UNIVERSITÄT ZU LÜBECK

From the Institute of Medical Informatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. habil. Heinz Handels

Multimodal Sensor Data Analysis for the Investigation of Physical and Mental States Using Machine Learning

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences and Technical Engineering

Submitted by

Muhammad Tausif Irshad
from Rahim Yar Khan, Pakistan

Lübeck 2023

First referee: Prof. Dr.-Ing. Marcin Grzegorzek

Second referee: Prof. Dr. phil. André Calero Valdez

Date of oral examination: 09.04.2024

Approved for printing: 11.04.2024, Lübeck

Acknowledgements

This dissertation is the culmination of several years of research work carried out in both the (no longer existing) Research Group for Pattern Recognition of the University of Siegen (Germany) and the Medical Data Science Team of the University of Lübeck (Germany). This journey would not have been possible without the support of many people, whom I would like to thank here.

First of all, I would like to thank my PhD supervisor Prof. Dr.-Ing. habil. Marcin Grzegorzek, who accepted and integrated me into his research group(s) both in Siegen and Lübeck. He has guided me with his advice through my research and my career. I am impressed not only by the expertise I learned from him but also by his optimistic thinking and worldview. Whenever I encountered difficulties or had questions about my research or writing, he always provided me with the right guidance, unwavering support, and tireless care. Without his guidance, I would not have been able to complete my dissertation.

I would like to thank all my colleagues who have accompanied me on this journey, both in Siegen and in Lübeck. An effective working environment for research and the exchange of ideas cannot be achieved without everyone's contribution. I am delighted to have had the opportunity to get to know you, and I appreciate the collaboration with you very much. In particular, I would like to thank Dr.-Ing. Xinyu Huang and Dr.-Ing. Muhammad Adeel Nisar, who provided scientific guidance and supervision for my research.

Finally, I would like to express my special gratitude to my family. This work would not have been possible without their patience and support during this long and not-always-easy period of time.

Abstract

Health plays an important role in our lives, which determines the peculiarities of modern society. It is considered to be the most significant social wealth, and it can be divided into physical and mental health. Physical and mental health conditions are intrinsically linked, and the presence of a physical health condition increases the likelihood of developing a mental health issue and vice versa. When such health conditions are not adequately treated, they affect the whole healthcare system — from primary care to hospital care. However, with the advancement in sensor technology, machine learning algorithms, and the computational power of machines, it has become possible to deploy machine learning-based systems to diagnose such health conditions or to assist doctors or other health professionals in making their decisions. Nevertheless, some physical and mental states are still not objectively recognized with non-invasive physiological measurements using machine learning approaches, such as hunger and satiety and human flow experience states. In this context, this thesis presents state-of-the-art machine learning-based approaches to recognizing those aforementioned health states using physiological time-series data acquired from non-invasive wearable sensor technology.

The first part of this thesis (Chapter 2) focuses on hunger and satiety state recognition. The perception of hunger and satiety is crucial in maintaining a healthy body weight and avoiding chronic diseases such as obesity, underweight, or deficiency syndromes due to malnutrition. A chronic loss of this perception characterizes numerous health conditions. Based on the literature highlights, these conditions are challenging to detect using non-invasive measurements. To fulfill this objective and obtain a highly precise system, this chapter presents a hierarchical framework that leverages signal processing and pattern recognition techniques to detect hunger and satiety non-invasively. Three devices, such as an *Empatica E4* wristband, a *Biosignalplux RespiBAN* wearable, and *JINS MEME* smart glasses, are used to capture physiological signals from five healthy, normal-weight subjects inactively sitting on a chair in a state of hunger and satiety. The investigations of this study turned to fill the scientific gap by proposing a non-invasive multimodal system capable of remarkably discriminating between hunger and satiety states. In addition, this study lays the grounds for the following future research direction. First, It is feasible to recognize these states through non-invasive physiological measurements. Second, the most discriminative features come from three specific sensor modalities: *Electrodermal activity* (EDA), *infrared Thermopile* (Tmp),

and *Blood Volume Pulse* (BVP). These sensor modalities are part of the Empatica E4 wristband, the most influential device in this study, that can be used as a standalone wearable for non-invasive hunger and satiety recognition in the future. Third, feature learning approaches do not necessarily perform well, particularly in the case of limited training examples. Last, the *Random Forest* (RF) classifier is the most reliable for such hunger state recognition and feature selection.

In the second part of this thesis (Chapter 3), a novel multimodal system is presented to identify the individual's flow experience during work activities. Flow experience is a specific positive and affective state of mind that occurs when humans are completely absorbed in an activity and forget everything else. This state can lead to high performance, well-being, and productivity at work. Few studies have been conducted to recognize the human flow experience using physiological wearable devices; however, these studies lack objectiveness from the implementation point of view. Moreover, studies have yet to explore how to address the data scarcity problem in this domain and how to use emotional data to enhance flow recognition performance. In this study, physiological data is collected from 25 subjects with multimodal sensing devices — the *Empatica E4* wristband, the *Emotiv Epoc X* Electroencephalography (EEG) headset, and the *Biosignalplux RespiBAN* wearable when participants were doing arithmetic and reading tasks. Experiments are performed to probe the discrimination between flow and non-flow states using feature engineering and deep feature learning approaches. Experiments are also conducted to investigate the connection between emotions and flow experience by testing transfer learning techniques involving an emotion recognition-related task on the source domain. The results of this study suggest that effective and objective discrimination between flow and non-flow states is possible with multimodal sensor data. The success of transfer learning using the *DEAP* emotion dataset as a source domain indicates that emotion-related dimensions, particularly arousal, and flow, are connected and that emotion recognition can be used as a latent task to enhance the performances of flow detection. This finding could help to circumvent the data scarcity problem in this domain.

