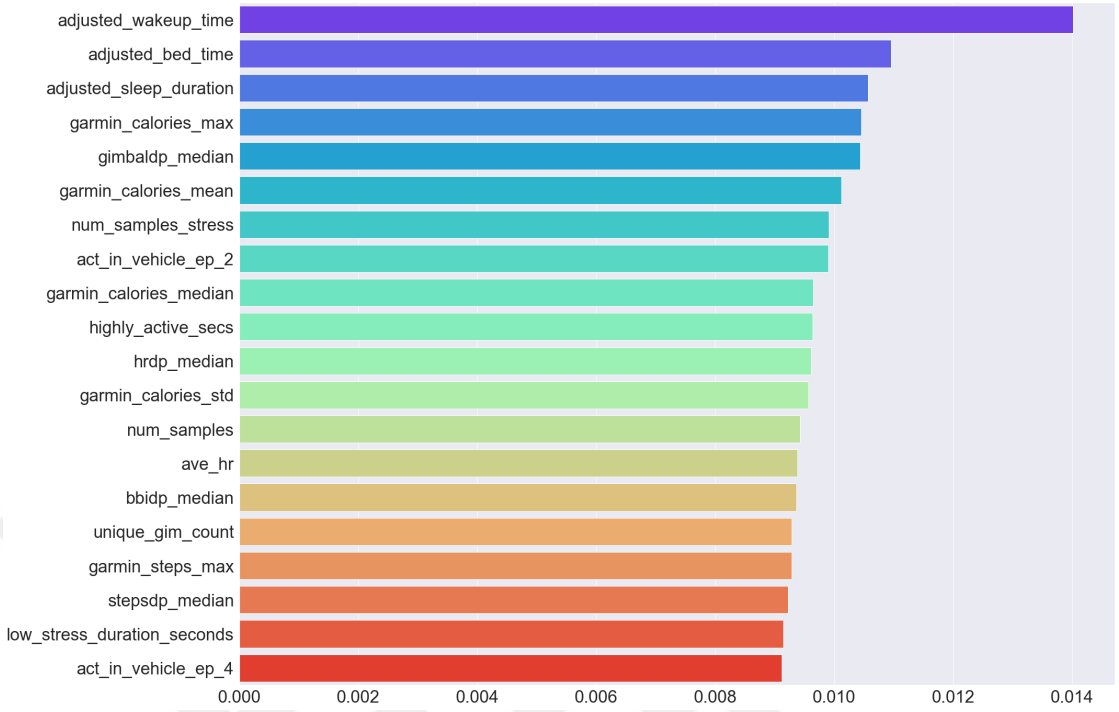

Figure 5.2. Feature ranking with class imbalance without personality.

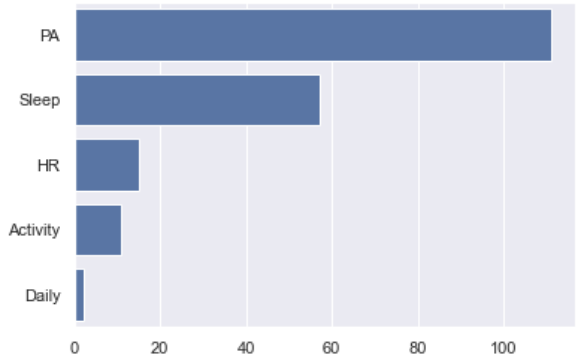

Figure 5.3. With class imbalance.

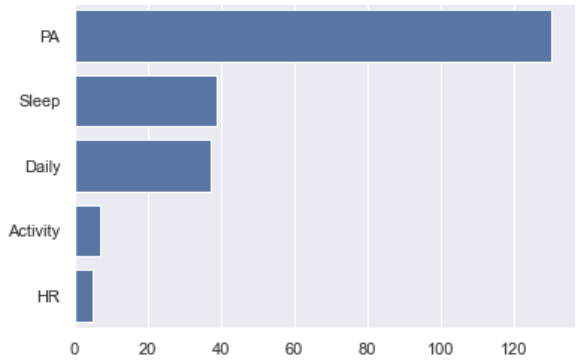

Figure 5.4. After SMOTE application.

Table 5.1. Without personality with class imbalance.

	precision	**recall**	**F1-score**	**support**
1	0.65	0.71	0.68	2244
2	0.52	0.72	0.60	2205
3	0.70	0.35	0.46	1611
4	0.97	0.12	0.21	251
5	0.92	0.25	0.39	44
accuracy			0.60	6355
macro avg	0.75	0.43	0.47	6355
weighted avg	0.63	0.60	0.58	6355

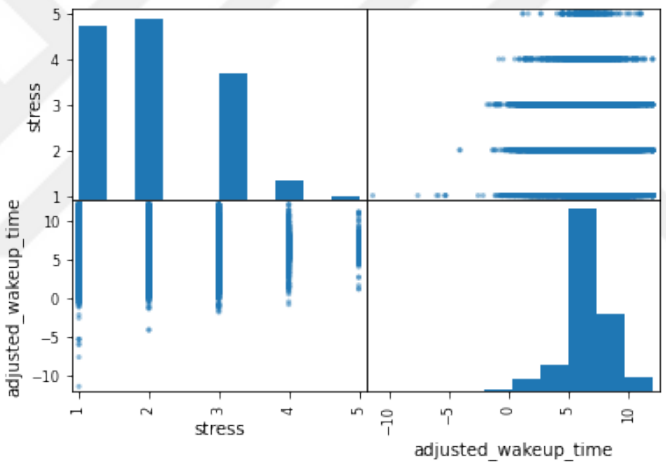

Figure 5.5. Wake-up time vs stress.

5.5.1.2. Performance after Applying SMOTE. To resolve the class imbalance, we applied the SMOTE technique. The idea is to produce synthetic data by keeping the same distribution of each class. Before applying SMOTE, the number of instances for each stress class was 10991, 11430, 7935, 1197, and 219 for classes 1 to 5 respectively; after SMOTE, we have 11430 instances for each class. Thus, the number of rows increases to 57150 from 31772.

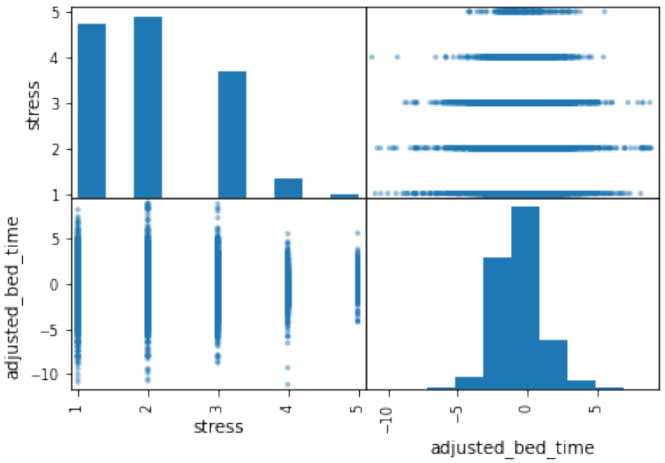

Figure 5.6. Bed time vs stress.

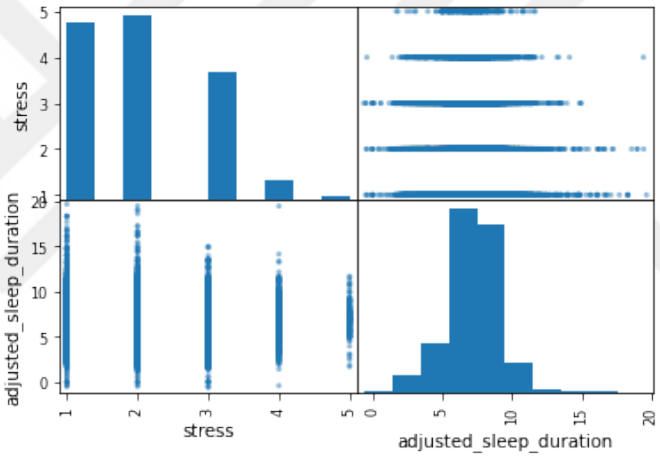

Figure 5.7. Sleep duration vs stress.

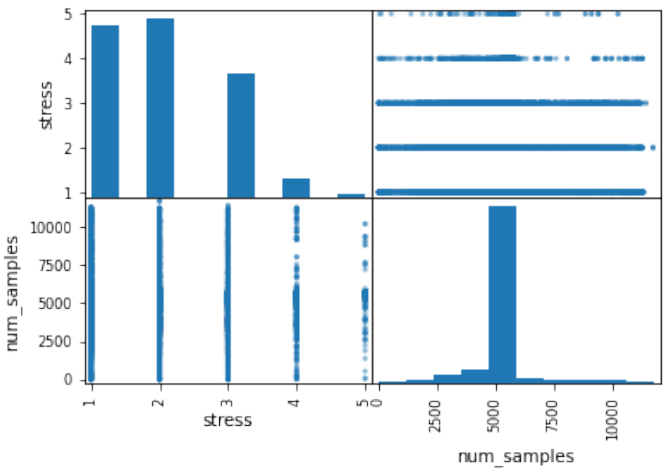

Figure 5.8. HR samples vs stress.

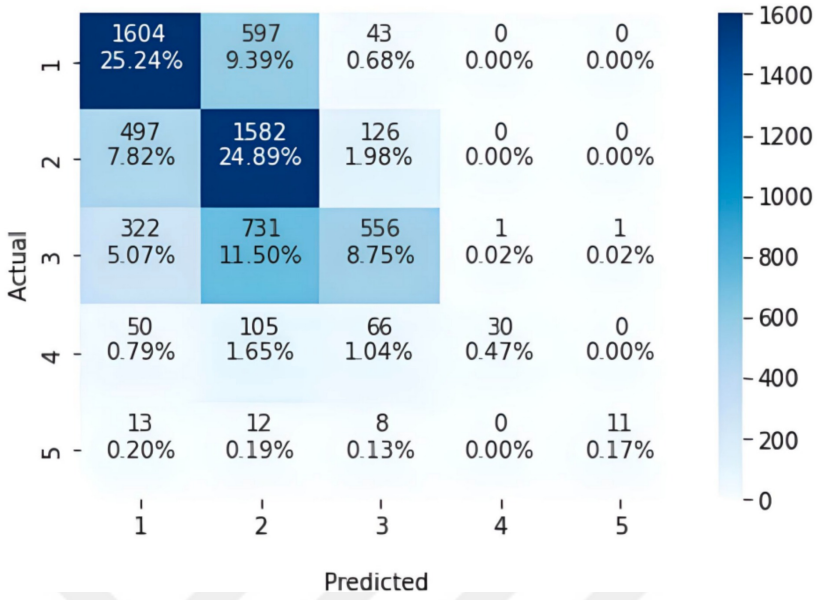

Figure 5.9. Confusion matrix without personality with class imbalance.

- *Most important digital biomarkers*

  Similarly, we applied RF feature selection and extracted most important 20 digital biomarkers in Figure 5.10. Even though the features selected in the imbalanced version are different, as we saw in Figure 5.4 that the ranking of the modalities is the same for the first ones, which are phone activity and sleep modalities. However, there is a slight change in the ranking of daily, activity, and HR modalities.

- *Classification*

  Classification result and confusion matrix can be seen in Table 5.2 and Figure 5.11. We observe the positive effect of balancing the classes. Overall class accuracy performance increased from 0.60 to 0.77. However, when we examine it in detail, we notice that this improvement comes from the $4th$ and $5th$ classes which have lower instances than other classes. The performance of the remaining classes was approximately the same except for the $3rd$ class, which has slightly increased.

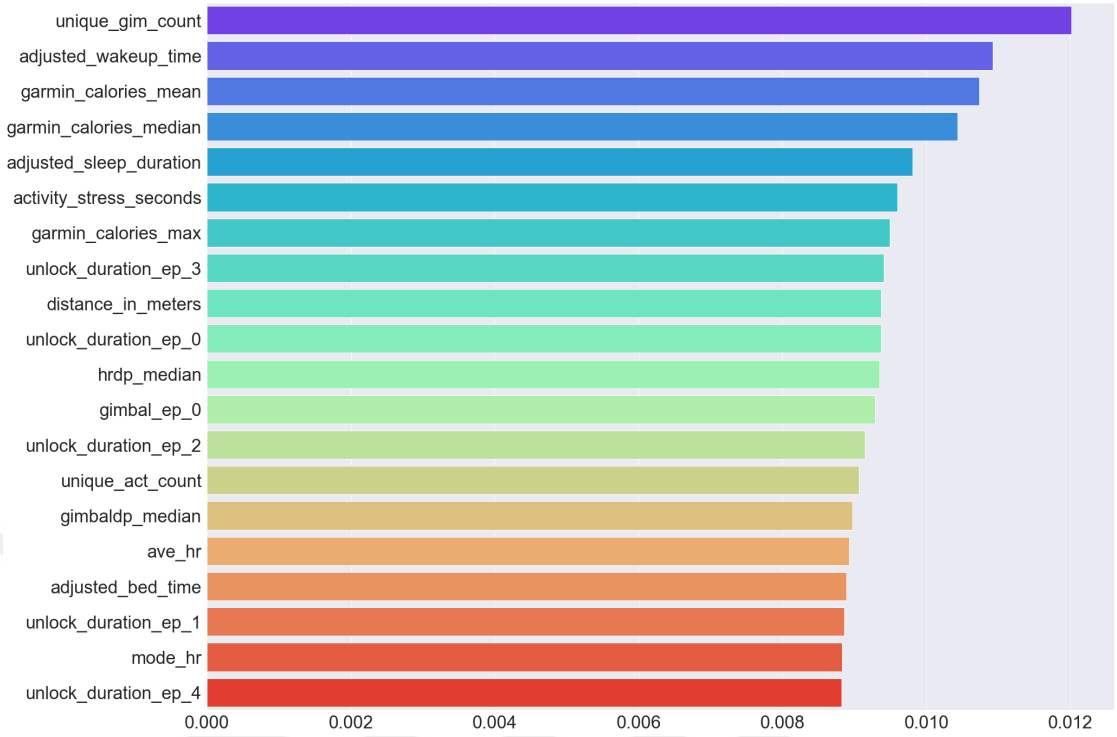

Figure 5.10. Feature ranking after SMOTE without personality.

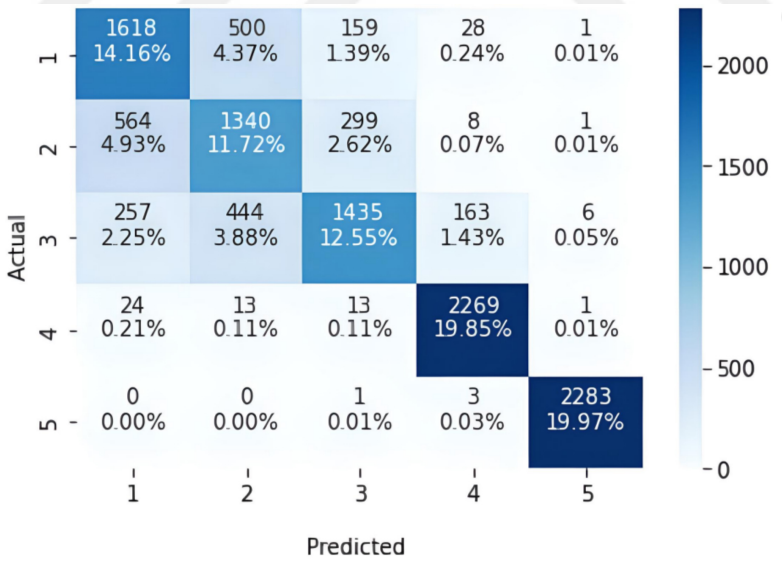

Figure 5.11. Confusion matrix without personality balanced data.

5.5.2. With Personality Dataset Results

In this experiment, we incorporate personality parameters into our analysis. We repeat the process with 194 independent variables, targeting the stress variable.

Table 5.2. Without personality after SMOTE.

	precision	**recall**	**F1-score**	**support**
1	0.64	0.69	0.66	2306
2	0.56	0.59	0.58	2212
3	0.74	0.60	0.66	2305
4	0.91	0.97	0.94	2320
5	1.00	1.00	1.00	2287
accuracy			0.77	11430
macro avg	0.77	0.77	0.77	11430
weighted avg	0.77	0.77	0.77	11430

5.5.2.1. With Class Imbalance.

- *Digital biomarkers*

  Since we added new parameters after feature selection, we observed them in the new feature ranking in Figure 5.12. We observed 7 over 10 recently added parameters among the most important digital biomarkers. The two overwhelming important digital biomarkers are neuroticism and wake-up time over stress.

  The general ranking of the modalities is shown in Figure 5.13. We observe that sleep and phone activity modalities keep their places of importance, and recently added features occur in our ranking. In Figure 5.15, 5.16, 5.17, 5.18, we share the *anxiety, agreeableness, consciousness, neuroticism* versus stress distribution. There are slightly visible effects between stress classes, especially for the distinction of the $5th$ class from the rest. For instance, we can see an effect where the range of these features become narrow when stress level increase.

- *Classification*

  In Table 5.3, we observe that even with a class imbalance on the target variable, we got higher F1-scores compared to the *without personality* dataset. Thus, we can say that these parameters improve the overall result accuracy from 0.60 to 0.62. The detailed confusion matrix can be seen in Figure 5.19.

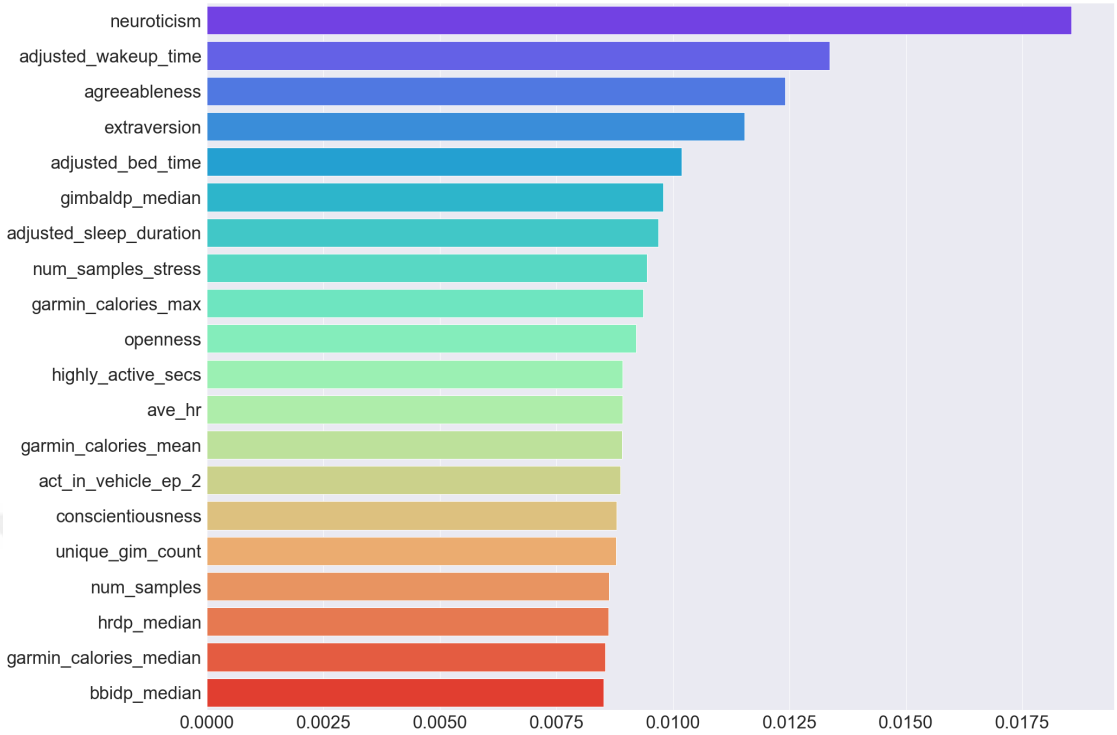

Figure 5.12. Feature ranking with class imbalance with personality.

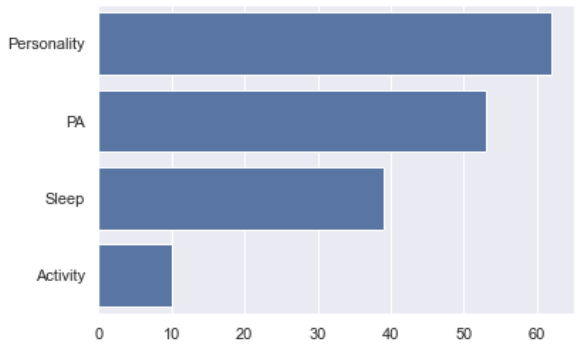

Figure 5.13. With class imbalance.

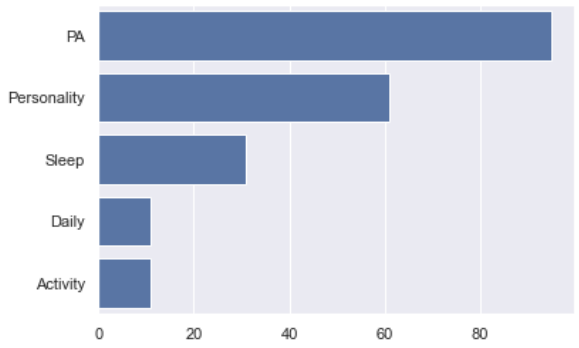

Figure 5.14. After SMOTE application.

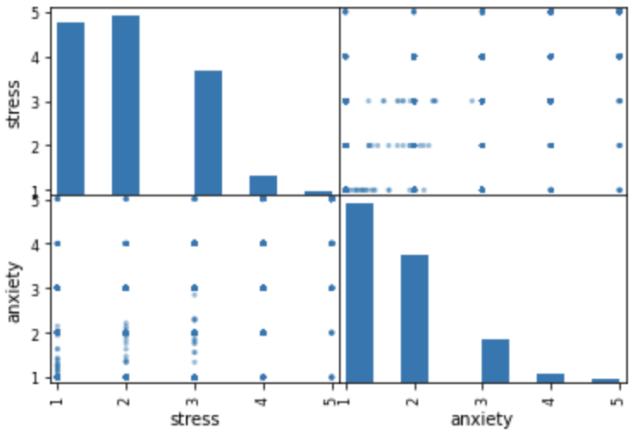

Figure 5.15. Anxiety vs stress.

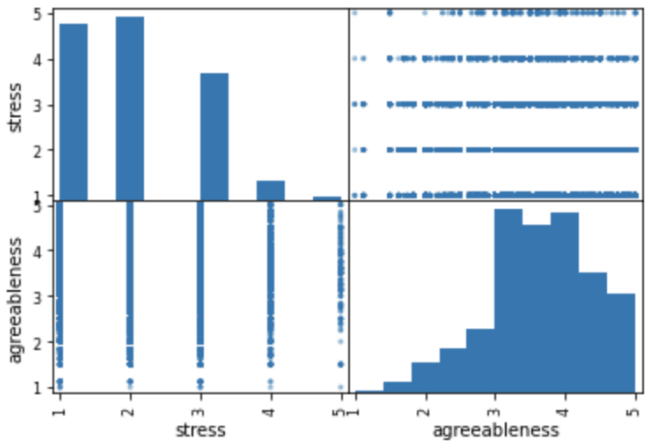

Figure 5.16. Agreeableness vs stress.

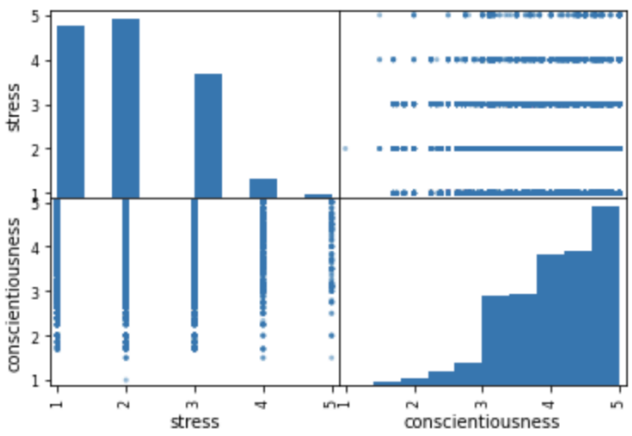

Figure 5.17. Conscientiousness vs stress.

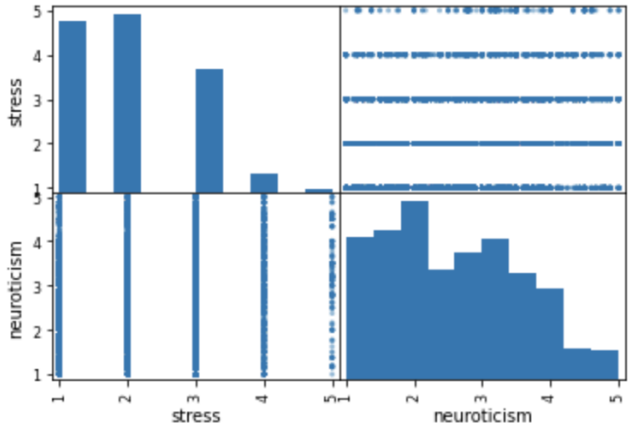

Figure 5.18. Neuroticism vs stress.

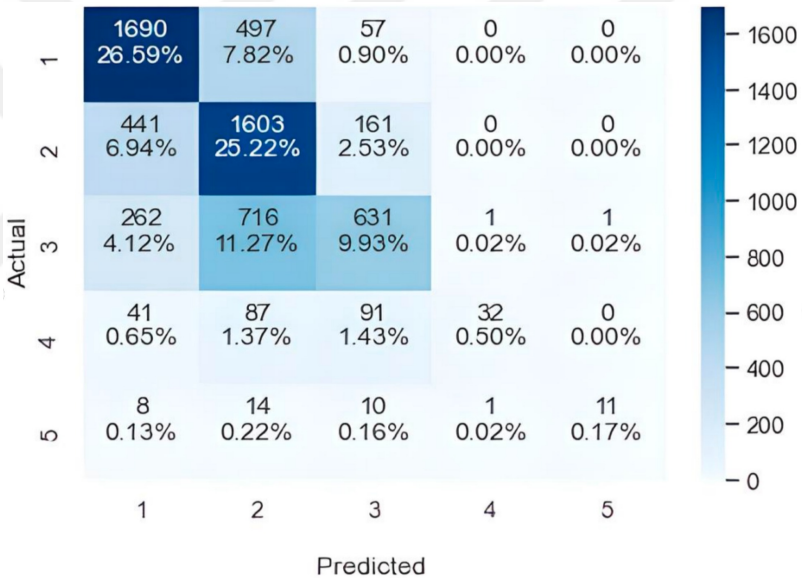

Figure 5.19. Confusion matrix with personality with class imbalance.

5.5.2.2. Performance after Applying SMOTE.

- *Digital biomarkers*

  We apply SMOTE again to address the stress class imbalance. The only difference from the *without personality* dataset is the parameter space, with the same row numbers after SMOTE. Extracted digital biomarkers can be seen in Figure 5.20. Here, the importance of order changed. The details are in Figure 5.14. Phone-related features rank up. Sleep features are important with different ratios.

Table 5.3. With personality with class imbalance.

	precision	**recall**	**F1-score**	**support**
1	0.69	0.75	0.72	2244
2	0.55	0.73	0.63	2205
3	0.66	0.39	0.49	1611
4	0.94	0.13	0.22	251
5	0.92	0.25	0.39	44
accuracy			0.62	6355
macro avg	0.75	0.45	0.49	6355
weighted avg	0.65	0.62	0.61	6355

- *Classification*

  We see the effect of balanced classes. Overall performance in accuracy is increased from 0.62 to 0.80. However, when we examine in detail, we notice that this improvement comes from the $4th$ and $5th$ classes, which have lower instances than other classes similar to *without personality dataset results* in Section 5.5.1.2 after SMOTE application. The other classes' performance remained quite the same except for the $3rd$ one, which increased from 0.49 to 0.70 F1-score. Performance details can be seen in Table 5.4 and confusion matrix in Figure 5.21.

