DOCTORAL DISSERTATION

École Doctorale Informatique, Télécommunications et Électronique de Paris (ED130)

AI for e-Health: Robust Artificial Intelligence Models for the Analysis of Wearable Medical Devices Data

Disseration submitted to

Sorbonne Université

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Hava Chaptoukaev

EURECOM, Data Science Department

Defense planned on March 31st, 2025

Before a committee composed of:

Dr Julie Josse, Inria	President
Dr Maria Vakalopoulou, CentraleSupélec	Examiner
Prof. Gustavo Carneiro, University of Surrey	Reviewer
Prof. Marc Niethammer, University of California, San Diego	Reviewer
Dr Maria A. Zuluaga, EURECOM	Thesis advisor

SORBONNE UNIVERSITÉ

EURECOM
Sophia Antipolis

L'IA pour la e-Santé: Conception d'Outils Robustes d'Intelligence Artificielle pour l'Analyse de Données de Dispositifs Médicaux Portables

Thèse présentée pour l'obtention du grade de
Docteur de Sorbonne Université

Hava Chaptoukaev
EURECOM, Département Data Science

Soutenance prévue le 31 Mars 2025
Devant un jury composé de:

Dr Julie Josse, Inria	Presidente
Dr Maria Vakalopoulou, CentraleSupélec	Examinatrice
Prof. Gustavo Carneiro, University of Surrey	Rapporteur
Prof. Marc Niethammer, University of California, San Diego	Rapporteur
Dr Maria A. Zuluaga, EURECOM	Directrice de thèse

SORBONNE UNIVERSITÉ

EURECOM
Sophia Antipolis

Acknowledgments

I would first like to profusely thank my supervisor Dr Maria A. Zuluaga for making my PhD experience such a great one: not only was the road so fun, but I have also learned from her how to be a researcher along the way! I feel lucky to have been supervised by someone I got along with so well, and that I can truly consider as a role model for the rest of my career.

Thank you to all members of my PhD jury – Prof. Gustavo Carneiro, Prof. Marc Niethammer, Dr Maria Vakalopoulou and Dr Julie Josse. I am truly honored and grateful they reviewed my work, providing valuable insights and many interesting discussions.

Thank you to my team: to Francesco who was always ready to lend a hand; to Lisa who always made any conference we went to so much better; to Vincenzo for the countless stimulating conversations we had (and the laughs too! research CAN be fun); to Bianca, Daniele, Luisa, Valentin, Mobin, Eleonora and everyone else for making my time at EURECOM so great.

Thank you to all my friends, Tamila, Madina, Medina, Omar and Gheed for being with me during this journey – especially for your presence and your amazing support through the most stressful day of my life (so far)!

Thank you to my parents for giving me the opportunity to pursue a PhD – I am so grateful they have granted me and my brothers the privilege to make it this far in our lives.

Lastly, thank you to my husband Josquin for EVERYTHING – support, brainstorming, reviewing, editing... I'm happy to share this success with him: he made these years so easy! 12/10.

Abstract

Artificial Intelligence (AI) based analysis of multimodal data collected using inexpensive and accessible wearable sensors is emerging as a promising opportunity to democratise access to healthcare. It would facilitate the prevention of various health problems and reduce the need for expensive clinical examinations that are difficult to access for a large portion of the population. However, concerns remain about the reliability and robustness of AI algorithms that are frequently overlooked in healthcare research. Yet, these aspects are crucial to the deployment of AI in medical applications. On one hand, most existing algorithms are trained on data that is hardly representative of the real world, and on the other, their architectures make them vulnerable to various perturbations commonly encountered in real data, once deployed. The aim of this PhD project is to develop innovative, robust and reliable methodologies for the analysis of wearable sensor data – with a particular focus on robustness to missing data. Our aim is to design novel multimodal methodologies that are evaluated on real sensor data and designed to be transposed and adapted to various medical applications – ranging from physiological signals analysis to the analysis of multimodal imaging data. We propose 5 contributions to achieve this goal. (1) We introduce `StressID`, a new dataset specifically designed for stress identification from unimodal and multimodal data, that we made publicly available for researchers. It contains videos, audio recordings, and physiological signals collected in ambulatory settings using wearable sensors. As it is collected from 65 participants, it includes a wide range of participant's responses. As such, it is a valuable support for building reliable and robust applications for stress identification. (2) We propose an open-source suite of baseline models for the analysis of `StressID`, that is representative of the current state-of-the-art in the domain, and facilitates future contributions to this domain by providing a starting point for researchers who wish to

use the dataset. We investigate the next steps needed to ensure reliability and robustness of existing models, and identify robustness to missing data as an essential aspect to exploiting the benefits of real-life multimodal datasets. (3) We explore whether the rich existing literature on missing values in tabular data can be leveraged to address this limitation. We conduct a comprehensive evaluation of existing methods for dealing with missing data, and assess their reliability within healthcare applications. This enables us to identify the strengths and limitations of existing approaches, to ultimately derive a set of guidelines to properly and responsibly handle missing values in healthcare applications. (4) Based on the considerations thus identified, we propose PicMi, an end-to-end imputation-free model designed for supervised learning with missing values in tabular data, that uses a permutation-invariant architecture to handle inputs of varying dimensions; integrates missing value patterns as a condition in its objective function to ensure robustness to various missing values scenarios; and is locally interpretable. (5) We extend our approach to multimodal learning with missing modalities, and introduce HyperMM, a framework designed for handling missing modalities without using reconstruction before training – as opposed to existing solutions. We introduce a novel strategy for training a universal feature extractor using a conditional hypernetwork, and propose a permutation invariant neural network that can handle inputs of varying dimensions to process the extracted features, in a two-phase task-agnostic framework. Our method is end-to-end and can be used in various applications, and thus contributes to the development of more reliable and robust AI systems in healthcare.

Keywords: Healthcare data, Multimodal data, Missing data, Supervised learning, Stress Identification.

Résumé

L'analyse basée sur l'intelligence artificielle (IA) des données multimodales collectées à l'aide de capteurs portables, peu coûteux et accessibles, apparaît comme une opportunité prometteuse pour démocratiser l'accès aux soins de santé. Elle faciliterait la prévention de divers problèmes de santé et réduirait la nécessité d'examens cliniques coûteux, difficilement accessibles à une grande partie de la population. Toutefois, la fiabilité et la robustesse des algorithmes d'IA, souvent négligées dans la recherche sur les soins de santé, suscitent encore des inquiétudes. D'une part, les algorithmes existants sont entraînés sur des données peu représentatives du monde réel, et d'autre part, leurs architectures les rendent vulnérables aux diverses perturbations couramment rencontrées une fois déployés. L'objectif de ce doctorat est de développer des méthodes innovantes, robustes et fiables pour l'analyse des données de capteurs portables – avec un accent particulier sur la robustesse aux données manquantes. Nous voulons proposer des méthodes multimodales évaluées sur données réelles et conçues pour être adaptées à diverses applications médicales – allant de l'analyse des signaux physiologiques à l'analyse d'imagerie multimodale. Nous proposons 5 contributions pour atteindre cet objectif. (1) Nous présentons `StressID`, un jeu de données conçu pour l'identification du stress à partir de données unimodales et multimodales, que nous avons rendu public. Il contient des vidéos, des enregistrements audio et des signaux physiologiques collectés à l'aide de capteurs portables. Il comprend les données de 65 participants, et donc un large éventail de réponses. Il s'agit d'un support précieux pour la création d'applications fiables et robustes pour l'identification du stress. (2) Nous proposons une suite de modèles pour l'analyse de `StressID`, qui est représentative de l'état de l'art actuel dans le domaine, et a pour but de faciliter les contributions futures en fournissant un point de départ pour les chercheurs. Nous analysons les limitations des modèles existants, et identifions la

robustesse aux données manquantes comme un aspect essentiel pour exploiter les avantages des données multimodales. (3) Nous examinons si la riche littérature existante sur les valeurs manquantes peut être exploitée pour remédier à ce problème. Nous étudions les méthodes existantes pour traiter les données manquantes, et évaluons leur fiabilité dans les applications de soins de santé. Cela nous permet d'identifier les forces et les limites des approches existantes, pour finalement dériver un ensemble de recommandations pour traiter correctement et de manière responsable les valeurs manquantes dans les applications de santé. (4) Sur la base des considérations ainsi identifiées, nous proposons PicMi, un modèle sans imputation pour l'apprentissage supervisé avec des valeurs manquantes dans des données tabulaires. Il utilise une architecture invariante par permutation pour traiter des entrées de dimensions variables ; intègre les motifs de valeurs manquantes comme condition dans son apprentissage pour assurer la robustesse à divers scénarios de valeurs manquantes ; et est localement interprétable. (5) Nous étendons notre approche à l'apprentissage multimodal avec des modalités manquantes et présentons HyperMM, une méthode conçue pour traiter les modalités manquantes sans utiliser de reconstruction – contrairement aux solutions existantes. Nous proposons une stratégie pour l'entraînement d'un extracteur de caractéristiques universel en utilisant un *hypernetwork* conditionnel, et proposons un réseau neuronal invariant par permutation qui peut gérer des entrées de dimensions variables pour traiter les caractéristiques extraites. Notre méthode est intégrée de bout-en-bout et peut être utilisée dans diverses applications, contribuant ainsi au développement de systèmes d'intelligence artificielle plus fiables et plus robustes dans le domaine des soins de santé.

Mots-clés: Données de santé, Données multimodales, Données manquantes, Apprentissage supervisé, Identification du stress.

Publications and Awards

Conference Papers

H. Chaptoukaev, V. Strizhkova, M. Panariello, B. Dalpaos, A. Reka, V. Manera, S. Thümmler, E. Ismailova, N. Evans, F. Bremond, M. Todisco, M. A. Zuluaga & L. M. Ferrari. (2023) "StressID: a multimodal dataset for stress identification." In *Advances in Neural Information Processing Systems (NeurIPS)*. Vol 36, p. 29798-29811.

H. Chaptoukaev, V. Marcianó, F. Galati, & M. A. Zuluaga (2024). "HyperMM: Robust Multimodal Learning with Varying-sized Inputs." In *5th International Workshop on Multiscale Multimodal Medical Imaging (MMMI), In conjunction with Medical Image Computing and Computer Assisted Intervention (MICCAI)*. In press.

V. Strizhkova, H. Kachmar, **H. Chaptoukaev**, R. Kalandadze, N. Kukhilava, T. Tsmindashvili, N. Abo-Alzahab, M. A. Zuluaga, M. Balazia, A. Dantcheva, F Bremond & L. M. Ferrari. (2024) "MVP: Multimodal Emotion Recognition based on Video and Physiological Signals." In *7th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW), In conjuction with European Conference on Computer Vision (ECCV)*. In press.

H. Chaptoukaev, V. Marcianó & M. A. Zuluaga. (2025) "PicMi: Imputation-free Supervised Learning in the Presence of Missing Values in Healthcare Data." In preparation.

Journal Papers

F. Lareyre, **H. Chaptoukaev**, S. C. Kiang, A. Chaudhuri, C. A. Behrendt, M. A. Zuluaga, & J. Raffort. (2022) "Telemedicine and digital health applications in vascular surgery". *In Journal of Clinical Medicine*. Vol. 11, no 20, p. 6047.

F. Capitano, M. Kuchenbuch, J. Lavigne, **H. Chaptoukaev**, M. A. Zuluaga, M. Lorenzi, R. Nabbout & M. Mantegazza. (2024) "Preictal dysfunctions of inhibitory interneurons paradoxically lead to their rebound hyperactivity and to low-voltage-fast onset seizures in Dravet syndrome". In *Proceedings of the National Academy of Sciences*. Vol. 121, no 23, p. e2316364121.

H. Chaptoukaev, M. Antonelli, & M. A. Zuluaga. (2025) "How to handle missing values in healthcare data?" In preparation.

Conference Abstracts

H. Chaptoukaev, M. Beurey, J. Raffort & M. A. Zuluaga. (2022) "Assessing Multiple Imputation of Missing Values for Robust Analysis of Telehealth Kiosk Data." In *IA et santé : approches interdisciplinaires, Centre de mathématiques Henri Lebesgue.*

H. Chaptoukaev, M. A. Zuluaga. (2024) "Approche Pour l'Apprentissage Multimodal Supervisé Sans Reconstruction de Modalités Manquantes". In *IABM 2024: Colloque Français d'Intelligence Artificielle en Imagerie Biomédicale.*

Awards

A *NeurIPS Scholar Award* was awarded at the NeurIPS 2023 conference for the paper: "StressID: a multimodal dataset for stress identification."

A *Best Presentation Award* was awarded at the MMMI workshop of the conference MICCAI 2024 for the paper: "HyperMM: Robust Multimodal Learning with Varying-sized Inputs."

Contents

Acknowledgments **i**

Abstract **ii**

Résumé **iv**

Publications and Awards **vi**

1 Introduction **1**
1.1 Overview 1
1.2 Medical Context 2
1.2.1 Emergence of Wearable Medical Devices and e-Health 3
1.2.2 Challenges of e-Health 6
1.3 Objectives and Contributions 8
1.4 Thesis Organization 10

2 StressID : A Multimodal Dataset for Stress Identification **12**
2.1 Introduction 13
2.2 Related Work 15
2.3 Design of the StressID Dataset 17
2.3.1 Experimental Protocol 18
2.3.2 Sensors 21
2.3.3 Recruitment and Recording 22
2.4 Dataset Description 23
2.4.1 Contents and Formats 23
2.4.2 Data Annotations 25
2.5 Intended Uses of StressID 29
2.6 Discussion 30

3 Stress Identification from Physiological Signals, Videos and Audio Data **32**

3.1 Introduction . . . 33
3.2 State-of-the-Art . . . 33
3.3 Baseline Models for Stress Identification . . . 35
3.3.1 Unimodal Models . . . 36
3.3.2 Multimodal baselines . . . 41
3.4 Main Limitations . . . 43
3.4.1 Missing Data . . . 44
3.4.2 Gender Imbalance . . . 44
3.5 Discussion . . . 45

4 How to Handle Missing Values in Healthcare Data? 47
4.1 Introduction . . . 48
4.2 Problem Formulation, Notations and Definitions . . . 49
4.3 State-of-the-Art . . . 51
4.4 Assessing the Reliability of Existing Approaches within Healthcare Applications 54
4.4.1 Overview of the Methodology . . . 55
4.4.2 Datasets Generation and Fingerprints Extraction . . . 56
4.4.3 Benchmark and Evaluation Criteria . . . 60
4.4.4 Analysis of the Results and Model Selection . . . 65
4.4.5 Decision tree-based Approach for Model Choice . . . 72
4.5 Guidelines for Handling Missing Values in Health Data . . . 74
4.5.1 Main Takeaways: a Practical Guide with Flowcharts . . . 74
4.5.2 Illustration on Healthcare Datasets . . . 77
4.6 Application to StressID . . . 79
4.7 Discussion . . . 80

5 PicMi: Imputation-free Supervised Learning in the Presence of Missing Values in Health Data 83
5.1 Introduction . . . 84
5.2 Related Work . . . 86
5.3 Method . . . 87
5.3.1 Permutation-invariant Architecture . . . 87
5.3.2 Conditioning Module . . . 89
5.3.3 Attention Module . . . 90
5.3.4 Theoretical Guarantees and Optimization . . . 91
5.4 Experiments and Results . . . 92
5.4.1 Comparison with State-of-the-Art Methods . . . 92
5.4.2 Robustness to Complex Scenarios . . . 97
5.4.3 Ablation Study and Implementation Details . . . 98
5.5 Application to StressID . . . 99
5.6 Discussion . . . 101

6 HyperMM : Robust Multimodal Learning with Varying-sized Inputs 103
6.1 Introduction . . . 104
6.2 Related Work . . . 105
6.3 Method . . . 106
6.3.1 Overview of the Method . . . 106
6.3.2 Universal Feature Extractor . . . 106
6.3.3 Permutation Invariant Multimodal Classifier . . . 108
6.4 Experiments and Results . . . 109
6.4.1 Alzheimer's Disease Detection . . . 109
6.4.2 Breast Cancer Classification . . . 113
6.5 Discussion . . . 115

7 Conclusion and Perspectives 117
7.1 Conclusion . . . 117
7.2 Perspectives . . . 122

Appendices 125

A StressID: A Multimodal Dataset for Stress Identification 125
A.1 Experimental Protocol . . . 125
A.2 Calibration and Synchronization of the Sensors . . . 126
A.3 Human Subject Considerations . . . 127
A.4 Ethical Considerations . . . 128
A.5 Dataset Accessibility . . . 129

B Stress Identification from Physiological Signals, Videos and Audio Data 130
B.1 Additional Experiments: Emotion Recognition . . . 130

C How to Handle Missing Values in Healthcare Data? 132
C.1 Decision Trees . . . 132

Bibliography 136

Chapter 1

Introduction

Contents

1.1 Overview 1
1.2 Medical Context 2
1.2.1 Emergence of Wearable Medical Devices and e-Health 3
1.2.2 Challenges of e-Health 6
1.3 Objectives and Contributions 8
1.4 Thesis Organization 10

1.1 Overview

Currently, 4.5 billion people worldwide have no access to essential health services [1]. In France alone, nearly 8 million people live in medical deserts [2]. Access to healthcare for everyone, everywhere is, therefore, more than ever at the heart of our concerns, and the search for innovative and competitive solutions on the global market is actively promoted by the French [3] and international governments. The unprecedented rise in access to wearable medical sensors and devices, our growing ability to collect, store, and process an abundance of data, and advances in Artificial Intelligence (AI) have all fostered the emergence of digital health solutions, or *e-health*, in recent years. In this context, AI analysis of multi-source, or multimodal, data collected using inexpensive and accessible wearable sensors is emerging as a

[1] WHO, Universal health coverage factsheet, (2023)
[2] Rapport de la commission de l'aménagement du territoire et du développement durable, (2020)
[3] Stratégie d'accélération Santé Numérique, (2021)

promising opportunity to democratize access to healthcare. It would facilitate the prevention of various health problems, and reduce the need for expensive clinical examinations that are difficult to access for a large proportion of the population. However, concerns remain about the reliability and robustness of AI algorithms. Although crucial to the use of AI in medical applications, these aspects are frequently overlooked. On one hand, most existing algorithms are trained on data that is hardly representative of the real world, and on the other, their architectures make them vulnerable to various perturbations once deployed. In particular, traditional algorithms are not equipped to handle missing values – although they are prone to occur frequently in increasingly complex real-world datasets. These limitations underline the need of innovative, robust and reliable methodologies to foster the potential benefits of AI in e-health applications.

In this thesis, I design a methodology to ensure the development of robust and reliable AI models for the analysis of wearable sensor data. This introduction provides the necessary context. To start, the opportunities offered by the emergence of wearable medical devices and the challenges associated with it are identified it in Section 1.2. In particular, the terms of reliability and robustness are clearly defined, which enables their association with several current limitations in e-health research. My thesis contributions to address these issues, through the design of a real-life dataset collected with wearable sensors and innovative methodologies to handle missing data, are described in Section 1.3. Finally, the organization of this thesis is outlined in Section 1.4.

1.2 Medical Context

The World Health Organization (WHO) defines *e-health* as the cost-effective and secure use of information and communication technologies in support of health and health-related fields, including healthcare services, health surveillance, health literature, and health education, knowledge and research. In other terms, e-health can be defined as the use of new technologies to support and improve healthcare efficiency, accessibility, quality, and management – while empowering patients to take an active role in managing their health. By integrating technology into healthcare systems, e-health also aims to reduce costs and bridge gaps in access to care, especially for under-served populations or those in remote areas. Multiple studies have shown the positive impact of new solutions like telemedicine (Armaignac et al., 2018; Lapointe et al., 2020), e-health mobile applications (Arsad et al., 2023), remote monitoring of patients via wearable devices (Huhn et al., 2022), or data-driven AI systems (Haleem et al., 2019; Ting et al., 2018) on the delivery of healthcare around the world. This thesis places particular emphasis on the possibilities offered by the increasing availability of wearable medical devices.

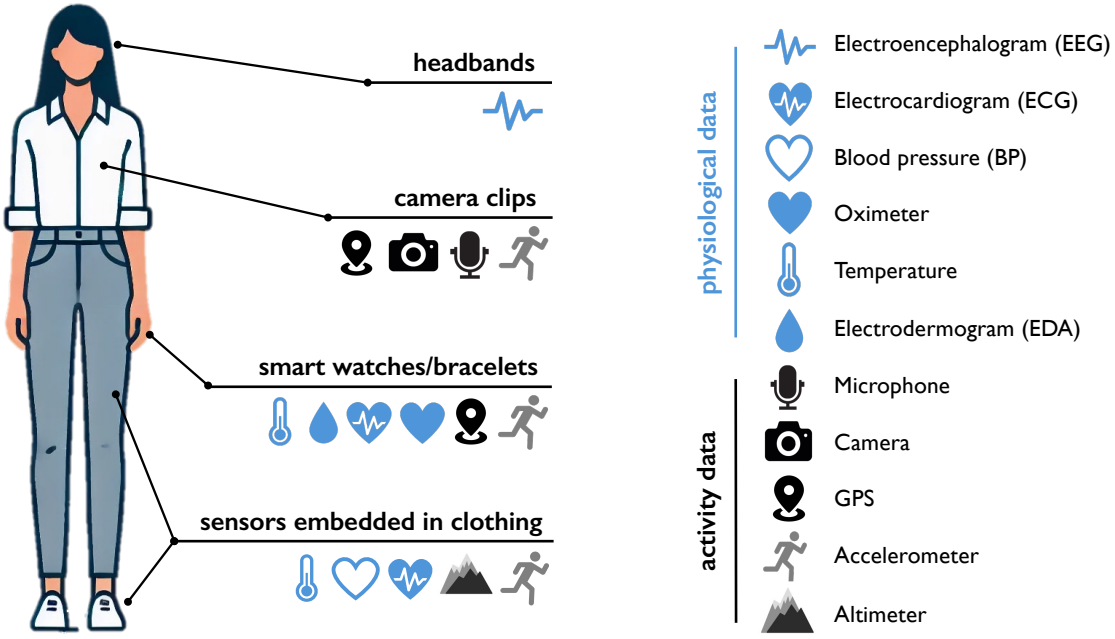


Figure 1.1: Examples of data collected by wearable devices. They can capture a variety of activity metrics and environment data, along with physiological signals providing insights into the wearer's health indicators. (reproduced from Piwek et al. (2016), the illustration of the female figure was produced using DALL·E 3)

1.2.1 Emergence of Wearable Medical Devices and e-Health

Wearable devices are compact and portable technologies designed to be worn on the body. These devices are equipped with sensors and digital interfaces enabling the continuous and real-time data collection and monitoring. As illustrated in Figure 1.2, these devices can capture a variety of activity and environmental data, such as movement through accelerometers, elevation changes via altimeters, GPS location, audio and video – offering insights into the wearer's surroundings.

However, their most significant impact lies in their ability to measure various health-related parameters in the form of physiological data. Traditionally, physiological data such as cardiac activity, cerebral activity or galvanic skin response (i.e. electrodermal activity) was collected in controlled laboratory or clinical environments, often limiting the scope of data to short-term and artificial settings. The increasing portability of wearable devices, as seen in Figure 1.2, has revolutionized this process, making it possible to monitor health metrics in real-world, ambulatory contexts where people live and work. Moreover, most wearable sensors available today have the ability to simultaneously collect multiple complementary

signals. For instance, among wearable devices, smartwatches, connected wristbands and bracelets have become popular options. They offer the ability to simultaneously record signals like heart rate, single-lead electrocardiograms (ECG), electrodermal activity (EDA), temperature, or respiration rates – providing rich multi-source insights in the wearer's health indicators. As such, these devices can benefit health monitoring and preventive care by making it easier to track changes in physiological patterns over time.

Applications of wearable devices. The increasing availability of wearable devices opens up a wide range of opportunities for applications that could significantly enhance the quality of life of our society, trough applications including:

- **Fitness and general health:** the use of wearable devices can promote healthy behaviors (e.g. regular exercise, better sleep), and enhance patient engagement in their own care – which in time can reduce the prevalence of lifestyle-related diseases.
- **Mental health monitoring:** continuous and real-time monitoring of symptoms of stress, anxiety and depression can enable early detection of mental health disorders and intervention.
- **Early detection and preventive care:** wearable devices can identify subtle changes in vital signs, such as heart rate irregularities or oxygen level drops, which might indicate the early onset of medical conditions.
- **Personalized care:** by tracking individual health metrics over time, wearable sensors facilitate personalized care plans tailored to a wearer's physiology, lifestyle, and medical history – which in turn can improve treatment outcomes.
- **Chronic disease management:** for conditions like diabetes, hypertension, or heart disease, wearable sensors can provide patients and healthcare professionals with real-time data to manage symptoms and treatment effectively.
- **Access to healthcare:** wearable sensors enable healthcare professionals to monitor patients remotely, reducing the need for frequent in-person visits. This is especially valuable as it can reduce the need for expensive clinical examinations that are difficult to access for a large proportion of the population.
- **Health research:** the analysis of the vast amounts of wearable devices data can help researchers better understand multitudes of health issues through AI and multimodal analyses.

The unprecedented rise in access to wearable medical devices and advances in AI analysis of large amounts of data have the potential to make healthcare more accessible, efficient, and personalized, ultimately improving health outcomes and quality of life for individuals worldwide.

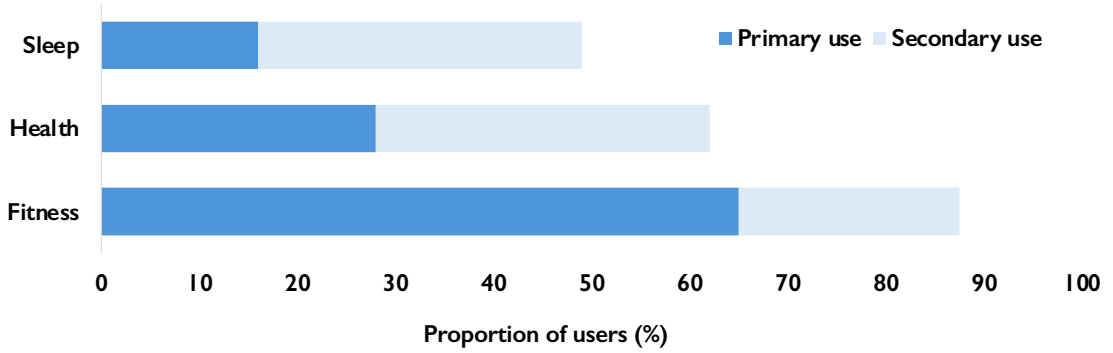


Figure 1.2: Usages of wearable devices reported by owners surveyed in Shandhi et al. (2024).

Usage of wearable devices. Today, medical wearable sensors represent a substantial portion of the e-health market, and this sector is only expected to grow in the coming years (Vaghasiya et al., 2023). Nagappan et al. (2024) analyzed survey responses collected from 2020 to 2022 from 23,974 US participants. Among the respondents, 45% owned at least one wearable device. They have found that healthcare and wellness were among the top *primary use* of wearable devices owners: more than half reported using them for general fitness and heath monitoring (i.e. tracking physical activity, fitness training, losing weight, and improving sleep); more than one-third reported using their devices to manage diagnosed medical conditions; and approximately one-third reported using them for mental health monitoring. Similarly, Shandhi et al. (2024) have conducted an online survey around smart device ownership and usage, and reported that among 1368 respondents, 59% owned at least one wearable device. Approximately 87% of users reported using their wearable devices for fitness and workout monitoring (i.e. heart rate, step tracking, jogging. etc.); more than 60% of owners use them for tracking health indicators tracking (i.e. blood oxygen, heart rhythm, ovulation tracking, etc.); and approximately 50% of users reported using them for sleep monitoring. These studies highlight that digital health monitoring functionalities of wearable devices are among the primary reasons for users owning and using them. As so, wearables are poised to transform healthcare delivery, and their usage could lead to overall improvements in public health outcomes, as they empower individuals and healthcare providers alike with actionable, real-time insights.

More so, AI analysis of multimodal data collected using inexpensive and accessible wearable sensors appears as a promising opportunity to democratize access to healthcare. Although many research studies combine AI and wearable sensors data to address diverse health issues ranging from chronic disease management (Xie et al., 2021; Ahmed et al., 2022) to mental health monitoring (Nahavandi et al., 2022; Abd-Alrazaq et al., 2023) – the use of AI-driven analysis methods is yet limited in real-world applications. A more widespread adoption

of AI systems is needed to foster the all the potential benefits of wearable devices usage. However, major concerns about the trustworthiness of AI remain today.

1.2.2 Challenges of e-Health

In 2019, the European Commission's High-Level Expert Group (HLEG) on Artificial Intelligence released the *European Ethics Guidelines for Trustworthy AI* [4]. This document highlights robustness as one of three fundamental requisites for the development, deployment, and use of trustworthy AI systems – along with adherence to the law and ethics. In particular, these guidelines state that trustworthy AI should be robust from a technical perspective while also taking into account its social environment. From the social perspective, robustness becomes entwined with ethics and the principles of fairness and reliability:

> *"Unfair bias must be avoided, as it could could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination."*

From the technical perspective, it calls for the adoption of a preventive approach to risks in the development of AI systems to ensure they perform as intended. Specifically, the HLEG put forward practical requirements that AI systems should meet:

> *"AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented."*

Currently, both aspects are often overlooked in research, significantly setting back the deployment of trustworthy AI systems (Galati et al., 2022; Zhang and Zhang, 2023; Bürger et al., 2024). Bürger et al. (2024) identify the lack of rigorous definitions for trustworthiness as a cause. Core pillars of trustworthiness, such as reliability and robustness, remain vaguely defined to this day and are often used interchangeably. Following (Galati et al., 2022), this thesis adheres to the definitions from the IEEE Standard Glossary of Software Engineering Terminology [5].

Definition 1.2.1. (Reliability) The ability of a system to perform its required functions under some stated conditions for a specified period of time.

Definition 1.2.2. (Robustness) The degree to which a system can function correctly in the presence of invalid inputs, i.e. those that fall outside some given specifications in which the system is developed.

[4] HLEG, Ethics guidelines for trustworthy AI, (2019)
[5] IEEE Standard Glossary of Software Engineering Terminology, (1990)

In e-health, a major obstacle to the development of trustworthy AI systems resides in the lack of high quality datasets that are representative of the real world. Most of current research in healthcare is conducted on datasets that are collected in controlled clinical environments with limited populations, and that lack diversity in terms of pathologies or patient responses they carry. As a result, subsequent models become unreliable when deployed in real-life, as they are not able to perform their required functions (Definition 1.2.1). This can translate into significant consequences in healthcare:

- **Lack of generalization:** such systems can result in poor generalization to more diverse real-world scenarios, causing errors in diagnosis. Overfitting to skewed data reduces model reliability, and in turn, AI systems may fail to detect critical anomalies or provide inaccurate recommendations (Gulrajani and Lopez-Paz, 2020; Wang et al., 2022b).
- **Fairness issues:** AI models trained on data that is not representative of the real-world can result in biased outcomes, and inequitable treatment. If the training data over-represents certain populations or lacks diversity, models may perform poorly for under-represented groups, exacerbating healthcare disparities (Yang et al., 2024).
- **Ethical issues:** additionally, unrepresentative data can raise ethical concerns, as models may not meet safety or fairness standards. This in turn undermines trust and adoption of AI healthcare applications.

Moreover, currently used datasets are often recorded in optimal conditions that are not met in real-life. For instance, there is a significant difference in the signal-to-noise ratio between physiological data collected in clinical environments and data obtained from wearable sensors. In ambulatory settings, noise and artifacts are more likely to occur. Models trained under the clinical *ideal* conditions lack the robustness needed to handle such perturbations in input data effectively (Definition 1.2.2).

Even more so, one of the most common source of invalid inputs, and thus lack of robustness (Definition 1.2.2), is the presence of missing values in datasets. Most datasets used for research are complete. But in reality, as the size and complexity of real-world healthcare datasets increase, missing values are prone to occur frequently. This can be due to a variety of factors, such as complex data collection processes, the aggregation of multiple data sources, sensor failures, or refusals to answer questions in surveys. Missing values pose significant challenges in AI, and especially in healthcare applications:

- **Technical limitations**: datasets with missing values make supervised learning a challenging task, as they prevent the straightforward use of traditional supervised methods. AI algorithms often rely on complete data to learn patterns and make accurate predictions, and missing values disrupt this process by introducing perturbations in the information available for training.

– **Bias:** they can compromise the quality of models. If not handled properly, missing values can lead to biased or inaccurate models, as certain patterns may become over-represented or ignored entirely (Jakobsen et al., 2017; Nijman et al., 2022).

– **Ethical issues:** in sensitive applications like healthcare, improperly addressing missing values can lead to ethical concerns, unreliable outcomes, and a loss of trust in AI systems. Therefore, developing robust strategies to handle missing values is essential for creating deployable AI solutions.

The aforementioned challenges play important roles in why healthcare professionals and patients may be hesitant to adopt e-health solutions today. They illustrate well the concerns that currently set back the deployment of digital health solutions, and underline how the development of reliable and robust strategies is essential for democratizing access to healthcare worldwide.

1.3 Objectives and Contributions

The the main objectives of this thesis are to develop novel solutions that tackle the challenges introduced in Section 1.2.2, i.e. to guarantee robustness of AI models for healthcare applications – with a particular focus on robustness to missing data. Specifically, I aim to propose innovative methodologies evaluated on real wearable sensor data, and designed to be transposed and adapted to various medical applications. While the works presented here are mainly focused on wearable sensors data analysis, they are designed to be used for various health applications, ranging from physiological signal analysis to clinical studies using multimodal imaging data. To this end, I develop a methodology including all the steps involved in creating an e-health application – from multimodal sensor data collection to model evaluation. The contributions of my thesis are summarized below.

StressID: a novel dataset collected with wearable devices. In the context of a collaboration between EURECOM, Inria Centre Université Côte d'Azur, and the Cognition Behavior Technology (COBTEK) institute, I have contributed to the design and collection of `StressID`: a multimodal dataset for stress identification. It includes 65 participants for whom we collected videos of facial expressions and audio recordings, as well as synchronized electrocardiography (ECG), electrodermal activity (EDA) and respiration signals – recorded using wearable sensors. We designed the `StressID` experimental protocol to increase participants' cognitive load across 11 tasks carefully chosen to induce different levels of stress in them. Each task is associated with 4 self-evaluation questions to assess participants' stress and emotional state. As access to high-quality datasets that are representative of the real world is essential to support the development of reliable and robust applications, we have made `StressID` available for researchers. It is a valuable resource for multiple fields and has the potential to improve the quality of life of our society by supporting the design

of e-health solutions to help prevent stress-related issues. This work has been presented at the NeurIPS 2023 conference, where I received the *NeurIPS Scholar Award*.

Baseline models for the analysis of wearable sensors data. In addition to the dataset, I have proposed a suite of methods for unimodal and multimodal analysis of `StressID`. Through experiments, I effectively show that combining multiple modalities carrying complementary information through multimodal learning has considerable benefits for stress identification. I identify the next steps needed to ensure reliability and robustness of models built on `StressID` – including robustness to missing data. An open-source implementation of these models is also made available. The aim of this initiative is to facilitate future contributions to this domain, by providing reference models for stress identification that can represent a starting point for researchers who wish to use the `StressID` dataset in their work. Even more so, the framework I have designed is representative of the state-of-the-art in physiological signal analysis, and stress identification from multimodal inputs. As such, it has been used for the related study of emotion recognition from multimodal data. This work resulted in a separate co-authored publication at an ECCV 2024 workshop.

Novel guidelines for handling missing values in healthcare. Having identified robustness to missing data as a major obstacle to the development of reliable and deployable AI systems for e-health, I have investigated whether the rich existing literature on missing values can be leveraged to this end. Although widely used in practice, the strategy of imputing missing values before learning can introduce bias in the training data and impact subsequent prediction models – which can lead to severe consequences in sensitive applications. In the context of a collaboration between EURECOM and King's college London, I have developed a framework tailored to assessing these methods' reliability within healthcare applications. In particular, I have designed a *tree*-based approach to help determine how to choose the best model for dealing with incomplete entries, given multiple characteristics of a health-related dataset. By identifying the strengths and limitations of state-of-the-art algorithms, I have derived a set of guidelines to responsibly handle missing values in healthcare data.

PicMi: a robust method for handling missing values. I have proposed a novel method to palliate the limitations of existing approaches to handling missing values in healthcare. Specifically, I have introduced PicMi, an end-to-end *imputation-free* model designed for supervised learning with missing values that uses a permutation-invariant architecture to handle inputs of varying dimensions; integrates missing value patterns as a condition in its objective function to ensure robustness to various missing values scenarios; and is locally interpretable. I have evaluated the method on multiple real-life healthcare datasets with missing values, and demonstrated its advantages over state-of-the-art algorithms.

HyperMM: a robust method for handling missing modalities. I have extended my previous contribution to the task of multimodal learning (MML) with missing modalities. While most works assume modality completeness in the input data, in clinical practice it is common to have incomplete modalities. Existing solutions that address this issue rely on complex, computationally costly modality reconstruction strategies. Instead, I have proposed HyperMM, an end-to-end framework designed for learning with varying-sized inputs. Specifically, it focuses on the task of supervised MML with missing imaging modalities without using reconstruction before training. I have introduced a novel strategy for training a *universal* feature extractor using a conditional hypernetwork, and proposed a permutation-invariant neural network that can handle inputs of varying dimensions to process the extracted features, in a two-phase *task-agnostic* framework. I experimentally demonstrated how the proposed approach can be transposed and adapted to various medical applications. Specifically, I have shown its effectiveness in two tasks: Alzheimer's disease detection and breast cancer classification from multimodal images. I have presented this work at a MICCAI 2024 workshop, where I have received the *Best Presentation Award*.

In addition to the work directly relevant to this thesis, I have contributed to other studies related to the use of AI in e-health and the analysis of physiological data.

1.4 Thesis Organization

This chapter provides an overview of the requirements for the development of reliable and robust AI applications in healthcare. It summarizes the opportunities and challenges of e-health, and the objectives and contributions of this thesis.

Chapter 2 introduces `StressID`, a dataset specifically designed to support the development of robust and reliable applications for the identification of stress from multimodal inputs – including physiological signals collected using wearable sensors, videos and audio recordings.

In Chapter 3, I provide an overview of the state-of-the-art in physiological signal analysis, and stress identification. Building on that, I develop an open-source implementation of unimodal and multimodal models for the analysis of `StressID` – and more generally sensor data – that can be used as a starting point for future researchers.

Chapter 4 explores whether existing state-of-the-art approaches for handling missing values in supervised learning can be reliably used in healthcare applications. To answer this question I propose a qualitative analysis of the performances of a large range of methods. I propose a decision tree-based approach to identify the strengths and limitations of the literature, and ultimately derive interpretable guidelines for addressing this issue.

In Chapter 5, I introduce PicMi, a novel method for handling missing values in healthcare applications that palliate the previously identified limitations of existing approaches. I

propose an approach that bypasses the need for imputation, is robust to diverse missing values scenarios, and is locally interpretable.

Chapter 6 transitions to a MML setting: I propose HyperMM, a framework that extends the previously developed method to the task of MML with missing modalities. I experimentally demonstrate the advantages of the approach on `StressID`, and several applications using multimodal images as inputs.

Finally, Chapter 7 concludes this thesis, and discusses future works and research lines.

Chapter 2

StressID : A Multimodal Dataset for Stress Identification

Contents

2.1 Introduction 13
2.2 Related Work 15
2.3 Design of the StressID Dataset 17
2.3.1 Experimental Protocol 18
2.3.2 Sensors 21
2.3.3 Recruitment and Recording 22
2.4 Dataset Description 23
2.4.1 Contents and Formats 23
2.4.2 Data Annotations 25
2.5 Intended Uses of StressID 29
2.6 Discussion 30

Abstract. The object of this thesis is the development of robust and reliable algorithms for the analysis of wearable device data. Access to high-quality datasets that are representative of the real world is an essential aspect of this task. Yet, the availability of such datasets remains limited. Therefore, it appears as essential to design and collect a dedicated dataset to support the development of such applications. In this chapter, we introduce `StressID`, a new dataset specifically designed for stress identification from physiological signals recorded with wearable sensors, facial expressions, and audio recordings. The work presented in this chapter is the first part of a conference paper published in NeurIPS 2023 (Chaptoukaev et al., 2023).

2.1 Introduction

In this chapter we introduce `StressID`, a new dataset designed and collected to support the development of robust tools for the analysis of wearable sensors data, with a focus on the task of stress identification. Our choice is motivated by both relevance and practical considerations. (1) The increasing prevalence of wearable devices has facilitated the monitoring of physiological signals – including cardiac or electrodermal activity, both highly correlated with stress. Additionally, the field of stress identification has advanced significantly in the last years due the pivotal role played by machine learning and deep learning. (2) The experimental setup for the collection of a dataset focused on stress identification is convenient, cost-efficient, and realistic, as it can be readily approved by research institutes' ethical committees. (3) The research on the effects of stress on health has seen considerable interest lately. While a healthy amount of stress is necessary for functioning in daily life, it can rapidly begin to negatively impact health and productivity when it exceeds an individual's coping level. Indeed, negative stress can be a triggering or aggravating factor for many diseases and pathological conditions (Dimsdale, 2008), and frequent and intense exposures to stress can cause structural changes in the brain with long-term effects on the nervous system (Bremner, 2022). Monitoring of stress levels could play a major role in the prevention of stress-related issues, and early stress detection can be vital in patients exhibiting emotional disorders, or working in high-risk jobs such as surgeons, pilots or long-distance drivers.

In practice however, building robust and reliable models for stress identification requires; (1) understanding and integrating the differences between subgroups of the population to ensure bias free applications, (2) integrating the relationships between physical and physiological responses to stress, (3) studying responses to various categories of stressors, as the perception of stress differs strongly from one individual to another. An essential element to such analyses is high-quality and versatile multimodal datasets that include varied categories of stressors, and are recorded on large and diverse populations. However, existing datasets do not answer these needs. They are generally restricted in size (i.e. a few dozen of participants) and a majority is focused on a single source of data (i.e. physiological signals, video or audio) – although multimodal datasets have considerable advantages (Huang et al., 2021; Jung et al., 2019). Moreover, existing datasets often provide imbalanced subject responses, due to both an inability of the recording protocol to elicit strong reactions, and a lack of diversity in the stimuli – making it difficult to deploy deriving analyses to real-life applications.

To address these limitations, we designed `StressID`, a novel multimodal dataset focused on stress-inducing tasks. We made the dataset available for research purposes at `https://project.inria.fr/stressid/`. `StressID` consists of recordings from 65 participants who performed 11 tasks: a guided breathing task; watching 2 video clips; 7 different interactive tasks; and a relaxation task – as well as the corresponding self-reported ratings of perceived relaxation, stress, arousal, and valence levels (Figure 2.1 presents a summary of the dataset contents and size). As illustrated in Figure 2.2, `StressID` uses a collection

StressID Dataset Facts

Dataset StressID

Motivation

Summary A multimodal dataset for stress identification from video, speech and physiological data from wearable sensors.
Example Use Cases Stress identification, emotion recognition, task classification
Original Authors H. Chaptoukaev, V. Strizhkova, M. Panariello, B. D'Alpaos, A. Reka, V. Manera, S. Thümmler, E. Ismailova, N. Evans, F. Bremond, M. Todisco, M. A. Zuluaga, L. M. Ferrari

Metadata

URL	`https://project.inria.fr/stressid/`
Keywords	Stress recognition, multimodal, wearable sensors
Format	.csv, .txt, .mp4, .wav
Ethical review	Approved by CER/CERNI
Licence	Proprietary
First release	2023

Sensors

ECG	BioSignalsPlux ECG sensor
EDA	BioSignalsPlux EDA sensor
Respiration	BioSignalsPlux Piezoelectric chest-belt
RGB Camera	Logitech QuickCam Pro 9000 RGB
Audio	QuickCam Pro 9000 integrated microphone

Data Annotations

Self-assessments	Stress, relax, arousal, valence
Labels for supervised learning	Binary stress, 3-class stress

Annotated Tasks

Relaxing	Guided breathing, relaxation
Audiovisual	Video clips
Interactive stressors	Cognitive tasks, public speaking, multi-tasking

Participants

Count	65
Gender	72%Male, 28%Female
Age	29 ± 7 years
Background	32%Master students, 20%PhD students, 48%Tertiary

Dataset Size

Total size	5.29GB
Physiological total duration across subjects and across tasks	1119 min
Video total duration across subjects and across tasks	918 min
Audio total duration across subjects and across tasks	385 min

Figure 2.1: A dataset summary card for `StressID`, constructed based on Bandy and Vincent. (2021); Gebru et al. (2021).

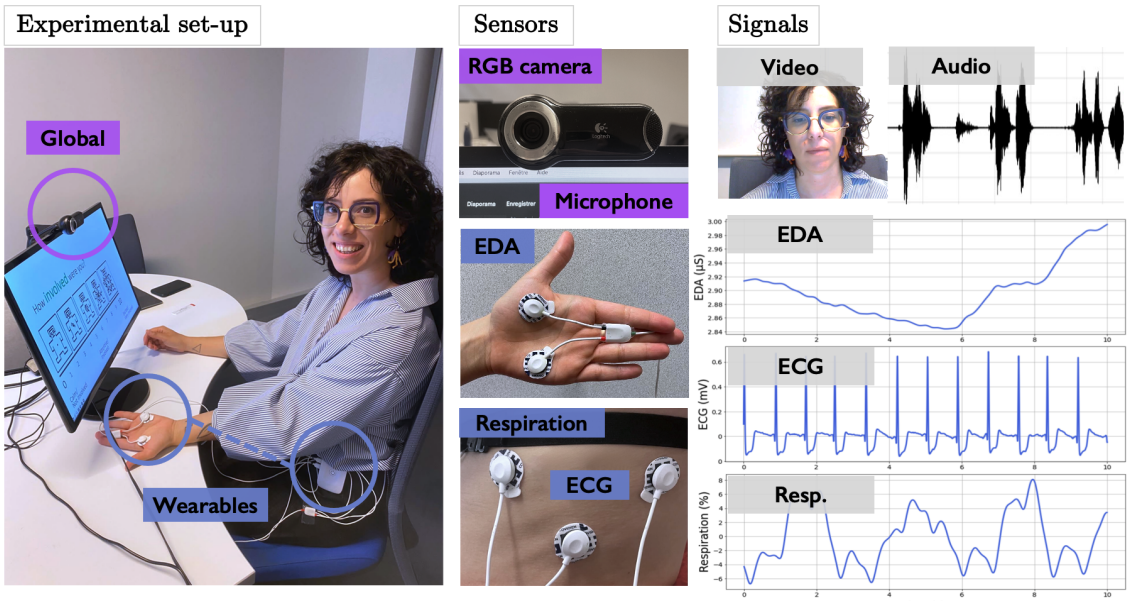


Figure 2.2: Data collection set-up of `StressID`.

of wearable sensors to record the physiological responses of the participants, namely, an electrocardiogram (ECG), an electrodermal activity (EDA) sensor, and a respiration sensor. The data is coupled with synchronized facial video and audio recordings. `StressID` is one of the largest datasets for stress identification, representing more than 39 hours of annotated data in total.