Zusammenfassung

Gesundheit spielt eine wichtige Rolle in unserem Leben, die die Besonderheiten der modernen Gesellschaft bestimmt. Sie gilt als wichtigstes soziales Gut und kann in körperliche und geistige Gesundheit unterteilt werden, welche untrennbar miteinander verbunden sind. Das Vorhandensein einer körperlichen Erkrankung erhöht die Wahrscheinlichkeit der Entwicklung eines psychischen Problems und umgekehrt. Wenn solche Gesundheitszustände nicht angemessen behandelt werden, wirken sie sich auf das gesamte Gesundheitssystem aus - von der Primärversorgung bis zur Krankenhausbehandlung. Mit den Fortschritten in der Sensortechnologie, den Algorithmen des maschinellen Lernens und der Rechenleistung von Maschinen ist es jedoch möglich geworden, auf maschinellem Lernen basierende Systeme einzusetzen, um solche Gesundheitszustände zu diagnostizieren oder Ärzte und andere Gesundheitsfachkräfte bei ihren Entscheidungen zu unterstützen. Dennoch sind einige körperliche und geistige Zustände immer noch nicht objektiv mit nicht-invasiven physiologischen Messungen unter Verwendung von maschinellen Lernansätzen zu erkennen, wie z. B. Hunger,Sättigung und Zustände des menschlichen Flow-Erlebens. In diesem Zusammenhang werden in dieser Arbeit moderne, auf maschinellem Lernen basierende Ansätze zur Erkennung der oben genannten Gesundheitszustände anhand von physiologischen Zeitreihendaten vorgestellt, die mit nicht-invasiver, tragbarer Sensortechnologie erfasst wurden.

Der erste Teil dieser Arbeit (Kapitel 2) befasst sich mit der Erkennung von Hunger- und Sättigungszuständen. Die Wahrnehmung von Hunger und Sättigung ist entscheidend für die Aufrechterhaltung eines gesunden Körpergewichts und die Vermeidung von chronischen Krankheiten wie Übergewicht, Untergewicht oder Mangelerscheinungen aufgrund von Fehlernährung. Ein chronischer Verlust dieser Wahrnehmung kennzeichnet zahlreiche Gesundheitszustände. Wie aus der Literatur hervorgeht, ist es schwierig, diese Erkrankungen mit nicht-invasiven Messungen zu erkennen. Um dieses Ziel zu erreichen und ein hochpräzises System zu erhalten, wird in diesem Kapitel ein hierarchisches System vorgestellt, das Signalverarbeitungs- und Mustererkennungstechniken einsetzt, um Hunger und Sättigung nicht-invasiv zu erkennen. Drei Geräte, das Empatica E4 Armband, das Biosignalplux RespiBAN Wearable und die JINS MEME Smart Glass, werden verwendet, um physiologische Signale von fünf gesunden, normalgewichtigen Probanden zu erfassen, die ruhig auf einem Stuhl sitzen während sie hungrig oder gesättigt sind. Die Untersuchungen dieser

Studie zielen darauf ab, die oben genannte wissenschaftliche Lücke zu schließen, indem ein nicht-invasives multimodales System vorgeschlagen wird, das in der Lage ist, zwischen Hunger- und Sättigungszuständen zu unterscheiden. Erstens ist es möglich, diese Zustände durch nicht-invasive physiologische Messungen zu erkennen. Zweitens stammen die unterscheidungsfähigsten Merkmale von drei spezifischen Sensormodalitäten: Elektrodermale Aktivität (EDA), Infrarot-Thermometer (Tmp) und Blutvolumenpuls (BVP). Diese Sensormodalitäten wurden vom Empatica E4 Armband aufgenommen, dem wichtigsten Gerät in dieser Studie, das in Zukunft als eigenständiges Wearable für die nicht-invasive Hunger- und Sättigungserkennung verwendet werden kann. Drittens sind die Ansätze des Merkmalslernens nicht unbedingt leistungsfähig, insbesondere bei begrenzten Trainingsbeispielen. Schließlich ist der Random Forest (RF) Klassifikator der zuverlässigste für die Erkennung von Hungerzuständen bei dieser Auswahl sinnvoller Merkmale.

Im zweiten Teil dieser Arbeit (Kapitel 3) wird ein neuartiges multimodales System vorgestellt, mit dem das Flow-Erlebnis des Einzelnen während der Arbeitstätigkeit ermittelt werden kann. Das Flow-Erlebnis ist ein spezifischer positiver und affektiver Geisteszustand, der eintritt, wenn der Mensch völlig in einer Tätigkeit aufgeht und alles andere vergisst. Dieser Zustand kann zu hoher Leistung, Wohlbefinden und Produktivität bei der Arbeit führen. Es wurden nur wenige Studien durchgeführt, um das menschliche Flow-Erlebnis mit Hilfe von physiologischen, tragbaren Geräten zu erkennen; diesen Studien mangelt es jedoch an Objektivität bei der Durchführung. Darüber hinaus muss noch untersucht werden, wie das Problem der Datenknappheit in diesem Bereich angegangen werden kann und wie emotionale Daten genutzt werden können, um die Flow-Erkennungsleistung zu verbessern. In dieser Studie werden physiologische Daten von 25 Probanden mit multimodalen Messgeräten - dem Empatica E4 Armband, dem Emotiv Epoc X Elektroenzephalographie (EEG) Headset und dem Biosignalplux RespiBAN Wearable - erhoben, während die Teilnehmer Rechen- und Leseaufgaben lösen. Es werden Experimente durchgeführt, um die Unterscheidung zwischen Flow- und Non-Flow-Zuständen mit Hilfe von Feature-Engineering und Deep-Feature-Learning-Ansätzen zu untersuchen. Außerdem wird die Verbindung zwischen Emotionen und Flow-Erfahrung mittels Transfer-Learning-Techniken untersucht, indem ein Modell zur Emotionserkennung vortrainiert und das generierte Wissen auf die Flow-Erkennung übertragen wird. Die Ergebnisse dieser Studie legen nahe, dass eine effektive und objektive Unterscheidung zwischen Flow- und Nicht-Flow-Zuständen mit multimodalen Sensordaten möglich ist. Der Erfolg des Transferlernens unter

Verwendung des DEAP-Emotionsdatensatzes zum Vortrainieren der Modelle deutet darauf hin, dass emotionsbezogene Dimensionen, insbesondere Erregung und Flow, miteinander verbunden sind und dass die Erkennung von Emotionen als latente Aufgabe verwendet werden kann, um die Leistung der Flow-Erkennung zu verbessern. Diese Erkenntnis kann dazu beitragen, das Problem der Datenknappheit in diesem Bereich zu umgehen.