- Causality

  In addition to feature ranking, we employed causal analysis to identify potential digital biomarkers by estimating the causal effects of various features on the target variable. Using the DoWhy library [8] in Python, we constructed a basic DAG where each feature was assumed to be directly related to the target variable. This simplified approach facilitated the computation of causal effects, allowing for a clearer understanding of how each feature influences the outcome. By quantifying the causal impact of each feature, we were able to highlight those that have a significant effect, providing a robust basis for selecting digital biomarkers. The causal analysis conducted on the dataset reveals a diverse range of causal effects,

[8] `https://www.pywhy.org/dowhy/v0.11.1/`

highlighting the complex relationships between features and the target variable. Features such as 'max_pressure' and 'max_visibility' show strong positive causal effects (Figure 5.22). This indicates that increases in these feature values are have positive effect on the target outcome. For instance, higher atmospheric pressure and better visibility might be associated with improved environmental conditions, which could positively influence health-related metrics or other target outcomes. Behavioral features like 'act_walking_ep_0' (walking episodes) also have positive causal effects, suggesting that increased physical activity positively impacts the outcome, which could be related to health or well-being.

Features such as 'floors_climbed' and 'total_snowcm' exhibit significant negative causal effects (Figure 5.23). This implies that higher values of these features are detrimental to the target outcome. For example, climbing more floors might be associated with increased physical strain or risk, negatively impacting the health outcome. Similarly, adverse weather conditions indicated by 'total_snowcm' (total snowfall) might negatively impact outdoor activities or psychological well-being, thus reducing the target outcome's value. The comparison between the feature ranking and causal analysis results provides critical insights into the underlying relationships between the variables and the target outcome. Notably, features such as neuroticism, openness, and unique_gim_count are consistently highlighted by both methods, indicating their strong and direct influence. However, several discrepancies arise, particularly with adjusted_wakeup_time, agreeableness, and various Garmin calorie metrics. Although these features are ranked highly, their causal effects are negative, suggesting that their apparent importance may stem from correlations or confounding factors rather than direct causation. This misalignment underscores the necessity of interpreting feature rankings with caution and considering causal analysis to uncover true drivers of the outcome. For instance, agreeableness and extraversion, despite their high rankings, have negative causal impacts, indicating potential indirect effects or confounding variables at play. Similarly, metrics like adjusted_sleep_duration and unique_act_count, with negligible causal effects, prompt a reevaluation of their direct relevance. Therefore, focusing on features with positive causal effects, such as neuroticism and

openness, alongside exploring the indirect pathways of misaligned features, will enable more accurate and actionable insights for model refinement and intervention strategies.

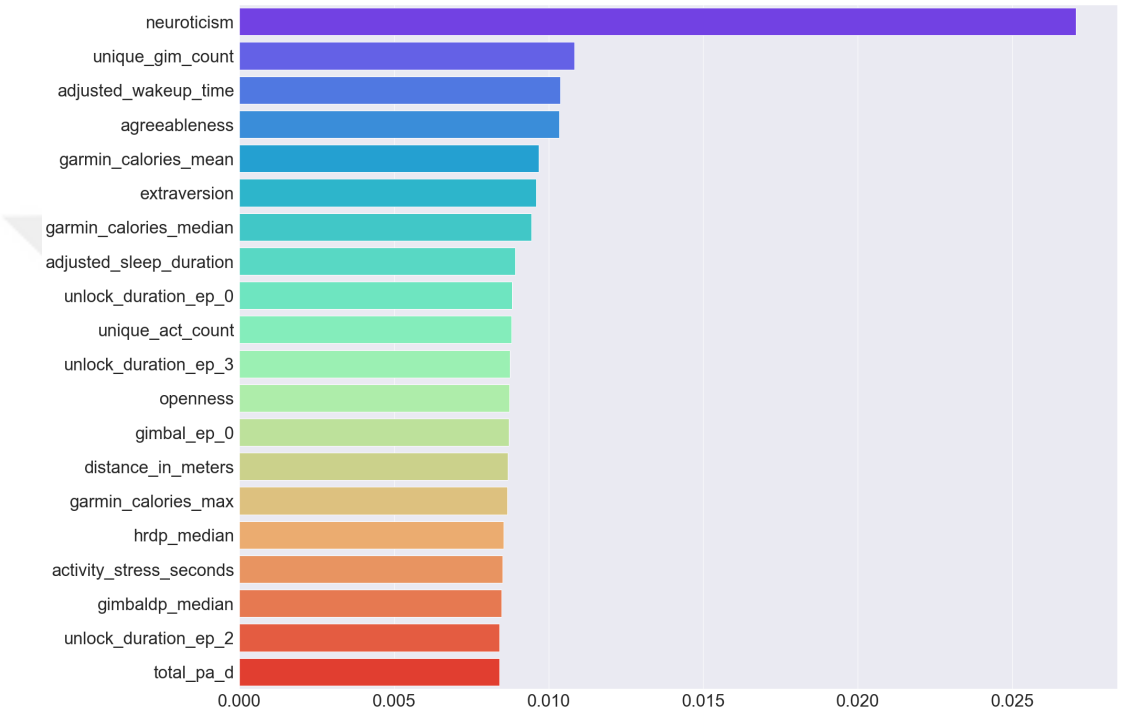

Figure 5.20. Feature ranking after SMOTE with personality.

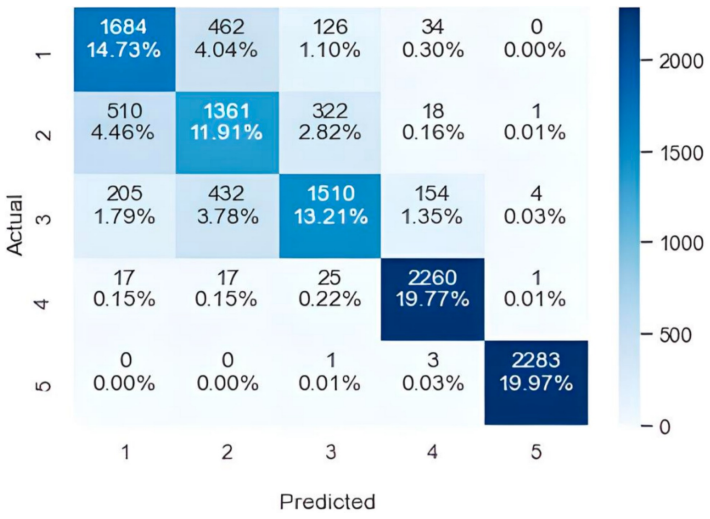

Figure 5.21. Confusion matrix with personality balanced data.

Table 5.4. With personality after SMOTE.

	precision	recall	F1-score	support
1	0.70	0.73	0.71	2306
2	0.60	0.62	0.61	2212
3	0.76	0.66	0.70	2305
4	0.92	0.97	0.94	2320
5	1.00	1.00	1.00	2287
accuracy			0.80	11430
macro avg	0.79	0.79	0.79	11430
weighted avg	0.80	0.80	0.79	11430

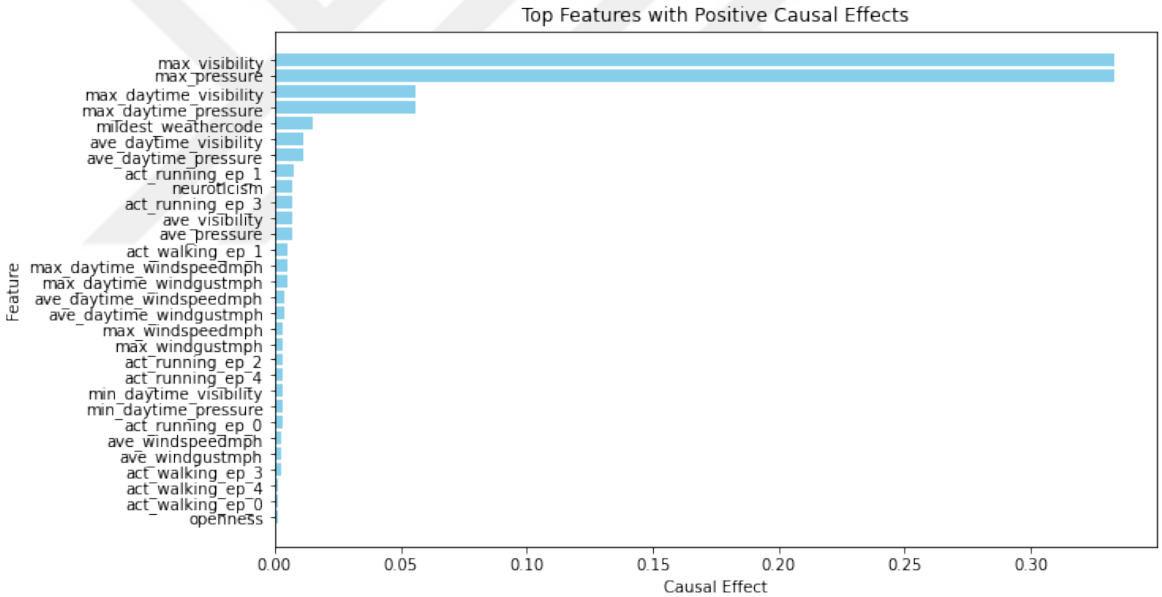

Figure 5.22. Positive causal effect.

5.5.3. Discussion and Comparison

Considering our two different experiments and their sub-scenarios, we obtained higher classification results by including ground truth parameters about personality in our parameter space. These parameters are extraversion, openness, neuroticism, agreeableness, consciousness, and mood-related parameters, such as positive and neg-

ative affect. In addition, we observe that they are selected among the most important parameters after feature ranking impacting the stress state and increasing the classification scores (RQ3). For instance, we obtain 0.60 accuracy (Table 5.1) when they are not included the analysis, while we have 0.62 accuracy (Table 5.3) when we add them (RQ5). The classification performances are even higher when we apply SMOTE and solve the class imbalance problem, which is 0.77 (Table 5.2) when these parameters are not considered, and 0.80 accuracy is achieved (Table 5.4) when they are taken into account. The most important ones are anxiety and negative affection. Intuitively, these are highly related to stress levels. Therefore, their usage for further analysis may be a helpful approach. Besides the personality-related parameters, sleep and phone activity-related markers are the most important ones in all cases (Figure 5.3, 5.4, 5.13, 5.14). In [154], authors found a strong relationship between sleep and stress. Also, in [155], health-related factors such as sleep patterns are found among the important ones affecting stress. Moreover, in study [156], researchers reveal the relationship between positive affect and well-being with sleep. Thus, they are also related to stress due to the relationship between sleep and stress. In [157], again, sleep is found to be related to the mood state. Furthermore, the relationship between exercise and anxiety decrease is examined and found related. Thus, our results aligned with the literature (RQ6). For the best performing case, we examined causal relations and compared them with the biomarker rankings. Although, some features appeared to be in both rankings at high positions, some others are found as having negative causal effects suggesting that indirect effects or confounding factors.

We also observed significant performance improvement after applying SMOTE since the number of instances per stress class became similar. We can better recognize the classes with lower instances. When we look in detail, we observe that the F1-score from the classes of lower instances is lower than the other classes. This is due to the class imbalance problem. Since we have fewer instances than other classes, the model fails to recognize them well. After we apply SMOTE to solve the class imbalance problem by generating data similar to the current measurements, we obtain higher final class F1-scores. As it was not the case for the classes with higher instances, they

had lower precision rates than the others, and their final class scores did not change much after the SMOTE application.

Comparing classifier performance with other stress level classification studies in the literature is also important. In [64], the authors prepared a detailed survey on stress detection in daily life scenarios using wearables. They present the performance in the literature along with the devices, sensors, and methods used. We observe that we achieve similar and even better results considering the ambulatory scenarios in the literature [64] even with a higher number of classes. In [60], they used speech data, applied GMM, and obtained 80.5% accuracy in the stress and relaxed states classification. Another high result is presented in [167]. They received 76% accuracy using *blood volume pulse, skin temperature, electro-dermal activity, heart rate variability, heart rate* parameters applying RF on two classes (stressed, relaxed) problem.

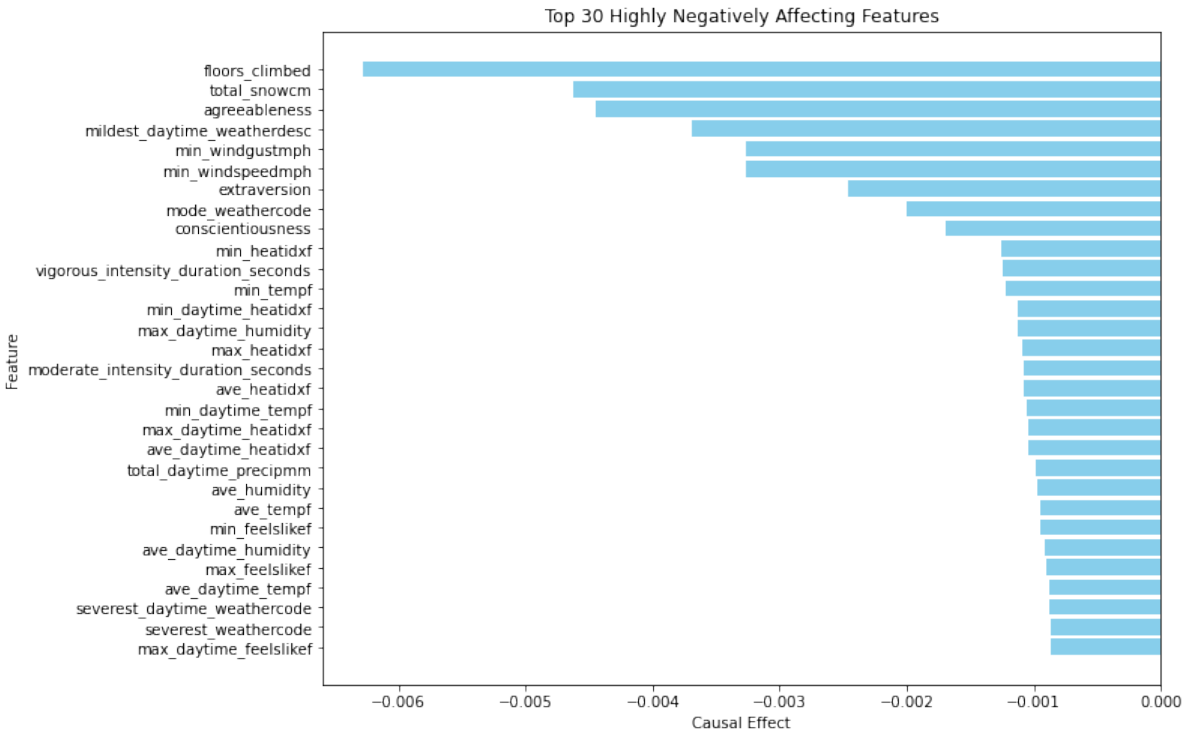

Figure 5.23. Negative causal effect.

5.6. Conclusion

In this section, we examined the digital biomarkers related to stress state since they become valuable information for well-being. We used the Tesserae dataset, collected from office workers. There have been several different modalities. We used the Machine Learning technique to examine these markers, specifically the RF algorithm. We extracted them by dealing with multi-modal data. Also, we made classification with selected digital biomarkers which refer to high importance in *stress* class. We applied SMOTE technique to solve the imbalance issue in the target class. We also analyzed the impact of using personality-related parameters. In all cases, phone activity and sleep modalities-related markers are the most relevant to *stress*. Highest classification score in terms of F1-score is 80% using *with personality dataset* after applying SMOTE. Our findings confirm the literature that examines the statistical relationship between sleep, well-being, stress, positive affect, mood, and physical activity modalities [154–157]. We contribute by working all these modalities in one section in a multi-modal manner by applying machine learning methods over wearable and survey data collected in an unrestricted environment.

6. TIME-BASED WELL-BEING PREDICTION ON THE TESSERAE DATASET

Wearables are able to continuously monitor various physiological signals and capture temporal patterns in real-time. This wealth of temporal data provides valuable insights into behavior analysis and opens up new possibilities for personalized, context-aware applications that can continuously monitor and classify temporal patterns to support various domains, including healthcare, fitness, and productivity enhancement.

Human well-being monitoring and prediction [168] is one of the example application areas focusing on an individual's overall state of physical, mental, and social health, reflecting their sense of contentment, happiness, and fulfilment in life. WHO defines a well-being index; WHO-5 [17] with 5 factors based on the answers given to the following questions: 'I have felt cheerful and in a good spirit', 'Calm and relaxed', 'Active and vigorous', 'Woke up fresh and rested', 'Daily life interests me'. The answers range from 'All of the time' to 'At no time' with six possible inputs.

In recent years, traditional machine-learning algorithms, such as logistic regression, decision trees, and ensemble methods, have been extensively utilized in classifying or predicting well-being using data collected from wearables and questionaries. However, the time aspect of the temporal data, in other words, how previous data impact future well-being levels, is often overlooked in related studies. Wearables offer a unique opportunity to collect longitudinal data, enabling a deeper understanding of how stress, anxiety, and emotional states fluctuate over time [169, 170]. Time-based prediction techniques, when combined with wearable devices, can offer a unique opportunity to capture and analyze temporal patterns. This integrated approach can provide valuable insights into individuals' emotional states over time, facilitating early detection and timely interventions for improved mental well-being.

In addition, deep-learning architectures have emerged as powerful tools for time-based prediction problems. RNNs and CNNs have gained attraction in well-being research [27], [130], [171–174]. RNNs, particularly LSTM variants, have demonstrated their ability to capture long-term dependencies and temporal dynamics in well-being-related data, while CNNs extract hierarchical features from wearable sensor data, enabling more accurate prediction. Furthermore, the combination of deep-learning architectures with wearable devices has led to the development of hybrid models that leverage both wearable sensor data and contextual information. These models incorporate multimodal inputs, such as physiological signals, accelerometer/activity data, and contextual features like time of day, location, social interactions, and self-reported mood states. By fusing multiple data sources that may impact an individual's well-being levels, the model's performance can be improved.

As well as deep-learning algorithms, ensemble methods are also powerful tools. Ensemble models combine multiple base classifiers to make predictions, leveraging the strengths of different models and reducing the impact of individual model weaknesses. Techniques like Bagging, Boosting, and RF have been successfully employed in time-based classification tasks, yielding improved performance and robustness.

In this section, specifically, we focus on workplace well-being prediction using the Tesserae dataset [16], which was collected over an extended period from office workers (details were explained in Section 2). The primary objective is to predict the stress, anxiety, and positive and negative affect levels by analyzing the impact of past measurements on their well-being. There is also the sleep factor in the definition of the well-being index based on the WHO-5 well-being questionnaire [17]. Although the sleep response was recorded in the dataset, once we applied the same preprocessing steps to compare each target value's results fairly, we were left with a limited amount of data due to the lack of intersection in missing values in parameter space and target value, i.e., sleep which is collected three to four times in a week even though they are classified in daily surveys. Thus, we could not include this fifth factor in the section.

Given the fact that the dataset is a time stream, we conducted experiments using prediction algorithms, namely RF), XGBoost as example ensemble algorithms, and LSTM as an example deep-learning algorithm, in their conventional versions as well as in time-based versions. Time-based versions use the concept of time lags to consider the time aspect of the data. In order to create the time-based versions, each data point is associated with previous time steps (i.e., days) to incorporate historical information. We aimed to observe the impact of previous, such as daily, weekly, and longer periods, well-being factors on the upcoming ones. We focus on the RQ4, RQ6, and RQ7 stated in Section 1.

6.1. Background and Motivation for a Time-Based Approach

In this section, we deal with temporal patterns and dependencies within the data using methods such as RF, XGBoost, and LSTM. Based on past data, these algorithms can effectively predict stress levels, anxiety states, and variations in positive and negative affect. Importantly, these algorithms' lagged (time-based) versions, incorporating various time windows, have shown promise in capturing nuanced temporal trends.

To the best of our knowledge, none of the well-being studies considered these four factors simultaneously. Furthermore, there is no time-based method application using the current dataset. Considering the intrinsic aspect of data collection and the nature of the states, we show that temporal dependencies have an effect. Even though RF was reported to be the most performing algorithm [28], [74], [126] in this dataset's scope, we are also exploring the performance with XGBoost and LSTM alongside RF. In addition, in our previous study [28], we extracted highly related digital biomarkers only for the classification of the stress variable. In this section, we expand by adding three more well-being factors in a prediction context rather than classification.

Ultimately, this research aims to advance the field of time-based classification, particularly in the context of workplace well-being prediction. By addressing challenges, exploring novel methodologies, and leveraging wearable technology, we strive

to contribute to developing intelligent systems that can effectively analyze and interpret temporal data for personalized applications in the workplace.

Time-based approaches can be used in well-being detection and management by leveraging physiological and contextual data collected over time. The data streams provide rich information that can be leveraged to develop accurate and personalized stress, anxiety, and mood prediction models.

We aim to answer the effect of time-based algorithms compared to their conventional versions and extract the related digital biomarkers (RQ4). In prediction tasks, the data format can significantly impact the modeling approach and the insights gained. Two primary data formats, time-based and conventional predictions, offer distinct characteristics and utility.

Time-series data are organized chronologically, where each observation is tied to a specific timestamp. This format introduces a crucial temporal dimension, capturing how the studied phenomenon evolves. Notable characteristics include the temporal sequencing of data points, time lags and dependencies necessitating feature selection, and the option to aggregate data over different time intervals. These features allow for modeling temporal patterns and dependencies, making time-based prediction particularly valuable in scenarios where understanding how variables change over time is essential, such as stock price forecasting, weather prediction, or user behavior analysis.

Conversely, the usual prediction data format lacks an inherent temporal component, treating each observation as independent and disregarding the order of data points. Key characteristics include the independence of observations, a focus on feature engineering that primarily considers static attributes, and the use of cross-sectional analysis, which provides a snapshot of the data at a particular time. This format is for scenarios where the sequence of observations is irrelevant or temporal aspects hold no significance in the predictive modeling process. It is often employed in areas like classification tasks and static attribute-based predictions.

6.2. Related Studies

Well-being, a multidimensional concept encompassing various facets of emotional and psychological health, has become a focal point of research [168], [175–178] and early health intervention by continuous monitoring [179, 180]. Within the spectrum of well-being, stress, anxiety, positive affect, and negative affect play pivotal roles in determining an individual's mental and emotional state [181, 182]. Stress and anxiety can significantly impact one's overall well-being, and understanding and managing these factors are essential for promoting mental health [181], [183, 184].

Studies have explored the integration of wearable sensors, such as heart rate monitors, electrodermal activity sensors, accelerometers, and even Bluetooth beacons, to capture physiological and behavioral signals associated with stress and emotional well-being [185, 186]. These sensors provide non-intrusive and continuous data streams that offer insights into individuals' emotional experiences and their positive and negative affect.

Nevertheless, the most studied factor is stress in the literature [71], [87], [126], [166], [187–189]. In recent years, stress prediction and assessment have witnessed a surge in innovative research aimed at harnessing diverse data sources and cutting-edge machine-learning techniques. These efforts have collectively contributed to a deeper understanding of stress patterns across different contexts and data modalities, paving the way for more effective stress management strategies.

As also discussed in Section 3, Breideband et al. [20] studied the impact of sleep patterns on remote teamwork and mood during the COVID-19 pandemic. Over ten weeks, they monitored 71 information workers across 22 teams using wearable sensors and daily questionnaires. Their findings revealed significant changes in sleep behavior due to remote work, such as delayed sleep times, longer sleep durations, and shifted schedules, all of which affected individual mood and team dynamics. The study found a complex relationship between sleep patterns and team dynamics. Surprisingly, more

sleep was linked to a decrease in positive affect and teamwork perceptions, whereas misaligned sleep schedules among team members led to improved positive affect and teamwork. These results challenge conventional assumptions about the straightforward benefits of longer sleep on productivity and collaboration.

These insights emphasize the importance of considering sleep patterns in remote team dynamics. Understanding how sleep influences team performance and well-being can help develop strategies to support a healthy work environment and enhance virtual collaboration. Further research is needed to explore the intricate relationship between sleep, mood, and teamwork in remote work settings.

Kang et al. [190] highlight the significant contributions of datasets to emotion intelligence technology with their introduction of the K-EmoPhone dataset. This dataset has played a crucial role in advancing affective computing by facilitating the analysis of real-world behavioral patterns despite facing certain limitations. It provides labeled data on emotions, stress, and attention, supporting machine learning models in predicting various affective states and contributing to attention management studies. However, the study acknowledges challenges such as the discontinuation of devices used for data collection, such as the MS Band 2, which may impact the scalability and reproducibility of research findings. Additionally, the absence of real-time monitoring during data collection raises concerns about the quality and accuracy of the captured information. Overall, the K-EmoPhone dataset represents a valuable resource for researchers in the field of emotion intelligence technology, enabling advancements in understanding and analyzing human emotions and behaviors despite the inherent challenges it faces.

Individual mood, encompassing emotions such as happiness, sadness, and joy, significantly influences personal lives [191]. Research has explored the connection between distinct activity patterns and depressive symptoms [192]. Mobile phone apps have emerged as valuable tools for inferring symptoms such as depressive mood, anxious mood, or fatigue using parameters like heart rate variability and skin conductance [193, 194]. Studies, such as those involving the Moodscope application [195],

have identified associations between decreasing calls, SMSs, and location entropy with feelings of sadness and depression, using EMA questionnaire labels as ground truth.

The proposal of personalized models for mental health monitoring by Paromita et al. [196] introduces individualized approaches to estimating stress and anxiety among healthcare workers. Despite limitations such as a small dataset, their models aim to contribute to personalized healthcare interventions, highlighting the importance of tailored approaches in mental health monitoring.

The paper [196] proposes personalized models for estimating stress, anxiety, and affect among healthcare workers using ambulatory-based multimodal signals. These models outperform non-personalized techniques, highlighting the importance of individualized approaches in mental health monitoring. The study aims to reduce inter-individual variability while preserving behavioral construct variability, contributing to personalized healthcare interventions. The study's use of a small dataset and predefined similarity threshold may limit the generalizability of the proposed models. Additionally, the reliance on self-reported surveys for ground truth data introduces the potential for response bias.

Veena et al. [21] explore the practical applications of Fitbit data for maintaining a healthy lifestyle, employing unsupervised clustering to classify subjects based on sleep and activity metrics. Despite challenges related to data accuracy, reliability, and security, their study aims to investigate anomalies and trends in Fitbit healthcare data, contributing to the understanding of temporal behavior detection using raw physiological signals.

The GLOBEMDataset [197] provides multi-year datasets for studying human routines and well-being metrics, supporting the development of longitudinal behavior modeling algorithms. Despite privacy risks, this dataset enables advancements in behavioral science and machine learning, fostering a deeper understanding of daily human behaviors and well-being.

Investigating the feasibility of stress prediction based on behavioral data, researchers have delved into smartphone activity as a potential stress indicator [87]. Leveraging machine-learning algorithms, this study dissects smartphone behaviors to uncover stress-indicative patterns, therefore contributing to a deeper understanding of stress dynamics in the digital age. By focusing on digital behavior, the study presents a unique perspective on stress prediction, emphasizing the significance of smartphone interactions in capturing stress-related cues.

Wearable technology emerges as another avenue for stress prediction. Employing physiological sensors, studies have endeavored to predict and visualize work-related stress through wearable sensing [188]. Heart rate and skin conductance are monitored to develop models that provide real-time insights into stress levels, thus empowering individuals and organizations with tools for effective stress management. By focusing on physiological markers, this research offers a unique approach to stress prediction, emphasizing the importance of wearable technology in monitoring and mitigating stress.

