ARISTOTLE UNIVERSITY OF THESSALONIKI

Testing (Un)fairness in Personalized Machine Learning for Stress Detection in Ubiquitous Computing

by

Alexia Ntantouri - Student ID: 3871

Pre-graduate thesis

in the
Faculty of Sciences
School of Informatics

Supervising Professor: Dr. Athena Vakali

June 2024

Declaration of Authorship

I, Alexia Ntantouri, declare that this thesis titled, "Testing (Un)fairness in Personalized Machine Learning for Stress Detection in Ubiquitous Computing" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:
Alexia Ntantouri

Date:
June 2024

"Great things are not done by impulse, but by a series of small things brought together."

Vincent Van Gogh

Abstract

This thesis explores the potential of personalized machine learning (ML) models for stress detection using physiological data from wearable devices, focusing on the critical balance between accuracy and fairness. The motivation for this research stems from the negative impact of stress on health and well-being and the opportunity that wearable technology provides for real-time monitoring and personalized interventions. However, the effectiveness of these technologies relies heavily on the accuracy and fairness of the underlying ML models. This study aims to fill gaps in the existing literature by systematically analyzing biases in stress detection datasets and exploring models that achieve both high accuracy and fairness.

The research methodology involves the application of personalized ML algorithms to datasets collected in both controlled laboratory environments and real-world settings, treating stress detection as a binary classification problem. Various ML models, including foundational and personalized models, were benchmarked to assess their performance in terms of accuracy and fairness. A comprehensive approach was taken to evaluate data bias, utilizing various plots to visualize biases and identify demographic imbalances within the datasets.

The analysis revealed significant biases in the datasets. The LifeSnaps dataset exhibited sampling and representation biases across gender and age, along with measurement biases due to missing values and imbalanced label distributions. Similarly, the SWELL-KW dataset showed pronounced biases, particularly in age representation. These findings underscore the necessity of developing more diverse and representative datasets to ensure the creation of equitable stress detection models.

Another key finding of this study is that it is indeed possible to achieve both accuracy and fairness in stress detection models. Personalized ML models demonstrated high F1-scores while maintaining fairness metrics within acceptable thresholds. For example, a personalized model applied to a specific dataset achieved an accuracy of 82.89% while adhering to fairness standards, indicating that equitable stress detection is feasible.

Future work should focus on mitigating biases throughout the entire AI lifecycle. This includes developing methods to ensure diverse and representative datasets and implementing fairness metrics to continuously assess and improve the equity of stress detection systems. Establishing standardized protocols for data collection can significantly enhance the consistency and quality of datasets, facilitating the development and comparison of stress detection models across different studies and populations. Additionally, adopting frameworks like the AI Fairness 360 toolkit, which provides comprehensive metrics for datasets and models to identify and mitigate biases, is crucial. Engaging diverse teams throughout the AI development lifecycle, including

problem framing, model development, validation, and deployment, ensures that models are designed with fairness at their core.

The importance of addressing bias in Artificial Intelligence (AI) systems cannot be overstated. Creating accurate and fair stress detection models is crucial for their trustworthiness and effectiveness in real-world applications. Accurate and fair models can be instrumental in various use cases, such as personalized treatment plans for individuals with high-stress levels, improving workplace wellness programs, and enhancing public health interventions. These models must be transparent and interpretable, ensuring that stakeholders can understand and trust their outputs. By focusing on fairness, AI systems can better serve all segments of the population, reducing disparities and promoting equity in technology applications.

In conclusion, this thesis underscores the importance of addressing biases in stress detection datasets and developing ML models that balance accuracy with fairness. By highlighting the demographic imbalances and biases present in stress detection datasets, the research advocates for more inclusive data collection practices and comprehensive fairness assessments. The findings demonstrate that accurate and fair stress detection models are feasible, paving the way for personalized interventions that can significantly improve individual well-being and contribute to equitable healthcare solutions. By addressing the limitations in current practices and exploring new directions for future research, this thesis contributes to the development of robust, fair, and user-centric stress detection systems. These advancements hold the potential to transform how stress is monitored and managed, offering tailored solutions that cater to diverse populations and promoting overall health equity.

Περίληψη

Το άγχος είναι ένα διαδεδομένο πρόβλημα στη σύγχρονη κοινωνία, επηρεάζοντας σημαντικά την υγεία και την ευημερία των ατόμων. Οι εργασιακές πιέσεις, οι προσωπικές σχέσεις και οι καθημερινές ευθύνες είναι μόνο μερικοί από τους λόγους που το άγχος κυριαρχεί στην καθημερινότητά μας. Γι΄ αυτό, η έγκαιρη ανίχνευση και διαχείριση του άγχους είναι ζωτικής σημασίας για τη διατήρηση της ψυχικής και σωματικής μας υγείας. Οι τεχνολογικές εξελίξεις, ιδιαίτερα στα έξυπνα ρολόγια (smartwatches), προσφέρουν νέες ευκαιρίες για συνεχή και μη παρεμβατική παρακολούθηση του άγχους. Αυτές οι συσκευές μπορούν να συλλέγουν δεδομένα, όπως καρδιακούς παλμούς ή πίεση του αίματος, προσφέροντας νέες μεθόδους για την ανίχνευση και διαχείριση του άγχους. Αυτή η διατριβή εξετάζει το τομέα της ανίχνευσης άγχους στην πανταχού παρούσα υπολογιστική (Ubiquitous Computing), εντοπίζοντας υπάρχοντα κενά και προτείνοντας μεθοδολογίες για την αντιμετώπιση των προκαταλήψεων (biases) στα δεδομένα και στα μοντέλα που χρησιμοποιούνται για την ανίχνευση άγχους.

Στην ανίχνεσυη άγχους, σπουδαίο ρόλο παίζει η Τεχνητή Νοημοσύνη (Artificial Intelligence - AI) με την ικανότητα να μαθαίνει από δεδομένα που συλλέγονται από φορητές συσκευές και να προβλέπει το άγχος μέσω μοντέλων μηχανικής μάθησης (Machine Learning - ML). Η εξατομίκευση (personalization) είναι επίσης ένας σημαντικός τομέας στην ανίχνευση άγχους, όπου τα μοντέλα προσαρμόζονται σε μεμονωμένους χρήστες για να βελτιώσουν την ακρίβεια και την αποτελεσματικότητά τους. Παράγοντες όπως ο τρόπος ζωής και οι προσωπικές εμπειρίες επηρεάζουν το πώς διαφορετικοί άνθρωποι βιώνουν και αντιμετωπίζουν το άγχος και με την εξατομίκευση των μοντέλων ανίχνευσης άγχους, είναι δυνατό να ληφθούν υπόψη αυτές οι ατομικές διαφορές, οδηγώντας σε πιο ακριβή και αποτελεσματική παρακολούθηση και διαχείριση του άγχους.

Ωστόσο, η αποτελεσματικότητα αυτών των τεχνολογιών εξαρτάται σε μεγάλο βαθμό από την ακρίβεια και το πόσο δίκαια είναι αυτά τα μοντέλα. Αυτή η μελέτη στοχεύει να καλύψει τα κενά στη υπάρχουσα βιβλιογραφία, αναλύοντας τις προκαταλήψεις στα δεδομένα ανίχνευσης άγχους και στην ανάλυση μοντέλων που επιτυγχάνουν τόσο υψηλή ακρίβεια όσο και δίκαιο αποτελέσματα για όλες τις ομάδες χρηστών. Πιο συγκεκριμένα, οι προκαταλήψεις που είναι εγγενείς στα σύνολα εκπαίδευσης δεν έχουν διερευνηθεί επαρκώς στις εφαρμογές ανίχνευσης άγχους. Αυτή η διατριβή αντιμετωπίζει αυτό το ζήτημα προσφέροντας λεπτομερείς οπτικοποιήσεις μέσω διαγραμμάτων για τον εντοπισμό και την κατανόηση των διαφορετικών τύπων προκαταλήψεων που υπάρχουν στα δεδομένα εκπαίδευσης. Υπάρχει επίσης αξιοσημείωτη έλλειψη αμφίδρομης αξιολόγησης σχετικά με την ακρίβεια και το πόσο δίκαια είναι τα εξατομικευμένα μοντέλα. Αυτή η διατριβή διερευνά τους συμβιβασμούς μεταξύ ακρίβειας και δικαιοσύνης στα θεμελιώδη και εξατομικευμένα μοντέλα μηχανικής μάθησης για την ανίχνευση άγχους, προσφέροντας πολύτιμες πληροφορίες για την πρακτική τους εφαρμογή.

Η μεθοδολογία που ακολουθείται περιλαμβάνει την εφαρμογή εξατομικευμένων αλγορίθμων μηχανικής μάθησης σε σύνολα δεδομένων που συλλέγονται τόσο σε ελεγχόμενα εργαστηριακά περιβάλλοντα όσο και σε πραγματικές συνθήκες, αντιμετωπίζοντας την ανίχνευση άγχους ως πρόβλημα δυαδικής ταξινόμησης. Ακολουθήθηκε μια ολοκληρωμένη προσέγγιση για την αξιολόγηση της προκατάληψης των δεδομένων, χρησιμοποιώντας διάφορα διαγράμματα για τον εντοπισμό δημογραφικών ανισοτήτων στα σύνολα δεδομένων. Η ανάλυση αποκάλυψε σημαντικές προκαταλήψεις στα σύνολα δεδομένων. Συγκεκριμένα, το σύνολο δεδομένων LifeSnaps παρουσίασε προκαταλήψεις στην δειγματοληψία και την αντιπροσώπευση αναφορικά με το φύλο και την ηλικία των συμμετεχόντων. Παρομοίως, το σύνολο δεδομένων SWELL-KW έδειξε έντονες προκαταλήψεις, ιδιαίτερα στην αντιπροσώπευση της ηλικίας των ατόμων. Αυτά τα ευρήματα υπογραμμίζουν την αναγκαιότητα ανάπτυξης πιο ποικίλων και αντιπροσωπευτικών συνόλων δεδομένων για να εξασφαλιστεί η δημιουργία δίκαιων μοντέλων ανίχνευσης άγχους.

΄Οσων αφορά τα μοντέλα, διάφορα μοντέλα μηχανικής μάθησης, συμπεριλαμβανομένων των θεμελιωδών και εξατομικευμένων μοντέλων, αξιολογήθηκαν για την απόδοσή τους. Χρησιμοποιήθηκε η βιβλιοθήκη aif360 της Python εφαρμόζοντας διάφορες μετρικές δικαιοσύνης για την αξιολόγηση των προκαταλήψεων των μοντέλων. Η ανάλυση της προκατάληψης των μοντέλων αποκάλυψε ότι είναι πράγματι δυνατό να επιτευχθούν τόσο η ακρίβεια όσο και η δικαιοσύνη στα μοντέλα ανίχνευσης άγχους. Τα εξατομικευμένα μοντέλα μηχανικής μάθησης έδειξαν ικανοποιητική ακρίβεια, ενώ διατηρούσαν τις μετρικές δικαιοσύνης εντός αποδεκτών ορίων. Για παράδειγμα, ένα εξατομικευμένο μοντέλο που εφαρμόστηκε σε συγκεκριμένο σύνολο δεδομένων πέτυχε f1-score 82,89% ενώ τηρούσε τα πρότυπα δικαιοσύνης, υποδεικνύοντας ότι είναι εφικτή η δίκαιη ανίχνευση άγχους.

Σχετικά με τη μελλοντική δουλειά που προτείνεται να διεξαχθεί σε αυτόν τον τομέα, είναι σημαντικό να γίνει προσπάθεια μετριασμού των προκαταλήψεων σε όλο τον κύκλο ζωής της μηχανικής μάθησης. Αυτό περιλαμβάνει την ανάπτυξη μεθόδων για να εξασφαλιστούν ποικίλα και αντιπροσωπευτικά σύνολα δεδομένων και την εφαρμογή μετρικών δικαιοσύνης για τη συνεχή αξιολόγηση και βελτίωση της δικαιοσύνης των συστημάτων ανίχνευσης άγχους. Η καθιέρωση τυποποιημένων πρωτοκόλλων για τη συλλογή δεδομένων μπορεί να ενισχύσει σημαντικά τη συνέπεια και την ποιότητα των συνόλων δεδομένων, διευκολύνοντας την ανάπτυξη και τη σύγκριση μοντέλων ανίχνευσης άγχους σε διάφορες μελέτες και πληθυσμούς. Επιπλέον, η υιοθέτηση πλαισίων όπως το AI Fairness 360 toolkit, το οποίο παρέχει ολοκληρωμένες μετρικές για σύνολα δεδομένων και μοντέλα για την αναγνώριση και τον μετριασμό των προκαταλήψεων, είναι κρίσιμη.

Η σημασία της αντιμετώπισης των προκαταλήψεων στα συστήματα AI δεν μπορεί να υποτιμηθεί. Η δημιουργία ακριβών και δίκαιων μοντέλων ανίχνευσης άγχους είναι κρίσιμη για την αξιοπιστία και την αποτελεσματικότητά τους σε πραγματικές εφαρμογές. Αυτά τα μοντέλα

πρέπει να είναι διαφανή και κατανοητά, διασφαλίζοντας ότι οι ενδιαφερόμενοι μπορούν να κατανοήσουν και να εμπιστευτούν τα αποτελέσματά τους. Με την εστίαση στη δικαιοσύνη, τα συστήματα AI μπορούν να εξυπηρετούν καλύτερα όλους τους πληθυσμούς, μειώνοντας τις ανισότητες και προωθώντας την ισότητα στις τεχνολογικές εφαρμογές.

Συμπερασματικά, αυτή η διατριβή υπογραμμίζει τη σημασία της αντιμετώπισης των προκαταλήψεων στα σύνολα δεδομένων ανίχνευσης άγχους και της ανάπτυξης μοντέλων μηχανικής μάθησης που ισορροπούν την ακρίβεια με τη δικαιοσύνη. Επισημαίνοντας τις δημογραφικές ανισότητες και τις προκαταλήψεις που υπάρχουν στα σύνολα δεδομένων, η έρευνα συνηγορεί υπέρ των πιο περιεκτικών πρακτικών συλλογής δεδομένων και ολοκληρωμένων αξιολογήσεων δικαιοσύνης. Τα ευρήματα της έρευνας δείχνουν ότι είναι εφικτή η ανάπτυξη ακριβών και δίκαιων μοντέλων ανίχνευσης άγχους, ανοίγοντας τον δρόμο για εξατομικευμένες παρεμβάσεις που μπορούν να βελτιώσουν σημαντικά την ατομική ευημερία και να συμβάλουν σε δίκαιες λύσεις υγειονομικής περίθαλψης. Με την αντιμετώπιση των περιορισμών στις τρέχουσες πρακτικές και την εξερεύνηση νέων κατευθύνσεων για μελλοντική έρευνα, αυτή η διατριβή συμβάλλει στην ανάπτυξη ακριβών και δίκαιων προς τον χρήστη συστημάτων ανίχνευσης άγχους. Αυτές οι εξελίξεις έχουν τη δυνατότητα να μεταμορφώσουν τον τρόπο με τον οποίο παρακολουθείται και διαχειρίζεται το άγχος, προσφέροντας προσαρμοσμένες λύσεις που καλύπτουν διαφορετικούς πληθυσμούς και προωθώντας τη συνολική ισότητα στην υγεία.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Athena Vakali, whose guidance and support have been invaluable throughout my academic journey. Her mentorship has profoundly shaped this thesis.

I am also indebted to Eva Paraschou, whose expertise and dedication contributed significantly to the completion of this thesis. Her constructive criticism and assistance were pivotal in conducting my research and improving the quality of my work.

I would also like to extend my appreciation to Sofia Yfantidou, who has been a major support on this journey giving me valuable insights into writing this thesis. I am also grateful to Christina Karagianni, who provided valuable guidance during the initial stages of this thesis and her early support helped set the foundation for this project. Additionally, I want to thank Vasiliki-Chrisovalanto Parousidou, since this work wouldn't have been possible without her assistance and previous work on the domain of this thesis.

Lastly, I would like to acknowledge my deepest gratitude to my family and friends. Their unwavering support and encouragement have been a constant source of strength throughout my academic journey and beyond. I am truly grateful for their love and belief in me.

Contents

List of Figures

List of Tables

Abbreviations

ADA	**Ada**Boost
AI	**A**rtificial **I**ntelligence
CNN	**C**onvolutional **N**eural **N**etwork
DIR	**D**isparate **I**mpact **R**atio
DT	**D**ecision **T**ree
ECG	**E**lectro**c**ardio**g**ram
EDA	**E**lectro**d**ermal **A**ctivity
EEG	**E**lectroen**e**phalo**g**ram
EHR	**E**lectronic **H**ealth **R**ecord
ER	**E**rror **R**ate
ERD	**E**rror **R**ate **D**ifference
ET	**E**xtra **T**rees
EU	**E**uropean **U**nion
FCM	**F**uzzy **C**-**M**eans
FOR	**F**alse **O**mission **R**ate
FORD	**F**alse **O**mission **R**ate **D**ifference
GBC	**G**radient **B**oosting **C**lassifier
kNN	**k**-**N**earest **N**eighbor **D**ifference
LDA	**L**inear **D**iscriminant **A**nalysis
LGB	**L**ight **G**radient **B**oosting
LGBM	**L**ight **G**radient **B**oosting **M**achine
LR	**L**ogistic **R**egression
MAFL	**M**ulti-**A**ttribute **F**airness **L**oss
ML	**M**achine **L**earning
NB	**N**aive **B**ayes

NPVD	**N**egative **P**redicted **V**alue **D**ifference
PCA	**P**rincipal **C**omponent **A**nalysis **M**achine
PSQ	**P**erceived **S**tress **Q**estionnaire
QDA	**Q**uadratic **D**iscriminant **A**nalysis
RC	**R**idge **C**lassifier
RF	**R**andom **F**orest
SPD	**S**tatistical **P**arity **D**ifference
SVM	**S**upport **V**ector **M**achine
UbiComp	**Ubi**quitous **Comp**uting

Chapter 1

Introduction

In the modern age, stress has become an integral part of our daily lives. As we navigate through the complexities of work, relationships, and daily responsibilities, the ability to monitor and manage stress has emerged as a critical aspect of proactive healthcare. Nowadays, advancements in technology offer promising solutions for this challenge. Specifically, mobile and wearable devices equipped with sensors can now capture various physiological and behavioral data, opening doors for continuous and unobtrusive stress monitoring.

In this chapter, we first present the state of the art regarding the subject of our thesis and provide the problem statement (Section 1.1). Then, we outline the current gaps and our contributions (Section 1.2 and lastly, we outline this dissertation's structure (Section 1.3).

1.1 State-of-the-art and Problem statement

Artificial Intelligence (AI) has become a transformative force across various domains, profoundly impacting both technology and society. AI's ability to learn from data, recognize patterns, and make decisions has enabled a wide range of applications that enhance our daily lives.

For example, in the automotive industry, AI powers self-driving cars by processing sensor data to navigate safely and efficiently. In finance, AI algorithms are used for fraud detection, analyzing transactions in real-time to identify suspicious activities. Moreover, in the retail sector, AI-driven recommendation systems personalize shopping experiences by analyzing consumer behavior and preferences data.

In some AI applications, data are coming from wearable devices such as smartwatches and smartphones. These devices continuously monitor physiological and environmental data, leveraging AI to interpret this information and offer actionable insights. For instance, smartwatches can track heart rate, sleep patterns, and physical activity, using AI algorithms to provide personalized health recommendations and alerts. This integration between AI and wearable devices exemplifies the seamless collection and analysis of data, enabling AI to provide context-aware and real-time solutions.

Such applications are part of the Ubiquitous Computing (UbiComp) ecosystem. UbiComp [9] refers to the integration of computational capabilities into everyday objects and environments, making technology pervasive and interactive. This integration allows for the seamless collection and analysis of data, enabling AI to provide context-aware and real-time solutions. The convergence of AI and UbiComp has further amplified the impact of advanced technologies, making technology pervasive and interactive. By embedding sensors and computing devices in our surroundings and personal items, UbiComp facilitates continuous data gathering and processing, leading to smarter and more adaptive systems. This integration is crucial for providing context-aware and real-time solutions that enhance user experience and improve quality of life.