The remainder of this chapter is organized as follows. We provide an overview of the existing datasets for stress recognition in Section 2.2. We then introduce the experimental protocol we have designed for recording behavioral and physiological responses to diverse triggers, using wearable and global sensors in Section 2.3, and describe the contents of the resulting dataset in Section 2.4. We discuss the intended usages of `StressID`, and possible applications in Section 2.5. Finally, we discuss contributions and limitations of our work, and future directions.

2.2 Related Work

Table 2.1 places `StressID` in the context of related stress recognition datasets. The SUS datasets (Steeneken and Hansen, 1999) gather the recordings of 35 subjects collected during aircraft communication training. This unimodal collection of datasets only features audio recordings without self-assessments or external annotation and employs an uncommon elicitation task. SADVAW (Tran et al., 2021) is a dataset composed of 1236 video clips from

Table 2.1: Comparison of `StressID` to related datasets.

Dataset	**#Subjects**	**Modalities**	**Stressors**	**Data annotations**
SUS	35	Speech	Aircraft communication training	Stressor-based
SADVAW	-	Video	-	External annotations
DriveDB	9	EMG, EDA, ECG, HR, Respiration	Driving tasks	Stressor-based
WeSAD	15	ECG, EDA, EMG, BVP, Respiration, Temperature, Acceleration	TSST, Audiovisual	Stressor-based, PANAS (Watson et al., 1988), STAI (Spielberger et al., 1971), SAM (Bradley and Lang, 1994)
CLAS	62	ECG, PPG, EDA, Acceleration	Cognitive load, Audiovisual	SAM
MuSE	28	EDA, HR, Breath rate, Temperature, Audio, Face and upper body video	Public speaking, Audiovisual	PSS (Lee, 2012), SAM, External annotations
SWELL-KW	25	ECG, EDA, Posture, Computer logging, Face and upper body video	Office work with interruptions and time pressure	NASA task load (Hart, 1986), SAM, Stress assessment
Distracted Driving dataset	68	EDA, HR signal, Respiration, Face video, Driving performances	Simulated driving with distractions	Stressor-based, NASA task load, SAM
`StressID`	**65**	**EDA, ECG, Speech, Respiration, Face video**	**Cognitive load, Public speaking, Audiovisual**	**SAM, Stress assessment**

41 Korean movies, making the setting closer to the real world and including a broader range of responses. However, it features video recordings exclusively, restricting deriving applications to computer vision systems only. Among the works investigating the physiological aspect of stress, DriveDB (Healey, 2000) collects physiological data from 9 subjects exposed to driving-related tasks. The lack of self-assessment or external annotations significantly limits the accuracy of measuring stress. In addition, the dataset is collected in the very specific setting of driving, with a narrow range of stressors – considerably restricting its usage. WeSAD (Schmidt et al., 2018) and CLAS (Markova et al., 2019), two of the most widely explored datasets for stress recognition, contain physiological data from 15 and 62 subjects respectively, collected using wearable devices. The participants partake in various tasks, combining perceptive stressors in the form of audiovisual stimuli, with several variations of the Trier Social Stress Test (TSST) (Allen et al., 2017). However, they do not include any behavioral modalities.

There exist a few multimodal datasets for stress recognition, such as MuSE (Jaiswal et al.,

2020) and SWELL-KW (Koldijk et al., 2014). They feature a broader set of modalities and are collected in laboratory environments imitating real-life activities. MuSE participants are elicited through audiovisual and public speaking tasks. SWELL-KW participants perform office work on several topics designed to elicit different emotions. These datasets are limited in size with recordings of respectively 28 and 25 subjects. Finally, the distracted driving dataset (Taamneh et al., 2017) gathers recordings of 68 subjects in the setting of simulated driving with stress-inducing distractions. The lack of diversity in the stimuli restricts subsequent applications to the setting of driving. Moreover, cardiac activity is acquired in terms of heart rate, which does not allow the extraction of heart rate variability (HRV) measures, a key measure in the identification of stress (Kim et al., 2018).

`StressID` aims to fill the gap in the existing related work. It features both physiological and behavioral modalities, includes a large number of participants, exploits varied stimuli, and includes participants' replies to 4 self-assessment questions providing insights on the subject's emotional state. Although CLAS (Markova et al., 2019) and WeSAD (Schmidt et al., 2018) present similar experimental set-ups, they focus on physiological modalities and do not include behavioral data. Instead, `StressID` features three types of modalities: video, audio, and physiological signals capturing complementary information. While MUSE (Jaiswal et al., 2020) and SWELL-KW (Koldijk et al., 2014) are also multimodal datasets recorded in similar conditions, they are very limited in size. On the contrary, with 65 subjects recorded `StressID` is one of the largest datasets designed for stress identification. Finally, although the size and modalities of the distracted driving dataset (Taamneh et al., 2017) are comparable to `StressID`, it relies on very environment-specific stressors, whereas `StressID` includes emotional video-clips, cognitive tasks, and social stressors based on public speaking, which represents a key aspect to guarantee the collection of a wide range of responses. To summarize, `StressID` is the first multimodal dataset for stress identification that is recorded on a large number of participants but also features a wide range of stimuli ensuring more versatility in deriving applications.

2.3 Design of the StressID Dataset

`StressID` is designed specifically for the identification of stress from different triggers. It uses a collection of wearable and global sensors to record the physiological and physical responses of 65 participants to 11 varied stress-inducing stimuli – such as emotional video clips, cognitive tasks including mathematical or comprehension exercises, and public speaking scenarios, designed to trigger a diverse range of emotional responses. Each task is associated with 4 different annotations: scores from a self-assessment rating perceived stress and relaxation, along with valence and arousal based on the Self-Assessment Manikin (SAM) (Bradley and Lang, 1994).

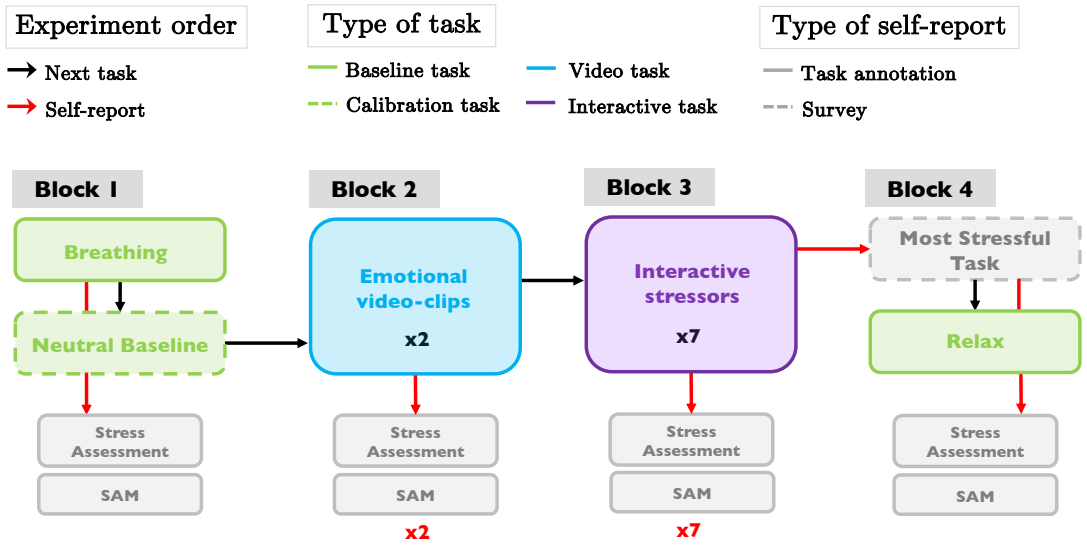


Figure 2.3: Overview of the experimental protocol of `StressID`. The experiment consists of 11 tasks divided into four blocks: a guided breathing task, 2 emotional video clips, 7 interactive stressors, and a relaxation task.

2.3.1 Experimental Protocol

All the stressors used in `StressID` are *mental* stressors. They are based on established clinical methods to induce stress in subjects (Bali and Jaggi, 2015), such as : the Mental Arithmetic Task (MAT) – one of the most commonly used stimuli for inducing stress, designed to increase the mental workload by having the subject perform a series of arithmetic operations with a varying range of difficulty, aloud. This stimulus is easy to implement and does not requires any special instrument Arsalan et al. (2022a); or, the Stroop Color-Word Test (SCWT) – a neuropsychological test that is extensively used for both experimental and clinical purposes. In the most common version of the SCWT, subjects are required to read named color-words printed in an inconsistent color ink (Stroop, 1935).

In addition, the choice and design of stressors is based on several considerations. (1) Tasks have been designed to elicit 3 different categories of responses: stimulate the audiovisual cortex of the participants; increase the cognitive load by soliciting attention, comprehension, mental arithmetic or multi-tasking abilities; and elicit psycho-social stress leveraging on public speaking as a stressor. (2) The tasks of the experiment are overall short – which allows the participants to perform several tasks in a row without tiring or losing acuity by the end of the experiment. (3) All interactive tasks are designed to leverage time restriction as a stressor by having a strict requirement for a response in 1 minute – thus, after receiving instructions on the screen, the subjects see a ticking 1-minute clock during the execution of

each task. (4) The order of the stressors is designed to be unexpected to the participants. Therefore the experiment alternates between subgroups of tasks (e.g. all mental arithmetic tasks do not come all at once). (5) The level of detail provided in the instructions as well as the duration of the instruction were also carefully thought to maximize levels of stress in the experiment, by preventing participants from preparing for the coming task. (6) Finally, all stimuli are easy to implement and do not require any special setup (Arsalan et al., 2022b).

The experimental protocol used to collect `StressID`, illustrated in Figure 2.3, was designed to have a total duration of 35 minutes. It consists of 11 tasks separated by self-assessments and grouped into 4 blocks: guided breathing, watching emotional video clips, a sequence of interactive tasks, and a relaxation phase.

Guided breathing. The first block of the protocol consists of the single task of *Breathing*. The participants watch a guided breathing video of 3 minutes. It aims to relax and reset to neutral the emotional state of the subjects. This recording is used as a baseline for the non-verbal neutral state of each participant. After the breathing task, the participants count forward for 1 minute.

Emotional video-clips. This block consists in watching 2 emotional videos clips, retrieved from the FilmStim database (Schaefer et al., 2010). These videos have been selected to elicit specific emotional responses.

- *Video1 :* an extract from the movie *There's something about Mary*, selected to elicit low arousal and positive valence in the participants.
- *Video2 :* an extract from the movie *Indiana Jones and the Last Crusade*, selected to elicit high arousal and negative valence.

Interactive tasks. This block consists of a sequence of 7 interactive stressors based on well-established clinical methods to induce stress (Bali and Jaggi, 2015). All the tasks have a strict requirement for response in 1 minute and the order of the stressors is designed to be unexpected to the participants.

- *Counting1 :* a MAT designed to increase the participants' cognitive load through arithmetic operations with a varying range of difficulty. In this task, the participants receive the instructions to count backwards from 100 subtracting 3 as fast as they can.
- *Counting2 :* another MAT of increased difficulty. Participants are asked to count backward from 1011 subtracting 7 as fast as they can.
- *Stroop :* a variant of the Stroop Color-Word Test (Stroop, 1935), selected to increase the cognitive load by soliciting the attention and reactivity of the participants.

- *Speaking :* a Social Evaluative Task (SET), leveraging public speaking as a social stressor. The subjects are instructed to explain their strengths and weaknesses, emulating stressful job interview conditions.
- *Math :* a task designed to increase the mental workload. The participants are asked to resolve 20 mathematical problems in one minute.
- *Reading :* a task composed of 2 phases and designed as a TSST variation. Participants have to read a text, in the first step, and then explain what they read, in the next step, thus simultaneously soliciting comprehension abilities and using speaking as a stressor.
- *Counting3 :* a MAT with added difficulty. Participants are instructed to count backwards from 1152 subtracting 3, as fast as they can, while repeating an independent hand movement. This task is designed to increase the mental workload by soliciting participants' multi-tasking abilities.

At the end of the third block, the participants are asked to designate the task perceived as most stressful.

Relaxation. The last block of the experimental protocol is solely composed of the *Relax* task. It consists of a 2 minute and 30 seconds long relaxation part, where participants are instructed to watch a relaxing video (Gros et al., 2017).

Each of the 11 tasks is followed by 4 self-assessment questions. The first 2 questions establish the participant's perceived stress and relaxation levels on a 0-10 scale. The following 2 questions are based on the SAM (Bradley and Lang, 1994), and establish the participants' valence and arousal on a 0-10 scale. Psychological evidence suggests that these two dimensions are intercorrelated (Kuppens et al., 2013). More so, research suggests relaxation and stress conditions can be described in different quadrants of the arousal-valence space. For instance, high arousal and negative valence are characteristics of emotional stress induced by threatening stimuli (Christianson, 1992), while low arousal and positive valence are characteristics of a calm and relaxed state (McManus et al., 2019).

The counting forward baseline section is not defined as a task, but is designed to keep the participants in a neutral affective state, therefore it is not coupled with any self-assessment. Additionally, participants answer a survey question at the end of the experiment and indicate the task they considered most stressful. Examples of instructions, tasks and self-assessments questions presented to the participants are provided in Appendix A.1.

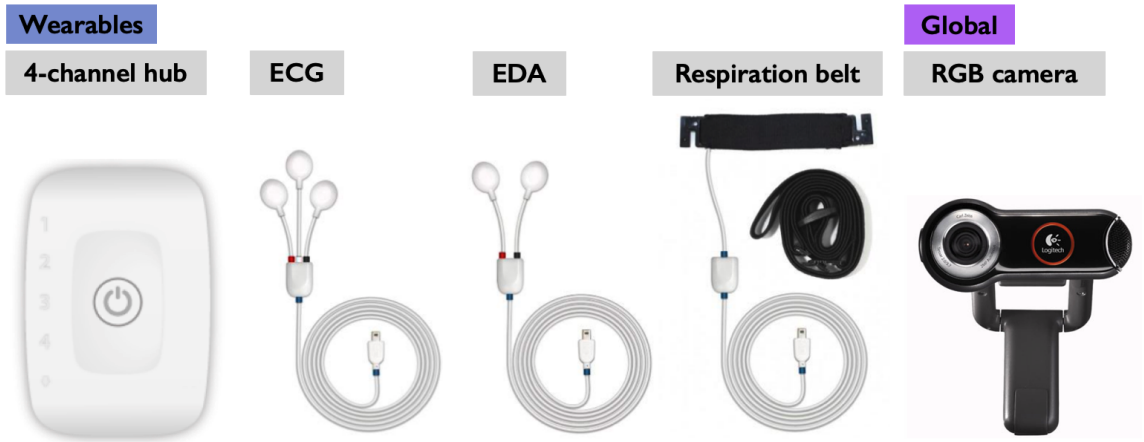


Figure 2.4: Sensors used in `StressID`. A 4 channels Biosignalsplux hub is used with a single lead ECG sensor, an EDA sensor and a respiration belt, to record physiological signals. An RGB camera with integrated microphone is used to record video and audio signals.

2.3.2 Sensors

All three physiological signals collected in `StressID` are recorded using the BioSignalsPlux acquisition system[1]. The BioSignalPlux kit consists of a 4-channel hub communicating via Bluetooth with the OpenSignals (R)evolution platform for data visualization and acquisition, connected to an ECG, EDA, and a respiration sensor. The devices used are illustrated in Figure 2.4. The hub ensures the synchronized recording of up to 4 sensors simultaneously. The ECG is acquired with 3 Ag/AgCl electrodes located on the ribs of the non-dominant side of the subjects. The EDA is measured with 2 Ag/AgCl electrodes attached to the palm of the non-dominant hand. The respiration is measured through a chest belt with an integrated piezoelectric sensing element. The selected devices have a high signal-to-noise ratio (PLU, 2020, 2021a,b), and all physiological signals are acquired with a sampling rate of 500 Hz and resolution of 16 bits per sample.

The data is coupled with synchronized facial video and audio recordings. The video and audio are acquired using a Logitech QuickCam Pro 9000 RGB camera with an integrated microphone. The video is acquired with a 720p resolution and a rate of 15 frames per second. The audio is recorded at a sampling rate of 32kHz and a resolution of 16 bits per sample. Additional details on the calibration and synchronization of the sensors are available in Appendix A.2.

[1] biosignalsplux, PLUX wireless biosignals S.A. (Lisbon, Portugal)

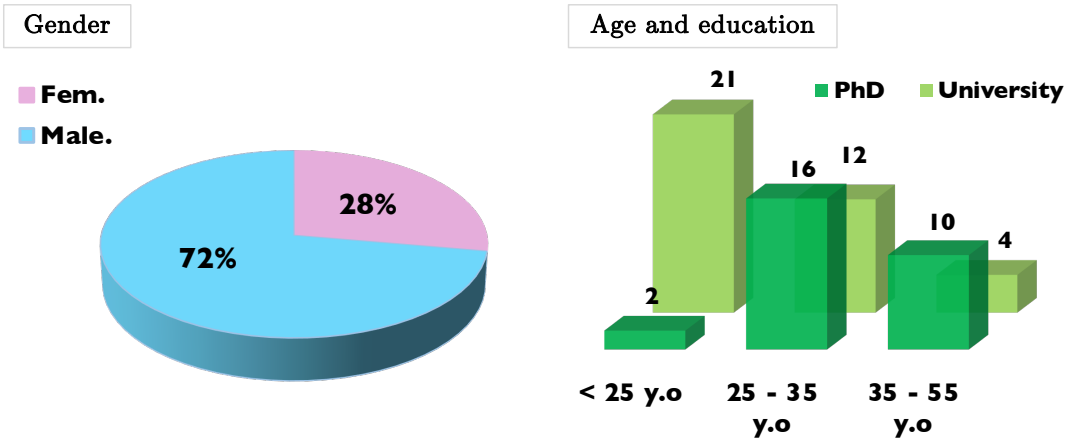


Figure 2.5: Demographics of the `StressID` dataset.

2.3.3 Recruitment and Recording

In total, 65 healthy participants were recruited on a voluntary basis, without compensation. Subjects were recruited by email, and word of mouth primarily. Each participant was recorded in a single session, lasting approximately 50 minutes including preparing sensors, calibration, and the 35 minutes long experiment. They were instructed not to smoke, intake caffeine, or exercise 3 hours before the experiment. At the beginning of each session, they were introduced to the purpose and content of the study. The experiments are conducted entirely in English. The experimental protocol was identical for all participants, and the experimenter was always present in the room during the recording. The participants could either consent to, **Option A**: research use and public release of all their recorded data, including identifying data (i.e. physiological, audio, and video); or **Option B**: research use of all their recorded data, but no public release of identifying data (i.e. only physiological and audio data, but no video). Among the 65 participants, 62 opted for option A and 3 opted for option B. We discuss in more details the considerations of working with human data in Appendix A.3.

Most of the participants of `StressID` are Science, Technology, Engineering, and Mathematics (STEM) students and workers. The demographics of the participants are illustrated in Figure 2.5. The participants included 18 women and 47 men of ages ranging between 21 and 55 years old (29y.o. $\pm$ 7). Among the participants, 32% were master students and interns, 20% PhD students, and the remaining 48% represented diverse tertiary professions. All subjects were required to have sufficient proficiency in English.

2.4 Dataset Description

Following data collection, we split each recorded session into individual tasks. In total, the final task-split dataset amounts to approximately 19 hours of annotated physiological data, 15 hours of annotated video data, and 6 hours of annotated audio data, thus amounting to more than 39 hours of data in total.

2.4.1 Contents and Formats

For each modality, we split the 35 minutes long recordings into 11 individual tasks: one 3 minutes breathing recording (block 1), 2 recordings corresponding to the watching of the video clips of respectively 2 and 3 minutes (block 2), 7 separate 1-minute recordings of the interactive tasks (block 3), and a 2 minute and 30 seconds long relaxation recording (block 4). As the guided breathing, the video clips and the relaxation parts do not carry meaningful audio, the audio part of the dataset consists of the 7 talking tasks only. During the acquisitions, due to camera malfunctions, video and audio recordings of 8 participants were damaged. In addition, while we keep them in the dataset, the EDA recordings of 7 subjects were damaged during certain tasks (total of 19 tasks), as illustrated in Figure 2.6. After splitting, `StressID` is composed of 711 distinct annotated recordings of the physiological modalities, 587 annotated videos, and 385 annotated audio recordings. Figure 2.7 visualizes the proportions of missing data per modality.

Each task is identified in the dataset by `subjectname_task`, where the task names are as described in Section 2.3.1. This convention facilitates different types of analyses, whether subject-specific or task-specific. For each modality, all tasks are grouped by subject into separate repositories. For each task, data from all wearable sensors is organized into a single `.txt` file. Each file contains 3 synchronized entries corresponding to the ECG, EDA, and respiration data respectively. In a similar fashion, for each task, the video data from the Logitech QuickCam Pro 9000 RGB camera is contained in a `.mp4` generated video file. Audio data is represented in the dataset as uncompressed `.wav` files.

Along with the unprocessed self-assessments provided by the participants, we propose 2 discrete labels that can be used to train supervised models: a 2-class label and a 3-class one. The 2-class label is computed using the stress self-assessment of each task by splitting the 0-10 scale at 5: self-assessment of stress below 5 is considered **not stressed** (0) while equal or above 5 is **stressed** (1). The 3-class label is based on the results outlined by Christianson (1992); McManus et al. (2019), which are in line with the observations drawn from Figure 2.8. It allows the prediction of **relaxed** vs. **neutral** vs. **stressed**. We considered a subject to be **relaxed** (0) for a task where they reported a valence rating above 5, an arousal rating below 5, and a perceived relaxation rating above 5. Similarly, we label tasks with arousal levels above 5, valence levels below 5, and perceived stress levels above 5 as **stressed** (2), and **neutral** (1) otherwise. We provide 3 `.csv` files containing self-assessments and labels.

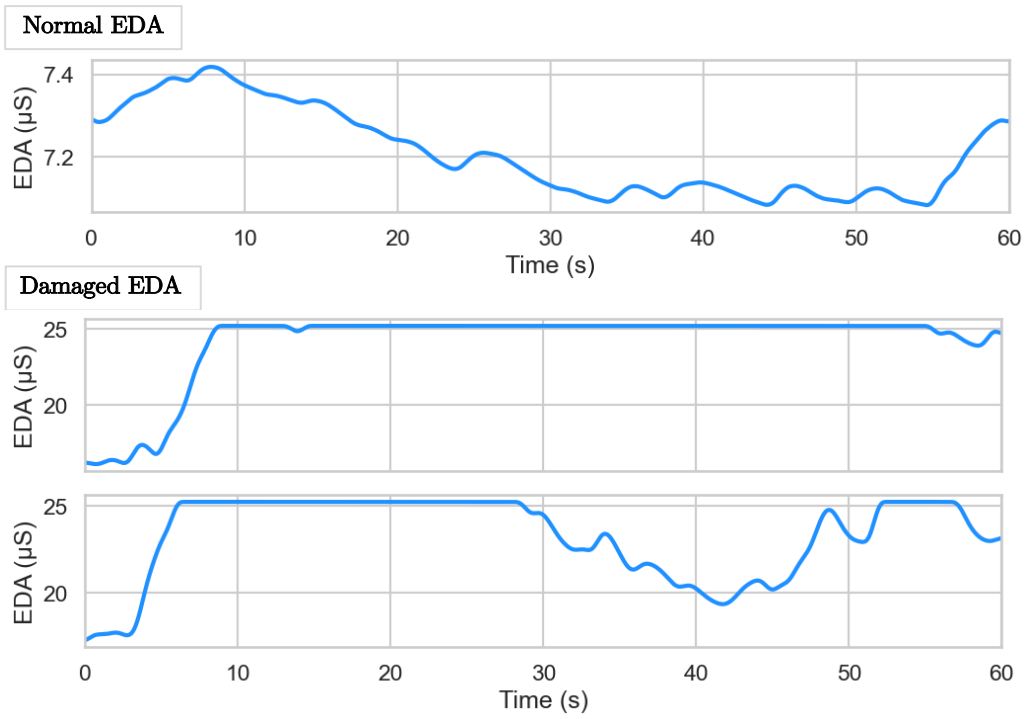


Figure 2.6: Illustrations of EDA recordings damaged due to sensor satuation. For several participants, the recorded signal exceeds the maximum value that the sensor can measure (25μS).

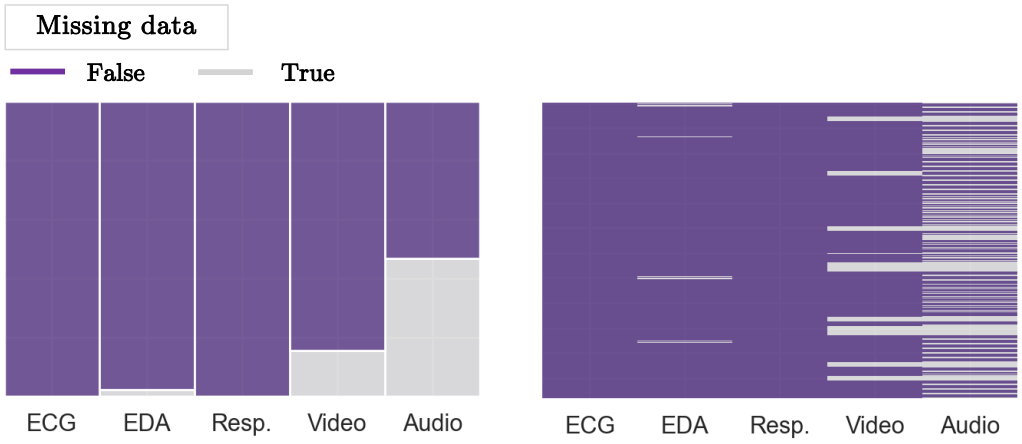
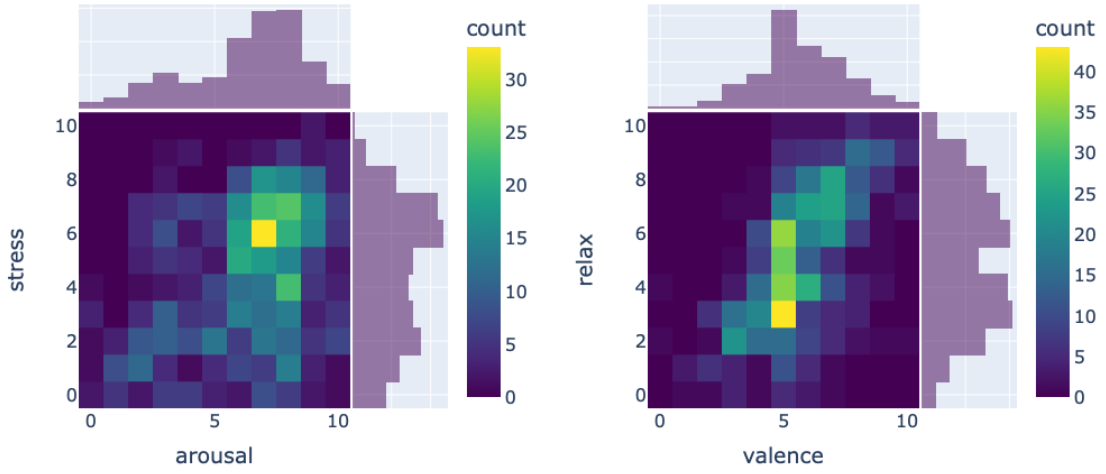


Figure 2.7: Proportions and repartition of missing modalities in the `StressID` dataset.



(a) Sress/Arousal

(b) Relax/Valence

Figure 2.8: Distribution of the self-assessment answers. (left) Joint and marginal distributions of stress and arousal. (right) Joint and marginal distributions of the relax and valence ratings.

2.4.2 Data Annotations

An overview of the distributions of the self-assessments is reported in Figure 2.8. The analysis of the distributions highlights a positive correlation between stress and arousal, as well as relax and valence. This suggests that across subjects and tasks, a high arousal is associated with a higher level of stress, and a positive valence corresponds to a higher level of relaxation. In addition, the marginal distributions of stress and relax ratings highlight a balance in the perceived stress and relaxation levels of the participants across the whole experiment, suggesting that the experimental protocol of `StressID` can arouse proportional instances of stress and relaxation. Furthermore, the distribution of arousal is significantly skewed towards a high rating across the dataset, while valence is centered around a neutral value, highlighting the ability of the protocol to create a high involvement in the participants and elicit strong responses.

Most stressful task. Fig. 2.9 shows the distribution of the answers to the question survey at the end of the experiment i.e. which task was perceived as most stressful for each subject. Approximately 30% of the participants of `StressID` designated the task *Counting2* as most stressful, 20% designated the task of public *Speaking*, 15% designated the task *Counting3*, while the remainder 35% chose between *Stroop*, *Math*, *Reading*, and *Counting1*. Although

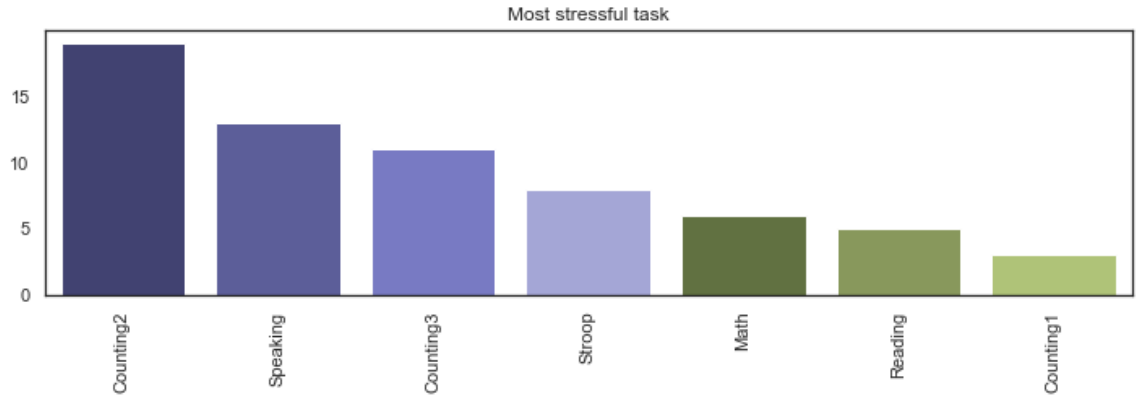


Figure 2.9: Most stressful tasks as designated by participants of `StressID`.

a majority of participants agreed on *Counting2* as the strongest stressor, this analysis highlights the advantages of relying on multiple and diverse stressors in an experimental protocol designed for stress identification. Perception of stress and relaxation can vary a lot from one participant to another – and more so, the effectiveness of a stressor can vary from one subject to another; while an arithmetic task can be a strong stressor for one individual, it can be an uneventful task for another.

Participant-specific distributions of `StressID` annotations. We analyze the distributions of the stress, relaxation, arousal, and valence self-assessments for each participant of `StressID`. To have a global vision of the dataset, for each self-assessment question we represent on a single figure all subject-specific Kernel Density Estimate (KDE) plots in Fig. 2.10. The KDE plot, analogous to a histogram, represents the distribution of self-assessment data – only using a continuous probability density curve. Several observations can be drawn from Fig. 2.10. First, for all 4 self-assessment questions, the participant-specific distributions are rather heavy-tailed, with the exception of a few subjects. This suggests that each participant of `StressID` gave a broad range of self-assessed scores across the experiment, affirming the ability of the `StressID` protocol to elicit varied responses. Second, the perceived stress and relaxation levels of the participants across the experiment are well balanced, suggesting the experimental protocol enabled the creation of a dataset with proportional instances of stress and relaxation across tasks. Finally, we observe that the distribution of arousal scores is significantly skewed towards higher ratings across the dataset, highlighting the protocol's ability to create a high involvement in the participants and elicit strong responses – whether stress or amusement.

Joint distributions of `StressID` annotations. We analyze the pair-wise joint distributions of the `StressID` annotations in Fig. 2.11. The analysis highlights a linear relation

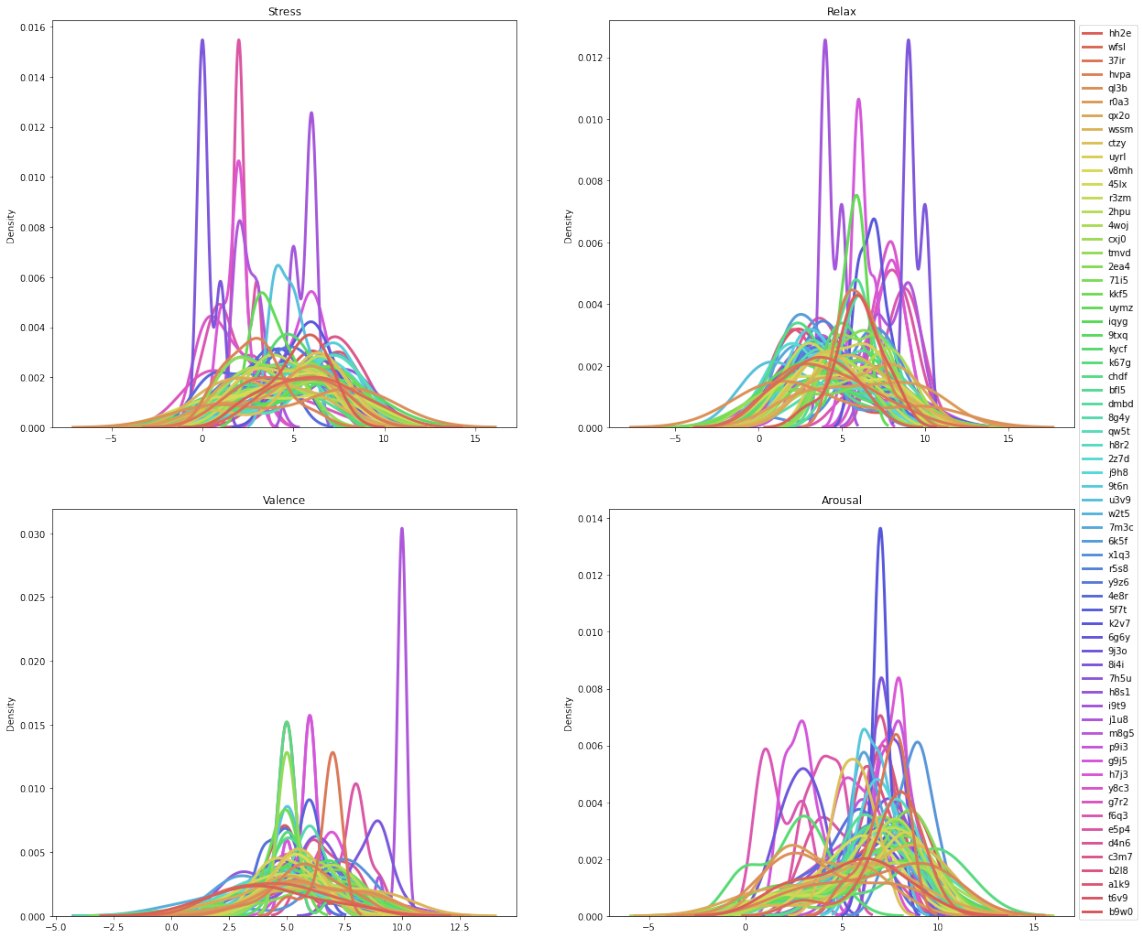


Figure 2.10: Participant-specific KDE plots for each of the self-assessment questions.

between stress and relaxation levels. In our experimental protocol, the participants' perceived levels of relaxation and stress associated with each task are mutually exclusive – globally, a subject cannot be both relaxed and stressed during a task. In addition, Fig. 2.11 highlights a positive correlation between stress and arousal, and a negative correlation between stress and valence – suggesting that across subjects and tasks, high arousal and low valence are associated with a higher level of stress. Similarly, relaxation is positively correlated to valence, and negatively correlated to arousal – suggesting low arousal and positive valence corresponds to higher levels of relaxation. These observations are consistent with psychological studies Christianson (1992); McManus et al. (2019) describing stress on the circumplex model of affect Russell (1980), thus once again affirming the coherence of the `StressID` dataset.

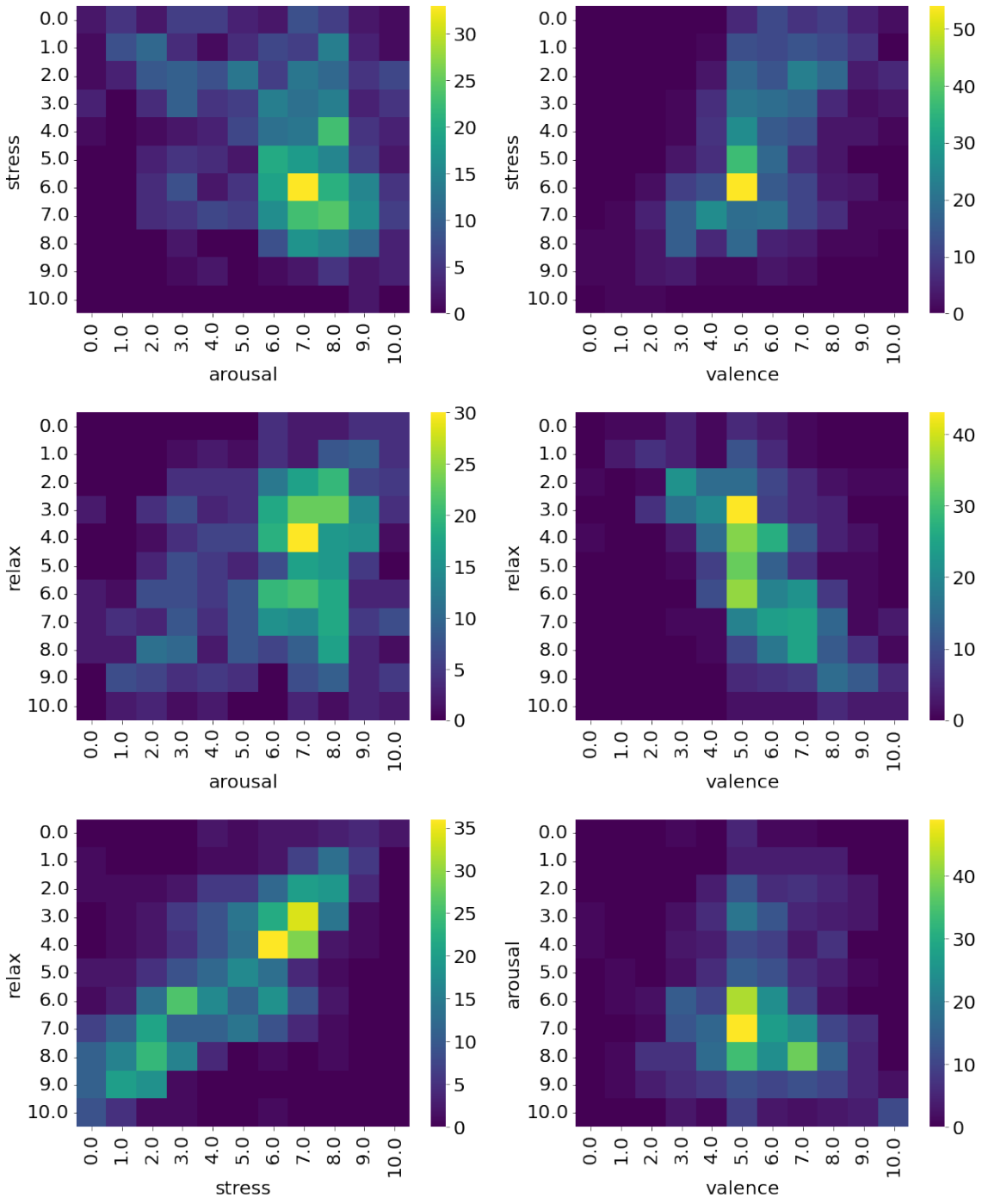


Figure 2.11: Joint distributions of pairs of self-assessment answers.

2.5 Intended Uses of StressID

`StressID` is conceived to support the development of robust, reliable and versatile automated stress recognition algorithms. Rather than focusing on a single task, `StressID` features responses to several categories of stress-inducing stimuli to account for the variability of responses from one individual to another. In addition, the dataset is recorded in an ambulatory setting using wearable devices. Although this makes the data susceptible to noise, artifacts and missing data due to sensors failure, it also makes it more representative of the real world. These aspects enable the design and validation of models that are robust to perturbations encountered in real-life. The large number of participants of the `StressID` also makes it possible to analyze the demographics associated with stress, based on factors such as gender or age, thus advancing towards the design of reliable and bias-free algorithms that integrate these differences.

In addition, the multimodal nature of `StressID` offers a large set of possible uses cases and applications. Diverse modalities carry complementary information that can be jointly exploited: video and audio capture the behavioural component of emotions – the reactions that are visible from outside, while the physiological signals capture valuable internal states not visible on camera such as cardiac activity, or skin sweating. By providing access to multiple synchronized modalities, `StressID` enables cross-modal analyses that have the potential to improve the understanding of the relationships between video, audio, and physiological responses to stress.

Moreover, the design of the `StressID` dataset supports a variety of learning pipelines, by offering possibilities for the analysis of subject-specific, task-specific and modality-specific responses to stress. Various use cases include extracting characteristics of stress from each modality, analyzing correlations between various modalities, analyzing how the modalities relate to specific tasks, training learning pipelines for the identification of stress in diverse verbal and non-verbal tasks, and training pipelines to discriminate between audiovisual stimuli, stressors designed to increase the cognitive load or stressors based on public speaking. More so, it can be used to advance research in multiple other fields, such as emotion recognition, affect understanding, or multimodal deep learning.

Ultimately, `StressID` is designed as a resource for improving the monitoring, modeling, and understanding of the mechanisms of human stress conditions. All intended applications have the potential to improve the quality of life of the population by helping prevent stress-related issues. However, recording and usage of human activity data is associated with high ethical implications, including privacy, bias, and impact on society. We further discuss these considerations in Appendix A.4.

2.6 Discussion

We present StressID, a dataset for stress identification featuring three categories of data modalities and three different types of stimuli. The experimental protocol designed to collect the StressID dataset is easy to replicate and can be adapted to additional sensors or stressors. The equipment used for the data collection is affordable, and the selected devices guarantee low noise in the recordings.

However, it is important to mention some limits that StressID suffers from. (1) This dataset is recorded in a relatively controlled environment specifically designed to elicit responses, where the process of attaching electrodes to the participants may be stressful in itself. Experiments conducted in laboratory settings do not take into consideration the external factors that contribute to the psychological mental state of participants and typically assume a stress reaction is an isolated occurrence. In reality, human emotions are complex and are influenced by combinations of factors. (2) Relying on self-assessed scales for data annotation is a participant-subjective process, and can lead to bias in subsequent analyses. Perception of stress and relaxation can vary a lot from one participant to another. Nevertheless, analyses described in Section 2.4.2 highlight a coherent distribution of the self-reported annotations across participants and the whole experiment. (3) Although all participants recruited for the study are proficient in English, the act of speaking English itself can be stress-inducing for non-native speakers. (4) Although the distribution of the self-assessments across the dataset is reasonably balanced, the audio component of the dataset suffers from an uneven distribution of labels, as the verbal tasks are associated with higher levels of stress. (5) StressID suffers from missing modalities for some participants, as discussed in Section 2.4.1. This makes learning from multimodal inputs a challenging task as it prevents the straightforward use of traditional supervised learning methods. (6) Finally, participants for the data collection were included in our dataset without restrictions on gender, race, age, or education level – instead favoring sample size. As so, StressID presents a gender imbalance representative of the female/male ratio in STEM studies and workforce (UNESCO Institute for Statistics, 2019). This is a limitation StressID shares with competitor datasets (Koldijk et al., 2014; Taamneh et al., 2017; Schmidt et al., 2018; Markova et al., 2019; Jaiswal et al., 2020), and a common issue in human data collection, in general (D'Mello et al., 2022; Pinho-Gomes et al., 2022). Therefore, systems that use the dataset for modeling and understanding the mechanisms of human stress conditions need to be aware of the potential imbalance in representation in the dataset.

Nonetheless, StressID is a valuable resource for multiple research fields. First, it has the potential to improve the understanding of the sources, demographics, and both physical and physiological mechanisms of stress responses. It is designed for the development of reliable algorithms for stress identification that can improve the quality of life of our society by helping prevent stress-related issues. For instance, early stress recognition can be beneficial for people suffering from neurological or developmental disorders with emotion deregulation,

such as autism, for whom the increase of stress can cause disruptive behaviors. Second, `StressID` can help improve affect understanding, as it offers the possibility to analyze and understand the correlation patterns between the distributions of perceived stress and emotion, how these correlations relate to different categories of stimuli, or how they impact subsequent stress and emotion recognition algorithms. Finally, `StressID` is useful to the machine learning and deep learning communities as well, as it can be used to further evolve multimodal learning algorithms, to develop strategies for learning with unevenly represented modalities, or to study how to make algorithms learning with human data more reliable.

In chapter 3, we introduce and implement a suite of reference models for the identification of stress from the unimodal and multimodal inputs of `StressID` – a natural next step to the work presented here. Doing so, we aim to assess the limitations of existing works and outline the necessary steps to ensure robustness and reliability in stress identification algorithms.

Chapter 3

Stress Identification from Physiological Signals, Videos and Audio Data

Contents

3.1 Introduction 33
3.2 State-of-the-Art 33
3.3 Baseline Models for Stress Identification 35
3.3.1 Unimodal Models 36
3.3.2 Multimodal baselines 41
3.4 Main Limitations 43
3.4.1 Missing Data 44
3.4.2 Gender Imbalance 44
3.5 Discussion 45

Abstract. A crucial step in the development of robust methods for the analysis of wearable sensors data is identifying where existing methods fail, in order to improve these aspects. In this spirit, we implement a set of unimodal and multimodal models that are representative of the state-of-the-art in stress recognition, and apply them to the `StressID` dataset to try and understand their strengths and weaknesses. The work presented in this chapter is closely related to Chapter 2, and corresponds to the second part of a conference paper published in NeurIPS 2023 (Chaptoukaev et al., 2023).

3.1 Introduction

In this chapter, we introduce a suite of unimodal and multimodal models for the study of stress using the `StressID` dataset. The objectives of this initiative are threefold. (1) We aim to define what currently constitutes the state-of-the-art in the domain of automated stress identification from physiological signals, and thus provide a clear reference for the research community. (2) We intend to facilitate future contributions to this domain by providing an open-source implementation of reference models for stress identification, that can represent a starting point for researchers who wish to use the `StressID` dataset in their work. (3) Lastly, we aim to identify key limitations of current approaches. We put a particular focus on their ability to handle perturbations commonly encountered in real-world applications using sensor data, such as the presence of noise and artifacts, missing data, or representation bias. Handling these aspects is critical for advancing the field towards real-world usability. By providing a discussion of the strengths and weaknesses of the current state-of-the-art, we aim to encourage the development of more innovative, robust and reliable solutions that can be deployed in real-world settings.