Moreover, the analysis of physiological signals has yielded significant strides in stress detection under real-life conditions [71]. By scrutinizing heart rate variability and electrodermal activity, researchers provide valuable insights into physiological stress markers, enhancing our understanding of stress dynamics in natural environments. This study's focus on physiological markers in real-life settings offers a nuanced perspective on stress detection, providing insights into stress responses beyond controlled environments.

Mobile sensing has also paved the way for predicting stressful life events by analyzing sensor data such as location and physical activity [189]. This novel approach enhances our comprehension of stress-inducing contexts and contributes to users' overall well-being by identifying potential stress triggers. This approach shifts the focus from immediate stress levels to predicting stress-inducing situations, highlighting the potential of preemptive interventions.

Intriguingly, multimodal approaches have been explored by combining speech and wearable sensor data for stress detection [126]. By integrating speech features and physiological signals, researchers have showcased the potential of using multiple data sources to achieve more accurate stress prediction models. The study's emphasis on combining speech and physiological data offers a comprehensive approach to stress detection, leveraging multiple modalities to enhance prediction outcomes.

Comprehensive stress and sleep prediction strategies have incorporated physiological signals and smartphone data [84], [166]. The fusion of heart rate variability, accelerometer data, and self-reported stress levels yields enhanced accuracy, demonstrating the potential of integrating diverse data sources for holistic stress assessment. This research stands out for its comprehensive integration of physiological signals and smartphone data, offering a multifaceted perspective on stress prediction [166]. In [84], the approach of combining wearables and phone sensors enables more accurate sleep detection by leveraging the benefits of both streams: combining wearable movement detection with mobile phone technology usage detection is employed. They showed that the combination of phone activity and wearables might produce better models of self-reported sleep than either stream alone on the Tesserae dataset.

There are also recent studies focusing on monitoring and predicting stress levels in the workplace. In a recent study [187], researchers developed a mobile app to collect a comprehensive dataset from 30 workers over 8 weeks. The app prompted users to complete a questionnaire three times daily, assessing stress, sleep quality, work abandonment, energy levels, and mood.

6.3. Data Preprocessing

As we found in Section 5, personality attributes are highly related with stress value, we employed these parameters in our input space. These parameters are extraversion, agreeableness, conscientiousness, neuroticism, and openness. Specifically, in this dataset, these personality values are collected approximately 4–5 times during the

period that we are concentrating on, i.e., two months. The selection of input columns and the construction of the final dataset is explained extensively in the "Dataset Construction Section" in our study [28] and a descriptive table (Table 3) is presented in Section 2. More detailed feature names can be accessed in Section 2.5.1.2.

We examined missing values among the independent, i.e., input, variables. Some attributes were never collected from some participants. Thus, an imputation could not be applied to these attributes. Among them, if there is a lack of many participants' data, their removal by row leads to a high decrease in our dataset. Instead, we deleted those columns. These are *act (activity) still*, *light mean*, *garmin hr (heart rate) min*, *garmin hr max*, *garmin hr median*, *garmin hr mean*, *garmin hr std*, *ave (average) hr at work*, *ave hr at desk*, *ave hr at desk*, *ave hr not at work*, *call in num (number)*, *call in duration*, *call out num*, *call out duration*, *call miss num*, and their derivatives according to time episodes. However, when we have data columns with missing values at the person level, we compute them by applying the mean operation for each participant. Even after this operation, there were some missing values due to the lack of column values at a person level. We had to remove them row by row, resulting in the removal of around 5000 rows (from $36,284$ to $31,772$). During this process, we observed that missing values at the person level for one parameter were highly correlated with the presence of missing values for other parameters as well. In addition, we applied data scaling using a scale mapper.

We proceeded with the same preprocessing steps for each target factor and achieved the same parameter space. In the end, we have 194 columns with input and target parameters and around $31,700$ rows due to changes in missing values in the target column. More precisely, it is $31,772$ for stress, $31,731$ for anxiety, $31,703$ for positive affect, and $31,702$ for negative affect. To give an idea about the dataset, we present some columns with their corresponding values and participant ID in Table 6.1 for the positive affect target variable.

Table 6.1. A snapshot of the data for positive affect.

ID	**Timestamp**	**Active Secs**	...	**Resting Hr**	**Steps**	**Stress Duration Seconds**	...	**Positive Affect**
4188a14d4edd8eb1cac1a146d9f88aee	21 February 2018	2,417	...	57.0	6,634	34,560.0	...	14
4188a14d4edd8eb1cac1a146d9f88aee	22 February 2018	2,557	...	55.0	6,008	21,960.0	...	14
4188a14d4edd8eb1cac1a146d9f88aee	23 February 2018	5,849	...	55.0	9,008	26,880.0	...	15
...	...	...	...	...	...	...	...	...
4188a14d4edd8eb1cac1a146d9f88aee	27 April 2018	3,595	...	72.0	4,837	12,240.0	...	15
...	...	...	...	...	...	...	...	
c30291318f6d680bd65666c183f6bb5e	11 January 2018	4,988	...	57.0	10,550	19,020.0	...	11
...	...	...	...	...	...	...	...	...
c30291318f6d680bd65666c183f6bb5e	13 March 2018	5,109	...	50.0	11,002	46,140.0	...	9

6.3.1. Creation of Lagged (Time-Based) Dataset

When dealing with time-series data in healthcare, it is essential to consider temporal dependencies and trends. The traditional version of the indicated algorithms (RF, XGBoost) may not capture these aspects effectively, as it does not inherently account for the sequential nature of time-series data. This is why we created lagged versions of the dataset. Also, even though LSTM considers the time aspect inherently, its lagged version, often referred to as Time-Lagged LSTM, extends the capabilities of traditional LSTMs by explicitly incorporating lagged features into the model. They excel at capturing complex temporal relationships and can provide valuable insights for time-series analysis in healthcare.

A lagged version introduces the concept of time lags to consider the time aspect in the data. It involves the creation of lagged versions of the features, where each data point is associated with previous time steps to incorporate historical information. This approach can significantly enhance the predictive performance, especially when dealing with time-series data. The model better understands how data evolves over time, making it a valuable tool for time-sensitive healthcare applications.

To explain this concept, consider an employee with a certain level of well-being (W) in terms of stress, anxiety, positive and negative affect separately (either one of them depending on the scenario) during a certain period t are

$$W(t) = [w_{t-n}, w_{t-(n-1)}, \ldots, w_{t-1}, w_t], \tag{6.1}$$

where W_t represents the well-being parameter value W over a period t, on a range of 1 to 5 for stress and anxiety, 5 to 25 for positive and negative affect. $w_{t-n}, w_{t-(n-1)}, \ldots, w_{t-1}$, w_t are the individual measurements of the well-being parameter W at different time points within the time window. Therefore, $w_{t-n} \in [1, \ldots, 5]$ or $[5, \ldots, 25]$ depending on the scenario of well-being, and n is the size of the time window, indicating the number of past measurements considered in the measurements of well-being parameters. In this section, we employed $1, 3, 7, 15, 30$ as look-back (n) sizes to consider daily, weekly,

fortnightly, and monthly results. Additionally, we wanted to observe the change between daily and weekly performance, so we also employed a 3 day look-back roughly to state half of the week. We could have considered longer periods, such as yearly and seasonal periods. However, as we have only 56 days of data, these approaches are not applicable due to the dataset constraints.

6.4. Prediction Algorithms

6.4.1. Random Forest

RF is an ensemble learning method that combines multiple decision trees to make more accurate predictions or classifications. It is particularly effective for various tasks, including regression and classification. It works by creating a multitude of decision trees during training. Each tree is constructed based on a random subset of the dataset (bootstrapping), and at each node, a random subset of features is considered. This randomness helps in reducing overfitting and improving generalization. During prediction, each tree provides an output (class label or numerical value), and the final prediction is determined by averaging (for regression) or taking a majority vote (for classification) of these individual tree outputs.

6.4.2. XGBoost

XGBoost, short for Extreme Gradient Boosting, is another robust machine-learning algorithm commonly used in various domains, including healthcare. Again, it is an ensemble learning technique that combines the predictions of multiple decision trees to achieve high predictive accuracy. XGBoost is known for its efficiency, scalability, and ability to handle various data types, making it a popular choice for predictive modeling tasks.

6.4.3. Long Short-Term Memory

LSTM is a RNN architecture that handles sequential data. It is well-suited for time-series analysis, natural language processing, and other applications. LSTMs are known for their ability to capture long-range dependencies in sequences while avoiding the vanishing gradient problem, which can hinder the training of traditional RNNs. LSTM models can be used in healthcare for patient monitoring, disease prediction, medical image analysis, and more. They excel in scenarios where the order and temporal relationships between data points are critical.

We used the RF algorithm in our analyses because it is an ensemble method and performs better among the other used methods in the literature in this domain [198], XGBoost as it is indicated to perform better compared to deep-learning techniques and even better once combined with deep-learning compared to alone version [199], and LSTM as it considers time aspects of data inherently. We also have their lagged versions for $1, 3, 7, 15, 30$ days look-back to consider the contribution of past data points and $1, 3, 7$ days lookup to understand up to how many days we can reach reasonable errors and which one's performance is better.

6.5. Performance of Well-Being Factors Prediction

6.5.1. Implementation Details

To be fair, in comparisons, as the parameter space is 99% is similar (in Section 6.3, the total number of instances per target value was given), and only the target column changes for each scenario specific to different well-being targets. We used the same preprocessing steps and hyper-parameter values for all algorithms. After hyper-parameter optimization, we found that *n_estimators* 1000, criterion *gini*, min sample split 2, min samples leaf 1, max features sqrt combination reveal the best performance for RF, 50 layered LSTM with epochs 50, batch size 72, adam optimizer with learning rate 0.001, hyperbolic tangent as an activation function and MAE as a loss function,

and for XGBoost colsample bytree 0.8482, gamma 4.34, max depth 3, min child weight 7.0, reg alpha 107.0, reg lambda 0.8336, *n_estimators* 20.

We employed 80% and 20% train and test dataset sizes, respectively, as in [28]. However, in this section, we also considered a person's ID information for splitting even though it is not personalized yet. When we split the dataset sharply to obtain 80% and 20%, it corresponded to splitting one person's data, some parts in the training set and some in the test. To solve this issue, we assigned 443 person's data in the training set and 100 person's data in the test set to reach roughly 80% and 20% splitting. In the context of neural networks, it is a common practice to have three different sets: train, test, and validation. Here, we did not employ a three-set division. The decision to exclude a separate test set was motivated by two key factors. First, training was stopped when no improvement between epochs was observed. This practice aligns with the common approach to halt training when the validation metric ceases to improve, indicating potential diminishing returns and the risk of overfitting. Second, the availability of a large and diverse dataset played a significant role in this decision. A substantial dataset can mitigate overfitting risks by providing a robust measure of the model's generalization capability without the need for a dedicated test set, which was not observed during our experiments.

6.5.2. Digital Biomarkers of Well-Being Factors

Before examining the performance of time-based versus non-time-based predictions, we extracted the most effective parameters, i.e., digital biomarkers, on the target well-being factors. We employed RF for feature importance. For each target variable, top 20 features are presented in Figure 6.1, 6.2, 6.3, 6.4, corresponding to stress, anxiety, positive affect, and negative affect, respectively. As these explain the health-related components, in the literature, they are called digital biomarkers [8] (also discussed in Section 2). By doing this, we aim to answer RQ4, mentioned in Section 6.1. In a previous study [28], concentrating on a stress classification task, we have already found that stress has two overwhelming digital biomarkers: anxiety and negative affection.

This was also the case in this prediction study. As these two factors are also our target variables in the scope of this work, we excluded the other examined factors while targeting one from the set of well-being factors. That is why they are not listed in the most important digital biomarkers lists.

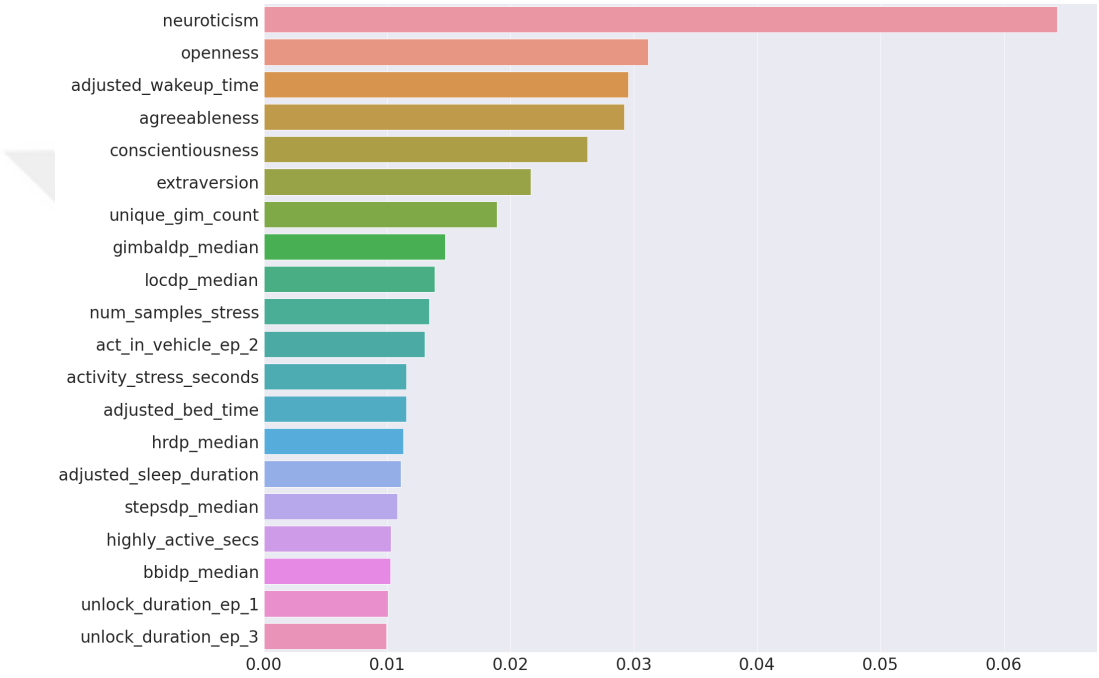

Figure 6.1. Stress.

Similar to the results presented in the stress classification in Section 5, the personality factors, which are extraversion, agreeableness, conscientiousness, neuroticism, and openness, are all found among the most important factors, almost always in the top 5 again (Figure 6.1, 6.2, 6.3, 6.4) in this section, for all target variables in a separate manner for a regression problem. As well as the personality attributes, we notice that some others, such as sleep-related parameters, including bedtime and wakeup time, are the common attributes in all factors. Similarly, the number of stress samples, highly active seconds, and activity stress seconds are also common parameters. In addition, phone unlock duration, which comes from phone activity, is common in three target factors, i.e., stress, anxiety, and negative affect.

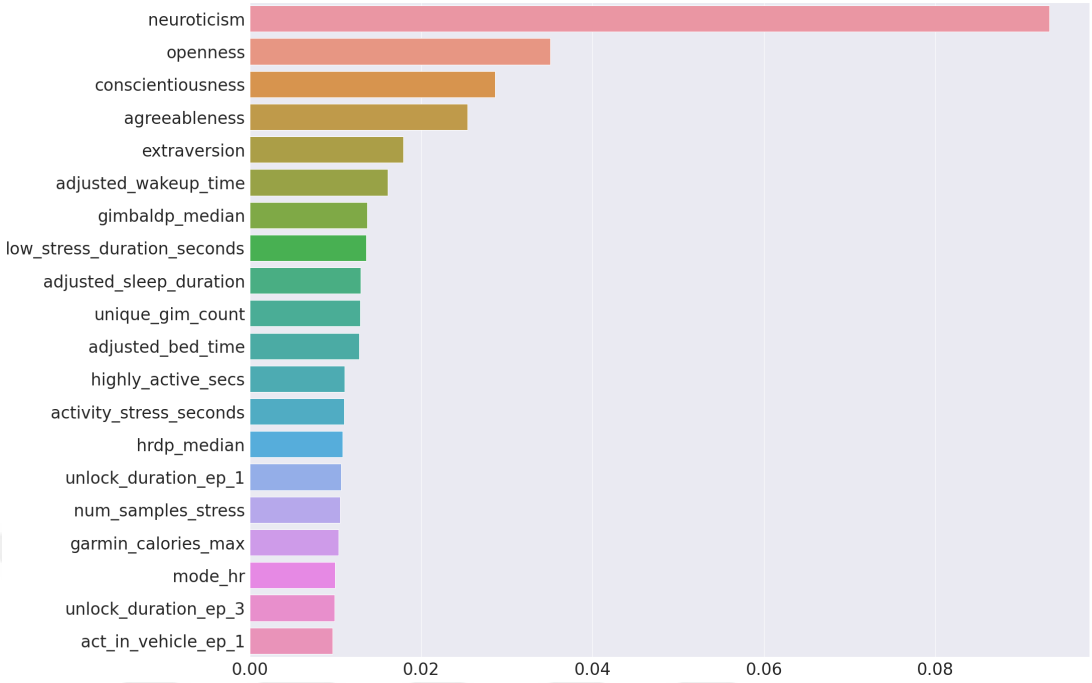

Figure 6.2. Anxiety.

As we are dealing with multiple types of inputs, i.e., digital biomarkers, coming from different modalities, i.e., measurements from different devices (phone, smartwatch) and personality surveys, we wanted to understand the importance of the modalities for future studies to decide on what types of data should be collected to predict the well-being factors. The modalities are phone activity, personality surveys, and the ones coming from the Garmin watch: sleep, stress, daily (this includes a summary of daily recordings, such as distance in meters), activity, and heart rate.

We rank the digital biomarkers according to their corresponding modality set. Since we have extracted the most important 20 features, rankings are from 20 to 1, 20 as the most important one. We sum the rankings of the digital biomarkers when computing the importance of the modality they belong to. For example, bedtime, sleep duration, and wakeup time digital biomarkers are under the sleep modality from the watch. If these were found among the most important 20 features, we sum their ranking values for the sleep modality.

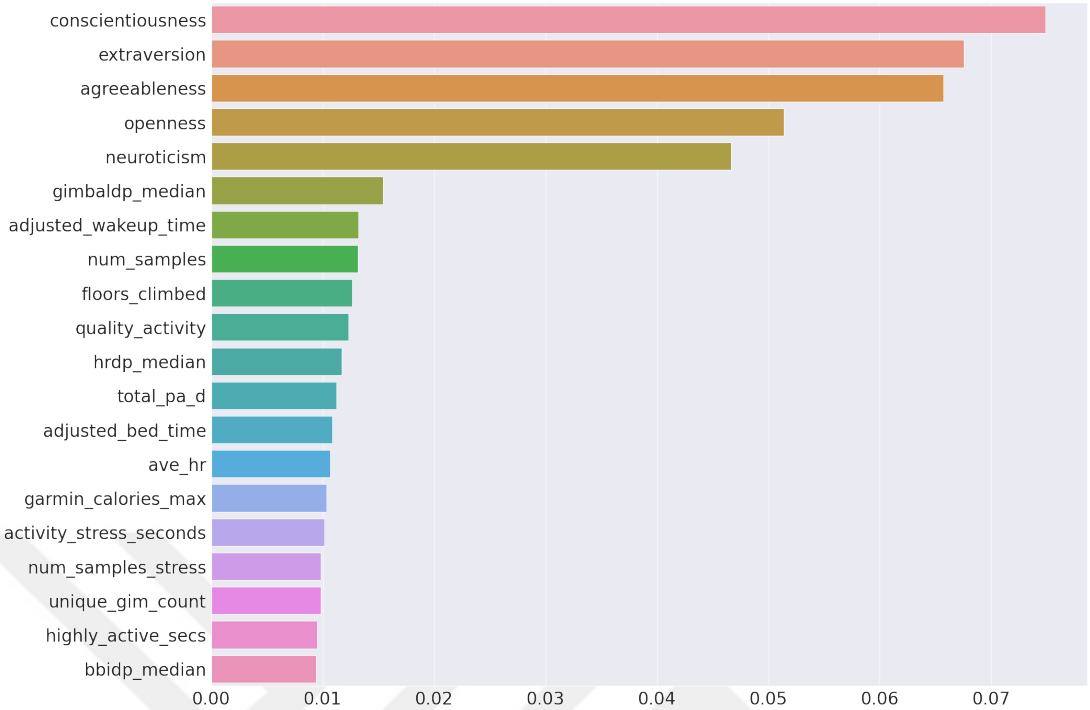

Figure 6.3. Positive affect.

In Figure 6.5, we provide the rankings of the digital biomarkers according to their modalities from the phone, the wearable watch (Garmin), and surveys (personality). We observed a similar ranking as *personality*, *phone activity*, *sleep*, *daily activity* for each target factor. The rest of the rankings are as follows *activity* for stress; *activity*, *heart rate* for anxiety; *heart rate, stress, activity* for positive affect; *activity*, *stress*, *heart rate* for negative affect (Figure 6.5). In addition, this ranking is the same for each target variable, which shows the transitivity of these modalities across targets.

As well as wearable and phone devices' various measurements to relate to the psychology literature in answering our second research question (RQ6), we focus on the parameters coming from sleep, stress, heart rate, and activity modalities. Please note that contrary to target variables from daily survey data, these parameters come from the device measurements. This is why we encounter stress sample numbers among the important features for stress targets. Observing these parameters on the parameter space validates device measurements with the subjective values.

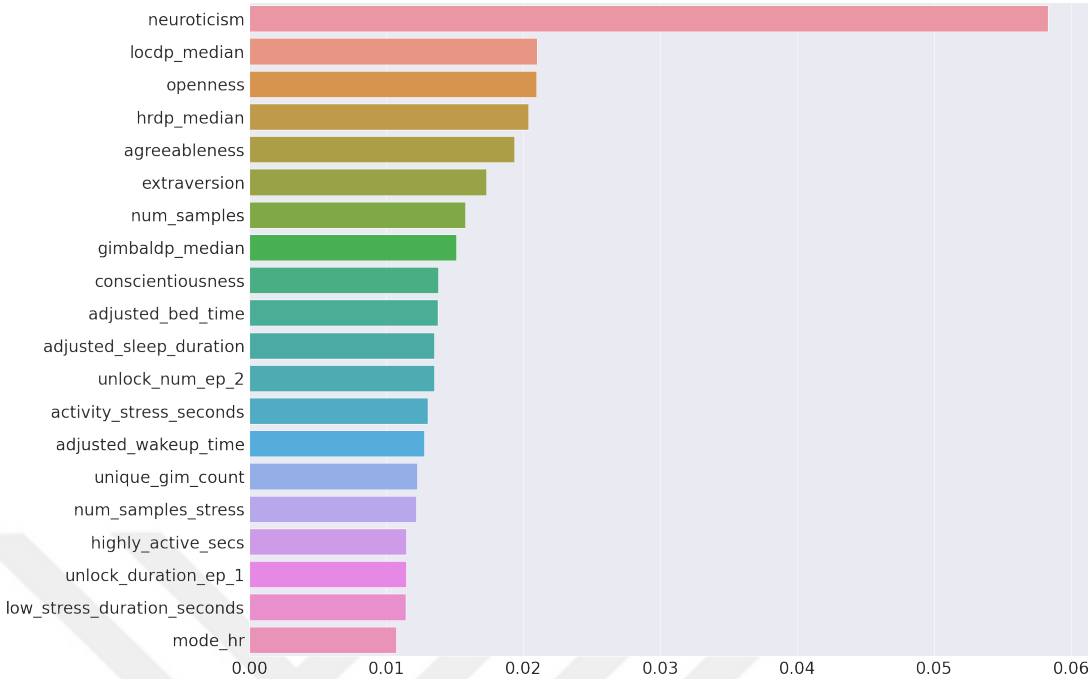

Figure 6.4. Negative affect.

It is observed that sleep and activity occur commonly among all target important variables. In the literature, it is stated that there is a strong relationship between sleep and stress [154], [200, 201], sleep and anxiety [183], [202, 203], sleep and mood (positive and negative affect) [156]. In addition, it is found that sleep patterns are among the most important health-related factors [155]. This section also validates these results (Figure 6.5) by having sleep in the upper ranking of importance compared to activity, stress, and heart rate for each one of the target variables.

In [184], [204], researchers found that there is a relation between physical activity, anxiety, and stress. It is stated that with the increase of physical activity levels, there is less subsequent stress and negative affect, as well as more positive affect [205]. In our results, activity-related parameters are found towards the end of the importance rankings in Figure 6.1, 6.2, 6.3, 6.4, yet the results shown are the most important among the whole parameter space. Thus, again, the results aligned with the psychological literature.

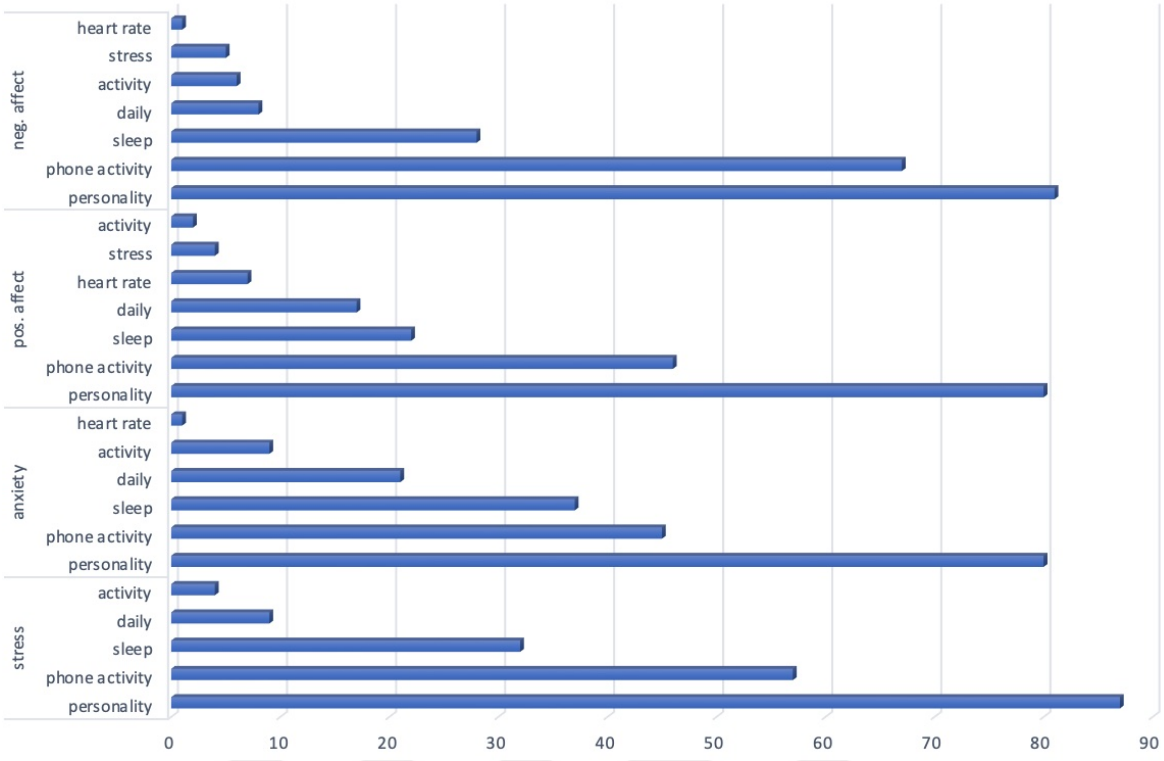

Figure 6.5. Modality rankings.