One of the promising applications of AI in UbiComp is stress detection. Stress detection is important because chronic stress can lead to severe health problems, including cardiovascular diseases, depression, and anxiety [22]. Recognizing this, several companies such as Fitbit [1] (Figure 1.1) and Garmin [2] (Figure 1.2) have integrated stress monitoring features into their products. For instance, Fitbit's Sense smartwatch measures electrodermal activity (EDA) to monitor stress levels, providing users with insights into their stress patterns and offering relaxation recommendations. Similarly, Garmin's wearable devices track stress by analyzing heart rate variability, helping users understand their stress responses and manage their well-being more effectively.

A vital part of effective stress detection is personalization, as individuals have diverse physiological and psychological responses to stress. UbiComp can support this need for personalization in stress detection through its seamless integration of AI, gathering a wealth of information about an individual's physiological responses, activities, and environmental factors in real-time and analyzing this data using personalized algorithms.

[1] https://www.fitbit.com/global/nl/technology/stress
[2] https://www.garmin.com/en-US/garmin-technology/health-science/stress-tracking/

FIGURE 1.1: Fitbit stress tracking.

FIGURE 1.2: Garmin stress tracking.

However, this personalization can lead to unfair outcomes. This can happen if, for example, the training data predominantly represent certain demographics or lifestyles, the resulting models may be less accurate for underrepresented groups. Additionally, without adequate validation across diverse populations, personalized AI applications may perform well for some individuals but poorly for others.

There have been circumstances where AI has not been fair, resulting in biased outcomes that disproportionately affect certain groups. For instance, a case involved a healthcare algorithm that was less likely to refer black patients for additional care than white patients with the same health conditions. This bias was attributed to the training data, which used healthcare costs as a proxy for health needs, inadvertently reflecting existing disparities in access to healthcare [23]. Similarely, as reported by the Associated Press, racial bias has been found in medical diagnosis, where an AI system for lung disease diagnosis likely underdiagnosed black male

patients exacerbating health disparities [3]. Another example refers to the criminal justice system. In 2016, ProPublica reported racial bias in COMPAS, a widely used algorithm assessing recidivism risk in the criminal justice system. It disproportionately flagged black defendants as future criminals compared to white defendants. This led to harsher sentences for black individuals, perpetuating racial disparities [4].

Such instances highlight the need for rigorous evaluation and mitigation of bias in AI systems to ensure equitable and just outcomes for all users. Not taking biases into account when developing personalized ML algorithms can exacerbate existing disparities in stress management and healthcare access. This could result in unequal access to effective stress management tools, amplifying health inequalities.

The importance of addressing fairness issues in AI applications is also highlighted by the European Union (EU) which is actively working to regulate AI to ensure fairness, among other ethical and trustworthy aspects. The EU's proposed AI Act [21] aims to establish a legal framework that addresses the risks associated with AI while promoting innovation and trust in AI technologies. This regulation includes provisions for mandatory bias assessments, transparency requirements, and robust monitoring mechanisms. By setting these standards, the EU seeks to mitigate the risks of AI bias and protect fundamental rights, ensuring that AI systems are developed and deployed in a manner that is fair and beneficial to all members of society.

Having described the current technological landscape in terms of stress detection in UbiComp, and highlighted the importance of addressing fairness concerns regarding personalization for stress detection, in the next section we outline the current research gaps and our contributions in this field.

1.2 Current Gaps and Contributions

In recent years, the intersection of UbiComp and AI has paved the way for innovative solutions aimed at monitoring and managing stress in our daily lives. However, despite the promise of these advancements, significant challenges persist in ensuring the accuracy, fairness, and inclusivity of stress detection models. This problem originates by the fact that biases in the training data are not extensively explored for stress detection applications and bidirectional

[3] https://apnews.com/article/black-racial-bias-lung-medical-diagnosis-e1f73be6d00f17091600b6f21f20264d

[4] https://www.benton.org/headlines/machine-bias-theres-software-used-across-country-predict-future-criminals-and-its-biased

assessment of personalized models' accuracy and fairness is lacking for stress detection applications. So, we highlight two key limitations that hinder the development of equitable stress detection systems:

- **(L1) Ingested hidden biases in the training data are not extensively explored for stress detection applications:** Existing works regarding stress detection often do not adequately explore and address biases in the training data, leading to the development of models that provide skewed results and unfair outcomes. Without a clear understanding and visualization of biases in the data, it is challenging to develop models that are both accurate and fair across different demographics and user groups.
- **(L2) Bidirectional assessment of personalized models' accuracy and fairness is lacking for stress detection applications:** Current works give emphasis on achieving high accuracy in stress detection models and these models might perform well on average but potentially overlook biases that disproportionately impact specific demographics. This focus on accuracy without consideration for fairness can lead to biased outcomes and limit the generalizability of the models for real-world applications.

This study aims to propose a more equitable approach to stress detection systems by addressing the aforementioned limitations. Current stress detection models are hindered by hidden biases in training data and a lack of focus on fairness alongside accuracy. By analyzing data biases in two UbiComp datasets through visualizations and benchmarking the trade-off between accuracy and fairness in foundational and personalized ML models for stress detection, we aim to address the current research gaps. So, our key contributions in this domain are the following:

- **(C1) A multifaceted analysis of data biases in two UbiComp datasets for enhanced bias awareness through visualizations:** We examine two UbiComp datasets, the LifeSnaps and the SWELL-KW dataset, to identify various types of data biases that can influence stress detection models, like sampling bias, representation bias or measurement bias. In order to identify data biases we create clear and comprehensive visualizations including bar plots and heatmaps, that depict the distribution of different demographic groups within the datasets and the distribution of the label (stressed vs not stressed) in those datasets. By making the biases in the training data explicit, we enable better-informed decisions during the model training process.

- **(C2) A benchmarking for accuracy and fairness trade-off in foundational and personalized ML models for stress detection:** We establish a benchmarking framework that evaluates the trade-off between accuracy and fairness in both foundational and personalized ML models. This framework includes multiple performance metrics, such as accuracy and F1-score, and fairness metrics like Error Rate Difference (ERD), False Omission Rate Difference (FORD), Statistical Parity Difference (SPD) and Negative Predicted Value Difference (NPVD). Using those metrics, we quantify the performance of the models in terms of accuracy and fairness across different demographic groups and we examine if the models can be both accurate and fair with minimum biases.

In conclusion, this thesis makes significant contributions to the field of personalized machine learning for stress detection by addressing critical limitations in terms of fairness evaluation. Through a multifaceted analysis of data biases in stress detection datasents and bidirectional assessment of personalized models' accuracy and fairness, we provide a robust framework for developing equitable stress detection models.

1.3 Thesis Structure

The rest of the thesis is organized as follows: In Chapter 2, we provide the fundamentals of stress in UbiComp and Machine Learning (ML) techniques that are used for stress detection, as well as the basics of fairness in ML and how it can be measured. In Chapter 3, we present the related work, in terms of data collection techniques for stress detection applications, ML and personalized ML algorithms and finally fairness evaluation of ML algorithms. In Chapter 4, we describe our methodology, in terms of the datasets and the ML models that were used and how we measure the accuracy and the fairness of the models. In Chapter 5, we present the results of our experimentation in terms of performance of the personalized ML models, data bias and model bias evaluation and in terms of the trade-Off between accuracy and fairness of the personalized ML models. Lastly, in Chapter 6, we conclude our work by providing the key findings and the limitations of research followed by suggesting future work that can be done in the domain of our research.

Chapter 2

Fundamentals

This chapter establishes the foundation for understanding stress detection in UbiComp. First, we explore UbiComp's growing role in mental health and wellness, with a focus on stress detection (Section 2.1). Then, we delve into how stress is measured, examining the use of wearable devices to capture relevant physiological data in UbiComp environments (Section 2.2). We also introduce key ML concepts for stress detection, including classification and personalization techniques (Section 2.3). Finally, we address fairness considerations in stress detection models to ensure equitable outcomes (Section 2.4).

2.1 Ubiquitous Computing

UbiComp, also known as pervasive computing, refers to a concept where computing technology is seamlessly integrated into the everyday environment and becomes an integral part of people's lives. UbiComp has become more prevalent with advancements in device miniaturization, wireless communication, and sensor technologies. Wearable devices are considered a significant part of UbiComp. They are designed to be worn or carried by individuals, providing a continuous and unobtrusive connection to computing resources and services.

UbiComp offers diverse applications for promoting (mental) health and well-being, with the most popular ones listed below, among others.

- **Sleep Monitoring:** Wearable devices can track sleep patterns, including sleep duration, sleep stages (light, deep, REM), and sleep disruptions. This data can be analyzed to identify potential sleep problems and inform personalized recommendations for improving sleep habits or optimizing sleep quality.
- **Activity Tracking:** Fitness trackers and smartwatches can monitor steps taken, distance traveled, calories burned, and activity intensity. This data can motivate individuals to engage in physical activity and maintain a healthy lifestyle.
- **Chronic Disease Management:** UbiComp can be used to develop systems for monitoring chronic health conditions such as diabetes or heart disease. Smart devices can collect vital signs, medication adherence data, and other relevant information, allowing for remote patient monitoring and improved disease management.
- **Stress Detection:** Sensors that capture physiological data like heart rate monitors, EDA sensors, and respiration monitors can be integrated into wearable devices. This data can be analyzed to detect stress markers and provide real-time feedback or trigger stress management interventions.

Stress Detection in Ubicomp is crucial to further examine since stress is a prevalent issue in modern society, impacting individuals' physical and mental well-being.

2.2 Stress Detection in UbiComp

According to the World Health Organization [36], stress can be defined as a state of worry or mental tension caused by a difficult situation. While moderate levels of stress can motivate and enhance performance, chronic or excessive stress can have detrimental effects on mental and physical health, leading to symptoms such as anxiety, fatigue, irritability, and impaired cognitive function. Thus, effective stress detection and management are crucial for maintaining overall well-being. In Subsection 2.2.1, we will explore various physiological and psychological measures of stress, followed by a discussion on wearable technologies for stress detection in Subsection 2.2.2.

2.2.1 Stress Measurements

Stress can be assessed through a variety of physiological signs and psychological questionnaires. Physiological signs and measures may include adrenal assessment, heart rate, heart rate variability, blood pressure, electroencephalography (EEG), breathing patterns, skin conductance, temperature, and sleep tracking, among others. Adrenal assessment examines stress hormones like adrenaline and cortisol, while heart rate and heart rate variability reflect changes in the autonomic nervous system. Blood pressure tends to rise under stress, and EEG can indicate mental states. Breathing patterns, skin conductance, temperature, and sleep tracking also provide insights into stress levels.

Psychological questionnaires, such as the Perceived Stress Scale (PSS) [7], the Perceived Stress Questionnaire (PSQ) [20], and the Ardell Wellness Stress Test [1], measure perceived stress based on self-reported feelings and experiences. While physiological measures offer objective data, questionnaires capture subjective perceptions of stress, setting them as an information-rich combination for further analysis.

Having explained the different types of stress measurements, in the next subsection, we describe the devices that can be used to collect these stress measurements.

2.2.2 Wearables for Stress Detection

UbiComp technologies, particularly those leveraging wearable physiological sensors and ML algorithms, play a significant role in enabling real-time stress monitoring and management.

Modern smartwatches, like the E4 wristband [1] [4], feature sensors capable of capturing high-resolution physiological data in real-time. These include a Photoplethysmogram (PPG) sensor for heart rate variability, an EDA sensor, a 3-axis accelerometer for motion tracking, and an infrared thermopile for peripheral skin temperature.

Similarly, the Fitbit Sense [2] boasts a range of sensors, including accelerometers, gyroscopes, heart rate trackers, skin temperature sensors, and ambient light sensors. These devices measure parameters closely linked to stress levels, such as heart rate variability, skin conductance, and activity levels.

[1] https://www.empatica.com/research/e4/
[2] https://www.fitbit.com/global/nl/products/smartwatches/sense2

By leveraging these measurements effectively, wearable devices enable real-time stress monitoring in everyday settings. Data scientists can exploit this rich data through ML algorithms to identify patterns and correlations between physiological responses and stress levels. This knowledge can be used to develop personalized stress management strategies and provide real-time feedback to users, empowering them to take preventative measures and improve their overall well-being.

In Section 2.3 we will discuss the integration of ML and personalization techniques for stress detection, highlighting how these approaches enhance the effectiveness of UbiComp systems.

2.3 Machine Learning and Personalization for Stress Detection

Building upon the physiological and wearable technologies discussed in Section 2.2, this section delves into the application of ML algorithms and personalization techniques in stress detection systems. By analyzing the rich data collected from wearables, ML algorithms can identify patterns and make predictions about an individual's stress levels, enabling more accurate and timely interventions.

Subsection 2.3.1 will provide an overview of classification algorithms used in stress detection, explaining how they categorize different stress levels based on physiological data. Subsection 2.3.2 will discuss clustering techniques, which group similar stress responses together without predefined labels, offering insights into natural stress patterns. Subsection 2.3.3 will focus on the personalization of stress detection models, detailing how personalized models are tailored to individual users to improve accuracy and relevance.

Classification algorithms are essential for identifying and categorizing different stress levels, providing a clear framework for understanding stress responses. Clustering techniques are valuable for uncovering natural groupings in the data, which can reveal underlying patterns and relationships. Personalization techniques are necessary to account for individual differences in stress responses, making the systems more user-specific and effective. These algorithms are used in order to build personalized ML models for stress detection. The models that are used in our work are explained in the Methodology chapter (Section 4.2).

2.3.1 Classification

Classification is a fundamental task in supervised ML where the goal is to assign predefined labels or categories to input data. In the context of stress detection, classification algorithms can be trained in order to detect if a person is stressed or not based on various input features. These features may include physiological signals (e.g. heart rate variability, skin conductance), behavioral data (e.g. activity levels, sleep patterns), and contextual information (e.g. time of day, location).

Some popular ML classifiers that have been proven effective for stress detection are listed below.

Logistic Regression [19] is a simple yet powerful algorithm that works well for binary classification problems (two classes). It models the relationship between features and a binary outcome using a linear function. The outcome is typically the probability of belonging to a specific class (e.g., stressed vs. not stressed). Its appeal lies in its simplicity and interpretability, coupled with computational efficiency, rendering it a preferred choice for various classification tasks and serving as a dependable baseline model for more sophisticated algorithms. Nevertheless, logistic regression's applicability is limited to binary classification tasks, and its performance may falter in scenarios involving multi-class problems or non-linear relationships between features and outcomes, particularly with highly complex data requiring nonlinear models.

Support Vector Machines (SVMs) [3] aim to find a hyperplane (decision boundary) in the feature space that best separates the data points of different classes with the maximum margin. SVMs can be used for both linear and non-linear classification by applying kernel functions. Their effectiveness lies in handling high-dimensional data and small datasets while efficiently dealing with non-linear data patterns. However, SVMs can be computationally expensive for large datasets, and interpreting their decisions can pose challenges due to the complex nature of the separating hyperplane.

Decision Trees [31] work by splitting the data based on a series of rules (decision rules) learned from the training data. Each split in the tree is based on the feature that best separates the data points according to their class labels. They are easy to interpret, they are robust to irrelevant features and they can handle both categorical and numerical data. On the other hand, they are prone to overfitting if not properly pruned and can be unstable with small changes in the data.

In the next Subsection, we will explore clustering techniques, which group similar data points together without predefined labels.

2.3.2 Clustering

Clustering is a fundamental task in ML and data analysis aiming to partition a dataset into groups or clusters such that data points within the same cluster are more similar to each other than to those in other clusters.

Clustering algorithms are essential in the UbiComp since they can be used for personalization, which we discuss in the next subsection. By grouping users with similar preferences, behaviors, or characteristics into groups, clustering algorithms enable personalized experiences and recommendations tailored to individual user needs and preferences.

Some clustering algorithms that have been proven suitable to group users into groups are:

K-Means [37] is one of the most commonly used clustering algorithms. It partitions the dataset into K clusters by iteratively assigning data points to the nearest centroid and updating the centroids based on the mean of the points in each cluster. Its strengths lie in its simplicity, making it easy to use, especially for big datasets, and its clear interpretation of cluster centers. However, it needs us to decide the number of clusters (k) upfront, and it might struggle with clusters that don't have a spherical shape. Also, outliers can throw off its results, making it sensitive to extreme data points.

Fuzzy C-Means (FCM) [11] extends the K-Means algorithm by introducing membership degrees for each data point, representing the degree to which the point belongs to each cluster. The objective function of FCM minimizes the distance between data points and cluster centroids weighted by the membership degrees. This makes FCM suitable for data with overlapping or unclear cluster boundaries. However, interpreting the results of FCM can be more intricate compared to traditional clustering methods, as data points may belong to multiple clusters simultaneously, adding complexity to the analysis.

In the next Subsection, we will explore personalization techniques, which tailor ML models to individual users or specific groups, enhancing the effectiveness of stress detection systems.

2.3.3 Personalization

Personalization plays a critical role in enhancing the effectiveness of stress detection in UbiComp environments, as we also discuss in the next chapter. Personalized ML refers to the development and application of ML models that are tailored to individual users or specific groups of users. The goal is to create a personalized experience by understanding and adapting to the unique characteristics, preferences, and behaviors of each user.

In traditional ML approaches for stress detection, models are trained on a large dataset that represents a diverse population and they might not capture the unique stress response patterns of individual users. However, personalized ML recognizes that individuals within a population can have different preferences, habits, and needs. By incorporating user-specific data, personalized ML models can provide more accurate predictions and recommendations for individual users or groups of users.

Below we mention some personalization techniques that leverage user-specific data and user characteristics to create more tailored stress detection models, ultimately leading to more effective stress management interventions.

- **User-Based Splitting** [24] involves building a separate ML model for each individual user. These models are trained using the user's own data, allowing them to capture the unique stress response patterns of that particular individual.
- **Attribute Splitting** [24] involves constructing separate models for different groups based on a single attribute or multiple attributes. This approach entails training various ML algorithms on the data of each group and selecting the best-performing algorithm for each group.
- **Fuzzy Splitting:** [24] uses fuzzy clustering algorithms like FCM to assign partial membership degrees to each user in multiple clusters. A user can have varying degrees (between 0 and 1) of belonging to each cluster. This allows for more nuanced groupings that better reflect individual characteristics.

In the pursuit of enhancing stress detection within Ubicomp environments, personalization emerges as a pivotal strategy, as elaborated in this section. But when we're dealing with personal data, like this, we have to be careful to treat it respectfully to ensure fairness and mitigate the risk of perpetuating biases in ML models, a topic we delve into further in the next section.

2.4 Fairness in UbiComp

Fairness in ML refers to the ethical consideration of ensuring that the use of data, ML models, and algorithms does not lead to biased or discriminatory outcomes, particularly with respect to sensitive attributes such as race, gender, ethnicity, age, or other characteristics. Understanding and identifying biases is crucial for effectively measuring and mitigating them in ML systems in order to develop and deploy applications that treat individuals fairly and do not perpetuate or exacerbate existing historical, social, or other types of biases.

It is important to note that addressing fairness and mitigating bias are fundamentally intertwined; ensuring fairness in stress detection models inherently involves identifying and reducing biases in the data and algorithms.

With that said, different types of biases can be introduced in each step of the ML pipeline. More specifically, biases can arise during data collection (Subsection 2.4.1) and also during model building and implementation (Subsection 2.4.2). Let's examine each step separately, adopting phases from Yfantidou et al. [39] and definitions from Srinivasan and Chander [32].

We will later explain in the Methodology chapter, how we leverage this knowledge to address the current research gaps and develop our contributions, which are mentioned in the Introduction chapter (Section 1.2).

2.4.1 Biases During Data Collection

Firstly, biases can arise during the creation of the datasets, namely sampling bias, historical bias, measurement bias, label bias and negative set bias.

Sampling bias occurs when a dataset is formed by favoring certain types of instances over others, leading to an inaccurate representation of the real-world diversity. For example, a face-recognition algorithm may be fed with more photos of light-skinned faces than dark skinned faces, thereby leading to poor performance in recognizing darker skinned faces.

Historical bias occurs when data reflect historical prejudices, stereotypes, or inequalities, even if the data are sampled correctly. For example, a language model trained on historical texts may associate certain professions predominantly with one gender due to past societal norms, such as linking "nurse" with women and "engineer" with men, despite contemporary efforts toward gender equality in these fields.