Contents

1 Introduction **1**

1.1 Fundamentals . . . 2
1.1.1 Ubiquitous Sensors . . . 2
1.1.2 Machine Learning for Time-Series Classification . . . 4
1.1.3 Pattern Recognition Pipeline . . . 7
1.1.4 Deep Learning . . . 12
1.2 Machine Learning in Healthcare . . . 14
1.2.1 Physical Health Assessment . . . 15
1.2.2 Mental Health Assessment . . . 17
1.2.3 Limitations and Challenges . . . 18
1.3 Motivations . . . 20
1.4 Own Contributions . . . 21
1.5 Overview . . . 24

2 Hunger and Satiety Recognition **26**

2.1 Introduction . . . 27
2.1.1 Background . . . 27
2.1.2 Current Challenges . . . 28
2.1.3 Research Motivation . . . 29
2.2 Related Work . . . 30
2.3 Own Contribution . . . 34
2.4 Dataset Description . . . 34
2.5 Data Preprocessing . . . 38
2.6 Feature Extraction . . . 40
2.6.1 Feature Engineering . . . 40
2.6.2 Feature Learning . . . 45
2.7 Classification . . . 48
2.8 Experiments and Results . . . 50

2.8.1 Feature Engineering . . . 51
2.8.2 Feature Learning . . . 54
2.8.3 Comparison with the Literature . . . 55
2.9 Scientific Discussion . . . 57
2.10 Summary . . . 60

3 Human Flow Experience Recognition **61**
3.1 Introduction . . . 62
3.1.1 Background . . . 62
3.1.2 Current Challenges . . . 63
3.1.3 Research Motivation . . . 64
3.2 Related Work . . . 66
3.3 Own Contribution . . . 70
3.4 Datasets Description . . . 71
3.4.1 Emotion Recognition Dataset: DEAP . . . 71
3.4.2 Flow Recognition Dataset: PhySF . . . 72
3.5 Data Preprocessing . . . 75
3.5.1 DEAP . . . 75
3.5.2 PhySF . . . 76
3.6 Feature Extraction . . . 77
3.6.1 Feature Engineering . . . 77
3.6.2 Feature Learning . . . 78
3.6.3 Transfer Learning . . . 79
3.7 Classification . . . 80
3.8 Experiments and Results . . . 83
3.8.1 Feature Engineering . . . 83
3.8.2 Feature Learning . . . 84
3.8.3 Transfer Learning . . . 85
3.8.4 Comparison with the Literature . . . 86
3.9 Scientific Discussion . . . 88
3.10 Summary . . . 91

4 Conclusion and Future Work **93**
4.1 Summary . . . 93
4.2 Scientific Findings . . . 96
4.3 Limitations . . . 98
4.4 Future Work . . . 99

Bibliography **101**

List of Own Publications **121**

List of Abbreviations **123**

List of Figures **125**

List of Tables **127**

Chapter 1

Introduction

People are more health conscious than ever in the twenty-first century because our health affects so many aspects of our lives. For example, it may affect our employment opportunities and income and may have an impact on our social activities, mood, and overall well-being [1]. Therefore, people's subjective well-being increases when they are in good health and decreases when they are ill or in poor health [2, 3].

According to the *World Health Organization* (WHO), health is a condition of complete physical, mental, and social well-being without any disease or illness. However, diagnosing a particular illness or disorder sometimes becomes difficult for physicians. Because physical and mental health are deeply interconnected and influence each other via various pathways, but little is known about their relationship [4]. In addition, the frequency and severity of medical conditions are also increasing at a rapid rate, making it challenging to identify the disease and provide certainty that it exists. Thus, using machine learning in healthcare is a potential solution to this problem to assist physicians or caregivers in making diagnostic decisions. Furthermore, with ubiquitous sensor technologies such as smartphones, smartwatches, smart glasses, and others, collecting patients' data non-invasively and unobtrusively has become possible. Therefore, machine learning with ubiquitous technology can help to assess physical and mental health states and improve disease detection and clinical decision-support system efficiencies [5, 6].

The remaining chapter is structured as follows: Section 1.1 provides an overview of the fundamental concepts related to this dissertation. Section 1.2 presents the most common applications of machine learning for physical and mental health assessment and the

limitations and challenges of wearables in assessing these states. Section 1.3 defines the motivations behind this dissertation. Section 1.4 explicitly states the contributions of this dissertation. Finally, Section 1.5 illustrates the overall layout of this thesis.

1.1 Fundamentals

This section provides a concise summary of this dissertation's core concepts. For instance, Section 1.1.1 presents the most commonly used ubiquitous or wearable sensors for detecting human health states. Section 1.1.2 provides the basic concept of machine learning for time-series classification. Section 1.1.3 explains the primary steps generally followed to recognize human health states using machine learning. Finally, Section 1.1.4 briefly describes the concept of deep learning.

1.1.1 Ubiquitous Sensors

Ubiquitous sensors are critical for developing a health recognition system using machine learning algorithms [7]. Integrated into devices or wearables, these sensors enable the continuous acquisition of diverse types of health-related measurements, such as heart rate (HR), blood pressure, temperature, physical activity, and sleep patterns. The analysis of such data can provide a more extensive understanding of human health conditions beyond conventional clinical measures and facilitate early disease detection, enabling timely interventions and proactive health management. The selection of sensors for such an application depends on the health parameters to be monitored. Furthermore, it is common in the research field to use a combination of multiple sensors to collect comprehensive data and investigate which specific measurements are more capable of detecting a particular health condition than others. The following are some of the most common types of sensors that can be used for various health assessments:

- **Inertial Sensors:** Inertial sensors, also known as motion sensors, consist of accelerometers, gyroscopes, and magnetometers. These sensors facilitate measuring acceleration, angular velocity, and sometimes orientation, which enable detecting and quantifying the movements of humans or patients [8]. They are playing an influential role, particularly in health-related applications, by

enabling movement monitoring, gait analysis, and fall detection and contributing to the development of various assistive applications for healthcare.

- **Visual Sensors:** In the context of health-related applications, visual sensors refer to imaging devices such as cameras that capture visual information and can support diagnostics, monitoring, rehabilitation, and the overall improvement of healthcare services [9]. Imaging devices include regular cameras, depth-sensing cameras, thermal cameras, and more. However, their use depends on the application's specific requirements or the problem to be solved. The visual information acquired by using such devices provides valuable insights for healthcare systems.