We focus on the tasks of supervised stress identification using the 2-class and 3-class labels defined in Section 2.4.1. While the primary focus of this thesis is the analysis of physiological signals collected from wearable sensors, we recognize the value of video and audio data in the stress identification domain. Therefore, in this chapter we propose baseline models for all modalities of `StressID`. In particular, we provide unimodal models that focus exclusively on individual modalities (i.e. physiological modalities, video, and audio), along with multimodal models that combine data from the physiological and physiological modalities. All the baselines introduced hereafter are selected to be representative of the state-of-the-art in the domain (Gedam and Paul, 2021; Garg et al., 2021; Vos et al., 2023). All implementations are available at `https://github.com/robustml-eurecom/stressID`.

The remainder of this chapter is organized as follows. In Section 3.2, we provide an overview of the state-of-the-art in automated signal analysis. In Section 3.3, we present the multiple baselines we have identified for stress detection, and illustrate them on the `StressID` dataset. We then discuss the current limitations of the existing approaches and identify the necessary steps to ensure robustness and reliability of applications using wearable devices data in Section 3.4. Finally, we summarize our work and discuss future directions.

3.2 State-of-the-Art

The general framework for automated signal analysis, illustrated in Figure 3.1, involves three key steps: (1) pre-processing of the raw data, which usually includes steps like denoising, downsampling or windowing the raw data; (2) processing – which typically consists of feature extraction and feature selection; (3) analysis or prediction using machine and deep learning algorithms. Current research in this area typically highlights three main approaches

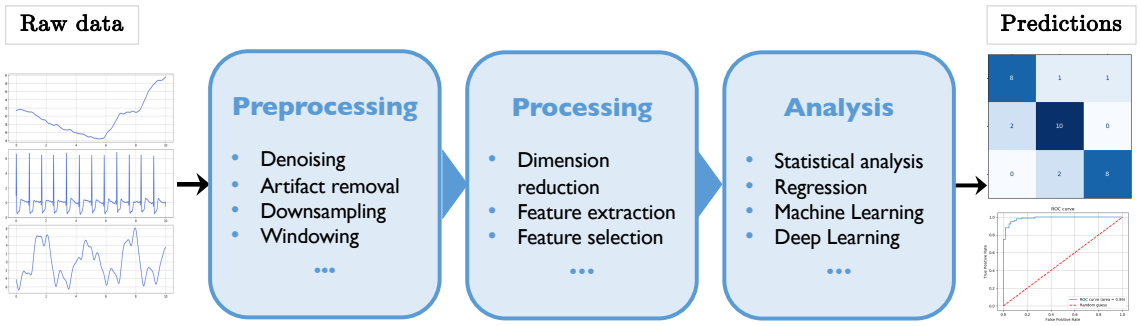


Figure 3.1: Overview of the general pipeline for automated signal analysis.

following this framework: machine learning (ML)-based methods, deep learning (DL)-based methods and hybrid methods. We focus on the literature on the analysis of physiological signals hereafter. Nonetheless, state-of-the-art models for stress identification from video and audio modalities follow the same approach (Aigrain et al., 2016; Giannakakis et al., 2020; Jaiswal et al., 2020; Arsalan et al., 2022b; Ahmed et al., 2023).

In ML-based methods, relevant features with discriminative values for different classes must be manually extracted from pre-processed physiological data before training a prediction model. This process amounts to transforming complex signals into sets of discrete, meaningful descriptive measures refered to as *handcrafted* (HC) features. Creating a tabular database of HC features extracted from signals is a widely used approach in the analysis of electroencephalography (EEG) (Boonyakitanont et al., 2020; Wang and Wang, 2021; Ein Shoka et al., 2023), ECG (Neha et al., 2021; Mir and Singh, 2021; Singh and Krishnan, 2023), or wearable sensors data (Sánchez-Reolid et al., 2020; Giannakakis et al., 2019; Arsalan et al., 2022b). The computational complexity of these features varies, ranging from simple statistical measures in the time domain (e.g., mean, variance) to more complex non-linear features (e.g., entropy measures), as well as features from the frequency and time-frequency domains. Such approaches have obvious advantages: they considerably decrease the complexity of the input data; many tools including the `Neurokit2` (Makowski et al., 2021) or `MNE` (Gramfort et al., 2013) python libraries are openly available for feature extraction; and HC features can be processed using traditional ML algorithms such as support vector machines (SVM), random forests (RF), k-nearest neighbours (kNN) or gradient boosted decision trees – which are known to outperform elaborated deep learning methods on tabular data (Grinsztajn et al., 2022). However, ML-based approaches rely heavily on domain expertise for feature engineering, and their performance can be sensitive to the quality of the extracted features.

In contrast, DL-based methods are end-to-end. They directly handle pre-processed data and integrate the steps of feature extraction, selection, and prediction, thus eliminating the need for manual feature engineering. In particular, Convolutional neural networks

(CNN) are increasingly used to analyze ECG (Strodthoff et al., 2020; Merdjanovska and Rashkovska, 2022; Rahman et al., 2022) data due to their ability to automatically identify spatial hierarchies and patterns, making them suitable for tasks like classification or anomaly detection. Recurrent neural networks (RNN), especially long short-term memory (LSTM) networks, are also used for monitoring physiological signals over time (Faust et al., 2018; Rim et al., 2020; Mao and Sejdić, 2022), which is crucial for applications like seizure or anomalous heart beat detection. Most recent advancements in DL, notably the use of attention mechanisms and transformers (Vaswani et al., 2017), have also allowed to achieve state-of-the-art performances on some tasks like detection of atrial fibrillation (Mousavi et al., 2019; Merdjanovska and Rashkovska, 2022). However, they have not yet become standard in physiological signal analysis. This is largely due to their reliance on substantial amounts of training data. To address this limitation, transfer learning techniques are increasingly being employed (Wan et al., 2021; Weimann and Conrad, 2021), enabling the adaptation of pre-trained models developed on large datasets to smaller, task-specific datasets.

In practice, while DL approaches have demonstrated efficiency in analyzing ECG signals – where distinctive patterns of the heartbeat are relatively easy to identify – the end-to-end analysis of more complex data like EEG signals presents greater challenges due to their inherently chaotic nature. It is difficult for DL models to directly learn from the raw signal, and instead operate on spectral density features, wavelet derived features, or Fourier feature maps for instance (Faust et al., 2018; Craik et al., 2019), resulting in hybrid models (Jafari et al., 2023). These models typically integrate the use of manual feature extraction, which is then followed by DL models for prediction. Many approaches Jaiswal et al. (2020); Ahmed et al. (2023) also inversely rely on DL for extracting features, which are then classified using traditional ML methods, effectively combining the strengths of both paradigms.

3.3 Baseline Models for Stress Identification

We implement several unimodal and multimodal baselines models for `StressID`. Following the general framework described in Section 3.2, all proposed models include a pre-processing phase, a feature extraction phase, and a classification phase. For all baselines, we have evaluated several combinations of feature selection algorithms and classifiers and selected the best-performing ones for our baseline results. For feature selection, we evaluated a Recursive Feature Elimination (RFE) algorithm, an L1 regularisation, and Principal-Component Analysis (PCA) for dimension reduction, as well as no feature selection. For the classification models, we have considered a large range of classical classifiers with different hyper-parameterizations. The exhaustive list is reported in Table 3.1.

We train the models to perform 2-class classification, i.e. binary discrimination between stressed and not stressed, as well as 3-class classification. In all the experiments, we generate 10 random splits, using 80% of the tasks for training, and 20% for testing for each split. The

Table 3.1: List of tested classifiers and corresponding grid search of hyper-parameters.

Model	Hyper-parameters	Grid search values
Support Vector Machines	kernel	'linear', 'rbf', 'sigmoid'
	C	0.1, 1.0, 10.0
	gamma	'scale', 'auto'
K-Nearest Neighbors	n_neighbors	3, 5, 10, 20
	weights	'uniform', 'distance'
	algorithm	'auto', 'ball_tree','kd_tree', 'brute'
Random Forests	n_estimators	100, 150, 200
	criterion	'gini','entropy'
	max_depth	3, 5, 7
	min_samples_split	2, 4, 6
	min_samples_leaf	1, 2, 3
	max_features	'auto', 'sqrt', 'log2'
	class_weight	None, 'balanced', 'balanced_subsample'
Multi Layer Perceptron	layer_depth	2,3,4
	layer_width	64, 128, 256
	activation	'logistic', 'tanh', 'relu'
	alpha	0.0001, 0.001, 0.01
	solver	'lbfgs', 'adam'
	learning_rate	'constant', 'invscaling', 'adaptive
	momentum	0.7, 0.8, 0.9
	early_stopping	True, False
Gradient Boosting Classifier	loss	'deviance', 'exponential
	n_estimators	100, 150, 200
	learning_rate	0.1, 0.5, 1.0
	max_depth	3, 5, 7
	min_samples_split	2, 4, 6
	max_features	'sqrt', 'log2

results are averaged over the 10 repetitions. To ensure robustness to potential imbalance resulting from the train-test splits, the results are assessed using the weighted F1-score and the balanced accuracy on the test data.

3.3.1 Unimodal Models

Physiological data. In line with the literature on stress recognition from physiological signals (Arsalan et al., 2022b; Gedam and Paul, 2021; Giannakakis et al., 2019), we propose a ML-based approach for the physiological baselines, including pre-processing of the signals, extraction of HC features, and classification using traditional ML algorithms. In a first step, the ECG, EDA, and respiration signals are filtered with Butterworth filters to reduce high-frequency noise and baseline wander. Precisely, we use a 0.5 Hz high-pass Butterworth

Table 3.2: Exhaustive list of physiological features extracted.

Domain	Features	Total
	ECG	
Time	MeanHR, minHR, maxHR, stdHR, modeHR, nNN, meanNN, SDSD, CVNN, SDNN, pNN50, pNN20, RMSSD, medianNN, q20NN, q80NN, minNN, maxNN, triHRV	19
Frequency	Total power of the signal, LF, HF, LF/HF, ULF, VLF, VHF, rLF, rHF, peakLF, peakHF	11
Non-linear	SD1, SD2, SD1SD2, ApEn, SampEn	5
	EDA	
Statistical	MinEDA, maxEDA, meanEDA, std, skeweness, kurtosis, median, dynamical range, minSCR, maxSCR, meanSCR, stdSCR, minSCL, maxSCL, stdSCL, slopeSCL	15
Time	nSCRpeaks, area under SCR, mean amplitude SCR (meanAmp), maxAmp, mean response SCR (meanResp), sumAmp, sumResp	8
	Respiration	
Time	MeanRR, minRR, maxRR, stdRR, nBB (breath to breath), meanBB, SDSD, SVNN, SDNN, RMSSD, minBB, maxBB, meanTT (trhough to through), SDTT, minTT, maxTT, meanBA (breath amplitude), SDBA, minBA, maxBA, meanBW (breath width), SDBW, minBW, maxBW	25
Frequency	Total power, LF, HF, VLF, VHF, LF/HF, rLF, rHF, peakLF, peakHF	10
Non-linear	SD1, SD2, SD1SD2, ApEn, SampEn	5

filter of order 5 for the ECG, a 5Hz low-pass Butterworth filter of order 4 for the EDA, and a 0.1-0.35 Hz bandpass Butterworth filter of order 2, followed by a constant detrending for the respiration signal. We use the `neurokit2` python package for all pre-processing. Then, 35 ECG features, 23 EDA features, and 40 respiration features are extracted. These include HRV features in the time domain including the number of R to R intervals (RR) per minute, the standard deviation of all NN intervals (SDNN), the percentage of successive RR intervals that differ by more than 20ms and 50ms (pNN20 and pNN50), or the root mean square of successive RR interval differences (RMSSD), as well as frequency-domain, and non-linear HRV measures. We have extracted statistical features of the Skin Conductance Level (SCL) and Skin Conductance Response (SCR) components of the EDA, including the slope and dynamic range of the SCL, along with time domain features including the number of SCR peaks per minute, the average amplitude of the peaks, and average duration of SCR responses. In addition, we have extracted Respiration Rate Variability (RRV) features in the time and frequency domains. Figure 3.2 illustrates an example of basic ECG, EDA, and respiration features visualized using `neurokit2`. An exhaustive list of the features used in our baselines is provided in Table 3.2. The resulting handcrafted (HC) features are then classified using RF classifiers with hyperparameters chosen by Cross-Validation (CV).

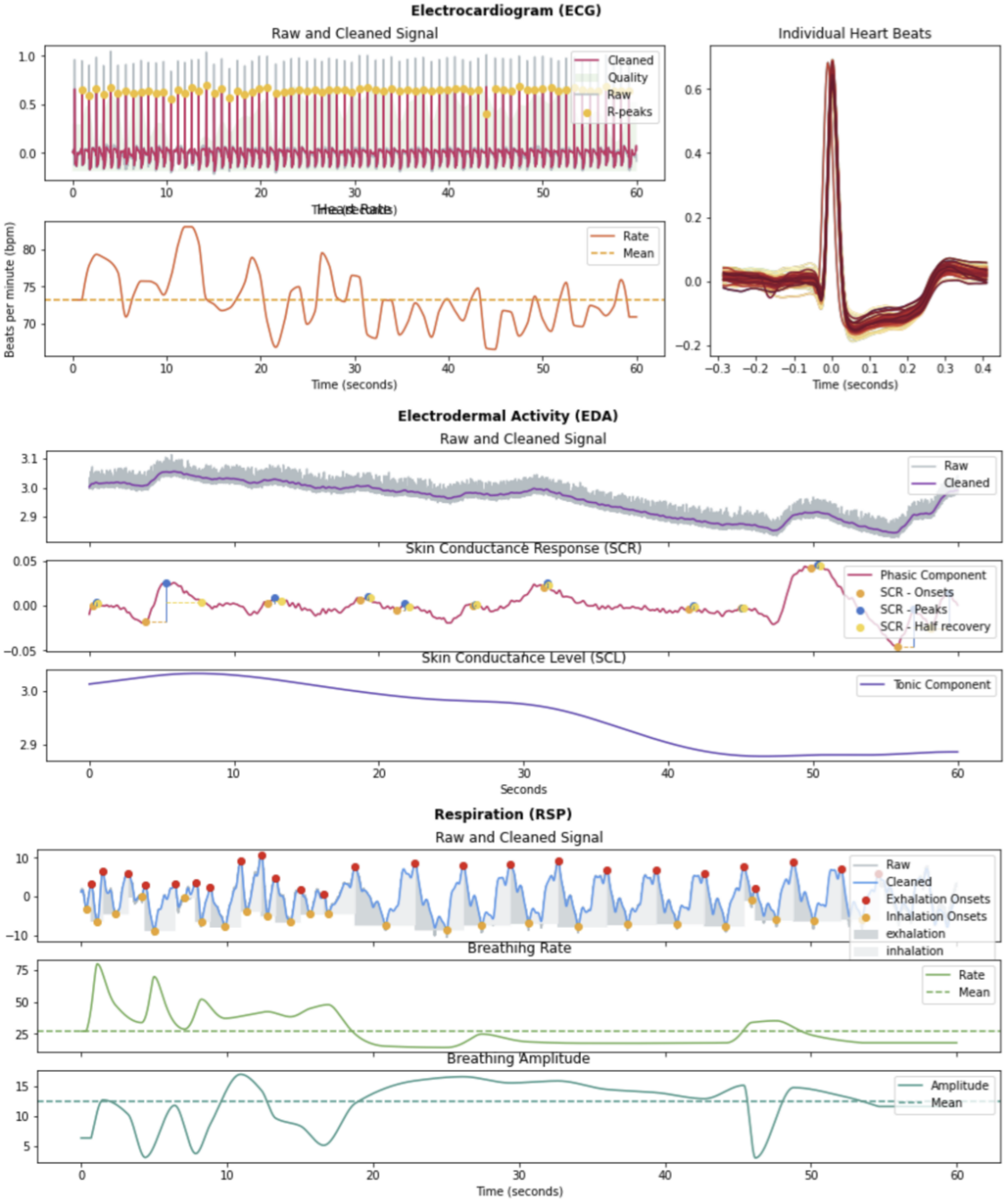
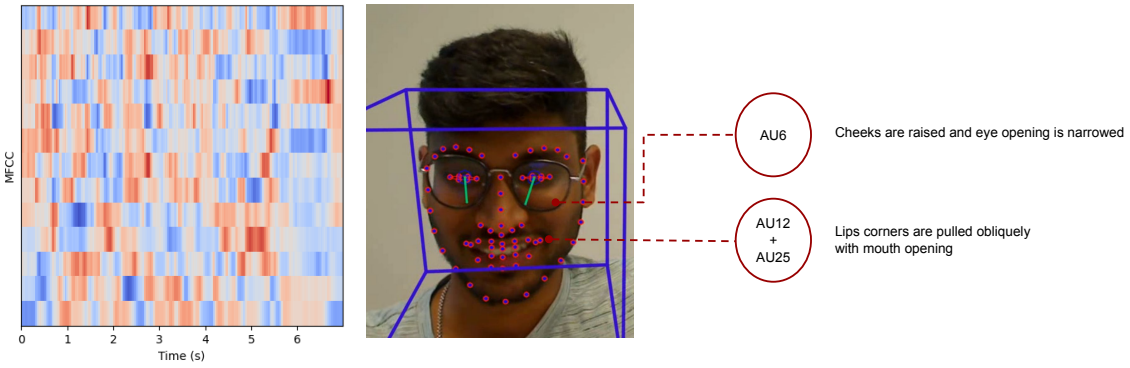


Figure 3.2: Example of features extracted from the ECG, EDA and respiration signals of `StressID`.



(a) Audio features

(b) Video features

Figure 3.3: Example of features extracted from `StressID`. (left) MFCCs features extracted from an audio extract. (right) Example of AUs extracted from a video extract.

Audio data. We propose two baselines for speech signals: the first employs HC features and ML algorithms, and the second one is a hybrid approach built on the Wav2Vec 2.0 (W2V) model (Schneider et al., 2019; Baevski et al., 2020). Both techniques involve downsampling from the original 32 kHz audio to 16 kHz, and the application of amplitude-based voice activity detection (VAD) (Kinnunen and Li, 2010) prior to feature extraction to eliminate non-speech segments. The first baseline relies on a plethora of specific audio features (Sahidullah et al., 2015; Allik et al., 2016) widely used in the literature on emotion recognition from speech (Ahmed et al., 2023; Arsalan et al., 2022b; Malla et al., 2020). These include Mel Frequency Cepstral Coefficients (MFCCs) and their first and second derivatives, which characterize the short-term power spectrum and its dynamics. Figure 3.3(a) shows an example of MFCCs extracted on a subject of `StressID`. The spectral centroid, bandwidth, contrast, flatness, and roll-off, which together provide a rich statistical representation of the spectral shape, are extracted. Harmonic and percussive components are also extracted, with tonal centroid features being computed for the harmonic component. The zero-crossing rate is a simple measure of the rate of sign changes; the rate of zero-crossings relates directly to the fundamental frequency of the speech signal. Last, we include tempogram ratio features (Peeters, 2005) which represent local rhythmic information. We compute the mean and standard deviation over time for all features, thereby resulting in feature vectors for each, which are then concatenated to form a comprehensive feature vector of 140 components, and used as input for ML algorithms. The handcrafted (HC) features are extracted using the `libROSA` python package (McFee et al., 2015), and the extraction is done using a GeForce RTX 3090 graphic card. The second baseline employs a large, pre-trained W2V model. The W2V 2.0 model produces features capturing a wealth of information relevant to diverse tasks including emotion recognition (Catania, 2023; Sharma, 2022; Chen and Rudnicky, 2023).

Features are extracted every 20 ms and averaged over time to obtain a single 513-component embedding per utterance, and are then classified using a linear classification layer optimized with Adam, cross-entropy loss, and an initial learning rate of 1e−3, until convergence.

Video data. We propose an hybrid approach for the video baseline employing Action Units (AU) and eye gaze for the classification of stress. AUs are commonly used as features in stress recognition applications (Giannakakis et al., 2020; Jaiswal et al., 2020; Aigrain et al., 2016). They are fine-grained facial muscle movements (Ekman and Friesen, 1978), each relating to a subset of extracted facial landmarks (Perveen and Mohan, 2020). Each AU is described in two ways: presence, if the AU is visible in the face, and intensity, indicating how intense the AU is on a 5-point scale (minimal to maximal). After downsampling the recordings to 5 frames per second, we use the OpenFace library (Baltrusaitis et al., 2018) to extract eye gaze and AUs from each video frame. We extract the following AUs: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45. As eye gaze features, we use two gaze direction vectors computed individually for each eye by detected pupil and eye location. Figure 3.3(b) shows an example of AUs extracted on a subject of `StressID`. The averages and standard deviations of each AU and eye gaze directions are computed across time frames. The feature extraction from 587 videos is done in 3 hours 42 minutes using two Dual CPU Intel Xeon E5-2630 v4 processors. The resulting 84-component vector is used as input to an MLP with 4 layers of width 256. In line with (Jaiswal et al., 2020), the number of layers and layer width of the MLP are chosen by CV in {2,3,4} and {64, 128, 256} respectively. We use ReLU activation and the MLP is trained for 100 epochs with cross-entropy loss optimized using Adam (Kingma and Ba, 2015) with an initial learning rate of 1e−3.

Each unimodal baseline is trained and tested on all available tasks of the corresponding modality. i.e. 692, 711, 587, and 385 tasks respectively for the EDA, ECG and respiration, video, and audio baseline. We remove from these analyses the damaged physiological modalities described in Section 2.4.1, and therefore consider them missing. The obtained performances for the unimodal baselines are reported in Table 3.3. Several conclusions can be drawn from the results. First, the performances highlight that overall the physiological modalities, especially the ECG and respiration signals, carry the most valuable information for the classification of 2-class stress. This suggests that unimodal physiological modalities are particularly useful for learning to reliably recognize a response to stress-inducing stimuli. Second, all the unimodal baselines achieve comparable results for the classification of 3-class stress, suggesting that the discrimination between positive and negative, or short-term and long-term stress is a more sensitive task. Nonetheless, it can be noted that the baselines on the ECG and audio modalities outperform other models in terms of accuracy. This suggests that physiological and audio modalities are more susceptible of carrying information allowing to discriminate between different states of stress, than video data.

Table 3.3: Performances (mean ± std) of unimodal baselines for the classification of stress. Each baseline is trained and tested on all available tasks of the corresponding modality.

		2-class		**3-class**	
	#tasks	F1-score (↑)	Accuracy (↑)	F1-score (↑)	Accuracy (↑)
Physiological modalities					
ECG	711	**73.2 ± 2.1**	**72.7 ± 2.9**	55.9 ± 2.9	**55.2 ± 2.8**
EDA	692	70.1 ± 3.5	70.2 ± 3.9	53.8 ± 2.8	54.1 ± 2.2
Respiration	711	72.8 ± 3.1	72.3 ± 3.1	54.9 ± 3.1	53.1 ± 3.1
Physical modalities					
Audio (HC + kNN)	385	67.9 ± 6.1	62.9 ± 4.5	53.1 ± 4.1	52.2 ± 3.9
Audio (W2V 2.0 + MLP)	385	70.1 ± 2.1	66.2 ± 2.9	**56.1 ± 3.8**	52.3 ± 3.9
Video	587	70.2 ± 3.6	70.2 ± 3.8	54.8 ± 2.6	54.6 ± 2.9

3.3.2 Multimodal baselines

Multimodal learning is an emerging field of machine learning combining modalities from various sources that depict a single subject from multiple views, and thus providing both shared and complementary information. It has shown considerable advantages in multiple domains (Baltrušaitis et al., 2018; Xu et al., 2023). While the existing literature is very rich and continuously expanding, three primary categories of multimodal models can be identified based on their modality fusion approach: *feature-level* fusion models, that combine low-level features from all modalities early at the input level and learns them together; *mid-level* fusion models, that learn modality-specific representations first, and fuse them later during learning to leverage both independent and combined information; and *decision-level* fusion models, that learn from modalities independently to generate separate outputs, which are then merged for a final decision.

Multimodal fusion strategies. In line with the state-of-the-art in stress identification, we propose fusion models combining all the extracted features using the most prominent fusion methods in the literature: *feature-level* and *decision-level* fusion (Ahmed et al., 2023; Middya et al., 2022). The differences between the unimodal approach and the two proposed multimodal approaches are illustrated in Figure 3.4. For feature-level fusion, unimodal HC features are combined into a single high-dimensional feature vector, used as input for learning algorithms. Similarly to (Jaratrotkamjorn and Choksuriwong, 2019; Chaparro et al., 2018), we evaluate feature-level fusion combined with ML classifiers. For decision-level fusion, following (Xu and Wang, 2018; Rao et al., 2019), we train independent SVMs for each modality using the HC features as input, and integrate the results of the individual classifiers at the decision level, i.e. the results are combined into a single decision using an average rule fusion.

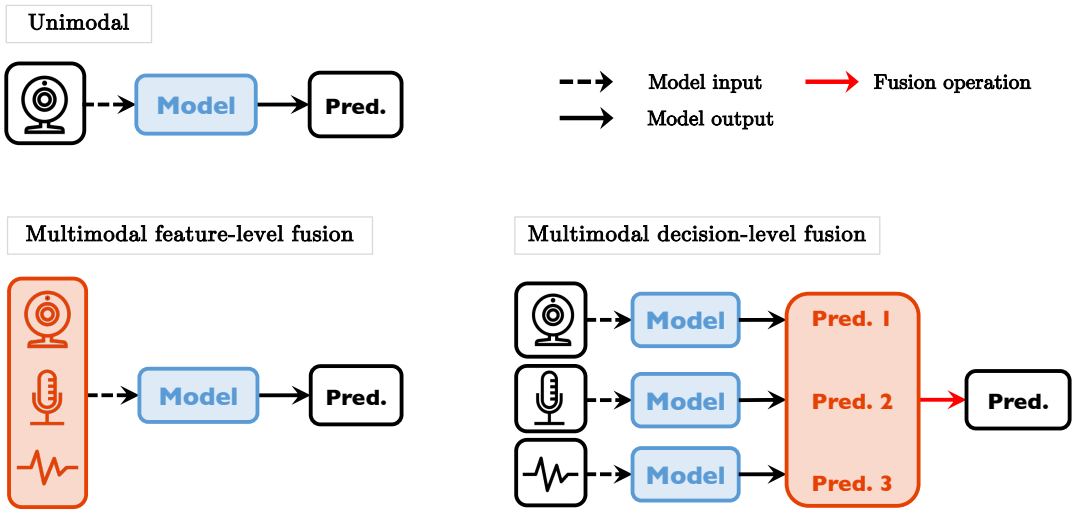


Figure 3.4: Comparison of the unimodal, feature-fusion multimodal and decision-level fusion multimodal baselines.

We propose to compare the performances of unimodal baselines with those of two categories of multimodal models: models combining the *3 physiological modalities* only, and models combining *all 5 modalities*, i.e. physiological audio, and data. The multimodal baselines of `StressID` are evaluated on the tasks that feature all modalities only, i.e. 355 tasks, to avoid learning with severely missing values. This subset of `StressID` is composed of talking tasks exclusively, i.e. all tasks without the audio modality are excluded. In this setting, the dataset presents a strong imbalance in the labels (70% stress). We use Minority Over-sampling Techniques (SMOTE) (Chawla et al., 2002) to balance the training set in each of the 10 repetitions, and leave the test sets untouched. The results for all multimodal baselines for the 2-class and 3-class classification are reported in Table 3.4. To ensure fairness in the comparison, the unimodal baselines are also trained on the modality-complete subset of 355 tasks.

Several observations can be made. First, it can be noted that most unimodal baselines computed on the modality-complete subset do not perform as well as the unimodal models computed on all available tasks (see Table 3.3). This can be due both to the smaller sample size used for training, and the label imbalance present in this new subset. Second, we can observe that the feature-level fusion multimodal model combining just the physiological modalities already significantly improves the performances of unimodal models trained on ECG, EDA and respiration signals independently. This highlights the benefits of multimodal learning, suggesting that the different modalities may carry complementary information useful to stress identification. Lastly, the results unsurprisingly show that combining all

Table 3.4: Performances (mean ± std) of multimodal baselines for the classification of stress, compared to unimodal models. All baselines are trained and tested only on tasks featuring all modalities, i.e. 355 tasks.

	2-class		**3-class**	
	F1-score (↑)	Accuracy (↑)	F1-score (↑)	Accuracy (↑)
Unimodal baselines				
ECG	65.9 ± 5.7	61.4 ± 5.3	52.2 ± 4.9	50.4 ± 6.2
EDA	61.1 ± 4.4	61.0 ± 3.2	47.1 ± 3.2	45.2 ± 4.2
Respiration	58.4 ± 2.5	54.8 ± 3.3	43.8 ± 3.2	44.8 ± 4.4
Video	67.7 ± 3.9	62.2 ± 4.2	58.3 ± 5.0	56.4 ± 4.3
Audio	67.7 ± 4.8	62.1 ± 4.2	56.2 ± 6.0	54.3 ± 6.1
Multimodal baselines (physio.)				
Feature fusion	**68.5 ± 5.1**	**63.1 ± 5.0**	**54.2 ± 4.9**	**51.8 ± 4.4**
Decision fusion	60.1 ± 4.7	54.5 ± 3.9	52.7 ± 5.1	49.3 ± 3.0
Multimodal baselines (all)				
Feature fusion	66.4 ± 4.3	61.2 ± 3.7	55.5 ± 6.2	51.4 ± 5.3
Decision fusion	**72.9 ± 4.8**	**65.2 ± 4.9**	**63.1 ± 5.1**	**58.6 ± 7.3**

modalities using decision-level fusion multimodal models considerably improves classification performances of all unimodal models separately. It also improves the 3-class classification results – which further highlights the benefits of combining multiple complementary sources of data, i.e. physiological and physical, for discriminating between different types of stress.

Remark 3.3.1. Both the unimodal and multimodal baselines highlight the difficulty of predicting the 3-class label defined in Chapter 2. This label was intended as an example, and further experiments suggest that focusing on other prediction tasks may be more appropriate. Therefore, to better showcase the potential of the `StressID` dataset and the proposed models to analyze it, we focus on a binary classification task for the remainder of this thesis.

3.4 Main Limitations

We have highlighted several limitations of `StressID` in Section 2.6. Namely, several modalities are missing for some participants across the dataset, and the dataset presents a strong imbalance in gender across the subjects. In this section, we discuss how these aspects can limit the robustness and reliability of state-of-the-algorithms algorithms used for stress identification.

3.4.1 Missing Data

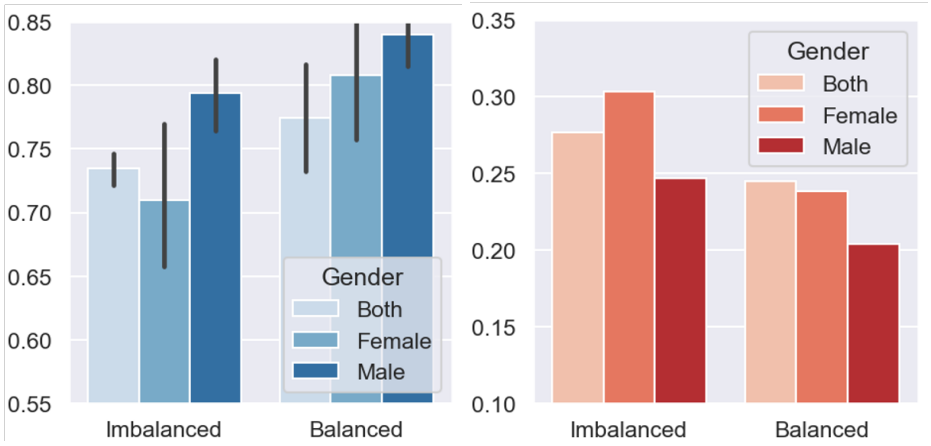
State-of-the-art algorithms for stress identification from physiological signals and multimodal inputs are predominantly hybrid or ML-based. Currently, these models are not designed to inherently handle missing data, requiring additional adaptations to effectively manage such scenarios. In our experiments, we have opted to use a subset of the `StressID` dataset where all modalities are complete for all participants. This decision was made primarily to ensure a fair and consistent comparison between unimodal and multimodal baselines by working with a uniform dataset. However, as demonstrated by the difference of performances reported in Table 3.3 and Table 3.4, this approach significantly reduces the efficiency of subsequent algorithms. Moreover, excluding samples due to missing modalities is not an ideal solution, as it prevents the use of all available data, which is counterproductive to the goal of maximizing information from multimodal inputs. More so, missing data itself can hold implicit information, and disregarding it entirely may introduce bias into the models, ultimately affecting their performance and reliability. This highlights the need for more robust methods that can handle incomplete data while taking advantage of the plurality of its sources.

3.4.2 Gender Imbalance

We analyse the effects of gender imbalance in training data on traditional learning algorithms used for stress identification from physiological data. We evaluate the `StressID` multimodal baselines combining all physiological modalities on two subsets of `StressID`. **Subset A:** we select the recordings from all 18 female participants, and randomly select 18 male participants – thus resulting in a subset composed of 36 subjects with a balanced ratio of female and male subjects. **Subset B:** we randomly select a subset of 36 subjects, preserving the female-to-male ratio of the original dataset, i.e. 9 female and 27 male subjects. We compare the performances of the `StressID` baselines on the 2-class classification task on the two subsets.

The results, averaged over 5 random repetitions, are reported in Figure 3.5. First, we observe that the baseline built on the subset A (balanced) outperforms the baseline using subset B. This suggests that more balanced datasets can improve the global performances of subsequent models, and thus already highlights potential bias induced by imbalanced representation in data. Second, we can observe that the difference in classification error rate between female and male participants is considerably decreased when working with balanced subsets. This is to be expected, as training on well-balanced data decreases the risk for a model to overfit – which in the case of gender imbalance can be translated as learning on male subjects mainly during the training phase and performing poorly during the testing phase on female subjects, less seen during training.

An important conclusion can be drawn from this experiment: a balanced dataset is crucial



(a) F1-score (b) Error rate

Figure 3.5: Comparison of performances between models trained on balanced and imbalanced subsets of `StressID`. F1-scores (left) and error rates (right) of the multimodal baseline combining all physiological modalities are reported.

for performing bias-free analyses and minimizing the risk of bias in algorithm development. Imbalances in gender, race, age, or background of the participants can limit the development of fair and equitable applications. Researchers need to be aware of this aspect and take the appropriate steps to build equitable systems before their use in real-life applications.

3.5 Discussion

We have established the state of the art in stress identification using physiological signals, video, and audio data. Building on the models identified in the literature, we have implemented a suite of unimodal and multimodal baselines for `StressID`, that we have publicly released, providing a valuable resource for researchers interested in working with the dataset.

Most current approaches to stress identification rely on ML-based and hybrid methods, combining feature extraction via DL with traditional ML algorithms for classification, or inversely, combining features extracted via handcrafted techniques, with DL classifiers. These methods offer several advantages: they leverage widely available feature extraction tools and libraries; they considerably reduce input data complexity; they enable the use of diverse ML models for the classification of tabular feature datasets; and their low-complexity make them particularly suited for real-time, closed-loop data processing necessary in wearable sensor applications. Furthermore, we have shown in Section 3.3.2 that even simple multimodal strategies, such as feature-level of tabular features, can significantly outperform unimodal models trained on individual modalities in the `StressID` dataset.

Nonetheless, we have identified several limitations in current state-of-the-art approaches. We demonstrated in Section 3.4.2 that models trained on real-world data are prone to bias, such as that caused by imbalanced gender representation in the dataset. This highlights the importance of researchers taking appropriate measures to ensure systems are reliable before deploying them in real-world applications. Equipping existing models to handle such issues could be done with solutions as simple as processing subpopulations of the dataset separately, or using conditional models. More critically, we highlighted in Section 3.4.1 that models intended for real-world use must be equipped to handle practical challenges, such as missing data, which can arise for many reasons, including simple sensor failures. Currently, most state-of-the-art models are not inherently designed to address this issue. This underscores the need to either develop innovative solutions or adapt existing models to make them robust to missing values. While adapting decision-level multimodal models to handle this problem can be relatively straightforward, the task becomes more challenging for feature-level fusion approaches or advanced multimodal models that rely on mid-level fusion – which has been shown to offer significant advantages in various domains (Baltrušaitis et al., 2018; Guarrasi et al., 2024; Strizhkova et al., 2025).

In the following chapters, we explore whether the extensive existing literature on handling missing values in tabular datasets can be leveraged to enhance the performance of state-of-the-art approaches for the binary classification of stress using multimodal data, including physiological signals. To do so, we first propose to evaluate the reliability and suitability of state-of-the-art techniques for addressing missing data within healthcare applications, in Chapter 4. Indeed, most of existing techniques heavily rely on imputation, which may be challenging to apply in sensitive domains like healthcare. Tackling this issue is a essential step towards developing both robust and reliable unimodal and multimodal models for analyzing wearable sensor data.

Chapter 4

How to Handle Missing Values in Healthcare Data?

Contents

4.1 Introduction 48
4.2 Problem Formulation, Notations and Definitions 49
4.3 State-of-the-Art 51
4.4 Assessing the Reliability of Existing Approaches within Healthcare Applications 54
4.4.1 Overview of the Methodology 55
4.4.2 Datasets Generation and Fingerprints Extraction 56
4.4.3 Benchmark and Evaluation Criteria 60
4.4.4 Analysis of the Results and Model Selection 65
4.4.5 Decision tree-based Approach for Model Choice 72
4.5 Guidelines for Handling Missing Values in Health Data 74
4.5.1 Main Takeaways: a Practical Guide with Flowcharts 74
4.5.2 Illustration on Healthcare Datasets 77
4.6 Application to StressID 79
4.7 Discussion 80

Abstract. In Chapter 3, we identified the handling of missing data as a significant challenge in developing robust and deployable models for analyzing wearable sensor data. We propose exploring whether the extensive existing literature on handling missing values in tabular datasets can be leveraged effectively to address this issue. Specifically, we aim to evaluate the reliability

of these techniques in healthcare applications, focusing on aspects such as imputation quality and impact on interpretability of subsequent models; and assess whether they can be applied to the `StressID` dataset.

4.1 Introduction

In this chapter, we investigate existing methods for handling missing values in supervised learning on tabular healthcare data. As discussed in Chapter 1, data loss is common in wearable sensors data and, more generally, missing values are prone to occur increasingly frequently as the size and complexity of real-world datasets increase. This phenomenon can be due to a variety of factors, such as, complex data collection processes, the aggregation of multiple data sources, sensors failures, or refusals to answers questions in surveys. Datasets with missing values make supervised learning a challenging task since traditional algorithms cannot be applied directly to incomplete data. In Chapter 3, we have adopted the easiest and cheapest solution to circumvent this problem – consisting of removing instances or features of the dataset containing missing values. However, this can be problematic in many domains ranging from clinical and medical studies, to finance and economics (Kang, 2013). Many current state-of-the-art solutions involve imputing the missing data, i.e. replacing the unavailable values in a dataset with substituted values. These techniques range from simple imputation by the mean, to using more elaborate generative models, before training a supervised learning model. Yet, this approach can be highly limiting when applied to real scenarios, as it may accentuate the biases present in already unrepresentative data and can have a major impact on the interpretability of the models developed. Therefore, such solutions can pose reliability issues in sensitive applications such as healthcare. Works like Perez-Lebel et al. (2022) have benchmarked existing methods on healthcare data. Yet, they have focused on prediction performances, and have not addressed aspects such as quality of imputation and impact on the interpretability of subsequent supervised learning models. While Shadbahr et al. (2023) have analyzed the importance of measuring imputation quality, their study is conducted on a limited amount of data, and no clear directives on how to appropriately choose a model are put forward.

For these reasons, we propose to conduct a comprehensive evaluation of existing methods for handling missing data and assess their reliability within healthcare applications. Specifically, we have developed a framework tailored to evaluating these methods based on the specific needs of such applications. In particular, we: (1) review existing methods designed for missing values and identify 3 main categories among them; (2) evaluate the performances of these models in many diverse settings of missing data, that can occur in real-life scenarios and are often overlooked in research; (3) analyze the quality and reliability of imputations produced by these models, and study how data distributions can be impacted by imputation; (4) investigate how imputation alters feature interaction in datasets; (5) evaluate the interpretability of these models; (6) investigate how the characteristics of a dataset can

impact the performances of these different methods; (7) propose a *tree-based* approach to help understand how to chose a model, given a dataset. This comprehensive analysis enables us to identify the strengths and limitations of existing approaches, in order to ultimately derive a set of guidelines to properly and responsibly handle missing values in healthcare applications. To ensure a robust and reliable study, we aim to encompass a wide range of models, and evaluate them on multiple datasets drawn from diverse real-world scenarios.

The remainder of this chapter is organized as follows. We first formally introduce the problem of supervised learning with missing values in tabular datasets and provide useful background in Section 4.2. We provide an overview of the existing approaches for handling missing values in Section 4.3. We describe the methodology we propose to assess their reliability in healthcare applications in Section 4.4, and illustrate our finding on several real-life datasets. We then translate our study into a set of guidelines for handling missing values in healthcare in Section 4.5. Lastly, we evaluate whether state-of-the-art methods are suited for `StressID` in Section 4.6, and we summarize our work and future directions.

4.2 Problem Formulation, Notations and Definitions

Many studies have addressed the issue of handling missing data in inferential frameworks (i.e. distribution modeling and parameters estimation) (Rubin, 1976; Dempster et al., 1977; Little, 1992; Little and Rubin, 2019; Nazábal et al., 2018; Ma et al., 2018b; Mattei and Frellsen, 2019; Collier et al., 2020; Gong et al., 2021). However, the problem of supervised learning with missing values is distinct from distribution estimation, and fewer studies have focused on the prediction of a target variable in presence of missing values. Before exploring existing resources on the problem, we provide a formal framework for supervised learning with missing values in this section. We briefly remind the classical setting for supervised learning, and introduce useful notations and definitions to formulate the problem of handling missing data in this scenario. Throughout this chapter, capital letters refer to random variables while lower-case letters denote realizations.

Supervised learning with missing values. Let us consider random independent input and output pairs (X, Y) drawn from a distribution P, where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. The goal in supervised learning is to predict Y given X. Formally, this corresponds to finding a function $f : \mathbb{R}^d \to \mathbb{R}$ that minimizes $\mathbb{E}[\ell(f(X), Y)]$ given a loss function $\ell : \mathbb{R} \times \mathbb{R}$. The optimal prediction function f^* is given by:

$$f^* \in \underset{f:\mathbb{R}^d \to \mathbb{R}}{\arg\min}\, \mathbb{E}[\ell(f(X), Y)]. \tag{4.1}$$

In practice, given a dataset $\mathcal{D}$ of $n \in \mathbb{N}$ training examples $\mathcal{D} := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, a learning process is used to estimate an approximation $\hat{f}_n$ of f^*. Given that the learning

process operates on a finite number of samples, and not the distribution P, it optimizes the empirical risk $\sum_i \ell(f(X_i), Y_i)$ rather than the expected risk defined in Eq. 4.1.

Ultimately, a learning process can be defined as the following optimization problem:

$$\hat{f}_n \in \underset{f:\mathbb{R}^d \to \mathbb{R}}{\arg\min} \left(\frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \right). \tag{4.2}$$

However, in presence of missing values, we do not observe the complete vector X. Therefore, it is necessary to define precisely the data in this new setting.

Let us introduce the indicator vector $M \in \{0,1\}^d$ to denote the positions of missing values in X such that $M_j = 1$ if and only if X_j is missing (where j denotes the j-th element of one observation X_i in $\mathcal{D}$). For realizations m of M, we denote by $obs(m)$ the indices of the observed variables of X, and by $X_{obs(m)}$ the vector of observed elements of X, such that $X_{obs(m)_j} = \texttt{na}$ if $m_j = 1$. The observed data $X_{obs(m)}$ can be written as

$$X_{obs(m)} = (1 - M) \odot X + M \odot \texttt{na}, \tag{4.3}$$

where $\odot$ is the term-by-term product, with the convention that, for all one-dimensional x, $\texttt{na} \odot x = \texttt{na}$. As a result, for a given realization $x = (1, 7, -2.9, 9, -2.5)$ and its associated missing indicator $m = (0, 1, 0, 0, 1)$, the observed data is

$$x_{obs(m)} = (1, \texttt{na}, -2.9, 9, \texttt{na}).$$

In this setting, the learning goal becomes the prediction of Y given $X_{obs(m)}$. In particular, the new objective becomes the optimization of the empirical risk over the set of measurable functions that map observed $x_{obs(m)}$ realizations to y, such that:

$$\hat{f}_n \in \underset{f:\mathbb{R}^d \to \mathbb{R}}{\arg\min} \left(\frac{1}{n} \sum_{i=1}^{n} \ell(f(X_{obs(m)_i}), Y_i) \right). \tag{4.4}$$

In practice however, the mixed nature of $X_{obs(m)}$ (see Eq. 4.3) makes supervised learning with missing values challenging. It prevents the straightforward use of traditional ML algorithms that require fixed dimensional vectors as input. As a result, the optimization problem described in Eq. 4.4 is hard to solve, and additional steps need to be taken to use standard learning algorithms.

Missing values mechanisms. In his pioneer work, Rubin (1976) introduces different data scenarios leading to three missing data mechanisms that can be defined in terms of the probability distribution of $M|X$. He distinguishes three missing values mechanisms based on the relationships between the observed variables and the missing patterns: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR).

Definition 4.2.1. (Missing Completely At Random) The data is MCAR if the probability of a value being missing is independent from the data, such that

$$\mathbb{P}(M = m|X) = \mathbb{P}(M) \quad \forall\, m \in \mathbf{M}$$

Definition 4.2.2. (Missing At Random) The data is MAR if the probability of a value being missing depends on the observed variables, such that

$$\mathbb{P}(M = m|X) = \mathbb{P}(M|X_{obs(m)}) \quad \forall\, m \in \mathbf{M}.$$

Definition 4.2.3. (Missing Non At Random) The data is MNAR if the probability of a value being missing depends on the full vector X, i.e. both observed variables and missing variables themselves.

To illustrate those definitions, let us imagine a data set of customer records with a variable *age* that is missing for some of them. The data is MCAR if the missing values occur completely at random, with no relationship to the variables (e.g. due to a technical error in data collection). If the missing values are related to another observed variable of the data such as the *gender* for example, the data is MAR (e.g. because male customers are less likely to provide their age). If the missing information is related to the variable *age* itself, the data is MNAR (e.g. because older customers may be less likely to disclose their age).