Measurement bias arises when errors occur in human measurements or due to inherent habits in how people collect data. For example, in case of the creation of image and video datasets, the images or videos may reflect the techniques used by the photographers. Some photographers might tend to take pictures of objects in similar ways and as a result, the dataset may contain object views from certain angles only.

Label bias happens when there are inconsistencies in the labeling process. For instance, different annotators could assign differing labels to the same type of object, e.g. grass and lawn, painting and picture.

Lastly, **negative set bias** or **representation bias** occurs in a dataset as a consequence of not having enough samples representative of "the rest of the world." Datasets define something by what it is but also by what it is not. Insufficient representation in the negative set can lead to poor performance of learned classifiers, particularly in detecting negative instances.

A set of these biases are quantified through various plots in order to detect biases in the LifeSnaps and the SWELL-KW dataset, as explained in the Methodology chapter (Section 4.3). The results of the data bias analysis are presented in the Experimentation & Results chapter (Section 5.2). Next, we will explore biases that can emerge during the model-building phase, shedding light on how these biases influence the performance and fairness of stress detection systems.

2.4.2 Biases During Model Building

As stated, biases may also arise during model building, specifically confounding bias and design-related bias.

Confounding bias occurs when the model learns incorrect relationships by either neglecting certain information in the data or failing to capture the essential connections between features and target outputs. This bias stems from shared factors that influence both the input variables and the output predictions, leading to a misalignment between the model's learned associations and the true underlying patterns in the data.

Omitted bias is a special type of confounding bias where relevant features are missing from the analysis. This is similar to model underfitting, where the model itself is too simple to capture all the important influences.

Indirect bias, meaning that even if decision-making does not explicitly consider sensitive variables like race and gender, other variables used in the analysis might act as "proxies" for these sensitive factors. For instance, zip code could be indicative of race, as certain racial groups might predominantly reside in specific neighborhoods.

There are also other types of biases like algorithm bias, and ranking bias that are not further discussed in this study. Having explored the different types of biases, we use certain fairness metrics in order to quantify those biases. These fairness metrics are discussed in the next subsection. In the Methodology chapter (Section 4.3) we explain how we quantify those biases in order to identify biases hidden in the datasets and in the Experimentation & Results (Section 5.2) chapter we showcase the results of the data bias evaluation through various visualizations.

2.4.3 Fairness Metrics

Evaluating the fairness of ML algorithms quantifying the above-described biases requires the use of appropriate fairness metrics tailored to assess disparate impacts across demographic groups. These metrics provide insights into various aspects of fairness and help identify and mitigate biases in model predictions. Some fairness metrics [8] commonly used in the evaluation of ML models and are also used in our work, are described below.

Error Rate Difference (ERD)

The Error Rate (ER) measures the proportion of incorrect predictions made by the model. It is calculated as:

$$\text{ER} = \frac{\text{Number of Incorrect Predictions}}{\text{Total Number of Predictions}} \tag{2.1}$$

The Error Rate Difference (ERD) measures the difference in error rates between two groups. It is calculated as:

$$\text{ERD} = \frac{\text{ER}_{\text{unprivileged}} - \text{ER}_{\text{privileged}}}{2} \tag{2.2}$$

False Omission Rate Difference (FORD)

The False Omission Rate (FOR) measures the proportion of false negatives among the negative predictions made by the model. It is calculated as:

$$\text{FOR} = \frac{\text{Number of False Negatives}}{\text{Number of Negative Predictions}} \tag{2.3}$$

In our case, a negative predicted instance correlates to having stress, while a positive prediction would be a prediction of not having stress.

The False Omission Rate Difference (FORD) measures the difference in false omission rates between two groups. It is calculated as:

$$\text{FORD} = \text{FOR}_{\text{unprivileged}} - \text{FOR}_{\text{privileged}} \tag{2.4}$$

Statistical Parity Difference (SPD)

The Statistical Parity Difference (SPD) measures the difference in positive outcome rates between two groups. It is calculated as:

$$\text{SPD} = \Pr(Y = 1|D = \text{unprivileged}) - \Pr(Y = 1|D = \text{privileged}) \tag{2.5}$$

Negative Predicted Value Difference (NPVD)

The Negative Predicted Value (NPV) is the proportion of true negative instances among the total predicted negative instances. It is calculated as:

$$\text{NPV} = \frac{\text{True Negatives}}{\text{Number of Negative Predictions}} \tag{2.6}$$

The Negative Predicted Value Difference (NPVD) measures the difference in negative predicted values between two groups. It is calculated as:

$$\text{NPVD} = \text{NPV}_{\text{unprivileged}} - \text{NPV}_{\text{privileged}} \tag{2.7}$$

By analyzing these metrics, we can gain insights into potential biases within the models and strive to develop fairer algorithms. In the Methodology chapter (Section 5.3) we explain how we utilize those metrics to evaluate the fairness of ML models and in the Experimentation & Results chapter (Section 5.3) we showcase the results of the model bias evaluation, meaning the values of these metrics.

In this chapter, we explored the fundamentals of UbiComp, ML for stress detection, and the various biases that arise in the ML pipeline. These foundational concepts set the stage for understanding the complexities involved in developing equitable and effective ML systems for

stress detection. In the next chapter, we delve into related work on this dissertation's subject, exploring existing research and identifying potential areas for further contribution. By reviewing the state-of-the-art in stress detection and fairness in ML, we aim to build a comprehensive understanding of the field and highlight the gaps that our work seeks to address.

Chapter 3

Related Work

In this chapter, we present relevant research papers and related work to this dissertation's subject. First, we discuss works that deal with data collection for stress monitoring in different domains (Section 3.1). Then, we focus on those using generic ML algorithms for stress detection (Section 3.2). Afterwards, we present works that apply personalized ML techniques (Section 3.3) and lastly, works that evaluate the fairness of ML algorithms for stress detection (Section 3.4).

As shown in Figure 3.1, the wearable data for stress detection are collected in various environments, like during driving conditions, academic or office environments, hospital working environments and diverse environments meaning in everyday settings. For the ML algorithms for stress detection traditional ML techniques have been used, along with ensemble methods and deep learning techniques. In terms of personalization, various methods have been employed like Long Short-Term Memory (LSTM) models, Convolutional Neural Networks (CNNs), single-task neural networks (ST-NN) and multi-task neural networks (MT-NN), Self-Organizing Maps (SOMs) and others. Finally, for fairness evaluation current works mention fairness toolkits, such as FairLearn, aif360, and Aequitas for implementations of pre-, in-, and post-processing bias mitigation algorithms and fairness metrics.

3.1 Wearable Data for Stress Detection

Data is a crucial part of the ML pipeline and there have been numerous studies regarding data collection for stress prediction in various environments. We need to explore various types of

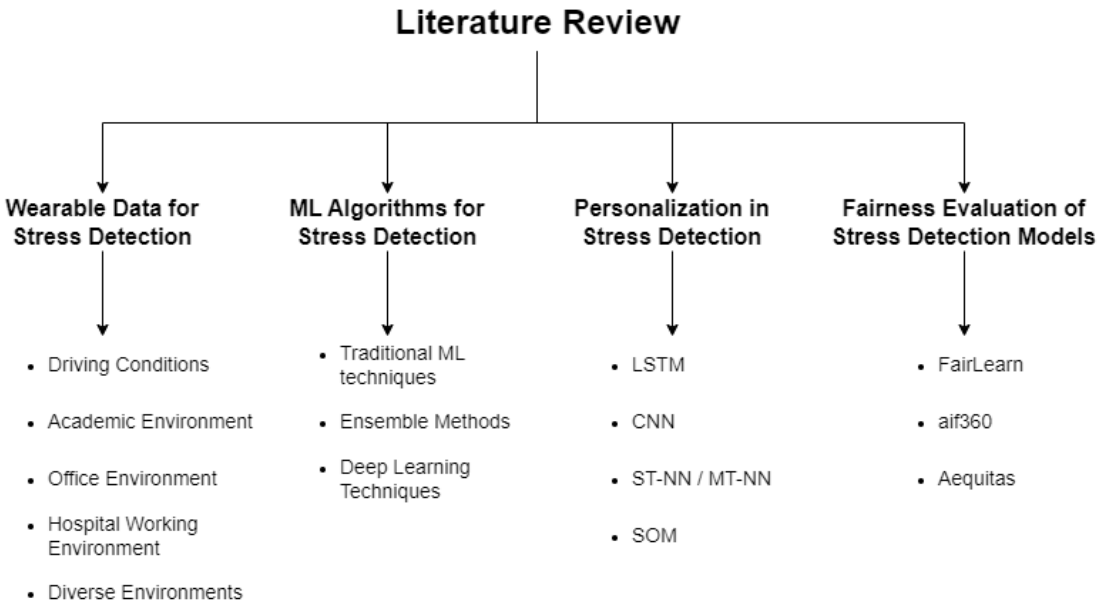

FIGURE 3.1: Pipeline of the literature review.

datasets to see if the biases differ depending on the different characteristics of the datasets. This way we address the first limitation mentioned in the Introduction chapter (Section 1.2), which is that hidden biases in the training data are not extensively explored for stress detection applications and our contribution is a multifaceted analysis of data biases in two UbiComp datasets for enhanced bias awareness through visualizations.

Stress Detection in Different Driving Conditions. Healey and Picard [13] collected and analyzed physiological data during real-world driving tasks to determine a driver's relative stress level. Data from 24 drives of at least 50-min duration were collected for analysis. Similarly, Haouij et al. [12] studied the effect of driving conditions on the stress levels of 10 drivers and created the AffectiveRoad dataset. This study was conducted with drivers taking a 1 hour 26 minute driving test. Two types of wireless physiological sensors were used to monitor the EDA, the heart rate, the skin temperature, the respiration, and the hand movement of the driver. Moreover, a sensor network was developed allowing to capture the ambient temperature, humidity, pressure, and luminosity.

Stress Detection in Academic Environment. Sano et al. [27] detected stress in academics by collecting extensive subjective and objective data using mobile phones, surveys, and wearable sensors worn day and night from 66 participants, for 30 days each, totaling 1,980 days of data. Similarly, Schmidt et al. [28] created the WESAD dataset by studying the stress of 15 students watching a movie and taking a TSST test [17]. On the same note, MDPSD (multi-modal

dataset for psychological stress detection) [6] is a comprehensive multimodal stress detection dataset on university students using EDA of skin and photoplethysmography (PPG) signals while performing different tests (e.g., Trier Social Stress Test TSST [17]). 120 participants of different genders and ages were recruited from universities to participate in the experiment.

Stress Detection in Office-Like Working Environment. Castaldo et al. [5] acquired data from 42 students using a 3-lead electrocardiogram (ECG) on two different days: the first recording was performed during an ongoing university verbal examination (stress phase), while the second one was taken in controlled resting condition (rest phase) after a vacation. Similarly, Koldijk et al. [18] created the SWELL Knowledge Work Dataset for Stress and User Modeling Research. The dataset was collected in an experiment, in which 25 people performed typical knowledge work, like writing reports, making presentations, reading e-mail, searching for information. Their working conditions were manipulated with the stressors, like email interruptions and time pressure. A varied set of data was recorded: computer logging, facial expression from camera recordings, body postures from a Kinect 3D sensor and heart rate (variability) and skin conductance from body sensors.

Stress Detection in Hospital Working Environment. Hosseini et al. [14] collected biometric data of nurses during the COVID-19 outbreak. Data was gathered for approximately one week from 15 female nurses working regular shifts at a hospital. The age of the nurses ranged from 30 to 55 years.

Stress Detection in Diverse Environments. Yfantidou et al. [38] created LifeSnaps, a multimodal dataset containing anthropological data collected unobtrusively from 71 participants in the course of more than 4 months. The participants contributed their data through validated surveys, ecological momentary assessments, and a Fitbit Sense smartwatch.

In conclusion, this section showcased a variety of studies across diverse environments that successfully captured wearable data for stress detection. These studies collected valuable physiological and behavioral information, paving the way for stress monitoring. However, a significant limitation in these works, as mentioned in the Introduction chapter (Section 1.2) is that they do not adequately explore and address potential biases in the training data. This oversight can lead to the development of models that provide skewed results and unfair outcomes. Our proposed approach addresses these limitations by leveraging data from multiple datasets, in the lab and in the wild, and provides a data bias analysis, giving a clearer picture about the accuracy and fairness in stress prediction.

In the next section, we are going to explore how different works utilized collected data in order to detect stress and other mental states.

3.2 Machine Learning Algorithms for Stress Detection

As demonstrated in the previous section, various studies have delved into wearable data collection across diverse environments for stress detection. This data captures valuable physiological and behavioral information, offering opportunities for real-time and unobtrusive stress monitoring.

However, extracting meaningful insights from this rich data stream requires powerful analytical tools. Here's where ML algorithms come into play. In this section, we dive into the diverse ML algorithms employed in order to detect stress and other mental states, like depressed mood or pain. Gedam and Paul [10] presents a comprehensive review which focuses on stress detection using wearable sensors and ML techniques.

Several studies have adopted **traditional ML techniques** to detect stress and other mental states. Logistic Regression, Decision Trees, KNNs and SVMs have been commonly used for these types of classification problems. For example, Saeed and Trajanovski [26] used Logistic Regression for driver stress detection using physiological signals, while Sultana et al. [33] used Logistic Regression and Decision Trees for pain assesement through wearables. Attaran et al. [2] and Tazarv et al. [34] applied the KNN algorithm in order to detect stress. Karagianni et al. [15] exploited Linear Regression and Decision Tree in order to infer user states from passively collected sensing data. Lastly, multiple previous studies have utilized SVMs for mental state inference ([29], [30], [33], [2], [26] and [34]). Beyond traditional ML techniques, other studies have experimented with different approaches like a Naive Bayes classifier and Elastic Net Regression. More precisely, Naive Bayes model can be found in the work of Sultana et al. [33] and have been used for pain assessment. Furthermore, Shah et al. [29] applied Elastic Net regression, to combine both L1 regularization (lasso) and L2 regularization (ridge) in order to overcome some of the limitations associated with each of these regularization techniques when used independently, to detect depressed mood.

Ensemble methods combine multiple weaker models to create a more robust and accurate prediction system. Random Forest, Gradient Boosting, AdaBoost, XGBoost, and Voting Regressor are popular choices for mental state prediction. Specifically, Karagianni et al. [15] exploited

Random Forest and Ada Boost Classifier to infer user states from passively collected sensing data. Shah et al. [29] used Random Forest, Gradient Boosting, Adaboost and Voting Regressor to predict depressed mood using wearables. Sultana et al. [33] applied XGBoost for pain assessment, while [34] applied XGBoost and Random Forest for stress monitoring using wearable sensors.

Lastly, **deep learning techniques**, particularly artificial neural networks (ANNs), have gained traction due to their ability to learn complex non-linear relationships within data. Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs) are two commonly used architectures. Tazarv et al. [34] employed a multilayer perceptron (MLP) for stress monitoring, while Yfantidou et al. [40] employed a multilayer perceptron (MLP) and a 2-layered CNN, for physical activity prediction.

While these studies demonstrate the effectiveness of various ML algorithms in stress detection, they primarily focus on achieving high accuracy. However, this emphasis on accuracy often overlooks the trade-off between accuracy and fairness, leading to models that might perform well on average but disproportionately impact specific demographics. This focus on accuracy without consideration for fairness can lead to biased outcomes and limit the generalizability of the models for real-world applications, which is a limitation mentioned in the Introduction chapter (Section 1.2). Our work incorporates multiple ML algorithms, which are described in the Methodology chapter (Section 4.2) using multiple datasets, including data in the lab and in the wild (Section 4.1), to later assess the fairness of these datasets and models and explore the trade-off between accuracy and fairness.

Having explored the diverse ML algorithms used for stress detection, the next section delves into the concept of personalization and how it can further enhance stress and other mental states detection models, ultimately leading to more effective solutions tailored to each user.

3.3 Personalization in Stress Detection

In the previous sections we talked about ML algorithms that predict various mental states without taking into account that each user is different.

UbiComp refers to technologies that are seamlessly integrated into our daily lives, such as wearable devices, smart watches etc and it offers many useful functionalities, especially in the

field of digital health. These functionalities require the personalization of services for each user. The experimental results by Parousidou et al. [24] demonstrate the superiority of personalized models over "one-size-fits-all" approaches, underscoring the importance of personalization. The experimental results demonstrate the superiority of personalized models over "one-size-fits-all" approaches, underscoring the importance of personalization. Parousidou et al. [24] experimented with various ML models and the results demonstrate the superiority of personalized models over "one-size-fits-all" approaches, underscoring the importance of personalization.

As such, many works, move past generic ML algorithms and use personalized ML techniques to predict various mental health states using sensing data. Personalized ML techniques are approaches that tailor models and predictions to individual users based on their specific characteristics, preferences, and behavior.

For instance, Yfantidou et al. [40] utilized a deep Long Short-Term Memory (LSTM) model to estimate engagement levels from face images of children with Autism Spectrum Condition. They trained on data from all users, froze the network parameters, and then fine-tuned the last layer to each user group separately based on protected attributes (e.g., health condition, hypertension, joint issues, diabetes, race, BMI, gender, age). Similarly, Sultana et al. [33] created a Multi-attribute Fairness Loss (MAFL) based CNN model that accounts for sensitive attributes included in the data and fairly predicts patients' pain status while attempting to minimize discrepancies between privileged and unprivileged groups. Moreover, Shah et al. [29] modeled individual depressed mood ratings using various modalities of data (neurocognitive data, MindLog EMA data, and smartwatch lifestyle data) and employed supervised ML regression models fine-tuned using a nested CV scheme. The pipeline compares multiple ML strategies for each subject including random forest, gradient boost, AdaBoost, elastic net, SVM, and poisson regressor. The voting regressor was also used that employs the best model from all the other strategies.

Several studies have emphasized the effectiveness of personalization through different methodologies, particularly in stress detection. Saeed and Trajanovski [26] employed both single-task neural networks (ST-NN) and multi-task neural networks (MT-NN) for personalized driver stress detection. Tazarv et al. [34] performed person-based splitting and demonstrated that personalization improves prediction performance, evidenced by a higher macro-F1 score. Tervonen et al. [35] presented multiple personalized models based on an unsupervised algorithm,

the Self-Organizing Map (SOM), considering three different levels of personalization: fully personal, semi-personal, and general. Parousidou [25] explored a variety of learning methods for personalized stress detection, including user-based, single-attribute-based, multi-attribute-based, group-based models, and fuzzy group-based models.

While these studies highlight the potential of personalized models, they often focus predominantly on achieving high accuracy. This approach may overlook the critical trade-off between accuracy and fairness, resulting in models that perform well on average but fail to account for biases that disproportionately affect specific demographics. This limitation can lead to biased outcomes and restrict the generalizability of the models for real-world applications, as mentioned in the Introduction chapter (Section 1.2). Our work incorporates multiple personalized ML algorithms, which are described in the Methodology chapter (Section 4.2) using multiple datasets, including data in the lab and in the wild (Section 4.1), to later assess the fairness of these datasets and the accuracy and fairness of these models.

Having explored the utilization of personalized ML algorithms to enhance stress detection accuracy and effectiveness, the next crucial step is evaluation. Evaluating the fairness of these personalized approaches ensures their real-world feasibility and identifies areas for further improvement.

3.4 Fairness Evaluation of Stress Detection Models

When using personalized ML techniques in UbiComp for stress detection concerns about the bias and fairness of these algorithms arise. That's why fairness evaluation within the ML pipeline should be addressed [16].

Yfantidou et al. [40] centers around uncovering bias in personal informatics systems, powered by smartphones and wearables. More specifically, they conducted comparative fairness assessments between three deep learning models for physical activity prediction. They utilized Disparate Impact Ratio (DIR), which is the ratio of base or selection rates between unprivileged and privileged groups, assuming equal ability to perform physical activity across demographics. They recommend utilizing fairness toolkits, such as FairLearn, aif360, and Aequitas for implementations of pre-, in-, and post-processing bias mitigation algorithms and fairness metrics.

Moreover, Sultana et al. [33] worked on unbiased pain assessment through wearables and electronic health record (EHR) data. More specifically, they introduced Multi-attribute Fairness Loss (MAFL), a novel loss function-based CNN model that incorporates either single or multiple protected attributes. By minimizing the disparities between privileged and unprivileged groups, MAFL-CNN effectively mitigates bias in classification tasks.