- **Wearable Sensors:** Wearable sensors are devices that include miniaturized sensors integrated into wearable items such as clothing, smart glasses, wristbands, headbands, rings, or other wearables. These devices facilitate the collection and monitoring of various types of signals from the body [10]. For example, visual sensor-integrated wearables like smart glasses can display real-time data and support healthcare professionals during procedures. Nevertheless, the *Empatica E4* wristband, *Jins MEME* smart glasses, and *Emotiv Epoc X* mobile brainwear are some of the most popular wearable devices. These wearables enable measuring HR, respiratory rate, skin temperature, electrodermal response, and other parameters that can be used to detect various types of health conditions.

- **Biochemical Sensors:** Biochemical sensors are devices particularly designed to detect specific biomarkers or chemical parameters of the human body. For example, chemical parameters present in bodily fluids such as sweat, saliva, blood, urine, tears, and respiratory gases can be measured with these sensors, thus providing valuable insights into an individual's metabolism and overall health condition [11, 12]. Furthermore, continuous monitoring with these sensors can facilitate healthcare professionals and researchers to gain critical insights into an individual's health status and specific physiological responses. However, some types of biochemical sensors have not been commercialized yet [13, 14].

- **Radar Sensors:** Radar sensors emit electromagnetic signals in the form of radio waves and can capture the reflections or echoes to gather information about the surrounding environment. Traditionally, they were designed for applications such as automotive, aviation, and security. However, the researcher has started

investigating their use in assistive healthcare to detect unusual movements, especially in elderly care [15, 16].

- **Sleep Sensors:** Sleep sensors streamline detecting and analyzing specific biological and environmental factors during sleep. These sensors come in various forms, including wearable devices, smart mattresses, and standalone sleep monitoring systems, which facilitate monitoring diverse parameters associated with sleep patterns, duration, and quality [17]. In clinical settings, sleep is usually evaluated in a clinical environment through the *Polysomnography* (PSG) study. During this examination, patients sleep while wearing various physiological sensors, which measure brain waves, eye movements, electrical activity of muscles, blood oxygen saturation, and audio sounds [18].

It is worth mentioning that the sensor system or wearable device selection depends on the application domain, problem, or the parameters to be measured.

1.1.2 Machine Learning for Time-Series Classification

Machine learning is a subfield of *Artificial Intelligence* (AI) and computer science that focuses on using data and mathematical algorithms to learn and perform tasks that generally require a high degree of abstract reasoning to be solved. The primary way machine learning proceeds towards this objective is by analyzing the most relevant data to the task intended to train the mathematical model. After the model's successful training using the relevant data (also called training data), it is usually given to the machine to be reused and tested with actual data (also called test data), i.e., the data that does not match with the data used to train the model. In light of the literature, machine learning models learn well when there is enough data because the model's generalization depends on the dataset's diversity, and a well-generalized model with diverse training data is more likely to make correct predictions for the test or unseen data [19].

Machine learning has gained a lot of popularity in recent decades because it has substantial overlaps with various emerging fields such as mathematical optimization, algorithmic, data science, and statistics and can solve various problems in many other application domains. For example, it is widely used in ubiquitous computing-related

studies to give computers the "intelligence" they need to assist their users in making decisions.

Machine learning approaches can be divided into various families depending on the data type required to train their models and how they use it. The three fundamental families of machine learning approaches that are mainly distinguished are:

1. **Supervised learning:** In supervised learning, models are trained using labeled data — the input features and corresponding target values are usually provided. The model learns to map the input features to the target labels using the posterior probability $p(y \mid x)$. Once trained on the training data, a model can classify test data x into y classes.

2. **Unsupervised learning:** Models of unsupervised learning work with unlabeled data in both the training and testing phase. Their main objective is to discover patterns, structures, or relationships, usually with similarity measures between train and test samples and without predefined target labels. Clustering and dimensionality reduction are standard unsupervised learning techniques.

3. **Semi-supervised learning:** This approach lies between supervised and unsupervised learning to solve their key challenges. It usually uses a few labeled samples (data) for training an initial model and then iteratively applies it to the more significant number of unlabeled samples. Primarily, the model learns from labeled samples and generalizes this knowledge to make predictions for unlabeled data.

In machine learning, the quality of the labels is crucial and can lead to better predictions. Nevertheless, labeling data seems simple and inexpensive, but in practice, it can be expensive and complicated because of a reasonable investment of time or resources. One popular method for containing labels for large datasets is crowd-sourcing, which can be used to collect a large number of labels in a short time and with little money. Unfortunately, the labels compiled through crowd-sourcing are usually highly erroneous because most workers in the crowd are not experts [20]. Although unsupervised learning does not require labels, supervised learning remains the most commonly used category of machine learning approaches because of its notable performances compared to semi-supervised and unsupervised learning in many application domains until now.

Supervised machine learning employs a dataset and its associated labels to solve a given problem for a given application. Most supervised learning methods transform a given problem into a regression or a classification problem. Regression approaches attempt to predict continuous values such as price, salary, age, etc. On the other hand, classification approaches find discrete values (or class labels), such as hunger or satiety, flow or non-flow, spam or non-spam, etc. In terms of available algorithms, both classification and regression approaches aim to approximate a function ϕ : x $\rightarrow$ y, which maps input data or samples x to continuous or discrete values y, respectively. This approximation is usually done by using computed features, which are the values that can generally be computed with the help of the domain experts, or by using simple mathematical, statistical, or frequency-related functions on the input data. Features represent input data abstractly for the given problem to solve. They allow associating each data sample x with a feature vector $f(x) = \{f_1(x), f_2(x), \cdots, f_n(x)\} \in R^n$ where $n \in N^*$ is number of computed features and R^n is the feature space. A machine learning model is then trained in feature space R^n to map each feature vector f (x) with its associated label y.

In general, classification and regression approaches vary in their degree of difficulty in training the model. With classification approaches, models are generally easier to train than regression ones. Nevertheless, each machine learning approach requires large amounts of training data and associated labels — regression approaches require more training data than classification approaches to approximate their continuous target values accurately. Therefore, classification is the prevailing approach in most machine learning-based applications.

Sensor modalities are a common source of data collection for many applications. Section 1.1.1 describes the most common sensors that can be used for health assessment applications. Nevertheless, machine learning approaches remain the same regardless of the data type used in the training set. In practice, a split has emerged over the past decade between researchers working with image modalities and those working with other sensor modalities. Researchers working on ubiquitous computing tend to fall into the second category because data from images or visual sensors can raise privacy concerns. Therefore, wearable sensors that provide data in a temporal sequence, also known as time-series data, are preferred for ubiquitous computing-based applications. In this context, time-series data classification has become an important topic for sensor-based applications, especially in the context of health assessment.