While statistical analysis with missing values has been abundantly studied under the MAR assumption, or the more restrictive MCAR assumption (Emmanuel et al., 2021; Lin and Tsai, 2020; Little and Rubin, 2019; Mayer et al., 2019; Josse and Reiter, 2018; Van Buuren, 2018), fewer works study the harder to address MNAR mechanism (Chen et al., 2018; Yang et al., 2018; Ipsen et al., 2021).

4.3 State-of-the-Art

The challenges of supervised learning in the presence of missing values are different from those of inference and imputation, as the principal goal becomes to minimize a prediction error given incomplete data. While numerous works have been developed for missing values imputation (Van Buuren and Groothuis-Oudshoorn, 2011; Stekhoven and Bühlmann, 2012; Gondara and Wang, 2018; Mattei and Frellsen, 2019) and distribution estimation in presence of missing values (Rubin, 1976; Dempster et al., 1977; Little, 1992; Robins et al., 1994;

Jones, 1996; Nazábal et al., 2018; Ma et al., 2018a,b; Mattei and Frellsen, 2019; Collier et al., 2020; Gong et al., 2021), fewer studies have focused on the prediction of a target variable in the presence of missing values in both the training and the test sets (Josse et al., 2019; Le Morvan et al., 2021).

The easiest and cheapest solution to circumvent the limitation introduced in Section 4.2 is list-wise deletion, consisting in removing incomplete samples for the training data. However this approach is suboptimal: it is highly limiting in applications like healthcare where training data is often scarce; outside of the MCAR scenario, the presence of missing values in itself can be meaningful, as such deletion can lead to a considerable loss of information; lastly, while this approach can allow the training of supervised models on complete data, it is not robust to missing values in testing data, and therefore not adapted to trustworthy real-life applications. Therefore, learning systems should be specifically equipped to handle missing values. We classify the existing solutions in the literature that address this problem intro three categories: *impute-then-regress* methods; *impute-and-regress* methods; and *imputation-free* methods.

Impute-then-regress. The most popular solution, as its name suggests, consists of first learning an imputation model, using it to fill in the missing values of a dataset, before fitting a supervised model on the imputed dataset (Bertsimas et al., 2021; Le Morvan et al., 2021). A considerable advantage of this approach lies in that it can be used to adapt existing algorithms and learning pipelines to the presence of missing values. Josse et al. (2019); Le Morvan et al. (2021) have shown that good predictions can be obtained using strategies as simple as imputing by the mean of the observed features, given a powerful enough prediction model. Impute-then-regress strategies can also take advantage of the many more elaborate imputation models available in the literature. Among these, MICE (Van Buuren and Groothuis-Oudshoorn, 2011) and MissForests (Stekhoven and Bühlmann, 2012), are well-known solutions commonly used in both research and practice. They iteratively impute each variable with missing values by learning to model them conditionally on the other variables in the dataset. Moreover, these methods can be used in a multiple imputation set-up (MI), consisting in generating multiple possible values for the missing values and combining the results to incorporate uncertainty. Another popular solution is KNN imputation (Troyanskaya et al., 2001), consisting in imputing each missing value as the average or most frequent value (for continuous and categorical data respectively) among the non-missing values of the k most similar neighbours. More recent methods based on DL approaches such as denoising autoencoders (Gondara and Wang, 2018), deep latent variable models (Mattei and Frellsen, 2019), or generative adversial networs Yoon et al. (2018) have reached state-of-the-art performances in terms of quality of the imputation in missing at random (MAR) or not at random (MNAR) settings (Ipsen et al., 2021). However, Josse et al. (2019); Le Morvan et al. (2021) have highlighted that optimal imputations do not necessarily lead to optimal prediction performances.

Impute-and-regress. Le Morvan et al. (2021) have shown that the difficulty of the prediction task strongly depends on the choice of imputation strategy. They suggest that the imputer and predictor should be adapted to one another, which can be difficult to ensure in practice, as choosing the right imputation function can be time-consuming and computationally costly. To circumvent this problem, Le Morvan et al. (2021) have suggested jointly learning imputation and prediction instead. They have proposed to use NeuMiss networks (Le Morvan et al., 2020a) that handle missing values using multiplication by the indicators M as nonlinearities to capture the conditional links across observed and unobserved variables, for imputing missing values; and a multilayer perceptron (MLP) for prediction. Building on this idea, Ipsen et al. (2022) have proposed a joint model for imputation and supervised learning in MAR settings by marginalizing over missing covariates and mimicking multiple imputation. Joint learning approaches, however, can be impractical in real-life, as they rely on complex optimization processes or require large numbers of samples to reach good performances ($n > 1e5$ in Le Morvan et al. (2021)).

Imputation-free. Another branch of solutions consists in end-to-end methods that do not rely on imputation, and instead predict from observed variables only. A first solution is learning independent estimators for each missing-values pattern, but as the number of possible patterns grows with the number of features d, in practice 2^d sub-models are required to fit the Bayes-consistent predictors (Le Morvan et al., 2020b). Ayme et al. (2022) have proposed a thresholded pattern-by-pattern linear estimator to palliate this limitation. However, it is designed in the specific context of linear models, and adapts poorly to classification problems or more complex datasets. Another solution is to directly use expectation maximization (EM) to compute maximum likelihood estimates from an incomplete dataset (Dempster et al., 1977). However, such approaches rely on strong assumptions on the missing values. Alternatively, the discrete nature of decision trees allows them to handle the mixed nature of $X_{obs(m)}$, and thus missing data, natively (Friedman, 2001; Kapelner and Bleich, 2015; Jeong et al., 2022). These approaches can operate by using surrogate splits when a missing value is encountered, or more commonly, by incorporating the missing data in attributes (MIA) (Twala et al., 2008). In particular, MIA methods are a good choice when the presence of missing values in a dataset is informative, as the splitting criteria are computed with respect to whether a feature is missing or not. MIA approaches represent a good alternative to imputation-based methods, they work well in diverse scenarios, and have shown great performances in multiple comparative studies (Josse et al., 2019; Perez-Lebel et al., 2022; Ipsen et al., 2022). Nonetheless, they can suffer from slow convergence (Josse et al., 2019), or fail to generalize when new missing patterns are introduced in the test samples.

4.4 Assessing the Reliability of Existing Approaches within Healthcare Applications

Several works have addressed the problem of supervised learning from medical datasets with missing values (Perez-Lebel et al., 2022; Nijman et al., 2022). Some have studied the impact of imputation on downstream classification performances (Campos et al., 2015; Jäger et al., 2021). However, less attention has been given to assessing whether the imputed data actually reflects the underlying features distributions accurately. Yet, Van Buuren (2018) have highlighted that optimal scores on metrics such as the mean square error (MSE) or the mean absolute error (MAE) – commonly used in studies to assess the quality of an imputation model by comparing the imputed values with the ground truth, can be achieved even when the distributions of imputed data are far from the true distributions. Building on this observation, Shadbahr et al. (2023) have explored and found that these metrics are actually uncorrelated from downstream classification performances, whereas more elaborate metrics that quantify distributional discrepancies are. All the more, the authors have shown that even a classifier built on data with a poor imputation quality can reach satisfactory performances. Similarly, Josse et al. (2019); Le Morvan et al. (2021) have demonstrated that good prediction performances on certain tasks can be achieved even with suboptimal imputations. However, this can be problematic in sensitive domains such as healthcare, where the reliability of the imputed data is of paramount importance. Poorly imputed data can introduce significant bias in data distributions, and create spurious relations in the data. This in turn can compromise the interpretability of subsequent prediction models: it can lead to assigning spurious importance to particular features and thus, to incorrect conclusions about the impact of a feature on an outcome. Perez-Lebel et al. (2022) have found in their evaluation of multiple state-of-the-art approaches that not only features with few missing values have important impacts in the outcomes of models, but features with high levels of missing values too. This can be problematic in medical applications, where it leads to interpretations relying on values that were not genuinely recorded. Therefore, in healthcare, the quality and reliability of imputations and interpretability are crucial factors to consider to ensure the deployment of trustworthy models.

We have developed a framework tailored to assessing the reliability of existing methods for handling missing data based on the specific needs of healthcare applications. Precisely, we systematically and carefully address the following open questions: (1) Are state-of-the-art models able to generate reliable imputations (when applicable) that reflect the true underlying distribution of the data? (2) Are these models able to capture and preserve the interactions between features found in the original dataset? (3) Are the interpretability mechanisms of these models affected by the quality of imputation, i.e. how reliable are the interpretations that these models provide? (4) Do the characteristics of a dataset impact the reliability of the different state-of-the-art methods? To answer these questions, we introduce a novel framework, illustrated in Figure 4.1.

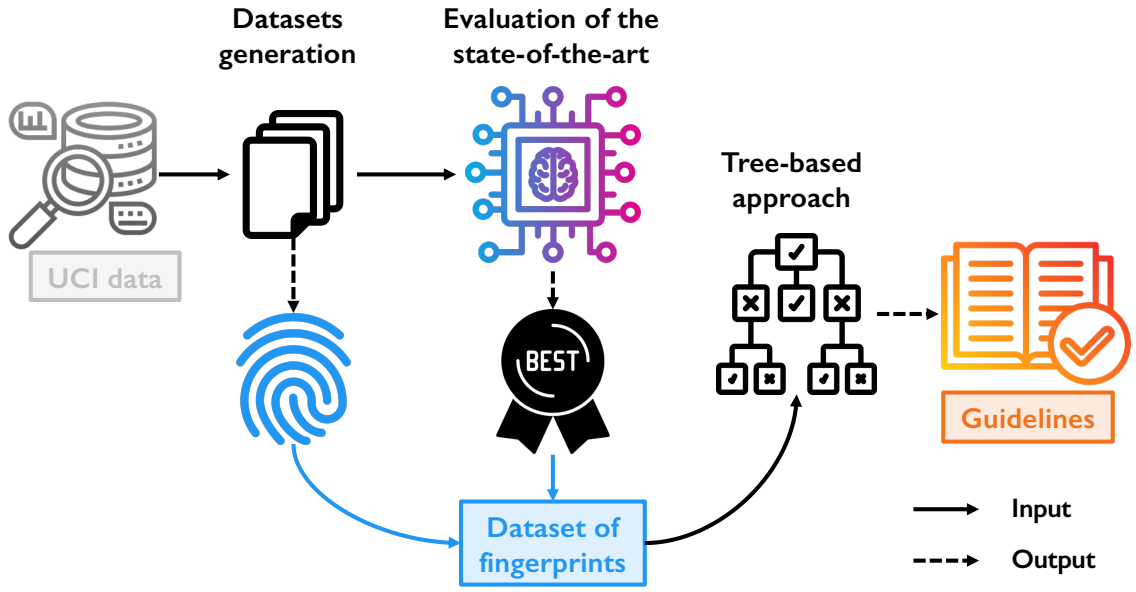


Figure 4.1: Overview of our methodology. First, we generate 384 datasets by introducing different scenarios of missing values in 14 complete datasets from the UCI repository. We then evaluate 5 models on the 384 dataset, using 10 different criteria, and determine the best choice for each. Lastly, we train a decision-tree to analyze the results and help us derive a set of guidelines to choose a model given the fingerprints (i.e. characteristics) of a dataset.

4.4.1 Overview of the Methodology

In a first step, we select 14 publicly available health-related datasets from the UCI repository (Dua et al., 2017) that do not contain any missing data. They are chosen to be as diverse as possible in terms of data content, size, and prediction tasks. For each dataset, we generate multiple versions where we artificially introduce missing values using various mechanisms and amounts. As a result of this step, we obtain 384 different databases.

In a second step, we evaluate 5 state-of-the-art models taken from each category of approaches (see Section 4.3) on the 384 datasets. In addition to prediction performances, we measure 10 criteria to assess their reliability. In particular, we compare the feature distributions of datasets imputed with each model to those of the complete data (14 original UCI datasets). Similarly, we compare the features interactions between the complete and imputed datasets to check for the presence of new spurious relations. Lastly, we compare the feature rankings of models trained on complete datasets with those of models trained on imputed ones.

In a last step, we investigate how the characteristics of a dataset can impact the reliability of different methods. We propose a *decision tree-based* approach to analyze the results of our

benchmark. We extract *fingerprints* (i.e. data characteristics) from each of the 384 datasets evaluated; select the *best model* based on a combination of the 10 evaluation criteria; and use this information as input for tree models to extract sets of *rules*. We ultimately use these rules to help us derive a set of guidelines on how to choose the most appropriate model, given a specific dataset.

4.4.2 Datasets Generation and Fingerprints Extraction

4.4.2.1 UCI Datasets

To evaluate the performances of state-of-the-art models focusing on more than prediction performances, we select multiple datasets from the UCI repository Dua et al. (2017) with varied characteristics, both numerical and categorical variables, and that contain no missing values. We consider the 14 following datasets:

1. **CDC Diabetes Health Indicators:** that contains 21 features that correspond to healthcare statistics and lifestyle survey information of 70,712* patients along with their diagnosis of diabetes. The prediction task on this dataset is binary classification.
2. **National Health and Nutrition Health Survey**: that contains 7 features collected from interviews, physical examinations, and laboratory tests for 748* individuals. The prediction task is the binary classification of the age group of the participants.
3. **AIDS Clinical Trials Group Study:** that contains 23 features representing healthcare statistics and categorical information about 2139 patients who have been diagnosed with AIDS. The prediction task on this dataset is the binary classification of death of the patient within a certain window of time.
4. **Glioma Grading Clinicial and Mutation Features:** that contains most frequently mutated 20 genes and 3 clinical features for 839 patients. The task on this dataset is the binary classification of glioma grade.
5. **National Poll on Healthy Aging:** that contains 14 features giving insights on the health, sleep issues affecting 714 Americans aged 50 and older. The prediction task on this dataset is 3-class classification of the frequency of doctor visits of the participants.
6. **Differentiated Thyroid Cancer Recurrence:** that contains 13 clinicopathologic features for 383 patients. The prediction task is binary classification of recurrence of well differentiated thyroid cancer.
7. **Estimation of Obesity Levels:** that contains 16 features corresponding to eating habits and physical condition for 2111 individuals. The prediction task is 7-class classification of the obesity levels of the patients.
8. **Heat Failure Clinical Records:** that contains 12 clinical features of 299 patients

who had heart failure, collected during their follow-up period. The prediction task on this dataset is the binary classification of death of the patient.

9. **Parkinson's Telemonitoring:** that contains 19 features corresponding a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease, measured during a six-month period of telemonitoring. The dataset contains 5875 distinct samples. We define two regression tasks: prediction of motor UPDRS score, and total UPDRS score.
10. **Infrared Thermography Temperature:** that contains 33 features corresponding to gender, age, ethnicity, ambiant temperature, humidity, distance, and various temperature readings from the thermal images for 1020 individuals. The prediction task on this dataset a regression to predict the oral temperature of each patient using other features.
11. **EEG Eye State:** that contains 14,980 entries of 14 features extracted from one continuous EEG measurement. The prediction task on this dataset is binary classification of the eye state (i.e. open or closed).
12. **Minimal Sepsis Records:** that contains 16,278* admissions of patients diagnosed with diagnosed with infections, systemic inflammatory response syndrome, sepsis by causative microbes, or septic shock. The prediction task is binary classification of whether a patient survived in the 9 days after their admission, using only 3 features.
13. **Hepatitis C Virus:** that contains 28 features corresponding to demographics, and laboratory results of 1385 patients who underwent treatment dosages for Hepatitis C virus. The prediction task is multiclass classification of liver fibrosis.
14. **Thoracic Surgery Data:** that contains 16 features collected during post-operative clinical examinations of 470 patients with lung cancer. The prediction task on this dataset is binary classification of survival.

Most ML models are not inherently robust to imbalanced datasets, and most of the time learning on imbalanced data leads to failure: despite a high accuracy score, models simply predict the majority class at inference time. In this analysis, we have under-sampled (Pereira and Saraiva, 2020) the datasets that present a strong class imbalance in the labels. They are denoted by a star (*) above. Their reported numbers of observations correspond to the downsampled sizes.

4.4.2.2 Missing Values Generation

For each of the 14 UCI datasets, we generate several versions where we artificially introduce missing data, resulting in 384 augmented variants. The missing data is generated using different missing values mechanisms (see Section 4.2). We consider the following settings, as described in Mayer et al. (2019):

- **MCAR.** The variables X and M are independent. The pattern M is generated according to a homogeneous Bernoulli distribution, such that $M \sim \mathcal{B}(p)$, where p is the probability for a value to be missing, and p is uniform across the dimensions of X.
- **MAR.** The missing values pattern M depends on the observed values. The data X is separated into two randomly selected subsets $X^{(1)}$ and $X^{(2)}$, such that $X^{(1)}$ is always fully observed. The missing pattern M associated to $X^{(2)}$ is generated according to a logistic model parametrised by β, such that $\mathbb{P}(M = 1|X^{(1)}) = \sigma(\beta X^{(1)})$, where $\sigma(.)$ is the sigmoid function, and the weights vector β is drawn randomly and re-scaled to attain the desired proportion of missing values p on $X^{(2)}$.
- **MNAR logistic.** The missing values pattern M depends on both the observed values and the missing values. The missing values probabilities are computed according to a logistic model parametrized by random weights β, taking all variables as inputs, such that $\mathbb{P}(M = 1|X) = \sigma(\beta X)$. That way, values that are inputs to the logistic model can also be missing.
- **MNAR self-masked.** Whether a variable X_j is missing or not only depends on X_j itself, hence the denomination of self-masking. The missing values pattern M is generated according to a self-masking logistic model parametrized by random weights β such that $\mathbb{P}(M = 1|X_j) = \sigma(\beta X_j)$. In other terms, a variable X_j have missing values probabilities given by a logistic model taking the same variable X_j as input.
- **MNAR with quantile censorship.** The missing values are generated on the q-quantiles. A subset $X^{(1)}$ of variables which will have missing values is randomly selected. Then, missing values are generated on the q-quantiles of $X^{(1)}$ only, such that $M \sim \mathcal{B}(p)$. As such, whether a variable has missing values depends on quantile information, that is masked. Additionally, 10% missing values generated completely at random are added on top.

For each UCI dataset, multiple scenarios under *each* of the mechanisms described above are generated: we randomly select the number of variables on which to introduce missing values, as well as the amount of missing values to generate – resulting in 384 different versions in total. Lastly, on each generated dataset, we make sure that no column or row is left without any values. Our implementation is based on the code provided by Muzellec et al. (2020).

4.4.2.3 Fingerprints Extraction

In order to analyze how the characteristics of a dataset impact the performances of state-of-the-art models, we extract a set of 15 descriptive characteristics from the 384 datasets, that we call *fingerprints*. In particular, we extract characteristics related to the size and content type of the dataset. Namely: the number of observations in the dataset n; the number of dimension in the dataset d; the sample-to-dimension ratio n/d; the proportion of categorical variables in the dataset d_c/d; and the proportion of numerical variables in the dataset d_n/d.

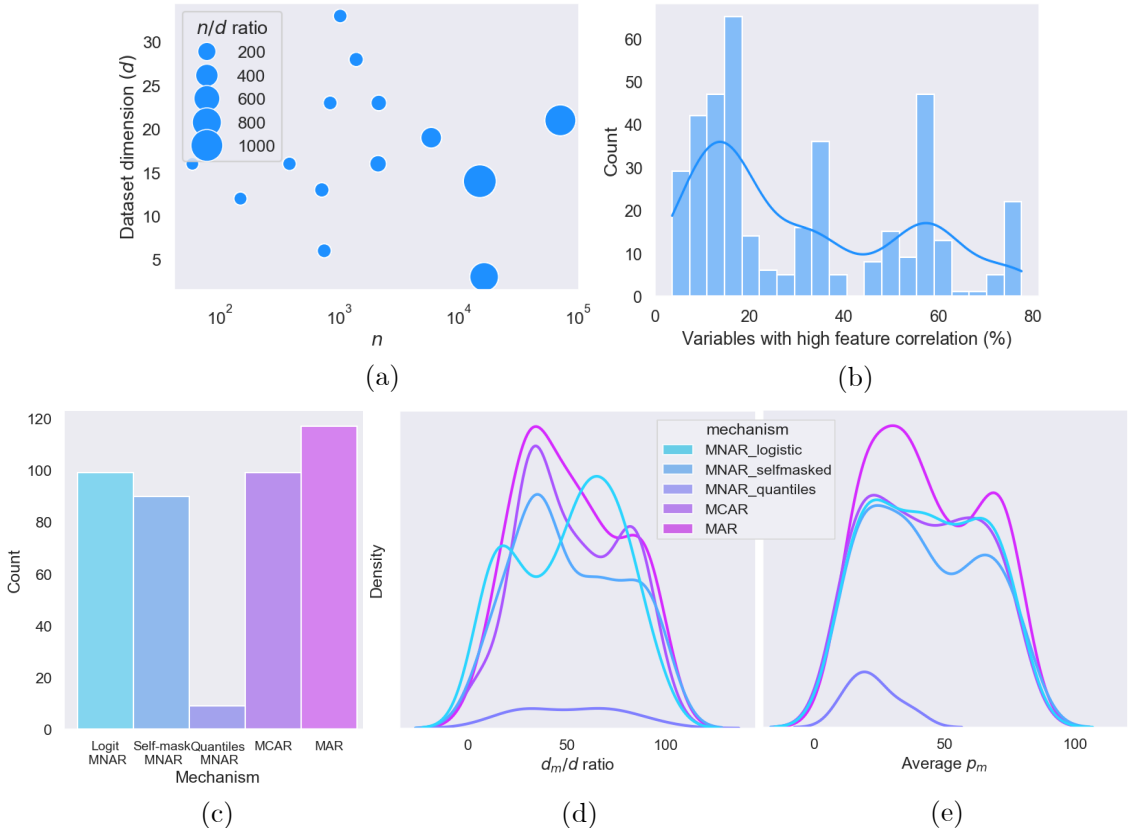


Figure 4.2: Visualization of the characteristics of the 384 augmented datasets in our collection. Dataset sizes (a) are reported, along with the distributions of $corr_{30}$ (b); the missing values mechanisms (c); d_m (d); and mean p_m (e) across the collection.

In addition, we extract measures describing the missing values in the dataset: the proportion of variables with missing values in the dataset d_m/d; the proportion of categorical variables with missing values among the total number of variables with missing values d_{c_m}/d_m; the proportion of numerical variables with missing values among the total number of variables with missing values d_{n_m}/d_m; the global proportion of missing values in the whole dataset p; the average proportion of missing values in the variables with missing entries *mean* p_m; the minimal proportion of missing values in the variables with missing entries *min* p_m; the maximal proportion of missing values in the variables with missing entries *max* p_m; the total number of distinct patterns of missing values across the dataset; and the mechanism used to generate the missing values. Lastly, we extract $corr_{30}$, corresponding to the average number of variables correlated to other variables with an absolute correlation coefficient larger than 30% (as described in Perez-Lebel et al. (2022).

The fingerprints of the 384 augmented datasets are visualized in Figure 4.2. Overall, the datasets obtained for our study are varied in terms of size and content. In addition, our collection comprises datasets with both very low and very high correlation between features. The amounts of missing values across the collection are also diverse, allowing us to encompass a large range of different scenarios in our study. Lastly, while the distribution of the datasets in MAR, MCAR, slef-masked MNAR and logistic MNAR is rather uniform, we have generated less datasets in the MNAR with quantile censorship setting for two reasons: due to the difficulty of the task; and to ensure realistic scenarios in our collection.

4.4.3 Benchmark and Evaluation Criteria

We use the 384 generated datasets to benchmark existing state-of-the-art approaches to handle missing values in supervised learning. In addition to prediction performances, we measure 8 additional criteria to assess their reliability within healthcare applications.

4.4.3.1 Methods

We evaluate 5 different state-of-the-art models taken from each category of approaches identified in Section 4.3. In particular, we evaluate the following methods.

Mean impute-then-regress. The missing values are imputed by the mean of each feature. Predictions are made using gradient boosted trees, taking as input the imputed data concatenated with the missing patterns mask M. While imputation by the mean is frowned upon in statistical practice, Josse et al. (2019); Le Morvan et al. (2021) have demonstrated that in the supervised setting, powerful enough models (such as boosted trees) trained on datasets imputed by the mean can achieve good prediction performances. Following the protocol presented by the authors, in our experiments we impute the test sets using the constants learned on the training data.

MICE impute-then-regress. The missing values are first imputed using the conditional imputation algorithm MICE. We implement this approach using the `IterativeImputer` function of `scikit-learn` (Pedregosa et al., 2011), that uses linear models to learn the impute the missing values. We impute the test sets using the model learned on the training data. Predictions are made using gradient boosted trees, taking as input the imputed data concatenated with the missing patterns mask M.

KNN impute-then-regress. The missing values are first imputed using conditional KNN imputation. The implementation is done using the `KNNImputer` of `scikit-learn`. As for the other impute-then-regress methods, we impute the test sets using the model learned on the training data. Predictions are made using gradient boosted trees, taking as input the imputed data concatenated with the missing patterns mask M.

supMIWAE. The imputation and prediction functions are jointly learned using the sup-

MIWAE model (Ipsen et al., 2022). For datasets where data is missing under a MAR or MCAR assumption, we use a MIWAE imputer (Mattei and Frellsen, 2019), following the original paper. For datasets in MNAR scenarios, we use a not-MIWAE imputer (Ipsen et al., 2021). Having encountered some issues with the code that authors have made available, we have implemented our version in `Pytorch`. We use the same hyperparameters as the original implementation for the depth and width of the imputer and predictor. We train the imputer for 200 epochs, and the predictor for 200, and select the best epoch. For both cases we use a batch size of 256, and an Adam optimizer with an initial learning rate set at 1e−3.

Gradient-boosted regression trees (GBRT). We use a decision tree-based approach to handle missing data without imputation. We use the `HistGradientBoostingRegressor` of `scikit-learn` to implement histogram-based gradient boosting (Friedman, 2001). In this implementation, the missing values are handled using a MIA approach.

The implementation choices made for the impute-then-regress baselines can be justified as follows: (1) we have chosen to use gradient boosted trees for the prediction tasks, as they have been shown to achieve state-of-the-art performances on tabular datasets Grinsztajn et al. (2022); Shwartz-Ziv and Armon (2022); (2) Josse et al. (2019); Sperrin et al. (2020) have shown that adding the mask M as input to the classifier helps the prediction task, by providing information discriminating between values that are observed and values that are imputed; (3) lastly, we impute the test sets using the models learned on the training data (rather than the whole dataset) to avoid data leakage, and ensure that we report accurate performances.

Little to no pre-processing is applied to the datasets before benchmarking the models. For each dataset, categorical features are transformed into a numerical representation. The features are standardized to the same range, defined on the training set and applied to the test set before imputation (as suggested by Karpievitch et al. (2012)). No feature selection is performed. A 80-20 ratio is used to split the data into training and testing sets. Each model is evaluated on 5 random repetitions for each dataset. We consider the averages over 5 repetitions to report the performances

4.4.3.2 Evaluation Criteria

Rather than focusing on prediction performances only, we compute 10 additional criteria to assess the quality of imputation of the models (when applicable), their impact on features interactions, and interpretability. First, we compare the prediction performances of models trained on imputed data, with those trained on the complete datasets (14 original UCI datasets). We then compare the feature distributions of datasets imputed with each model to those of the complete data, computing various robust discrepancy scores between the complete dataset and the imputed dataset. Shadbahr et al. (2023) have highlighted three distinct classes of discrepancy scores used to assess imputation quality: (1) sample-wise scores, computing the error between imputed values and the corresponding ground truths,

widely used in much of the literature; (2) feature-wise distribution discrepancy scores, that quantify the reconstruction quality of individual features distributions. Thurow et al. (2021); Shadbahr et al. (2023) analyze many of these scores in more detail in their works; and lastly (3) dataset-wise discrepancy scores that assess the differences between the whole distributions of the complete and imputed datasets. We compute scores from all three categories in our analysis. Additionally, we compare the features interactions between the complete and imputed datasets to check for the presence of new spurious relations. Lastly, we compare the feature importance of models trained on complete datasets with those of models trained on imputed ones.

Relative prediction performances. We compute the *relative prediction score* of each method on each dataset, defined as:

$$\Delta score = score(f_{ref}(X), Y) - score(f(\tilde{X}), Y), \tag{4.5}$$

where X denotes the complete data, f_{ref} denotes the model used to compute the reference performances on the complete dataset, f denotes the considered model to handle missing values, and $\tilde{X}$ denotes the imputed dataset. In all our experiments, we use gradient boosted trees for f_{ref}. As *score*, we select the accuracy for classification tasks, and the MSE for regression tasks. For the particular case of the GBRT baselines that do not rely on imputation, the relative score is given by $\Delta score = score(f_{ref}(X), Y) - score(f(X_{obs}), Y)$.

Imputation MSE. To measure imputation quality, we first compute the *average sample-wise imputation error* for each imputed dataset, defined as:

$$MSE_{imp} = \frac{1}{d}\frac{1}{n}\sum_{j=1}^{d}\sum_{i=1}^{n}(X_{i,j} - \tilde{X}_{i,j})^2, \tag{4.6}$$

where $\tilde{X}$ denotes the imputed dataset, and the indices i, j refer to the j-th feature of the i-th sample of the dataset.

Goodness-of-fit tests. We perform univariate statistical tests between the features of the complete data, and the imputed features. Specifically, we test *goodness-of-fit* between each X_j and $\tilde{X}_j$ using two-sample Kolmogorov-Smirnov tests on numerical features, and χ^2 tests on categorical ones. For each method, on each dataset, we report the *proportion* of variables for which the null hypothesis (i.e. that the two samples come from the same distribution) is not rejected, defined as:

$$t_{ratio} = \frac{1}{d}\sum_{j=1}^{d}\mathbb{1}_{\mathcal{H}_0:X_j,\tilde{X}_j\sim\mathcal{D}} \tag{4.7}$$

where $\mathcal{D}$ corresponds to a common distribution that X_j and $\tilde{X}_j$ should follow, and $\mathbb{1}_{\mathcal{H}_0}$ denotes the acceptance of this hypothesis.

Average energy distance. We then compute a first feature-wise discrepancy measure: the *average energy distance*, defined as

$$E_{dist} = \frac{1}{d}\sum_{j=1}^{d} D(P_j, \tilde{P}_j), \tag{4.8}$$

where P_j and $\tilde{P}_j$ are the univariate cumulative distribution functions of features X_j and $\tilde{X}_j$ respectively, and $D(P, \tilde{P})$ is the energy distance, that characterizes the equality of the distributions. For two random vectors X and Y with cumulative distributions F and G, the energy distance between the two distributions F and G is given by

$$D(F, G) = (2\mathbb{E}|X - Y| - \mathbb{E}|X - X'| - \mathbb{E}|Y - Y'|)^{1/2}, \tag{4.9}$$

where $\mathbb{E}$ denotes the expected value, and $|\cdot|$ is the length of a vector. A low energy distance indicates a high similarity in the compared distributions.

Average W_1-Wasserstein. Then, we compute a second feature-wise metric: the *average* $1d$ W_1*-Wasserstein distance*, defined as

$$W_{1_{dist}} = \frac{1}{d}\sum_{j=1}^{d} W_1(P_j, \tilde{P}_j), \tag{4.10}$$

where P_j and $\tilde{P}_j$ are the univariate probability distributions of features X_j and $\tilde{X}_j$ respectively, and where $W_1(P, \tilde{P})$ is the W_1-Wasserstein metric. For two probability distributions P and Q, the W_p-Wasserstein distance is defined as

$$W_p(P, Q) = \inf_{\gamma \in \Gamma(P,Q)} (\mathbb{E}_{(x,y)\sim\gamma} d(x, y)^p)^{1/p}, \tag{4.11}$$

where $\Gamma(P, Q)$ is the set of all joint distributions whose marginals are P and Q. Again, a low W_1-Wasserstein distance indicates a high similarity in the compared distributions.

Earth-mover's distance. We compute a final metric for evaluating the imputation quality: the d-dimensional W_1-Wasserstein distance between X and $\tilde{X}$ (see Eq 4.11), also known as the earth mover's distance. This metric is a dataset-wise measure, and as such compares the whole distributions of the complete datasets to the whole distributions of the imputed datasets.

Remark 4.4.1. As the GBRT does not rely on imputation, all imputation-related metrics described above are computed to compare X with X_{obs} directly (instead of $\tilde{X}$) in this case. As such, we are evaluating whether handling the missing values natively allows to preserve a data distribution that is close to the reel underlying distribution, rather than assessing the quality of imputation.

Frobenius norm. In addition to imputation quality, we want to assess whether new spurious relationships have been introduced in the datasets. To do so, we compare the features correlation on the complete datasets, to the ones computed on the imputed datasets. We compute the Frobenius norm between the two matrices, given by

$$||(C - \tilde{C})||_f = \sqrt{\sum_{i,j} |c_{i,j} - \tilde{c}_{i,j}|^2}, \tag{4.12}$$

where C and $\tilde{C}$ correspond to the correlation matrices computed on the complete and imputed datasets respectively. A high Frobenius norm indicates a large difference in the compared matrices.

Correlation matix distance (CMD). We additionally compute the correlation matrix distance described in Herdin et al. (2005), defined as

$$d_{corr}(C, \tilde{C}) = 1 - \frac{tr(C \cdot \tilde{C})}{||C||_f ||\tilde{C}||_f} \quad \in [0, 1], \tag{4.13}$$

where $||\cdot||_f$ denotes the Frobenius norm. This metric goes to zero if the compared correlation matrices are equal up to a scaling factor, and goes to one if they differ completely.

Rank-biased overlap score (RBO). Then, we compare the feature importance of models trained on complete datasets against models trained on imputed ones to assess the reliability of the interpretability of state-of-the-art approaches. To do so, we start by using the SHAP framework (Lundberg, 2017) to derive feature importance rankings: for each feature of a dataset, SHAP assigns an importance to each individual sample, quantifying its impact on the model outcome. We define the feature importance vector of a model $R_{SHAP} := (r_1, \ldots, r_d)$ as the feature-wise averaging of the SHAP values across all the observations of the dataset, such that:

$$R_{SHAP} := (r_1, \ldots, r_d) = \left(\frac{1}{n} \sum_{i=1}^{n} c_{i,1}, \ldots, \frac{1}{n} \sum_{i=1}^{n} c_{i,d} \right), \tag{4.14}$$

where the $c_{i,j}$ coefficient correspond to the SHAP value associated to the i-th observation of the j-th feature in a dataset. Ultimately, we evaluate the interpretability of a model f by comparing its feature importance vector to the one of the reference models f_{ref} computed on the complete data, using the *rank-biased overlap score* (Webber et al., 2010). The RBO is a similarity measure between incomplete, top-weighted and indefinite rankings, as such it is perfectly adapted to compare feature importance rankings. For two rankings R_{ref} and R, it is defined as:

$$RBO(R_{ref}, R) = (1-p) \sum_{d_p=0}^{\infty} p^{d_p-1} \cdot A_{d_p}, \tag{4.15}$$

where d_p corresponds to the depth of the ranking to be examined, $A_{d_p} = X_{d_p}/d_p$ is the agreement between R_{ref} and R, $X_{d_p} = |R_{ref:d_p} \cap R_{:d_p}|$ is the size of the overlap between R_{ref} and R up to depth d_p, and $p \in [0, 1]$ is a parameter used to determine the contribution of the top d_p ranks to the final value of the RBO. In our experiments we set $d_p = d$ to ensure that the whole feature importance rankings are evaluated, and $p = 0.4$ such that the first 40% of features have more weight in the final RBO. A RBO value of zero indicates the lists are completely different, and a RBO of one means completely identical.

Computation time. Lastly, and less critically, we consider the computational time of each approach, consisting of the time required to train the imputer; impute the missing values; and train the downstream prediction model.

4.4.4 Analysis of the Results and Model Selection

As expressed in the introduction of Section 4.4, in this study we aim to assess whether state-of-the-art models are able to generate reliable imputations; whether these models are able to capture and preserve the interactions between features of the underlying true distributions; and whether the interpretations that these models provide are reliable. We are particularly interested in understanding if the characteristics of a dataset impact the reliability of these methods. As such, after evaluating 5 different state-of-the-art models on 384 varied datasets using 10 distinct evalutation criteria, we analyze the outcomes of the benchmark to identify any insights into the questions.

4.4.4.1 Exploratory Analysis of the Results

ANOVA analysis. One of the key aims of our study is to identify and quantify the influence of different dataset characteristics on the different facets of the performance of state-of-the-art models. As a first step, we perform multiple one-way analysis of variance (ANOVA) for each of the obtained 10 evaluation criteria, across all methods, to determine the impact of each *fingerprint* described in Section 4.4.2.3 on it. The results of the analysis, reported in Figure 4.3, highlight several strong – yet unsurprising – influential factors: the prediction performances are strongly impacted by the amounts of missing values in the

Figure 4.3: One-way ANOVA analyses for the 10 evaluation criteria computed on our benchmark. The results are reported across all models combined. Each row corresponds to a response variable of separate ANOVA analyses (i.e. the criteria being analyzed), and columns correspond to the single explanatory variables, or impact factors being tested (i.e. fingerprints). Factors not significant at the 5% level are denoted in white.

dataset; the feature-wise distributions metrics – especially the average energy distance and W_1-Wasserstein distance – are strongly impacted by the mechanism generating missing values in the data; the metrics related to feature interaction, as well as interpretability, are greatly impacted by the amount of feature correlation in the true distributions of the datasets.

All these outcomes are coherent and in line with research on missing values. First, Shadbahr et al. (2023) have already highlighted that prediction performances of state-of-the-art models are most strongly impacted by the amounts of missing values in the datasets. Second, it is expected that the imputation quality performances are impacted by missing values mechanisms: in MNAR setting where the missing values are unrelated from the observed values, reconstructing the real underlying features distributions is not a trivial task. Lastly, it is natural that the amount of correlation between features in the true distributions of the data have an impact on the ability of different models to recover these relationships. In particular, approaches leveraging conditional-imputation are specifically designed to take advantage of this type of data settings.

Our end-goal is to understand how to choose the most reliable way to handle missing values. Thus, we build on these observations to explore the performances of each of the 5 models.

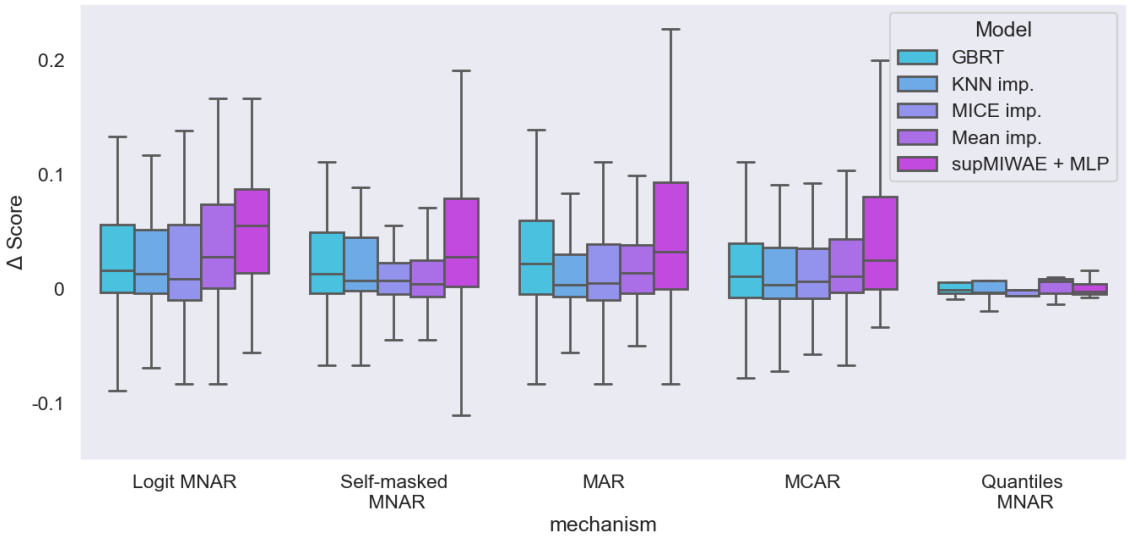


Figure 4.4: Relative prediction performances of the state-of-the-art models with respect to the missing values mechanism in the datasets. For visualization purposes, only relative accuracy scores on classification tasks are reported here.

Prediction performances. In Figure 4.4, we show how each state-of-the-art model performs in terms of relative prediction score. The results are reported across the whole collection of datasets, with respect to the mechanism generating missing values in each of the 384 datasets. Overall, there is no clear patterns that strongly sets a model apart, as most approaches perform rather equivalently. Nonetheless: (1) approaches based on conditional imputation (i.e. KNN and MICE) appear to yield slightly best prediction performances, consistently across the collection. In particular, as expected they perform considerably better than other models in MAR scenarios. Along with mean-impute, they also perform remarkably well in self-masked MNAR settings. They are, however, less adapted to the logistic MNAR scenario. (2) They are closely followed by the imputation-free model (GBRT). However, unlike conditional-imputation approaches, GBRT performs consistently well in all scenarios. (3) In contrast, the impute-and-regress model (supMIWAE) demonstrate a high variability across the collection – highlighting that the approach is not adapted for all datasets and tasks. (4) Despite the low amount of variability in the MNAR with quantile censorship scenarios (due to the small number of datasets in our collection), it appears that MICE-based models generally perform better than other approaches.

Remark 4.4.2. Missing values models occasionally perform better than the reference models (i.e. negative tails on the boxplots), suggesting that on some tasks the presence of missing values is informative and helps improve prediction performances.

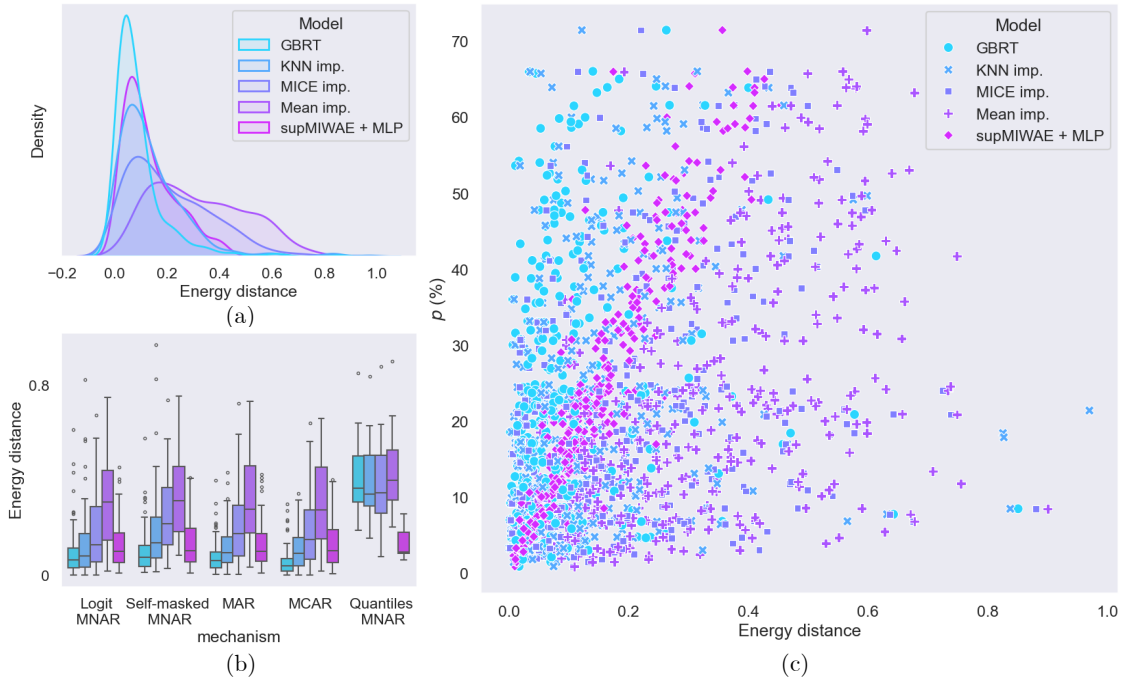


Figure 4.5: Impact of different data characteristics on the average energy distance achieved by the models in our benchmark. We report the global distributions of the performances of each model across the whole collection of datasets (a); the performances of each model with respect to the missing values mechanism of the datasets (b); and the performances of each model with respect to the global amount of missing values p in the datasets (c).

Quality of imputation. In Figure 4.5, we show the reliability of imputation (when applicable) for each state-of-the-art model. Specifically we analyze the performances in feature-wise distribution reconstruction, using the average energy distance achieved by the models in our benchmark on each dataset. As highlighted in Figure 4.5(a), GBRT models perform globally better across the datasets in the collection. Figures 4.5(b) and 4.5(c) also suggest that GBRT demonstrates a clear advantage in performances with respect to the missing values mechanisms and the global missing rate p respectively. This superiority is largely due to the fact that GBRT does not rely on imputation; instead, the distances being computed reflect the comparison between the true underlying distributions of features and the observed distributions, rather than imputed ones. This highlights the significant advantage of inherently handling missing values, as it ensures that downstream analyses and tasks rely on data that remains closer to the true underlying data structure. Nonetheless, while methods like MICE and mean-imputation fall notably short in comparison, KNN-imputation and supMIWAE perform reasonably well too.

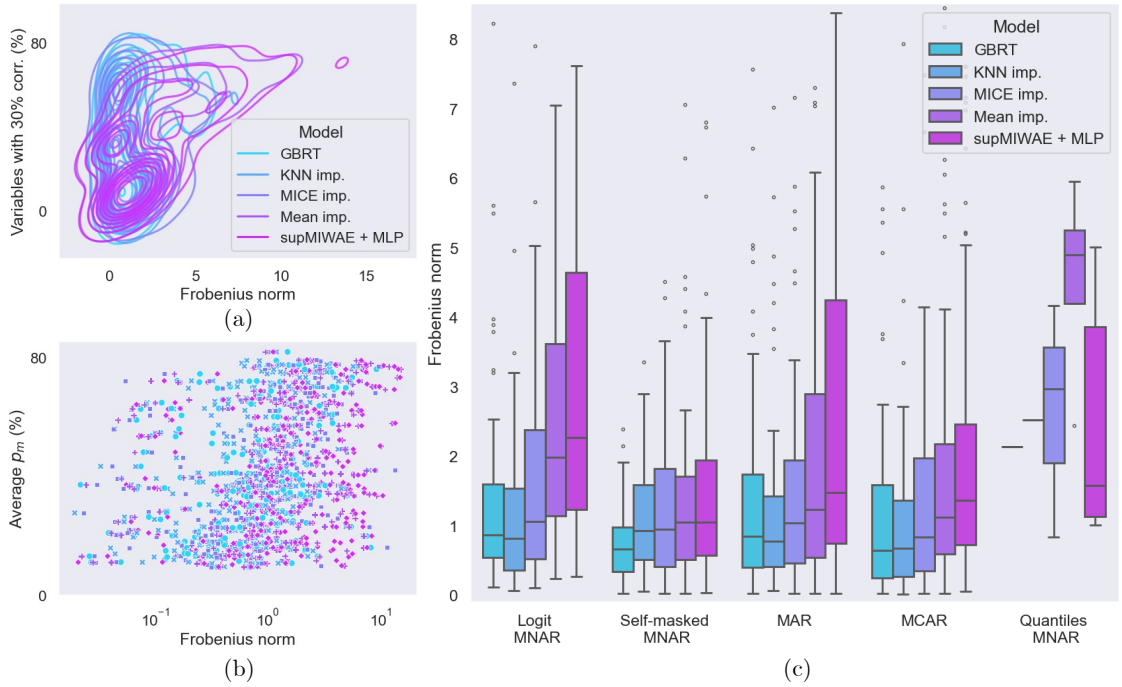


Figure 4.6: Impact of different data characteristics on the Frobenius norm. We report the distributions of the performances of each model with respect to: the amount of correlation in true data distribution (a); the average missing values rates p_m in the datasets (b); and the missing values mechanisms in the datasets (c).