These works pave the way for fairness assessment in personalized ML. In our work, we utilize the aif360 [1] python library to measure the bias of personalized ML algorithms for stress detection, using various data, in the lab and in the wild. The way we perform the fairness evaluation, both for the data and the ML algorithms, is described in the Methodology chapter (Section 4.3 and 4.4) and the results of the fairness evaluation are presented in the Experimentation & Results chapter (Section 5.2 and 5.3).

As mentioned in the Introduction chapter (Section 1.2), current stress detection models are hindered by hidden biases in training data and a lack of focus on fairness alongside accuracy. Our contributions consist of analyzing data biases in two UbiComp datasets through visualizations and benchmarking the trade-off between accuracy and fairness in foundational and personalized ML models for stress detection in order to address the current research gaps.

Table 3.1 provides a comprehensive overview of the features addressed in each paper reviewed in this chapter, comparing data sources, algorithms, and evaluation criteria.

	Data		Algorithms		Evaluation
	Data in the lab	**Data in the wild**	**Personalized AI**	**Mental Health State Inference**	**Bias and Fairness in AI**
Yfantidou et al. [40]		✓	✓		✓
Parousidou et al. [25]	✓	✓	✓	✓	
Gedam et al. [10]	✓	✓		✓	
Shah et al. [29]	✓	✓	✓	✓	
Shi et al. [30]	✓		✓	✓	
Sultana et al. [33]		✓	✓		✓
Attaran et al. [2]	✓		✓	✓	
Tervonen et al. [35]	✓	✓	✓	✓	
Saeed et al. [26]	✓	✓	✓	✓	
Tazarv et al. [34]		✓	✓	✓	
Our work	✓	✓	✓	✓	✓

Table 3.1: Features per paper

[1] https://aif360.readthedocs.io/en/stable/

The table underscores the limited focus on bias and fairness in AI across existing studies, underscoring the need for further research. Our work aims to fill this gap by analyzing data biases in two UbiComp datasets through visualizations and benchmarking the trade-off between accuracy and fairness in foundational and personalized ML models for stress detection, as mentioned in the Introduction chapter (Section 1.2).

Overall, this chapter presented a comprehensive review of stress detection using wearable data and ML. We examined various data collection environments and highlighted the limitations in generalizability due to domain-specific focuses. We also reviewed ML algorithms ranging from traditional methods to deep learning architectures, discussing their unique benefits and challenges. Furthermore, we explored the concept of personalization in stress detection, demonstrating its effectiveness over-generalized approaches. Finally, we presented works that deal with fairness assessment in UbiComp.

Having explored the related work to this dissertation's subject, in the next chapter, we will showcase the methodology we follow in order to get the results of our research.

Chapter 4

Methodology

In this chapter, we explain the methodology of our research. More specifically, we describe the datasets (Section 4.1) and the personalized ML models (Section 4.2) we use for accurate stress detection. Furthermore, we showcase how we detect biases included in the datasets (Section 4.3) and lastly, we explain how we evaluate the fairness of the ML models (Section 4.4).

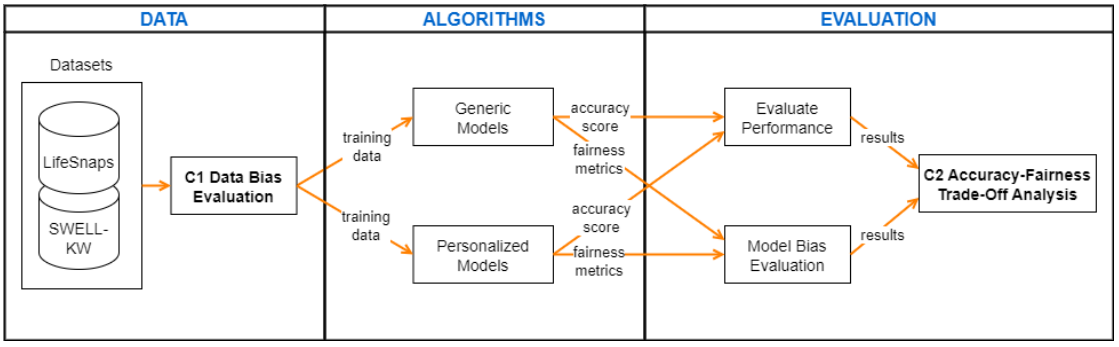

FIGURE 4.1: Pipeline of the our work.

Figure 4.1 showcases the comprehensive pipeline of our work, which addresses current limitations in stress detection research by focusing on data bias evaluation and the trade-off between accuracy and fairness of ML models. First, we take two UbiComp datasets, LifeSnaps and SWELL-KW, and analyze data biases that are hidden in the datasets through visualization techniques like bar plots and heatmaps. Next, using the data from these two datasets separately, we train both generic and personalized ML models, evaluating their performance using metrics, like accuracy and F1-score and fairness metrics, like ERD, FORD, SPD and NPVD. Finally, we analyze the trade-off between accuracy and fairness for the ML models. Through data bias exploration and trade-off between accuracy and fairness analysis for the ML models,

this pipeline aims to contribute to the development of equitable and effective stress detection systems.

4.1 Datasets in UbiComp

Following the work of Parousidou [25], the two datasets we employ for stress detection are the LifeSnaps dataset and the SWELL-KW dataset. We use these two datasets to enhance the generalizability of our claims, since the LifeSnaps dataset includes data from participants in real-world, everyday environments, providing a diverse and naturalistic view of stress-inducing situations, while the SWELL-KW dataset includes data from participants in a controlled laboratory setting. Table 4.1 presents a comprehensive comparison of the datasets used in this study in terms of number of participants, the duration of the studies, the number of instances, the device used, the environment in which the studies took place and the features of the datasets.

Dataset	**LifeSnaps**	**SWELL-KW**
Participants	71	25
Duration	4 months	3 hours
Instances	7,410	3,140
Device	Fitbit	TMSi
Environment	In-the-wild	Lab Setting
Features	Physiological	Heart Rate, Heart Rate Variability, Skin Conductance

Table 4.1: Dataset Details

Overall, the datasets include physiological data captured by wearables and some additional features for each participant. In Subsection 4.1.1 we describe the Lifesnaps dataset and in Subsection 4.1.2 the SWELL-KW dataset in more detail.

4.1.1 The LifeSnaps Dataset

The LifeSnaps dataset [38] represents a comprehensive, openly accessible collection of multimodal data captured in real-world settings, utilizing ubiquitous self-tracking technologies. Over a period of 4 months, physiological metrics such as physical activity, sleep patterns, heart rate, and temperature were unobtrusively recorded from 71 participants. These individuals, originating from diverse geographic locations including Greece (25), Cyprus (12), Italy (10), and Sweden (24), consisted of 42 men and 29 women.

Data acquisition was facilitated through a combination of Fitbit Sense smartwatches, validated surveys, and ecological momentary assessments. Participants were encouraged to wear their Fitbit devices consistently, while continuing their daily routines. In addition to physiological data, participants provided insights into their mood states ("Relaxed", "Tired", "Neutral", "Happy", "Anxious", "Alert", "Sad" and locations ("Home", "Work/School", "Outdoors", "Home Office", "Transit", "Entertainment", "Gym", "Other") through a mobile application.

Furthermore, participants completed a survey evaluating their personality traits based on the Big Five model, encompassing Conscientiousness, Extraversion, Agreeableness, Stability, and Intellect. This model, widely accepted for characterizing personality traits, has been linked to variations in exposure to stressors and the perceived severity of stressful experiences. Notably, research suggests that individuals with higher levels of conscientiousness, agreeableness, and extraversion tend to exhibit lower stress levels [19].

The dataset's features, described in Table 4.2, encompass a total of 7,410 instances, providing a rich resource for studying various aspects of human behavior and well-being.

Before applying ML algorithms, we preprocess the data following the below-described techniques.

- **Feature extraction.** The first step involves feature engineering to extract meaningful characteristics from existing data. In the LifeSnaps dataset, we utilize the "date" feature to derive new features such as "day of the month," "day of the week," "month," "week," and "year" for each data point. To capture the cyclical nature of these temporal features, we apply sine and cosine transformations. This transformation creates a representation where temporally close values, like weekdays 0 (Sunday) and 6 (Saturday), are mapped closer together in the new feature space.
- **Data imputation.** This is a crucial step when datasets contain missing values. In our case, the LifeSnaps dataset is the only one that requires imputation. We address missing values using "Single Imputer", which replaces missing numerical features with the mean value of their respective columns and missing categorical features with the most frequent value observed in each category.

Feature	Description
ID	The participant's ID.
Date	The date of the measurements.
Nightly Temperature	The night skin temperature of the user during sleep.
Heart Rate Variability	The variation of time between each heartbeat.
Daily SpO2	Daily blood oxygen saturation levels of the user.
Respiratory Rate	Data about sleep breathing.
Temperature Variation	The difference between the user's temperature and baseline.
VO2 Max	It shows how well the user's body uses oxygen.
Heart Rate	User's daily average beats per minute.
Mindfulness Session	It indicates if the user conducted a session voluntarily.
Skin Conductance	The average change in the sweat level of the skin.
Resting Heart Rate	The number of heartbeats per minute when at rest.
Sleep	10 features related to user's sleep.
Badge Type	The rewards the user earned for completing milestones.
Calories	The number of calories burnt by the individual.
Distance	Daily distance walked by the user.
Activity Type	The types of exercise the user has performed.
Active Minutes	Daily minutes the user was active.
Sedentary Minutes	Daily minutes the user was seated.
Steps	Daily number of steps walked by the individual.
Time in HR Zones	Time spent in the fat burn, cardio, or peak heart-rate zones.
Profile	User's age, gender, and BMI.
Step Goals	4 features related to the user's daily step goal range.
Mood	User's mood.
Place	User's location.
Personality	User's big five personality scores.

Table 4.2: LifeSnaps Features

- **Normalization.** This is a critical preprocessing step aimed at scaling our data to a standardized range, facilitating more accurate comparisons. In our approach, we utilize "Standard Scaler" to normalize the data. This technique standardizes features by eliminating the mean and scaling to unit variance.

- **Dimensionality Reduction.** This step becomes essential when dealing with a large number of features. In our approach, we employ Principal Component Analysis (PCA), a statistical technique that projects data onto a lower-dimensional space.

4.1.2 The SWELL-KW Dataset

This multimodal dataset was gathered as part of the SWELL Project by Koldijk et al. [18] and encompasses an experimental setting involving 25 participants (8 females and 17 males) engaging in typical knowledge work for a duration of three hours. Tasks included activities such

as report writing, presentations, email correspondence, and information retrieval. During the experiment, participants encountered two stress-inducing stimuli: time pressure and email interruptions.

Data recording encompassed facial expressions, body postures, computer usage logs, and physiological measurements, including heart rate, heart rate variability, and skin conductance, captured using a Mobi device (TMSI). The dataset comprises both raw and preprocessed data, with extracted features publicly accessible. Notably, participants refrained from consuming caffeine or smoking three to four hours prior to the experiment to mitigate potential confounding effects.

Additionally, participants completed validated questionnaires before and after the experiment. The pre-experiment questionnaire covered personality traits, habits, and demographic information such as age, gender, dominant hand, occupation, and medical history. Participants also provided information on their recent smoking, caffeine, alcohol consumption, exercise, and stress levels. Furthermore, participants filled out the "Internal Control Index" questionnaire, which assesses the extent to which individuals perceive control over events affecting their lives. This index may influence stress perception and behavioral responses.

The experiment consisted of three one-hour blocks with varying stress conditions (neutral, time pressure, interruption) separated by relaxation breaks. Each condition involved participants completing similar tasks under different constraints (e.g., no time limit vs. time pressure). The dataset comprises 3140 examples and includes ground truth labels indicating whether individuals were in a stressed state or not, along with numerical values for the recorded modalities.

Table 4.3 provides a review of the features contained in the SWELL-KW dataset.

No additional preprocessing or feature engineering was performed, as the SWELL-KW dataset was already well-prepared and suitable for immediate use

Having explored the two datasets, the next step is to train ML models on these datasets to detect stress.

4.2 Models for Stress Detection

As mentioned, we use the datasets described above to train a variety of ML models in order to find the best algorithm for accurate stress detection. To do that, we build on the work of

Feature	Description
ID	The participant's ID.
Condition	The participant's condition (relaxed, tense etc.).
Heart Rate	The participant's heart rate.
RMSSD	The participant's Root Mean Square of Successive Differences.
SCL	The participant's Skin Conductance Level.
Age	The participant's age.
Gender	The participant's gender.
Dominant hand	The participant's dominant hand.
Occupation	The participant's occupation.
Internal Control Index	The participant's Internal Control Index
Wear glasses	If the participant wears glasses.
Have heart disease	If the participant has a heart disease.
Take medicine	If the participant takes medicine.

Table 4.3: SWELL-KW Features

Parousidou [25], who benchmarked personalized ML algorithms for stress detection. The focus of your work is not to accurately predict stress but rather to explore the trade-off between the accuracy and the the fairness of these models.

In this work several categories of ML models have been leveraged, including Generic ML models (Subsection 4.2.1), User-Based splitting models (Subsection 4.2.2), Single-Attribute (Subsection 4.2.3) and Multi-Attribute Splitting (Subsection 4.2.4) models and lastly Fuzzy Splitting models (Subsection 4.2.5). The performance results of these models in terms of accuracy and F1-score are given in Section 5.1.

4.2.1 Generic ML Models

Parousidou [25] utilized the the "scikit-learn" [1] and the Pycaret python library [2] to implement the stress detection task. So, the generic ML models reproduced in this study are the ones that are part of the pycaret library. The algorithms that are tested are the following:

- **Linear Algorithms**
  - **Linear Discriminant Analysis (LDA).** A statistical technique that reduces the dimensionality of data while preserving class separation. It finds a linear combination of features that best separates different classes in the dataset.

[1] https://scikit-learn.org/stable/
[2] https://pycaret.gitbook.io/docs

  - **SVM - Linear Kernel (SVM).** An algorithm that finds the optimal hyperplane to separate data points of different classes with the maximum margin.
  - **Logistic Regression (LR).** An algorithm that estimates the probability of an input belonging to a certain class.
  - **Ridge Classifier (RC).** An algorithm that mitigates overfitting by adding a penalty term to the standard linear regression loss function.

- **Non-linear algorithms**
  - **Decision Tree Classifier (DT).** An algorithm that partitions the feature space into regions and predicts the class label of an observation by traversing the tree from the root node to a leaf node.
  - **Naive Bayes (NB).** A probabilistic classification algorithm based on Bayes' theorem, assuming independence among features.
  - **K Neighbors Classifier (kNN).** An algorithm that assigns a class label to a data point based on the majority vote of its k nearest neighbors in the feature space.
  - **Quadratic Discriminant Analysis (QDA).** A classification method that expands on LDA to accommodate non-linear decision boundaries.

- **Boosting Algorithms**
  - **Random Forest Classifier (RF).** An ensemble learning method that builds multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees.
  - **Extra Trees Classifier (ET).** A method similar to Random Forests, but it randomly selects subsets of features at each split, making it computationally more efficient.
  - **AdaBoost Classifier (ADA).** An ensemble learning method that builds a strong classifier by combining multiple weak classifiers in a sequential manner. It assigns higher weights to misclassified points in each iteration to focus on difficult instances.
  - **Gradient Boosting Classifier (GBC).** An ensemble learning technique that builds a strong classifier by sequentially adding weak learners, each correcting the errors made by its predecessor. It uses gradient descent to minimize the loss function.
  - **Light Gradient Boosting Machine (LGBM).** A variant of gradient boosting that uses a tree-based approach. It employs leaf-wise tree growth and histogram-based

algorithms to achieve faster training speed and lower memory usage compared to traditional gradient boosting methods.

To train and test the models we split the datasets into training and testing sets based on unique user IDs, ensuring that data from the same user does not appear in both sets. More, specifically, 70% of the users are included in the training set, and the remaining 30% are included in the testing set. This method ensures the model's performance is evaluated on completely unseen users.

The best model is selected based on the F1-score, a metric that is particularly useful when dealing with imbalanced datasets like ours, since the "stressed" label is underrepresented in comparison to the "not stressed" label. The best model can be found in Section 5.1, where the accuracy and fairness scores are analyzed.

A universal model, though, may not effectively capture individual differences. Thus in Subsection 4.2.2, we describe a different model specialized to each user.

4.2.2 User-Based Splitting

To capture individual differences, we adopt a personalized approach by building separate models for each individual based on their unique data. We split the dataset by participant ID and train various ML algorithms on each individual's data. For each participant, we apply 75% training and 25% testing data split and we select the algorithm that achieves the highest performance in terms of F1-score. The ML algorithms that are tested are the same as the ones for the generic model, utilizing again the "scikit-learn" and "Pycaret" libraries.

However, this approach, while effective, faces challenges when data from individual users is insufficient, leading to potential overfitting and the "cold-start" problem. Thus in Subsection 4.2.3, we describe a different personalized model, which aims to solve these problems.

4.2.3 Single-Attribute Splitting

To not rely only on methods that introduced the cold-start problem, we expand our approach, in which users are grouped based on personality traits, allowing for predictions to be made using pre-trained models for each personality group, eliminating the need for extensive labeled data for new users.

To group the users, the following steps were taken:

1. **Load Personality Features:** Personality data like the Internal Control Index from the SWELL-KW dataset or the Big Five personality traits from the LifeSnaps dataset were collected.
2. **Preprocessing:** Categorical data was converted to numerical labels using LabelEncoder [3], and all features were scaled for consistency using StandardScaler [4].
3. **Group Selection:** Techniques like the Elbow Method and Silhouette Score were used to determine the optimal number of personality-based groups (k).
4. **User Splitting:** K-means clustering was used to assign users to the k groups based on their personality features.

After grouping the users, we build a separate model for each group, similarly to the User-based Splitting method described in Sub-section 4.2.2. This involves training various ML algorithms on the data of each group and selecting the algorithm that achieves the best performance based on f1-score. The train and test set splitting method, along with the tested algorithms are the same as those used in the generic approach detailed in Section 4.1, utilizing the "scikit-learn" and "Pycaret" libraries.

While the Single-Attribute Splitting model offers promising results, we can also take more attributes into account when grouping the users. So, in Subsection 4.2.4, we describe the Multi-Attribute Splitting, which is trained on groups of users with multiple attributes in common.

4.2.4 Multi-Attribute Splitting

In addition to personality, other factors influence stress levels, prompting us to explore splitting individuals based on all available features.

To group the users, the following steps were taken:

1. **Load and Clean Features:** Additional features and physiological data statistics (mean, min, std. deviation) are loaded and irrelevant features are removed.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

2. **Combine and Preprocess:** Additional features and physiological data statistics are combined. Categorical data is encoded, numerical data is scaled, and dimensionality reduction (PCA) is applied for the LifeSnaps dataset.

3. **Group Selection:** Similar to the personality-based approach, the Elbow Method and Silhouette Score are used to determine the optimal number of groups (k).

4. **User Splitting:** K-means clustering then assigns users to these groups based on their features.

After grouping the users, we build a separate model for each group, similarly to the User-based Splitting method described in Sub-section 4.2.2 and Sub-section 4.2.4. This involves training various ML algorithms on the data of each group and selecting the algorithm that achieves the best performance based on F1-score. The train and test set splitting method, along with The tested algorithms are the same as those used in the generic approach detailed in Section 4.1, utilizing the "scikit-learn" and "Pycaret" libraries.

All the above approaches assume that there are no overlaps between the users. However, to further examine if our datasets tend to include such overlaps we examine the Fuzzy splitting model, which suggests that a user can belong in more that one group.

4.2.5 Fuzzy Splitting

To address the potential overlap of users across multiple stress-related attribute clusters, we employ "fuzzy" clustering, specifically utilizing the FCM algorithm. Unlike traditional clustering methods, FCM assigns users membership degrees for each cluster, ranging from 0 to 1, where the sum of all degrees equals 1. The clustering process utilizes the same features as described in Sub-section 4.2.4, with similar preprocessing steps. However, instead of the K-means algorithm, users are divided using FCM and instances are split based on each user's membership degree to each cluster, enabling personalized stress analysis. For instance, a percentage of instances belonging to each user are allocated to respective cluster datasets based on their membership degrees.

Following fuzzy clustering, we construct individual models for each group similar to the Multi-Attribute Splitting method. This involves training various ML algorithms on the data of each group and selecting the algorithm that demonstrates the highest performance. The train and

test splitting method, along with the ML algorithms are the same ones considered in the generic approach. Finally, to evaluate the overall performance of this approach, we calculate the average accuracy and F1-score across all groups.