1.1.3 Pattern Recognition Pipeline

Developing a machine learning-based health assessment system follows a standardized sequence of actions using sensor technology and machine learning algorithms [21]. This process is generally referred to as the *Pattern Recognition Pipeline* (PRP) [22]. Figure 1.1 graphically presents this PRP, which includes each step from data acquisition to classification. These steps are explained in the following sections:

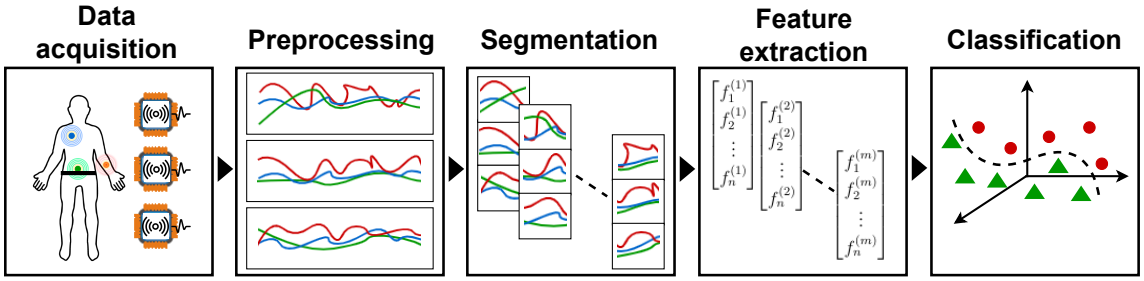

Figure 1.1: The primary stages involved in the design of a machine learning-based health assessment system. In the first step, raw data is usually acquired from the sensors. After the necessary preprocessing of the data, segments are extracted (Segmentation), and values relevant to the problem to be solved are calculated from them (Feature Extraction). Lastly, a classifier is trained and evaluated (Classification).

Data Acquisition

It is the first and most significant step in the PRP because sensor selection is decisive at this stage and usually depends on the targeted application type. Furthermore, ethical approval is also mandatory at this stage to collect data from the study probands or subjects. Section 1.1.1 lists the ubiquitous sensors generally used to acquire input data for developing various machine learning-based health assessment applications.

Preprocessing

Data preprocessing involves transforming the raw data into an appropriate format for further analysis. It may include sensor calibration, noise reduction, normalization, or the cleanup of corrupted data from possible hardware errors or data transmission issues. However, synchronizing data becomes crucial when various multimodal wearable devices with varying sampling frequencies are involved in the data acquisition.

Segmentation

The segmentation process depends on the type of data in hand. For example, in image segmentation, the segmentation algorithm usually segments an image into fragments and assigns each fragment a label. However, the time-series data segmentation techniques

divide time-series data belonging to each class into discrete segments depending on the implemented approach, such as sliding window, top-down, and bottom-up.

Feature Extraction

Feature Extraction methods transform each data segment into a set of meaningful and representative feature values that can be used as input to a machine learning algorithm. The selection of a feature extraction method depends on several factors, including the type and structure of the data, the specific domain, and the available computing power. Nevertheless, carefully extracted features can significantly improve the performance of machine learning models by reducing noise, improving interpretability, and capturing relevant patterns in the data. There are two leading families of feature extraction methods: feature engineering and feature learning.

1. **Feature engineering:** Feature engineering refers to the manual crafting of features. Such features are generally known as *Hand-Crafted Features* (HCFs) and are either based on experts' knowledge or simple transformation functions to the data segments. The HCFs can be computed in both the time and frequency domains. The most commonly used engineered features include but are not limited to mean, maximum, minimum, percentile, standard deviation, spectral energy, and spectral entropy. They are discussed in Chapter 2, Section 2.6.1.

2. **Feature learning:**

   Feature learning is the process of employing deep learning models to learn distinguishing features fully automatedly (see Section 1.1.4). The most commonly used deep learning-based models for feature learning include *Multilayer Perceptrons* (MLPs), *Convolutional Neural Networks* (CNNs), and *Recurrent Neural Networks* (RNNs). The MLPs contain fully connected layers and perform feature learning by processing data through multiple hidden layers, where each layer learns and transforms the input data through weighted operations and activation functions. This process allows the network to learn higher-level feature representations progressively. The CNNs include convolution layers and apply convolutional operations to small patches of the input data segment, extracting features concerning local data patterns. The RNNs form a chain of repeating neural network modules and are typically employed for long input data sequences. Their computation depends on the previous computations at each step. Therefore, they are usually considered computationally very expensive

in the literature. Feature learning using MLPs and CNNs is discussed in Chapter 2, Section 2.6.2.

Classification

In this step, a classifier is trained based on the extracted features. A classifier is a machine learning algorithm that classifies the input data segments into predefined classes. The most commonly used classifiers in health assessment applications are described below:

- **Naïve Bayes (NB):** This is a probabilistic classifier based on *Bayes'* theorem. The NB is called "naive" because it assumes all features are independent. This strong assumption is not always valid, but NB often achieves good results in practice despite this simplification. It is easy to implement, requires only a small amount of training data, and makes predictions much faster than most well-known classifiers [23].

- **Support Vector Machine (SVM):** A supervised machine learning classifier suitable for linear and non-linearly separable data. The main objective of an SVM is to find a hyperplane that maximally separates the elements of the two classes [24]. The same principle is applied in multi-class classification after decomposing the multi-class classification problem into several binary classification problems. In the case of linearly separable data, the hyperplane is a straight line. However, in the case of non-linearly separable data, the SVM uses a technique known as the kernel trick to transform the data into a higher-dimensional space where linear separation is possible [25]. It has been used mainly with feature engineering approaches and provides high classification performance. Nevertheless, it can also successfully classify features learned using feature-learning techniques.