Features interaction. In Figure 4.6, we analyze how well each model preserves the feature interactions present in the true data distributions. We analyze the performances in the Frobenius norm between the correlation matrices of the true data, and the imputed (or observed) data. Figure 4.6(a) shows that in datasets where a large proportion of variables exhibit more than 30% correlation with other features, conditional-imputation models achieve the lowest values of the norm, i.e. minimal difference in feature interaction between the real and imputed data. This aligns with the nature of these models, which excel in highly correlated datasets by imputing missing values conditionally on other observed variables. Figure 4.6(a) demonstrates that GBRT also consistently maintains a low value of the norm, underscoring again the benefits of natively handling missing values. Figure 4.6(b) highlights a clear impact of the average proportion of missing values p_m in incomplete variables on feature interaction – suggesting that high levels of missingness may introduce spurious relationships into the datasets. Finally, Figure 4.6(c) highlights the difficulty of the logistic MNAR and MNAR with quantile censorship scenarios, where accurately recovering the original feature interactions in the data becomes nearly impossible.

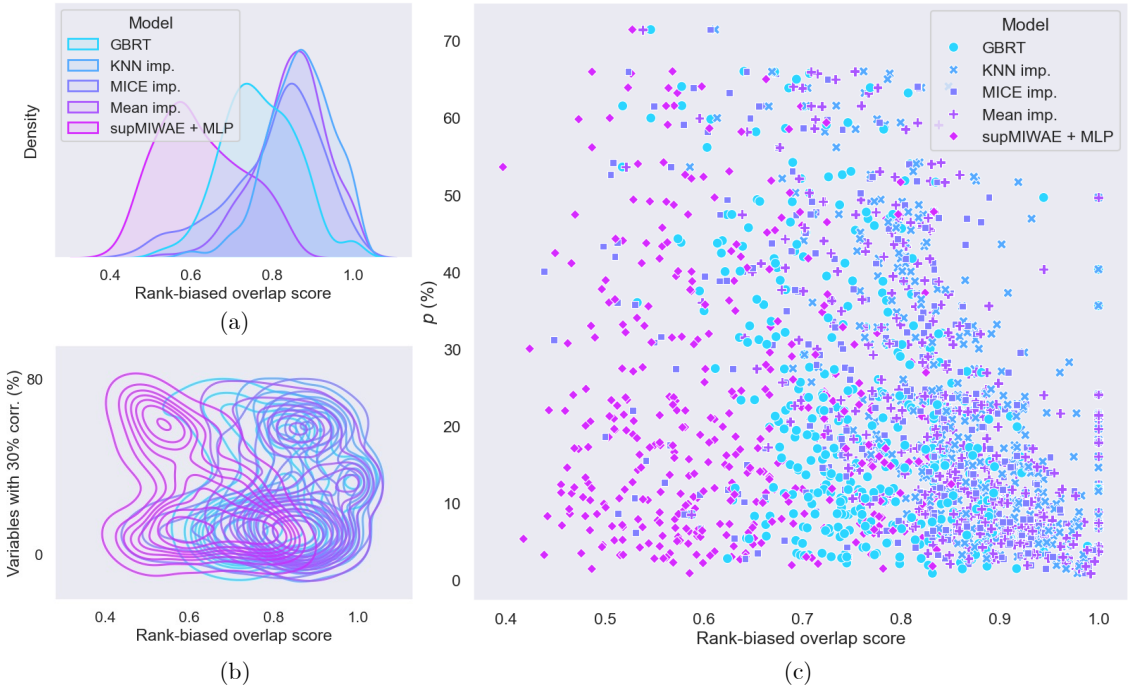


Figure 4.7: Analysis of the impact of different data characteristics on RBO. We report the global distributions of the performances of each model across the whole collection of datasets (a); the distributions of the performances of each model with respect to the amount of correlation in true data distribution (b); and the performances of each model with respect to the global amount of missing values p in the datasets (c).

Reliability of the interpretability. Lastly, we analyze in Figure 4.7, the impact of missing data on the interpretability of each model, using the RBO score computed between the feature importance vectors of the reference models f_{ref} and the models in our benchmark. Figure 4.7(a) reveals that mean, KNN and MICE achieve the highest overall similarity scores. This observation is further supported by Figures 4.7(b) and 4.7(c), which show that conditional-imputation models consistently yield the highest RBO scores when considering the amount of correlation in the datasets $corr_{30}$ and the global proportion of missing values p, respectively. Two hypotheses can explain these results: (1) in healthcare, the presence of a value itself can carry intrinsic information. In these cases, a good imputation may enhance interpretability of subsequent prediction models by recovering meaningful missing values, as noted by Perez-Lebel et al. (2022); (2) inversely, in MAR and MCAR settings, the presence of missing values is not necessarily informative. This scenario is incidentally where GBRT, that handles missing values using MIA – and therefore gives importance to the missingness, is limited and is prone to creating spurious conclusions. As a result, while

GBRT exhibits reasonable performance across the analyses, they remain inferior to other approaches. Lastly, supMIWAE shows the lowest performances, suggesting that its jointly learned imputation compromises the interpretability of the prediction model.

4.4.4.2 Model Selection

Our ultimate goal is to determine the most reliable approach for handling missing values in health-related datasets. However, the previous analyses reveal that no single model consistently outperforms *all* others across *all* the criteria we have defined. More so, the results emphasize how the choice of model is very much dependent on the characteristics of the dataset and the trade-offs between accuracy, bias, and interpretability. To further our analysis, and identify the importance of different dataset characteristics on model choice, we propose an approach to determine the best model on each of the 384 datasets of our collection.

Specifically, we propose to use a linear combination of the ranking of each model on each criteria. On each dataset of the collection, the procedure is as follows:

1. We rank the performances of the 5 considered models on each of the 10 criteria separately, such that for each model i we obtain a vector of 10 rank positions called $R_i := (r_1, \ldots, r_{10})$, where $r_k \in \{1, 5\}$.
2. For each model i, we then compute the average ranking $\bar{r}_i$ across the 10 criteria, such that $\bar{r}_i = \sum_{r_k \in R_i} w_k \cdot r_k$, where w is a vector of weights such that $w_k \in [0, 1]$.
3. Finally, we rank the resulting average model rankings $\bar{r}_i$ among them to determine the model that ensures the best trade-off between accuracy, bias, and interpretability.

Figure 4.8 compares the model selections obtained with 4 different sets of weights w (defined according to the following order of criteria: $\Delta Score$, E_{dist}, $W_{1_{dist}}$, EMD_{dist}, t_{ratio}, MSE_{imp}, $Frob$, d_{corr}, RBO and time).

Scenario 1: in Figure 4.8(a) the model selection considers the prediction performance as sole criteria. **Scenario 2:** in Figure 4.8(b) the linear combination of performances corresponds to an unweighted average of the 10 criteria we have defined. As 5 of these criteria are related to imputation quality, GBRT shows an unsurprising advantage over other models. We have demonstrated in previous experiments the benefits of learning prediction tasks on the obeserved data X_{obs} directly, rather than imputed datasets. **Scenario 3:** in Figure 4.8(c) the weights are re-scaled such that prediction performances, imputation quality, feature interaction and interpretability are equally important in the model selection (whereas the weight associated to computational time is reduced). Under this scenario, model selection is more nuanced. No single model outperforms all other across the collection of datasets, and the best choices are divided between GBRT and conditional imputation. **Scenario 4:** lastly, in Figure 4.8(d) the weights are re-scaled such that imputation quality, feature interaction

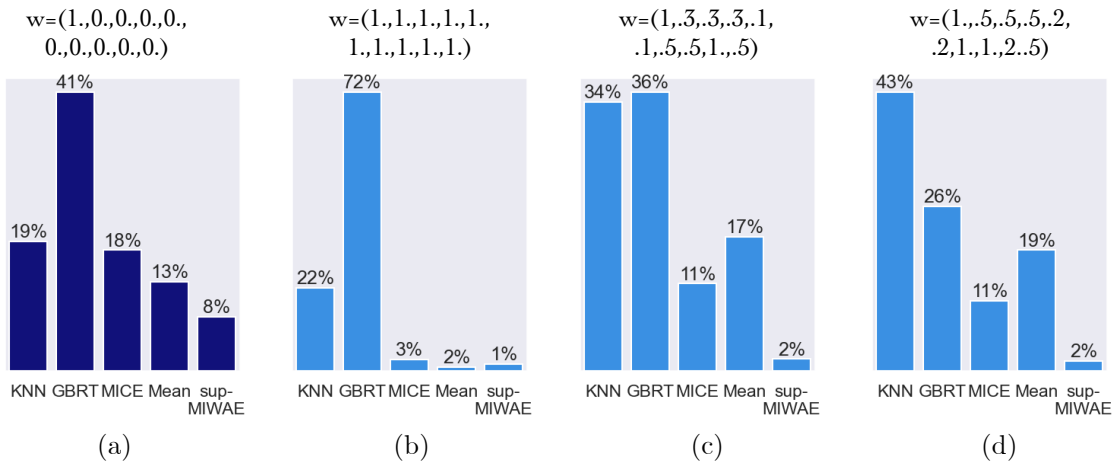


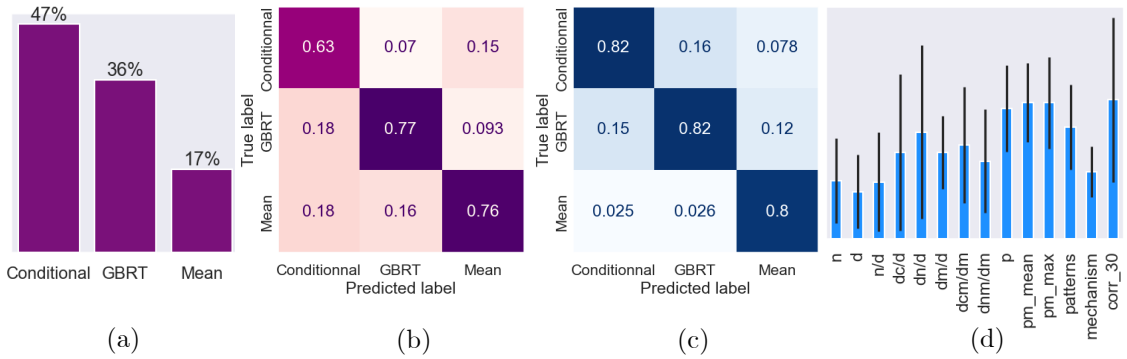
Figure 4.8: Model selection across the collection of 384 datasets, performed with 4 different sets of weights. The selection is performed using: prediction performances solely (a); an unweighted average of the 10 evaluation criteria (b); weights giving equal importance to prediction performances, imputation quality, feature interaction and interpretability (c); and weights giving twice more importance to imputation quality, feature interaction and interpretability than prediction performances (d).

and interpretability are assigned twice the importance of prediction performances. Again, in this scenario the process of model selection is more refined, highlighting that almost all methods show advantages on certain datasets.

The considerable variation in the best model choices across the four scenarios emphasizes that the selection of a model should not rely solely on prediction performance, as is often the case in studies evaluating imputation methods (Jäger et al., 2021; Perez-Lebel et al., 2022). Instead, it is crucial to ensure that models are reliable in terms of imputation quality and interpretability. Finding a good trade-off may sometimes require giving less importance to prediction performance or, at the very least, avoiding model selection based exclusively on it.

4.4.5 Decision tree-based Approach for Model Choice

In the last part of our study, we investigate how to choose the most reliable way to handle missing values, given the characteristics of a dataset. We propose a tree-based approach to analyze the results of our benchmark and extract decision rules for model choice. We aim to ultimately use these rules to help us derive a set of guidelines for learning with missing data in health-related datasets. We use the fingerprints extracted in Section 4.4.2.3, and the results of the model selection processes described in Section 4.4.4.2 to constitute a *dataset of*



(a) (b) (c) (d)

Figure 4.9: Performances of the tree based models with the highest accuracy. We report the class distribution in the target (a); the normalized confusion matrix of the DT (b); the confusion matrix of the RF (c); and the feature importance rankings extracted from the RF (d).

fingerprints. This dataset is used for training multiple tree-based models to predict the best approach to handle missing values, using dataset characteristics as inputs. We consider the following classification task: predicting the best model choice as identified using the scenario 3 of model selection (see Section 4.4.4.2). Due to their low occurrences, we exclude the instances where the best model is supMIWAE. Additionally, we regroup the instances where the best model is KNN or MICE under a single class that we name *conditional imputation*. As a result, the tasks is a 3-class classification problem.

We train a random forest (RF) classifier and a decision tree (DT). Our interest in tree-based models mainly lies in their interpretability properties. Specifically, we aim to extract feature importance rankings from the trained RF to identify which fingerprints have the greatest influence on model selection. Additionally, we seek to retrieve interpretable decision rules from the DT for choosing a model based on its characteristics. Every tree model is trained and hyper-parametrized using 10-fold cross validation and an 80-20 train-test split. Additionally, due to the strong imbalance of labels in the task, we over-sample the training sets using SMOTE to enhance the performances of the trees. We select the models from the 3 splits that achieve the highest accuracies for further analysis of their properties. We obtain RF classifiers achieving 81% accuracy, and DTs reaching 72% accuracy. The lower performances on the DTs are due to the fact that we put the constraint of a *maximal tree depth of 5* to facilitate ulterior interpretations. The obtained DTs can be found in Appendix C.1. Additional performance metrics, along with the feature importance rankings extracted from the RF are reported in Figure 4.9.

Preliminary analyses of the trained RF (see Figure 4.9(d)) suggest that key factors in choosing the most appropriate model for handling missing values include the amount of

feature correlation $corr_{30}$ in data, the missing value rates p_m, and the features types, i.e. the proportion of categorical features present in the data. Surprisingly, the feature importance rankings of the RF also suggest that the missing values mechanisms are less critical in the choice of model. This can be explained by the fact that all the models evaluated in our experiments can perform sufficiently well in MNAR settings: no model in our benchmark shows clear advantage over *all* other baselines across *all* datasets. Le Morvan et al. (2021); Perez-Lebel et al. (2022) have suggested that the practice of adding the missing values pattern M as input in state-of-the-art approaches can make them robust even to MNAR scenarios.

Additionally, we draw several conclusions from the analysis of the obtained DTs. (1) In highly correlated datasets, impute-then-regress approaches using conditional-imputation provide the most reliable predictions. (2) When the amounts of missing values are high, the imputation-free approach, i.e. the GBRT, consistently delivers the best (i.e. reliable) performances. (3) The GBRT model scales well to large datasets. (4) In smaller datasets with low feature correlation, imputation by the mean makes reliable predictions, and can be chosen over other approaches. (5) In datasets where the patterns of missing values are few, imputation by the mean even performs better than other approaches in MNAR scenarios. These observations highlight once again that there is no *one-fits-all* model to reliably handle missing values in health data, and instead, the choice of model is very much specific to the characteristics of each dataset.

4.5 Guidelines for Handling Missing Values in Health Data

By combining the analyses of the outcomes of our benchmark of 5 state-of-the-art approaches on 384 datasets (see Section 4.4.4), and the analysis of the properties of the tree models in Section 4.4.5, we summarize the key findings of our study. Additionally, we propose novel and clear directives for handling missing values in health-related datasets.

4.5.1 Main Takeaways: a Practical Guide with Flowcharts

Imputation-free methods reduce the introduction of bias in the data. As demonstrated in Section 4.4.4.1, the imputation-free approach exemplified by the GBRT in our experiments, minimizes the introduction of bias into the data distributions. Specifically, in most cases the distributions of the observed unaltered (i.e. not imputed) data X_{obs} is closest to the true underlying distribution of X and thus, preserves best the integrity of the original data. As a result, eliminating the need for imputation ensures that the data reflects the underlying patterns more accurately. In contrast, all imputation-methods introduce some kind of distortion, particularly when the imputation method does not fully capture the complexity of the missing data, e.g. in MNAR scenarios.

Conditional imputation may help interpretability of downstream predictors. As highlighted in Section 4.4.4.1, conditional-imputation methods are particularly effective in highly correlated datasets. By definition, these models leverage on the interactions between variables to estimate missing values while accounting for the dependencies between features. As such, they may help recover the true state of feature correlations, in turn improving the interpretability of downstream prediction models. In contrast, GBRT models handle missing values using MIA, i.e. using the missing values to compute the splitting criterion itself in the decision tree. While this enhances predictive performance and can be particularly useful in MNAR scenarios, it can also lead to creating spurious conclusions in MAR and MCAR settings by overemphasizing the importance of the presence of missing values in the model's outcome.

Reliability is a trade-off between accuracy, bias and interpretability. Overall, GBRT and approaches relying on conditional imputation yield the best predictive performance. However, prediction accuracy alone is not a marker of reliability – as each approach has its own drawbacks. A key takeaway from this study, emphasized in Section 4.4.4.2, is that model selection in health applications should not be based solely on prediction performance. It is equally important to ensure that models also provide high-quality imputations and maintain reliable interpretability properties. As such, achieving the right balance may sometimes mean giving less importance to performance. Ultimately, ensuring that a model is reliable involves a trade-off between accuracy, bias, and interpretability.

Choosing a reliable model is dependent on dataset characteristics. In Section 4.4.5, we have identified several key factors influencing model choice. These include the amount of feature correlation in the data, missing values rates and variable types, among others. As such, our most important recommendation is to first analyze the characteristics of a dataset to select the most appropriate model, tailored to the application of interest.

More so, our analyses have provided valuable insights that allow us to derive clear directives for choosing a reliable method for handling missing values according to a specific datasets characteristics. These guidelines are summarized in a flowchart, illustrated in Figure 4.10. We have designed these to be general, and usable for any application. For this reason, we focus on quantifiable, known characteristics that can be computed for any dataset. As such, we do not make the choice of model dependent on the exact knowledge of the underlying mechanisms behind the missing values in a dataset, that are often hard to determine in real-world settings. Our conclusions not only align with existing research and theoretical properties of the models considered, but they also include novel answers supported by thorough experimental validation.

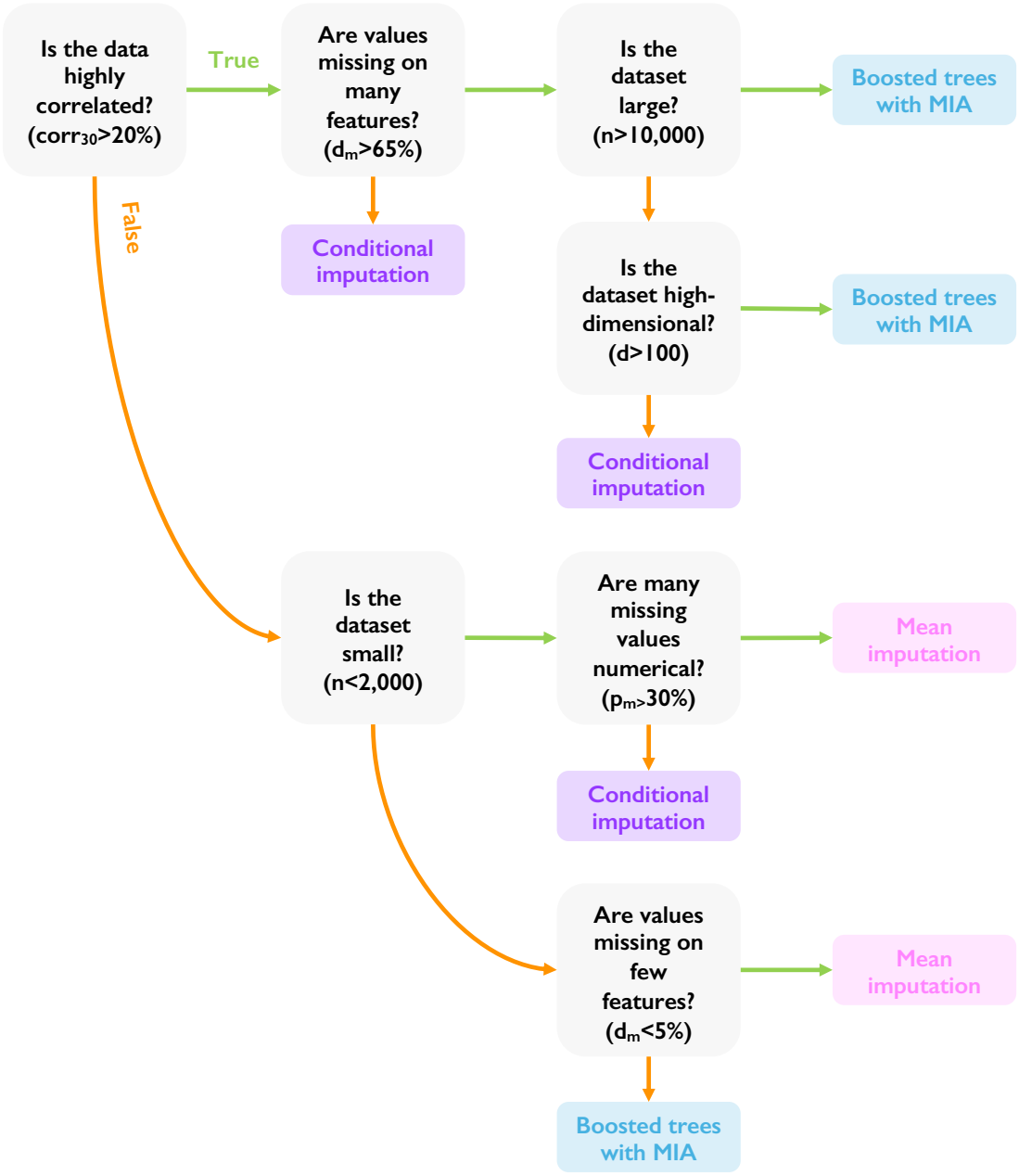


Figure 4.10: Guidelines for handling missing values in healthcare. This selection process ensures a trade-off between prediction performances, minimizing imputation bias, and preserving reliable interpretability. The choice is dependent on measurable quantities only, and can be applied to any dataset.

4.5.2 Illustration on Healthcare Datasets

To further illustrate the benefits of choosing a model based on its reliability rather than prediction performances, we evaluate state-of-the-art models on 7 real-life problems with missing values. In these tasks, the true data X is not known, and instead only X_{obs} is available.

Datasets. We consider 7 classification problems on 3 different databases. (1) MIMIC III (Johnson et al., 2016): the dataset, developed by the MIT Lab for Computational Physiology, includes demographics, vital signs, laboratory tests, and more features associated with about 60,000 intensive care unit (ICU) admissions. We focus on 2 tasks. **M1:** predicting 30-days mortality in patients with sepsis, as described in Hou et al. (2020); and **M2:** predicting in-hospital mortality in patients with heart failure, as described in Li et al. (2021). In both tasks, we use the same features and preprocessing as the authors. (2) Breast Cancer (Razavi et al., 2018): the dataset is derived from an oncology database collected at Memorial Sloan Kettering Cancer Center. It contains genomic profiling tumor samples with detailed clinical variables and outcomes for each patient and the therapy administrated over the time of treatment. We focus on the task of classifying types of breast cancer **(BC)** using features described in Shadbahr et al. (2023). (3) UK Biobank (Sudlow et al., 2015): the dataset is a prospective epidemiology cohort with biomedical measurement recorded on 500,000 participants. We define 4 tasks on this dataset. **UKB1:** diagnosing breast cancer using the features described in Läll et al. (2019); **UKB 2:** diagnosing cardiovascular diseases from diverse genetic biomarkers using the features and preprocessing described in Widen et al. (2021); **UKB3:** diagnosing diabetes using the same subset; and **UKB4:** diagnosing liver problems using the same features. To ensure a robust analysis, datasets with strong class imbalance in the labels are under-sampled. The characteristics after preprocessing of each of the datasets are visualized in Figure 4.11.

Baselines. In addition to the 5 state-of-the-art baselines considered in our previous experiments, we evaluate MIDA imputation (Gondara and Wang, 2018). MIDA is a multiple imputation model based on deep denoising autoencoders, falling in the category of impute-then-regress approaches. We also evaluate an impute-and-regress method proposed by Le Morvan et al. (2021) that uses a NeuMiss block (Le Morvan et al., 2020a) for imputation, chained with an MLP for prediction.

Results. Table 4.1 reports the prediction performances in terms of area under curve (AUC) for the 7 models on the 7 real-life problems we have defined. The results are averaged over 10 random cross validation loops. For each dataset, we compare the best model in terms of AUC with the expected most reliable model, according to our guidelines. Two insights can be drawn from Table 4.1. (1) The divergence between the models with highest AUC scores and the models recommended through the guidelines (i.e. M1, M2, UKB2, UKB4) highlights

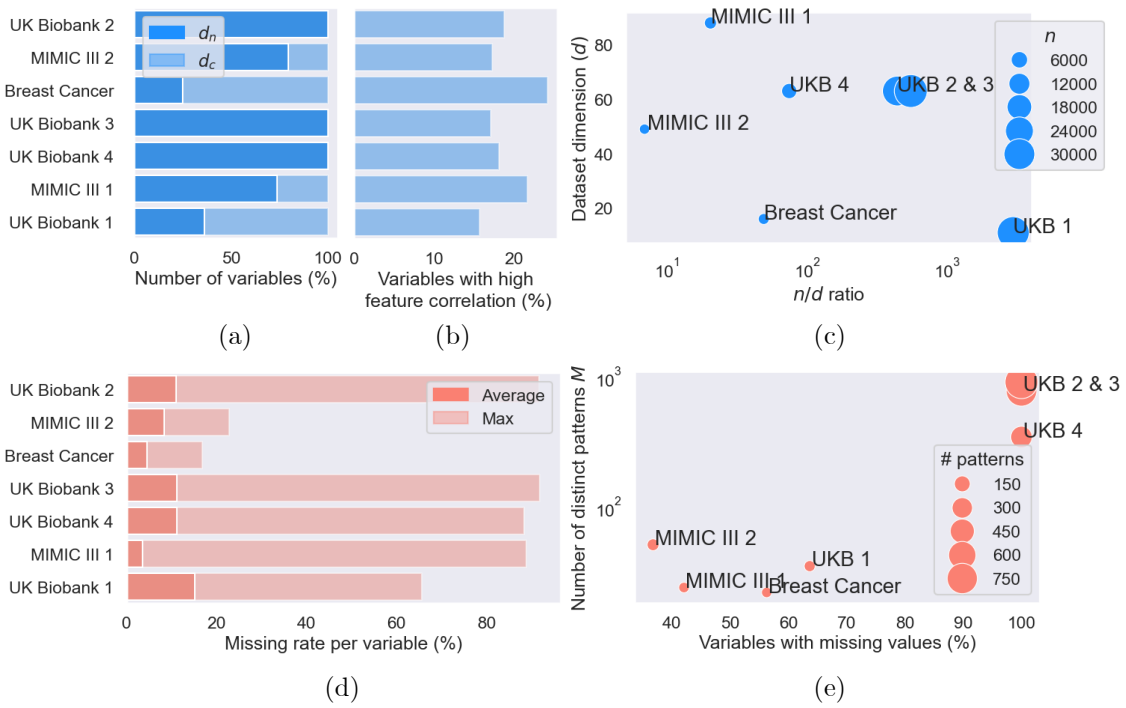


Figure 4.11: Characteristics of the 7 datasets defined on the MIMIC III, Breast Cancer and UK Biobank databases. We report the feature types (a); the amount of feature correlation (b); the sizes (c); the missing values rates per variable (d); and the proportion of variables with missing values and number of distinct missing values patterns M (e) for each dataset.

Table 4.1: Test AUC (mean±std) over 10 random repetitions on the Breast Cancer, MIMIC III and UK Biobank datasets. **Bold** values denote the best prediction performances. Starred* values denote performances that significantly outperform the mean-imputation baseline. Values highlighted in blue denote the most reliable performances according to our guidelines.

	M1	**M2**	**BC**	**UKB1**	**UKB2**	**UKB3**	**UKB4**
Imp.-then-reg.							
Mean	79.7±2.1	72.1±4.6	64.9±4.5	**78.4±0.5**	76.3±0.5	86.0±0.3	**64.2±1.5**
KNN	79.6±2.7	72.2±3.7	64.2±4.6	62.3±0.7	76.1±0.5	85.4±0.3	63.1±1.5
MICE	79.6±1.9	73.4±4.2*	**67.6±3.3***	55.8±11.8	**76.5±0.4**	86.0±0.3	63.0±1.4
MIDA	**79.9±2.1**	72.0±4.0	66.1±2.4*	66.1±12.0	76.1±0.4	85.6±0.4	61.6±1.5
Imp.-and-reg.							
NeuMiss	75.4±2.5	73.5±4.4*	64.0±3.2	51.5±1.0	71.9±1.0	82.5±0.7	57.1±3.2
supMIWAE	76.8±3.0	71.6±4.0	66.7±5.7*	76.1±4.5	73.1±2.1	83.8±1.1	58.5±1.2
Imp.-free							
GBRT	**79.9±2.3**	**75.2±5.2***	65.7±3.1*	**78.4±0.5**	76.3±0.3	**86.2±0.3**	62.9±1.3

again that good prediction performances do not necessarily imply reliable predictions. (2) On several tasks (e.g. M1, UKB2, UKB3), and many real-world datasets, there is little to no significant difference in the prediction performances of various models. This lack of distinction makes the usual model selection process all the more challenging, as there is no clear criterion left to guide the choice. In such cases, focusing on the reliability of the models rather than solely on performance metrics provides a more informed approach. Prioritizing reliability ensures that the selected model carries minimal bias and remains interpretable, offering a more conscious alternative to purely performance-driven decision-making.

4.6 Application to StressID

As a reminder, `StressID` is a multimodal dataset comprising five distinct modalities. However some are missing for certain participants. In Chapter 3, we have developed multimodal models employing feature-level and decision-level fusion, that integrate handcrafted and DL-based features in the form of a tabular dataset. As a result, in our applications, missing modalities translate to missing values within this tabular representation. Previously, we have opted to handle missing entries using list-wise deletion. However, this approach is highly restrictive, particularly in scenarios with limited data, and can be counter-intuitive as it excludes potentially valuable information rather than leveraging all available data. Therefore, we are interested in assessing whether state-of-the-art approaches for handling missing values can be leveraged to enhance the performances of the feature-level fusion multimodal models previously used for the analysis of `StressID`.

Preprocessing and fingerprint extraction. Before evaluating the 7 models defined in Section 4.5.2, we apply feature selection to the dataset. Precisely, we use univariate ANOVA tests to determine the 20 features with the highest impact on the outcome for each modality. After fusion, we obtain a tabular dataset of size $n = 711$ and $d = 100$, where values are missing on 60% of the features, with an average missing rate per variable with missing values $p_m = 20\%$, and a global missing rate $p = 12.5\%$. The dataset contains numerical features solely, and 14% of the variables have an absolute correlation of 20% or more with other variables.

Results. Table 4.2 reports the performances on `StressID`, of each state-of-the-art model for handling missing values. They are compared to the reference feature-fusion and decision-fusion multimodal baselines obtained in Chapter 3, where missing values have been handled using list-wise deletion. Several conclusions can be drawn from the results. (1) Nearly all seven considered models improve the performances of the two multimodal models using list-wise deletion, particularly in terms of accuracy. (2) Since the models evaluated in this chapter employ feature-level fusion, the most relevant comparison is with the previous multimodal model that also used feature-level fusion. In this case, performance improves

Table 4.2: Test F1-scores and accuracies (mean ± std) of state-of-the-art methods for the classification of stress. **Bold** values denote the best prediction performances. Starred* values denote performances that significantly outperform the reference feature-level fusion model. Values highlighted in blue denote the most reliable performances according to our guidelines.

		2-class	
	#tasks	F1-score (↑)	Accuracy (↑)
Multimodal (ref)			
Feature fusion	355	66.4 ± 4.3	61.2 ± 3.7
Decision fusion	355	72.9 ± 4.8	65.2 ± 4.9
Impute-then-regress			
Mean	711	**74.8±2.1***	73.7±2.7*
KNN	711	73.7±3.2*	72.8±2.9*
MICE	711	73.8±5.5*	73.4±5.5*
MIDA	711	74.3±3.1*	73.7±2.7*
Impute-and-regress			
NeuMiss	711	68.4±5.1*	58.0±4.7
supMIWAE	711	**74.8±4.0***	**73.8±3.9***
Imputation-free			
GBRT	711	73.9±2.8*	73.2±2.8*

significantly, highlighting the importance of properly handling missing values and ensuring stress identification models are robust to this challenge. (3) Lastly, according to our guidelines, the model that guarantees the best trade-off between prediction performance, bias reduction, and interpretability is the one based on mean imputation. While it does not achieve the highest accuracy score on the binary classification of stress, its F1-score and accuracy remain significantly better than the reference multimodal model using feature-fusion, demonstrating that missing values can be managed reliably without sacrificing predictive performance.

4.7 Discussion

In this chapter, we have studied the rich existing literature on missing values in tabular datasets. We have designed a framework specifically tailored to evaluating the reliability of state-of-the-art methods in health applications. We have benchmarked 5 approaches from 3 different categories of methods on 384 datasets – using 10 evaluation criteria focusing on aspects like imputation quality, impact on feature interaction and impact on interpretability of downstream predictors. We have extensively analyzed its outcomes and have investigated how the characteristics of a dataset can impact the reliability of these different models – leveraging on tree-based classifiers.

Several aspects of our study could be improved in further analyses. (1) Most datasets used for evaluation are classification tasks. Including more regression tasks would ensure a more comprehensive analysis. Additionally, while the largest dataset considered contains 70,000 samples, many are on a smaller scale; larger datasets would provide better insights into scalability and performance. (2) Our benchmarking includes five different approaches from three different categories. The current set-up involves training each model on 384 datasets, using 5 fold cross validation each. Due to the computational costs of an analysis of this scale, we have not evaluated any multiple-imputation approach, although Perez-Lebel et al. (2022) have illustrated their competitiveness in their work. (3) Model selection in the evaluation of the benchmark relies on a weighted average of the 10 criteria, using deterministic weights. Developing an automated method to assign weights based on the importance of each criterion would be a valuable improvement. (4) The tree models designed to derive guidelines could be enhanced to achieve higher accuracy. Their current performances are limited by their reliance on highly aggregated input features (i.e. fingerprints), which may lack the granularity needed for more accurate predictions. Addressing this issue by incorporating more detailed characteristics in the fingerprints could help.

Nonetheless, the obtained results have provided valuable insights into the reliability of state-of-the-art methods. As a result, we have derived a set of clear guidelines for choosing the most reliable approach for dealing with incomplete entries, given a specific dataset. We have identified that key factors include the amount of feature correlation in the data, missing value rates, and datasets sizes. Surprisingly, we have also found that missing value mechanisms are less critical in model choice, as most approaches perform equivalently in our study (notably due to the practice of using the missing values pattern M as input to the downstream predictors, as highlighted by many studies (Josse et al., 2019; Le Morvan et al., 2021; Perez-Lebel et al., 2022)). Additionally, we found that imputation-free and conditional imputation-based approaches achieve comparable prediction performances. However, on one hand the imputation-free approach introduces the least bias in the datasets by far, as it avoids imputation entirely and instead processes unaltered observed data; on the other hand, the boosted-tree approach (GBRT) we have evaluated significantly alters interpretability, whereas in contrast, conditional imputation methods excel in highly correlated datasets – even helping recover the interpretability of models trained on complete data. Still, relying on imputed values for interpretation remains problematic and should be approached cautiously. Ultimately, the most important conclusion of this analysis is that no single method is superior to others across *all* aspects, making the choice dependent on dataset characteristics and trade-offs between accuracy, bias, and interpretability – and therefore underlining the necessity of having clear directives to help choose the most reliable method. Even more so, we have shown that focusing on the reliability of the models rather than solely on performance metrics offers an informed approach when there is little to no significant difference in the prediction performances of various models. Using the derived guidelines, we have also evaluated state-of-the-art models on `StressID`, demonstrating that they can be

leveraged to ensure the robustness and reliability of baseline models for stress identification.

Our analyses have highlighted the considerable advantages of avoiding imputation and training models directly on the observed data X_{obs}: the dataset remains close to the true distribution X, and preserves original feature correlations. Building on this observation, in Chapter 5 we propose to leverage these benefits. We also address the interpretability challenges of GBRTs by investigating how to ensure that interpretations rely on observed data solely, avoiding biases from missing or imputed values. By doing so, we aim to bridge a critical gap in the current state-of-the-art and offer a more reliable and robust approach that aligns with key criteria for healthcare applications, where trustworthiness is critical.

Chapter 5

PicMi: Imputation-free Supervised Learning in the Presence of Missing Values in Health Data

Contents

5.1 Introduction 84
5.2 Related Work 86
5.3 Method 87
5.3.1 Permutation-invariant Architecture 87
5.3.2 Conditioning Module 89
5.3.3 Attention Module 90
5.3.4 Theoretical Guarantees and Optimization 91
5.4 Experiments and Results 92
5.4.1 Comparison with State-of-the-Art Methods 92
5.4.2 Robustness to Complex Scenarios 97
5.4.3 Ablation Study and Implementation Details 98
5.5 Application to StressID 99
5.6 Discussion 101

Abstract. In Chapter 4 we have identified that robustness to high missing rates, and reliability of imputations and interpretations are key aspects in ensuring trustworthiness of AI models handling missing values in health-related data. However, currently no available method meets *all* of these criteria. In this chapter, we introduce PicMi, a novel model specifically designed to address this

issue. As it does not rely on imputation; is agnostic to the different missing value mechanisms; and is locally interpretable, it is equipped to meet the all the demands of developing robust and reliable healthcare applications.

5.1 Introduction

As highlighted in Chapter 4, supervised learning with missing values faces additional significant challenges specific to healthcare applications. Aspects such as robustness of the models to different missing values scenarios, and reliability of imputation and interpretability mechanisms are crucial factors to consider, for several reasons. (1) In practice the mechanisms behind missing values are not always known, and the misuse of models designed for specific data settings can result in poor imputations and predictions. (2) Imputation can introduce significant bias in data distributions, and subsequent supervised models. We further illustrate this claim in Figure 5.1, that depicts distributions of multiple features from the UCI Heart Disease dataset (Janosi et al., 1989). It compares the features imputed with different methods to the initial distribution without missing values, and the observed distribution (i.e. missing values, no imputation) under several missing values settings. In all cases, the imputed distributions differ strongly from the original ones. This underlines the difficulty of imputing values accurately, and the potential reliance on inaccurate data in ensuing predictors. In contrast, the observed distributions remain the most similar to the original ones, highlighting the advantages of imputation-free models that are able to avoid introducing additional bias in subsequent prediction tasks. (3) Perez-Lebel et al. (2022) have shown that even features with high amounts of missing values have important impacts on the outcomes of models. Yet, poorly imputed data can compromise the interpretability of subsequent classifiers. We illustrate this limitation in Figure 5.2. We compare the feature importance of random forests (RF) trained on the complete Heart Disease dataset (and averaged over 10 repetitions on the same split), to those of RFs trained on data imputed by different methods, using the same missing values scenarios as in Figure 5.1. The rankings of imputed data strongly differ from the initial one. This highlights how poor imputations can lead to incorrect conclusions about the impact of a feature on an outcome. This is all the more problematic in medical applications, where it leads to interpretations relying on values that were not genuinely recorded (i.e. diagnosing a patient on the basis of a feature that was not recorded for them). We have shown in Chapter 4 that no state-of-the-model is superior to another across all these aspects. In particular, we have found that using imputation-free approaches, such as tree-based ones using MIA to handle missing values ensure learning on the most reliable data distributions. Yet, their interpretability remains a limitation. However, models designed for healthcare applications should ideally meet all these criteria to ensure their reliability once deployed.

For this reason, in this chapter we propose an alternative to state-of-the-art works for handling missing values, that is specifically designed to address the limitations that can arise

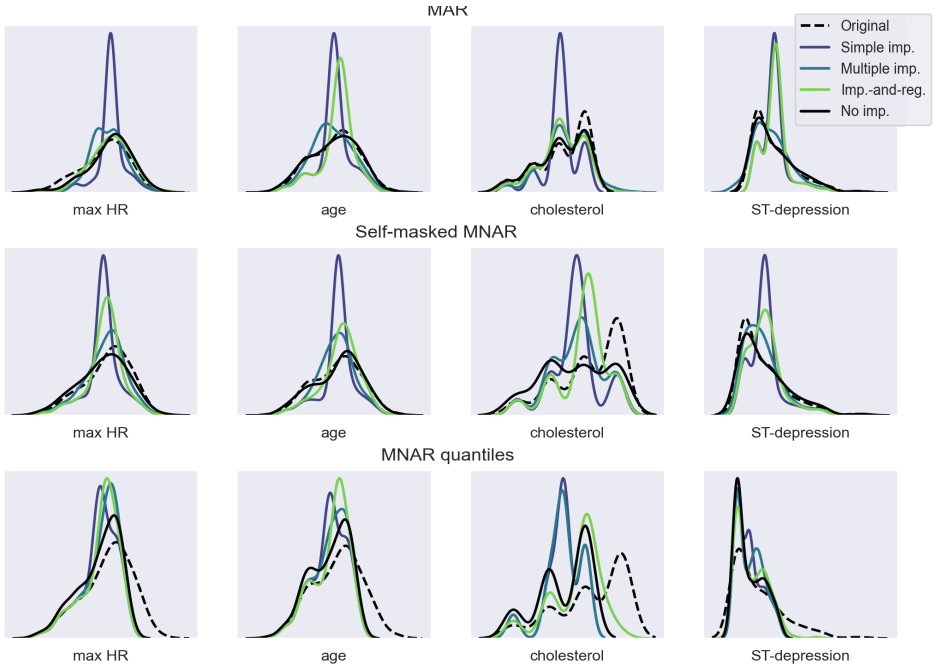


Figure 5.1: Impact of imputation on feature distribution. The distributions of fully observed features from the Heart Disease dataset are compared to: no imputation; imputed with simple imputation (mean); multiple imputation (MICE (Van Buuren, 2018)); and jointly learned imputation model (supMIWAE (Ipsen et al., 2022)). The missing values settings correspond to MAR with a global missing rate of $r = 0.5$, self-masked MNAR with $r = 0.5$, and MNAR with quantile censorship with $r = 0.5$ in the upper half of the data.

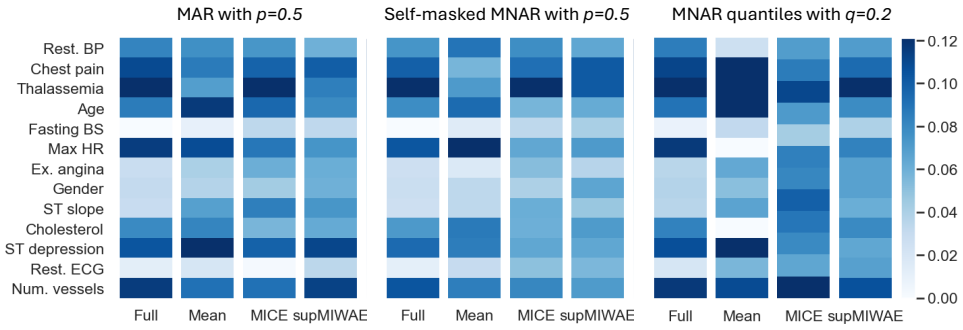


Figure 5.2: Impact of imputation on importance rankings. The ranking of a model trained on fully observed features is compared to the rankings of models trained with data imputed using the same models and under the same scenarios as in Figure 5.1.

in healthcare application. Specifically, we aim to leverage the benefits of avoiding imputation without compromising the reliability of the predictor's interpretability. We propose PicMi, a **P**ermutation-**I**nvariant **C**onditional model for supervised learning with **Mi**ssing values. Our model does not rely on imputation; is agnostic to the different missing value mechanisms; and is locally interpretable. Specifically, PicMi bypasses the need to impute missing values by reformulating the problem of predicting inputs with missing values as one of predicting *sets* of observations, using a permutation-invariant architecture (Zaheer et al., 2017). This relaxes the requirement of fixed-dimensional data inputs of traditional models, and results in a very natural data representation where missing values are simply not represented anymore. Robustness to different complex missing data mechanisms and high amounts of missing values is ensured by introducing a conditioning module enabling the learning of a sparse representation of the input data that integrates the missing patterns as observed variables. Lastly, the reliability of interpretability is achieved by using self-attention mechanisms that quantify the individual weights of each *observed* elements of a set only.

The remainder of this chapter is organized as follows. In Section 5.2, we briefly discuss related work leveraging on permutation-invariant architectures to handle incomplete data. In Section 5.3, we describe our permutation-invariant framework, introduce a conditioning module that makes our model robust to different missing data mechanisms, and describe how we enable our model to make interpretable predictions using self-attention mechanisms. We then demonstrate the benefits of our model through experiments on 11 different health-related databases in Section 5.4. We evaluate PicMi on `StressID` in Section 5.5. Finally, we summarize our work and discuss future directions.

5.2 Related Work

We have discussed in Chapter 4 the existing solutions for handling missing values in supervised learning. We have identified three categories of methods: (1) impute-then-regress methods, that consist of using an imputation model (Van Buuren and Groothuis-Oudshoorn, 2011; Troyanskaya et al., 2001; Stekhoven and Bühlmann, 2012; Mattei and Frellsen, 2019) to fill the missing entries of a dataset before training a prediction model on it; (2) impute-and-regress methods, that propose to jointly learn the imputer and the prediction model to ensure that they are adapted to one another (Le Morvan et al., 2021; Ipsen et al., 2022); and (3) imputation-free methods that handle missing values directly in their design (Le Morvan et al., 2020b; Ayme et al., 2022; Twala et al., 2008). PicMi fits in the latter category. Using a permutation-invariant architecture, it bypasses the need to impute incomplete datasets, and instead allows varying sized-inputs.