In conclusion, the above ML models are the ones we utilize in order to accurately detect stress. They are later going to be used to examine the trade-off between accuracy and fairness for stress detection, which is our second contribution mentioned in the Introduction Chapter (Section 1.2). But before we evaluate the model bias, we first quantify the data bias, as described in the next Section.

4.3 Data Bias Analysis

As mentioned in the Fundamentals chapter, biases can arise during the creation of a dataset. So it is important to be able to quantify those biases through various plots. In this section, we present how we quantify biases in the LifeSnaps and the SWELL-KW datasets. To do that, we first define the protected attributes, meaning characteristics of individuals that are legally protected from discrimination and bias. These attributes can include age, gender, race, ethnicity, disability status, and more and in our case, the protected attributes are age and gender. We also need to define the privileged and the unprivileged group for each protected attribute. In our case, the unprivileged group is the female group and the age group over 30 years old, since they have the least amount of instances in the datasets.

4.3.1 Representation of the real world population

Sampling bias, as mentioned in the Fundamentals Chapter (Section 3.4) may arise when the composition of the dataset differs significantly from that of the real-world population it aims to represent. To assess this, we compare the distribution of attributes in the dataset to that of the general population. We focus on the years 2021 and 2012, corresponding to the release years of the LifeSnaps and SWELL-KW dataset, to establish the real-world distributions for gender and age. The real-world distributions for gender in Europe are 48.16% male and 51.84% female in 2021, and 48.23% male and 51.77% female in 2012. For age, the distributions are 31.79% under 30 years old and 68.21% 30 years old and above in 2021, and 34.54% under 30 years old and 65.46% 30 years old and above in 2012. The results are showcased in Subsection 5.2.1 in the Experimentation & Results Chapter.

4.3.2 Underrepresented populations

As mentioned in the Fundamentals Chapter (Section 3.4), representation biases can occur when sampling methods lead to underrepresenting populations. This occurs when certain groups are inadequately represented, leading to imbalances in specific protected attributes. For example, if a dataset contains 75% samples from a privileged group but only 25% from an unprivileged group, the underrepresented population faces significant imbalance. So, these imbalances are evident as disparities between the number of samples in the privileged group compared to those in the unprivileged group, which is the female group and the age group over 30 years old in our case. The results are showcased in Subsection 5.2.2 in the Experimentation & Results Chapter.

4.3.3 Label Distribution for Protected Attribute Groups

As mentioned in the Fundamentals Chapter (Section 3.4), measurement biases can occur when choosing, collecting, and calculating features and labels for the prediction problem. The analysis of label distribution across protected attribute groups sheds light on potential differences in the prevalence of the target label, such as stressed vs not stressed, among different demographic categories. By examining groups delineated by protected attributes like gender and age, we can discern whether certain segments of the population are disproportionately affected by the target outcome. For instance, disparities may emerge if one gender or age group exhibits a higher proportion of stressed individuals compared to others. This analysis provides valuable insights into how the target label is distributed across various demographic groups. The results are showcased in Subsection 5.2.3 in the Experimentation & Results Chapter.

4.3.4 Label distribution for missing values of protected attributes

This Subsection also refers to measurement biases. The examination of label distribution within subsets of data where the gender and age columns are missing provides insights into how the target labels, stressed versus not stressed, are distributed in these instances. This analysis allows us to discern whether there are any notable patterns or disparities in the distribution of stressed and not stressed labels within these subsets, offering valuable insights into how missing values may impact the overall representation of the target outcome. By examining the label distribution for missing values, we can better understand how these gaps affect the

preprocessing steps. For instance, if the distribution of stressed and not stressed labels in the missing data subsets differs significantly from the overall distribution, it may indicate a bias that needs to be addressed during data imputation or other preprocessing techniques. Conversely, if the label distribution in the missing values is similar to the overall distribution, it suggests that the missing data may not introduce significant bias. The results are showcased in Subsection 5.2.4 in the Experimentation & Results Chapter.

As mentioned in the Introduction Chapter (Section 1.2), a current gap in the domain of our research is that biases in the training data are not extensively explored for stress detection applications and we address this with our first contribution (Section 1.2), which is by having clear understandable visualizations in order to identify different types of biases in the training data. In the next Section, we describe how we identify biases in the ML models.

4.4 Model Bias Analysis

As mentioned in Fundamentals Chapter (Subsection 2.4.2), biases may also arise during model building in the ML pipeline. So it is important to not only analyze the bias of the datasets but also evaluate the fairness of the ML algorithms. In our work, we used the aif360 [5] library to assess the fairness of our models concerning demographic attributes such as age, gender, BMI, and occupation. The fairness metrics we utilize were also introduced in detail in Subsection 2.4.3 and they are the following:

- **Error Rate Difference (ERD):** A value close to 1 indicates that the unprivileged user group systematically receives more wrong results compared to the privileged user group.
- **False Omission Rate Difference (FORD):** A value close to 1 indicates that the unprivileged user group is systematically receiving more wrong “stress” predictions compared to the privileged user group.
- **Statistical Parity Difference (SPD):** A negative value indicates that the unprivileged user group systematically receives fewer “not stressed” predictions compared to the privileged group.

[5] https://aif360.readthedocs.io/en/latest/

- **Negative Predicted Value Difference (NPVD):** A value close to 1 indicates that the unprivileged user group is systematically receiving fewer correct "stressed" predictions compared to the privileged.

Specifically, we assess fairness in generic models, user-based models, single-attribute splitting models, multi-attribute splitting models, and fuzzy splitting models. The fairness evaluation results are interpreted to understand the implications for model performance and ethical considerations (Section 5.3). Discrepancies in fairness metrics are investigated to determine whether certain groups are disproportionately affected by model predictions. By doing that, we can examine the accuracy and fairness trade-off in foundational and personalized ML models for stress detection, which is our second contribution as mentioned in the Introduction Chapter (Section 1.2). Our contribution addresses the current limitation also mentioned in Section 1.2, which is that bidirectional assessment of personalized models' accuracy and fairness is lacking for stress detection applications.

Overall, in this chapter, we described the datasets we employ for stress detection along with the personalized ML models that are trained on these datasets. We also explained how we conduct the data bias analysis and how we interpret the fairness metrics for the model bias analysis. Having described our methodology, in the next chapter, we present the results of our research.

Chapter 5

Experimentation & Results

In this Chapter we provide the results of our research and analyze them. First, we delve into the results regarding the performance of the ML models in terms of accuracy and F1-score (Section 5.1). Then, we showcase the results of the data bias analysis, which includes various plots in order to visualize the biases hidden in the datasets (Section 5.2). Lastly, we present the results of the fairness evaluation of the models using various fairness metrics (Section 5.3) and examine the trade-off between accuracy and fairness for these models (Section 5.4).

5.1 Performance Evaluation

In this Section we provide the performance results of the ML models that we use to detect stress in terms of accuracy and F1-score. The ML models that we employ are generic models, user-based splitting models, single-attibute and multi-attribute splitting models and fuzzy splitting models as described in the Methodology Chapter (Section 4.2)

The datasets that they are trained on are the LifeSnaps dataset and the SWELL-KW dataset 4.1 and for the SWELL-KW dataset we train the models both with and without including protected attributes in the training process to see how including protected attributes affects the performance results.

As mentioned before, the ML models we use are based on Parousidou [25]'s work who benchmarked personalized ML algorithms for stress detection and are not built from scratch.

Table 5.1 displays the accuracy and Table 5.2 the F1-score attained in each dataset using the top-performing ML model from each proposed approach. The top-performing model is different for each case, based on the data that it is trained on. Across the two tables, in the case of the Generic method, the algorithms yielding the highest performance in each dataset are denoted within parentheses. Also, for each column in the tables, the highest value is noted in bold.

In more detail, the results in Table 5.1 show that the Generic method achieved an accuracy of 0.8741 with the LightGBM (LGB) model on the LifeSnaps dataset, while achieving 0.5295 with the Quadratic Discriminant Analysis (QDA) model and 0.5373 with the Logistic Regression (LR) model on the SWELL-KW dataset taking and not taking protected attributes into account for training, respectively.

For the LifeSnaps dataset, the Multi-Attribute Splitting method outperformed others, achieving an accuracy of 0.9617. For the SWELL-KW dataset, the User-based Splitting method outperformed others, achieving an accuracy of 0.8289 and 0.8163 with and without protected attributes used for training, respectively.

Remarkably, removing protected attributes from SWELL-KW for training resulted in slightly lower accuracy across all splitting methods compared to using protected attributes for training.

Method/Dataset	**LifeSnaps**	**SWELL-KW with protected attributes**	**SWELL-KW without protected attributes**
Generic	0.8741 (LGB)	0.5295 (QDA)	0.5373 (LR)
User-based Splitting	N/A	**0.8168**	**0.8055**
Single-Attribute Splitting	0.956	0.7431	0.7188
Multi-Attribute Splitting	**0.9617**	0.7631	0.7541
Fuzzy-Clustering Splitting	0.8186	0.76354	0.74058

Table 5.1: Comparative table in terms of Accuracy using the best ML model along with hyper-parameter tuning.

Table 5.2 shows the comparative F1-score results. The Generic method yielded an F1-score of 0.3836 with the LGB model on the LifeSnaps dataset, while achieving 0.6924 with the QDA model and 0.5353 with the LR model on the SWELL-KW dataset with and without protected attributes for training, respectively.

For the LifeSnaps dataset, the Single-Attribute Splitting method outperformed others, achieving an accuracy of 0.5667. The User-based Splitting method again demonstrated superior performance, achieving an F1-score of 0.8289 and 0.8163 on the SWELL-KW dataset with and without protected attributes used in training, respectively.

Again, removing protected attributes from SWELL-KW for training resulted in slightly lower accuracy across all splitting methods compared to using protected attributes for training.

Method/Dataset	**LifeSnaps**	**SWELL-KW with protected attributes**	**SWELL-KW without protected attributes**
Generic	0.3836 (LGB)	0.6924 (QDA)	0.5753 (LR)
User-based Splitting	N/A	**0.8289**	**0.8163**
Single-Attribute Splitting	**0.5667**	0.7664	0.7496
Multi-Attribute Splitting	0.55	0.7643	0.7571
Fuzzy-Clustering Splitting	0.4156	0.7743	0.7512

TABLE 5.2: Comparative table in terms of F1-score using the best ML model along with hyper-parameter tuning.

In conclusion, the top-performing personalized ML model based on the F1-score and trained on the LifeSnaps dataset is the Single-Attribute Splitting model. For the SWELL-KW dataset, the top-performing model is the User-Based Splitting model both in the case of including protected attributes and in the case not including protected attributes in the training process. Having examined the models in terms of F1-score, we continue with the evaluation of biases hidden in the datasets.

5.2 Data Bias Evaluation

In this Section we quantify the biases hidden in the LifeSnaps and the SWELL-KW datasets. As mentioned in the Introduction Chapter (Section 1.2), one of our contributions is the analysis of data biases in these two datasets for enhanced bias awareness through visualizations, since we have addressed the limitation that hidden biases in the training data are not extensively explored for stress detection applications in current works.

As explained in the Methodology Chapter (Section 4.3), first we will examine if the real-world population is well represented in the datasets (Subsection 5.2.1) and then, if there are under-represented populations in the datasets (Subsection 5.2.2). Lastly, we will explore the label distribution for protected attribute groups (Subsection 5.2.3) and the label distribution in case of missing values for protected attributes (Subsection 5.2.4).

5.2.1 Is the real-world population well represented in the datasets?

In the Methodology Chapter (Subsection 4.3.1), we stated that by visualizing the distribution of the population in the real world and in the datasets, we quantify **sampling bias** and this way we can check if the real-world population is well represented in the datasets.

From Figure 5.1 we observe that the Lifesnaps dataset does not accurately represent the real-world population regarding the "gender" protected attribute, since the real gender ratio in Europe in 2022 was 0.93 with the number of females being greater than the number of males, while the LifeSnaps gender ratio is 1.56 and 1.73 after preprocessing, with the number of males being greater than the number of females.

This means that there is **sampling bias** in the LifeSnaps dataset regarding the "gender" protected attribute, since the female group, which is the unprivileged group, is underrepresented. This discrepancy can cause fairness issues, as models trained predominantly on male data may perform poorly for females, leading to unfair outcomes. The consistency in gender ratio between the preprocessed and post-processed datasets suggests that preprocessing did not address this imbalance, perpetuating inherent biases.

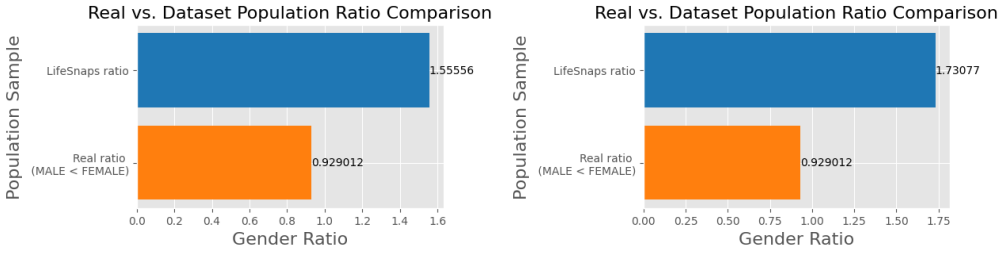

FIGURE 5.1: Gender distribution between real-world's and dataset's populations before (right) and after (left) preprocessing for the LifeSnaps dataset.

Regarding the "age" protected attribute for the LifeSnaps dataset (Figure 5.2), we observe again that the dataset does not accurately represent the real-world population, since the real age ratio in Europe in 2022 was 0.47 with the number of people over 30 being greater than the number of people under 30 years old, while the LifeSnaps gender ratio is 1.19 and 1.37 after preprocessing, with the number of people under 30 being greater than the number of people over 30 years old.

This means that there is **sampling bias** in the LifeSnaps dataset regarding the "age" protected attribute, since the group over 30 years old, which is the unprivileged group, is underrepresented. This discrepancy can cause fairness issues, as models trained predominantly on data of

people under 30 years old may perform poorly for people over 30 years old, leading to unfair outcomes. The consistency in age ratio between the preprocessed and post-processed datasets suggests that preprocessing did not address this imbalance, perpetuating inherent biases.

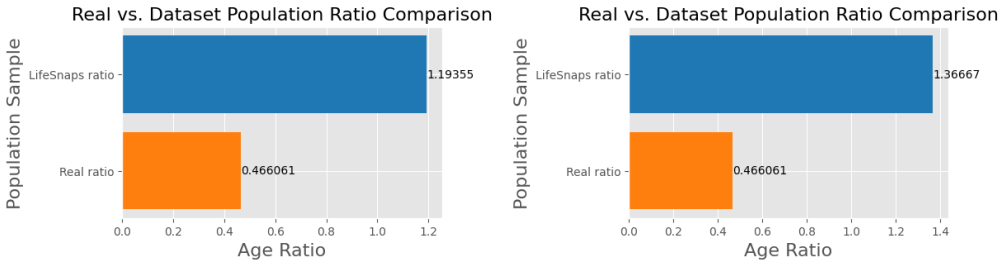

FIGURE 5.2: Age distribution between real-world's and dataset's populations before (right) and after (left) preprocessing for the LifeSnaps dataset.

From Figure 5.3, we observe that the SWELL-KW dataset also does not accurately represent the real-world population regarding the "gender" protected attribute, since the real gender ratio in Europe in 2012 was 0.93 with the number of females being greater than the number of males, while the SWELL-KW gender ratio is 2.125 with the number of males being significantly greater than the number of females.

This indicates a **sampling bias** in the SWELL-KW dataset regarding the "gender" protected attribute, with the female group, which is the unprivileged group, being underrepresented. This discrepancy can cause fairness issues, as models trained predominantly on male data may perform poorly for females, leading to unfair outcomes and the consistency in age ratio between the preprocessed and post-processed datasets suggests that preprocessing did not address this imbalance, perpetuating inherent biases.

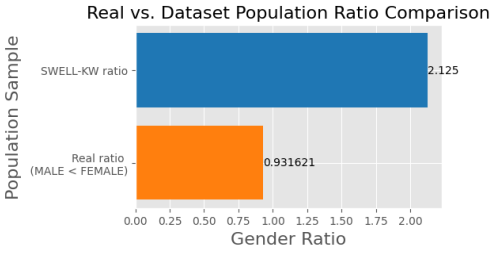

FIGURE 5.3: Gender distribution between real-world's and dataset's populations for the SWELL-KW dataset.

Regarding the "age" protected attribute for the SWELL-KW dataset (Figure 5.4), we observe that the dataset again does not accurately represent the real-world population. The real age ratio in Europe in 2012 was 0.69 with the number of people under 30 being less than the number

of people over 30 years old, while the SWELL-KW age ratio is 24, indicating a substantial overrepresentation of people under 30 years old.

This means that there is **sampling bias** in the SWELL-KW dataset regarding the "age" protected attribute, with the group over 30 years old, which is the unprivileged group, being underrepresented. This discrepancy can cause fairness issues, as models trained predominantly on data of people under 30 years old may perform poorly for people over 30 years old, leading to unfair outcomes and the consistency in age ratio between the preprocessed and post-processed datasets suggests that preprocessing did not address this imbalance, perpetuating inherent biases.

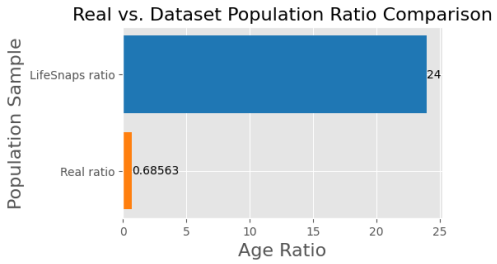

FIGURE 5.4: Age distribution between real-world's and dataset's populations for the SWELL-KW dataset.

Overall, we can say that there is **sampling bias** in the data since the ratio of females and males and the ratio of people over and under 30 years old is different in the datasets in comparison to the ratio in the real world. Knowing if there is sampling bias is important to ensure fairness, accuracy, and generalizability of the model. In the next Subsection we explore if there are underrepresented populations in the data.

5.2.2 Are there underrepresented populations?

In the Methodology Chapter (Subsection 4.3.2), we stated that by visualizing the number of samples for both the privileged and the unprivileged groups in the datasets, we quantify **representation bias** and this way we can check if all groups are adequately represented.

Fron Figure 5.5, we observe the number of samples for the privileged and the unprivileged group for the protected attributes "gender" and "age" for the LifeSnaps dataset. For the gender attribute, it is clear that the number of females, which is the unprivileged group is smaller than the number of males, which is the privileged group. More specifically, 42 males and 27 females before preprocessing and 45 males and 26 females after preprcessing. Similarly, the

number of participants over 30 years old, which is the unprivileged group, is smaller than the number of participants under 30 years old, which is the privileged group. More specifically, 37 participants under 30 years old and 31 participants over 30 years old before preprocessing and 41 participants under 30 years old and 30 participants over 30 years old after preprocessing.

The disparity in the number of samples between privileged and unprivileged groups indicates **representation bias**. This is problematic because it can lead to unfair outcomes, meaning models trained on biased data may perform poorly for underrepresented groups. For example, if the model is trained predominantly on male data, it may not generalize well to female users, resulting in less accurate predictions and potential disadvantages for females. The fact that the number of samples is more or less the same between the preprocessed and post-processed LifeSnaps datasets suggests that preprocessing did not address the representation bias. This perpetuation of bias means that any model trained on this data is likely to inherit and possibly exacerbate the existing biases, leading to continued unfair treatment of unprivileged groups.