- **Decision Tree (DT):** It is a classification and regression analysis technique that learns simple decision rules from labeled training data to make predictions for unseen instances. It consists of internal nodes, branches, and leaf nodes. Each internal node represents a decision, while the branches represent the possible outcome of that decision. The leaf nodes represent the conclusion of the predictions. In the training phase, the DT classifier analyzes the features of the training data and selects the most informative features to create decision nodes. To make predictions for unseen data, the DT classifier follows the decision rules

at each internal node as it traverses the tree from the root node to the leaf node. It assigns the class label associated with the reached leaf node to the input instance as its predicted class [26].

- **Random Forest (RF):** It is an ensemble learning method that constructs multiple DTs called learners at training time, where each learner is trained on a different subset of data. The randomness of the samples helps to avoid overfitting and improve generalization. For the final results, the predictions of each learner are counted, and the class with the majority of votes becomes the final prediction. This method has several advantages, including handling large datasets and missing data values. It can overcome the limitations of a single DT and provide more accurate and robust predictions since it is less prone to overfitting compared to a single DT classifier. In addition, it can be used to measure the importance of features to find out which features have the most significant impact on the prediction [27].

- **Adaptive Boosting (AdaBoost):** This is also an ensemble learning method originally developed to improve the performance of binary classification tasks. It combines the predictions of multiple weak learners (usually DTs) to create a robust classifier. AdaBoost iteratively trains a group of weak learners, with each successive learner focusing on the examples that the previous learner had difficulty with. Each iteration assigns a higher weight to the misclassified examples to emphasize their importance and force the weaker learners to focus on them. In this way, weak learners gradually improve their performance. However, it is sensitive to outliers and can be very intensive computationally due to its iterative nature [28].

- **Extreme Gradient Boosting (XGBoost):** XGBoost is an ensemble learning method that combines the predictions of multiple weak learners to build an accurate predictive model. The key concept behind it is gradient boosting, in which an ensemble of trees is built sequentially, where each DT is trained to correct the previous errors in the ensemble, focusing on the examples that were difficult to predict correctly. The DTs are added iteratively at each iteration, and the model attempts to optimize a particular loss function (i.e., the gradient of a differentiable loss) [29]. The XGBoost has attracted much attention due to its scalability, flexibility, and high performance and is now widely used for classification tasks in various domains [30].

The selection of a classifier depends on various factors (e.g., data type, size, and resources available to process the data). Usually, these classifiers are implemented in the training and testing phases. The raw data is first acquired from the sensors in the first phases. After the necessary preprocessing and segmentation of the data, suitable features are extracted to create feature vectors. The feature vectors are then split into training and testing datasets. In the training phase, the training dataset is used to tune the inner parameters of the classifier to reduce the variance between the predicted and actual labels. This process is referred to as training the model. In the testing phase, the trained model is used to predict the labels of the test dataset. Finally, the model's performance is evaluated using various metrics on the test dataset.

Evaluation

Evaluating health assessment models is an essential step in assessing their performance and determining their effectiveness. In other words, it helps to understand how well your model is performing. In machine learning, the number of correctly predicted positive labels (t_p), the number of incorrectly predicted positive labels (f_p), the number of correctly predicted negative labels (t_n), and the number of incorrectly predicted negative labels (f_n) serve as indicators of the model's performance.

To evaluate the classification performance of all implemented models, the accuracy, sensitivity (recall), and specificity were computed, as given in Equations 1.1 to 1.3, respectively.

$$\textit{Accuracy} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{1.1}$$

$$\textit{Sensitivity} = \textit{Recall} = \frac{t_p}{t_p + f_n} \tag{1.2}$$

$$\textit{Specificity} = \frac{t_n}{t_n + f_p} \tag{1.3}$$

Since class imbalance can affect the overall accuracy, the average F1 score, also referred to as the macro Averaged F1 (AF1) score, was computed. The AF1 score is the average of all c classes of F1 scores, whereas each class F1 score is the harmonic mean of the considered class precision and recall Equation 1.5.

$$\textit{Precision} = \frac{t_p}{t_p + f_p} \tag{1.4}$$

$$\textit{F1 score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{1.5}$$

$$\textit{AF1 score} = \frac{1}{c} \sum_{i=1}^{c} \text{F1 score}_i \tag{1.6}$$

In Equation 1.6, c represents the no. of classes, and F1 score$_i$ represents the F1 score for the ith class.

1.1.4 Deep Learning

Dechter first used the term "deep learning" in 1986 [31] to describe the use of deep ANNs in machine learning applications. The fundamental building blocks of deep learning, the ANNs, derive their design principles primarily from the functioning of biological neurons of the human brain. They comprise interconnected artificial neurons, where each neuron serves as a fundamental non-linear computational unit [32], as illustrated in Figure 1.2.

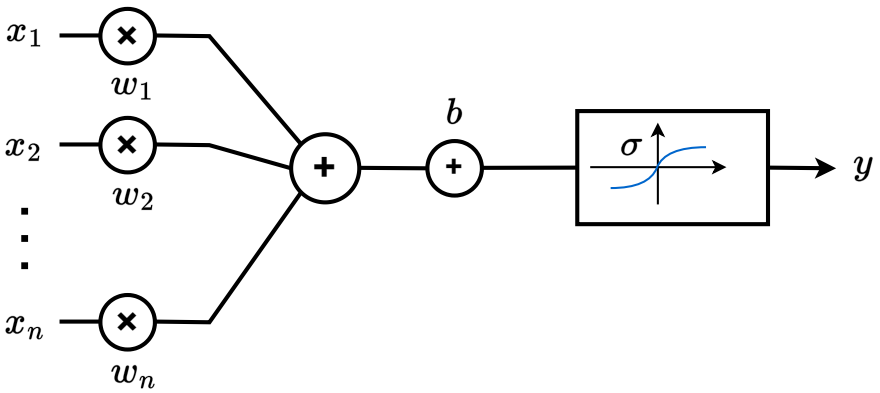

Figure 1.2: Illustrates the function of an artificial neuron: $n \in N^*$; the inputs $\{x_1, x_2, \ldots, x_n\} \in R^n$ undergo a process where they are multiplied by respective weights $\{w_1, w_2, \ldots, w_n\} \in R^n$, summed together, combined with a bias $b \in R$, and then passed through a non-linear function (σ), ultimately producing an output value $y \in R$.

The artificial neurons of an ANN process the given inputs and generate an output using the following equation.

$$y = \sigma(\sum_{k=1}^{n} w_k x_k) + b \tag{1.7}$$

Where σ denotes a non-linear function, known as the *activation function*, $\forall k \in \{1, 2, \ldots, n\}, w_k \in R$ represent the internal learnable parameters, also called *neural weights*, $x_k \in R$ represents the input, and $b \in R$ is an offset parameter referred to as neural bias.