Multiple works have leveraged on permutation-invariant architectures (Zaheer et al., 2017; Qi et al., 2017) to handle incomplete datasets. Horn et al. (2020) have used them to handle irregularly sampled multivariate time series by encoding temporal information as a set of

observations. Ma et al. (2018a,b) proposed to deal with missing values in MAR and MCAR settings, using permutation-invariant models for distribution estimation, imputation and image generation. Leveraging on these works, we propose an approach specifically designed for supervised learning. Ipsen et al. (2022) have already evaluated a primitive adaptation of the works of Ma et al. (2018a,b) for prediction tasks. We have developed it further and propose a model that is robust to all missing data mechanisms thanks to an encoder conditioned on the missing data patterns; is interpretable and yields improved performances thanks to the use of attention mechanisms that quantify the importance of each elements of a set in the final outcome.

5.3 Method

We briefly remind the problem formulation introduced in Chapter 4. We consider $n \in \mathbb{N}$ independent input and output pairs $\{(X_1, y_1), \ldots, (X_n, y_n)\}$ where $X \in \mathbb{R}^d$ and $y \in \mathbb{R}$. For simplicity, we denote a single observation X_i as x in the remainder. An indicator vector $m \in \{0, 1\}^d$ is used to denote the positions of missing values in x such that $m_j = 1$ if and only if x_j is missing. For realizations of m, we denote by $obs(m)$ the indices of the observed variables of x, and by $x_{obs(m)}$ the vector of observed elements of x, such that $x_{obs(m)_j} = \texttt{na}$ if $m_j = 1$. The observed data $x_{obs(m)}$ can be written as $x_{obs(m)} = (1 - m) \odot x + m \odot \texttt{na}$ where $\odot$ is the term-by-term product.

Our learning goal is to predict y given $x_{obs(m)}$ and m. We propose to do so by using a permutation-invariant architecture conditioned on m, taking sets of varying-size as input. The overview of our model is illustrated in Figure 6.1. In our framework, all elements of $x_{obs(m)}$ are encoded individually, and a sum of the encodings weighted by self-attentions weights is passed to a classification network.

5.3.1 Permutation-invariant Architecture

Standard machine learning models are built to handle data inputs of a fixed size. Therefore, they are not directly applicable when data have missing values, as working with fixed-size vectors implies that the missing values present in the data need to be replaced by something else (e.g. 0, mean, more complex imputations). In contrast, we aim to alleviate this constraint, and intend on learning without the use of any form of imputation – thus, with entries of different dimension where the missing values are simply not represented. To do so, we propose to learn a *permutation-invariant* function f operating on sets. This formulation relaxes the requirement of fixed-dimensional data. We introduce the following definition to describe our model.

Definition 5.3.1. (Set representation of an observation with missing values) Let $x_{obs(m)}$ of instance i be an entry with missing values. Its set representation is given by S_i of $p = |S_i|$ observed elements s_k, such that $S_i = \{s_1, \ldots, s_p\}$, where $p \leq d$. Each observation s_k is

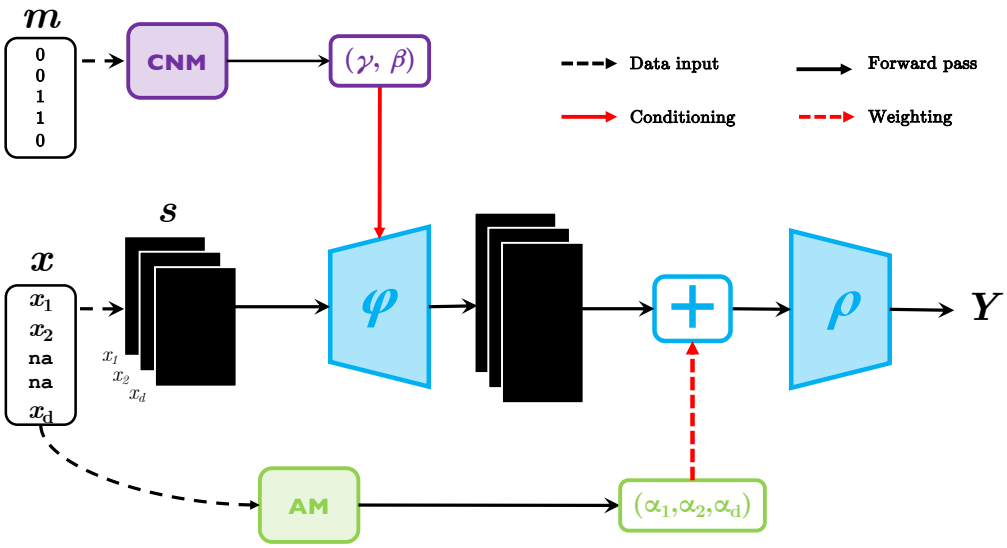


Figure 5.3: Overview of PicMi. A sample with missing values is reformulated as a set with observed values only. Each element of the set is individually encoded through φ, that is conditioned on m via the CNM network. The $\varphi(x)$ encodings are then summed using attention-weights produced by the AM network. Finally, the aggregation is processed by ρ to make predictions.

represented as a tuple (x_j, e_j) consisting of an observed value $x_j \in \mathbb{R}$, and a variable e_j describing its identity.

The identity variable e_j can be defined in various ways, including as a positional embedding, a one-hot embedding of the d variables of x, or an unknown embedding to be optimized during training. Similarly, there are different ways to construct s_k, a common choice being the simple concatenation of x_j and e_j as proposed by Qi et al. (2017). Def. 5.3.1 is deliberately left flexible to allow observations of varying dimensions. Thereby, it does not require nor expects all observations to have the same number of elements and it fully allows observations with missing values. A d-dimensional vector $x_{obs(m)}$ containing `na` values can simply be expressed as a set S of size $p \leq d$ where the `na` values are not represented anymore.

We leverage on the findings of Zaheer et al. (2017), who proposed a learning framework that considers permutation invariant functions operating over sets. The structure of such functions is characterized as follows.

Theorem 5.3.1. *Zaheer et al. (2017) A function f operating on a set X having elements in a countable universe, is a valid set function if and only if there exist functions $\varphi : \mathbb{R} \to L$*

and $\rho : L \to \mathbb{R}$ *such that*

$$f(X) = \rho\Big(\sum_{x \in X} \varphi(x)\Big)$$

In other words, f is invariant to the permutations of its input, if it is sum-decomposable via a latent space L of dimension d_L.

Following Def. 5.3.1 and Theorem 5.3.1, we reformulate our learning goal as one of learning a conditional set function f of the form

$$f(x_{obs(m)}|m) = \rho\left(\sum_{s_k \in S} \alpha_k \varphi(s_k|m)\right) \tag{5.1}$$

where S is the set representation of vector $x_{obs(m)}$ of instance i from Def. 5.3.1, s_k is a single element of the instance S, and α_k is the attention weight vector associated to s_k. The functions $\varphi : \mathbb{R} \times \{0,1\}^d \to \mathbb{R}^{d_l}$ and $\rho : \mathbb{R}^{d_l} \to \mathbb{R}$ are implemented as neural networks, and $d_l \in \mathbb{N}^+$ is the dimension of the latent space.

In practice, a given observation vector with missing values $x_{obs(m)}$ is encoded as a as a set of unordered measurements S where no new information (e.g. imputed values) is added (Def. 5.3.1). Each element $s_k \in S$ is then transformed into a representation $\varphi(s_k|m)$ through the network φ, conditioned by the indicator vector m denoting the missing elements of x. The representations $\varphi(s_k|m)$ are weighted by an attention weight vector α, and aggregated using a permutation invariant operation such as the sum, the mean or the maximum. By transforming individual elements s_k of S at a time, and then aggregating the transformations, our network encodes sets of varying sizes into a fixed representation $\sum \alpha_k \varphi(s_k|m)$. This aspect is particularly relevant, as it is what enables handling a dataset with missing values as an unordered set. Finally, the aggregation is processed through the network ρ, which allows to predict the target y corresponding to the input $x_{obs(m)}$.

5.3.2 Conditioning Module

As missing values mechanisms are often unknown in real-life datasets, robustness of models to different data scenarios is crucial for generalization. As suggested by Le Morvan et al. (2021); Perez-Lebel et al. (2022), and further demonstrated in Chapter 4, adding the missing values indicator mask m as input is key to making state-of-the-art imputation-based models robust to MNAR scenarios, as it provides additional information about the patterns of the underlying missing values mechanisms. However, in our case, this task is not so trivial; due to the nature of PicMi itself, we cannot concatenate a mask of size d with an input of size $p \leq d$. Instead, we propose to ensure this property by conditioning our model on missing scenarios. Following the hypothesis of Collier et al. (2020), we consider the missing values in the input x to be the result of a corruption process m, resulting in the corrupted version

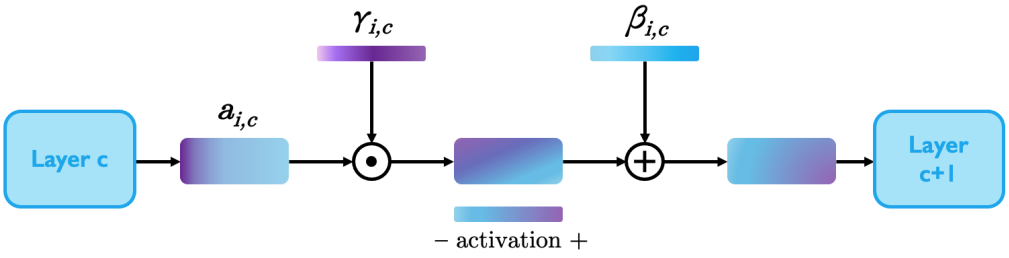


Figure 5.4: A single conditionned layer of φ. The dot denotes a Hadamard product.

$x_{obs(m)}$ of x. By considering m as an additional observed variable, we are able to integrate its structure into the learning of a sparse representation of the input data, in the form of a condition on the φ network, as illustrated in Figure 6.1.

To do so, we introduce a conditional normalization module (CNM) inspired by De Vries et al. (2017); Perez et al. (2018) that learns to adaptively modulate the activations of φ, by applying an affine transformation on the network's intermediate features. Specifically, the CNM learns arbitrary functions g and h (implemented by neural networks) that, for each observation of instance i, take the associated condition m_i as input, and output $\gamma_{i,c} = g(m_i)$ and $\beta_{i,c} = h(m_i)$ respectively, for each layer c of φ. Concretely, φ is conditioned by transforming each layer's activations as follows:

$$\text{CNM}(\mathbf{a}_{i,c}|\gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c} \odot \mathbf{a}_{i,c} + \beta_{i,c}$$

where $\mathbf{a}_{i,c}$ corresponds to the activations of the i-th instance at the c-th layer of the network φ. Figure 5.4 illustrates the conditionning of a single layer.

The encoder network being thus conditioned on m, coerces the model into being agnostic to the missing values mechanisms, and learning to make predictions whether the data is MCAR, MAR or MNAR. This aspect also makes the network robust to datasets with large numbers of missing values patterns.

5.3.3 Attention Module

Understanding how a model makes decisions based on its input is essential in the healthcare domain. However, in the presence of missing values, general interpretability (i.e. feature importance) of simpler models can be biased by missing or imputed values, and neural-network based models ones are not inherently explainable. We propose to address this limitation by enabling our model to make interpretable predictions using self-attention mechanisms that quantify the individual weights of each observed elements of a set, in a similar fashion to Horn et al. (2020). In particular, we introduce an attention module

(AM) that learns an attention vector $\alpha = \{\alpha_1, \ldots, \alpha_p\}$ for each observation S of our dataset. Then, α can be used to compute the reinforced aggregation $\sum \alpha_k \varphi(s_k)$, as illustrated in Figure 6.1. We define our attention weights following Amekoe et al. (2023). Given an input set S of size p, the attention vector α is defined as:

$$\alpha = \frac{\sum_{dk} Q \odot K}{d_k}$$

where $K, Q \in \mathbb{R}^{p \times d_k}$, defined respectively as $K = [k_1, \ldots, k_p]^T$ and $Q = [q_1, \ldots, q_p]^T$, are projections of the input data into p keys of dimension d_k. In practice, K and Q are generated using two separate MLPs with sigmoid activations, bounding all their elements in $[0, 1]$.

This allows our model to identify the features that contributed the most to its output, but also take into account potential strong feature interactions, unlike many state-of-the-art post-hoc interpretability tools (Chen and Guestrin, 2016; Lundberg, 2017), and Horn et al. (2020) that operate on single data points. In addition, the sum or mean aggregation functions can be sensitive to extreme values and redundant information in the data, and the influence of a single observation can decrease as the size of the set increases. Our approach palliates these limitations, as the weighted aggregation favors the most relevant elements of the input sets, over irrelevant ones. Ultimately, the AM of PicMi represents a valuable tool for healthcare applications: while some other approaches for handling missing values can offer general interpretability, none provide patient-specific interpretations. In contrast, our model computes the weights of each observed feature in the final diagnosis, at the patient-level.

5.3.4 Theoretical Guarantees and Optimization

The Deep Set model introduced by Zaheer et al. (2017) is a theoretically sound framework, and the additions we brought to the architecture do not change any of the guarantees of the original model: (1) conditioning the model has already been covered and discussed by the authors in their original paper; (2) it is trivial that weighting a sum or average preserves the permutation-invariant nature of the operation.

The optimization of a set function f that predicts a target variable from the set representation of the incomplete observation $x_{obs(m)}$ given the missing pattern m, simply accounts to solving the following optimization problem

$$f^* \in \underset{f:\Omega \to \mathbb{R}}{\arg\min} \quad \mathbb{E}[\ell(y, f(x_{obs(m)}|m))],$$

where $\ell(\cdot)$ is a task-specific loss function, and f^* is the estimated set function operating on observations with missing values. We optimize a cost function defined as

$$\mathcal{L}(\theta_1, \theta_2) = \mathbb{E}_{(S,Y)\in\mathcal{D}}\Big[\ell\Big(Y, \rho_{\theta_2}\Big(\sum_{s_k \in S} \varphi_{\theta_1} \alpha_{\theta_3 k}(s_k|m)\Big)\Big)\Big].$$

5.4 Experiments and Results

We first evaluate our method on health datasets, and compare it to multiple state-of-the art models. We then study the performances of PicMi with respect to missing values mechanisms and rates to further highlight its robustness to diverse scenarios.

5.4.1 Comparison with State-of-the-Art Methods

Datasets. To evaluate the performances of our models and compare our approach to current state-of-the-art, we select a large range of health-related datasets with varied characteristics. We consider classification problems on the 10 following datasets:

1. **MIMIC III** (Johnson et al., 2016): The dataset, developed by the MIT Lab for Computational Physiology, includes demographics, vital signs, laboratory tests, and more features associated with about 60,000 intensive care unit (ICU) admissions. Here, we use features described in Li et al. (2021) to predict in-hospital mortality in ICU patients with heart failure. After processing, we obtain a balanced subset of 338 samples and 49 features, with missing values in more than 35% of the variables.

2. **Stroke prediction:** The dataset consists of 10 variables representing various health measurements such as BMI, average glucose level or smoking status, and a target variable corresponding to stroke occurrence. It is publicly available on Kaggle. We select a balanced subset of 502 samples, containing 10% missing values on BMI measurements.

3. **Pima Indians Diabetes Database** (Smith et al., 1988): The dataset originates from the National Institute of Diabetes, and Digestive and Kidney Diseases. It gathers 8 health measurement to use to diagnose whether or not the patients have diabetes. We select a balanced subset of 556 samples, containing up to 48% missing values in 5 variables.

4. **MIMIC II:** The dataset corresponds to 43 variables extracted from MIMIC-II (Goldberger et al., 2000), including demographics and clinical observations collected during patients' first ICU stays. The outcome is mortality on the 28th day. We select a balanced subset of 586 samples that contains up to 28% missing values in 16 variables.

5. **Breast Cancer** (Razavi et al., 2018): the dataset is derived from an oncology dataset collected at Memorial Sloan Kettering Cancer Center. It contains genomic profiling tumour samples with detailed clinical variables and outcomes for each patient and the therapy administrated over the time of treatment. We use the features described in Shadbahr et al. (2023) in our analyses. We select a balanced subset of 784 samples, containing missing values in more than 50% of the features.

6. **Myocardial Infarction Complications** (Golovenkin and Voino-Yasenetsky, 2020):

The dataset is available in the UCI repository (Dua et al., 2017). It gathers 111 features from 1700 subjects, with missing values in almost all variables.

7. **Covid-19:** The dataset gathers 20 features corresponding to symptoms, status and medical history of Covid-19 patients to predict whether they are high risk. It was provided by the Mexican government and is currently available on Kaggle. We select a balanced subset of 2120 samples containing up to 56% missing values in 5 features.

8. **Support 2**: The dataset is available in the UCI repository. It comprises 42 features from critically ill patients across 5 United States medical centers, used to predict 6-month survival rates based on several physiologic, demographics, and disease severity information. It consists of 5,826 samples and has missing values in more than 70% of the variables.

9. **Diabetes 130-US Hospitals** (Clore and Strack, 2014): The dataset is available in the UCI repository. It consists of 47 features corresponding to hospital records of patients diagnosed with diabetes collected at 130 US hospitals. The goal is to predict the early readmission of the patient. We select a balanced subset of 20,000 samples, that have missing values in approximately 20% of the features.

10. **UK Biobank** (Sudlow et al., 2015): The dataset is a prospective epidemiology cohort with biomedical measurement recorded on 500,000 participants. Here, we use features described in Läll et al. (2019) to diagnose breast cancer. After processing, we obtain a dataset of 36,642 and 11 features, with missing values in more than 60% of the variables.

The characteristics related to the size of the dataset, as well as summary statistics on the missing values in each datasets after pre-processing are summarized in Table 5.4.1. Figure 5.5 provides a visual representations of these attributes. It is useful to note that most ML models are not inherently robust to imbalanced datasets. As this is not the focus of our work, we have simply used a widespread method to remedy this issue: undersampling (Pereira and Saraiva, 2020).

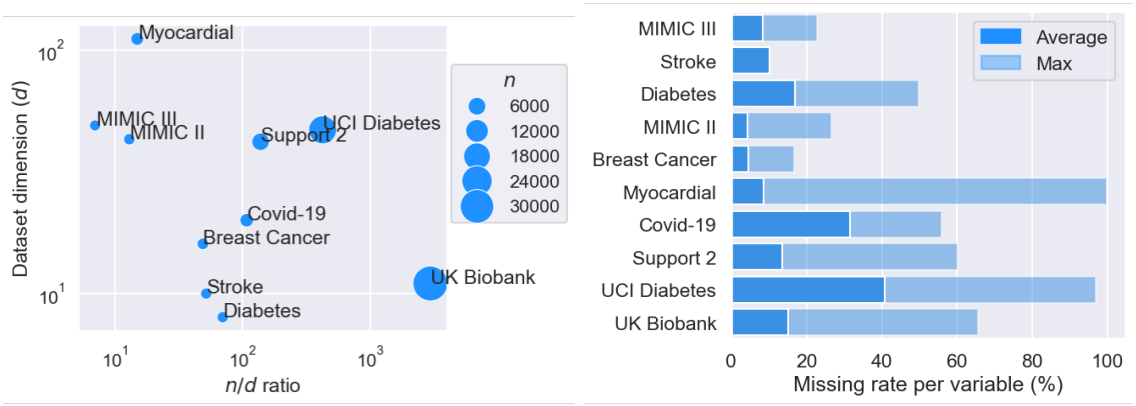
Baselines. We compare our method against models for supervised learning with missing values from each category presented in Section 5.2. More specifically, we evaluate:

– **Impute-then-regress** strategies where mean, MICE (Van Buuren, 2018) imputations averaged over 20 repetitions, and KNN (Troyanskaya et al., 2001) are performed before fitting gradient-boosted trees. MICE and KNN imputations are performed using the `IterativeImputer` of scikit-learn (Pedregosa et al., 2011).

– **Impute-and-regress** strategies where imputation and prediction functions are learned jointly using NeuMiss+MLP (Le Morvan et al., 2021), and supMIWAE (Ipsen et al., 2022) networks.

Table 5.1: Description of the datasets. The global proportion of missing values, the proportion of variables with missing values, the average rate of missing values and the maximal rate of missing values in those variables are denoted respectively as r, d_m, mean p_m and max p_m.

	n	d	n/d **ratio**	r (%)	d_m (%)	**mean** r_m (%)	**max** r_m (%)
MIMIC III	338	49	7	3.1	36.7	8.4	22.7
Stroke	518	10	52	1.1	10.0	10.2	10.2
Diabetes	556	8	70	12.7	75.0	16.9	49.8
MIMIC II	586	43	13	1.7	39.5	4.2	26.6
Breast Cancer	784	16	49	2.5	56.2	4.5	16.8
Myocardial	1,700	111	15	8.5	99.0	8.5	99.7
Covid-19	2,170	20	108	7.8	25.0	31.5	55.8
Support 2	5,826	42	138	11.7	76.2	13.5	60.1
UCI Diabetes	20,000	47	425	7.8	19.1	40.8	96.8
UK Biobank	32,642	11	2967	9.6	63.6	15.1	65.6



(a) Dataset sizes

(b) Missing values amounts

Figure 5.5: Visualization of the characteristics of the datasets used in our experiments. Dataset sizes (left) as well as average and maximal p_m rates per variable with missing data (right) are reported.

– **An Imputation-free** strategy where predictions are made using gradient-boosted regression trees (GBRT) (Friedman, 2001) that handle missing data directly, without the need for imputation. The method is performed using the `HistGradientBoostingRegressor` of scikit-learn.

As many studies (Josse et al., 2019; Le Morvan et al., 2021; Perez-Lebel et al., 2022) have highlighted the benefits of using the mask m concatenated as input in impute-then-regress methods, we have adopted this approach in all baselines.

Results. We report the performances of all models in 2 tables, based on the dimensions d of the datasets. Table 5.2 reports the average accuracies over 10 repetitions on the datasets with $d < 40$, namely the Covid-19, Stroke, Diabetes, Breast Cancer and UK Biobank datasets Table 5.3 reports the performances on the MIMIC III, MIMIC II, Support 2, UCI Diabetes and Myocardial datasets. We performed significance tests in all our experiments. We assess the significance of all models with respect to the baseline (mean impute). The results highlight the competitiveness of our model. First, they highlight that PicMi performs particularly well on datasets with high numbers of available samples n: it outperforms all baselines on the Support2, UCI Diabetes, and the UK Biobank datasets. Second, our model is also very efficient on datasets with high amounts of missing values: along with MICE and KNN, it reaches the best accuracy on the Covid-19 dataset, and it outperforms all other baselines on the Diabetes dataset (in addition to Support2, UCI Diabetes, and the UK Biobank datasets that also suffer from large missing rates). Third, on the Stroke dataset, PicMi outperforms every competitor. On this task, the values are missing on the BMI features, which is an important factor in cardiovascular diseases. The lower performances of other baselines highlight how imputation of such an important values can be problematic. Fourth, PicMi yields good performances in cases where the available n/d ratio are greater than 20 samples per dimension (i.e. Stroke, Diabetes, Breast Cancer, Covid-19, Support2, UCI Diabetes and UK Biobank). We remark nonetheless, that all neural-network based models (i.e. ours, NeuMiss and supMIWAE) are less efficient when learning with low n/d ratios – especially when the total amount of available samples n is small, and are outperformed by impute-then-regress strategies on those datasets (i.e. MIMIC II and III, and Myocardial). This suggesting that neural-network based approaches are not the best choice for high-dimensional datasets with few samples.

The benefits and utility of our method do not only lie in prediction performances. Our method represents a competitive imputation-free alternative for handling missing data, that performs as well as, or outperforms state-of-the-art methods, while eliminating the limitations discussed in Section 5.1. Our proposed approach eliminates the issue of classical imputation-based approaches introduce non-negligible bias in data distributions, and consequently impact the interpretability of downstream classifiers.

Table 5.2: Test accuracies (mean±std) over 10 random repetitions on the Covid-19, Stroke, Diabetes, Breast Cancer and UK Biobank datasets. Bold values denote the best performances. Starred values denote models that are significantly better than the baseline.

	Stroke	**Diabetes**	**Breast Cancer**	**Covid-19**	**UK Biobank**
Impute-then-regress					
Mean	75.2±2.9	76.4±2.1	64.3±4.4	90.2±0.7	78.3±0.3
MICE	75.3±2.2	75.0±2.0	67.5±3.4*	**90.9±0.7**	55.8±11.7
KNN	74.3±3.3	76.1±4.5	64.2±4.6	90.6±1.0	62.3±0.7
MIDA	74.1±2.2	72.7±3.3	66.0±2.4*	90.6±1.2	66.1±12.9
Impute-and-regress					
NeuMiss	72.4±2.9	75.5±6.2	63.8±3.1	84.7±6.6	51.5±4.4
supMIWAE	75.0±2.5	76.1±2.5	66.6±5.6*	89.8±2.4	76.1±4.4
Imputation-free					
GBRT	72.7±4.9	78.2±1.0*	65.6±3.1	90.1±0.2	78.3±0.4
PicMi (ours)	**79.1±1.9***	**83.2±3.0***	**69.4±0.8***	90.9±1.7	**79.6±0.7***

Table 5.3: Test accuracies (mean±std) over 10 random repetitions on the MIMIC III, MIMIC II, Support 2, UCI Diabetes and Myocardial datasets. Bold values denote the best performances. Starred values denote models that are significantly better than the baseline. "Not def." means that the algorithm could not be computed, because not defined.

	MIMIC III	**MIMIC II**	**Support 2**	**UCI Diabetes**	**Myocardial**
Impute-then-regress					
Mean	72.3±4.8	96.1±1.2	75.7±0.9	61.2±0.2	53.7±7.6
MICE	73.5±4.1	96.2±0.7	75.4±1.3	62.2±0.3	64.0±2.4*
KNN	72.3±3.9	96.1±0.4	75.5±0.8	62.3 ±0.3	67.1±1.1*
MIDA	72.2±4.0	96.4±0.5	75.4±1.7	Not def.	Not def.
Impute-and-regress					
NeuMiss	73.5±4.6	93.5±1.2	73.7±0.9	65.9±0.1*	57.4±5.7
supMIWAE	71.9±4.1	94.3±0.9	73.8±1.4	65.2±0.5*	64.2±3.4*
Imputation-free					
GBRT	**75.3±5.1***	**98.1±0.5***	**76.5±1.1***	65.9±0.3*	**67.9±2.1***
PicMi (ours)	62.4±0.8	93.5±1.0	**76.5±0.6***	**66.4±0.2***	64.8±0.7*

5.4.2 Robustness to Complex Scenarios

To further highlight the robustness of PicMi, we conduct additionale synthetic experiments. To illustrate the flexibility of our model, we evaluate it under various complex scenarios. We randomly generate missing patterns m to introduce missing values on the UCI Heart Disease dataset (Janosi et al., 1989) that initially contains no missing data. It is composed of 13 attributes corresponding to various physical and physiological measurements for 303 patients. The task is the prediction of heart disease in patients (i.e. binary classification).

Missing values mechanisms. We consider the following settings for m, as described in Mayer et al. (2019):

- **MAR:** A subset of 2 variables that are always observed are randomly selected. The remaining variables have missing values probabilities given by a logistic model with random weights taking the observed variables as inputs;
- **MNAR self-masked:** Variables have missing values probabilities given by logistic models taking themselves as input. Whether a variable has missing values or not only depends on itself, hence the denomination of self-masking;
- **MNAR with quantile censorship:** The missing values are generated on the q-quantiles. Whether a variable has missing values depends on quantile information, that is masked.

Our experiments have been designed to cover various missing data mechanisms that are currently identified in the literature to demonstrate that our model works well in diverse scenarios. We have evaluated the models under both the mechanisms where state-of-the-art algorithms perform best (MAR) and more complex mechanisms often overlooked in the literature (various types of MNAR).

Results. Figure 5.6 reports the average test accuracies over 10 repetitions for our model, mean and MICE imputation, NeuMiss and GBRT. The models are evaluated under missing rates $r = [0.25, 0.75]$ in MAR and self-masked MNAR settings. 25 and 75% missing values are generated on the upper quantiles $q = 0.5$ of each feature in the MNAR with quantile censorship scenario, resulting in rates $r = [0.35, 0.50]$. The results show the effectiveness of PicMi in complex missing values scenarios. It systematically outperforms all competitors under higher missing rates: while all other models show a significant drop from their performances when the rates of missing values increase, our method remains consistent. In addition, while competitors performances decrease as soon as we step out of the MAR scenario, PicMi is robust to all mechanisms and appears to be the best choice in most MNAR scenarios. This demonstrates the robustness of our model to various difficult data settings – which is an essential element for real-life healthcare applications, where the underlying mechanisms behind missing values are often unknown.

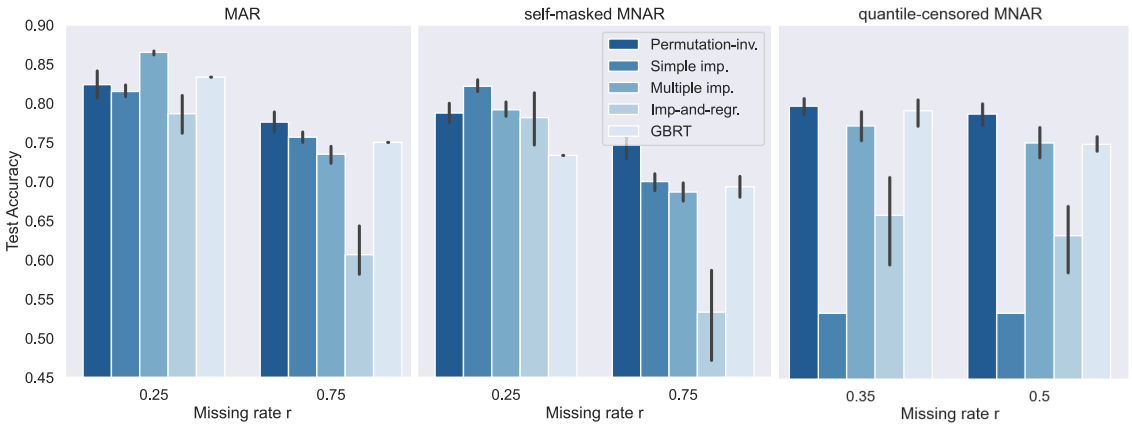


Figure 5.6: Robustness to complex mechanisms and high amounts of missing values. The test performances over 10 random repetitions are reported for our method (permutation-inv.); mean imputation (simple imp.); MICE imputation (multiple imp.); supMIWAE (imp-and-regr.); and imputation-free tree-based prediction (GBRT).

Table 5.4: Ablation study. Performances (mean±std) over 5 random splits.

	Stroke	**Diabetes**	**Support 2**
Perm-inv.	61.9±11.0	64.2±12.7	51.6±.5
Perm-inv. + condition	76.7±.1	76.8±1.4	76.1±1.0
PicMi (condition + attention)	**78.4±1.9**	**82.9±3.0**	**76.4±1.2**

5.4.3 Ablation Study and Implementation Details

Ablation study. We have performed an ablation study to analyze the roles of the conditioning (CNM) and attention (AM) modules in our model. The results are reported in Table 5.4. They highlight that the CNM is an indispensable element of PicMi, as it is what allows our model to reach state-of-the-art level performances. Additionally, the high variance of performances achieved by the *bare* model suggest that the conditional normalization helps stabilize our model. Lastly, the AM slightly improves the performances on 2 of the 3 tasks we have evaluated, suggesting that while its main role resides in introducing interpretability to the model, it also helps the model to better understand interactions in the data.

Implementation details. We built the sets s_k in PicMi as a concatenation of the observed values x_j and their corresponding identity variables e_j, similarly to Qi et al. (2017). We define e_j as an unknown embedding trained during optimization. The size of the embedding e is chosen by CV. The φ, and ρ networks are implemented as *multilayer perceptrons*

Table 5.5: Grid search of hyper-parameters for each evaluated model.

Model	Hyper-parameters	Grid search values	Best choice
NeuMiss	`neumiss_depth`	3, 5, 10	5
	`mlp_depth`	2, 3, 5	2
	`mlp_width`	d, 10, 20, 50	d
supMIWAE	`miwae_depth`	1, 3, 5	3
	`miwae_width`	10, 20, 50	10
	`mlp_width`	10, 20, 50	50
	`n_samples`	10, 100, 1000	100
GBRT	`max_iter`	100, 200	100
	`learning_rate`	0.0001, 0.001, 0.01	0.001
PicMi (ours)	`phi_depth`	1, 2, 3	2
	`rho_depth`	1, 2	1
	`width`	$0.5d$, d, $2d$	d
	`e_size`	2, 5, 10	5
	`attention_dim`	5, 10, 15	5
	`aggregation_func`	`'sum', 'mean', 'max'`	`'sum'`

(MLPs). The number of hidden layers and hidden units of each network are chosen by CV. We note that in practice, the encoder φ needs to project the input set to a higher or of equal dimension space to ensure a good learning of the set function f, as highlighted by Wagstaff et al. (2019) who have demonstrated that a universal function representation of set functions can only be guaranteed if $d_l \geq max_i|S_i|$ is satisfied. We use ReLU activation functions in our implementation. We use the sum operation for the aggregation of the outputs of the network φ. However, any other permutation-invariant pooling operation, such as the sum or the maximum, would also be appropriate. We train our model for 200 epochs, using a batch size of 128, and an Adam optimizer with an initial learning rate set at 1e−3.

The exhaustive list of the different hyper-parameterisations tested for our model, and all other baselines is reported in Table 5.5. We perform 3-fold CV on all tasks presented in the paper, and select the best performing combinations for each model by voting across tasks. All final performances are obtained using the thus chosen hyper-parameters. All experiments are run on a MacBook Air with a 1,6 GHz Intel Core i5 dual core processor. All models are implemented using `PyTorch` and `scikit-learn`, using train-test splits with a 80-20 ratio.

5.5 Application to StressID

Finally, we evaluate PicMi on `StressID`, and compare it's performances with state-of-the-art methods for handling missing values, as well as the reference multimodal models for stress

Table 5.6: Test F1-scores and accuracies (mean ± std) of state-of-the-art methods for the classification of stress. **Bold** values denote the best prediction performances. Starred* values denote performances that significantly outperform the reference feature-level fusion model. Values highlighted in blue denote the most reliable performances according to the guidelines derived in Chapter 4.

		2-class	
	#tasks	F1-score (↑)	Accuracy (↑)
Multimodal (ref)			
Feature fusion	355	66.4 ± 4.3	61.2 ± 3.7
Decision fusion	355	72.9 ± 4.8	65.2 ± 4.9
Impute-then-regress			
Mean	711	74.8±2.1*	73.7±2.7*
KNN	711	73.7±3.2*	72.8±2.9*
MICE	711	73.8±5.5*	73.4±5.5*
MIDA	711	74.3±3.1*	73.7±2.7*
Impute-and-regress			
NeuMiss	711	68.4±5.1*	58.0±4.7
supMIWAE	711	74.6±4.0*	**73.8±3.9***
Imputation-free			
GBRT	711	73.9±2.8*	73.2±2.8*
PicMi	711	**75.0±1.0***	72.7±1.5*

identification (see Chapter 3).

Preprocessing and fingerprint extraction. We follow the same set-up as in Chapter 4: we select the 20 features with the highest impact on the outcome for each modality, using univariate ANOVA tests. After fusion, the dataset contains 711 samples and 100 features, with missing on 60% of them. The average missing rate per variable with missing values $p_m = 20\%$, and a global missing rate $p = 12.5\%$.

Results. Table 5.6 reports the performances of PicMi on `StressID`, along with state-of-the-art approaches for handling missing values. They are compared to the reference feature-fusion and decision-fusion multimodal baselines obtained in Chapter 3, where missing values have been handled using list-wise deletion. Several observations can me made: (1) PicMi significantly outperforms both initial fusion multimodal models used to analyze `StressID`; (2) it outperforms all missing values state-of-the-art approaches on the F1-score – including the most reliable one; (3) it achieves competitive performances in terms of accuracy, yet does not outperform other models. A plausible explanation for this performance is the size of the `StressID` dataset. As discussed in Section 5.4, PicMi performs optimally in large datasets, and especially when the sample to dimension ratio n/d is high – whereas $n/d = 7$ in `StressID`.

Nonetheless, while the mean-imputation baseline (identified as the most reliable one) introduces a considerable bias in the data distributions, PicMi eliminates the imputation step and learns from the observed data directly. In addition, its interpretability is also unbiased, as it leverages on observed values solely. As such, PicMi offers an alternative that achieves comparable prediction performances, all the while eliminating the main limitations of the impute-then-regress approach.

Remark 5.5.1. We have tried to enhance the benefits of PicMi on `StressID` by further reducing the dimension of the dataset – selecting fewer features to obtain a n/d ratio of 20. However, we have found that this considerably decreases the prediction performances.

5.6 Discussion

We have introduced PicMi, an end-to-end framework for supervised learning in the presence of missing values that offers many interesting properties. (1) It relaxes the requirement of fixed-dimensional datasets of traditional models – and thus, does not suffer from the complexity added by the need to choose the right imputation model, as it eliminates altogether the need to fill in missing entries. (2) Our conditional architecture allows our model to integrate the structure of the missing values patterns directly into its learning objective to take into account the *corruption processes* behind the missing values. This aspect is particularly important in practice, as the underlying mechanism behind missing values in real-world data is often unknown and hard to determine. (3) Our approach offers unbiased local interpretability, a highly desirable property in healthcare applications that is unique to our model. Through experiments, we have demonstrated the advantages of PicMi showing that it outperforms several state-of-the-art strategies on real-life tasks. We additionally show that it is also robust to various missing data mechanisms, including in difficult MNAR settings, and high missing rates. This is especially relevant since we have shown in Chapter 4 that the performance of the state-of-the-art methods are highly affected by the amount of missing values.

While our model achieves good performances, it has a few limitations. PicMi outperforms other baselines on MNAR data with high missing rates. However, it is not the best-suited choice for MAR scenarios with moderate amounts of missing values, where multiple imputation methods still outperform competitors in terms of prediction performances. Moreover, PicMi performs optimally on datasets with high n/d ratios, but is less effective with higher-dimensional datasets where $n/d < 20$. In our experiments, this limitation is shared by other DL based approaches (i.e. supMIWAE, NeuMiss). In addition, the experimental set-up we have proposed in this chapter could be improved in the following aspect: although we have proposed an extensive framework to compare PicMi to competitor methods, our study lacks an efficient method for the evaluation of the interpretability of our model. Indeed, our model does not provide a feature ranking (FR), but simply

computes weights to obtain a reinforced latent representation of an observation, in the form of the aggregation $\sum \alpha_k \phi(s_k)$ that is passed to a ρ network for prediction. As so, directly computing summary statistics on the weights and comparing them with FRs of other models would not be a reasonable approach and bear no meaning (even more-so as the whole goal of computing weights is to obtain patient-specific interpretations). Similarly, using post-hoc methods based on perturbation such as SHAP would not be significant on data with NaNs. To the best of our knowledge, there is no efficient way identified in the literature to compare attention weights with feature importance rankings.

Nevertheless, we have demonstrated that PicMi is a competitive alternative for handling missing values in tabular data, that performs as well as, or outperforms state-of-the-art methods, while eliminating the limitations they suffer from. It does not rely on imputation, which is particularly relevant in sensitive fields such as healthcare, where using imputed data can raise concerns about the trustworthiness of algorithms. In addition, PicMi is locally interpretable, unlike competitors that provide general FRs, or rely on post-hoc tools. More so, while the interpretability of other baselines is greatly impacted by imputation (see Figure 5.2), our model provides weights for observed elements only and is not biased by attributing importance to *fake* values. Lastly, we have evaluated our model on `StressID` and have shown that not only PicMi significantly outperforms multimodal baselines obtained using list-wise deletion, but also achieves prediction performances that are comparable to state-of-the-art approaches for handling missing values, while eliminating all their limitations.

Using a permutation-invariant architecture to represent observations of varying sizes as sets where the missing values are simply not represented anymore is a very natural way to handle missing values. Therefore, we propose to take further advantage of this type of architectures, and explore in Chapter 6 whether our approach can be extended to handle missing modalities. In particular, we investigate whether permutation-invariant architecture can be used to design DL-based classification models that operate on the modality inputs directly rather than tabular databases of extracted features, and that are reliable and robust.

Chapter 6

HyperMM : Robust Multimodal Learning with Varying-sized Inputs

Contents

6.1 Introduction 104
6.2 Related Work 105
6.3 Method 106
6.3.1 Overview of the Method 106
6.3.2 Universal Feature Extractor 106
6.3.3 Permutation Invariant Multimodal Classifier 108
6.4 Experiments and Results 109
6.4.1 Alzheimer's Disease Detection 109
6.4.2 Breast Cancer Classification 113
6.5 Discussion 115

Abstract. In Chapter 5, we have demonstrated that using a permutation-invariant architecture offers a natural, efficient and robust way to handle inputs with missing entries. In this chapter we propose to extend this approach to multimodal learning with missing modalities. We propose HyperMM, a model that directly operates on modality inputs, and offers a robust DL-based approach for learning from varying numbers of modalities using mid-level fusion. It is particularly relevant for the development of trustworthy applications using wearable devices data, and more generally medical applications, where it is common to have incomplete modalities in practice. The work presented in this chapter is based on a conference paper published in a workshop in MICCAI 2024 (Chaptoukaev et al., 2024).

6.1 Introduction

As discussed in Chapter 1, multimodal learning (MML) is a promising avenue for the AI-driven analysis of wearable devices data, as it allows to combine modalities from various sources that depict a single subject from multiple views, thus providing both shared and complementary information. We have shown the potential of such models for predicting stress from multimodal inputs in Chapter 3. In reality, MML has shown considerable advantages in multiple other domains (Baltrušaitis et al., 2018; Xu et al., 2023). For instance, multimodal imaging techniques are widely used both in clinical practice and medical research. Simultaneous acquisition and analysis of multiple imaging modalities, such as Emission Tomography (PET), Computed Tomography (CT), or Magnetic Resonance Imaging (MRI), has shown to be beneficial in the diagnosis of Alzheimer's disease (Teipel et al., 2015), or detection of cancers (Tempany et al., 2015), among others. Accordingly, DL methods designed to learn from multimodal medical images, and more generally multimodal health-related data (Sun et al., 2023), have seen rapid growth. However, most current multimodal models assume completeness of the training and testing data, which is rare in real-world applications: due to their ambulatory nature, systems using wearable sensors data are particularly prone to data loss due to factors like sensor malfunction or user non-compliance; in routine clinical practice obtaining several modalities for the same subject is not a standard, for multitudes of reasons, including unavailability of acquisition material (Gallach et al., 2020), or simply patient refusal to partake in specific examinations. As a result, having varying numbers of modalities per patient is common in real-life, which results in multimodal datasets where one or more modalities can be missing. This makes MML challenging as it prevents the straightforward use of the existing methods, as highlighted in Chapter 3. More so, multimodal models trained on complete datasets become unusable (without complex additional processing steps) if modalities are missing at testing time, which severely restricts their usage to complete samples only. Therefore, the robustness of multimodal models to missing modalities is of paramount importance for the use of MML in real-life applications.

Extending the work we have introduced in Chapter 5, we now address the issue of supervised MML with missing modalities by proposing an end-to-end *reconstruction-free* strategy – as opposed to many existing solutions that rely on complex and computationally costly modality reconstruction models. Building on conditional hypernetworks (Ha et al., 2016), we formulate a novel strategy for training a *universal* modality-agnostic feature extractor using pre-trained networks. We then reformulate the problem of predicting multimodal observations with missing modalities as one of predicting *sets* of observations of varying size. We implement this approach through a permutation-invariant neural network (Zaheer et al., 2017), allowing the mid-level fusion of varying-sized multimodal inputs, hence eliminating the need to reconstruct missing modalities. By combining these elements into a two-step training framework, we formulate HyperMM, a novel *task* and *model-agnostic* strategy for MML from incomplete datasets.

The remainder of this chapter is organized as follows. We first provide an overview of related work for handling missing data in multimodal inputs in Section 6.2. We introduce HyperMM and describe the separate elements of our strategy in Section 6.3. We then show through experiments how our approach is suited for various medical applications, illustrating its benefits on 2 different multimodal imaging tasks in Section 6.4. Finally, we summarize our work and discuss future directions.

6.2 Related Work

MML aims to build models that process and combine information from multiple sources (Baltrušaitis et al., 2018), i.e. multiple modalities. The most prominent way to combine multi-source information resides in fusion methods that can be classified in three categories: early fusion, mid-level fusion, and decision-level fusion of modalities (Xu et al., 2023). In practice, summation and averaging are common and straightforward techniques used for fusion. However, when modalities are missing, these operations are impossible for early and mid-level fusion in classical multimodal architectures. They are usually not designed to handle varying-sized inputs and fail to account for missing data.

A vast majority of existing solutions to missing modalities in supervised learning consists of first training a generative model on a complete dataset, and using it to reconstruct missing modalities before learning a discriminative model for prediction (Cai et al., 2018; Kim and Chung, 2020; Sun et al., 2021; Zhang et al., 2023b). This approach has considerable limitations in practice. Firstly, an unreasonable number of samples may be needed for training a good missing-modality reconstruction model. For instance, generative adversarial networks (GANs) (Isola et al., 2017; Zhu et al., 2017), often used for image generation and reconstruction, can typically require up to 10^6 samples for efficient training (Karras et al., 2020). This considerably limits their uses in medical applications where data is often scarce. In addition, the complexity of the prediction model strongly depends on the choice of the reconstruction model. The imputer and predictor networks need to be adapted to each other (Le Morvan et al., 2021; Lu, 2024), which can be difficult to ensure in practice.

Some studies (Suo et al., 2019; Wang et al., 2023a) address this limitation by focusing on jointly learning imputation of the latent modality representations and prediction tasks, but these models rely on complex and computationally costly training strategies. Instead, our approach is simple to optimize thanks to its two-phase training strategy, and can integrate pre-trained models to further reduce computational costs.

Some recent works have also proposed handling missing data without using reconstruction of missing modalities (Parthasarathy and Sundaram, 2020; Zhou et al., 2023; Chen et al., 2024; Mordacq et al., 2024). Instead of directly imputing the missing modalities, they replace them with *dummy* inputs, such as a constant or generated data (e.g., zeros or Gaussian noise), and then learn to ignore these during training using masking strategies. In contrast,

we propose to simply learn with varying-sized inputs to avoid model degradation caused by poor reconstructions or the presence of *dummy* data.

While the methods discussed here-above are not an exhaustive list of existing solutions, they are a good representation of those most widely used in practice, clinical studies, and medical research. For a more comprehensive overview, Wu et al. (2024) provide an extensive survey and taxonomy of recent advancements in MML with missing modalities.

6.3 Method

We consider a dataset $\mathcal{D}$ of $n \in \mathbb{N}$ independent input and output pairs such that $\mathcal{D} := \{(X_1, Y_2), \ldots, (X_n, Y_n)\}$, and for which the goal is to predict Y given X. Each $X := \{x_1, \ldots, x_d\}$ corresponds to a d-modal observation, where each x_i represents one of the available modalities. Let us now introduce the indicator vector $v \in \{0, 1\}^d$ to denote the positions of missing modalities in X, such that $v_i = 1$ if x_i is missing, and 0 otherwise. The observed data of X can be expressed as $X_{obs} = (1 - v) \odot X + v \odot \texttt{na}$, where $\odot$ is the term-by-term product. In this setting, the learning goal becomes the prediction of Y given X_{obs}.