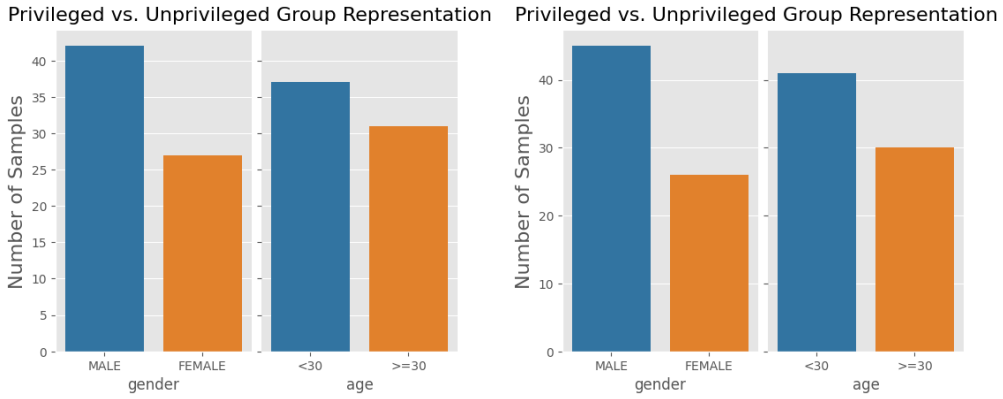

FIGURE 5.5: Underrepresented groups for each protected attribute before (left) and after (right) preprocessing for the LifeSnaps dataset.

Fron Figure 5.6, we observe the number of samples for the privileged and the unprivileged group for the protected attributes “gender” and “age” for the SWELL-KW dataset. For the gender attribute, it is clear that the number of females, which is the unprivileged group is smaller than the number of males, which is the privileged group. More specifically, 17 males and 8 females. Similarly, the number of participants over 30 years old, which is the unprivileged group, is smaller than the number of participants under 30 years old, which is the privileged group. More specifically, 24 participants under 30 years old and only 1 participants over 30 years old.

The disparity in the number of samples between privileged and unprivileged groups indicates **representation bias**. This is problematic because it can lead to unfair outcomes, meaning

models trained on biased data may perform poorly for underrepresented groups. In the context of the "gender" protected attribute, where males are the privileged group and females are the unprivileged group, a real-world application with this number of samples could result in gender-biased outcomes. Similarly, for the "age" protected attribute, where individuals under 30 years old are the privileged group and those over 30 years old are the unprivileged group, an application with this distribution could lead to age-biased results. The fact that the number of samples is more or less the same between the preprocessed and post-processed data suggests that preprocessing did not address the representation bias. This perpetuation of bias means that any model trained on this data is likely to inherit and possibly exacerbate the existing biases, leading to continued unfair treatment of unprivileged groups.

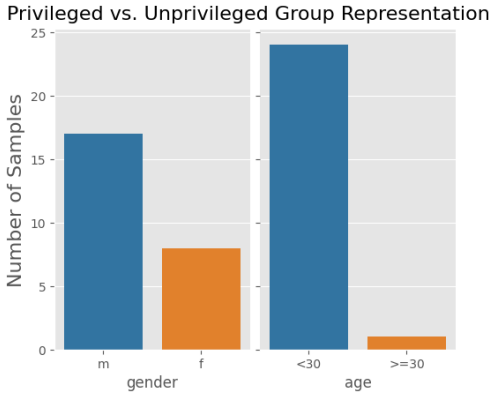

FIGURE 5.6: Underrepresented groups for each protected attribute for the SWELL-KW dataset.

From Figure 5.7 we observe the intersectional representation of protected attributes "gender" and "age" for the LifeSnaps dataset. There are 15 samples for females under 30, 22 samples for males under 30, 12 samples for females aged 30 and over, and 19 samples for males aged 30 and over before preprocessing. while after preprocessing, there are 14 samples for females under 30, 27 samples for males under 30, 12 samples for females aged 30 and over, and 18 samples for males aged 30 and over after preprocessing.

The implications of this intersectional **representation bias** are substantial. The underrepresentation of females, both under and over 30, suggests that any models trained on this dataset may perform poorly for female users, leading to gender-biased predictions and recommendations that could adversely affect female users' experiences. The overrepresentation of younger males means that the model may be overly tuned to this group, resulting in better performance for younger males compared to other groups and exacerbating biases. The significant underrepresentation of older females indicates that these individuals face both gender and age bias, potentially leading to doubly disadvantaged outcomes in model performance and fairness.

Comparing the preprocessed data with the post-processed data, we see that the preprocessing did not substantially change the distribution of samples among these intersectional groups and the slight adjustments do not address the inherent biases, meaning there is still **representation bias** after preprocessing.

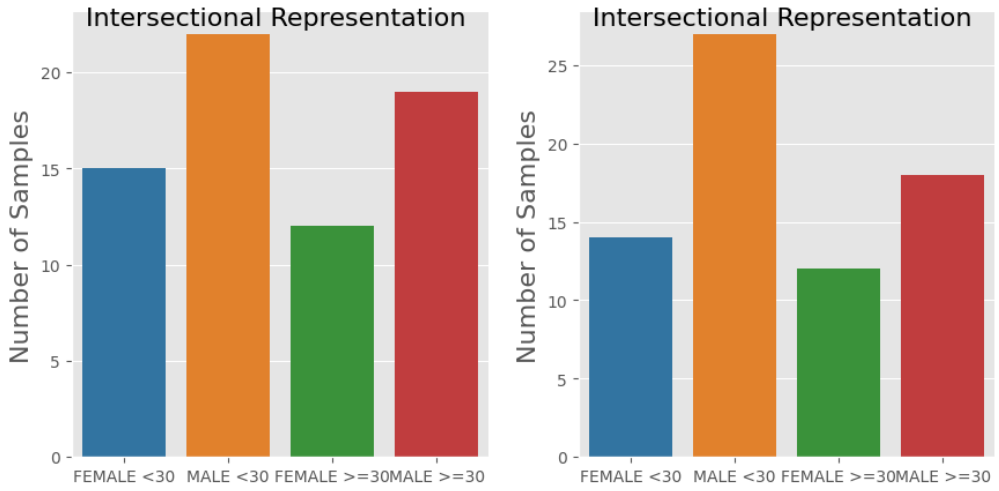

FIGURE 5.7: Intersectional representation of protected attributes before (left) and after (right) preprocessing for the LifeSnaps dataset.

Figure 5.8 shows the same information as Figure 5.7 but using a heatmap instead of a bar plot.

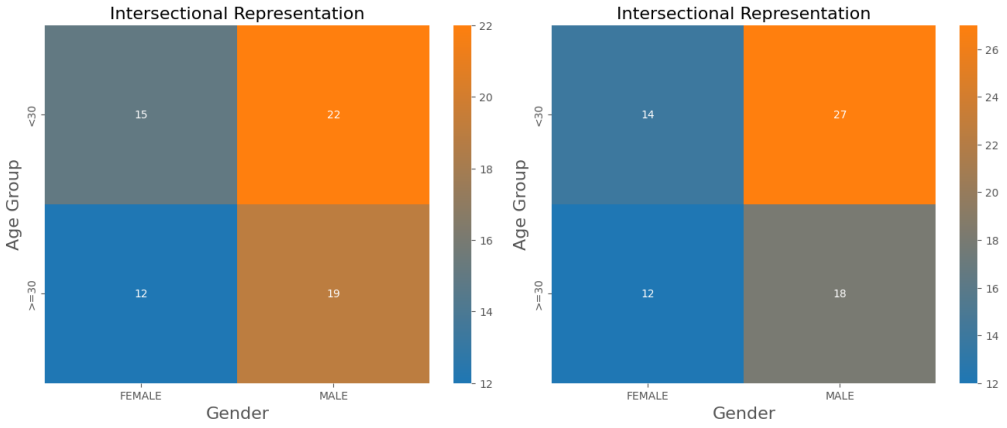

FIGURE 5.8: Intersectional representation of protected attributes before (left) and after (right) preprocessing for the LifeSnaps dataset.

Figure 5.9 shows the intersectional representation of protected attributes “gender” and “age” for the two datasets for the SWELL-KW dataset. There are 7 samples for females under 30, 17 samples for males under 30, and 1 sample for females aged 30 and over. Notably, there are no male participants aged 30 and over, indicating a complete absence of representation for this group.

This distribution reveals significant **representation bias**. The underrepresentation of females, both under 30 and especially those aged 30 and over, suggests that any models trained on this dataset may perform poorly for female users, leading to gender-biased predictions and recommendations that could negatively impact female users' experiences. The overrepresentation of younger males means that the model may be overly tuned to this group, resulting in better performance for younger males compared to other groups and exacerbating existing biases. The significant underrepresentation of older individuals, both male and female, indicates that these individuals face age bias, potentially leading to unfair outcomes in model performance and fairness. Furthermore, the absence of older male participants in the dataset is particularly concerning as it suggests that this demographic is not represented at all, leading to models that completely overlook the needs and behaviors of older males. This lack of representation can result in models that are not generalizable and fail to account for the diversity within the population.

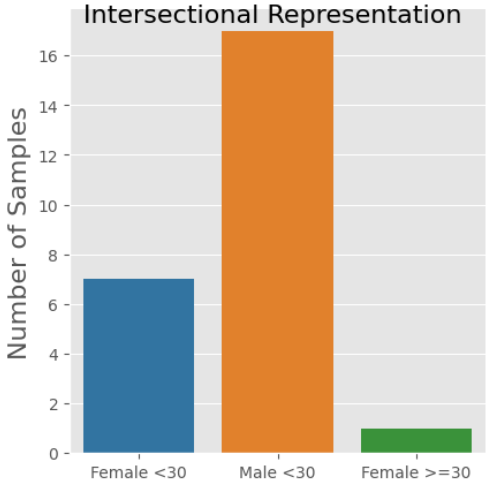

FIGURE 5.9: Intersectional representation of protected attributes for the SWELL-KW dataset.

Figure 5.10 shows the same information as Figure 5.7 but using a heatmap instead of a bar plot.

In conclusion, the analysis of the LifeSnaps and SWELL-KW datasets demonstrates significant **representation bias**, with unprivileged subgroups, like older females in LifeSnaps and older individuals entirely in SWELL-KW, being severely underrepresented. The findings underscore the critical need for more balanced and representative datasets and improved preprocessing techniques to ensure fair and equitable AI model performance across all demographic groups.

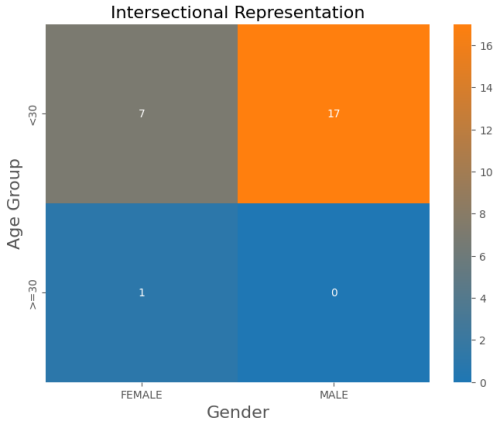

FIGURE 5.10: Intersectional representation of protected attributes for the SWELL-KW dataset.

5.2.3 What is the label distribution for protected attribute groups?

In the Methodology Chapter (Subsection 4.3.3), we stated that by visualizing the label distribution across protected attribute groups in the datasets, we quantify **measurement biases** and we can discern whether certain segments of the population are disproportionately affected by the target outcome.

Figure 5.11 shows the distribution of the target label (stressed vs. not stressed) across the different protected attributes, gender and age, for the LifeSnaps dataset. It is apparent that for the LifeSnaps dataset the number of "stressed" labels in the datasets are far less in comparison to the "not stressed" labels.

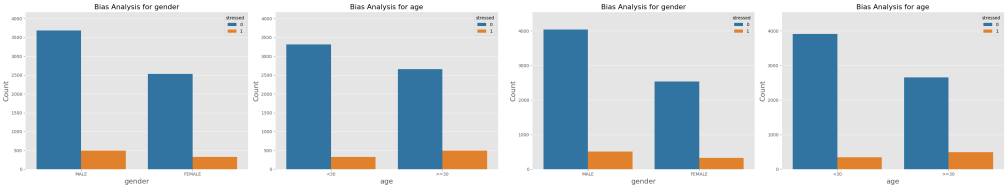

FIGURE 5.11: Label distribution across protected attribute groups before (left) and after (right) preprocessing for the LifeSnaps dataset.

Figure 5.12 shows the ratio of stressed and not stressed labels across gender and age for the LifeSnaps dataset.

These distributions indicate a **measurement bias** within the LifeSnaps dataset. The gender distribution shows that both females and males have a high percentage of "not stressed" labels, with males having slightly more "not stressed" labels compared to females. This minor difference suggests a relatively balanced distribution between the genders, though it is important to note the overall imbalance with a predominance of "not stressed" labels.

In terms of age distribution, participants under 30 years old have a much higher percentage of "not stressed" labels (91.05%) compared to those 30 and over (84.31%). Conversely, the percentage of "stressed" labels is higher in the 30 and over group (15.69%) compared to the under 30 group (8.95%). This indicates that older individuals are more likely to be labeled as "stressed" in the dataset, highlighting an age-related **measurement bias**.

Such biases can lead to unfair outcomes in stress detection models trained on this dataset, meaning that models may disproportionately identify older individuals as stressed compared to younger individuals, potentially leading to overestimations of stress levels in the older population. Similarly, the slight gender bias might affect the sensitivity of stress detection for females and males differently, though to a lesser extent compared to the age bias.

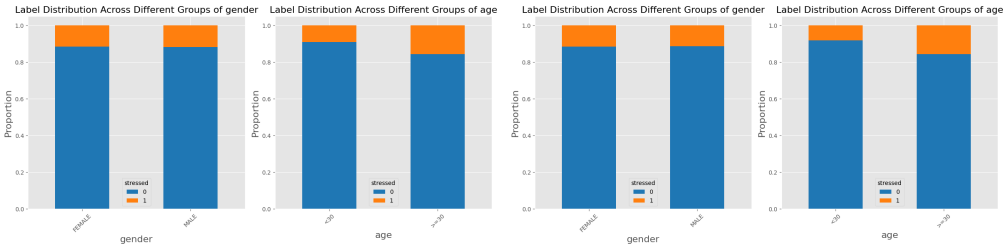

FIGURE 5.12: Label distribution across protected attribute groups before (left) and after (right) preprocessing for the LifeSnaps dataset.

Figure 5.13 shows the distribution of the target label (stressed vs. not stressed) across the different protected attributes, gender and age, for the SWELL-KW dataset. We observe that the number of stressed and not stressed labels are balanced with the number of "stressed" labels being slightly more than the "not stressed" ones.

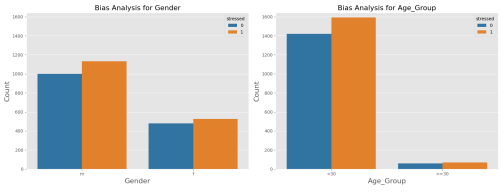

FIGURE 5.13: Label distribution across protected attribute groups for the SWELL-KW dataset.

Figure 5.14 shows the ratio of stressed and not stressed labels across gender and age for the SWELL-KW dataset. These distributions indicate that the SWELL-KW dataset has a relatively balanced distribution of the "stressed" and "not stressed" labels across both gender and age groups. The slight majority of "stressed" labels over "not stressed" labels in all groups suggests a potential bias towards labeling participants as stressed.

For gender, both females and males have nearly equal percentages of “stressed” and “not stressed” labels, with a marginally higher percentage of “stressed” labels in both groups. This balance suggests minimal **measurement bias** concerning gender in the dataset.

For age, participants under 30 and those 30 and over also show similar distributions, with slightly more “stressed” labels in both age groups. This indicates that the dataset does not exhibit significant age-related **measurement bias**, as both age groups have comparable ratios of stress labels.

However, the slight overrepresentation of “stressed” labels could lead to models being more likely to classify individuals as stressed. While this may enhance the model’s sensitivity to detecting stress, it could also result in higher false positive rates, where non-stressed individuals are misclassified as stressed.

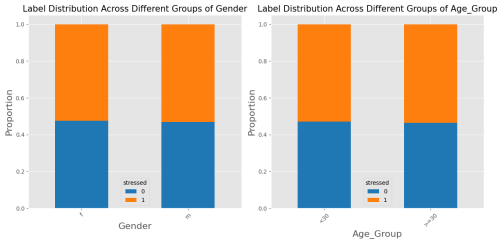

FIGURE 5.14: Label distribution across protected attribute groups for the SWELL-KW dataset.

In conclusion, while both datasets show some level of **measurement bias**, the LifeSnaps dataset exhibits more pronounced imbalances with the number of “not stressed” labels being far bigger than the number of “stressed” labels, which could significantly impact model fairness and accuracy. The SWELL-KW dataset, with its more balanced label distribution, having relatively the same number of “stressed” and “not stressed” labels, presents a lower risk of bias. In the next Subsection we explore the label distribution in case of missing values for the protected attributes.

5.2.4 What is the label distribution in the case of missing values for protected attributes?

In the Methodology Chapter (Subsection 4.3.4), we stated that by visualizing the label distribution in the case of missing values for protected attributes, we quantify **measurement biases** and we can discern how missing values may impact the overall representation of the target outcome.

Figure 5.14 shows the percentage for missing values for the LifeSnaps respectively. We observe that the dataset has missing values for the protected attributes of gender and age. More specifically, 5% of values for the gender attribute are missing and around 8% of values for the age attribute are missing. The SWELL-KW dataset has no missing values.

Missing values can introduce **measurement bias**, since the missing data is correlated with the target outcome. If we are missing values from certain gender or age groups, then the model might not learn to accurately predict stress for those groups, leading to biased predictions.

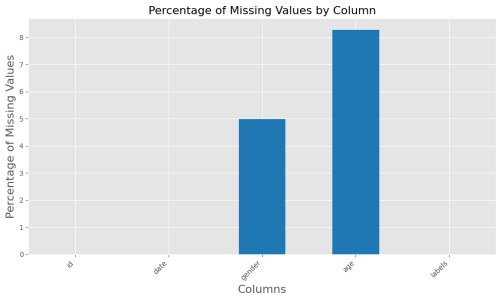

FIGURE 5.15: Distribution of missing values for the LifeSnaps dataset.

Figure 5.16 shows the distribution of the target label (stressed vs. not stressed) across the different protected attributes, gender and age, when their values are missing. In Figure 5.11 we saw the label distribution when these values are not missing. By comparing the two figures we can say that the ratio is pretty similar with the number of "stressed" labels being far less than the number of "not stressed" labels.

We can say that the label distribution for missing values of protected attributes for the LifeSnaps dataset is similar to the label distribution to the label distribution across protected attribute groups before and after preprocecing. The missing values suggest that there is again some **measurement bias** in the LifeSnaps dataset leading to unfair outcomes, which is not the case for the SWELL-KW dataset.

FIGURE 5.16: Label distribution for missing values of protected attributes for the LifeSnaps dataset.

Having no missing values, as in the SWELL-KW dataset, means that the model can leverage the full dataset for training. This reduces the risk of **measurement bias** and ensures that all demographic groups are adequately represented. With complete data, the models can learn more accurately from the entire population, leading to more reliable and fair predictions across different protected attributes.

Overall, the presence of missing values in the LifeSnaps dataset introduces some **measurement bias**, potentially affecting the fairness and accuracy of the stress detection models and handling these missing values effectively is crucial to mitigate these biases. On the other hand, the SWELL-KW dataset, with no missing values, avoids these issues, providing a more robust foundation for training fair and accurate models.

In conclusion, our analysis reveals that these datasets exhibit some biases. Specifically, the LifeSnaps dataset exhibits **sampling and representation biases** across the protected attributes "gender" and "age" and also **measurement biases**, since there were missing values for the protected attributes and the label distribution across protected attribute groups is not balanced. The SWELL-KW datasets appears to have **sampling and representation biases** across the protected attribute "gender" and these biases are even more pronounced for the protected attribute "age".

The results above regarding data bias address the limitation that existing works often do not adequately explore and address biases in the training data for stress detection, as stated in the Introduction Chapter (Section 1.2). By visualizing data bias, we aim to understand biases in the training data before training a model with the data. In the next Section we explore biases that can found in the ML models during model building.

5.3 Model Bias Evaluation

In this Section we quantify the biases hidden in ML models trained for stressed detection using the LifeSnaps and the SWELL-KW datasets. As mentioned in the Introduction Chapter (Section 1.2), one of our contributions is assessing the trade-off between accuracy and fairness, ensuring that the models we evaluate are not only accurate but also equitable. This way we address the limitation that bidirectional assessment of personalized models' accuracy and fairness is lacking for stress detection applications.

As explained in the Methodology Chapter (Section 4.4), the fairness metrics we use are the ERD, FORD, SPD and NPVD. First, we present the results for the Generic model (Subsection 5.3.1), then the User-Based Splitting model (Subsection 5.3.2), then the Single-Attribute (Subsection 5.3.3) and the Multi-Attribute Splitting model (Subsection 5.3.4) and lastly the Fuzzy Splitting model (Subsection 5.3.5).