Interestingly, most of the stress-related factors were found to be important for positive and negative affect. In [206, 207], it is reported that perceived stress level is associated with positive and negative affects' reactivity to current events while trait anxiety moderated reactivity of agitation. Also, it is stated that the intensity and duration are important for their relationship. This finding also overlaps with our findings.

In conclusion, this section provides the most important physiological parameters that align with the existing studies in the literature from the psychological perspective.

6.5.3. Conventional vs. Time-Based Prediction Performances

In well-being, four key factors available in this dataset provide insights into an individual's psychological state. Stress and anxiety, rated on a scale from 1 to 5, encompass the emotional responses to challenges and worries. Stress levels indicate the

intensity of pressures and challenges experienced, while anxiety reflects the degree of unease and fear. On the positive side, positive affect, rated from 5 to 25, measures the abundance of positive emotions, indicating the extent of happiness and enthusiasm. Conversely, negative affect, also rated from 5 to 25, gauges the presence of negative emotions, portraying the degree of distress and unhappiness. These measures collectively offer a comprehensive view of an individual's emotional state and overall well-being, from coping with stress and anxiety to experiencing positive and negative emotions.

As stated earlier, we are using RF, XGBoost, and LSTM algorithms and their lagged variations in this section. This section will refer to them as conventional and time-based versions to facilitate the understanding. The performances of the algorithms are measured in terms of MAE to be able to account for negative errors properly. The reason behind the choice of MAE comes from the observation of the model's under-prediction, as one may observe in Figure 6.6, 6.7, 6.8, 6.9. Thus, no sign difference suits the usage of MAE as an error metric. The X-axis represents the index ranging from 0 to 249, corresponding to the data points in the lagged dataset. Since the dataset is constructed with consideration for the time aspect, these indexes are also related to time.

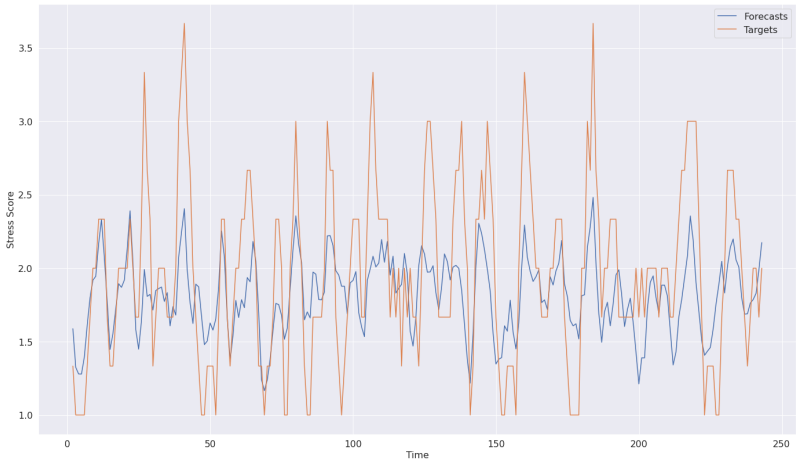

Figure 6.6. Stress.

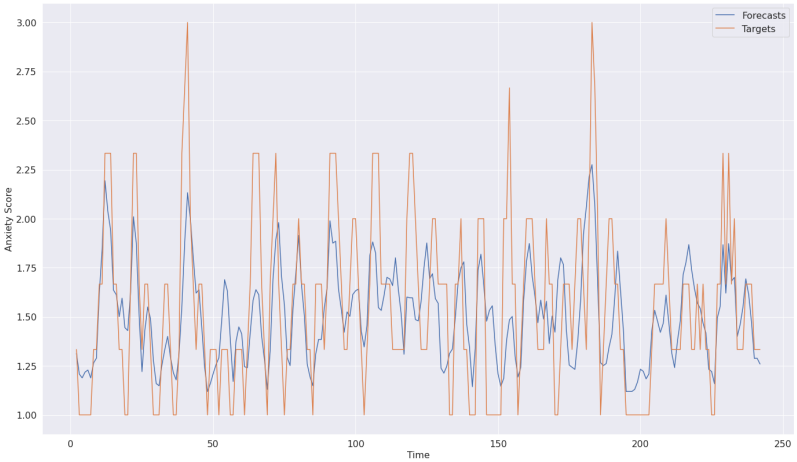

Figure 6.7. Anxiety.

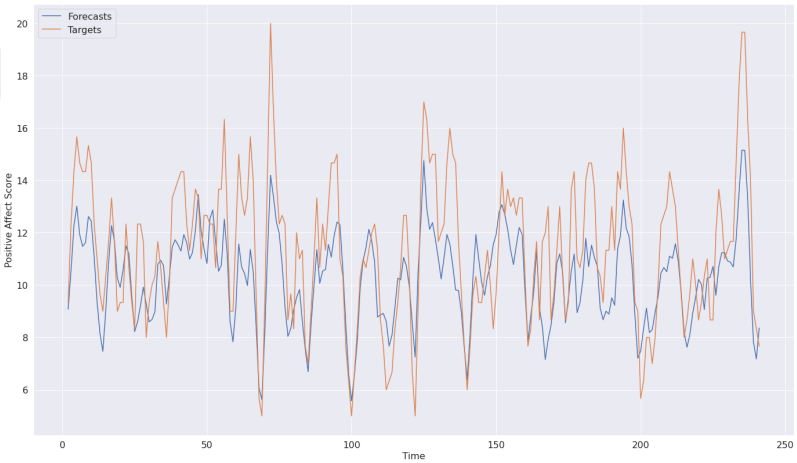

Figure 6.8. Positive affect.

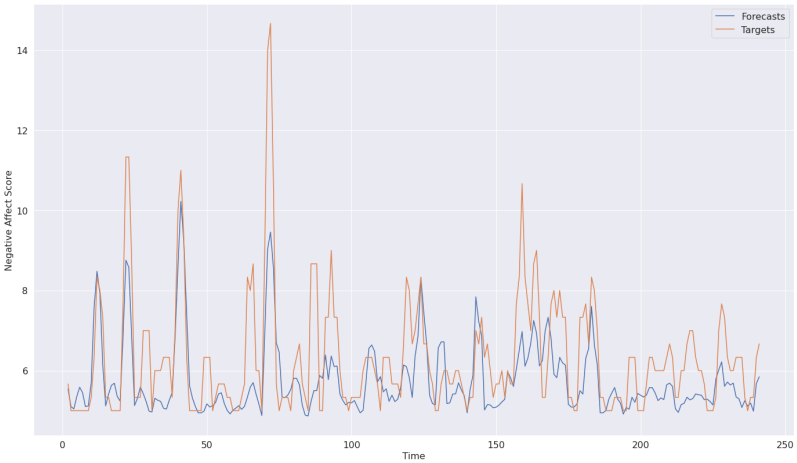

Figure 6.9. Negative affect.

The results are presented in Table 6.2. When we look at the results of the conventional versions of the algorithms, we see that RF and XGBoost results are closer, but LSTM is better since it captures the time aspect in the data intriguingly. On the other hand, we observed that XGBoost performs far better in terms of running time compared to RF. RF is a versatile algorithm often used in predictive modeling tasks. We used RF to predict outcomes based on historical data. However, RF may not capture temporal dependencies and patterns as effectively as time-based models. It performed reasonably well but did not outperform other models in capturing time-related factors. LSTM is a deep-learning model designed for sequential data. We used LSTM to capture intricate time dependencies for effective time-series forecasting.

In lagged (time-based) versions of data, we considered different time windows (1 day, 3 days, 7 days, 15 days, and 30 days look-back) and allowed for lookahead predictions (1 day, 3 days, and 7 days lookup) involved incorporating historical data. However, in order not to present an overwhelming table, we provided only some combinations of look-back and lookup parameter space.

When we compare conventional and time-based versions, we observe a clear improvement in performance with the time-based version regardless of the algorithm. The best-performing configuration for time-based versions belongs to LSTM. Particularly, the 3-day look-back and 1-day lookup likely excelled because the combination effectively captures recent trends while considering the influence of the previous days. This configuration balanced short-term adaptability with longer-term patterns, providing the most accurate results for almost all target variables. We observed better results with a 15-day look-back only for the stress target. We interpret it as the intriguing aspect of the data; it may represent the stressful period of work while collecting the dataset. However, we are not fully sure about this since we do not have ground-truth recordings for this issue, or in other words, the daily lives of the participants.

Even though LSTM was found to be the best-performing one, its results are very close to XGBoost, and its results are even better in some cases (e.g., 15 look-back-

1 lookup). However, XGBoost's execution time is shorter than LSTMs in all cases. Furthermore, the data ranges for positive and negative affect are different and wider compared to stress and anxiety. We observe higher error values for them.

To the best of our knowledge, this is the first study that employs different well-being factors at the same time and also focuses on their time-based results in the scope of a prediction problem. Nevertheless, to compare our findings with the literature where time-based techniques were utilized in other well-being datasets [166], we see that they employed daily and weekly resolutions. They employed different models such as LogR, DT, and ADA on different kinds of features, yet they also concluded an increase in performances from daily to weekly transition. More precisely, for the stress prediction study case, they achieved a nearly 59% and 56% F-score on their data and the StudentLife dataset correspondingly. These performances increased to 72% with their data and 61% in the case of StudentLife. In another paper [74], where the same dataset is used, the authors focused on creating a benchmark for predictive analysis by integrating several aspects of an individual's physical and psychological behavior, psychological states and traits, and job performance. They employed High Order Networks to include the time aspect of the data while combining the modalities. Although our research deals exclusively with well-being factors, this study may also provide insights into higher-level components. Additionally, in [28] and in Section 5, we focused solely on the stress variable and applied conventional techniques. However, as we were dealing with a classification problem in that study, it would not be appropriate to compare its results with the regression results found in the scope of this section.

To show the differences between predictions (forecasts) and ground-truth target values, we present detailed prediction visualization for XGBoost for 7 days look-back predictions in detail in Figure 6.6, 6.7, 6.8, 6.9 for stress, anxiety, positive affect, and negative affect. In these figures, the commonly observed pattern is that the predicted values are lower than the original values. Thus, models under-predict the values, but the patterns are being captured.

In conclusion, we found that time-based versions of the applied algorithms perform better than conventional ones in this section's scope, which answers our third question (RQ6). The best-performing one is LSTM, with a 3-day look-back and a 1-day lookup for almost all target variables.

6.6. Conclusions

We focused on the prediction of well-being factors, namely stress, anxiety, and positive and negative affect, using the Tesserae dataset collected from office workers. We answered three research questions related to digital biomarkers affecting these factors (RQ4), their alignment with the conventional psychology literature (RQ6), and time-based performances of applied methods (RQ7).

We found similar modality rankings for each target variable. Even though we are working on prediction problems rather than classification, these findings align with the one-factor stress classification results after imputing missing values [28].

In addition, we validated our findings by examining the conventional psychology literature where there is no wearable device usage. We found that device-measured modalities aligned with the paper-based studies. Furthermore, we found that we achieve better prediction performances when we consider the time aspect of the data. In particular, LSTM performs the best compared to other ensemble algorithms.

The inclusion of the time element in our methodology significantly decreases model error, thereby enhancing the predictive performance. Time-based approaches, such as the use of Long Short-Term Memory (LSTM) networks, capture temporal dependencies and patterns that static models might miss. Without incorporating the time aspect, models may fail to recognize the sequential nature of well-being data, leading to higher prediction errors and less reliable results. This temporal awareness allows the models to better understand and predict the fluctuations in stress, anxiety, and affect over time, offering a more accurate picture of an individual's well-being.

This section's findings reveal the potential for these models to play a significant role in various practical applications. These include early interventions and support to monitor and address periods of heightened stress and anxiety. Moreover, these models could facilitate continuous mental health monitoring through wearable devices, offering individuals valuable insights into their well-being. In workplace settings, integrating these models into well-being programs can help manage stress and anxiety levels among employees, contributing to healthier and more productive work environments. At a broader level, our models offer data-driven insights that can inform public health initiatives, educational programs, and policy decisions geared toward addressing mental health challenges. However, it is crucial to emphasize that implementing these models in real-world scenarios requires thorough validation, ethical considerations, and careful integration. Additionally, a more in-depth discussion on the limitations, challenges, and ethical aspects of applying predictive models to mental health prediction would be a valuable complement to this section.

Furthermore, it is important to note that these results are mainly based on the types of people covered in the datasets, namely office workers. The generalizability of these findings to other populations or settings may be limited, and further research is necessary to validate these models across diverse groups.

Table 6.2. Prediction performance in terms of MAE (best results shown in bold).

Type	Method	Stress	Anxiety	Positive Affect	Negative Affect
Conventional	Random Forest (RF)	0.6702	0.6508	3.2955	2.2088
	XGBoost	0.6910	0.6316	3.4269	1.7874
	LSTM	0.6443	0.6445	3.2325	1.5121
Time-Based	RF (1 day look back-1 day lookup)	0.5922	0.4700	2.4881	1.4999
	XGBoost (1 day look back-1 day lookup)	0.6319	0.4745	2.4947	1.4304
	LSTM (1 day look back-1 day lookup)	0.5377	0.3897	2.5075	1.3841
	RF (3 days look back-1 day lookup)	0.5907	0.4628	2.2741	1.3118
	XGBoost (3 days look back-1 day lookup)	0.5952	0.4653	2.1851	1.2466
	LSTM (3 days look back-1 day lookup)	0.5394	**0.3729**	**2.1387**	**1.0574**
	RF (7 days look back-1 day lookup)	0.5808	0.4575	2.3130	1.2717
	XGBoost (7 days look back-1 day lookup)	0.5791	0.4527	2.2623	1.0866
	LSTM (7 days look back-1 day lookup)	0.5111	0.3846	2.1923	**1.0522**
	XGBoost (15 days look back-1 day lookup)	0.6327	0.5736	2.4947	1.2982
	LSTM (15 days look back-1 day lookup)	**0.4718**	0.4752	2.3795	1.3159
	XGBoost (30 days look back-1 day lookup)	0.6518	0.6025	2.1583	1.3630
	LSTM (30 days look back-1 day lookup)	0.5004	0.4804	2.3929	1.4011
	RF (3 days look back-3 days lookup)	0.5950	0.5239	2.5205	1.3992
	XGBoost (3 days look back-3 days lookup)	0.6078	0.5356	2.3898	1.3983
	LSTM (3 days look back-3 days lookup)	0.5547	0.4600	2.3425	1.2564
	RF (7 days look back-3 days lookup)	0.6324	0.5605	2.6482	1.4781
	XGBoost (7 days look back-3 days lookup)	0.6122	0.5386	2.5328	1.4341
	LSTM (7 days look back-3 days lookup)	0.5480	0.4805	2.6318	1.2277
	XGBoost (15 days look back-3 days lookup)	0.6371	0.5754	2.3931	1.5495
	LSTM (15 days look back-3 days lookup)	0.5470	0.4931	2.6475	1.3904
	XGBoost (30 days look back-3 days lookup)	0.6542	0.6067	2.4716	1.6104
	LSTM (30 days look back-3 days lookup)	0.5576	0.5046	2.8832	1.3621
	XGBoost (7 days look back-7 days lookup)	0.6203	0.5450	2.5584	1.4378
	LSTM (7 days look back-7 days lookup)	0.5947	0.5023	2.6342	1.3411
	XGBoost (15 days look back-7 days lookup)	0.6396	0.5846	2.5653	1.5649
	LSTM (15 days look back-7 days lookup)	0.6018	0.5322	2.8423	1.5316
	XGBoost (30 days look back-7 days lookup)	0.6548	0.6125	2.5487	1.6158
	LSTM (30 days look back-7 days lookup)	0.5802	0.5106	2.9486	1.4353

7. MULTI-TASK LEARNING FOR MENTAL HEALTH

Mental health is a multifaceted aspect of human well-being, encompassing dimensions such as depression, anxiety, and stress. WHO indicated that in 2019, around 301 million people suffered from anxiety, and 280 million people lived with depression [9] . These two factors are known to be related to mental disorders and may also be referred to as mental health conditions [10] . Stress is another affecting factor for mental and physiological problems [11] .

To measure these three important factors for mental health, psychologists developed assessment techniques based on questionnaires. The most known was developed in 1995 by Lovibond, and Lovibond [50]. That self-report-based 42-item questionnaire has three dimensions: depression, anxiety, and stress (DASS-42). The main contribution of this new scale is to be a combined version of separate questionnaires for each factor. Traditional approaches to mental health assessment often rely on subjective self-reports and clinical interviews, which may not capture the full spectrum of these conditions. Accurate assessment and differentiation of these constructs have been long-standing challenges in the field. In parallel to Lovibond and Lovibond's pioneering work on the DASS-42 instrument for affective symptom assessment, we aim to enhance our understanding of these dimensions using cutting-edge computational techniques and wearable technology.

Recent advancements in wearable technology have revolutionized our capacity to monitor mental health in real-world settings. Wearable devices, including smartwatches and fitness trackers, provide a non-invasive means of collecting continuous data on individuals' physiological and behavioral patterns. This invaluable resource enables us to explore the intricate connections between mental health and wearable-derived

[9] https://www.who.int/observatories/global-observatory-on-health-research-and-development/analyses-and-syntheses/mental-health/global-strategic-direction

[10] https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[11] https://www.apa.org/topics/stress/health

digital biomarkers. Their computational capacity enables them to collect continuous data and even run machine learning models on them to extract meaningful, high-level, human-interpretable information.

Traditional machine learning algorithms, such as logistic regression, decision trees, and ensemble methods, have been extensively utilized for classifying or predicting well-being and mental health using data collected from wearables and questionnaires [208, 209]. However, many studies often overlook the temporal aspect of the data, not considering how past data impacts future mental health levels. Wearables provide a unique opportunity to collect longitudinal data, offering deeper insights into the fluctuations of factors like stress, anxiety, and emotional states over time. Integrating time-based prediction techniques with wearable devices provides a unique opportunity to capture and analyze temporal patterns, facilitating early detection and timely interventions to improve mental well-being. Deep learning architectures, including RNNs, particularly LSTMs, have emerged as powerful tools for addressing time-based prediction problems related to mental health.

Throughout this section, we utilize the NetHealth dataset [18], which was explained in Section 2 and specifically focus on the prediction of students' mental health, collected from undergraduate students at NDU spanning from Fall 2015 to Spring 2019, constituting 8 semesters corresponding to each semester. This rich dataset comprised information from various modalities, including social networks, physical activity, sleep data collected from Fitbit wearable devices, and ground truth data sourced from questionnaires.

In this section, we delve into the domain of multi-task learning, a powerful machine learning paradigm that enables the simultaneous modelling of multiple mental health dimensions, including depression, anxiety, and stress. It enhances predictive models by jointly considering multiple related target variables, leveraging shared information and improving overall performance. By jointly training models for these tasks, we aim to leverage shared information and dependencies among these constructs, poten-

tially leading to improved predictive accuracy and a more comprehensive understanding of mental well-being.

Our investigation encompasses a wide spectrum of predictive techniques. We delve into deep learning methodologies, including LSTM networks, to capture temporal dependencies in mental health data. Additionally, we incorporate conventional prediction algorithms such as RF, and XGBoost to provide a holistic view of mental health assessment. In addition, we consider their performances in the time-based (lagged) version of data.

Our primary objective is to predict stress, anxiety, and depression levels by analyzing the impact of past measurements on mental health. By leveraging wearable technology, multi-task learning, and advanced predictive techniques, we aim to enhance our understanding of the complex interplay between depression, anxiety, and stress. Ultimately, our objective is to improve the mental well-being and overall quality of life for individuals in their daily lives.

In a dataset where time plays a crucial role, we conducted experiments using various prediction algorithms, including RF, XGBoost, and LSTM, both in their conventional and time-based versions. The time-based models consider the temporal aspect of the data by incorporating time lags to assess how previous mental health factors impact future well-being levels. We observed that, despite the interrelated nature of our target values, the application of multi-task methods did not significantly improve baseline performances; the results were very close in both baseline and multi-task scenarios. Conversely, considering the time aspect in the time-based versions further enhanced baseline results. The best performance was achieved with a 1-day look-up while considering a 15-day look-back on RF. However, the performances of the other algorithms were closely aligned. Additionally, we applied multi-tasking on top of the best-performing time-based case. Although there were no significant differences between multi-tasking and baseline scenarios, we observed performance improvements when employing multi-tasking considering the time aspect of the data (time-based ver-

sion). In each case, the performances of the employed base algorithms were similar, with LSTM consistently performing the worst.

This section seeks to address the following research questions; RQ4 considering factors such as stress, anxiety, and depression, RQ6, RQ7, RQ8 stated in Section 1.

7.1. Related Studies on Time-Based and Multi-Task Learning for Mental Health

The pervasive impact of mental illness on emotions, reasoning, and social interactions underscores the urgent need for early detection to enable effective prevention and intervention. Recognizing the potential of machine learning as a tool in predicting mental health, researchers employ this technology to extract valuable insights and develop intelligent systems aimed at enhancing overall well-being [208, 209].

In the landscape of mental health research, the diagnosis of mental health problems necessitates a diverse repertoire of techniques, ranging from traditional interviews and questionnaires to the sophisticated monitoring of physiological signals. Common mental disorders, including schizophrenia, depression, bipolar disorder, anxiety disorders, and post-traumatic stress disorder, reveal themselves through distinct characteristics and symptoms that intricately impact cognition, emotion, and behavior. This multifaceted nature demands advanced methodologies for analysis, where data mining emerges as a pivotal player in knowledge discovery within databases. In conjunction, machine learning, harnessing the power of artificial intelligence to glean insights from experiential data, becomes indispensable. The commonly employed machine learning approaches encompass supervised learning and unsupervised learning, providing researchers with tools to delve into complex patterns and relationships within mental health datasets [208], [210, 211].

The gaps in the literature that have been found draw attention to issues including small sample sizes, poor validation, little deep learning research, and a lack of real-

world testing. Future research directions that have been suggested call for more in-depth investigation of deep learning, the collection of high-quality data, the creation of precise prediction tools, the explanation of models, the use of transfer learning, and adaptable algorithms [211].

In [212], researchers delve into the temporal dimension of health monitoring, specifically exploring the utilization of HRV data from wearable devices for predicting mental and general health outcomes. The research focuses on the feasibility of forecasting complex health measures, including stress, anxiety, depression, and overall well-being, based on HRV data collected over short durations. Employing advanced LSTM networks, a sub-type of RNNs are designed for sequence analysis, the study achieves notable classification accuracies of up to 85%. The emphasis on utilizing short-duration data streams, ranging from two to five minutes, underscores the potential for real-time and efficient health predictions. This research highlights the significance of temporal dimensions in health assessment by demonstrating the capacity to derive relevant health predictions from brief time periods. This maximizes the usefulness of wearable devices for ongoing health monitoring.

In another study [213], researchers introduce a machine learning model utilizing EHRs to predict mental health crises continuously over a 28-day period. The model, developed to address the impracticality of manual review in busy clinical settings, achieves a notable area under the receiver operating characteristic curve of 0.797. A subsequent 6-month study in clinical practice demonstrates the algorithm's value, proving effective in managing caseloads and mitigating crisis risks in 64% of cases. This research is unique in its continuous prediction of a wide range of mental health crises and explores practical applications in real-world clinical scenarios. The global mental health crisis, exacerbated by the COVID-19 pandemic, underscores the need for efficient mental health services, and this study presents a promising step toward leveraging machine learning for proactive mental health care decision-making. The model's performance is evaluated across disorders, age groups, and data availability, revealing consistent effectiveness and highlighting key predictive features such as his-

torical symptom severity, hospital interactions, and patient characteristics. Overall, the study contributes to automating large-scale mental health data review, potentially revolutionizing care decision processes and improving patient outcomes.

Transitioning to passive sensing approaches for mental health detection, Khoo et al. [214] conduct a systematic review highlighting the effectiveness of combining multiple modalities. Their insights into smartphone sensing and mental health symptoms offer recommendations for feature extraction and machine learning methodologies, addressing challenges and opportunities in passive sensing for mental health monitoring. The systematic review explores passive sensing approaches for mental health detection, emphasizing the effectiveness of combining multiple modalities. The paper highlights the correlations between smartphone sensing and mental health symptoms, recommending feature extraction, fusion, and machine learning methodologies. It provides insights into the challenges and opportunities in leveraging passive sensing technologies for mental health monitoring.

Concluding with a survey on mental health monitoring systems by Garcia et al. [215], the paper provides an overview of research in mental health care applications. Their discussion on classification taxonomies and sensing technologies aims to address challenges and opportunities in continuous sensor data collection, supporting advancements in personalized healthcare interventions. Another survey discusses MHMS using sensor data and machine learning algorithms. It provides an overview of research in mental health care applications, classification taxonomies, and sensing technologies. The study aims to address the challenges and opportunities in continuous sensor data collection for mental health monitoring, supporting advancements in personalized healthcare interventions.

The examination of changes in daily routines and physical activity levels among first-generation college students by Wang et al. [22] provides insights into the impact of COVID-19 concerns and gender differences on behavior. Despite limitations such as restricted sample size and reliance on self-reported metrics, the study sheds light on

changing behaviors and mental health resilience during the pandemic, offering valuable implications for college life. The study examines changes in daily routines and physical activity levels among first-generation college students throughout their first year, noting gender differences in activity levels and COVID-19 concerns. It explores behavioral variations between on-campus and off-campus students during the pandemic, shedding light on changing behaviors, resilience, and mental health in college life.

Transitioning to developing a home environment sensor kit for indoor air quality tracking by Wu et al. [216], the study proposes a conceptual framework for organizing human-centric data modalities. It develops a home environment sensor kit for indoor air quality tracking and proposes a conceptual framework for organizing human-centric data modalities. It collects data reflecting participants' mood, sleep, behavior, and living environment, aiming to understand and monitor various health aspects. Small sample size and measuring living environments alongside human-centered sensing data pose methodological challenges, potentially affecting data accuracy and interpretation.

The StudentLife project [68] explored modalities such as conversation frequency, duration, GPS, and Bluetooth usage, finding strong correlations with depression using the PHQ-9 as ground truth. By examining the relationship between depression and sleep, as demonstrated in [49], and considering social factors such as loneliness, education, and employment status, researchers have gained valuable insights into how various factors influence mood [7]. Overall, these studies contribute valuable insights into the intricate interplay between affect, mood, and various aspects of individuals' lives.

In our previous study [29], we explore the prediction of well-being factors, including stress, anxiety, positive affect, and negative affect, using data collected from office workers through wearable devices. The research compares the performance of machine learning algorithms, such as Random Forest, XGBoost, and LSTM, in conventional and time-based versions. The findings highlight the significance of time-based predictions, with LSTM showing the best performance. The study identifies digital biomarkers, in-

cluding personality factors, sleep-related parameters, and activity levels, as influential in predicting well-being. The results suggest practical applications in mental health monitoring, early interventions, and workplace well-being programs.

Moreover, [217] investigates the effects of providing information workers with a one-week forecast of their mental health on their well-being and job performance. While these studies offer valuable insights into the predictive power of mobile sensing data for workplace performance and well-being, there are areas that warrant further exploration. Understanding the mechanisms underlying the relationships between mobile sensing data, job performance, and well-being could inform the development of targeted interventions to support employee well-being and productivity in the workplace.

From the perspective of multi-task learning studies, [23] propose a deep learning model for predicting human moods based on calendar events, aiming to address challenges such as skewed data and overfitting. The model incorporates an event feature extractor using LSTM and introduces multi-task learning to enhance prediction accuracy. The research focuses on the impact of events on mood, particularly calendar events, as they reflect user-specific situations. The proposed model outperforms existing machine learning models, with multi-task learning helping prevent overfitting, especially on days with significant mood fluctuations. The paper suggests potential applications in mental health management and emphasizes the importance of considering external factors in mood prediction models. The evaluation demonstrates improvements in undertraining and prediction accuracy, highlighting the model's effectiveness in capturing variations in mood due to different events. Limitations include the standardized participant profile and the need for further validation with diverse populations and improved data quality control methodologies.