6.3.1 Overview of the Method

We intend on learning without the use of any form of reconstruction of missing modalities, and therefore, with entries of different dimensions. However, standard including MML models are built to handle data inputs of a fixed size. In contrast, we aim to learn a sum-decomposable function f of the form $f = \rho(\sum \varphi(x_i))$, operating on *sets* and thus relaxing the requirement of fixed-dimensional data. We propose a two-step framework that we call HyperMM to implement our method. Figure 6.1 presents an overview of our strategy. In a first step, we learn a neural network φ that can extract features from any modality present in $\mathcal{D}$. Then in a second step, we freeze the learned φ, use it to encode each element of X_{obs}, and feed the combination of the encoded inputs to a classifier ρ through a permutation-invariant architecture.

6.3.2 Universal Feature Extractor

A single network φ that can encode all observed modalities in $\mathcal{D}$ is a requirement for learning a set function as described in Sec. 6.3.1. We propose to achieve this by first learning such a universal feature extractor φ using a conditional hypernetwork (Ha et al., 2016). In this first step, we train a network on all available images x in the dataset, without any modality pairing. As illustrated in Figure 6.1, we introduce an *auxiliary* network h that takes as input m, the modality identifier corresponding to the image x, and generates conditional weights for the last layer of the encoder φ. By doing so, the last feature extraction step is different for each modality but still performed by the same network. Specifically, modality-specific

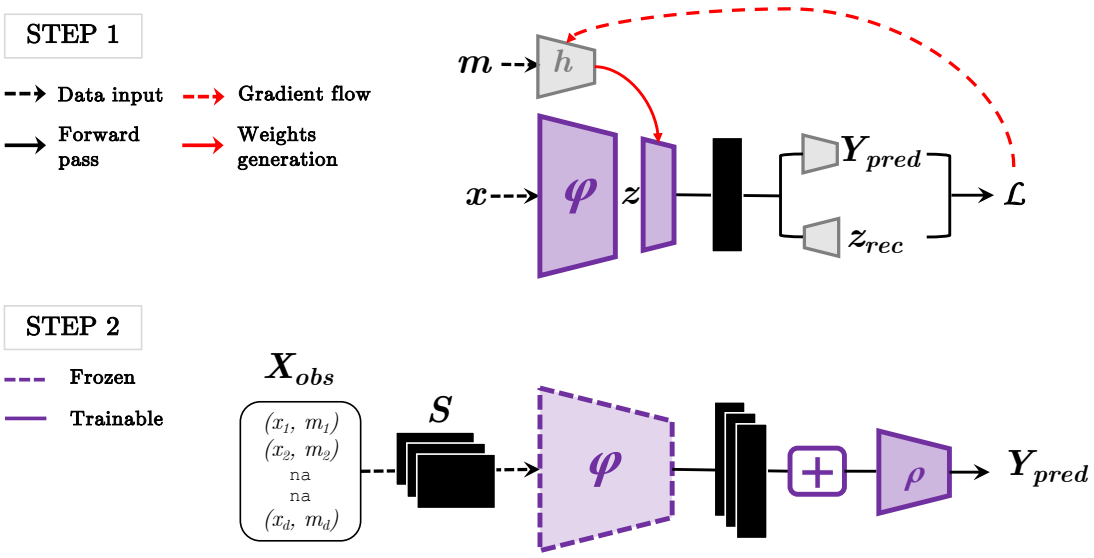


Figure 6.1: Overview of our HyperMM framework. A network φ is trained to extract features from any modality in $\mathcal{D}$ by jointly optimizing feature reconstruction and unimodal prediction (step 1). The learned φ network is frozen, and used to process multimodal inputs, the latent features are then aggregated and processed through a network ρ for prediction (step 2).

layers are generated through a common hypernetwork, which facilitates information sharing across modality-specific layers.

In practice, our universal feature extractor φ can be implemented using transfer learning and networks pre-trained on natural images such as VGGs (Simonyan and Zisserman, 2014). First, we use the pre-trained encoding layers of a VGG to extract features from our dataset. Then, we adapt the obtained general features into medical ones by training an additional layer on top of the VGG extractor, that is conditioned using the auxiliary network h. By stacking these elements together, we obtain our universal feature extractor φ that is adapted to the modalities of our dataset.

To ensure that the features learned by φ are relevant, the network is jointly trained to predict y from the single modality images (i.e. unimodal prediction), and reconstruct z, the features outputed by the second-to-last layer of φ. This is achieved by optimising a loss function of the form $\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{CE}$, where $\mathcal{L}_{MSE}$ denotes the mean squared error between z and z_{rec}, and $\mathcal{L}_{CE}$ the cross-entropy loss between y and y_{pred}. This optimisation loss has been chosen by cross-validation, as it yielded better performances than optimising on the classification or reconstruction only.

6.3.3 Permutation Invariant Multimodal Classifier

Once we have learned φ, we freeze it, and use it to implement a permutation invariant network for supervised MML with missing modalities. To do so, we define S, the set representation of the $q = |S|$ observed elements of X_{obs}, such that $S := \{s_1, \ldots, s_q\}$, with $q \leq d$. Each element s_j is represented as a tuple (x_i, m_i) consisting of an observed modality x_i, and the corresponding modality identifier m_i. This reformulation allows observations of varying dimensions. Thereby, it does not require nor expects all observations to have the same number of elements and it fully allows observations with missing modalities. A d-modal observation X_{obs} containing `na` values can simply be expressed as a set S of size $q \leq d$ where the `na` values are not represented anymore.

Using this definition, we leverage on the findings of (Zaheer et al., 2017), who proposed a learning framework that considers permutation invariant functions operating over sets. We reformulate our learning goal as one of learning a set function f of the form

$$f(X_{obs}) = \rho\left(\sum_{s_k \in S} \varphi(s_k)\right), \tag{6.1}$$

where the function $\varphi : \mathbb{R} \times \{r \times r\}^d \to \mathbb{R}^{d_l}$ corresponds the encoder obtained from the pre-training phase, the function $\rho : \mathbb{R}^{d_l} \to \mathbb{R}$ is implemented as neural network, r is the size of each image and $d_l \in \mathbb{N}^+$ is the dimensionality of the latent space of φ.

As illustrated in Figure 6.1, a given observation X_{obs} with missing modalities is encoded as a set S. Each element $s_k \in S$ is then transformed into a representation $\varphi(s_k) := \varphi(x_i|m_i)$ through the frozen network φ conditioned by the modality identifier m. The representations $\varphi(s_k)$ are aggregated using a permutation invariant operation such as the sum, the mean or the maximum. The aggregation is processed through the network ρ, which allows to predict the target Y corresponding to the input X_{obs}. The proposed architecture interprets each observation S of a dataset as a set of unordered modalities, where all information available in X_{obs} is conserved and no new information, such as imputed images, is added. By transforming individual elements s_k of S at a time and then aggregating the transformations, our network encodes sets of arbitrary sizes into a fixed representation $\sum \varphi(s_k)$. This aspect is particularly relevant and further justifies handling our dataset with missing modalities as unordered sets.

Our permutation invariant model is learned by optimising the loss function

$$\mathcal{L}(\theta) := \mathbb{E}_{(S,Y)\in\mathcal{D}}\Big[\ell\Big(Y, \rho_\theta\Big(\sum_{s_k \in S} \varphi(s_k)\Big)\Big)\Big], \tag{6.2}$$

where ρ is parametrised by θ, and ℓ is the cross-entropy loss. As φ is optimised in the

pre-training step, its weights are not updated in this step.

6.4 Experiments and Results

6.4.1 Alzheimer's Disease Detection

In a first application, we illustrate the performances of HyperMM and its robustness to missing modalities on the task of binary classification of Alzheimer's disease (AD) using multimodal images from the ADNI dataset (Mueller et al., 2005). We select a subset of 300 patients for which both T1-weighted MRIs and FDG-PET images are available, resulting in 165 cognitively normal (CN) and 135 AD observations. Before learning, all the samples are skull stripped using HD-BET (Isensee et al., 2019), resampled through bicubic interpolation to set an uniform voxel size, standardised, and normalised using min-max scaling.

Baselines. We first evaluate the advantages of our strategy for MML with complete data. We compare the performances of HyperMM against:

- **Uni-CNN:** unimodal CNNs as implemented by (Liang et al., 2021).
- **Multi-CNN:** a multimodal CNN as proposed by (Venugopalan et al., 2021).
- **Multi-VAE**, a multimodal VAE (Wu and Goodman, 2018) that we adapt for classification.

Then, we compare our method against state-of-the-art techniques for MML with missing modalities in two scenarios: complete MRIs +50% of PETs available for training and testing, and complete PETs +50% of MRIs available. Specifically we compare to:

- **pix2pix:** a strategy where an image-to-image translation model (Isola et al., 2017) is trained on the subset of the training data containing only modality-complete samples, is then used to impute the missing modality of the incomplete data, and once imputed the data is classified using a Multi-CNN.
- **cycleGAN:** the same strategy, only using a cycleGAN (Zhu et al., 2017) for reconstruction.

Implementation details. We randomly split the data into train, validation and test sets with a 6:1:3 ratio on the patient-level, and repeat all experiments 3 times. For simplicity and fairness, we use the same feature extraction strategy (Figure 6.2) in all baselines, following (Liang et al., 2021). Specifically, 3D MRI and PET images are processed as batches of 2D slices that are each fed to a pre-trained frozen VGG11 (Simonyan and Zisserman, 2014) feature extractor. We feed all 2D slices of a 3D volume to the VGG, and apply a 1D max pooling on the slice dimension to the resulting feature blocks to obtain a single block per 3D image. The resulting block is passed through a 1×1 convolution

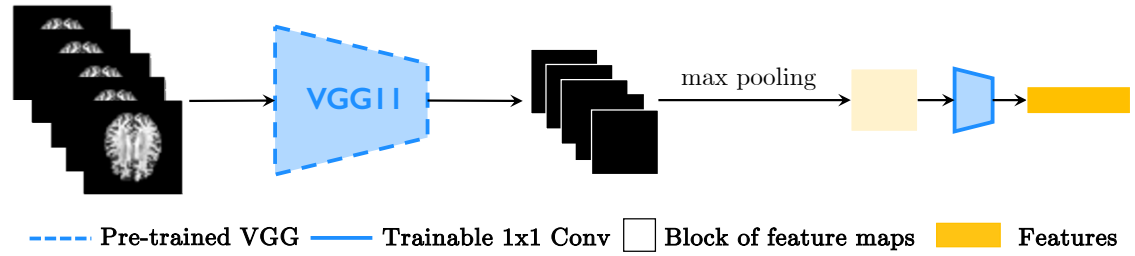


Figure 6.2: Feature extraction strategy used in the ADNI baselines (see (Liang et al., 2021)). All 2D slices of one 3D volume are fed to a VGG11. A 1D max pooling on the slice dimension is applied to the resulting feature blocks to obtain a single block per 3D image. The latter is passed through a 1×1 convolution layer to obtain AD-specific features that can then be fed to a classifier.

layer after the pre-trained VGG encoder, allowing us to adapt the pre-trained features into AD-specific ones.

This corresponds to the training of the φ network in the step 1 of our framework, where we simply make the last 1×1 convolution layer conditional. In step 2, the ρ network is implemented by 3 linears layers separated by ReLU activations. All models are implemented with PyTorch, and trained on an Nvidia TITAN Xp GPU for a maximum of 100 epochs using an early stopping strategy, where training stops after 10 iterations without a decrease in the validation loss. We use a batch size of 1 and an Adam optimiser with an initial learning rate of 1e$-$4.

Results. Performances of all models are reported in Table 6.1. Several observations can be drawn from these results. First, MML shows considerable improvements over unimodal baselines. In particular, HyperMM achieves the best performances for binary classification of AD using complete multimodal data and considerably improves the F1-score, recall metric, and precision/recall balance. Second, MML with missing modalities still achieves better results than unimodal models. Notably, HyperMM trained on MRIs available even for only 50% of the patients performs better than an unimodal model trained on PETs only. Inversely, having access to PETs for 50% of the patients improves the F1-score and recall of learning from MRIs only. Third, HyperMM outperforms state-of-the-art strategies on MML with missing modalities. While GAN-based strategies can handle missing PETs in the input data, they are considerably less efficient in terms of precision/recall balance when the missing modality is MRI. In this scenario, the missing high-resolution MRIs need to be translated from the available low-resolution PETs before learning. This limitation is further illustrated in Figure 6.4.1. While PET reconstruction yields realistic images, the imputed MRIs are of poor quality: they suffer from important structural deformations and a great loss of information (as highlighted by the SSIM and PSNR scores between

Table 6.1: Performances (mean±std) on the ADNI dataset. **Bold** values denote the best performing baselines.

	Acc. (↑)	AUC (↑)	F1 (↑)	Prec. (↑)	Rec. (↑)	Time (↓)
Complete unimodal						
Uni-CNN PET	0.61±.05	0.58 ±.05	0.58±.06	0.65±.06	0.31±.05	< 20 min
Uni-CNN MRI	0.71±.02	0.69±.02	0.58±.02	**0.85±.03**	0.43±.05	< 20 min
Complete multimodal						
Multi-VAE classifier	0.66±.03	0.65±.03	0.54±.04	0.74±.04	0.41 ±.03	< 30 min
Multi-CNN	0.70±.02	0.70±.01	0.67±.01	0.67±.02	0.68±.02	< 30 min
HyperMM w/o 2-steps (ours)	0.62±.03	0.61±.02	0.53±.02	0.61±.03	0.46±.03	< 20 min
HyperMM w/ 2-steps (ours)	**0.74±.02**	**0.73±.02**	**0.70±.01**	0.70±.02	**0.70±.02**	< 1 h
100% MRI + 50% PET						
pix2pix	0.65±.02	0.64±.02	**0.62±.02**	**0.62±.03**	**0.61±.02**	> 14+1 h
cycleGAN	0.62±.09	0.60±.07	0.57±.07	0.61±.08	0.54±.08	> 30+1 h
HyperMM (ours)	**0.67±.02**	**0.66±.02**	0.61±.03	0.61±.03	0.61±.03	< 1 h
100% PET + 50% MRI						
pix2pix	0.62±.04	0.62±.03	0.53±.03	0.61±.05	0.48±.05	> 14+1 h
cycleGAN	0.62±.09	0.59±.1	0.47±.07	0.60±.07	0.39±.07	> 30+1 h
HyperMM (ours)	**0.64±.02**	**0.63±.02**	**0.61±.02**	**0.61±.03**	**0.61±.03**	< 1 h

the reconstructions and the original images). In contrast, as HyperMM does not rely on any reconstruction, it performs well in both scenarios, and trains in significantly less time than competitors. Lastly, these results highlight the importance of the pre-training and conditioning step of the HyperMM framework.

In addition, the results illustrate how HyperMM tackles the main limitations of existing methods. First, as our model does not require training an reconstruction model prior to prediction, it does not call for the large amounts of data typically required for training GANs efficiently. The results observed in Table 6.1 highlight the poor performances of cycleGAN for translating PETs into MRIs, which could be due to insufficient training data. Second, our model is agnostic to the missing modality, whereas the prediction and reconstruction quality in other approaches strongly depends on it, as highlighted by our experiments. Indeed, because HyperMM bypasses the reconstruction step altogether, our approaches eliminates the need to ensure that the imputer and predictor are adapted to each other. This, in turn, leads to drastically reduced computing time and learning complexity. Lastly, as our method does not employ any imputed or dummy data, it avoids model degradation caused by poor reconstructions or noisy data.

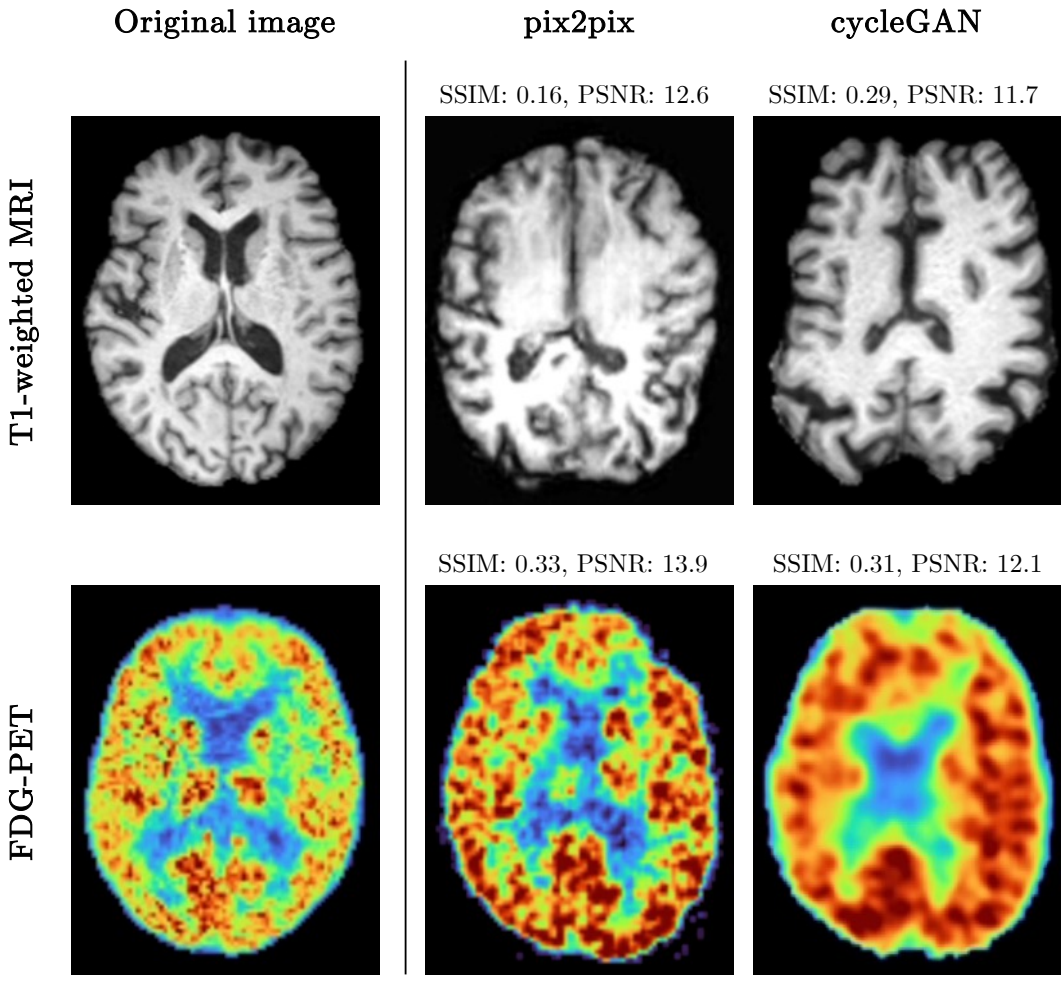


Figure 6.3: Examples of real and imputed slices of MRI and PET images for one patient. While the PET reconstructions (bottom right) translated from the corresponding MRI (top left) are reasonably similar to the original PET image (bottom left), the MRI reconstructions (top right) translated from the low-resolution PET (bottom left) are much less consistent with reality (top left).

6.4.2 Breast Cancer Classification

In a second application, we demonstrate the flexibility of HyperMM and its benefits for learning with varying-sized datasets, beyond the scenario of missing modalities. We investigate the usage of HyperMM for the slightly different task of analysing multi-resolution histopathological images. Because potential tumors are typically acquired at multiple magnification levels, the numbers of samples per patients in histopathology datasets are often highly varying. We perform binary classification of breast cancer using histopathological images from the BreaKHis dataset (Spanhol et al., 2015). BreaKHis contains multiple images per sample (i.e. patient) of benign or malignant tumors observed through different microscopic magnifications: 40×, 110×, 200×, and 400×. We select a balanced subset of the data composed of samples of 24 benign and 29 malignant tumors, resulting in 5,575 images in total. We use the images as they are for learning, and do not perform any pre-processing or data augmentation.

In clinical practice, pathologists combine the complimentary information present in images captured under different magnifications in order to make a patient-level decision. Nonetheless, most current learning approaches consist of magnification-specific models, due to the difficulty of processing images of different natures with a single model. Moreover, because the number of available images can vary a lot from one patient to another, traditional algorithms cannot be applied at the patient-level. Existing methods rather predict from individual images, and later combine the predictions in order to form a global decision. Instead, we propose to tackle this problem using HyperMM, conditioning the universal feature extractor on the different magnification levels. We classify tumors at patient-level by combining all available images during training directly.

Baselines. We evaluate the benefits of HyperMM for learning from histopathology data, and compare its performances with:

- **CNN**, a magnification-specific CNN is trained to classify tumor types from individual images, and patient-level prediction is obtained by averaging the classification scores of individual images (Spanhol et al., 2015).
- **Incremental-CNN**, in which a magnification-agnostic CNN is trained by incrementally updating its weights on successive batches of 40×, 100×, 200× then 400× magnifications, as proposed in (Mayouf and Dupin de Saint-Cyr, 2022). The patient-level decision is obtained similarly to the previous baseline.

The differences between our approach and traditional ones are further illustrated in Figure 6.4.

Implementation details. We randomly split the data into train-test with a 8:2 ratio at the patient-level, and repeat all experiments 5 times. We use a pre-trained VGG11 (Simonyan and Zisserman, 2014) feature extractor for all baselines, and adapt the features to our application by adding a 1×1 convolution block on top of the frozen VGG encoder. All

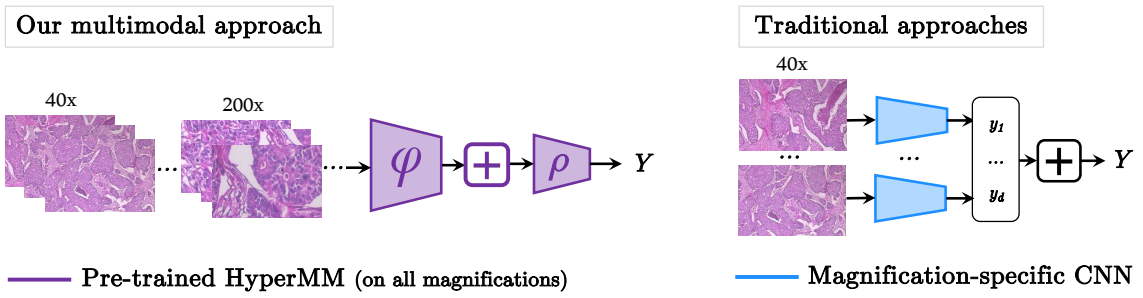


Figure 6.4: Comparison of decision strategies for patient-level tumor classification. Our method (left) enables the combination of a subject's available images during training, regardless of the magnification level to obtain a patient-level decision. In opposition, traditionnal approaches (right) make prediction on the image-level, and combine the final predictions to obtain a patient-level decision.

models are trained for a maximum of 50 epochs using an early stopping strategy such that training stops after 10 iterations without a decrease in the validation loss. We train the model with an Adam optimiser with an initial learning rate of 1e−4. We use a batch size of 16 for image-level baselines (i.e. CNN and Incremental-CNN) and 1 for HyperMM.

Results. All performances averaged over 5 repetitions are reported in Table 6.2. They underline the clear benefits of HyperMM for cancer classification from histopathological images. In particular, our method outperforms magnification-specific models, and is closely followed by Incremental-CNN, which highlights the benefits of combining the information carried by different magnifications. Moreover, while Incremental-CNN maximises the recall score of the task, HyperMM maximises precision, and overall improves upon Incremental-CNN. This shows that learning to predict an early latent combination of features (i.e. combining multiple images of a same patient during model training directly) yields better performances than combining predictions made on individual images.

While the analysis of multi-resolution images is not a multimodal application by definition, our method is designed to enable mid-level fusion of latent features of varying-sized inputs, and is therefore adapted for this use case. Because of the varying number of images per patient in histopathology datasets, traditional approaches are not equiped to combine multiple resolutions directly during training to make patient-level decisions, and instead rely on the late fusion of image-level decisions. In contrast, HyperMM offers this possibility. It opens a new and different way to classify cancer patients. Moreso, our experiments suggest that mid-level fusion even considerably improves the performances of existing late fusion models.

Table 6.2: Performances (mean ± std) on the BreaKHis dataset. **Bold** values denote the best performing baselines.

	Acc. (↑)	AUC (↑)	F1 (↑)	Prec. (↑)	Rec. (↑)
Magnification-specific					
CNN 40×	0.83±0.07	0.81±0.07	0.83±0.06	0.85±0.08	0.83±0.08
CNN 100×	0.85±0.08	0.85±0.08	0.87±0.06	0.85±0.07	0.90±0.07
CNN 200×	0.84±0.07	0.84±0.09	0.84±0.05	0.80± 0.11	0.90± 0.09
CNN 400×	0.83±0.09	0.83±0.09	0.85±0.10	0.88±0.11	0.83±0.15
Magnification-agnostic					
Incremental-CNN	0.89±0.11	0.88±0.12	0.90±0.10	0.88±0.12	**0.93±0.09**
HyperMM (ours)	**0.92±0.06**	**0.91±0.07**	**0.90±0.08**	**0.94±0.09**	0.88±0.10

6.5 Discussion

We have proposed HyperMM, an end-to-end framework designed for learning with varying-sized inputs – with a focus on supervised MML with missing modalities. We introduced a novel strategy for training a *universal* feature extractor using a conditional hypernetwork, and proposed a permutation-invariant neural network that can handle inputs of varying dimensions to process the extracted features, in a two-phase *task-agnostic* framework. We illustrated the relevance of our method in two multimodal imaging tasks: Alzheimer's disease detection and breast cancer classification – demonstrating that flexibility of our strategy allows it to handle varying-sized datasets beyond the scenario of missing modalities.

Although HyperMM is designed to handle multimodal inputs, its current implementation is limited to inputs of the same nature (e.g. multimodal images only, or multimodal physiological signals only). To fully foster its potential, it is crucial to extend the approach to accommodate the integration of diverse data sources, and enable the combination of imaging data with time-series data or textual information. Doing so would allow HyperMM to fully take advantage of multimodal data of `StressID`. This would enable our approach to address a broader range of applications that better reflect the complexity of real-world data. Additionally, the current feature fusion strategy used in HyperMM relies on simple aggregation methods, such as summation or averaging. While these operations are computationally efficient, they are not the best-suited for capturing and leveraging the complementary information present in multimodal inputs. As a result, valuable interactions between modalities may be underutilized. Therefore, a more efficient method to fuse multiple modalities needs to be investigated in order to (1) maximize the use of marginal information provided by each modality and (2) dynamically assign greater importance to the most relevant modalities depending on the context, using tools such as attention mechanism as proposed in Chapter 5.

Nonetheless, our method has considerable advantages: (1) Unlike many competitors, it is end-to-end: HyperMM eliminates the time-consuming steps of manually imputing the missing modalities using a previously trained reconstruction model, before finally training a prediction model. On the contrary, our two-step model is trained without interruption or human intervention. (2) As it eliminates the need to use complex and computationally costly reconstruction strategies, it significantly decreases model training times. More generally, it relies on a simple optimization problem making it computationally efficient. (3) Unlike competitors, its performances are not dependent on which modality is missing in the data. (4) by only using the observed modalities of the incomplete dataset, HyperMM avoids prediction bias caused by poor reconstruction. (5) Lastly, our strategy is *task-agnostic*, and can be easily used beyond the applications we have presented in this chapter. While we used pre-trained feature extractors in all our experiments for simplicity, HyperMM is also *model-agnostic* and adaptable to any neural network-based feature extractor or predictor.

Ultimately, HyperMM represents a significant step towards the development of trustworthy AI-driven multimodal healthcare applications. It is robust to one of the most common perturbations encountered in real-life datasets: missing data. As it does not really on generated data, it is also considerably more reliable than many current approaches that reconstruct missing modalities – which can raise concerns about the trustworthiness of AI and set back the adoption and deployment of AI models in real-life applications. As such, HyperMM represents a solid foundation for building solutions that can be effectively applied across diverse medical contexts and real-world scenarios. By addressing the limitations we have identified here, it has the potential to become an even more versatile and effective framework for multimodal applications combining heterogeneous data sources, beyond the case of wearable sensors data. In Chapter 7, we conclude this thesis and present some perspectives to tackle these aspects.

Chapter 7

Conclusion and Perspectives

Contents

7.1 Conclusion 117
7.2 Perspectives 122

7.1 Conclusion

The unprecedented rise in access to wearable medical devices, advances in data processing, and AI have driven the growth of e-health solutions. As such, AI analysis of multimodal data from affordable wearable sensors offers a promising way to improve access to healthcare – enabling prevention and reducing reliance on costly clinical examinations. However, concerns about the reliability and robustness of AI persist today, preventing the deployment and adoption of AI-driven healthcare research in real-life. In this thesis, we have contributed to the development of robust AI solutions for healthcare applications, focusing on innovative methodologies that are robust to missing data, and evaluated with real wearable sensor data. While centered on wearable data, these methods are adaptable to diverse health applications, from physiological signal analysis to clinical studies with multimodal imaging. The work presented in this thesis spans the entire process of developing e-health applications, with the main contributions summarized below.

StressID: a novel dataset collected with wearable devices. In Chapter 2, we have introduced `StressID`, a multimodal dataset for stress identification that we made available for research. The dataset aims to fill the gap in the existing related databases. It features both physiological and behavioral modalities and includes a large number of participants. It exploits varied stimuli (i.e. emotional video-clips, cognitive tasks, and social stressors

based on public speaking) to guarantee the collection of a wide range of responses and thus, ensure more versatility in downstream applications, and includes participants' replies to 4 self-assessment questions providing insights on the subject's emotional state. Moreover, analyses of the dataset highlight high engagement and a wide range of diverse responses to the stimuli, both between individuals and within the same individual. This highlights that the selected tasks are well-designed, offering sufficient variety and effectively challenging all participants.

Despite the careful design of `StressID`, some limitations can be noted. The dataset is recorded in a relatively controlled environment and does not take into consideration the external factors that contribute to the psychological mental state of participants; relying on self-assessed scales for data annotation is a participant-subjective process, and can lead to bias in subsequent analyses; the dataset suffers from missing modalities for some participants, making MML a challenging task as it prevents the straightforward use of traditional methods; it presents a gender imbalance representative of the female/male ratio in STEM studies and workforce, a common issue in human data collection.

Nonetheless, `StressID` is a valuable resource for research. It has the potential to improve the understanding of the sources, demographics, and both physical and physiological mechanisms of stress responses. It is designed for the development of reliable algorithms for stress identification that can improve the quality of life of our society by helping prevent stress-related issues. Lastly, it is useful to the machine learning and deep learning communities, as it can be used to further evolve multimodal learning algorithms, or to study how to make algorithms learning with human data more reliable.

Baseline models for the analysis of wearable sensors data. In Chapter 3, we have established the state of the art in stress identification using physiological signals, video, and audio data. Building on the models identified in the literature, we have proposed a suite of methods for unimodal and multimodal analysis of `StressID`. We made our implementations public, providing a valuable tool for researchers interested in working with the dataset. Through experiments, we have effectively shown that combining multiple modalities carrying complementary information through multimodal learning has considerable benefits for stress identification, and generally, wearable sensors data analysis.

We have identified several limitations in current state-of-the-art approaches, and the steps needed to ensure reliability and robustness of models built on `StressID`. We demonstrated that models trained on real-world data are prone to bias, such as ones caused by an imbalanced gender representation, highlighting the necessity of taking appropriate measures to ensure systems are reliable before deploying them in real-world applications; more critically, we highlighted that currently, most state-of-the-art models for stress identification are not inherently designed to handle missing values. This underlines the need to either develop novel innovative solutions or adapt existing models to make them robust to missing data.

Nonetheless, the baseline models we have introduced have considerable advantages. They are representative of the state-of-the-art in the domain, where most works rely on ML-based and hybrid methods (i.e. combining feature extraction, via handcrafted techniques or DL, with traditional ML algorithms for classification). These methods are attractive for several reasons. They considerably reduce input data complexity. Their low-complexity make them particularly suited for real-time data processing necessary in wearable sensor applications. Lastly, they enable the use of diverse ML models for the classification of tabular feature datasets and thus, the rich existing literature on handling missing values in tabular data.

Novel guidelines for handling missing values in healthcare. Having identified missing data as a major challenge in the development of reliable and deployable AI systems for e-health, in Chapter 4, we have studied the rich literature on missing values in tabular datasets. We have designed a framework tailored to evaluating the reliability of state-of-the-art methods within healthcare applications. Specifically, we have investigated how the characteristics of a dataset can impact the performances of these different models – focusing on aspects like the bias introduced in imputed data distributions, feature interaction and impact on interpretability of downstream predictors. We have evaluated 5 approaches from 3 different categories on 384 datasets, using 10 criteria to determine the best choice. Lastly, we proposed a decision tree-based approach to analyze the outcomes of this study.

Several aspects of our study could be improved in future work: most of the datasets used for the evaluation framework are classification tasks; many datasets contain less than 5,000 samples, and evaluating on larger datasets would improve scalability analysis; our analysis could benefit from adding more models in our benchmark; currently selection of the best model relies on a linear combination of 10 performance criteria using either unweighted averages or deterministic weights. Developing an automated weighting method would improve objectivity; lastly, the tree models designed for guideline derivation could be enhanced to achieve higher accuracy by incorporating more granulated dataset characteristics in its training data.

Nonetheless, the obtained results have provided valuable insights to derive guidelines on how to chose the most reliable method to handle missing values. We have identified that key factors include the amount of feature correlation in data, missing value rates, dataset size, and variable types. Overall, boosted tree-based approaches that inherently handle missing entries achieve the best prediction performances; they also introduce the least bias in data distributions as they avoid imputation entirely. However their interpretability is significantly altered. In contrast, conditional-imputation methods excel in highly correlated datasets, even helping recover the interpretability of models trained on complete data. We have also found that missing value patterns and mechanisms are less critical in choice of model as most approaches perform equivalently in our study. Ultimately, we have shown that no single method is superior across all aspects, making the choice dependent on dataset characteristics and the trade-offs between accuracy, bias, and interpretability. This further

highlights the necessity of having clear guidelines to help chose the model best adapted for a health-related dataset in order to ensure reliability and trustworthiness of the developed applications. More so, focusing on the reliability rather than solely on performance metrics offers an informed approach when no significant difference can be found in the prediction performances of various models. Lastly, we have found that state-of-the-art models for handling missing values can be reliably leveraged to improve the performances of multimodal baselines on `StressID`.

PicMi: a robust method for handling missing values. Motivated by our findings, in Chapter 5 we introduced PicMi, an end-to-end *imputation-free* model designed for supervised learning with missing values. It uses a permutation-invariant architecture to handle inputs of varying sizes. By relaxing the requirement of fixed-dimensional datasets of traditional models, PicMi eliminates altogether the need to impute missing entries. It uses a conditional architecture to integrate the structure of the missing values pattern directly into its learning objective, making it robust to diverse missing data scenarios. Lastly, using attention-weights,it offers local interpretability, a highly desirable property in healthcare applications that is unique to our model. Through experiments, we demonstrated the advantages of our method on 11 health datasets.

However, while PicMi achieves good performances we have uncovered several limitations: it is not the best-suited choice for MAR scenarios with moderate amounts of missing values, where multiple imputation methods still outperform competitors in terms of prediction performances; and it performs optimally with high n/d ratios but is less effective with high-dimensional datasets. In addition, although we have proposed an extensive framework to compare PicMi to competitor methods, our study lacks an efficient method for the evaluation of the interpretability of our model. To the best of our knowledge, there is no efficient way identified in the literature to compare attention weights with feature importance rankings.

Nevertheless, our experiments on `StressID`, and 11 other health datasets, have shown that PicMi is a competitive alternative for handling missing data, that performs as well as, or outperforms state-of-the-art methods, while eliminating the limitations they suffer from. It does not rely on imputation, which is particularly relevant in sensitive fields such as healthcare, where using *fake* (i.e. imputed) data can raise concerns about the trustworthiness of algorithms. In addition, PicMi is locally interpretable, and provides weights for observed elements only. Additionally, we have shown that our approach is robust to both various missing data mechanisms, including in difficult MNAR settings, and high missing rates. These aspects are particularly important in practice: the underlying mechanism behind missing values in real-world data is often unknown and hard to determine; and existing studies (Shadbahr et al., 2023), as well as our analyses in Chapter 4 have shown the performance of the prediction models in impute-then-regress methods are highly affected by the percentage of missing values in the data. By focusing on challenges overlooked in

current research, our model advances towards establishing more trustworthy AI-systems for healthcare applications.

HyperMM: a robust method for handling missing modalities. In Chapter 6, we have introduced HyperMM, an end-to-end framework designed for MML with missing modalities without using reconstruction before training. Many existing solutions for handling missing modalities rely on complex, computationally costly modality reconstruction strategies. Instead, we have introduced a novel strategy for training a *universal* feature extractor using a conditional hypernetwork, and proposed a permutation-invariant neural network that can handle inputs of varying dimensions to process the extracted features, in a two-phase *task-agnostic* framework. Additionally, our approach is *model-agnostic* i.e. can be transposed to many applications by adapting the backbone architecture used for feature extraction. Through experiments, we highlighted the benefits of HyperMM on multiple medical imaging analysis applications.

Although HyperMM is designed for multimodal inputs, it currently focuses on inputs of the same nature (e.g. multimodal images only, or multimodal physiological signals only). To fully foster its potential, it is essential to extended the approach to the combination of different sources of data. This will enable us, in the future, to fully take advantage of rich multimodal datasets such as `StressID`. Moreover, the current approach to feature fusion relies on simple aggregation methods, such as summation or averaging. This approach is not optimal for leveraging the complementary information lying in multimodal inputs. Therefore, a more efficient method to integrate multiple modalities needs to be investigated in order to: maximize the use of the marginal information from each; and assign greater importance to the most relevant modalities when necessary.

Still, HyperMM has many advantages: unlike reconstruction-based methods, our approach is end-to-end and eliminates the time-consuming steps of manually reconstructing the missing modalities using a previously trained reconstruction model; as such, it significantly decreases model training time; and unlike competitors, its performances are not dependant on which modality is missing in the data. In addition, we have shown that the flexibility of HyperMM alleviates the constraints usually met in applications with varying-sized datasets and opens up a whole new range of possible learning strategies, beyond the scenario of missing modalities. As so, our approach represents a significant step forward in advancing the development of robust AI-driven multimodal healthcare applications. It represents a solid foundation for building solutions that can be effectively applied across diverse medical contexts and real-world scenarios.

7.2 Perspectives

While the works presented in this thesis have provided several insights into the development of robust and reliable AI systems for healthcare applications, they have also highlighted several follow-up questions that warrant further investigation. We conclude this thesis by outlining a handful of possible future directions that could follow this work.

Improving multimodal learning through intelligent feature fusion. To fully take advantage of the framework we have proposed in Chapter 6, we aim to extend it to MML with inputs of different types, combining images with text or time series for instance. As such, our next step is to adapt this framework for the analysis of `StressID`. A clear direction for this task is the use of modality-specific encoders, as done in numerous MML approaches (Han et al., 2019; Aguilar et al., 2019; Mordacq et al., 2024; Wang et al., 2023b; Zhang et al., 2023a). However, efficiently fusing features from modalities of different types into a single latent space remains a challenge.

In Chapter 3, we introduced early and late fusion MML models for the analysis of `StressID`. While they have yielded good performances for stress identification, MML models relying on mid-level fusion have shown more advantages in many studies (Baltrušaitis et al., 2018; Guarrasi et al., 2024). As mid-level fusion MML models are not inherently robust to missing modalities, we introduced a solution with HyperMM in Chapter 6. While effective, our current fusion approach relies on simple aggregation operations, due to the nature of the permutation-invariant architecture we have used. Moving forward, we plan to investigate more elaborate fusion techniques, beginning with the use of attention mechanisms to compute weighted aggregations, as explored in Chapter 5. Attention-based approaches have been widely used by researchers for this task (Pan and Wang, 2022; Mordacq et al., 2024; Chen et al., 2024). We have also identified other alternatives in the current literature, that represent interesting perspectives. For instance, Zadeh et al. (2017) proposed to reformulate the problem of multimodal analysis as one of modeling intra-modality and inter-modality dynamics end-to-end, using innovative tensor product operations. Other methods have focused on mutual information (MI) maximization. Han et al. (2021), for example, proposed hierarchically maximizing MI within unimodal input pairs; and between the multimodal fusion output and unimodal inputs. Doing so, they ensure that relevant information is preserved during multimodal fusion. Other works like Wang et al. (2022a); Shen et al. (2024) have also explored complementary information (CI) learning. It is particularly relevant in multimodal medical imaging, where clinicians rely on multiple imaging modalities for segmentation and diagnosis due to the limitations of individual modalities. CI learning can effectively help model and mitigate the negative impact of inter-modal redundancy – which can lead to issues such as misjudging modality importance or ignoring specific modal information.

In future work, our focus will be on developing a fusion method that remains invariant

to permutations, such that it can be integrate into HyperMM and preserve robustness to missing values.

Leveraging multimodal learning for enhanced unimodal predictions. As highlighted by Lu (2024), in practice a model trained on multiple modalities can outperform a finely-tuned unimodal model on unimodal tasks. In future work, we aim to build on this observation and explore if we can develop supervised MML models trained on complete multimodal datasets that remain robust to unimodal inputs at inference. Our goal will be to leverage the relationships learned between modalities during training, to allow richer and deeper information to be extracted from unimodal inputs at inference, thus enabling enhanced unimodal predictions benefiting from multimodal knowledge.

Such an approach requires effectively learning a function *bridging* the modalities (Lu, 2024). Several existing works offer promising directions for this objective. Han et al. (2019) proposed a joint training model that implicitly fuses information from multiple modalities in the training procedure by using one modality-specific network per individual modality and one shared network to map cues of each modality into final predictions. Doing so, they take advantage of multiple modalities to train models that perform well in unimodal scenarios. Similarly, Wang et al. (2023a) designed an approach to take advantage of all available input modalities during training and evaluation by learning shared and specific features to better represent the input data. Aguilar et al. (2019) developed a multimodal representation that captures relevant information from multiple modalities during training but operates with a single modality during inference through disjoint models. In addition, while the main focus is handling missing modalities, several works have proposed approaches to learn a rich multimodal latent space that aligns well with our goals, by doing the equivalent of implicit imputation of latent features. For instance, Zhang et al. (2022) have proposed imputing the information from missing modalities directly in the latent space by leveraging on the data from similar observations. Lastly, several works (Wang et al., 2020, 2023b) have proposed MML approaches based on distillation knowledge. In particular, Wang et al. (2023b) have proposed a cross modal knowledge distillation model to adaptively identify important modalities and distill knowledge from them to enhance other modalities roles, which could be beneficial for our objectives.

If successful, leveraging multimodal learning for enhancing unimodal predictions could significantly impact the democratization of healthcare. For instance, models trained on multimodal datasets combining clinical 12-lead ECGs with wearable device single-lead ECGs could learn a rich representation of the mutual. Once trained, such a model could detect complex cardiac issues using only single-lead ECGs from wearable devices, replacing expensive and less accessible clinical exams with cost-effective e-health solutions.

Ensuring generalization and fairness of e-health applications. Lastly, we plan to conduct further analyses on StressID with a focus on fairness and reliability. As discussed in Chapters 2 and 3, research datasets often fail to represent real-world populations accurately. Even carefully designed data collection processes are susceptible to suffer from representation bias and lack of heterogeneity. Ignoring this aspect leads to models that fail to generalize, neglect underrepresented groups, and produce unreliable scientific discoveries. As such, we intend to explore methods for developing heterogeneity-aware models.

A potential area of interest for this objective is domain generalization. There exists several algorithms designed to perform well on distributions different from those seen during training. For example, distributionally robust optimization (DRO) (Sagawa et al., 2019) performs empirical risk minimization that increases the importance of domains with larger errors, thus preparing models for *worst-case scenario* once deployed and applied on new data. Invariant risk minimization (IRM) (Arjovsky et al., 2019), learns an invariant feature representation across domain seen in training to ensure that models generalize well to new environments. While Gulrajani and Lopez-Paz (2020) have shown that there is currently no *one-size-fits-all* solution for domain generalization, they have proposed a thorough framework for evaluating existing approaches. It offers valuable insights and could be highly beneficial for healthcare applications, which are often based on human data with inherent variability.

By addressing the lack of heterogeneity in training data, we aim to improve fairness, reliability, and trustworthiness of AI systems. Advancing these aspects is critical for the real-world adoption and deployment of healthcare applications, as they rely on accurate and equitable performance across diverse populations.

Appendix A

StressID: A Multimodal Dataset for Stress Identification

A.1 Experimental Protocol

Figure A.1 shows the self-assessments questions as presented to the participants. Figure A.2 shows examples of tasks participants of `StressID` were asked to partake in. They show the instructions presented to the participants, and the set time limit.

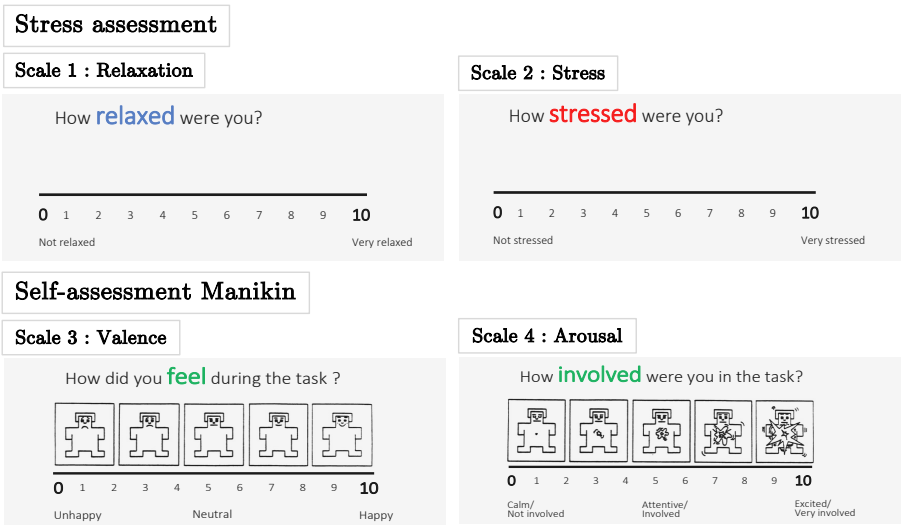


Figure A.1: Illustration of the four self-assessment questions used in `StressID`.