5.3.1 Generic Models Bias Assessment

For the LifeSnaps dataset the top-performing model in terms of F1-score was an LGB model as stated in Section 5.1. The results of the fairness analysis for this model are showcased in Table 5.3. The fairness evaluation metrics collectively indicate a **relatively balanced performance** across demographic groups with some **slight disparities**.

LifeSnaps	Age	Gender	BMI
Error Rate Difference (ERD)	-0.056570	-0.044389	-0.017106
False Omission Rate Difference (FORD)	-0.013629	0.017970	0.024426
Statistical Parity Difference (SPD)	-0.085896	-0.070938	-0.042530
Negative Predicted Value Difference (NPVD)	0.013629	-0.017970	-0.024426

Table 5.3: Fairness evaluation metrics for the generic model trained on the LifeSnaps dataset

More specifically, for the age attribute, the ERD of -0.057 suggests that the privileged group (those under 30 years old) systematically receives more wrong results compared to the unprivileged group (those over 30), while the FORD of -0.014 implies that the privileged group systematically receives more wrong "stressed" predictions compared to the unprivileged group (those over 30). The SPD of -0.086 shows that the unprivileged age group systematically receives fewer "not stressed" predictions, and the NPVD of 0.014 indicates that the unprivileged age group systematically receives fewer correct "stressed" predictions.

For the gender attribute, the ERD of -0.044 suggests that males receive more wrong results, while the FORD of 0.018 suggests that females receive more wrong "stress" predictions. The SPD of -0.071 and the NPVD of -0.018 indicate that females systematically receive fewer "not stressed" but more correct "stressed" predictions, respectively.

For the BMI attribute, the ERD of -0.017 suggests that the privileged group receives more wrong results, with the FORD of 0.024 indicating more wrong "stress" predictions for unprivileged individuals. The SPD of -0.043 and the NPVD of -0.024 show that the unprivileged BMI group

systematically receives fewer “not stressed” but more correct “stressed” predictions, respectively.

These results suggest that the generic model trained on the LifeSnaps dataset performs relatively well with some slight disparities.

The fairness evaluation metrics for the generic model (LR model) trained on the SWELL-KW dataset without protected attributes (Table 5.4) demonstrate a **less balanced performance** across demographic groups in comparison to the LifeSnaps dataset, but still the **disparities are not significant**.

SWELL-KW without protected attributes	Age	Gender	Occupation
Error Rate Difference (ERD)	-0.155497	-0.113362	-0.109225
False Omission Rate Difference (FORD)	-0.288138	-0.299172	-0.137787
Statistical Parity Difference (SPD)	0.277696	0.376754	0.031159
Negative Predicted Value Difference (NPVD)	0.288138	0.299172	0.137787

Table 5.4: Fairness evaluation metrics for the generic model trained on the SWELL-KW dataset without protected attributes

More specifically, for the age attribute, the ERD of -0.156 suggests that the privileged group (those under 30 years old) systematically receives more wrong results compared to the unprivileged group (those over 30). The FORD of -0.288 implies that the unprivileged group systematically receives more wrong “stressed” predictions compared to the unprivileged group. The SPD of 0.278 shows that the privileged age group systematically receives fewer “not stressed” predictions, and the NPVD of 0.288 indicates that the unprivileged age group systematically receives fewer correct “stressed” predictions.

For the gender attribute, the ERD of -0.113 suggests that males receive more wrong results and the FORD of -0.299 suggests that males also receive more wrong “stressed” predictions. The SPD of 0.377 indicates that the privileged gender group (males) systematically receives fewer “not stressed” predictions and the NPVD of 0.299 that the unprivileged gender group (females) systematically receives more correct “stressed” predictions, respectively.

For the occupation attribute, the ERD of -0.109 suggests that the privileged group systematically receives more wrong results, with the FORD of -0.138 indicating that this group also receives more wrong “stressed” predictions. The SPD of 0.031 and the NPVD of 0.138 show that the privileged occupation group systematically receives more “not stressed” predictions and more correct “stressed” predictions, respectively.

These results suggest that the generic model trained on the SWELL-KW dataset without protected attributes exhibits several biases, indicating that certain demographic groups are disproportionately affected by incorrect stress predictions.

The fairness evaluation metrics for the generic model (QDA model) trained on the SWELL-KW dataset with protected attributes (Table 5.5) indicate a **highly balanced performance** across demographic groups, with **minimal disparities** across demographic group.

SWELL-KW with protected attributes	Age	Gender	Occupation
Error Rate Difference (ERD)	-0.000549	-0.007252	0.00646
False Omission Rate Difference (FORD)	0.000000	0.000000	0.00000
Statistical Parity Difference (SPD)	0.000000	0.000000	0.00000
Negative Predicted Value Difference (NPVD)	0.000000	0.000000	0.00000

Table 5.5: Fairness evaluation metrics for the generic model trained on the SWELL-KW dataset with protected attributes

Specifically, for the age attribute, the ERD of -0.0005 suggests a very slight bias where the privileged group (those under 30 years old) receives more wrong results compared to the unprivileged group (those over 30). However, the FORD, SPD, and NPV) are all 0.000, indicating no measurable bias in terms of incorrect “stressed” predictions, the distribution of “not stressed” predictions, or the accuracy of “stressed” predictions between age groups.

For the gender attribute, the ERD of -0.007 indicates a slight bias where males receive more wrong results, but like the age attribute, the FORD, SPD, and NPVD values are all 0.000. This shows that there is no measurable bias in terms of incorrect “stressed” predictions, the distribution of “not stressed” predictions, or the accuracy of “stressed” predictions between gender groups.

For the occupation attribute, the ERD of 0.006 suggests a very minor bias where the unprivileged group receives slightly more wrong results. However, the FORD, SPD, and NPVD values are all 0.000, indicating no measurable bias in terms of incorrect “stressed” predictions, the distribution of “not stressed” predictions, or the accuracy of “stressed” predictions between occupation groups.

These results suggest that the generic model trained on the SWELL-KW dataset with protected attributes performs equitably across different demographic groups, exhibiting minimal to no bias in stress prediction outcomes. This highlights the importance of including protected attributes in the model training process to achieve fairer predictions.

In conclusion, the LGB model for the LifeSnaps dataset exhibits **slight disparities** across age, gender, and BMI, leading to potential unfair outcomes. The LR model trained on the SWELL-KW dataset without protected attributes also showed **significant biases** across demographic groups. However, when protected attributes were included in the SWELL-KW dataset, the QDA model demonstrated **minimal to no bias**, highlighting the effectiveness of considering protected attributes during training.

5.3.2 User-Based Splitting Bias Assessment

The fairness evaluation metrics for the User-Based splitting model trained on the SWELL-KW dataset without protected attributes (Table 5.6) indicate a **highly balanced performance** across demographic groups, with **minimal disparities** observed in error rates.

More specifically, the ERD values for age, gender, and occupation (0.064, 0.016, and 0.054, respectively) indicate minimal disparities in error rates. The negative FORD of -0.159 for age implies that the privileged age group (under 30) systematically receives more wrong "stressed" predictions. The SPD for age (0.147) and gender (-0.095) highlight discrepancies in the distribution of "not stressed" predictions between privileged and unprivileged groups, while the NPVD for age (0.159) indicates that the unprivileged age group receives fewer correct "stressed" predictions.

SWELL-KW	Age	Gender	Occupation
Error Rate Difference (ERD)	0.064336	0.015860	0.054119
False Omission Rate Difference (FORD)	-0.158730	0.048062	0.007735
Statistical Parity Difference (SPD)	0.147319	-0.094538	0.057318
Negative Predicted Value Difference (NPVD)	0.158730	-0.048062	-0.007735

Table 5.6: Fairness evaluation metrics for the user-based splitting model trained on the SWELL-KW dataset without protected attributes

For the model with protected attributes, the fairness metrics further reflect a **balanced performance** with **even fewer disparities**. The ERD values for age, gender, and occupation (0.049, 0.034, and -0.009, respectively) show minimal error rate differences. The positive FORD values for age, gender, and occupation (0.025, 0.045, and 0.025, respectively) indicate slightly more wrong "stressed" predictions for the privileged groups. The SPD for age (0.207) reveals a notable disparity in the distribution of "not stressed" predictions against the privileged group, while the NPVD values close to zero for all attributes indicate negligible differences in the correct "stressed" predictions between groups.

SWELL-KW	Age	Gender	Occupation
Error Rate Difference (ERD)	0.048951	0.033838	-0.008931
False Omission Rate Difference (FORD)	0.025253	0.045168	0.024527
Statistical Parity Difference (SPD)	0.206993	-0.088442	-0.010730
Negative Predicted Value Difference (NPVD)	-0.025253	-0.045168	-0.024527

Table 5.7: Fairness evaluation metrics for the user-based splitting model trained on the SWELL-KW dataset with protected attributes

Overall, the user-based splitting models trained on the SWELL-KW dataset demonstrate a generally balanced performance, with the inclusion of protected attributes slightly improving fairness across demographic groups.

5.3.3 Single-Attribute Splitting Bias Assessment

For the Single-Attribute Splitting model trained on the SWELL-KW dataset without protected attributes (Table 5.8), we see that the fairness evaluation metrics collectively indicate a **relatively balanced performance** across demographic groups with **minimum disparities**.

Specifically, the ERD values for age, gender, and BMI (0.160, -0.071, and 0.003, respectively) suggest slight variations in error rates among these attributes. The FORD values of -0.042 for age, -0.038 for gender, and -0.074 for BMI imply that the privileged groups in these categories receive more wrong "stressed" predictions. The SPD values (0.054 for age, 0.092 for gender, and 0.101 for BMI) show slight disparities in the distribution of "not stressed" predictions indicating that the privileged group receives fewer "not stressed" predictions, while the NPVD values (0.042 for age, 0.038 for gender, and 0.074 for BMI) indicate that the unprivileged groups receive fewer correct "stressed" predictions.

LifeSnaps	Age	Gender	BMI
Error Rate Difference (ERD)	0.160422	-0.071209	0.002979
False Omission Rate Difference (FORD)	-0.042148	-0.038495	-0.073856
Statistical Parity Difference (SPD)	0.053589	0.091709	0.100629
Negative Predicted Value Difference (NPVD)	0.042148	0.038495	0.073856

Table 5.8: Fairness evaluation metrics for the single attribute splitting model trained on the LifeSnaps

For the SWELL-KW dataset without protected attributes (Table 5.9), the metrics highlight **significant biases**, particularly for the age attribute. The ERD of -0.242 for age suggests that the privileged group (under 30) receives more wrong results. The FORD of -0.7 for age indicates a substantial disparity, with the privileged group receiving significantly more wrong "stressed"

predictions. The SPD of 0.021 for age further underscore the bias, revealing that the privileged age group systematically receives fewer "not stressed" predictions. The NPVD of 0.7 reveals that the unprivileged age group systematically receives fewer correct "stressed" predictions. In contrast, the metrics for gender and occupation exhibit minimal disparities, with ERD, FORD, SPD, and NPVD values close to zero, indicating a **more balanced performance.**

SWELL-KW without protected attributes	Age	Gender	Occupation
Error Rate Difference (ERD)	-0.241836	-0.058484	0.008288
False Omission Rate Difference (FORD)	-0.700000	-0.068848	0.039934
Statistical Parity Difference (SPD)	0.020814	-0.016993	-0.265376
Negative Predicted Value Difference (NPVD)	0.700000	0.068848	-0.039934

Table 5.9: Fairness evaluation metrics for the single attribute splitting model trained on the SWELL-KW dataset without protected attributes

For the SWELL-KW dataset with protected attributes (Table 5.10), a similar pattern is observed. The age attribute again shows **considerable bias**, with an ERD of -0.148 and a FORD of -0.520, indicating that the privileged group (udner 30) receives more wrong results and more wrong "stressed" predictions. The NPVD of 0.52 reveals that the unprivileged age group systematically receives fewer correct "stressed" predictions and the SPD of 0.021 for age suggests a slight disparity in the distribution of "not stressed" predictions. The metrics for gender and occupation reveal fewer disparities, with ERD, FORD, SPD, and NPVD values closer to zero, indicating a **balanced performance** across these attributes.

SWELL-KW with protected attributes	Age	Gender	Occupation
Error Rate Difference (ERD)	-0.148324	0.063463	-0.028413
False Omission Rate Difference (FORD)	-0.519737	0.061167	-0.011779
Statistical Parity Difference (SPD)	0.021303	0.077150	-0.175248
Negative Predicted Value Difference (NPVD)	0.519737	-0.061167	0.011779

Table 5.10: Fairness evaluation metrics for the single attribute splitting model trained on the SWELL-KW dataset with protected attributes

In summary, the single-attribute splitting model exhibits **minimal disparities** for the LifeSnaps dataset, while **significant biases** are observed for the age attribute in the SWELL-KW dataset, both with and without protected attributes. This suggests that the model's ability to predict stress accurately is affected by the age attribute, disproportionately impacting the privileged age group.

5.3.4 Multi-Attribute Splitting Bias Assessment

For the Multi-Attribute Splitting model trained on the LifeSnaps dataset (Table 5.11), the fairness evaluation metrics collectively indicate a **relatively balanced performance** across demographic groups with **minimum disparities**.

Specifically, the ERD values for age, gender, and BMI (-0.021, 0.045, and -0.197, respectively) suggest slight variations in error rates among these attributes. The FORD values of 0.007 for age, 0.028 for gender imply that the unprivileged groups in these categories receive more wrong "stressed" predictions, while the -0.014 for BMI implies that the privileged groups in these categories receive more wrong "stressed" predictions. The SPD values (-0.166 for age, -0.019 for gender, and 0.009 for BMI) show slight disparities in the distribution of "not stressed" predictions, while the NPVD values (-0.007 for age, -0.028 for gender, and 0.014 for BMI) indicate that the privileged groups receive fewer correct "stressed" predictions for the age and gender group and the unprivileged group receives fewer correct "stressed" predictions for the BMI group.

LifeSnaps	**Age**	**Gender**	**BMI**
Error Rate Difference (ERD)	-0.020544	0.044928	-0.196544
False Omission Rate Difference (FORD)	0.006762	0.028415	-0.013653
Statistical Parity Difference (SPD)	-0.166201	-0.019003	0.008756
Negative Predicted Value Difference (NPVD)	-0.006762	-0.028415	0.013653

Table 5.11: Fairness evaluation metrics for the multi attribute splitting model trained on the LifeSnaps dataset

For the Multi-Attribute Splitting model trained on the SWELL-KW dataset without protected attributes (Table 5.12), we observe **higher discrepancies** in comparison to the LifeSnaps dataset. Notably the "Age" group has the highest discrepancies. The "Occupation" group also has **high discrepancies**, while the "Gender" group has **the lowest discrepancies**.

More specifically, the ERD of 0.719 for age suggests that the unprivileged group (over 30) receives more wrong results, while the FORD of 0.595 indicates a substantial disparity, with the unprivileged group receiving significantly more wrong "stressed" predictions. The SPD of -0.237 for age further underscores the bias, revealing that the unprivileged age group systematically receives fewer "not stressed" predictions, while the NPVD of -0.595 reveals that the privileged age group systematically receives fewer correct "stressed" predictions. The metrics for occupation also show high discrepancies, with an ERD of 0.034, a FORD of 0.318, an SPD

of 0.562, and an NPVD of -0.318, indicating biases in the model's predictions for different occupational groups. The gender attribute exhibits the lowest discrepancies, with ERD, FORD, SPD, and NPVD values close to zero, suggesting a more balanced performance.

SWELL-KW without protected attributes	Age	Gender	Occupation
Error Rate Difference (ERD)	0.718921	0.005697	0.033792
False Omission Rate Difference (FORD)	0.595238	-0.014563	0.317622
Statistical Parity Difference (SPD)	-0.237175	-0.116691	0.561864
Negative Predicted Value Difference (NPVD)	-0.595238	0.014563	-0.317622

Table 5.12: Fairness evaluation metrics for the multi attribute splitting model trained on the SWELL-KW dataset without protected attributes

For the Multi-Attribute Splitting model trained on the SWELL-KW dataset with protected attributes (Table 5.13) there are **notable differences** in fairness evaluation metrics across demographic groups.

Specifically, the ERD of 0.434 for age suggests that the unprivileged group (over 30) receives more wrong results, while the negative ERD of -0.265 for gender indicates that the privileged group (males) receives more wrong results. The FORD of -0.257 for gender implies that males receive more wrong "stressed" predictions. The SPD of 0.424 for age and 0.399 for occupation suggests that the unprivileged groups in these categories are more likely to receive "not stressed" predictions, while the NPVD values of -1 for age and -0.395 for occupation indicate significant disparities, with the privileged groups receiving fewer correct "stressed" predictions.

SWELL-KW with protected attributes	Age	Gender	Occupation
Error Rate Difference (ERD)	0.434264	-0.264734	0.113599
False Omission Rate Difference (FORD)	0.000000	-0.256947	0.394910
Statistical Parity Difference (SPD)	0.424242	-0.156692	0.398623
Negative Predicted Value Difference (NPVD)	-1.000000	0.256947	-0.394910

Table 5.13: Fairness evaluation metrics for the multi-attribute splitting model trained on the SWELL-KW dataset with protected attributes

In summary, the multi-attribute splitting model trained on the LifeSnaps dataset exhibits **minimal disparities**, indicating a **relatively balanced performance**. However, for models trained on the SWELL-KW dataset, both with and without protected attributes, **significant biases** are observed, particularly for the age and occupation attributes, suggesting that these demographic groups are disproportionately affected by incorrect stress predictions.

5.3.5 Fuzzy Splitting bias assessment

For the Fuzzy Splitting model trained on the LifeSnaps (Table 5.14), the fairness evaluation metrics collectively indicate a **relatively balanced performance** across demographic groups with **minimal disparities**.

Specifically, the ERD values for age, gender, and BMI (0.039, 0.086, and -0.056, respectively) suggest slight variations in error rates among these attributes. The FORD values of -0.020 for age, and -0.036 for BMI imply that the privileged groups in these categories receive more wrong "stressed" predictions, while the 0.057 value for gender suggests that the unprivileged group receives more wrong "stressed" predictions. The SPD values (0.034 for age, 0.017 for gender, and -0.028 for BMI) show slight disparities in the distribution of "not stressed" predictions, while the NPVD values (0.020 for age, -0.057 for gender, and 0.036 for BMI) indicate slight disparities in the number of correct "stressed" predictions among demographic groups.

LifeSnaps	**Age**	**Gender**	**BMI**
Error Rate Difference (ERD)	0.039421	0.086097	-0.056072
False Omission Rate Difference (FORD)	-0.019609	0.057459	-0.035566
Statistical Parity Difference (SPD)	0.034156	0.016690	-0.027727
Negative Predicted Value Difference (NPVD)	0.019609	-0.057459	0.035566

Table 5.14: Fairness evaluation metrics for the fuzzy splitting model trained on the LifeSnaps dataset

For the fuzzy splitting model trained on the SWELL-KW protected attributes (Table 5.15), the fairness evaluation metrics collectively indicate a **relatively balanced performance** across demographic groups with **minimal disparities**.

Specifically, the ERD of 0.280 for age suggests that the unprivileged group (over 30) receives more wrong results, while the FORD of 0.159 indicates a disparity, with the unprivileged group receiving more wrong "stressed" predictions. The SPD of 0.130 for age and the NPVD of -0.159 underscore bias for the privileged age group, revealing that it systematically receives fewer "not stressed" predictions and fewer correct "stressed" predictions. The metrics for occupation and gender exhibit smaller discrepancies, with values closer to zero, indicating a more balanced performance for these attributes.

For the fuzzy splitting model trained on the SWELL-KW dataset with protected attributes (Table 5.16), the fairness evaluation metrics collectively indicate a **relatively balanced performance** across demographic groups with **minimal disparities**.