In [24], researchers introduce a MTL framework, comparing three formulations: MTL deep neural networks, Multi-Task Multi-Kernel Learning, and a Hierarchical Bayesian model for predicting mood, stress, and health outcomes. The primary focus is on personalization, allowing individuals to possess tailored models while still

leveraging data from the general population. MTL significantly enhances mood prediction performance compared to traditional machine learning models. The accuracy of MTL models, personalized for individual differences, ranges from 78% to 82%, an improvement over the 60-66% accuracy of one-size-fits-all models. The personalized MTL models outperform both STL and MTL models.

To the best of our knowledge, the only study in the literature that utilized both multi-task learning and time-series analysis for mood/well-being prediction is presented in [25]. The authors present a multi-task model trained to learn valence and arousal and show that it is more accurate than independent models. The data comprises self-reported data using an Android application and passive behavioural data from sensors on the phone (accelerometer, microphone, location, text messages, phone calls, etc.). While we follow a similar methodology in this section, we utilize a different dataset, consider different machine learning algorithms besides LSTM and identify the most important digital biomarkers in explaining three aspects of mental health: depression, anxiety and stress.

In the context of this section, we focus on temporal aspects and the integration of MTL techniques on student well-being prediction. Leveraging methods such as RF, XGBoost, and LSTM, we delve into the temporal patterns and dependencies present in the data to predict various well-being indicators, such as depression, anxiety, and stress, based on historical data. Our research simultaneously considers the interplay of these factors within a time-based and MTL framework, introducing lagged versions of both baseline and MTL algorithms that account for diverse time windows, showcasing their potential to capture temporal trends. This section extends the application of MTL to the NetHealth dataset, addressing a gap in current research.

7.2. Methodology

7.2.1. Time-Based Modeling

Time-based modeling in the scope of mental health represents a multifaceted approach that harnesses temporal data and techniques to gain deeper insights into, predict, and manage mental health conditions over time [218]. It is particularly valuable in the context of longitudinal studies, where data is collected over extended periods, enabling the exploration of how mental health issues evolve over time and the identification of influential factors. Time series analysis, focusing on fluctuations in symptoms, mood, and behaviors, aids in pattern recognition and the identification of triggers. Predictive models leverage historical data to forecast future mental health outcomes, potentially paving the way for proactive interventions. These models also facilitate the determination of optimal intervention timings, a crucial aspect of effective mental health care. Wearable technology, capable of monitoring physiological and behavioral markers, can be integrated into these models for real-time monitoring and feedback [219]. By identifying specific triggers, assessing treatment responses, and developing early warning systems, time-based modeling enhances the understanding and management of mental health conditions, ultimately leading to more personalized and effective treatment plans in both traditional and telehealth settings.

7.2.2. Multi-Task Learning

The aim of MTL is to train models for multiple similar tasks simultaneously [220]. Generally, it has been used to reduce overfit and improve model accuracy [221]. For instance, in the context of mental health, such as predicting depression, anxiety and stress, multi-task learning can be valuable strategy. Instead of training separate models for each mental health aspect (depression, anxiety, stress), one can use multi-task learning to build a single model that predicts all three simultaneously. MTL in the context of mental health involves leveraging shared information and data from multiple mental health-related tasks or domains to improve the accuracy, generalization,

and understanding of mental health-related outcomes. This approach recognizes that different aspects of mental health, such as depression, anxiety, and stress, are interconnected and can benefit from joint modeling. For instance, a multi-task learning framework may simultaneously predict and analyze various mental health conditions, taking into account their interdependencies. By jointly considering factors like self-reported symptoms, physiological data, social interactions, and treatment responses, multi-task learning can provide a more holistic view of an individual's mental health status and needs. Additionally, it can help identify shared risk factors or protective factors across different mental health conditions, contributing to a deeper understanding of the complexities of mental health. Ultimately, multi-task learning has the potential to enhance mental health assessment, intervention, and support by leveraging the synergy between multiple mental health-related tasks and domains.

7.2.3. Data Preprocessing

As mentioned, in this section, we utilize the NetHealth dataset which was introduced in Section 2.

7.2.3.1. Target Values. Even though we focus on DAS parameters, our dataset does not explicitly contain DASS values collected in a single questionnaire, as performed for DASS-21 or DASS-42; instead, we have separately collected values for depression, anxiety, and stress. For simplicity, we will use the term 'DAS' to represent all three target variables. In the following, we describe the target values and their characteristics specific to the current dataset.

- Depression: NetHealth contains two different survey responses for depression such as 'CES-D' [12] and 'BDI' [13] . Additionally, we have continuous scores ranging from 0 to 60 for CES-D and from 0 to 63 for BDI, with higher scores indicating greater

[12] https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/depression-scale

[13] https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/beck-depression

depressive symptoms. For CES-D, there is also a grouped version ranging from 0 to 1, representing not depressed and depressed. For BDI, the grouped version ranges from 0 to 3, representing minimal, mild, moderate, and severe levels of depression. Throughout the study, we employed the BDI with its corresponding classification levels.

- Anxiety: Similar to depression values, NetHealth also contains two different survey responses for anxiety: 'STAI' [14] and 'BAI' [15] . We have continuous scores ranging from 20 to 80 for STAI and from 0 to 63 for BAI, where higher scores indicate greater anxiety. STAI has a grouped version ranging from 0 to 1, representing not anxious and anxious, respectively. On the other hand, BAI ranges from 0 to 2, representing low, moderate, and severe anxiety levels. In the scope of this study, we utilized the BAI with its corresponding levels.
- Stress: In NetHealth, we have only 'PSS' [16] . We are only provided with its Likert-scale version ranging from 0 to 4, corresponding to never through very often. Given the absence of alternative options for the stress target, we utilized a class-based scale for the entire study. We made choices regarding the survey and value types for other target variables (raw continuous value or classes) by aligning them with the response type available for stress. This decision was made to ensure coherence among target variables despite the differing class scale ranges.

7.2.3.2. Missing Data. In the context of our study, spanning 8 semesters of data, we encountered challenges as certain questions were not consistently addressed or answered across all semesters. Despite having rich data over the semesters, the absence of commonalities prompted the need for a comprehensive approach to handling missing data.

To manage missing data effectively, we adopted the MICE method [222]. MICE provides a systematic framework for filling in data gaps by generating educated esti-

[14] https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/trait-state

[15] https://en.wikipedia.org/wiki/Beck_Depression_Inventory

[16] https://www.das.nh.gov/wellness/docs/percieved\%20stress\%20scale.pdf

mates based on the available data. The process involves iteratively imputing missing values for each variable using conditional distributions. In simpler terms, MICE breaks down the imputation process into a series of steps, where each variable's missing values are estimated based on the observed values of other variables. This iterative process continues until convergence is achieved, resulting in multiple complete datasets.

Upon examination, we identified that missing data from wearables were relatively minimal. However, substantial amounts of missing values were observed in survey data. For certain variables, such as gender and related survey questions, which were not consistently posed every semester, we addressed the issue by replicating a single collected response across all missing columns for each individual throughout different semesters. On the other hand, we employed MICE to fill in the missing values for missing values related to aspects like sleep, health, and mental health features, ensuring a more holistic dataset for our analysis. At the end of this process, we ended up with 310,967 rows and 74 columns.

7.2.3.3. Data Preprocessing for Time-Based Methods. To be able to encounter time aspect and temporal dependencies within the data, we employed time-lagged version of the data where each data point is associated with the previous data point. Even though, LSTM has its own time-dependency, in our previous study on another dataset [29], we found that LSTM on time-lagged data further improve model performance.

Consider a mental health parameter, like depression, anxiety, or stress (DAS) at a given time 't'. To be able to create lagged version of DAS, we present DAS parameter as follows

$$DAS(t) = [DAS(t-n), DAS(t-(n-1)), \ldots, DAS(t-1), DAS(t)]. \tag{7.1}$$

Here, DAS(t) represents one of the mental health parameter (e.g., depression level) at time 't,' ranging from, for instance, 0 to 4 for severity. DAS(t-n), DAS(t-(n-1)), ..., DAS(t-1), DAS(t) represent individual measurements of the mental health parameter at different time points within a specified time window. The 'n' in this context signifies

the size of the time window, indicating the number of past measurements considered when assessing the mental health parameter.

In this section, we employed various look-back (n) sizes, such as 1, 7, 15, and 30, a semester to consider daily, weekly, fortnightly, and monthly, and semesterly patterns in mental health data.

7.2.4. Models and Evaluation Metrics

We employed traditional RF, XGBoost, LSTM, and their lagged versions as prediction/regression models. Since one of our goals is to show the effect of MTL, we employed multi-task versions of baselines (RF, XGBoost, LSTM) and their lagged versions. Once we found the best-performing look-back and look-up combination with the baseline's time-lagged version, we employed the same combination, which is 1-day look-up and 15-day look-back, for the time-lagged version of the multi-task.

Since all of our tasks are regression-based for MTL, we employed multi-output-based multi-task learning. Multi-task learning and multi-output learning differ primarily because output variables in multi-output learning often share the same training data or features. In contrast, separate tasks may be learned in MTL on different training sets or features. Multi-task supervised learning becomes multi-label learning or multi-output regression when multiple tasks share the training set [223, 224]. Instead of running algorithms over the same data for different task predictions, by employing MTL, we aim to run each method once to measure their performance on different tasks.

The RF algorithm is employed due to its performances in the literature [29], [198]; XGBoost is employed as it is indicated to perform better compared to deep-learning techniques and even better once combined with deep-learning compared to the alone version [199], [213]; and LSTM is employed as it inherently considers the time aspects of data.

We also have their lagged versions for $1, 7, 15, 30, 90, 180$ days look-back to consider the contribution of past data points and $1, 7, 15, 30$ days lookup to understand up to how many days we can reach reasonable errors and which one's performance is better.

We used MAE as an error metric since we did not encounter a sign difference between the target and predicted variable (an example of this visualization can be seen in Figure 7.1). It belongs to the multitask-LSTM-anxiety case's visualization. The X axis represents the indexes corresponding to the data points in the testing dataset. Furthermore, to prevent over-fitting, we split the dataset into 5, 2, and 2 semesters for train, validation, and testing, respectively.

Please note that while the target variables in the original dataset were discrete, the MICE application resulted in obtaining continuous values for missing target entries. Considering our expertise, especially in the context of psychological questionnaires, we refrained from categorizing these values into specific classes. Instead, we chose to utilize models that align with the nature of the dataset, prioritizing prediction over classification.

Additionally, due to the same consideration of utilizing prediction models, there are instances where the predicted values fall outside the range of the original target values. Despite this, we opted not to remove such instances. Our rationale is rooted in the nature of our predictive modeling approach, where the goal is to estimate numerical values without imposing constraints based on predefined classes.

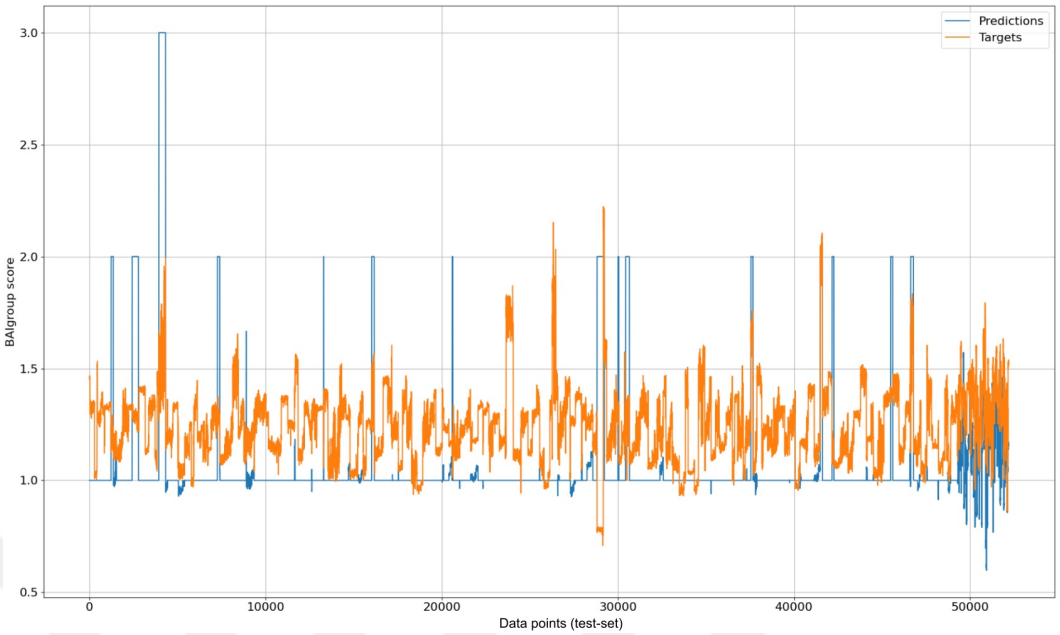

Figure 7.1. Visualization of multi-task results for anxiety over test data.

7.3. Results

7.3.1. Quantification of Digital Biomarkers Related to DAS

We employed SHAP values [225] to assess the influence of each feature on the primary model. Drawing inspiration from game theory's Shapley value, this technique gauges the individual contributions of each feature towards the model's outcomes for interpreting machine learning models.

In our examination of the mental well-being model using SHAP, we assessed the relative impact of the top 20 features based on mean absolute SHAP values for each target variable separately (Figure 7.2, 7.3, 7.4). The key contributors to the model's predictive power included SRQE_introj (introjective self-regulation), SRQE_Ext (external self-regulation), SelfEsteem (on the whole, I am satisfied with myself), PSQI (sleep quality index), Trust (most people can be trusted), selsa (loneliness) along with other education and origin related parameters. These findings highlight the crucial role of these factors in influencing the model's predictions related to mental well-being. Since

the effect of parameters get lower and lower at some point, to facilitate understanding, we only plot the first eleven parameters in Figure 7.5, 7.6, 7.7. By providing SHAP values, we answered our first research question (RQ4).

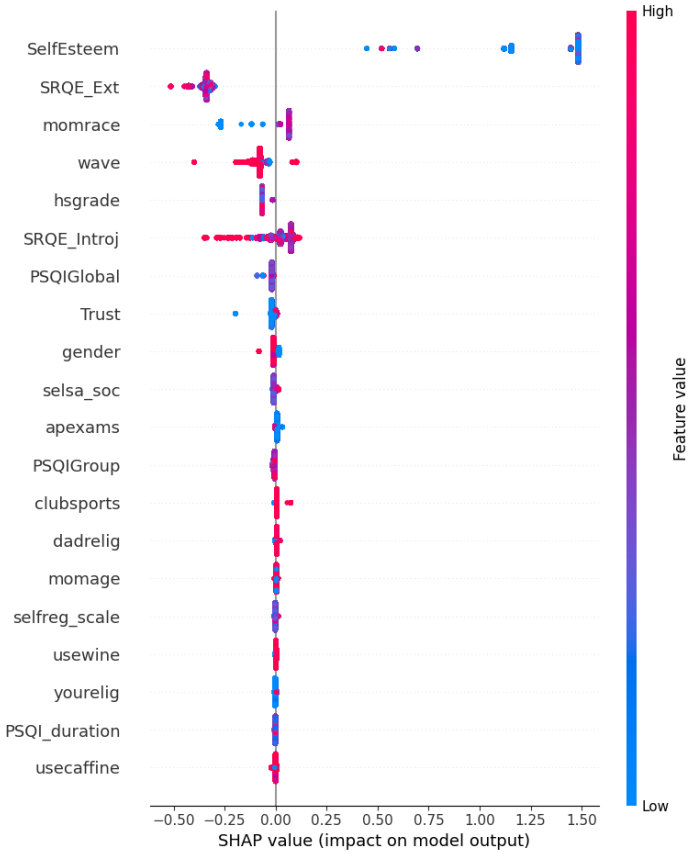

Figure 7.2. Waterfall plot with 20 values for depression.

In Section 6 which concentrated on the well-being of office workers, specifically examining stress, anxiety, and mood (positive and negative affect), our research identified personality factors such as extraversion, agreeableness, conscientiousness, neuroticism, and openness (Big-5 inventory) as among the most crucial contributors [29]. However, within the context of the NetHealth dataset, despite collecting these personality values, none of them emerged among the top 20 results according to SHAP analysis. We attribute this phenomenon to the specificity of the dataset, revealing that other factors beyond personality traits play a much more significant role in influencing outcomes.

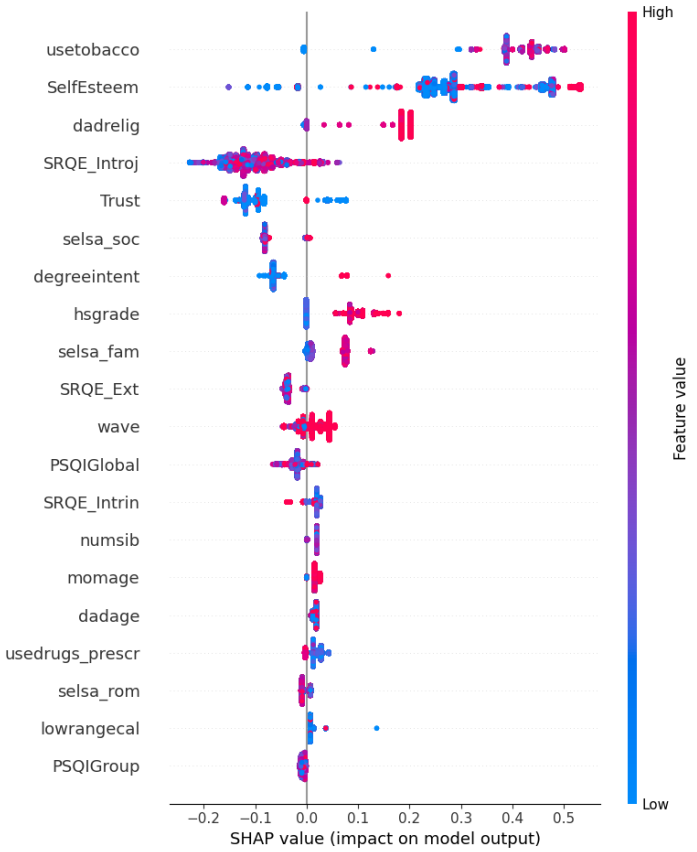

Figure 7.3. Waterfall plot with 20 values for anxiety.

Furthermore, in [226], researchers focused on adolescents' self-regulation and mental health relation. Engaging in exploration, formulation, implementation, and assessment is found to be positively associated with the mental health of adolescent boys and girls. This suggests that individuals who actively explore various options in their environment, consider alternatives, and maintain focus on tasks instead of getting distracted tend to evaluate events more positively. Consequently, this approach is linked to lower cognitive distortion, enhancing psychological functioning and higher self-esteem. These results correlate with our findings about the "Self Esteem" parameter as well (in stress SHAP results).

Different sleep sub-parameters, such as overall quality value in terms of the continuous value between 0-21 or its classified version under several classes indicating sleep quality or sleep duration, are observed in all of the target values' SHAP parameters.

It aligns with the literature. In [227], researchers found that greater improvements in sleep quality lead to more significant improvements in mental health. The discussion emphasizes the trans-diagnostic nature of sleep as a treatment target, noting that the benefits extend across various populations, whether individuals have clinically defined mental health difficulties or non-clinical experiences. Additionally, targeting sleep is seen as a preventive measure, potentially limiting the risk of developing or exacerbating mental health difficulties in the future. This finding also aligns with our previous study on office workers [29], where sleep-related parameters were found commonly among all of the target's most important parameters.

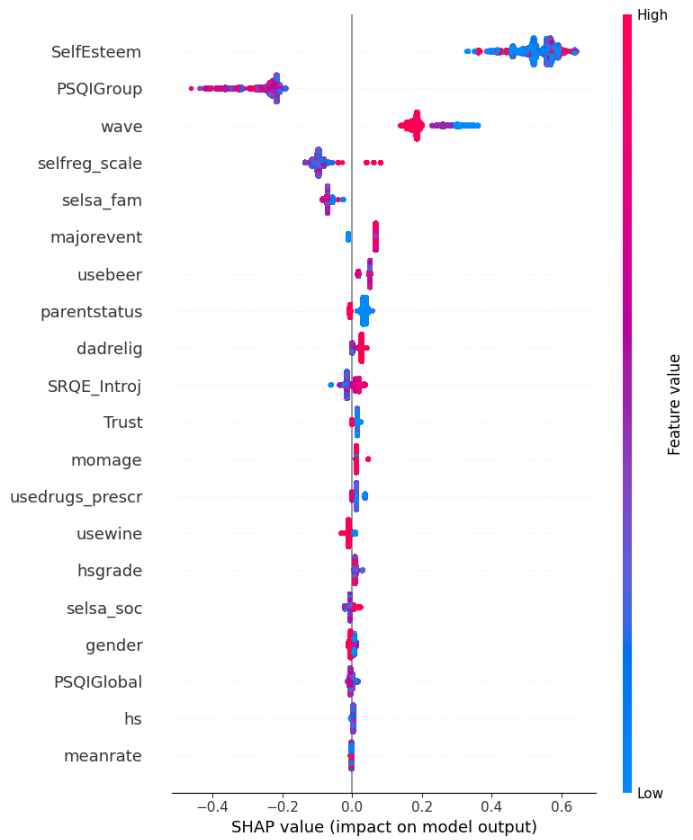

Figure 7.4. Waterfall plot with 20 values for stress.

Among the common parameters of all target variables, we found loneliness and its derivatives, such as romantic, family or social loneliness. In addition to loneliness, the trust parameter, which indicates the perspective of the participant over the

other people, whether they are trustable or not, is observed in the depression and stress target values SHAP list. The study [228] investigated how loneliness and trust issues, both in personal relationships and institutions, impact stress among expatriates during the early COVID-19 pandemic. It found a direct relationship between perceived stress, loneliness, and trust levels. Low trust, especially in interpersonal relationships, led to increased loneliness. Conversely, higher interpersonal trust was linked to lower perceived stress. Institutional trust had a more pronounced impact on stress than loneliness. The study highlighted the evolving trust dynamics during the pandemic, including strained relationships with academic institutions. Overall, the findings emphasized the interconnectedness of trust, loneliness, and mental well-being among expatriates during challenging times.

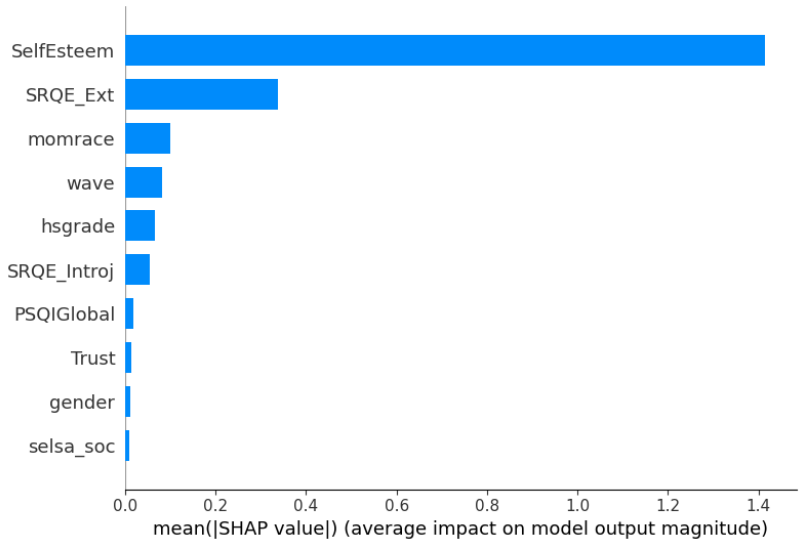

Figure 7.5. Bar plot with 10 values for depression.

In our SHAP values analysis, we observed notable contributions of “wave” parameter underscoring the importance of semester-specific data and the temporal dimension in predicting target values. We can validate temporal importance with time-based algorithm performance improvements in Section 7.3.2. It is important to note that the current section omits personalization, leaving this aspect for consideration in future research studies. In this way, we answered our second research question (RQ7).

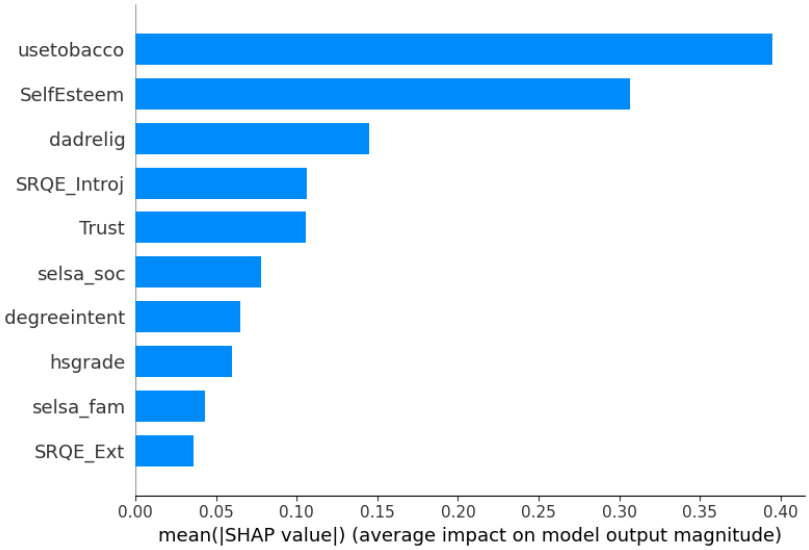

Figure 7.6. Bar plot with 10 values for anxiety.

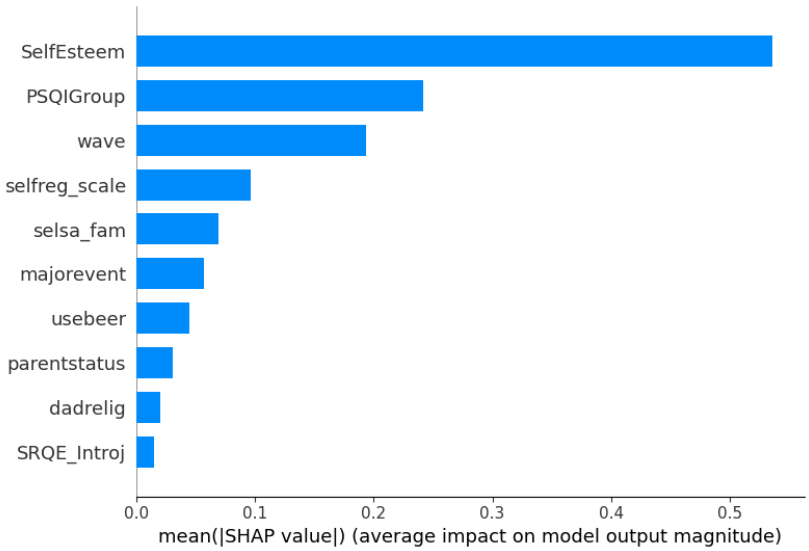

Figure 7.7. Bar plot with 10 values for stress.

In Figure 7.8, 7.9, 7.10, we provide modality rankings based on the SHAP rankings of each target variable. These rankings highlight the importance of various digital biomarkers in assessing different mental health conditions. The consistent patterns observed in these rankings for Depression, Anxiety, and Stress suggest the transitivity of these modalities across different target factors, indicating that similar types of data can be relevant for multiple aspects of mental health.