In an ANN, artificial neurons are structured into diverse layers — comprising the input, hidden, and output layers. They process the input and pass it to the neurons in the next layer. *Deep Neural Networks* (DNNs) are a specific type of ANNs that contains multiple hidden layers, allowing for the design of deep architectures. The hidden layers enable DNNs to learn and extract complex feature representations from given data by progressively transforming it through each layer, leading to more abstract feature representations.

The ANNs are adaptable and can be used for classification and regression tasks by adjusting the activation function and size of the last or output layer. For the classification task, the standard practice is to set the output layer neurons equal to the number of class labels of the given dataset — by using a softmax activation function [33]. This function normalizes these inputs into a probability distribution, ensuring that the output values fall between 0 and 1 and that the sum of all the output values across neurons in the output layer equals 1. The output values of the output layer neurons can be interpreted as the probability estimation of each class — the class associated with the highest probability is often considered the predicted class for the given input. For an ith class, this probability (p_i) can be calculated as:

$$p_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \quad for\ j = 1, \ldots, N \tag{1.8}$$

In Equation 1.8, x_i represents the input to the ith neuron, e is the base of natural logarithm (Euler's number), $j = 1, \ldots, N$ represents the number of neurons in the last

layer and the denominator represents sum of the exponential values of all inputs in the layer.

The training phase of an ANN enables it to learn from a designated (or training) dataset and enhances its ability to make accurate predictions on a separate, unseen test dataset. Throughout this process, the network fine-tunes its parameters by adjusting the weights and biases of all neurons across its layers. Neurons' weights and biases are generally randomly initialized using the Glorot initialization strategy [34]. This method selects random values for the weights and biases within intervals determined by the sizes of the current and subsequent layers. These learnable parameters are iteratively updated during the training phase of the network by a loss function. The loss function computes disparities between expected and actual neural output to minimize the loss using a mathematical optimization approach [35]. For an ANN or DNN, gradient descent is the most commonly used method to minimize the loss; it iteratively updates each weight and bias parameter during training in the opposite direction of the derivative of the loss function. Nevertheless, parameter updates that are allied to the other layers of the network are carried out by the backpropagation algorithm [36], which back-propagates the loss to the neurons of the previous layers using the chain rule of derivation [37].

Deep learning, as a field, primarily revolves around DNNs, focusing on training these deep architectures to automatically learn and represent data in hierarchical layers, making it particularly effective for handling large, unstructured datasets. In essence, DNNs are a type of ANN, and their multi-layered structure characterizes *deep learning* in handling complex tasks such as image recognition, natural language processing, and other pattern recognition problems.

1.2 Machine Learning in Healthcare

Recently, there has been an increasing interest in leveraging machine learning and ubiquitous sensing technologies for physical and mental health applications. These technologies allow for the automatic and continuous assessment of many health states, such as pain, involuntary and spontaneous movements in babies, depression, anxiety, stress levels, different stages of sleep, and even psychological flow states. Sections 1.2.1 and 1.2.2 describe the recently proposed and most common applications for assessing physical and mental health, respectively.

1.2.1 Physical Health Assessment

A physical health assessment involves assessing and tracking various aspects of an individual's physical health. It covers the systematic observation, measurement, and evaluation of various parameters associated with the body's functioning and general health status. The most commonly measured physiological parameters include blood pressure, HR, respiratory rate, body temperature, skin temperature, etc.

The main objective of a physical health assessment is to gain insight into an individual's physiological condition, identify any potential health problems or abnormalities, and take preventative measures to maintain or enhance their health. Nowadays, machine learning is being widely utilized in the medical industry due to its decision-making capabilities. The essence of a machine learning-based system is to employ classification algorithms to develop an intelligent decision-making model based on physiological data collected using ubiquitous sensors. Such models can successfully assess if a person has a sickness and what type of condition they have without needing external intervention. The following are the most common machine learning-based applications for physical health assessment:

Pain Recognition: The assessment of pain is a key area of research. Because pain can indicate various health problems and serve as a natural protective mechanism against harm. It is particularly important to assess it in medicine because it encompasses both symptoms and diseases. In recent years, several ideas have been put forward for automatic pain assessment systems that use machine learning models trained with data from sensors that measure behavior or physiological states [38].

Heartbeat Monitoring: In clinical contexts, there are numerous applications of *Electrocardiography* (ECG) signal analysis. Nevertheless, as wearable sensors become more ubiquitous, current research has dealt with the issue of enabling long-term monitoring of heartbeats. As heartbeat morphology varies with HR, models learned when resting must be adapted to analyze ECG signals recorded during daily activities. In a similar manner, Carrera et al. [39] proposed an online ECG monitoring system that models the normal heartbeats of users in sparse representations. When users' heartbeats do not match the sparse representation model, they are recognized as abnormal, and vice versa.

Gait Analysis: Gait analysis is another prominent area of research that focuses on studying human locomotion. This type of analysis can provide useful information not only for medical diagnosis but also for applications in security and biomechanics. After great success with wearable and visual sensors in this field, researchers are now focusing on acoustic signals for gait analysis [40]. In this context, Altaf et al. [41] introduce an acoustic gait profile obtained from temporal signal analysis of the sound of footsteps collected by microphones. They revealed how some spatiotemporal gait parameters can be extracted from their proposed acoustic gait profile, which can consistently and reliably estimate a subset of clinical and biometric gait parameters currently used for standardized gait assessments.

Body Composition Analysis: Body composition has been identified as a significant predictive marker for diseases such as cancer, cirrhosis of the liver, and critically ill patients. Most body composition assessments are based on *computed tomography* (CT) scan analysis. However, several recent articles have proven excellent accuracy compared to human raters for measuring different body composition parameters, such as skeletal muscle, visceral adipose, and subcutaneous adipose tissue from the lumbar vertebrae region, using deep learning. This indicates that DNNs can successfully automate body composition analysis. For example, Ma et al. [42] introduced a full-body 3D CT segmentation method to accurately segment bone, skeletal muscle, subcutaneous fat, and visceral fat. Ha et al. [43] presented a fully automatic deep learning system for L3 slice selection and segmentation of abdominal muscle areas. These studies confirm the feasibility of using deep learning algorithms for automated segmentation of pelvic muscles, fat, and bone from CT studies, enabling precise body composition assessment.