SWELL-KW without protected attributes	Age	Gender	Occupation
Error Rate Difference (ERD)	0.279528	0.040037	0.024031
False Omission Rate Difference (FORD)	0.159148	0.042099	0.073864
Statistical Parity Difference (SPD)	0.130249	-0.027545	0.064496
Negative Predicted Value Difference (NPVD)	-0.159148	-0.042099	-0.073864

Table 5.15: Fairness evaluation metrics for the fuzzy splitting model trained on the SWELL-KW dataset without protected attributes

Specifically, the ERD of 0.196 for age suggests that the unprivileged group (over 30) receives more wrong results, and the positive ERD of 0.024 for gender indicates that females receive more wrong results. The FORD of 0.119 for age and 0.047 for gender implies that the unprivileged groups receive more wrong "stressed" predictions. The SPD of 0.214 for age and 0.144 for gender suggests that the privileged groups are more likely to receive "not stressed" predictions. The NPVD values of -0.119 for age and -0.047 for gender indicate fewer correct "stressed" predictions for the privileged groups.

SWELL-KW with protected attributes	Age	Gender	Occupation
Error Rate Difference (ERD)	0.196194	0.024012	-0.034729
False Omission Rate Difference (FORD)	0.119048	0.047241	0.007476
Statistical Parity Difference (SPD)	0.213583	0.144122	-0.048682
Negative Predicted Value Difference (NPVD)	-0.119048	-0.047241	-0.007476

Table 5.16: Fairness evaluation metrics for the fuzzy splitting model trained on the SWELL-KW dataset with protected attributes

In summary, the fuzzy-splitting model trained on the LifeSnaps dataset exhibits **minimal disparities**, indicating a **relatively balanced performance**. For the SWELL-KW dataset, the model without protected attributes shows **more pronounced biases**, particularly for the age attribute, while the model with protected attributes exhibits a **more balanced performance** with **minimal disparities** across demographic groups.

Overall, we can say that fairness outcomes vary significantly depending on the dataset, splitting method, and whether protected attributes are included in the training data. In the next section, we will highlight the trade-off between accuracy and fairness for the ML models. This contribution addresses the limitation that current works regarding stress detection usually focus on accuracy without consideration for fairness, as stated in the Introduction Chapter (Section 1.2).

5.4 Trade-Off between Accuracy and Fairness

The quest for accurate and fair stress prediction models presents a complex challenge. While high accuracy ensures reliable stress detection, achieving fairness across different demographic groups is equally important. In this section we explore the interplay between these two crucial aspects, drawing insights from the comparative tables (Tables 5.1 and 5.2) that show the accuracy and the F1-score of the personalized ML models and the tables that show the results derived from the bias analysis of the ML models (Table 5.3 - 5.16).

For the generic models, the accuracy and fairness metrics show varying results across the datasets. For the LifeSnaps dataset, the generic model achieved an accuracy of 0.8741 and F1-score of 0.3836 using the LGB algorithm. Fairness evaluation metrics (Table 5.3) indicate slight disparities across demographic groups, but overall, the model maintains a relatively balanced performance. For the SWELL-KW with protected attributes, the generic model achieved an accuracy of 0.5295 and F1-score of 0.6924 using QDA, meaning lower accuracy but higher F1-score compared to the LifeSnaps dataset. The fairness metrics (Table 5.5) indicate a highly balanced performance. For the SWELL-KW without protected attributes, the generic model achieved an accuracy of 0.5373 and F1-score of 0.5753 using Logistic Regression (LR). The fairness metrics (Table 5.4) demonstrate less balanced performance with higher disparities across demographic groups compared to the model with protected attributes and the generic model trained to the LifeSnaps dataset.

Regarding the user-based splitting models, for the SWELL-KW dataset with protected attributes, the model achieved an accuracy of 0.8168 and F1-score of 0.8289, the highest performance for this dataset in terms of accuracy and F1-score. The fairness metrics (Table 5.6) indicate a balanced performance across demographic groups. For the SWELL-KW without protected attributes, the user-based splitting model achieved an accuracy of 0.8055 and F1-score of 0.8163, slightly lower than with protected attributes, but again the highest performance for this dataset in terms of accuracy and F1-score. The model still maintained minimal disparities in fairness metrics (Table 5.7).

Regarding the single-attribute splitting models, for the lifeSnaps dataset, the model achieved an accuracy of 0.956 and F1-score of 0.5667, indicating highest performance in terms of F1-score

for the this dataset. The fairness metrics (Table 5.8) also show good results with minimal disparities across demographic groups. For the SWELL-KW with protected attributes, the single-attribute splitting model achieved an accuracy of 0.7431 and F1-score of 0.7664, with fairness metrics (Table 5.10) indicating some significant disparities, particularly in the FORD and NPVD for Age. For the SWELL-KW without protected attributes, the single-attribute splitting model achieved an accuracy of 0.7188 and F1-score of 0.7496, slightly lower than with protected attributes, and the fairness metrics (Table 5.9) show again significant disparities in the FORD and NPVD for Age.

Regarding the multi-attribute splitting models, for the LifeSnaps dataset the model achieved an accuracy of 0.9617, the highest among all models and an F1-score of 0.55. Fairness metrics (Table 5.11) also indicate a good performance with minimal disparities across demographic groups. For the SWELL-KW with protected attributes, the multi-attribute splitting model achieved an accuracy of 0.7631 and F1-score of 0.7643, with fairness metrics (Table 5.13) indicating some notable differences across demographic groups, especially of Age and also Occupation. For the SWELL-KW without protected attributes, the model achieved an accuracy of 0.7541 and F1-score of 0.7571, with fairness metrics (Table 5.12) showing generally lower discrepancies across all demographic groups compared to the model trained with protected attributes.

Finally, for the fuzzy splitting models, for the LifeSnaps dataset, the model achieved an accuracy of 0.8186 and F1-score of 0.4156. Fairness metrics (Table 5.14) show a relatively balanced performance across demographic groups. For the SWELL-KW with protected attributes, the fuzzy-clustering splitting model achieved an accuracy of 0.76354 and F1-score of 0.7743, with fairness metrics (Table 5.16) indicating minimal disparities with higher disparities in the Age group. For the SWELL-KW without protected attributes, the model achieved an accuracy of 0.74058 and F1-score of 0.7512, and the fairness metrics (Table 5.15) show relatively balanced performance with some disparities similar to the ones for the SWELL-KW with protected attributes.

In conclusion, the LifeSnaps dataset exhibits generally higher accuracy in comparison to SWELL-KW, but SWELL-KW exhibits higher F1-scores. In terms of fairness metrics, the LifeSnaps dataset showcases in general a more balanced performance in comparison to the SWELL-KW dataset with and without protected attributes, which suggests that the SWELL-KW data might require more sophisticated fairness-aware training techniques. Finally, it is pleasantly surprising that the models that achieved the best performance in terms of accuracy and F1-score,

also achieved balanced fairness metrics. Nevertheless, there is no one-size-fits-all solution for achieving optimal accuracy and fairness. By understanding the trade-off and adopting appropriate strategies, we can strive for models that are reliable and equitable in identifying stress across diverse populations.

In conclusion, this chapter presents the results of our work in terms of model performance, data biases and model biases. Our experimentation results align with the contributions outlined in the Introduction chapter (Section 1.2). The first contribution involved a multifaceted analysis of data biases in two UbiComp datasets, which we achieved through detailed visualizations, revealing the hidden biases present in the training data. This way we address the limitation that ingested hidden biases in the training data are not extensively explored for stress detection applications. The second contribution focused on benchmarking the trade-off between accuracy and fairness in foundational and personalized ML models for stress detection addressing the gaps in current literature that bidirectional assessment of personalized models' accuracy and fairness is lacking for stress detection applications. The next and final chapter provides the conclusions of our research based on the contributions we stated in the Introduction chapter (Section 1.2), along with the limitations of our work and suggested future work.

Chapter 6

Conclusions & Future Work

In this thesis, we set out to explore the potential of personalized ML for stress detection using physiological measurements captured by wearables both in terms of accuracy and fairness. Through our research, we applied various personalized ML alogrithms to datasets collected both in the lab and in the wild, treating stress detection as a binary classification problem. Our work consisted of analyzing data biases in two UbiComp datasets, the LifeSnaps and the SWELL-KW dataset, through visualizations and benchmarking the trade-off between accuracy and fairness in foundational and personalized ML models for stress detection.

The key findings of our research are listed below aligning with our two research contributions mentioned in the Introduction chapter (Section 1.2).

- **Biases in Datasets.** Our analysis revealed that the datasets used in this study exhibited biases. Specifically, the LifeSnaps dataset exhibited sampling and representation bias across the protected attributes "gender" and "age" (Subsections 5.2.1 and 5.2.2). It also exhibited measurement biases, since there were missing values for the protected attributes and the label distribution across protected attribute groups was not balanced (Subsections 5.2.3 and 5.2.4). The SWELL-KW datasets appeared to have sampling and representation biases across the protected attribute "gender" and these biases were even more pronounced for the protected attribute "age". The results of the data bias analysis are presented in detail in the Experimentation & Results chapter, in Section 5.2. Addressing such biases is crucial for developing fair and equitable models.

- **Accuracy and Fairness in Stress Detection Models.** Our trade-off analysis between accuracy and fairness for the personalized ML models for stress detection revealed that the models can be both accurate and fair, though achieving this balance requires careful consideration. The personalized ML models demonstrated high F1-score while minimizing biases. For example, the user-based splitting model trained on the SWELL-KW dataset including protected attributes achieved an F1-score of 82.89% and maintained fairness metrics within acceptable thresholds (Subsection 5.3.2). This indicates that it is possible to develop stress detection systems that do not compromise on fairness for accuracy. We present the trade-off analysis in detail in the Experimentation & Results in Section 5.4.

It is also important to address the limitations of our research:

- **Limited Use of Fairness Assessment Tools.** In this study, we applied only one library, the aif360 python library and four specific metrics to assess the fairness of the ML models. While this approach provides valuable insights into biases included in the models, utilizing and comparing multiple fairness libraries could have offered a more comprehensive quantification of biases. Future research should consider applying various fairness assessment tools to better understand and mitigate biases in stress detection models.

- **Incomplete Fairness Assessment.** Our assessment of fairness is focused on quantifying the biases of the input data and output results of the models, which may overlook potential biases present during other phases of the model lifecycle, such as during problem formulation or validation and testing of the ML models. Assessing fairness throughout the entire ML lifecycle is crucial to identify and address additional biases that may arise at different stages. Future work should aim to evaluate and enhance fairness across all phases of model development.

To address bias in AI systems comprehensively, future work must focus on mitigating biases across the entire AI lifecycle. This involves not only ensuring diverse and representative datasets during data collection but also implementing fairness metrics to continuously assess and improve the equity of stress detection systems. Standardized protocols for data collection can significantly enhance the consistency and quality of datasets, facilitating the development

and comparison of stress detection models across different studies and populations. Additionally, adopting frameworks like the AI Fairness 360 toolkit, which provides comprehensive metrics for datasets and models to identify and mitigate biases, is crucial. Engaging diverse teams throughout the AI development lifecycle, including problem framing, model development, validation, and deployment, ensures that models are designed with fairness at their core.

Bias mitigation in AI systems is essential for creating accurate and fair models that can be trusted in real-world applications. Accurate and fair stress detection models can be instrumental in various use cases, such as personalized treatment plans for individuals with high-stress levels, improving workplace wellness programs, and enhancing public health interventions. These models must be transparent and interpretable, ensuring that stakeholders can trust their outputs. By focusing on fairness, AI systems can better serve all segments of the population, reducing disparities and promoting equity in technology applications.

Building on our findings, several potential scenarios and use cases illustrate the practical benefits of personalized stress detection models. In healthcare, these models can tailor stress management plans for patients with chronic stress, addressing data biases by ensuring diverse and representative training sets. Workplace wellness programs can leverage wearables to monitor employee stress, implementing personalized support and ensuring fairness across demographics. Public health agencies can utilize aggregated stress data to design targeted interventions in high-stress communities, benefiting from standardized data collection protocols. Educational institutions can support student mental health by providing timely, personalized interventions, ensuring inclusivity and effectiveness for all students. These scenarios demonstrate how our research contributions, which are analyzing data biases and exploring the balance between accuracy and fairness of personalized ML models, can lead to robust, equitable stress detection models that enhance individual and public health outcomes.

In conclusion, this thesis highlights the potential of personalized AI for stress detection using wearable technology. By addressing limitations in existing work and exploring future directions, we can pave the way for robust, fair, and user-centric stress detection systems. This has the potential to significantly improve individual well-being and contribute to advancements in personalized and equitable healthcare.

Bibliography

[1] Donald B. Ardell. 2018. The Ardell Wellness Stress Test Self-Assessment. https://premierespeakers.com/donaldardell/blog/2018/07/01/the-ardell-wellness-stress-self-assessment.

[2] Nasrin Attaran, Abhilash Puranik, Justin Brooks, and Tinoosh Mohsenin. 2018. Embedded low-power processor for personalized stress detection. *IEEE Transactions on Circuits and Systems II: Express Briefs* 65, 12 (2018), 2032–2036.

[3] Himani Bhavsar and Mahesh H Panchal. 2012. A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 1, 10 (2012), 185–189.

[4] Sara Campanella, Ayham Altaleb, Alberto Belli, Paola Pierleoni, and Lorenzo Palma. 2023. A method for stress detection using empatica E4 bracelet and machine-learning techniques. *Sensors* 23, 7 (2023), 3565.

[5] R. Castaldo, W. Xu, P. Melillo, L. Pecchia, L. Santamaria, and C. James. 2016. Detection of mental stress due to oral academic examination via ultra-short-term HRV analysis. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 3805–3808. https://doi.org/10.1109/EMBC.2016.7591557

[6] Wei Chen, Shixin Zheng, and Xiao Sun. 2021. Introducing mdpsd, a multimodal dataset for psychological stress detection. In *Big Data: 8th CCF Conference, BigData 2020, Chongqing, China, October 22–24, 2020, Revised Selected Papers 8*. Springer, 59–82.

[7] Sheldon Cohen, Tom Kamarck, Robin Mermelstein, et al. 1994. Perceived stress scale. *Measuring stress: A guide for health and social scientists* 10, 2 (1994), 1–2.

[8] AI Fairness 360 Contributors. 2024. Fairness Metrics — aif360 0.6.1 documentation. https://aif360.readthedocs.io/en/latest/modules/metrics.html Accessed: 2024-05-30.

[9] Michael Friedewald and Oliver Raabe. 2011. Ubiquitous computing: An overview of technology impacts. *Telematics and Informatics* 28, 2 (2011), 55–65.

[10] Shruti Gedam and Sanchita Paul. 2021. A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques. *IEEE Access* 9 (2021), 84045–84066. https://doi.org/10.1109/ACCESS.2021.3085502

[11] Soumi Ghosh and Sanjay Kumar Dubey. 2013. Comparative analysis of k-means and fuzzy c-means algorithms. *International Journal of Advanced Computer Science and Applications* 4, 4 (2013).

[12] Neska El Haouij, Jean-Michel Poggi, Sylvie Sevestre-Ghalila, Raja Ghozi, and Mériem Jaïdane. 2018. AffectiveROAD system and database to assess driver's attention. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. 800–803.

[13] J.A. Healey and R.W. Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems* 6, 2 (2005), 156–166. https://doi.org/10.1109/TITS.2005.848368

[14] Seyedmajid Hosseini, Raju Gottumukkala, Satya Katragadda, Ravi Teja Bhupatiraju, Ziad Ashkar, Christoph W Borst, and Kenneth Cochran. 2022. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Scientific Data* 9, 1 (2022), 255.

[15] Christina Karagianni, Eva Paraschou, Sofia Yfantidou, and Athena Vakali. 2023. MINDSET: A benchMarking suIte exploring seNsing Data for SElf sTates inference. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 1–10.

[16] Christina Karagianni, Eva Paraschou, Sofia Yfantidou, and Athena Vakali. 2023. MINDSET: A benchMarking suIte exploring seNsing Data for SElf sTates inference. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*. 1–10. https://doi.org/10.1109/DSAA60987.2023.10302638

[17] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. 1993. The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 1-2 (1993), 76–81.

[18] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A Neerincx, and Wessel Kraaij. 2014. The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction*. 291–298.

[19] Michael P LaValley. 2008. Logistic regression. *Circulation* 117, 18 (2008), 2395–2399.

[20] Susan Levenstein, Cosimo Prantera, Vilma Varvo, Maria L Scribano, Eva Berto, Carlo Luzi, and Arnaldo Andreoli. 1993. Development of the Perceived Stress Questionnaire: a new tool for psychosomatic research. *Journal of psychosomatic research* 37, 1 (1993), 19–32.

[21] Tambiama Madiega. 2021. Artificial intelligence act. *European Parliament: European Parliamentary Research Service* (2021).

[22] Agnese Mariotti. 2015. The effects of chronic stress on health: new insights into the molecular mechanisms of brain–body communication. *Future science OA* 1, 3 (2015).

[23] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[24] Vasiliki Parousidou, Sofia Yfantidou, Christina Karagianni, and Athena Vakali. 2023. Stress Beats: a continuum of learning methods for personalized stress detection. In *2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 40–47.

[25] Vasiliki-Chrysovalanto Parousidou. 2023. *Personalized Machine Learning Benchmarking for Stress Detection*. Ph. D. Dissertation. ARISTOTLE UNIVERSITY OF THESSALONIKI.

[26] Aaqib Saeed and Stojan Trajanovski. 2017. Personalized driver stress detection with multi-task neural networks using physiological signals. *arXiv preprint arXiv:1711.06116* (2017).

[27] Akane Sano, Andrew J. Phillips, Amy Z. Yu, Andrew W. McHill, Sara Taylor, Natasha Jaques, Charles A. Czeisler, Elizabeth B. Klerman, and Rosalind W. Picard. 2015. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. 1–6. https://doi.org/10.1109/BSN.2015.7299420

[28] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 400–408.

[29] Rutvik V Shah, Gillian Grennan, Mariam Zafar-Khan, Fahad Alim, Sujit Dey, Dhakshin Ramanathan, and Jyoti Mishra. 2021. Personalized machine learning of depressed mood using wearables. *Translational psychiatry* 11, 1 (2021), 1–18.

[30] Yuan Shi, Minh Hoai Nguyen, Patrick Blitz, Brian French, Scott Fisk, Fernando De la Torre, Asim Smailagic, Daniel P Siewiorek, Mustafa Al'Absi, Emre Ertin, et al. 2010. Personalized stress detection from physiological measurements. In *International symposium on quality of life technology*. 28–29.

[31] Yan-Yan Song and LU Ying. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry* 27, 2 (2015), 130.

[32] Ramya Srinivasan and Ajay Chander. 2021. Biases in AI Systems: A survey for practitioners. *Queue* 19, 2 (may 2021), 45–64. https://doi.org/10.1145/3466132.3466134

[33] Sharmin Sultana, Md Mahmudur Rahman, Atqiya Munawara Mahi, Shao-Hsien Liu, and Mohammad Arif Ul Alam. 2023. Unbiased Pain Assessment through Wearables and EHR Data: Multi-attribute Fairness Loss-based CNN Approach. *arXiv preprint arXiv:2307.05333* (2023).

[34] Ali Tazarv, Sina Labbaf, Stephanie M Reich, Nikil Dutt, Amir M Rahmani, and Marco Levorato. 2021. Personalized stress monitoring using wearable sensors in everyday settings. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 7332–7335.

[35] Jaakko Tervonen, Sampsa Puttonen, Mikko J Sillanpää, Leila Hopsu, Zsolt Homorodi, Janne Keränen, Janne Pajukanta, Antti Tolonen, Arttu Lämsä, and Jani Mäntyjärvi. 2020. Personalized mental stress detection with self-organizing map: From laboratory to the field. *Computers in Biology and Medicine* 124 (2020), 103935.

[36] World Health Organization. 2023. Stress. https://www.who.int/news-room/questions-and-answers/item/stress Accessed: 2024-05-29.

[37] Jyoti Yadav and Monika Sharma. 2013. A Review of K-mean Algorithm. *Int. J. Eng. Trends Technol* 4, 7 (2013), 2972–2976.

[38] Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. 2022. LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data* 9, 1 (2022), 663.

[39] Sofia Yfantidou, Pavlos Sermpezis, Athena Vakali, and Ricardo Baeza-Yates. 2023. Uncovering Bias in Personal Informatics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–30.

[40] Sofia Yfantidou, Pavlos Sermpezis, Athena Vakali, and Ricardo Baeza-Yates. 2023. Uncovering Bias in Personal Informatics. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 139 (sep 2023), 30 pages. https://doi.org/10.1145/3610914