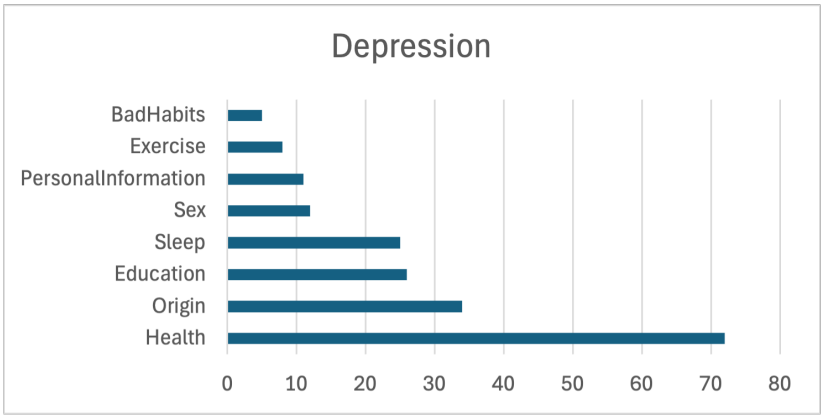

Figure 7.8. Depression modality.

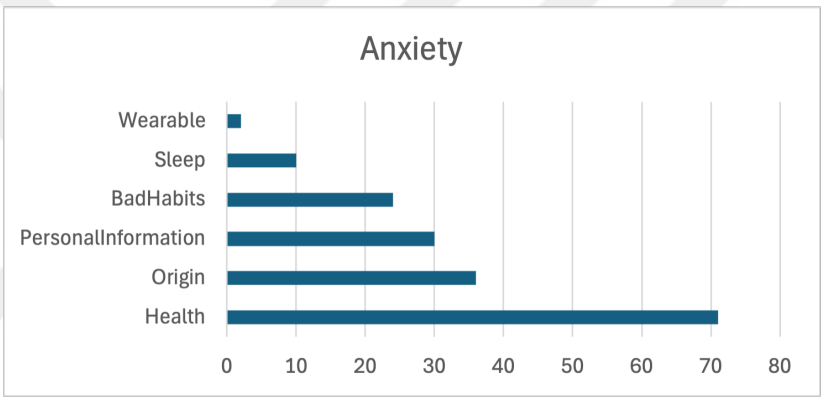

Figure 7.9. Anxiety modality.

7.3.1.1. Wearable-Only Digital Biomarkers for DAS. Our study examines the impact of survey questions and wearables on model performance in predicting DAS. While wearables are not identified as the most crucial factors in SHAP analyses, their specific ranking is explored by excluding survey parameters in this section. The model performance results indicate a slight increase in errors compared to including all parameters, suggesting that survey parameters contribute marginally more to model performance. Despite wearables being positioned towards the end of the parameter space, their influence is overshadowed by the minor contributions of other survey parameters. Detailed parameter contributions exclusively focusing on wearables are visually presented in Figure 7.11, Figure 7.12, and Figure 7.13 illustrating the role of wearable-related factors in the model predictions for depression, anxiety and stress respectively.

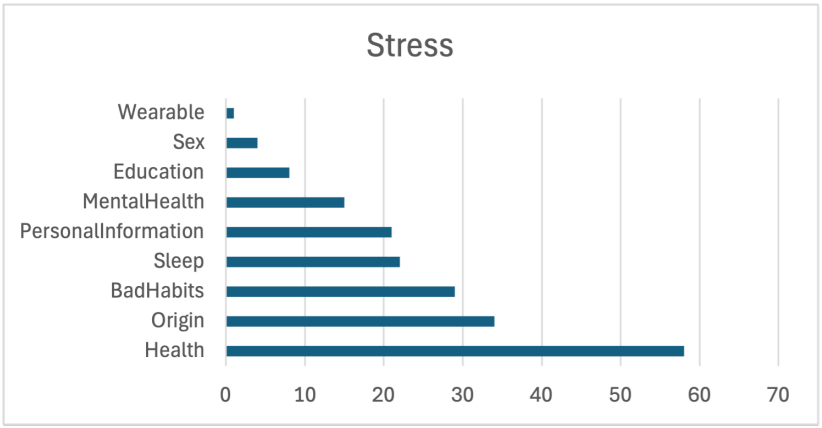

Figure 7.10. Stress modality.

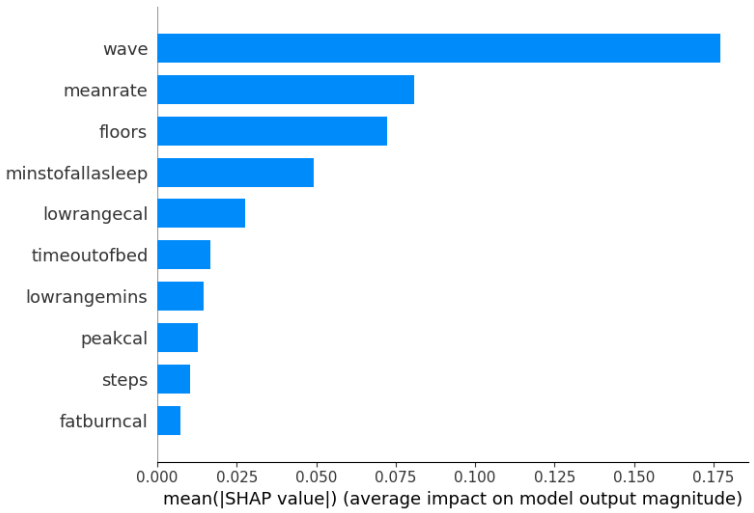

Figure 7.11. Depression.

7.3.2. Comparison of Prediction Performance: Conventional vs. Multi-Task vs. Time-Based vs. Multi-Task Time-Based DAS

We are dealing with three target variables: depression on a scale of 0 to 3, anxiety on a scale of 0 to 2, and stress on a scale of 0 to 4. We aim to predict these target values using RF, XGBoost, and LSTM. Additionally, we aim to empirically assess the impact of a lagged dataset on performance and investigate whether multi-tasking alone or in combination with the lagged version influences outcomes (RQ8, RQ9).

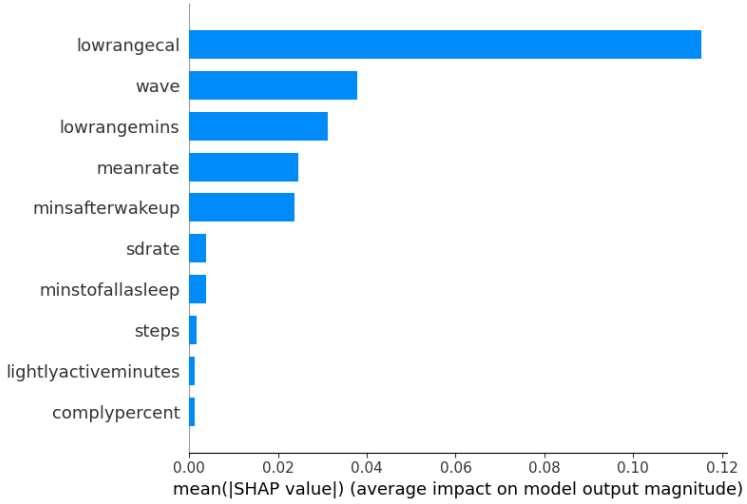

Figure 7.12. Anxiety.

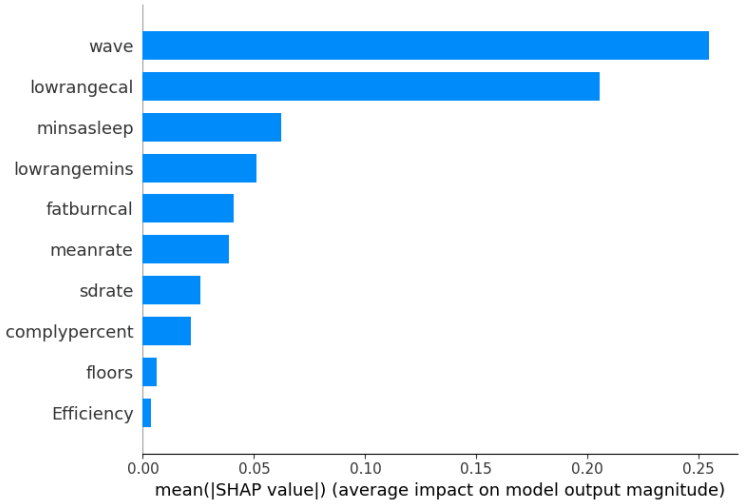

Figure 7.13. Stress.

The comprehensive outcomes are outlined in Table 7.1, with additional insights into time-based performances accessible in Table 7.2. Despite their minimal impact on the ultimate performance enhancement, these details afford a broad perspective on the consequences of diverse look-back and look-up combinations.

The baseline and MTL performances are closely aligned, with MTL showing no significant improvement. RF and XGBoost results also exhibit a minimal difference in both scenarios. Notably, incorporating the temporal aspect has led to substantial performance enhancements, with the highest outcome achieved using RF with a 15-

day look-back and 1-day lookup. While Table 7.1 outlines the time-based results for a 1-day lookup, Table 7.2 encompasses various combinations of look-backs and look-ups. In addition, when considering a longer look-back period in the time-based method, such as comparing 15 days to 1 month or 1 month to 1 semester, despite the increased computational and time burden, we did not observe any performance improvement. In the pursuit to refine time-based performances, we implemented the multi-task version on top of time-based methods. This strategy yielded improved outcomes for depression and stress target variables, although it did not enhance the anxiety target performances. RF and XGBoost performances remained closely aligned. An interesting observation was the consistently poorer performances with LSTM across all cases compared to others, though the difference is not high. The simple one-layer LSTM may explain why it underperformed, despite handling time aspects and improving with lagged data.

A closer look at Table 7.2 reveals a general trend of increasing error values (indicating worse performance) from top to bottom as we compare each algorithm with its counterpart in different combinations. However, this trend is not consistently observed in all cases. For instance, the performance of LSTM in the 7-day look-up and 1-semester look-back combination is improved when employing the 15-day look-up and 15-day look-back combination. Therefore, a consistent pattern that applies to all cases and algorithms cannot be conclusively determined. Interestingly, longer look-back periods did not result in performance improvements. Implementing MTL on time-based methods improved outcomes for depression and stress but did not enhance anxiety target performances. LSTM consistently exhibited poorer performances across all cases compared to RF and XGBoost. Analyzing Table 7.2 reveals an overall trend of increasing error values from top to bottom, which is inconsistent across all cases and algorithms. For example, the LSTM performance improved in the 7-day look-up and 1-semester look-back combination when employing the 15-day look-up and 15-day look-back combination, suggesting a lack of a universal pattern.

To the best of our knowledge, our study is the first to incorporate the time-based aspect of data among the studies utilizing NetHealth.

Table 7.1. Model prediction performances for DAS, presented as MAE, with best performances in bold.

Type		Method	Depression	Anxiety	Stress
Baseline		RF	1.1804	**0.1277**	0.2742
		XGBoost	**1.1537**	0.1588	**0.2722**
		LSTM	1.7342	0.1487	0.5199
Multi-task		RF	**1.1600**	0.1348	**0.2794**
		XGBoost	1.1771	**0.1347**	0.2798
		LSTM	1.7181	0.2584	0.6029
Time-based baseline (1 day look-up)	1 day look back	RF	0.0198	0.0047	0.0314
		XGBoost	0.0433	0.0060	0.0264
		LSTM	0.0379	0.0221	0.0653
	7 days look back	RF	0.0198	0.0036	0.0230
		XGBoost	0.0361	0.0060	0.0278
		LSTM	0.0373	0.0178	0.0646
	15 days look back	RF	**0.0194**	**0.0047**	**0.0228**
		XGBoost	0.0399	0.0072	0.0289
		LSTM	0.0424	0.0134	0.0703
	1 month look back	RF	0.0198	0.0047	0.0233
		XGBoost	0.0433	0.0059	0.0292
		LSTM	0.0462	0.0220	0.0741
	1 semester look back	RF	0.0234	0.0048	**0.0220**
		XGBoost	0.0371	0.0059	0.0270
		LSTM	0.0738	0.0286	0.0792
Time-based multi-task		RF	**0.0049**	0.0131	0.0166
		XGBoost	0.0070	**0.0083**	**0.0091**
		LSTM	0.0319	0.0352	0.0381

Unlike studies solely focusing on one target variable, we selected three related variables appropriately termed DAS in the literature. Furthermore, in addition to the time-based improvement, we utilized MTL. Since all three tasks involve regression,

we approached multi-tasking as a multi-output regression problem. This approach allowed us to fairly compare algorithm performances only while concentrating on their type changes rather than their architectures.

Table 7.2. Further results of the time-based predictions.

Type	Time	Details	Method	Depression	Anxiety	Stress
Time-based	7 days lookup	7 days look back	RF	0.0528	0.0105	0.0436
			XGBoost	0.1324	0.0190	0.0455
			LSTM	0.0392	0.0109	0.0704
		15 days look back	RF	0.0563	0.0114	0.0397
			XGBoost	0.1339	0.0200	0.0426
			LSTM	0.0492	0.0188	0.0739
		1 month look back	RF	0.0535	0.0097	0.0442
			XGBoost	0.1298	0.0187	0.0448
			LSTM	0.0389	0.0150	0.0721
		1 semester look back	RF	0.0774	0.0097	0.0407
			XGBoost	0.1305	0.0182	0.0426
			LSTM	0.0993	0.0322	0.1508
	15 days lookup	15 days look back	RF	0.1082	0.0192	0.0618
			XGBoost	0.2426	0.0368	0.0648
			LSTM	0.0905	0.0142	0.0759
		1 month look back	RF	0.1091	0.0185	0.0662
			XGBoost	0.2409	0.0359	0.0691
			LSTM	0.0608	0.0281	0.1524
		1 semester look back	RF	0.1430	0.0196	0.0670
			XGBoost	0.2394	0.0355	0.0641
			LSTM	0.1265	0.0353	0.1074
	1 month lookup	1 month look back	RF	0.1955	0.0314	0.1055
			XGBoost	0.4281	0.0644	0.1009
			LSTM	0.0838	0.0361	0.0794
		1 semester look back	RF	0.2581	0.0322	0.1113
			XGBoost	0.4279	0.0640	0.0974
			LSTM	0.2443	0.0586	0.1596

7.4. Conclusion

We focused on predicting mental well-being factors, including depression, stress, and anxiety, using the NetHealth dataset from college students. Addressing four research questions, we explored digital biomarkers' impact on these factors (RQ4), their alignment with conventional psychology literature (RQ6), the time-based performance of applied methods (RQ7), and since the target variables are correlated, we wanted to explore further performance improvements while implementing specific methods designed for this purpose such as MTL (RQ8).

Our findings revealed interesting modality rankings for each target variable, aligning with conventional psychology literature for each parameter. Additionally, we validated our results by comparing them with paper-based studies that lacked wearable device usage. We ended up with no parameter coming from wearable on the top 20 variables. That means their contributions to the final model performance are lower than the presented survey parameters in the SHAP figures. Device-measured modalities aligned with traditional studies, and we observed improved prediction performances when considering the temporal aspect of the data. We further improved by employing the multi-tasking technique while considering the time aspect of data.

The updated analysis of mental health multi-task results indicates that baseline and MTL performances align closely, with no significant improvement observed in the multi-task scenario. RF and XGBoost results exhibit minimal differences, and substantial performance enhancements are achieved by incorporating the temporal aspect, particularly with RF using a 15-day look-back and 1-day lookup.

It is worth noting that our findings are primarily derived from data collected from college students. Consequently, the applicability of these results to other demographics or environments might be restricted. Further research is needed to confirm the validity of these models across a wider range of populations.

8. CONCLUSION

This thesis comprises a diverse set of methodologies to advance the understanding of well-being through the lens of digital biomarkers derived from wearables and a wide range of questionnaires. Our exploration, initiated with the motivation detailed in the introductory section, has systematically addressed specific objectives that contribute significantly to the existing state-of-the-art knowledge.

In Section 2, we started by highlighting the necessary background information, describing wearables' function in health monitoring, and identifying digital biomarkers. These core ideas acted as a roadmap for our further research. Exploring Tesserae and launching the NetHealth datasets provided the baseline information regarding the datasets' specificity. In Section 3, we provided a wide range of related studies specifically related to the context of the thesis and highlighted the differences from the state-of-the-art studies in the discussion section. Section 4, centered on the NetHealth dataset, initiated the set of methodologies by delving into the factors affecting academic performance and sleep quality among college students. By predicting academic grades and understanding the relationship between wearable data and sleep patterns, we listed related digital biomarkers such as *sleep, big five, health, mental health* for the target variables. We employed different algorithms for academic grade performance classification and CNN architectures for sleep quality prediction.

Section 5 shifted the focus to stress among workers, employing the Tesserae dataset. Here, we aimed to extract novel stress-associated digital biomarkers utilizing diverse parameters from mobile phones and wearables. The Random Forest algorithm allowed us to assess the impact of essential parameters on stress prediction. Furthermore, we reported the importance of personality traits on model performances. Due to the class imbalance issues, we employed SMOTE and achieved the better performance in the current state-of-the-art literature over stress classification on this dataset.

Time-based well-being prediction was targeted in Section 6, where we explored temporal patterns to enhance our understanding of how well-being evolves over time. Leveraging baseline methods such as RF, XGBoost, and LSTM, we further improved their performance with time-lagged dataset versions. Similar to the stress classification discussed in Section 5, personality factors, i.e. extraversion, agreeableness, conscientiousness, neuroticism, and openness consistently emerged as among the most influential factors (Figure 6.1, 6.2, 6.3, 6.4). These findings were consistent across all target variables considered in our regression analysis. Furthermore, attributes such as sleep-related parameters (bedtime and wakeup time), number of stress samples, highly active seconds, activity stress seconds, and phone unlock duration (derived from phone activity) were consistently associated with various target factors, including stress, anxiety, and negative affect.

In Section 7, we explored the intricate relationships among various mental health factors using multi-task learning. We examined digital biomarkers for each target variable and compared the results with baseline methods and their time-based counterparts. Our findings suggest that while multi-task learning did not outperform baseline methods overall, it did yield significant enhancements when combined with time-lagged data. Specifically, implementing the RF model with a 15-day look-back and 1-day lookup showed substantial predictive improvements, underscoring the importance of temporal considerations in mental health analytics.

As we conclude this thesis, the collective findings from these sections contribute to a well-defined understanding of well-being, encompassing academic performance, sleep quality, stress, and various mental health factors (depression, anxiety, positive and negative affect). This incremental approach, combining diverse datasets, advanced methodologies, and predictive models, provides a foundation for future research in the dynamic intersection of wearables, digital biomarkers, and well-being studies. It underscores that while standalone wearable data or biased questionnaires may not suffice alone, their combination proves effective in advancing our understanding and applications. This synthesis positions our study to guide and inspire forthcoming inves-

tigations aimed at unraveling the complexities of human well-being through innovative, interdisciplinary approaches.

8.1. Future Work and Limitations

Regarding Section 4 and Section 4.1, exploring other semesters within the dataset, given its eight-semester structure, could enhance the availability of instances for each class. Moreover, considering that the dataset originates from a prestigious university, where higher grades such as A and A+ are more prevalent, extending the experimental setup to include students with lower academic performance could provide valuable insights. Such an expansion of the section's scope would contribute to a more comprehensive understanding of the factors influencing academic outcomes across diverse performance levels.

We acknowledge that the sleep section (Section 4.2) primarily addressed a prediction problem with a relatively narrow target value range (between 0.93 and 0.95), which showed little variation. Future studies could explore prediction challenges with greater variability in target values. Moreover, investigating individual differences and analyzing changes over specific periods, such as exam dates and holidays, could offer deeper insights into the dynamics of sleep efficiency. These avenues present exciting opportunities for further research in this field.

In future research, integrating additional data sources could provide a more comprehensive understanding of the long-term health impacts of sleep. Such an approach would facilitate a more detailed exploration of how various factors influence sleep patterns and health outcomes, potentially revealing insights that our initial analysis did not capture. By expanding the scope of data collection and analysis, future studies can contribute to a more holistic understanding of sleep's role in overall health and well-being.

Regarding Section 5, we excluded some of the collected wearable and mobile phone data due to the lack of associated ground truth. This exclusion encompasses a substantial amount of data, spanning almost a year for some participants. One potential approach to address this issue is to impute ground truth data corresponding to these records for further analysis. Additionally, we did not account for time-related aspects in this section, such as potential variations in stress levels according to different days of the week. It is commonly expected that stress levels may decrease during weekends and holidays. Exploring these temporal patterns could provide valuable insights. Also, for the causality analysis, we did not include confounding factors. Including proved confounding factors from psychology literature might provide positive effect on the relevance of the causality results and their interpretations.

Furthermore, future research could investigate the application of different deep learning methods, such as CNNs and LSTMs. These methodologies could potentially enhance the modelling of temporal dependencies and spatial features within stress data. Moreover, instead of employing SMOTE for class balancing in Section 4 and Section 5, alternative methods like transformer networks or self-supervised learning could be explored to further improve the robustness and effectiveness of the models.

In conclusion, future studies could enrich our findings by addressing missing ground truth data, exploring temporal variations in stress levels, and leveraging advanced deep learning techniques. These efforts would build upon our foundational work in digital health research and contribute to a more comprehensive understanding of stress dynamics.

In Section 6, while we have diligently worked on managing the vast amount of available data points by excluding missing ones, the potential benefit of imputation for enhancing precision and performance remains to be explored. Although we implemented person-based splitting during the train-test phase, our analysis did not incorporate a personalized approach, which is a priority for future studies.

It is crucial to note that the results presented in this section are based on predictions derived from a test set containing data from individual participants. However, our research objectives primarily focus on extracting aggregate insights rather than providing personalized predictions. This distinction influences how we interpret and utilize our findings. Despite our test set consisting of individual-level data, our main goal is to identify overarching patterns and trends in the data that hold across different individuals. This approach allows us to draw broader conclusions about the dataset and evaluate the model's effectiveness in capturing general trends. While individual-level predictions are valuable in specific contexts, this section contributes insights into collective behaviors and trends relevant to the predictive task under investigation.

In future work, we plan to explore the implementation of imputation techniques to handle missing data, which could enhance the precision and performance of our models. Additionally, incorporating personalized approaches will be a key focus to provide more tailored predictions for individual participants. Another promising avenue for future research is the application of recent deep learning-based time-series analysis methods, such as Lag-Llma. These advanced techniques have shown great potential in capturing complex temporal patterns and dependencies, and their integration could significantly improve the accuracy and robustness of our predictive models. By leveraging these cutting-edge methods, we aim to refine our models further and expand their applicability across diverse populations and settings.

In Section 7, personalization was not explored within the scope of that section, either, leaving it as a consideration for future studies. Nevertheless, significant challenges persist, such as the necessity for robust validation methodologies and ethical concerns surrounding data privacy and the interpretation of predictive models in mental health contexts [30]. Future research directions may involve implementing personalized techniques to enhance predictive accuracy, aligning with existing literature advocating for tailored approaches in mental health monitoring [28], [30].

REFERENCES

1. Castillejo, P., J.-F. Martinez, J. Rodriguez-Molina and A. Cuerva, "Integration of Wearable Devices in a Wireless Sensor Network for an E-health Application", *IEEE Wireless Communications*, Vol. 20, No. 4, pp. 38–49, 2013.

2. Pantelopoulos, A. and N. G. Bourbakis, "A Survey on Wearable Sensor-based Systems for Health Monitoring and Prognosis", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 40, No. 1, pp. 1–12, 2010.

3. Zhang, X., W. Kou, I. Eric, C. Chang, H. Gao, Y. Fan and Y. Xu, "Sleep Stage Classification Based on Multi-level Feature Learning and Recurrent Neural Networks via Wearable Device", *Computers in Biology and Medicine*, Vol. 103, pp. 71–81, 2018.

4. Li, B. and A. Sano, "Extraction and Interpretation of Deep Autoencoder-based Temporal Features from Wearables for Forecasting Personalized Mood, Health, and Stress", *Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 4, No. 2, pp. 1–26, 2020.

5. Centers for Disease Control and Prevention, "Health-Related Quality of Life (HRQOL)", `https://www.cdc.gov/hrqol/wellbeing.htm`, accessed on June 18, 2021.

6. 8fit Team, "Definition of Well-being", `https://8fit.com/lifestyle/the-definition-of-wellbeing-what-it-is-and-why-its-important/`, accessed on June 19, 2021.

7. Bai, J., L. Shen, H. Sun and B. Shen, "Physiological Informatics: Collection and Analyses of Data from Wearable Sensors and Smartphone for Healthcare",

Healthcare and Big Data Management, Vol. 1028, pp. 17–37, 2017.

8. Vasudevan, S., A. Saha, M. E. Tarver and B. Patel, "Digital Biomarkers: Convergence of Digital Health Technologies and Biomarkers", *Nature Partner Journals Digital Medicine*, Vol. 5, No. 1, pp. 36–38, 2022.

9. Jaques, N., S. Taylor, A. Sano and R. Picard, "Multi-task, Multi-kernel Learning for Estimating Individual Wellbeing", *Neural Information Processing Systems Workshop on Multimodal Machine Learning*, Vol. 898, pp. 3–13, Montreal, Canada, 2015.

10. Jaques, N., S. Taylor, A. Azaria, A. Ghandeharioun, A. Sano and R. Picard, "Predicting Students' Happiness from Physiology, Phone, Mobility, and Behavioral Data", *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 222–228, Xi'an, China, 2015.

11. Sano, A. and R. W. Picard, "Stress Recognition using Wearable Sensors and Mobile Phones", *Affective Computing and Intelligent Interaction*, pp. 671–676, Geneva, Switzerland, 2013.

12. Sano, A., W. Chen, D. Lopez-Martinez, S. Taylor and R. W. Picard, "Multimodal Ambulatory Sleep Detection using Lstm Recurrent Neural Networks", *IEEE Journal of Biomedical and Health Informatics*, Vol. 23, No. 4, pp. 1607–1617, 2018.

13. Suhara, Y., Y. Xu and A. Pentland, "Deepmood: Forecasting Depressed Mood based on Self-reported Histories via Recurrent Neural Networks", *26th International Conference on World Wide Web*, pp. 715–724, Perth, Australia, 2017.

14. Jaques, N., S. Taylor, A. Sano, R. Picard *et al.*, "Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation", *Workshop on Artificial Intelligence in Affective Computing*, pp. 17–33, Melbourne, Australia, 2017.

15. Wang, F. and A. Preininger, "AI in Health: State of the Art, Challenges, and Future Directions", *Yearbook of Medical Informatics*, Vol. 28, pp. 16–26, 2019.

16. Mattingly, S. M., J. M. Gregg, P. Audia, A. E. Bayraktaroglu, A. T. Campbell, N. V. Chawla, V. Das Swain, M. De Choudhury, S. K. D'Mello, A. K. Dey *et al.*, "The Tesserae Project: Large-scale, Longitudinal, In situ, Multimodal Sensing of Information Workers", *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–8, Glasgow, Scotland, United Kingdom, 2019.

17. Topp, C. W., S. D. Østergaard, S. Søndergaard and P. Bech, "The WHO-5 Well-Being Index: a Systematic Review of the Literature", *Psychotherapy and Psychosomatics*, Vol. 84, No. 3, pp. 167–176, 2015.

18. Purta, R., S. Mattingly, L. Song, O. Lizardo, D. Hachen, C. Poellabauer and A. Striegel, "Experiences Measuring Sleep and Physical Activity Patterns across a Large College Cohort with Fitbits", *Association for Computing Machinery International Symposium on Wearable Computers*, pp. 28—-35, Heidelberg, Germany, 2016.

19. Li, N., H. Zhang, L. Feng, Y. Ding and H. Li, "Analyzing and Identifying Predictable Time Range for Stress Prediction based on Chaos Theory and Deep Learning", *Health Information Science and Systems*, Vol. 12, No. 1, pp. 16–30, 2024.