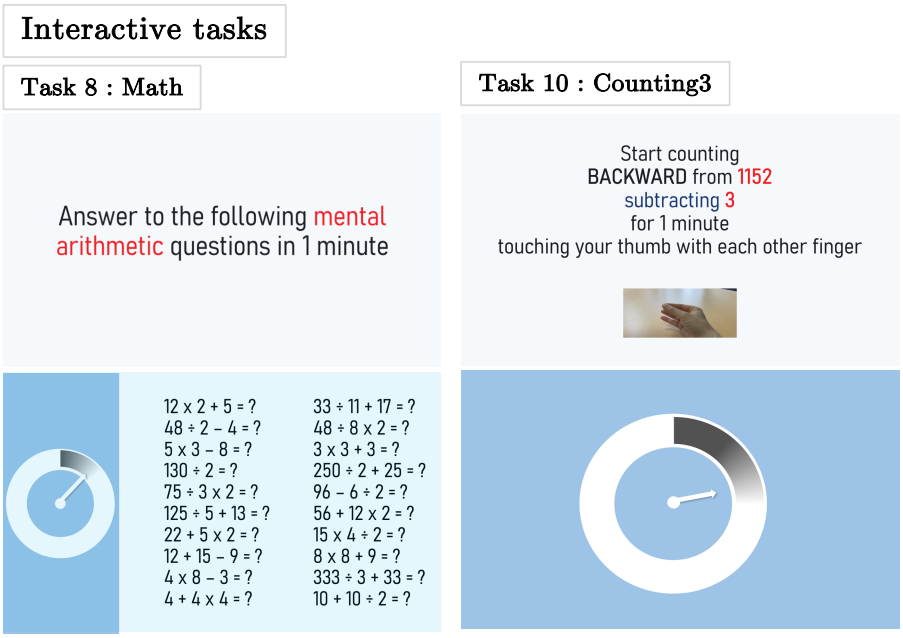


Figure A.2: Examples of tasks and stressors used in `StressID`.

A.2 Calibration and Synchronization of the Sensors

The wearable sensors are set-up first to enable good electrodes/skin interfacing, as the gel of the Ag/AgCl electrodes can take some minutes to correctly hydrate the skin. The BioSignalsPlux acquisition system is mounted with the ECG sensor, the EDA sensor, and the piezoelectric respiration belt. The experimenter starts by placing 3 Ag/AgCl electrodes on the ribcage of the subjects to capture the ECG signal, as the BioSignalsPlux ECG sensor is designed to record single lead ECG signals using 3 derivation configurations. Then, 2 Ag/AgCl electrodes are attached to the palm of the non-dominant hand of the subject to acquire the EDA signal. Finally, the experimenter helps the subjects put on the respiration chest-belt, and adjust it to their morphology – making sure the participants are as comfortable as possible wearing the sensors.

After setting up the electrodes, the device is connected to the OpenSignals (R)evolution platform for recording and streaming the physiological data, thus allowing the experimenter to observe a real-time reading of the signals. To ensure accurate and low-noise data, the experimenter checks the sensors' wires placement, as well as the posture and position of the subject before the start of the experiment. He adjusts and fixes the wires of the sensors using medical tape so that the presence of motion artifacts in the data during the collection is minimized.

Next, the Logitech QuickCam Pro 9000 RGB with integrated microphone is prepared. The camera is adjusted such that each subject is recorded in the middle of the frame with a neutral background. The participants sit approximately 50cm from the microphone. The start of the video/audio recording is marked on the OpenSignals (R)evolution platform using the event annotation plug-in.

Finally, once all devices are set up and the participants are installed, the experiment instructions are displayed on a screen placed in front of the participants. The beginning of the experiment is indicated by a beep sound. Another event annotation is added at the beep. This ensures the synchronization of the video, audio, and physiological signals for each task of the experiment.

A.3 Human Subject Considerations

The `StressID` project was approved by the Institutional Review Board (IRB) of Université Côte d'Azur, namely the Committee on Ethics for Non-Interventional Research (CERNI/CER). The project has been conducted under agreement n° 2021-033 for data collection, and n° 2023-016 for the publication of the dataset.

Safety risks included those associated with the wearable sensors used in `StressID`. Notably, the use of Ag/AgCl electrodes can cause discomfort or cutaneous irritations in subjects – however, using clinical grade electrodes during the data collection campaign, we did not encounter any issue of this type. In addition, the wearable devices used in `StressID` should not be used in patients with implanted electronic devices of any kind, including pacemakers, electronic infusion pumps, stimulators, defibrillators, or similar. All subjects were made aware of this fact, and could not participate in the experiment if they fell in any of the mentioned categories. The experiment presented no safety risks associated with tasks. Participants were informed they could stop the experiment at any time. The subjects were also informed that they can withdraw their consent at any time. In that case, the data collected prior to the creation of the database will be destroyed. If the database has already been created and the subjects have given consent to the use of physiological data or audio, as these are pseudo-anonymous, they cannot be deleted. Video data will not be shared with other people after the withdrawal request. However, data that has already been shared cannot be modified. Once the database has been shared with other authorized researchers, the subjects will no longer be able to exercise their right of withdrawal on that copy of the database.

Given the identifying nature of the videos, privacy was a primary concern in this project. Therefore, the data collection protocol of `StressID` considered the privacy risks for the participants as much as possible. Before the start of the experiment, they were introduced to the purpose and contents of the project, and public release modalities and privacy concerns were described. The participants explicitly consented to the recording of their

session, the dataset creation, and its release for research purposes following General Data Protection Rules (GDPR). The personal information (sex, age, education), and the acquired physiological and audio signals are pseudonymized, and an alphanumeric code is given for each participant. The goals and implications of publishing personally identifiable facial videos were clearly described to each participant, and a dedicated media release consent form was signed to acknowledge participants' willingness for their video to be part of the public release of the data. The participants could ultimately select between two options: **Option A:** research use and public release of all their recorded data, including identifying data (i.e. physiological, audio, and video). **Option B:** research use of all their recorded data, but no public release of identifying data (i.e. only physiological and audio data, but no video). The videos of the participants who selected option B were removed from the public version of the dataset. Among the 65 participants, 62 opted for option A and 3 opted for option B (2 women and 1 man). Although the participants explicitly consent to the recording of their session, the dataset creation, and its public release for research purposes, no attempts should be made to actively identify the subjects included in the dataset. The data should also not be modified or augmented in a way that further exposes the subjects' identities.

A.4 Ethical Considerations

In general, recording and usage of human activity data is associated with high ethical implications, including privacy, bias, and impact on society. If new projects use the `StressID` experimental protocol to replicate the study, using similar sensors and identifying modalities, the privacy of any new subjects should be protected, and the implications of the project clearly described to the participants. In addition, future applications that use the `StressID` protocol and/or dataset for building and training new learning pipelines, should consider the societal implications of their work. `StressID` is designed as a resource for improving the monitoring, modeling, and understanding of the mechanisms of human stress conditions. All intended applications have the potential to improve the quality of life of the population by helping prevent stress-related issues. However, researchers need to be aware of potential representation bias in their analyses. Indeed, `StressID` and subsequent analysis may present an imbalance in gender, race, age, or background of the participants – which could lead to unanticipated consequences. Additional information is provided about the participants' demographics along with the dataset and should be taken into account when developing new applications based on the `StressID` dataset.

We are aware that despite all the precautions, the dataset can be misused by bad-intentioned users. The authors declare that they bear all responsibility in case of any violation of rights during the collection of the data or other work, and will take appropriate action when needed, e.g., by removing data with such issues.

A.5 Dataset Accessibility

Given the identifying nature of the facial videos, the dataset is made accessible through open credentialized access only, for research purposes. Users are required to sign an end-user license agreement to request the data. Once validated, a link to the repository with a username and a password will be given to grant access. The`StressID` dataset represents 5.29 GB of data. It is hosted on Inria servers, using storage intended for long-term availability, and ensuring sufficient space to hold all collected data. This space is maintained by the INRIA infrastructure team. It is also easily accessible to the research team, allowing new data to be added as it is collected, or withdrawn if needed. This storage thus, allows the dataset to be both dynamic and persistent. The front-end website[1] describes the StressID project, access instructions for downloading the data, the adopted sensors, the recording framework, dataset composition details, and the baseline models. It is hosted on Inria servers intended for long-term persistent websites and also maintained by the infrastructure team. The website acts as a portal pointing to all relevant visualizations, data, code, and instructions. The code for the baselines and analyses uses an open-source 3-Clause BSD License Initiative. (2023), and is available on GitHub[2]. It includes `ReadMe` files describing the code structure, installation, and usage. In addition, third-party services for archival code repositories will be explored.

[1] `https://project.inria.fr/stressid/`
[2] `https://github.com/robustml-eurecom/stressID`

Appendix B

Stress Identification from Physiological Signals, Videos and Audio Data

B.1 Additional Experiments: Emotion Recognition

We report here additional experiments performed with binary labels extracted from the 4 self-assessments. We evaluate our learning pipeline on 4 binary classification tasks; namely discriminate between stressed (1) vs not stressed (0), relaxed (1) vs not relaxed (0), high valence (1) vs low valence (0), and high arousal (1) vs low arousal (0).

Each continuous value of the self-assessment is split as follows; if *value* is less than 5 then the label is 0, and if *value* is equal or greater than 5, then the label is 1. The created **stress** label is balanced and composed of 48% and 52% of class 0 and 1 respectively. Similarly, the **relax** label is composed of 54% and 46% of 0 and 1 respectively, and the **valence** label consists of 50% of each class. On the other hand, the **arousal** label is severely imbalanced and consists of 71% of high arousal (1) and 29% of low arousal (0).

The classification performances for all modalities and each label are reported in Table **??**. Our analysis confirms that the labels and the acquired data are coherent and meaningful, and the labels are predicted from the data with f1-scores well above the random.

Despite the different number of trials for each modality, some general observations can be highlighted. The valence appears here as the most difficult label to predict. This is especially true for audio and video, while physiological data seems to carry more useful information to discriminate between positive and negative valence. For the video, this can be related to the fact that a positive or negative valence in this set-up can be expressed with similar

Table B.1: Baseline f1-scores for different classification tasks. Each unimodal baseline is trained and tested on all available tasks of the corresponding modality (#tasks).

Data subset (#tasks)	**Binary stress**	**Binary relax**	**Binary arousal**	**Binary valence**
Physiological (711)	0.73 ± 0.04	0.67 ± 0.06	0.66 ± 0.06	0.64 ± 0.07
Video (587)	0.62 ± 0.04	0.62 ± 0.06	0.67 ± 0.10	0.54 ± 0.07
Audio-HC (385)	0.67 ± 0.04	0.62 ± 0.1	0.79 ± 0.09	0.55 ± 0.09

expressions. A person can smile because they are amused by the task or they can smile nervously. Recognizing a positive smile from a negative one is still a challenging task to this day in the field of emotion recognition.

On the other hand, the arousal is better predicted by the audio. This can be due to the fact that when people are more engaged in the task their tone of voice is incremented.

For the tasks of identifying stress and relaxation, the physiological signals appear as the most meaningful modality. Nonetheless, the results highlight good performances for all modalities, highlighting the strong correlations between the recorded data and the labels.

Appendix C

How to Handle Missing Values in Healthcare Data?

C.1 Decision Trees

We have trained multiple decision trees to retrieve interpretable decision rules for choosing a model based on its characteristics. Figures C.1, C.2 and C.3 visualizes the 3 best performing trees we have obtained, and used for further analysis.

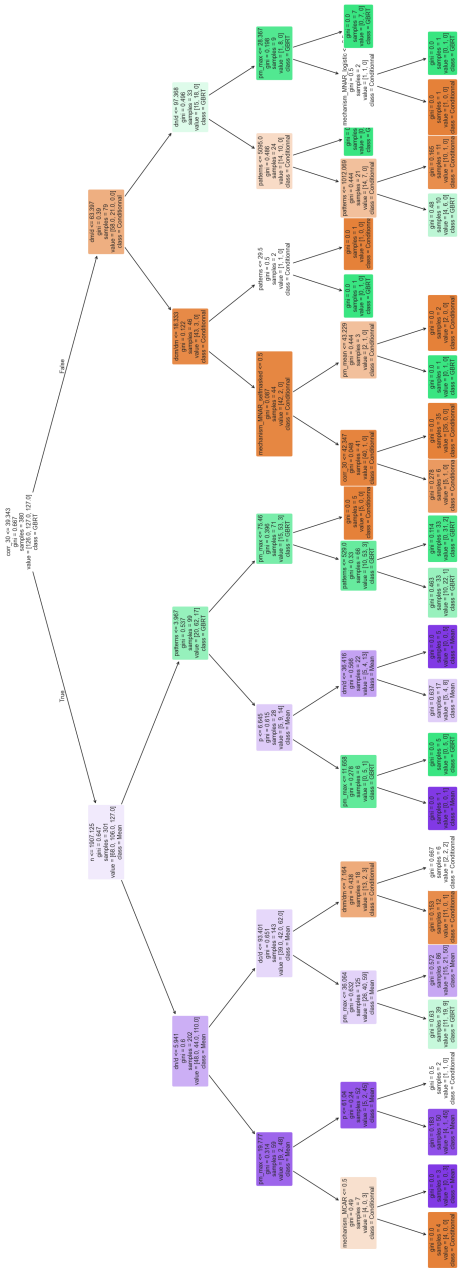


Figure C.1: Best performing decision tree.

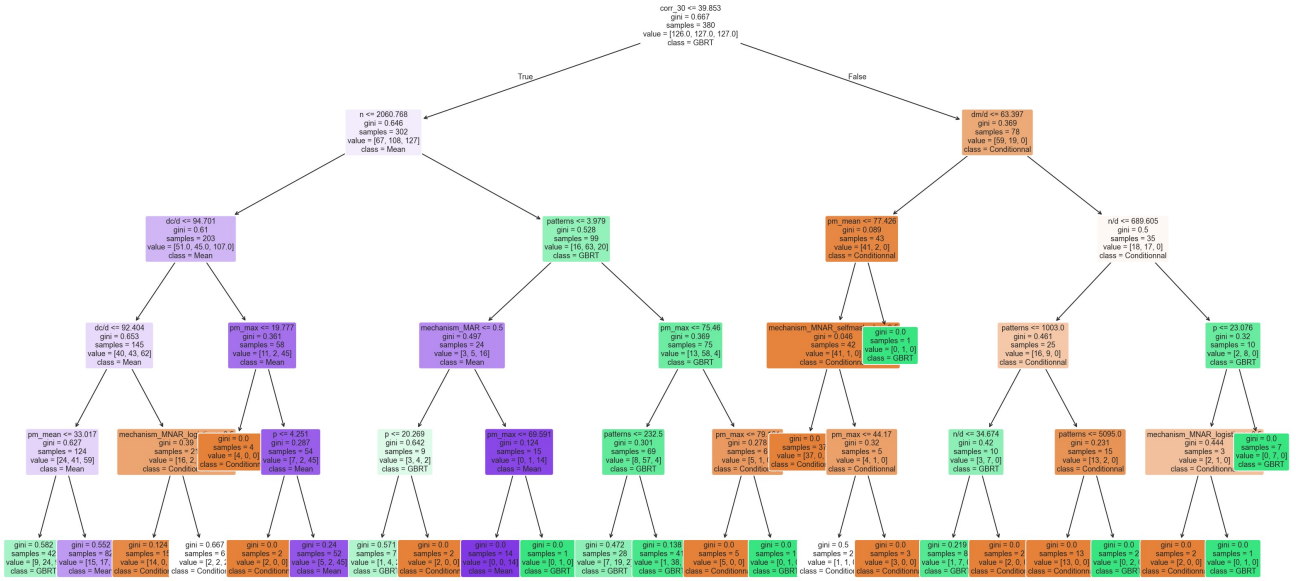


Figure C.2: Best performing decision tree.

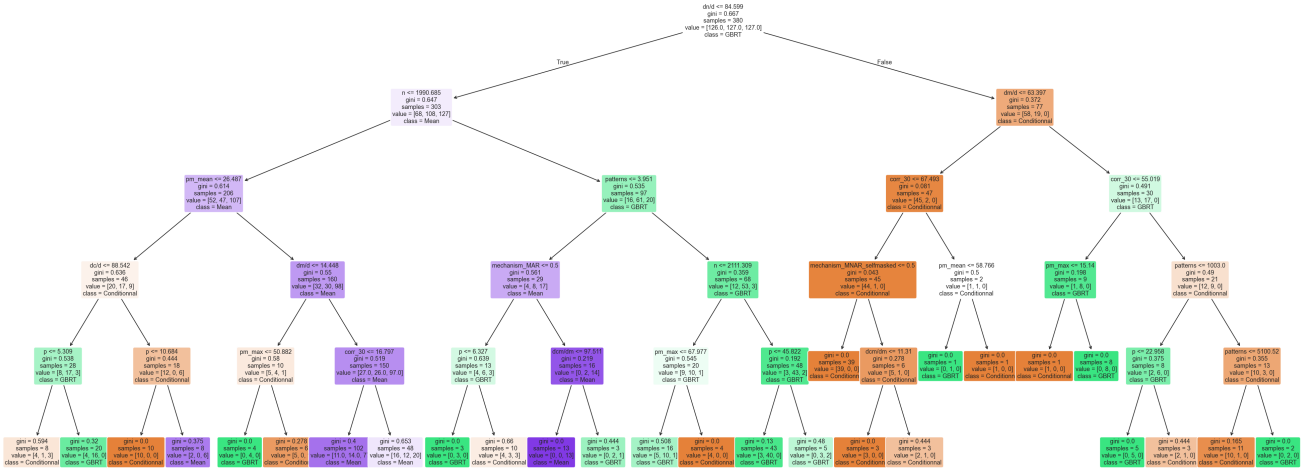


Figure C.3: Best performing decision tree.

Bibliography

A. Abd-Alrazaq, R. AlSaad, S. Aziz, A. Ahmed, K. Denecke, M. Househ, F. Farooq, and J. Sheikh. Wearable artificial intelligence for anxiety and depression: scoping review. *Journal of Medical Internet Research*, 25:e42672, 2023.

G. Aguilar, V. Rozgić, W. Wang, and C. Wang. Multimodal and multi-view models for emotion recognition. *arXiv preprint*, 2019.

A. Ahmed, S. Aziz, A. Abd-Alrazaq, F. Farooq, and J. Sheikh. Overview of artificial intelligence–driven wearable devices for diabetes: scoping review. *Journal of Medical Internet Research*, 24(8):e36010, 2022.

N. Ahmed, Z. Al Aghbari, and S. Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17:200171, 2023.

J. Aigrain, M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen, and M. Chetouani. Multimodal stress detection from multiple assessments. *IEEE Transactions on Affective Computing*, 9(4):491–506, 2016.

A. P. Allen, P. J. Kennedy, S. Dockray, J. F. Cryan, T. G. Dinan, and G. Clarke. The trier social stress test: principles and practice. *Neurobiology of stress*, 6:113–126, 2017.

A. Allik, G. Fazekas, and M. B. Sandler. An ontology for audio features. In *International Society for Music Information Retrieval Conference*, 2016.

K. M. Amekoe, M. D. Dilmi, H. Azzag, Z. C. Dagdia, M. Lebbah, and G. Jaffre. Tabsra: An attention based self-explainable model for tabular learning. In *ESANN 2023-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 199–204. Ciaco-i6doc. com, 2023.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

D. L. Armaignac, A. Saxena, M. Rubens, C. A. Valle, L.-M. S. Williams, E. Veledar, and L. T. Gidel. Impact of telemedicine on mortality, length of stay, and cost among patients in progressive care units: experience from a large healthcare system. *Critical care medicine*, 46(5):728–735, 2018.

F. S. Arsad, S. S. Syed Soffian, P. S. N. Megat Kamaruddin, N. R. Nordin, M. H. Baharudin, U. M. Baharudin, M. R. Hassan, A. Mohamed Nawi, and N. Ahmad. The impact of ehealth applications in healthcare intervention: a systematic review. *Journal of Health Research*, 37(3):178–189, 2023.

A. Arsalan, S. M. Anwar, and M. Majid. Mental stress detection using data from wearable and non-wearable sensors: A review. *CoRR*, abs/2202.03033, 2022a. URL `https://arxiv.org/abs/2202.03033`.

A. Arsalan, S. M. Anwar, and M. Majid. Mental stress detection using data from wearable and non-wearable sensors: a review. *arXiv preprint arXiv:2202.03033*, 2022b.

A. Ayme, C. Boyer, A. Dieuleveut, and E. Scornet. Near-optimal rate of consistency for linear models with missing values. In *International Conference on Machine Learning*, pages 1211–1243. PMLR, 2022.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

A. Bali and A. S. Jaggi. Clinical experimental stress studies: methods and assessment. *Reviews in the Neurosciences*, 26(5):555–579, 2015.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

J. Bandy and N. Vincent. Nutrition label template for dataset documentation. `https://www.overleaf.com/latex/templates/nutrition-label-template-for-dataset-documentation/gxzpbfmncyfp`, 2021.

D. Bertsimas, A. Delarue, and J. Pauphilet. Beyond impute-then-regress: Adapting prediction to missing data. *arXiv preprint arXiv:2104.03158*, 2021.

P. Boonyakitanont, A. Lek-Uthai, K. Chomtho, and J. Songsiri. A review of feature extraction and performance evaluation in epileptic seizure detection using eeg. *Biomedical Signal Processing and Control*, 57:101702, 2020.

M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1): 49–59, 1994.

J. D. Bremner. Traumatic stress: effects on the brain. *Dialogues in clinical neuroscience*, 2022.

V. K. Bürger, J. Amann, C. K. Bui, J. Fehr, and V. I. Madai. The unmet promise of trustworthy ai in healthcare: why we fail at clinical translation. *Frontiers in Digital Health*, 6:1279629, 2024.

L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166, 2018.

S. Campos, L. Pizarro, C. Valle, K. R. Gray, D. Rueckert, and H. Allende. Evaluating imputation techniques for missing data in adni: a patient classification study. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 20th Iberoamerican Congress, CIARP 2015, Montevideo, Uruguay, November 9-12, 2015, Proceedings 20*, pages 3–10. Springer, 2015.

F. Catania. Speech emotion recognition in italian using wav2vec 2.0 and the novel crowdsourced emotional speech corpus emozionalmente. 2023.

V. Chaparro, A. Gomez, A. Salgado, O. L. Quintero, N. Lopez, and L. F. Villa. Emotion recognition from eeg and facial expressions: a multimodal approach. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 530–533. IEEE, 2018.

H. Chaptoukaev, V. Strizhkova, M. Panariello, B. Dalpaos, A. Reka, V. Manera, S. Thümmler, E. Ismailova, M. Todisco, M. A. Zuluaga, et al. Stressid: a multimodal dataset for stress identification. *Advances in Neural Information Processing Systems*, 36:29798–29811, 2023.

H. Chaptoukaev, V. Marcianó, F. Galati, and M. A. Zuluaga. Hypermm: Robust multimodal learning with varying-sized inputs. *arXiv preprint arXiv:2407.20768*, 2024.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

B. Chen, J. Oliva, and M. Niethammer. A unified model for longitudinal multi-modal multi-view prediction with missingness. *arXiv preprint arXiv:2403.12211*, 2024.

J. Chen, C. Wang, M. Ester, Q. Shi, Y. Feng, and C. Chen. Social recommendation with missing not at random data. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 29–38. IEEE, 2018.

L.-W. Chen and A. Rudnicky. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

S.-Å. Christianson. Emotional stress and eyewitness memory: a critical review. *Psychological bulletin*, 112(2):284, 1992.

C. K. D. J. Clore, John and B. Strack. Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5230J.

M. Collier, A. Nazabal, and C. K. Williams. Vaes in the presence of missing data. *arXiv preprint arXiv:2006.05301*, 2020.

A. Craik, Y. He, and J. L. Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. *Advances in neural information processing systems*, 30, 2017.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

J. E. Dimsdale. Psychological stress and cardiovascular disease. *Journal of the American College of Cardiology*, 51(13):1237–1246, 2008.

A. M. D'Mello, I. R. Frosch, C. E. Li, A. L. Cardinaux, and J. D. Gabrieli. Exclusion of females in autism research: Empirical evidence for a "leaky" recruitment-to-research pipeline. *Autism Research*, 15(10):1929–1940, 2022.

D. Dua, C. Graff, et al. Uci machine learning repository. 2017.

A. A. Ein Shoka, M. M. Dessouky, A. El-Sayed, and E. E.-D. Hemdan. Eeg seizure detection: concepts, techniques, challenges, and future trends. *Multimedia Tools and Applications*, 82(27):42021–42051, 2023.

P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona. A survey on missing data in machine learning. *Journal of Big Data*, 8(1):1–37, 2021.

O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, 161:1–13, 2018.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

F. Galati, S. Ourselin, and M. A. Zuluaga. From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review. *Applied Sciences*, 12(8):3936, 2022.

M. Gallach, M. M. Lette, M. Abdel-Wahab, F. Giammarile, O. Pellet, and D. Paez. Addressing global inequities in positron emission tomography-computed tomography (pet-ct) for cancer management: a statistical model to guide strategic planning. *Medical science monitor: international medical journal of experimental and clinical research*, 26: e926544–1, 2020.

P. Garg, J. Santhosh, A. Dengel, and S. Ishimaru. Stress detection by machine learning and wearable sensors. In *Companion Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 43–45, 2021.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. I. au2, and K. Crawford. Datasheets for datasets. `https://arxiv.org/abs/1803.09010`, 2021.

S. Gedam and S. Paul. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access*, 9:84045–84066, 2021.

G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460, 2019.

G. Giannakakis, M. R. Koujan, A. Roussos, and K. Marias. Automatic stress detection evaluating models of facial action units. In *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)*, pages 728–733. IEEE, 2020.

A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

S. V. R. D. S. P. N. S. O. Y. Golovenkin, S.E. and V. Voino-Yasenetsky. Myocardial infarction complications. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C53P5M.

L. Gondara and K. Wang. Mida: Multiple imputation using denoising autoencoders. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*, pages 260–272. Springer, 2018.

Y. Gong, H. Hajimirsadeghi, J. He, T. Durand, and G. Mori. Variational selective autoencoder: Learning from partially-observed heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 2377–2385. PMLR, 2021.

A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*, 2022.

A. Gros, E. Chapoulie, R. Ramadour, V. Robert, J. de Stoutz, S. Guetin, D. Chaïma, E. Wyckaert, V. Manera, P. Robert, et al. Rel@ x: Sensory and virtual immersion to reduce the anxiety of patients consulting for the first time in nice memory center. *Alzheimer's and Dementia*, 13(7):P609–P610, 2017.

V. Guarrasi, F. Aksu, C. M. Caruso, F. Di Feola, A. Rofena, F. Ruffini, and P. Soda. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *arXiv preprint arXiv:2408.02686*, 2024.

I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

D. Ha, A. M. Dai, and Q. V. Le. Hypernetworks. *CoRR*, 2016.

A. Haleem, M. Javaid, and I. H. Khan. Current status and applications of artificial intelligence (ai) in medical field: An overview. *Current Medicine Research and Practice*, 9 (6):231–237, 2019.

J. Han, Z. Zhang, Z. Ren, and B. Schuller. Implicit fusion by joint audiovisual training for emotion recognition in mono modality. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5861–5865, 2019.

W. Han, H. Chen, and S. Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*, 2021.

S. G. Hart. Nasa task load index (tlx). 1986.

J. A. Healey. *Wearable and automotive systems for affect recognition from physiology*. PhD thesis, Massachusetts Institute of Technology, 2000.

M. Herdin, N. Czink, H. Ozcelik, and E. Bonek. Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels. In *2005 IEEE 61st Vehicular Technology Conference*, volume 1, pages 136–140. IEEE, 2005.

M. Horn, M. Moor, C. Bock, B. Rieck, and K. Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, pages 4353–4363. PMLR, 2020.

N. Hou, M. Li, L. He, B. Xie, L. Wang, R. Zhang, Y. Yu, X. Sun, Z. Pan, and K. Wang. Predicting 30-days mortality for mimic-iii patients with sepsis-3: a machine learning approach using xgboost. *Journal of translational medicine*, 18:1–14, 2020.

Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

S. Huhn, M. Axt, H.-C. Gunga, M. A. Maggioni, S. Munga, D. Obor, A. Sié, V. Boudo, A. Bunker, R. Sauerborn, et al. The impact of wearable technologies in health research: scoping review. *JMIR mHealth and uHealth*, 10(1):e34384, 2022.

O. S. Initiative. The 3-Clause BSD License. `https://opensource.org/license/bsd-3-clause/`, 2023.

N. B. Ipsen, P. Mattei, and J. Frellsen. not-MIWAE: Deep generative modelling with missing not at random data. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. URL `https://openreview.net/forum?id=tu29GQT0JFy`.

N. B. Ipsen, P.-A. Mattei, and J. Frellsen. How to deal with missing data in supervised deep learning? In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

M. Jafari, A. Shoeibi, M. Khodatars, S. Bagherzadeh, A. Shalbaf, D. L. García, J. M. Gorriz, and U. R. Acharya. Emotion recognition in eeg signals using deep learning methods: A review. *Computers in Biology and Medicine*, page 107450, 2023.

S. Jäger, A. Allhorn, and F. Bießmann. A benchmark for data imputation methods. *Frontiers in big Data*, 4:693674, 2021.

M. Jaiswal, C.-P. Bara, Y. Luo, M. Burzo, R. Mihalcea, and E. M. Provost. Muse: a multimodal dataset of stressed emotion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1499–1510, 2020.

J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, 17:1–10, 2017.

A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. Heart Disease. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C52P4X.

A. Jaratrotkamjorn and A. Choksuriwong. Bimodal emotion recognition using deep belief network. In *2019 23rd International Computer Science and Engineering Conference (ICSEC)*, pages 103–109. IEEE, 2019.

H. Jeong, H. Wang, and F. P. Calmon. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9558–9566, 2022.

A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

M. P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230, 1996.

J. Josse and J. P. Reiter. Introduction to the special section on missing data. *Statistical Science*, 33(2):139–141, 2018.

J. Josse, N. Prost, E. Scornet, and G. Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

T.-P. Jung, T. J. Sejnowski, et al. Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. *IEEE Transactions on Affective Computing*, 13(1): 96–107, 2019.

H. Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402–406, 2013.

A. Kapelner and J. Bleich. Prediction with missing data via bayesian additive regression trees. *Canadian Journal of Statistics*, 43(2):224–239, 2015.

Y. V. Karpievitch, A. R. Dabney, and R. D. Smith. Normalization and missing value imputation for label-free lc-ms analysis. *BMC bioinformatics*, 13:1–9, 2012.

T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry investigation*, 15(3): 235, 2018.

J.-C. Kim and K. Chung. Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. *IEEE Access*, 8:104933–104943, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40, 2010. doi: 10.1016/j.specom.2009.08.009.

S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij. The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction*, pages 291–298, 2014.

P. Kuppens, F. Tuerlinckx, J. A. Russell, and L. F. Barrett. The relation between valence and arousal in subjective experience. *Psychological bulletin*, 139(4):917, 2013.

K. Läll, M. Lepamets, M. Palover, T. Esko, A. Metspalu, N. Tõnisson, P. Padrik, R. Mägi, and K. Fischer. Polygenic prediction of breast cancer: comparison of genetic predictors and implications for risk stratification. *BMC cancer*, 19:1–9, 2019.

L. Lapointe, M.-H. Lavallee-Bourget, A. Pichard-Jolicoeur, C. Turgeon-Pelchat, and R. Fleet. Impact of telemedicine on diagnosis, clinical management and outcomes in rural trauma patients: a rapid review. *Canadian Journal of Rural Medicine*, 25(1):31–40, 2020.

M. Le Morvan, J. Josse, T. Moreau, E. Scornet, and G. Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. *Advances in Neural Information Processing Systems*, 33:5980–5990, 2020a.

M. Le Morvan, N. Prost, J. Josse, E. Scornet, and G. Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR, 2020b.

M. Le Morvan, J. Josse, E. Scornet, and G. Varoquaux. What's a good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34: 11530–11540, 2021.

E.-H. Lee. Review of the psychometric evidence of the perceived stress scale. *Asian nursing research*, 6(4):121–127, 2012.

F. Li, H. Xin, J. Zhang, M. Fu, J. Zhou, and Z. Lian. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the mimic-iii database. *BMJ open*, 11(7):e044779, 2021.

G. Liang, X. Xing, L. Liu, Y. Zhang, Q. Ying, A.-L. Lin, and N. Jacobs. Alzheimer's disease classification using 2d convolutional neural networks. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3008–3012, 2021.

W.-C. Lin and C.-F. Tsai. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2):1487–1509, 2020.

R. J. Little. Regression with missing x's: a review. *Journal of the American statistical association*, 87(420):1227–1237, 1992.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Z. Lu. A theory of multimodal learning. *Advances in Neural Information Processing Systems*, 36, 2024.

S. Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

C. Ma, W. Gong, J. M. Hernández-Lobato, N. Koenigstein, S. Nowozin, and C. Zhang. Partial vae for hybrid recommender system. In *NIPS Workshop on Bayesian Deep Learning*, volume 2018, 2018a.

C. Ma, S. Tschiatschek, K. Palla, J. M. Hernández-Lobato, S. Nowozin, and C. Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*, 2018b.

D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. A. Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior research methods*, pages 1–8, 2021.

S. Malla, A. Alsadoon, and S. K. Bajaj. A dfc taxonomy of speech emotion recognition based on convolutional neural network from speech signal. In *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, pages 1–10, 2020. doi: 10.1109/CITISIA50690.2020.9371841.

S. Mao and E. Sejdić. A review of recurrent neural network-based methods in computational physiology. *IEEE transactions on neural networks and learning systems*, 34(10):6983–7003, 2022.

V. Markova, T. Ganchev, and K. Kalinkov. Clas: A database for cognitive load, affect and stress recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*, pages 1–4. IEEE, 2019.

P.-A. Mattei and J. Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.

I. Mayer, A. Sportisse, J. Josse, N. Tierney, and N. Vialaneix. R-miss-tastic: a unified platform for missing values methods and workflows. *arXiv preprint arXiv:1908.04822*, 2019.

M. S. Mayouf and F. Dupin de Saint-Cyr. Curriculum incremental deep learning on breakhis dataset. In *Proceedings of the 2022 8th International Conference on Computer Technology Applications*, pages 35–41, 2022.

B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

M. D. McManus, J. T. Siegel, and J. Nakamura. The predictive power of low-arousal positive affect. *Motivation and Emotion*, 43:130–144, 2019.

E. Merdjanovska and A. Rashkovska. Comprehensive survey of computational ecg analysis: Databases, methods and applications. *Expert Systems with Applications*, 203:117206, 2022.

A. I. Middya, B. Nag, and S. Roy. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems*, 244:108580, 2022.

H. Y. Mir and O. Singh. Ecg denoising and feature extraction techniques–a review. *Journal of medical engineering & technology*, 45(8):672–684, 2021.

J. Mordacq, L. Milecki, M. Vakalopoulou, S. Oudot, and V. Kalogeiton. Adapt: Multimodal learning for detecting physiological changes under missing modalities. In *MIDL 2024-Medical Imaging with Deep Learning*, 2024.

S. Mousavi, F. Afghah, A. Razi, and U. R. Acharya. Ecgnet: Learning where to attend for detection of atrial fibrillation with deep visual attention. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE, 2019.

S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

B. Muzellec, J. Josse, C. Boyer, and M. Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR, 2020.

A. Nagappan, A. Krasniansky, and M. Knowles. Patterns of ownership and usage of wearable devices in the united states, 2020-2022: Survey study. *Journal of Medical Internet Research*, 26:e56504, 2024.

D. Nahavandi, R. Alizadehsani, A. Khosravi, and U. R. Acharya. Application of artificial intelligence in wearable devices: Opportunities and challenges. *Computer Methods and Programs in Biomedicine*, 213:106541, 2022.

A. Nazábal, P. M. Olmos, Z. Ghahramani, and I. Valera. Handling incomplete heterogeneous data using vaes. *CoRR*, abs/1807.03653, 2018. URL `http://arxiv.org/abs/1807.03653`.

Neha, H. Sardana, R. Kanwade, and S. Tewary. Arrhythmia detection and classification using ecg and ppg techniques: A review. *Physical and Engineering Sciences in Medicine*, 44(4):1027–1048, 2021.

S. Nijman, A. Leeuwenberg, I. Beekers, I. Verkouter, J. Jacobs, M. Bots, F. Asselbergs, K. Moons, and T. Debray. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of clinical epidemiology*, 142:218–229, 2022.

J. Pan and S. Wang. Cross-modal transformer gan: A brain structure-function deep fusing framework for alzheimer's disease. *arXiv preprint arXiv:2206.13393*, 2022.

S. Parthasarathy and S. Sundaram. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404, 2020.

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

G. Peeters. Rhythm Classification using spectral rhythm patterns. In *ISMIR*, pages –, London, United Kingdom, Sept. 2005. URL `https://hal.science/hal-01106133`. cote interne IRCAM: Peeters05b.

J. Pereira and F. Saraiva. A comparative analysis of unbalanced data handling techniques for machine learning algorithms to electricity theft detection. In *2020 IEEE congress on evolutionary computation (CEC)*, pages 1–8. IEEE, 2020.

E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

A. Perez-Lebel, G. Varoquaux, M. Le Morvan, J. Josse, and J.-B. Poline. Benchmarking missing-values approaches for predictive models on health databases. gigascience 11, giac013, 2022.

N. Perveen and C. K. Mohan. Configural representation of facial action units for spontaneous facial expression recognition in the wild. In *VISIGRAPP (4: VISAPP)*, pages 93–102, 2020.

A.-C. Pinho-Gomes, J. Gong, K. Harris, M. Woodward, and C. Carcel. Dementia clinical trials over the past decade: are women fairly represented? *BMJ Neurology Open*, 4(2), 2022.

L. Piwek, D. A. Ellis, S. Andrews, and A. Joinson. The rise of consumer health wearables: promises and barriers. *PLoS medicine*, 13(2):e1001953, 2016.

Electrocardiography (ECG) Sensor User Manual. PLUX wireless biosignals S.A., 2020. URL `https://support.pluxbiosignals.com/wp-content/uploads/2021/10/biosignalsplux-Electrocardiography-ECG-User-Manual.pdf`.

Electrodermal Activity (EDA) Sensor Datasheet. PLUX wireless biosignals S.A., 2021a. URL `https://support.pluxbiosignals.com/wp-content/uploads/2021/11/Electrodermal_Activity_EDA_Datasheet.pdf`.

Respiration (PZT) Sensor User Manual. PLUX wireless biosignals S.A., 2021b. URL `https://support.pluxbiosignals.com/wp-content/uploads/2021/11/Respiration_PZT_User_Manual.pdf`.

C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

A.-u. Rahman, R. N. Asif, K. Sultan, S. A. Alsaif, S. Abbas, M. A. Khan, and A. Mosavi. Ecg classification for detecting ecg arrhythmia empowered with deep learning approaches. *Computational intelligence and neuroscience*, 2022(1):6852845, 2022.

K. P. Rao, M. C. S. Rao, and N. H. Chowdary. An integrated approach to emotion recognition and gender classification. *Journal of Visual Communication and Image Representation*, 60:339–345, 2019.

P. Razavi, M. T. Chang, G. Xu, C. Bandlamudi, D. S. Ross, N. Vasan, Y. Cai, C. M. Bielski, M. T. Donoghue, P. Jonsson, et al. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer cell*, 34(3):427–438, 2018.

B. Rim, N.-J. Sung, S. Min, and M. Hong. Deep learning in physiological signal data: A survey. *Sensors*, 20(4):969, 2020.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161, 1980.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

M. Sahidullah, T. Kinnunen, and C. Hanilçi. A comparison of features for synthetic speech detection. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 2087–2091. ISCA, 2015. URL `http://www.isca-speech.org/archive/interspeech_2015/i15_2087.html`.

R. Sánchez-Reolid, M. T. López, and A. Fernández-Caballero. Machine learning for stress detection from electrodermal activity: A scoping review. 2020.

A. Schaefer, F. Nils, X. Sanchez, and P. Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and emotion*, 24(7):1153–1172, 2010.

P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition, 2019.

T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, M. Thorpe, R. V. Torné, E. Sala, P. Lió, et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3(1):139, 2023.

M. M. H. Shandhi, K. Singh, N. Janson, P. Ashar, G. Singh, B. Lu, D. S. Hillygus, J. M. Maddocks, and J. P. Dunn. Assessment of ownership of smart devices and the acceptability of digital health data sharing. *NPJ Digital Medicine*, 7(1):44, 2024.

M. Sharma. Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6907–6911. IEEE, 2022.

C. Shen, W. Li, H. Chen, X. Wang, F. Zhu, Y. Li, X. Wang, and B. Jin. Complementary information mutual learning for multimodality medical image segmentation. *arXiv preprint arXiv:2401.02717*, 2024.

R. Shwartz-Ziv and A. Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

A. K. Singh and S. Krishnan. Ecg signal feature extraction trends in methods and applications. *BioMedical Engineering OnLine*, 22(1):22, 2023.

J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association, 1988.

F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7): 1455–1462, 2015.

M. Sperrin, G. P. Martin, R. Sisk, and N. Peek. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of clinical epidemiology*, 125:183–187, 2020.

C. D. Spielberger, F. Gonzalez-Reigosa, A. Martinez-Urrutia, L. F. Natalicio, and D. S. Natalicio. The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican journal of psychology*, 5(3 & 4), 1971.

H. J. Steeneken and J. H. Hansen. Speech under stress conditions: overview of the effect on speech production and on system performance. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 4, pages 2079–2082. IEEE, 1999.

D. J. Stekhoven and P. Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

V. Strizhkova, H. Kachmar, H. Chaptoukaev, R. Kalandadze, N. Kukhilava, T. Tsmindashvili, N. Abo-Alzahab, M. A. Zuluaga, M. Balazia, A. Dantcheva, et al. Mvp: Multimodal emotion recognition based on video and physiological signals. *arXiv preprint arXiv:2501.03103*, 2025.

N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

J. R. Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779, 2015.

W. Sun, F. Ma, Y. Li, S.-L. Huang, S. Ni, and L. Zhang. Semi-supervised multimodal image translation for missing modality imputation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4320–4324, 2021.

Z. Sun, M. Lin, Q. Zhu, Q. Xie, F. Wang, Z. Lu, and Y. Peng. A scoping review on multimodal deep learning in biomedical images and texts. *Journal of Biomedical Informatics*, page 104482, 2023.

Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, and A. Zhang. Metric learning on healthcare data with incomplete modalities. In *IJCAI*, volume 3534, page 3540, 2019.

S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis. A multimodal dataset for various forms of distracted driving. *Scientific data*, 4(1):1–21, 2017.

S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, P. Skudlarski, E. Cavedo, G. B. Frisoni, W. Hoffmann, et al. Multimodal imaging in alzheimer's disease: validity and usefulness for early detection. *The Lancet Neurology*, 14(10):1037–1053, 2015.

C. M. Tempany, J. Jayender, T. Kapur, R. Bueno, A. Golby, N. Agar, and F. A. Jolesz. Multimodal imaging for improved diagnosis and treatment of cancers. *Cancer*, 121(6): 817–827, 2015.

M. Thurow, F. Dumpert, B. Ramosaj, and M. Pauly. Imputing missings in official statistics for general tasks–our vote for distributional accuracy. *Statistical Journal of the IAOS*, 37 (4):1379–1390, 2021.

D. S. Ting, Y. Liu, P. Burlina, X. Xu, N. M. Bressler, and T. Y. Wong. Ai for medical imaging goes deep. *Nature medicine*, 24(5):539–540, 2018.

T.-D. Tran, J. Kim, N.-H. Ho, H.-J. Yang, S. Pant, S.-H. Kim, and G.-S. Lee. Stress analysis with dimensions of valence and arousal in the wild. *Applied Sciences*, 11(11):5194, 2021.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

B. E. Twala, M. Jones, and D. J. Hand. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956, 2008.

F. UNESCO Institute for Statistics. Women in science: Fact sheet no. 55. `https://uis.unesco.org/sites/default/files/documents/fs55-women-in-science-2019-en.pdf`, 2019. Accessed: 2023-10-19.

J. V. Vaghasiya, C. C. Mayorga-Martinez, and M. Pumera. Wearable sensors for telehealth based on emerging materials and nanoarchitectonics. *npj Flexible Electronics*, 7(1):26, 2023.

S. Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

G. Vos, K. Trinh, Z. Sarnyai, and M. R. Azghadi. Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review. *International Journal of Medical Informatics*, 173:105026, 2023.

E. Wagstaff, F. Fuchs, M. Engelcke, I. Posner, and M. A. Osborne. On the limitations of representing functions on sets. In *International Conference on Machine Learning*, pages 6487–6494. PMLR, 2019.

Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu. A review on transfer learning in eeg signal analysis. *Neurocomputing*, 421:1–14, 2021.

D. Wang, T. Zhao, W. Yu, N. V. Chawla, and M. Jiang. Deep multimodal complementarity learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10213–10224, 2022a.

H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023a.

H. Wang, C. Ma, J. Zhang, Y. Zhang, J. Avery, L. Hull, and G. Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 216–226. Springer, 2023b.

J. Wang and M. Wang. Review of the emotional feature extraction and classification using eeg signals. *Cognitive robotics*, 1:29–40, 2021.

J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and S. Y. Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022b.

Q. Wang, L. Zhan, P. Thompson, and J. Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838, 2020.

D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

K. Weimann and T. O. Conrad. Transfer learning for ecg classification. *Scientific reports*, 11(1):5251, 2021.

E. Widen, T. Raben, L. Lello, and S. Hsu. Machine learning prediction of biomarkers from snps and of disease risk from biomarkers in the uk biobank. genes 12. issn: 2073-4425, 2021.

M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

R. Wu, H. Wang, H.-T. Chen, and G. Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

Y. Xie, L. Lu, F. Gao, S.-j. He, H.-j. Zhao, Y. Fang, J.-m. Yang, Y. An, Z.-w. Ye, and Z. Dong. Integration of artificial intelligence, blockchain, and wearable technology for chronic disease management: a new paradigm in smart healthcare. *Current Medical Science*, 41(6):1123–1133, 2021.

F. Xu and Z. Wang. Emotion recognition research based on integration of facial expression and voice. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE, 2018.

P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

L. Yang, Y. Cui, Y. Xuan, C. Wang, S. Belongie, and D. Estrin. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM conference on recommender systems*, pages 279–287, 2018.

Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, pages 1–11, 2024.

J. Yoon, J. Jordon, and M. Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, J. Wang, and J. Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2418–2428, 2022.

J. Zhang and Z.-m. Zhang. Ethics and governance of trustworthy medical artificial intelligence. *BMC medical informatics and decision making*, 23(1):7, 2023.

Y. Zhang, H. Doughty, and C. Snoek. Learning unseen modality interaction. *Advances in Neural Information Processing Systems*, 36:54716–54726, 2023a.

Y. Zhang, C. Peng, Q. Wang, D. Song, K. Li, and S. K. Zhou. Unified multi-modal image synthesis for missing modality imputation. *arXiv preprint arXiv:2304.05340*, 2023b.

Q. Zhou, H. Zou, H. Jiang, and Y. Wang. Incomplete multimodal learning for visual acuity prediction after cataract surgery using masked self-attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 735–744. Springer, 2023.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.