20. Breideband, T., G. J. Martinez, P. Talkad Sukumar, M. Caruso, S. D'Mello, A. D. Striegel and G. Mark, "Sleep Patterns and Sleep Alignment in Remote Teams during COVID-19", *Association for Computing Machinery on Human-Computer Interaction*, Vol. 6, No. CSCW2, pp. 1–31, 2022.

21. Veena, A. and S. Gowrishankar, "Processing of Healthcare Data to Investigate the Correlations and the Anomalies", *Fourth International Conference on I-SMAC*

(IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 611–617, Palladam, India, 2020.

22. Wang, W., S. Nepal, J. F. Huckins, L. Hernandez, V. Vojdanovski, D. Mack, J. Plomp, A. Pillai, M. Obuchi, A. Dasilva *et al.*, "First-gen Lens: Assessing Mental Health of First-generation Students across Their First Year at College using Mobile Sensing", *Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 6, No. 2, pp. 1–32, 2022.

23. Tateyama, N., R. Fukui and S. Warisawa, "Mood Prediction Based on Calendar Events Using Multi-task Learning", *IEEE Access*, Vol. 10, pp. 79747–79759, 2022.

24. Taylor, S., N. Jaques, E. Nosakhare, A. Sano and R. Picard, "Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health", *IEEE Transactions on Affective Computing*, Vol. 11, No. 2, pp. 200–213, 2017.

25. Spathis, D., S. Servia-Rodriguez, K. Farrahi, C. Mascolo and J. Rentfrow, "Sequence Multi-task Learning to Forecast Mental Well-being from Sparse Self-reported Data", *Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining International Conference*, pp. 2886–2894, New York, NY, USA, 2019.

26. Saylam, B., E. Y. Ekmekci, E. Altunoğlu and O. D. Incel, "Academic Performance Relation with Behavioral Trends and Personal Characteristics: Wearable Device Perspective", *Information Society Multiconference*, pp. 35–39, 2022.

27. Kilic, O., B. Saylam and O. Durmaz Incel, "Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning", *Proceedings of the 8th International Conference on Machine Learning Technologies*, pp. 116–120, Stockholm, Sweden, 2023.

28. Saylam, B. and Ö. Durmaz İncel, "Extracting Digital Biomarkers for Unobtrusive

Stress State Screening from Multimodal Wearable Data", *International Conference on Intelligent Edge Processing in the IoT Era*, pp. 130–151, Online, 2022.

29. Saylam, B. and Ö. D. İncel, "Quantifying Digital Biomarkers for Well-Being: Stress, Anxiety, Positive and Negative Affect via Wearable Devices and Their Time-Based Predictions", *Sensors*, Vol. 23, No. 21, pp. 8987–9005, 2023.

30. Saylam, B. and Ö. D. İncel, "Multitask Learning for Mental Health: Depression, Anxiety, Stress (DAS) Using Wearables", *Diagnostics*, Vol. 14, No. 5, pp. 501–518, 2024.

31. Birenboim, A., M. Dijst, F. E. Scheepers, M. P. Poelman and M. Helbich, "Wearables and Location Tracking Technologies for Mental-state Sensing in Outdoor Environments", *The Professional Geographer*, Vol. 71, No. 3, pp. 449–461, 2019.

32. Hammerla, N. Y., S. Halloran and T. Plötz, "Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables", arXiv preprint arXiv:1604.08880, accessed on June 8, 2018.

33. Lara, O. D. and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors", *Institute of Electrical and Electronics Engineers Communications Surveys & Tutorials*, Vol. 15, No. 3, pp. 1192–1209, 2012.

34. Wang, R., M. S. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, M. Merrill, E. A. Scherer *et al.*, "CrossCheck: Toward Passive Sensing and Detection of Mental Health Changes in People with Schizophrenia", *Association for Computing Machinery International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 886–897, Heidelberg, Germany, 2016.

35. McGinnis, R. S., E. W. McGinnis, J. Hruschak, N. L. Lopez-Duran, K. Fitzgerald, K. L. Rosenblum and M. Muzik, "Rapid Anxiety and Depression Diagnosis in Young Children enabled by Wearable Sensors and Machine Learning", *40th An-*

nual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3983–3986, Honolulu, Hawaii, 2018.

36. Van Til, K., M. G. McInnis and A. Cochran, "A Comparative Study of Engagement in Mobile and Wearable Health Monitoring for Bipolar Disorder", *Bipolar Disorders*, Vol. 22, No. 2, pp. 182–190, 2020.

37. Mohr, D. C., M. Zhang and S. M. Schueller, "Personal Sensing: Understanding Mental Health using Ubiquitous Sensors and Machine Learning", *Annual Review of Clinical Psychology*, Vol. 13, pp. 23–47, 2017.

38. Mahadevan, N., C. Demanuele, H. Zhang, D. Volfson, B. Ho, M. K. Erb and S. Patel, "Development of Digital Biomarkers for Resting Tremor and Bradykinesia using a Wrist-worn Wearable Device", *Nature Partner Journals Digital Medicine*, Vol. 3, No. 1, pp. 5–16, 2020.

39. Perez, M. V., K. W. Mahaffey, H. Hedlin, J. S. Rumsfeld, A. Garcia, T. Ferris, V. Balasubramanian, A. M. Russo, A. Rajmane, L. Cheung *et al.*, "Large-scale Assessment of a Smartwatch to Identify Atrial Fibrillation", *New England Journal of Medicine*, Vol. 381, No. 20, pp. 1909–1917, 2019.

40. Andrzejewski, K. L., A. V. Dowling, D. Stamler, T. J. Felong, D. A. Harris, C. Wong, H. Cai, R. Reilmann, M. A. Little, J. T. Gwin *et al.*, "Wearable Sensors in Huntington Disease: a Pilot Study", *Journal of Huntington's Disease*, Vol. 5, No. 2, pp. 199–206, 2016.

41. Saeb, S., E. G. Lattie, S. M. Schueller, K. P. Kording and D. C. Mohr, "The Relationship between Mobile Phone Location Sensor Data and Depressive Symptom Severity", *Peer-reviewed Scientific Mega Journal*, Vol. 4, pp. e2537–e2551, 2016.

42. Saeb, S., M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording and D. C. Mohr, "Mobile Phone Sensor Correlates of Depressive Symptom Severity in

Daily-life Behavior: an Exploratory Study", *Journal of Medical Internet Research*, Vol. 17, No. 7, pp. e175–e186, 2015.

43. Vrigkas, M., C. Nikou and I. A. Kakadiadis, "Classifying Behavioral Attributes using Conditional Random Fields", *Hellenic Conference on Artificial Intelligence*, pp. 95–104, Ioannina, Greece, 2014.

44. Song, Y., L.-P. Morency and R. Davis, "Multi-view Latent Variable Discriminative Models for Action Recognition", *Institute of Electrical and Electronics Engineers Conference on Computer Vision and Pattern Recognition*, pp. 2120–2127, Rhode Island, USA, 2012.

45. Vrigkas, M., C. Nikou and I. A. Kakadiaris, "A Review of Human Activity Recognition Methods", *Frontiers in Robotics and Artificial Intelligence*, Vol. 2, No. 28, pp. 1–28, 2015.

46. Hao, T., G. Xing and G. Zhou, "iSleep: Unobtrusive Sleep Quality Monitoring using Smartphones", *Association for Computing Machinery Conference on Embedded Networked Sensor Systems*, pp. 4–17, Roma, Italy, 2013.

47. PsychDB, "Polysomnography", `https://www.psychdb.com/neurology/polysomnography`, accessed on Dec 8, 2018.

48. ApexMedical, "ActiGraph", `https://apexmedical.co.nz/products/activity-monitoring/actigraph-activity-measurement-solutions/`, accessed on Dec 7, 2018.

49. Chen, Z., M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury and A. T. Campbell, "Unobtrusive Sleep Monitoring using Smartphones", *7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pp. 145–152, Venice, Italy, 2013.

50. Lovibond, P. F. and S. H. Lovibond, "The Structure of Negative Emotional States:

Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories", *Behaviour Research and Therapy*, Vol. 33, No. 3, pp. 335–343, 1995.

51. Conejero, I., E. Olié, R. Calati, D. Ducasse and P. Courtet, "Psychological Pain, Depression, and Suicide: Recent Evidences and Future Directions", *Current Psychiatry Reports*, Vol. 20, pp. 1–9, 2018.

52. Rochette, L. and C. Vergely, "Hans Selye and the Stress Response: 80 years after his "letter" to the Editor of Nature", *Annales de Cardiologie et d'Angeiologie*, Vol. 66, No. 4, pp. 181–183, 2017.

53. Fink, G., "Stress: Concepts, Definition and History. In Reference Module", *Reference Module in Neuroscience and Biobehavioral Psychology*, Vol. 1, pp. 1–9, 2017.

54. Bakhshian, F., A. Abolghasemi and M. Narimani, "Thought Control Strategies in the Patients with Acute Stress Disorder and PTSD", *Procedia-Social and Behavioral Sciences*, Vol. 84, pp. 929–933, 2013.

55. Cohen, S., D. A. Tyrrell and A. P. Smith, "Psychological Stress and Susceptibility to the Common Cold", *New England Journal of Medicine*, Vol. 325, No. 9, pp. 606–612, 1991.

56. Poirat, L., B. Gaye, M. Perier, F. Thomas, C. Guibout, R. Climie, L. Offredo, M. Tafflet, C. Lemogne, B. Pannier *et al.*, "Perceived Stress is Inversely Related to Ideal Cardiovascular Health: The Paris Prospective Study III", *International Journal of Cardiology*, Vol. 270, pp. 312–318, 2018.

57. Marcatto, F., L. Colautti, F. L. Filon, O. Luis, L. Di Blas, C. Cavallero and D. Ferrante, "Work-related Stress Risk Factors and Health Outcomes in Public Sector Employees", *Safety Science*, Vol. 100, No. 89, pp. 274–278, 2016.

58. Bogomolov, A., B. Lepri, M. Ferron, F. Pianesi and A. Pentland, "Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits", *Association for Computing Machinery International Conference on Multimedia*, pp. 477–486, Orlando Florida, USA, 2014.

59. Bakker, J., M. Pechenizkiy and N. Sidorova, "What's Your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data", *11th International Conference on Data Mining Workshops*, pp. 573–580, Vancouver, Canada, 2011.

60. Lu, H., D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez and T. Choudhury, "Stresssense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones", *Association for Computing Machinery Conference on Ubiquitous Computing*, pp. 351–360, Pittsburgh Pennsylvania, USA, 2012.

61. Samson, C. and A. Koh, "Stress Monitoring and Recent Advancements in Wearable Biosensors", *Frontiers in Bioengineering and Biotechnology*, Vol. 8, pp. 1037–1045, 2020.

62. Seshadri, D. R., R. T. Li, J. E. Voos, J. R. Rowbottom, C. M. Alfes, C. A. Zorman and C. K. Drummond, "Wearable Sensors for Monitoring the Physiological and Biochemical Profile of the Athlete", *Nature Partner Journals Digital Medicine*, Vol. 2, No. 1, pp. 1–16, 2019.

63. Dalmeida, K. M. and G. L. Masala, "HRV Features as Viable Physiological Markers for Stress Detection using Wearable Devices", *Sensors*, Vol. 21, No. 8, pp. 2873–2890, 2021.

64. Can, Y. S., B. Arnrich and C. Ersoy, "Stress Detection in Daily Life Scenarios using Smart Phones and Wearable Sensors: A survey", *Journal of Biomedical Informatics*, Vol. 92, pp. 103139–103160, 2019.

65. Goecks, J., V. Jalili, L. M. Heiser and J. W. Gray, "How Machine Learning will Transform Biomedicine", *Cell*, Vol. 181, No. 1, pp. 92–101, 2020.

66. Knight, A., J. Fouyaxis, G. Jarrad, K. Beski, G. Cho and N. Bidargaddi, "Systems to Harness Digital Footprint to Elucidate and Facilitate Ageing in Place", *IOS Press*, Vol. 246, pp. 91–101, 2018.

67. Kroenke, K., 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.

68. Wang, R., F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev and A. T. Campbell, "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones", *Association for Computing Machinery International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 3–14, Seattle, Washington, USA, 2014.

69. Striegel, A., S. Liu, L. Meng, C. Poellabauer, D. Hachen and O. Lizardo, "Lessons Learned from the Netsense Smartphone Study", *Association for Computing Machinery Special Interest Group on Data Communication Computer Communication Review*, Vol. 43, No. 4, pp. 51–56, 2013.

70. Aharony, N., W. Pan, C. Ip, I. Khayal and A. Pentland, "Social fMRI: Investigating and Shaping Social Mechanisms in the Real World", *Pervasive and Mobile Computing*, Vol. 7, No. 6, pp. 643–659, 2011.

71. Nepal, S., G. J. Martinez, S. Mirjafari, K. Saha, V. D. Swain, X. Xu, P. G. Audia, M. De Choudhury, A. K. Dey, A. Striegel *et al.*, "A Survey of Passive Sensing in the Workplace", arXiv preprint arXiv:2201.03074, accessed on Jan 18, 2020.

72. Lin, S., S. M. Mattingly, X. Han, P. Audia, A. Campbell, V. Das Swain, M. Choudhury, S. D'Mello, A. Dey, G. Gao *et al.*, "Sensing Personality to Predict Job

Performance", *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–9, Glasgow, Scotland, United Kingdom, 2019.

73. Das Swain, V., K. Saha, H. Rajvanshy, A. Sirigiri, J. M. Gregg, S. Lin, G. J. Martinez, S. M. Mattingly, S. Mirjafari, R. Mulukutla *et al.*, "A Multisensor Person-centered approach to Understand the Role of Daily Activities in Job Performance with Organizational Personas", *Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 3, No. 4, pp. 1–27, 2019.

74. Robles-Granda, P., S. Lin, X. Wu, G. J. Martinez, S. M. Mattingly, E. Moskal, A. Striegel, N. V. Chawla, S. D'Mello, J. Gregg *et al.*, "Jointly Predicting Job Performance, Personality, Cognitive Ability, Affect, and Well-being", *IEEE Computational Intelligence Magazine*, Vol. 16, No. 2, pp. 46–61, 2021.

75. Nepal, S., G. J. Martinez, S. Mirjafari, S. Mattingly, V. D. Swain, A. Striegel, P. G. Audia and A. T. Campbell, "Assessing the Impact of Commuting on Workplace Performance using Mobile Sensing", *IEEE Pervasive Computing*, Vol. 20, No. 4, pp. 52–60, 2021.

76. Mirjafari, S., K. Masaba, T. Grover, W. Wang, P. Audia, A. T. Campbell, N. V. Chawla, V. D. Swain, M. D. Choudhury, A. K. Dey *et al.*, "Differentiating Higher and Lower Job Performers in the Workplace using Mobile Sensing", *Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 3, No. 2, pp. 1–24, 2019.

77. Nepal, S., S. Mirjafari, G. J. Martinez, P. Audia, A. Striegel and A. T. Campbell, "Detecting Job Promotion in Information Workers using Mobile Sensing", *Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 4, No. 3, pp. 1–28, 2020.

78. Das Swain, V., M. D. Reddy, K. A. Nies, L. Tay, M. De Choudhury and G. D. Abowd, "Birds of a Feather Clock Together: A Study of Person-Organization Fit Through Latent Activity Routines", *Association for Computing Machinery on Human-Computer Interaction*, Vol. 3, No. CSCW, pp. 1–30, 2019.

79. Vhaduri, S. and C. Poellabauer, "Impact of Different Pre-sleep Phone use Patterns on Sleep Quality", *15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 94–97, Las Vegas, Nevada, USA, 2018.

80. Creswell, J. D., M. J. Tumminia, S. Price, Y. Sefidgar, S. Cohen, Y. Ren, J. Brown, A. K. Dey, J. M. Dutcher, D. Villalba *et al.*, "Nightly Sleep Duration Predicts Grade Point Average in the First Year of College", *Proceedings of the National Academy of Sciences*, Vol. 120, No. 8, pp. e2209123120–e2209123127, 2023.

81. Faust, L., K. Feldman, S. M. Mattingly, D. Hachen and N. V. Chawla, "Deviations from Normal Bed-times are associated with Short-term Increases in Resting Heart Rate", *Nature Partner Journals Digital Medicine*, Vol. 3, No. 1, pp. 39–47, 2020.

82. Mattingly, S. M., T. Grover, G. J. Martinez, T. Aledavood, P. Robles-Granda, K. Nies, A. Striegel and G. Mark, "The Effects of Seasons and Weather on Sleep Patterns Measured through Longitudinal Multimodal Sensing", *Nature Partner Journals Digital Medicine*, Vol. 4, No. 1, pp. 76–90, 2021.

83. Mattingly, S. M., G. Martinez, J. Young, M. K. Cain and A. Striegel, "Snoozing: an Examination of a Common Method of Waking", *Sleep*, Vol. 45, No. 10, pp. zsac184–zsac196, 2022.

84. Martinez, G. J., S. M. Mattingly, J. Young, L. Faust, A. K. Dey, A. T. Campbell, M. De Choudhury, S. Mirjafari, S. K. Nepal, P. Robles-Granda *et al.*, "Improved Sleep Detection through the Fusion of Phone Agent and Wearable Data Streams", *International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 1–6, Austin, TX, USA, 2020.

85. Kılıç, A. C., A. Karakuş and E. Alptekin, "Prediction of University Students' Subjective Well-Being with Sleep and Physical Activity Data using Classification Algorithms", *Procedia Computer Science*, Vol. 207, pp. 2648–2657, 2022.

86. Wang, C., O. Lizardo and D. S. Hachen, "Using Fitbit Data to Examine Factors that Affect Daily Activity Levels of College Students", *Public Library of Science One*, Vol. 16, No. 1, pp. e0244747–e0244766, 2021.

87. Martinez, G. J., T. Grover, S. M. Mattingly, G. Mark, S. D'Mello, T. Aledavood, F. Akbar, P. Robles-Granda and A. Striegel, "Alignment between Heart Rate Variability from Fitness Trackers and Perceived Stress: Perspectives from a Large-scale in situ Longitudinal Study of Information Workers", *JMIR Human Factors*, Vol. 9, No. 3, pp. e33754–e33774, 2022.

88. Faust, L., K. Feldman, S. Lin, S. Mattingly, S. D'Mello and N. V. Chawla, "Examining Response to Negative Life Events through Fitness Tracker Data", *Frontiers in Digital Health*, Vol. 3, pp. 659088–659100, 2021.

89. Das Swain, V., V. Chen, S. Mishra, S. M. Mattingly, G. D. Abowd and M. De Choudhury, "Semantic Gap in Predicting Mental Well-being through Passive Sensing", *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–16, New Orleans, LA, USA, 2022.

90. Bin Morshed, M., K. Saha, M. De Choudhury, G. D. Abowd and T. Plötz, "Measuring Self-esteem with Passive Sensing", *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pp. 363–366, Online, 2020.

91. Faust, L., D. Hachen, O. Lizardo and N. V. Chawla, "Quantifying Subjective Well-being using Trends in Weekend Activity", *IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 123–129, New York City, New York, USA, 2018.

92. Chan, L., V. D. Swain, C. Kelley, K. de Barbaro, G. D. Abowd and L. Wilcox, "Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being", *Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 2, No. 1, pp. 1–20, 2018.

93. Saha, K., L. Chan, K. De Barbaro, G. D. Abowd and M. De Choudhury, "Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments", *Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 1, No. 3, pp. 1–27, 2017.

94. Martinez, G. J., S. M. Mattingly, S. Mirjafari, S. K. Nepal, A. T. Campbell, A. K. Dey and A. D. Striegel, "On the Quality of Real-world Wearable Data in a Longitudinal Study of Information Workers", *International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 1–6, Austin, TX, USA, 2020.

95. Martinez, G. J., S. M. Mattingly, P. Robles-Granda, K. Saha, A. Sirigiri, J. Young, N. Chawla, M. De Choudhury, S. D'Mello, G. Mark *et al.*, "Predicting Participant Compliance with Fitness Tracker Wearing and Ecological Momentary Assessment Protocols in Information Workers: Observational Study", *JMIR mHealth and uHealth*, Vol. 9, No. 11, pp. e22218–e22238, 2021.

96. Faust, L., R. Purta, D. Hachen, A. Striegel, C. Poellabauer, O. Lizardo and N. V. Chawla, "Exploring Compliance: Observations from a Large Scale Fitbit Study", *Proceedings of the 2nd International Workshop on Social Sensing*, pp. 55–60, Pittsburgh, PA, USA, 2017.

97. Faust, L., P. Jiménez, D. Hachen, O. Lizardo, A. Striegel and N. V. Chawla, "Long-term Compliance Habits: What Early Data Tells Us", arXiv preprint arXiv:1804.04256, accessed on Jan 28, 2020.

98. Talkad Sukumar, P., T. Breideband, G. J. Martinez, M. Caruso, S. Rose,

C. Steputis, S. D'Mello, G. Mark and A. Striegel, "Designing an Interactive Visualization System for Monitoring Participant Compliance in a Large-scale, Longitudinal Study", *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–8, Yokohama, Japan, 2021.

99. Faust, L., P. Jiménez-Pazmino, J. K. Holland, O. Lizardo, D. Hachen and N. V. Chawla, "What 30 Days Tells Us about 3 Years: Identifying Early Signs of User Abandonment and Non-adherence", *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pp. 216–224, Trento, Italy, 2019.

100. Sukumar, P. T., G. J. Martinez, T. Grover, G. Mark, S. K. D'Mello, N. V. Chawla, S. M. Mattingly and A. D. Striegel, "Characterizing Exploratory Behaviors on a Personal Visualization Interface using Interaction Logs", *EuroVis 2020-Short Papers*, Vol. 0, pp. 1–5, 2020.

101. Sukumar, P. T., A. Dey, G. Mark, R. Metoyer and A. Striegel, "Triggers and Barriers to Insight Generation in Personal Visualizations", *Graphics Interface*, Montréal, Quebec, Canada, 2022.

102. Lin, S., L. Faust and N. V. Chawla, "Lan: Learning to Augment Noise Tolerance for Self-Report Survey Labels", *International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10, Kassel, Germany, 2021.

103. Karl, R., J. Takeshita, A. Mohammed, A. Striegel and T. Jung, "Cryptogram: Fast Private Calculations of Histograms Over Multiple Users' Inputs", *17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 25–34, Online, 2021.

104. Lin, S., X. Wu and N. V. Chawla, "Motif2vec: Semantic-aware Representation Learning for Wearables' Time Series Data", *8th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10, Porto, Portugal, 2021.

105. Lin, S., X. Wu, G. Martinez and N. V. Chawla, "Filling Missing Values on Wearable-sensory Time Series Data", *Proceedings of the SIAM International Conference on Data Mining*, pp. 46–54, Cincinnati, Ohio, USA, 2020.

106. Han, X., J. Hu and J. Ghosh, "A Novel Control-Variates Approach for Performative Gradient-Based Learners with Missing Data", *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Gold Coast, Australia, 2023.

107. Wu, X., S. Mattingly, S. Mirjafari, C. Huang and N. V. Chawla, "Personalized Imputation on Wearable-sensory Time Series via Knowledge Transfer", *Association for Computing Machinery International Conference on Information & Knowledge Management*, pp. 1625–1634, Online, 2020.

108. Saha, K., M. D. Reddy, V. das Swain, J. M. Gregg, T. Grover, S. Lin, G. J. Martinez, S. M. Mattingly, S. Mirjafari, R. Mulukutla *et al.*, "Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior", *8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 178–184, Cambridge, United Kingdom, 2019.

109. Saha, K., "Computational and Causal Examinations of Well-being in Situated Contexts by Leveraging Social Media and Multi-modal Data", *Conference on Computer Supported Cooperative Work and Social Computing*, pp. 147–152, Online, 2020.

110. Saha, K., P. Gupta, G. Mark, E. Kiciman and M. De Choudhury, "Observer Effect in Social Media Use", *Conference on Human Factors in Computing Systems*, pp. 1–20, Honolulu, USA, 2024.

111. Saha, K. and M. De Choudhury, "Examining Well-being in Situated Contexts with Computational Modeling of Social Media Data", *Mobile Sensing in Psychology: Methods and Applications*, Vol. 0, pp. 215–257, 2023.

112. Boero, R., P. T. Hraber, K. A. Kaufeld, E. A. Moore, E. Romero-Severson, J. J. Ambrosiano, J. L. Whitton and B. H. Sims, *Analysis of Multi-modal Wearable Sensor Data to Characterize Social Groups and Influence in Organizations*, Tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2019.

113. Saha, K., T. Grover, S. M. Mattingly, V. D. Swain, P. Gupta, G. J. Martinez, P. Robles-Granda, G. Mark, A. Striegel and M. De Choudhury, "Person-centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing", *Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 5, No. 1, pp. 1–32, 2021.

114. Meng, L., Y. Hulovatyy, A. Striegel and T. Milenković, "On the Interplay between Individuals' Evolving Interaction Patterns and Traits in Dynamic Multiplex Social Networks", *IEEE Transactions on Network Science and Engineering*, Vol. 3, No. 1, pp. 32–43, 2016.

115. Lin, S., L. Faust, P. Robles-Granda, T. Kajdanowicz and N. V. Chawla, "Social Network Structure is Predictive of Health and Wellness", *Public Library of Science One*, Vol. 14, No. 6, pp. e0217264–e0217280, 2019.

116. Wang, C., O. Lizardo and D. S. Hachen, "Neither Influence Nor Selection: Examining Co-evolution of Political Orientation and Social Networks in the NetSense and NetHealth Studies", *Public Library of Science One*, Vol. 15, No. 5, pp. e0233458–e0233477, 2020.

117. Liu, S., D. Hachen, O. Lizardo, C. Poellabauer, A. Striegel and T. Milenković, "Network Analysis of the NetHealth Data: Exploring Co-evolution of Individuals' Social Network Positions and Physical Activities", *Applied Network Science*, Vol. 3, pp. 1–26, 2018.

118. Liu, S., D. Hachen, O. Lizardo, C. Poellabauer, A. Striegel and T. Milenković, "The Power of Dynamic Social Networks to Predict Individuals' Mental Health",

Pacific Symposium on Biocomputing, pp. 635–646, Fairmont Orchid, Hawaii, USA, 2019.

119. Hachen, D., C. Wang, B. Sepulvado and O. Lizardo, "Generators or Diffusers? Examining Differences in the Dynamic Coupling of Context and Social Ties across Multiple Types of Foci", *Social Networks*, Vol. 77, pp. 151–165, 2022.

120. Sepulvado, B., M. L. Wood, E. Fridmanski, C. Wang, M. J. Chandler, O. Lizardo and D. Hachen, "Predicting Homophily and Social Network Connectivity from Dyadic Behavioral Similarity rajectory clusters", *Social Science Computer Review*, Vol. 40, No. 1, pp. 195–211, 2022.

121. Fridmanski, E., M. L. Wood, O. Lizardo and D. Hachen, "Clustering in a Newly Forming Social Network by Subjective Perceptions of Loneliness", *Journal of American College Health*, Vol. 70, No. 5, pp. 1326–1331, 2022.

122. Nurek, M., R. Michalski, O. Lizardo and M.-A. Rizoiu, "Predicting Relationship Labels and Individual Personality Traits From Telecommunication History in Social Networks Using Hawkes Processes", *IEEE Access*, Vol. 11, pp. 8492–8503, 2023.

123. Boleslaw, B., "Modeling Memory Imprints Induced by Interactions in Social Networks", *Social, Cultural, and Behavioral Modeling: 15th International Conference*, pp. 186–195, Pittsburgh, PA, USA, 2022.

124. Flamino, J. R., *Analyzing and Modeling Human Behavioral Dynamics in Social Networks*, Ph.D. Thesis, Rensselaer Polytechnic Institute, 2021.

125. Wang, C., "Smoking Prevalence, Core/periphery Network Positions, and Peer Influence: Findings from Five Datasets on US Adolescents and Young Adults", *Public Library of Science One*, Vol. 16, No. 3, pp. 1–16, 2021.

126. Booth, B. M., H. Vrzakova, S. M. Mattingly, G. J. Martinez, L. Faust and S. K.

D'Mello, "Toward Robust Stress Prediction in the Age of Wearables: Modeling Perceived Stress in a Longitudinal Study with Information Workers", *IEEE Transactions on Affective Computing*, Vol. 13, No. 4, pp. 2201–2217, 2022.

127. Phan, D.-V., C.-L. Chan and D.-K. Nguyen, "Applying Deep Learning for Prediction Sleep Quality from Wearable Data", *Proceedings of the 4th International Conference on Medical and Health Informatics*, pp. 51–55, Kamakura City, Japan, 2020.

128. Sano, A., A. J. Phillips, Z. Y. Amy, A. W. McHill, S. Taylor, N. Jaques, C. A. Czeisler, E. B. Klerman and R. W. Picard, "Recognizing Academic Performance, Sleep Quality, Stress Level, and Mental Health using Personality Traits, Wearable Sensors and Mobile Phones", *12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 1–6, Cambridge, Massachusetts, USA, 2015.

129. Arora, A., P. Chakraborty and M. Bhatia, "Analysis of Data from Wearable Sensors for Sleep Quality Estimation and Prediction using Deep Learning", *Arabian Journal for Science and Engineering*, Vol. 45, No. 12, pp. 10793–10812, 2020.

130. Sathyanarayana, A., S. Joty, L. Fernandez-Luque, F. Ofli, J. Srivastava, A. Elmagarmid, T. Arora, S. Taheri *et al.*, "Sleep Quality Prediction from Wearable Data using Deep Learning", *JMIR mHealth and uHealth*, Vol. 4, No. 4, pp. e6562–e6574, 2016.

131. Lacour, M. and L. D. Tissington, "The Effects of Poverty on Academic Achievement", *Educational Research and Reviews*, Vol. 6, No. 7, pp. 522–527, 2011.

132. Shephard, R. J., "Curricular Physical Activity and Academic Performance", *Pediatric Exercise Science*, Vol. 9, No. 2, pp. 113–126, 1997.

133. Vhaduri, S. and C. Poellabauer, "Multi-modal Biometric-based Implicit Authen-

tication of Wearable Device Users", *IEEE Transactions on Information Forensics and Security*, Vol. 14, No. 12, pp. 3116–3125, 2019.

134. Faust, L., C. Wang, D. Hachen, O. Lizardo, N. V. Chawla *et al.*, "Physical Activity Trend Extraction: a Framework for Extracting Moderate-vigorous Physical Activity Trends from Wearable Fitness Tracker Data", *JMIR mHealth and uHealth*, Vol. 7, No. 3, pp. e11075–e11088, 2019.

135. Wang, R., F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev and A. T. Campbell, "Student Life Dataset", http://studentlife.cs.dartmouth.edu/, accessed on March 23, 2020.

136. Wang, R., G. Harari, P. Hao, X. Zhou and A. T. Campbell, "SmartGPA: How Smartphones can Assess and Predict Academic Performance of College Students", *Association for Computing Machinery International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 295–306, Osaka, Japan, 2015.

137. Daniels, B. T. and E. K. Howie, "Relationships between Physical Fitness, Health Behaviors, and Occupational Outcomes in Students, Faculty, and Staff of an American University", *Journal of American College Health*, Vol. 0, No. 0, pp. 1–8, 2022.

138. Zhang, S., Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng and N. Alshurafa, "Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances", *Sensors*, Vol. 22, No. 4, pp. 1476–1518, 2022.

139. Chen, C.-Y., S. Vhaduri and C. Poellabauer, "Estimating Sleep Duration from Temporal Factors, Daily Activities, and Smartphone Use", *44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 545–554, Madrid, Spain, 2020.

140. Aledavood, T., I. Kivimäki, S. Lehmann and J. Saramäki, "A Non-negative Ma-

trix Factorization based Method for Quantifying Rhythms of Activity and Sleep and Chronotypes using Mobile Phone Data", arXiv preprint arXiv:2009.09914, accessed on Oct 18, 2022.

141. Aledavood, T., S. Lehmann and J. Saramäki, "Social Network Differences of Chronotypes Identified from Mobile Phone Data", *European Physical Journal Data Science*, Vol. 7, No. 1, pp. 46–58, 2018.

142. Akimova, E. T., R. Taiji, X. Ding and M. C. Mills, "Gene-x-environment Analysis supports Protective Effects of Eveningness Chronotype on Self-reported and Actigraphy-derived Sleep Duration among Those who always Work Night Shifts in the UK Biobank", *Sleep*, Vol. 46, No. 5, pp. zsad023–zsad034, 2023.

143. Allen, R. W., R. D. Shaw, C. P. Burney, L. E. Newton, A. Y. Lee, B. G. Judd and S. J. Ivatury, "Deep Sleep and Beeps II: Sleep Quality Improvement Project in General Surgery Patients", *Surgery*, Vol. 172, No. 6, pp. 1697–1703, 2022.

144. Osborn, S., *The Impact of Insufficient Sleep and Early Class Start Times on US Air Force Academy Cadet Health and Performance*, Master's Thesis, James Madison University, 2021.

145. Vincent, A., A. Gupta, R. Li, C. Shaw and S. Akhyani, "Data Acquisition and Visual Analytic Tool-set for Paediatric Sleep Data", *13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pp. 320–326, Trento, Italy, 2019.

146. Allen, R. W., C. P. Burney, A. Davis, J. Henkin, J. Kelly, B. G. Judd and S. J. Ivatury, "Deep Sleep and Beeps: Sleep Quality Improvement Project in General Surgery Patients", *Journal of the American College of Surgeons*, Vol. 232, No. 6, pp. 882–888, 2021.

147. Ding, X., D. M. Brazel and M. C. Mills, "Gender Differences in Sleep Disruption

during COVID-19: Cross-sectional Analyses from two UK Nationally Representative Surveys", *Open British Medical Journal*, Vol. 12, No. 4, pp. e055792–e55804, 2022.

148. Hossain, H. S., S. R. Ramamurthy, M. A. A. H. Khan and N. Roy, "An Active Sleep Monitoring Framework using Wearables", *Association for Computing Machinery Transactions on Interactive Intelligent Systems (TiiS)*, Vol. 8, No. 3, pp. 1–30, 2018.

149. Zeng, M., L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu and J. Zhang, "Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors", *6th International Conference on Mobile Computing, Applications and Services*, pp. 197–205, Austin, TX, USA, 2014.

150. Triwiyanto, T., I. P. A. Pawana and M. H. Purnomo, "An Improved Performance of Deep Learning based on Convolution Neural Network to Classify the Hand Motion by Evaluating Hyper Parameter", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 28, No. 7, pp. 1678–1688, 2020.

151. Xu, Y. and T. T. Qiu, "Human Activity Recognition and Embedded Application based on Convolutional Neural Network", *Journal of Artificial Intelligence and Technology*, Vol. 1, No. 1, pp. 51–60, 2021.

152. Ha, S. and S. Choi, "Convolutional Neural Networks for Human Activity Recognition using Multiple Accelerometer and Gyroscope Sensors", *International Joint Conference on Neural Networks (IJCNN)*, pp. 381–388, Vancouver, BC, Canada, 2016.

153. Jordao, A., L. A. B. Torres and W. R. Schwartz, "Novel Approaches to Human Activity Recognition based on Accelerometer Data", *Signal, Image and Video Processing*, Vol. 12, pp. 1387–1394, 2018.

154. Marcusson-Clavertz, D., M. J. Sliwinski, O. M. Buxton, J. Kim, D. M. Almeida and J. M. Smyth, "Relationships between Daily Stress responses in Everyday Life and Nightly Sleep", *Journal of Behavioral Medicine*, Vol. 45, No. 4, pp. 518–532, 2022.

155. Gedam, S. and S. Paul, "A Review on Mental Stress Detection using Wearable Sensors and Machine Learning Techniques", *IEEE Access*, Vol. 9, pp. 84045–84066, 2021.

156. Steptoe, A., K. O'Donnell, M. Marmot and J. Wardle, "Positive Affect, Psychological Well-being, and Good Sleep", *Journal of Psychosomatic Research*, Vol. 64, No. 4, pp. 409–415, 2008.

157. Pilcher, J. J. and A. I. Huffcutt, "Effects of Sleep Deprivation on Performance: a Meta-analysis", *Sleep*, Vol. 19, No. 4, pp. 318–326, 1996.

158. Cagney, D. N., J. Sul, R. Y. Huang, K. L. Ligon, P. Y. Wen and B. M. Alexander, "The FDA NIH Biomarkers, EndpointS, and other Tools (BEST) Resource in Neuro-oncology", *Neuro-oncology*, Vol. 20, No. 9, pp. 1162–1172, 2018.

159. Frank, R. and R. Hargreaves, "Clinical Biomarkers in Drug Discovery and Development", *Nature Reviews Drug Discovery*, Vol. 2, No. 7, pp. 566–580, 2003.

160. Fendrich, S. J., M. Balachandran and M. S. Patel, "Association between Behavioral Phenotypes and Sustained use of Smartphones and Wearable Devices to Remotely Monitor Physical Activity", *Scientific Reports*, Vol. 11, No. 1, pp. 21501–21512, 2021.

161. Mandryk, R. L., M. V. Birk, S. Vedress, K. Wiley, E. Reid, P. Berger and J. Frommel, "Remote Assessment of Depression using Digital Biomarkers from Cognitive Tasks", *Frontiers in Psychology*, Vol. 12, pp. 767507–767520, 2021.

162. Park, C., R. Mishra, J. Golledge and B. Najafi, "Digital Biomarkers of Physical

Frailty and Frailty Phenotypes using Sensor-based Physical Activity and Machine Learning", *Sensors*, Vol. 21, No. 16, pp. 5289–5300, 2021.

163. Sano, A., S. Taylor, A. W. McHill, A. J. Phillips, L. K. Barger, E. Klerman and R. Picard, "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, pp. e210–e229, 2018.

164. Akbar, F., G. Mark, S. Prausnitz, E. M. Warton, J. A. East, M. F. Moeller, M. E. Reed and T. A. Lieu, "Inbox work: In situ Measurement with Aearable Sensors", *JMIR Medical Informatics*, Vol. 9, No. 4, pp. e24014–e24026, 2021.

165. Maxhuni, A., P. Hernandez-Leal, L. E. Sucar, V. Osmani, E. F. Morales and O. Mayora, "Stress Modelling and Prediction in Presence of Scarce Data", *Journal of Biomedical Informatics*, Vol. 63, pp. 344–356, 2016.

166. Muñoz, S., C. Á. Iglesias, O. Mayora and V. Osmani, "Prediction of Stress Levels in the Workplace using Surrounding Stress", *Information Processing & Management*, Vol. 59, No. 6, pp. 103064–103075, 2022.

167. Gjoreski, M., H. Gjoreski, M. Luštrek and M. Gams, "Continuous Stress Detection using a Wrist Device: in Laboratory and Real Life", *Association for Computing Machinery International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pp. 1185–1193, Heidelberg, Germany, 2016.

168. Alhejaili, R. and A. Alomainy, "The Use of Wearable Technology in Providing Assistive Solutions for Mental Well-being", *Sensors*, Vol. 23, No. 17, pp. 7378–7397, 2023.

169. Neto, M., V. S. Carvalho, M. J. Chambel, S. Manuel, J. P. Miguel and M. de Fátima Reis, "Work-family Conflict and Employee Well-being over Time",

Journal of Occupational and Environmental Medicine, Vol. 58, No. 5, pp. 429–435, 2016.

170. Gupta, S., L. Goel, A. Singh, A. K. Agarwal and R. K. Singh, "TOXGB: Teamwork Optimization Based XGBoost Model for Early Identification of Post-traumatic Stress Disorder", *Cognitive Neurodynamics*, Vol. 16, No. 4, pp. 833–846, 2022.

171. Sathishkumar, B., T. Saranya, A. Srishuwetha and S. Thirumalini, "Stress Detection using RNN Algorithm", *8th International Conference on Communication and Electronics Systems (ICCES)*, pp. 877–811, Coimbatore, Tamil Nadu, India, 2023.

172. Raksha, M., K. Suresh and D. Anusha, "To Identify Patient Psychological Impacts based on Enhanced Recurrent Neural Network Technique", *3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 979–986, Coimbatore, India, 2022.

173. Gupta, M. V., S. Vaikole, A. D. Oza, A. Patel, D. P. Burduhos-Nergis and D. D. Burduhos-Nergis, "Audio-Visual Stress Classification Using Cascaded RNN-LSTM Networks", *Bioengineering*, Vol. 9, No. 10, pp. 510–531, 2022.

174. Madhusanka, B. and S. Ramadass, "Implicit Intention Communication for Activities of Daily Living of Elder/disabled People to Improve Well-being", *IoT in Healthcare and Ambient Assisted Living*, Vol. 1, pp. 325–342, 2021.

175. Trmčić, B. R., A. Labus, S. Mitrović, V. Buha and G. Stanojević, *Internet of Things in E-health: an Application of Wearables in Prevention and Well-being*, IGI Global, Hershey, Pennsylvania, USA, 2018.

176. Tikkanen, H., K. Heinonen and A. Ravald, "Smart Wearable Technologies as Resources for Consumer Agency in Well-being", *Journal of Interactive Marketing*,

Vol. 58, No. 2-3, pp. 136–150, 2023.

177. Zehrung, R., L. Huang, B. Lee and E. K. Choe, "Investigating Opportunities to Support Kids' Agency and Well-being: A Review of Kids' Wearables", arXiv preprint arXiv:2104.05979, accessed on Dec 28, 2021.

178. Andrushevich, A., M. Biallas, R. Kistler, L. Eusebi, J. Ronai and M. Valla, "Towards Smart Working spaces with Enhanced Well-being and Safety of Elderly Staff", *Global Internet of Things Summit (GIoTS)*, pp. 1–6, Geneva, Switzerland, 2017.

179. Smith, E. N., E. Santoro, N. Moraveji, M. Susi and A. J. Crum, "Integrating Wearables in Stress Management Interventions: Promising Evidence from a Randomized Trial", *International Journal of Stress Management*, Vol. 27, No. 2, pp. 172–207, 2020.

180. Arora, A., P. Chakraborty and M. Bhatia, "Intervention of Wearables and Smartphones in Real Time Monitoring of Sleep and Behavioral Health: an Assessment using Adaptive Neuro-fuzzy Technique", *Arabian Journal for Science and Engineering*, Vol. 47, No. 2, pp. 1–26, 2021.

181. Kaur, T., P. Ranjan, A. Chakrawarty, K. Kasi, P. Berry, S. Suryansh, A. Mazumder, M. Khan, A. D. Upadhyay, G. Kaloiya *et al.*, "Association of Sociodemographic Parameters with Depression, Anxiety, Stress, Sleep Quality, Psychological Trauma, Mental Well-being, and Resilience during the Second Wave of COVID-19 Pandemic: a Cross-sectional Survey from India", *Cureus*, Vol. 13, No. 7, pp. e16420–e16430, 2021.

182. Wunsch, K., N. Kasten and R. Fuchs, "The Effect of Physical Activity on Sleep Quality, Well-being, and Affect in Academic Stress Periods", *Nature and Science of Sleep*, Vol. 9, pp. 117–126, 2017.

183. Mellman, T. A., "Sleep and Anxiety Disorders", *Psychiatric Clinics*, Vol. 29, No. 4, pp. 1047–1058, 2006.

184. Taylor, A. H., "Physical Activity, Anxiety, and Stress", *Physical Activity and Psychological Well-being*, Vol. 1, pp. 22–52, 2003.

185. Doherty, S. T., C. J. Lemieux and C. Canally, "Tracking Human Activity and Well-being in Natural Environments using Wearable Sensors and Experience Sampling", *Social Science & Medicine*, Vol. 106, pp. 83–92, 2014.

186. Hänsel, K., "Wearable and Ambient Sensing for Well-being and Emotional Awareness in the Smart Workplace", *Association for Computing Machinery International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pp. 411–416, Heidelberg, Germany, 2016.

187. Hosseini, S., R. Gottumukkala, S. Katragadda, R. T. Bhupatiraju, Z. Ashkar, C. W. Borst and K. Cochran, "A Multimodal Sensor Dataset for Continuous Stress Detection of Nurses in a Hospital", *Scientific Data*, Vol. 9, No. 1, pp. 255–267, 2022.

188. Sükei, E., A. Norbury, M. M. Perez-Rodriguez, P. M. Olmos and A. Artés, "Predicting Emotional States using Behavioral Markers derived from Passively Sensed data: Data-driven Machine Learning Approach", *JMIR mHealth and uHealth*, Vol. 9, No. 3, pp. e24465–e24478, 2021.

189. Aledavood, T., I. Kivimäki, S. Lehmann and J. Saramäki, "Quantifying Daily Rhythms with Non-negative Matrix Factorization applied to Mobile Phone Data", *Scientific Reports*, Vol. 12, No. 1, pp. 5544–5553, 2022.

190. Kang, S., W. Choi, C. Y. Park, N. Cha, A. Kim, A. H. Khandoker, L. Hadjileontiadis, H. Kim, Y. Jeong and U. Lee, "K-emophone: A Mobile and Wearable Dataset with In-situ Emotion, Stress, and Attention Labels", *Scientific Data*,

Vol. 10, No. 1, pp. 351–371, 2023.

191. Murray, C. J., J. Abraham, M. K. Ali, M. Alvarado, C. Atkinson, L. M. Baddour, D. H. Bartels, E. J. Benjamin, K. Bhalla, G. Birbeck *et al.*, "The State of US Health, 1990-2010: Burden of Diseases, Injuries, and Risk Factors", *Journal of the American Medical Association*, Vol. 310, No. 6, pp. 591–606, 2013.

192. Kim, J., T. Nakamura, H. Kikuchi, T. Sasaki and Y. Yamamoto, "Co-variation of Depressive Mood and Locomotor Dynamics Evaluated by Ecological Momentary Assessment in Healthy Humans", *Public Library of Science One*, Vol. 8, No. 9, pp. e74979–e74990, 2013.

193. Shiffman, S., A. A. Stone and M. R. Hufford, "Ecological Momentary Assessment", *Annual Review of Clinical Psychology*, Vol. 4, pp. 1–32, 2008.

194. Runyan, J. D., T. A. Steenbergh, C. Bainbridge, D. A. Daugherty, L. Oke and B. N. Fry, "A Smartphone Ecological Momentary Assessment/Intervention "app" for Collecting Real-time Data and Promoting Self-awareness", *Public Library of Science One*, Vol. 8, pp. e71325–e71333, 2013.

195. Madan, A., M. Cebrian, D. Lazer and A. Pentland, "Social Sensing for Epidemiological Behavior Change", *12th Association for Computing Machinery International Conference on Ubiquitous Computing*, pp. 291–300, Copenhagen, Denmark, 2010.

196. Paromita, P., K. Mundnich, A. Nadarajan, B. M. Booth, S. S. Narayanan and T. Chaspari, "Modeling Inter-individual Differences in Ambulatory-based Multimodal Signals via Metric Learning: a Case Study of Personalized Well-being Estimation of Healthcare Workers", *Frontiers in Digital Health*, Vol. 5, pp. 1195795–1195813, 2023.

197. Xu, X., H. Zhang, Y. Sefidgar, Y. Ren, X. Liu, W. Seo, J. Brown, K. Kuehn,

M. Merrill, P. Nurius *et al.*, "GLOBEM Dataset: Multi-year Datasets for Longitudinal Human Behavior Modeling Generalization", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24655–24692, 2022.

198. Taelman, J., S. Vandeput, A. Spaepen and S. Van Huffel, "Influence of Mental Stress on Heart Rate and Heart Rate Variability", *4th European Conference of the International Federation for Medical and Biological Engineering: ECIFMBE*, pp. 1366–1369, Antwerp, Belgium, 2009.

199. Shwartz-Ziv, R. and A. Armon, "Tabular Data: Deep Learning is not All You Need", *Information Fusion*, Vol. 81, pp. 84–90, 2022.

200. Martire, V. L., D. Caruso, L. Palagini, G. Zoccoli and S. Bastianini, "Stress & sleep: A Relationship Lasting a Lifetime", *Neuroscience & Biobehavioral Reviews*, Vol. 117, pp. 65–77, 2020.

201. Partinen, M., "Sleep disorders and stress", *Journal of Psychosomatic Research*, Vol. 38, pp. 89–91, 1994.

202. Rosa, R. R., M. H. Bonnet and M. Kramer, "The Relationship of Sleep and Anxiety in Anxious Subjects", *Biological Psychology*, Vol. 16, No. 1-2, pp. 119–126, 1983.

203. Alfano, C. A., A. H. Zakem, N. M. Costa, L. K. Taylor and C. F. Weems, "Sleep Problems and Their Relation to Cognitive Factors, Anxiety, and Depressive Symptoms in Children and Adolescents", *Depression and Anxiety*, Vol. 26, No. 6, pp. 503–512, 2009.

204. Wijndaele, K., L. Matton, N. Duvigneaud, J. Lefevre, I. De Bourdeaudhuij, W. Duquet, M. Thomis and R. M. Philippaerts, "Association between Leisure Time Physical Activity and Stress, Social Support and Coping: A Cluster-Analytical Approach", *Psychology of Sport and Exercise*, Vol. 8, No. 4, pp. 425–

440, 2007.

205. Schultchen, D., J. Reichenberger, T. Mittl, T. R. Weh, J. M. Smyth, J. Blechert and O. Pollatos, "Bidirectional Relationship of Stress and Affect with Physical Activity and Healthy Eating", *British Journal of Health Psychology*, Vol. 24, No. 2, pp. 315–333, 2019.

206. Stone, A. A., J. M. Neale and S. Shiftman, "Daily Assessments of Stress and Coping and Their Association with Mood", *Annals of Behavioral Medicine*, Vol. 15, No. 1, pp. 8–16, 1993.

207. van Eck, M., N. A. Nicolson and J. Berkhof, "Effects of Stressful Daily Events on Mood States: Relationship to Global Perceived Stress.", *Journal of Personality and Social Psychology*, Vol. 75, No. 6, pp. 1572–1586, 1998.

208. Graham, S., C. Depp, E. E. Lee, C. Nebeker, X. Tu, H.-C. Kim and D. V. Jeste, "Artificial Intelligence for Mental Health and Mental Illnesses: an Overview", *Current Psychiatry Reports*, Vol. 21, pp. 1–18, 2019.

209. Bond, R. R., M. D. Mulvenna, C. Potts, S. O'Neill, E. Ennis and J. Torous, "Digital Transformation of Mental Health Services", *Nature Partner Journals Mental Health Research*, Vol. 2, No. 1, pp. 13–21, 2023.

210. Jordan, M. I. and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects", *Science*, Vol. 349, No. 6245, pp. 255–260, 2015.

211. Chung, J. and J. Teo, "Mental Health Prediction using Machine Learning: Taxonomy, Applications, and Challenges", *Applied Computational Intelligence and Soft Computing*, Vol. 2022, pp. 1–19, 2022.

212. Coutts, L. V., D. Plans, A. W. Brown and J. Collomosse, "Deep Learning with Wearable based Heart Rate Variability for Prediction of Mental and General Health", *Journal of Biomedical Informatics*, Vol. 112, pp. 103610–103617, 2020.

213. Garriga, R., J. Mas, S. Abraha, J. Nolan, O. Harrison, G. Tadros and A. Matic, "Machine Learning Model to Predict Mental Health Crises from Electronic Health Records", *Nature Medicine*, Vol. 28, No. 6, pp. 1240–1248, 2022.

214. Khoo, L. S., M. K. Lim, C. Y. Chong and R. McNaney, "Machine Learning for Multi-modal Mental Health Detection: A Systematic Review of Passive Sensing Approaches", *Sensors*, Vol. 24, No. 2, pp. 348–412, 2024.

215. Garcia-Ceja, E., M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard and J. Tørresen, "Mental Health Monitoring with Multi-modal Sensing and Machine Learning: A Survey", *Pervasive and Mobile Computing*, Vol. 51, pp. 1–26, 2018.

216. Wu, C., H. Fritz, S. Bastami, J. P. Maestre, E. Thomaz, C. Julien, D. M. Castelli, K. de Barbaro, S. K. Bearman, G. M. Harari *et al.*, "Multi-modal Data Collection for Measuring Health, Behavior, and Living Environment of Large-scale Participant Cohorts", *GigaScience*, Vol. 10, No. 6, pp. giab044–giab056, 2021.

217. Tateyama, N., R. Yokomura, Y. Ban, S. Warisawa and R. Fukui, "Planning the Future in a Longer Perspective: Effects of a One-week Forecast of Mental Health", *Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 8, No. 1, pp. 1–20, 2024.

218. Kaushik, S., A. Choudhury, P. K. Sheron, N. Dasgupta, S. Natarajan, L. A. Pickett and V. Dutt, "AI in Healthcare: Time-series Forecasting using Statistical, Neural, and Ensemble Architectures", *Frontiers in Big Data*, Vol. 3, pp. 4–20, 2020.

219. Seng, K. P., L.-M. Ang, E. Peter and A. Mmonyi, "Machine Learning and AI Technologies for Smart Wearables", *Electronics*, Vol. 12, No. 7, pp. 1509–1529, 2023.

220. Evgeniou, T. and M. Pontil, "Regularized Multi-task Learning", *Association for*

Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining International Conference, pp. 109–117, Seattle, WA, USA, 2004.

221. Ruder, S., "An Overview of Multi-task Learning in Deep Neural Networks", arXiv preprint arXiv:1706.05098, accessed on Sept 18, 2021.

222. Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, P. Solenberger *et al.*, "A Multivariate Technique for Multiply imputing Missing Values using a Sequence of Regression Models", *Survey Methodology*, Vol. 27, No. 1, pp. 85–96, 2001.

223. Xu, D., Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong and X. Shen, "Survey on Multi-output Learning", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 31, No. 7, pp. 2409–2429, 2019.

224. Zhang, Y. and Q. Yang, "An Overview of Multi-task Learning", *National Science Review*, Vol. 5, No. 1, pp. 30–43, 2018.

225. Lundberg, S. M. and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 1–10, 2017.

226. Shekhawat, N. and K. Jain, "Relationship of Self-Regulation with Mental Health Among Boy and Girl Adolescents", *International Journal of Indian Psychology*, Vol. 10, No. 4, pp. 1–5, 2022.

227. Scott, A. J., T. L. Webb, M. Martyn-St James, G. Rowse and S. Weich, "Improving Sleep Quality leads to Better Mental Health: A Meta-analysis of Randomised Controlled Trials", *Sleep Medicine Reviews*, Vol. 60, pp. 101556–101574, 2021.

228. Billah, M. A., S. Akhtar and M. N. Khan, "Loneliness and Trust Issues reshape Mental Stress of Expatriates during Early COVID-19: a Structural Equation Modelling Approach", *BioMed Central Psychology*, Vol. 11, No. 1, pp. 140–150, 2023.

APPENDIX A: TELİF HAKLARIYLA KORUNAN MATERYALİN KULLANIMI

Bu tez kapsamında ortaya çıkan ve yayınevine teslim edilen görseller, yayınevinin kendi internet sayfasında yer alan "yazarın kendi ürettiği metin ve grafiklerin tekrar kullanımı hakkında geçerli olan yayın politikası"na uygun olarak tez kitabında kullanılmıştır.