Human Activity Recognition: Human Activity Recognition (HAR) involves monitoring individuals with disabilities or specific conditions by analyzing physical activity. The most common application of HAR is recognizing daily activities (ADL), such as sitting, standing, walking, etc. Khan et al. [44] developed a system using accelerometer and pressure sensor data in this context. Nisar et al. [45] introduced a two-level hierarchical model to recognize ADLs. The model first finds atomic activities and then uses rank-pooling to handle transitions. Similarly, Bisio et al. [46] proposed a smartphone-based HAR system for remote health monitoring, classifying activities like sitting, standing, walking, or running using accelerometer signals.

1.2.2 Mental Health Assessment

Mental illness affects a person's emotions, reasoning, and social interactions. These problems show that mental illness has severe implications for society and requires the development of new prevention and intervention strategies. Typically, mental illness is diagnosed based on self-reported symptoms. This requires questionnaires that capture specific patterns of emotions or social interactions [47]. However, the rise of mental health problems and the need for effective medical healthcare have prompted research on the application of machine learning to mental illness. In recent years, the research community has shown great interest in determining mental health problems using wearable physiological signals and machine learning methods. The following are the most common applications of machine learning to mental health assessment:

General Movement Assessment: General movements (GMs) are spontaneous movements of infants up to five months after birth that involve the whole body and vary in sequence, speed, and amplitude. The detection of GMs is important because their assessment is considered a valuable tool for early predicting various neurological disorders in infants. In recent years, wearables and camera-based systems have been used to detect GMs [21, 48].

Anxiety and Depression Screening: Anxiety and depression are two critical disorders in the broad spectrum of mental illness. The number of disabilities due to these mental health problems steadily increases worldwide. Therefore, timely diagnosis of these problems is essential for people's health and well-being. In this context, Moshe et al. [49] introduced a new diagnostic method that predicts symptoms of depression and anxiety using smartphone and wearable device (i.e., Oura Ring) data. Their results indicate that these wearable devices can provide valuable data sources for predicting depression and anxiety symptoms, especially data related to general sleep measurements.

Flow Experience Recognition: Humans experience flow as a particular positive affective state that occurs when they become completely absorbed in an activity and forget everything else. This state can lead to high performance, well-being, and productivity at work. Thus, the recognition of flow is important for human well-being, and its scientific understanding becomes a prerequisite to contributing to improving human life. Furthermore, describing, explaining, and predicting this phenomenon can help improve behaviors. As a result, many ideas have been introduced to look into this

using machine learning models that use wearable sensors to track physiological states and automatically recognize interruption-free flow experiences [50].

Post-traumatic Stress Disorder Assessment: The assessment of Post-traumatic Stress Disorder (PTSD) is another critical area of research. It is a chronic mental health condition that can occur after a potentially traumatic event, such as being exposed to a death threat, serious injury, sexual violence, etc. The consequences of it include prolonged suffering, distress, impaired quality of life, and increased mortality. Assessment of PTSD is important for accurate diagnosis and treatment planning and to ensure that individuals with this disorder receive appropriate care and support, resulting in improved outcomes and quality of life. Several approaches have proposed automated assessments of PTSD using machine learning models trained with data from *Magnetoencephalography* (MEG) or other physiological sensors [51, 52].

Schizophrenia Detection: Schizophrenia is a chronic mental disorder that affects various aspects of cognition and perception, including how a person thinks, acts, expresses feelings, and perceives reality. The traditional diagnosis methods, which rely on patient history and mental health assessments, are considered inaccurate [53]. Buettner et al. [53] provide an alternative by employing a machine learning approach that analyzes *Electroencephalography* (EEG) recordings. The authors reportedly used 499 one-minute EEG recordings from an open neurological and psychiatric archive for their analysis. By applying an RF classifier, they attained a highly balanced accuracy rate of 96.77% in distinguishing between patients with and without schizophrenia.

Advancements in technology have led to the development of various wearable devices, mobile applications, and other digital tools that enable individuals to track and monitor their physical and mental health more conveniently. These tools provide real-time data and insights, allowing individuals to make informed decisions about their lifestyle choices, exercise routines, and overall health management.

1.2.3 Limitations and Challenges

Wearable sensors have gained significant popularity in recent years due to their capacity to facilitate the collection of diverse data types, including Electrooculography (EOG), Electromyography (EMG), Electrodermal Activity (EDA), EEG, ECG, HR, skin temperature, and more. Notably, they offer an advantage in preserving users' privacy

compared to visual sensors or cameras. Nevertheless, these sensors come with specific limitations and challenges that need to be considered, such as:

Sensor Placement: The precise positioning of sensors is crucial in determining the precision of wearable-based applications. Ensuring consistent and optimal sensor placement can present challenges, potentially resulting in less accurate results [54]. Furthermore, the comfort factor is pivotal for sustained wearable usage. When designing a study, considerations such as sensor size, weight, battery life, and the possibility of skin irritation must be considered.

Battery Life: Most commonly known sensors and devices rely on batteries, restricting their use. Sustaining continuous data collection with battery power can be challenging, as frequent recharging or battery replacement can disrupt the data collection process or irritate subjects [55].

Data Integrity: The integrity or reliability of wearable sensors' data can vary depending on factors such as the quality of the sensor, the placement of the sensor, and the calibration. Some sensors may provide inaccurate or inconsistent readings due to their quality or inappropriate placement, resulting in misleading or unreliable information [56].

Managing Data Confidentiality: Wearable sensors generally facilitate the acquisition of sensitive or health-related confidential data. Protecting such data from unauthorized access, breaches, or misuse is essential. In fact, maintaining such data's privacy and security can be challenging because wirelessly connected devices can transmit it to an external system [57].

Data Synchronization: Accurate synchronization of multimodal sensors' data can lead to excellent prediction performances [58]. However, accurate data alignment becomes a complex task when sensors have different clock rates or introduce delays during the recordings. To overcome this challenge, techniques or algorithms for accurate time synchronization must be implemented.

Addressing these challenges requires a combination of expertise. However, when it comes to data analysis, system design, and development, expertise in signal processing techniques, statistical methods, and advanced algorithms is particularly important. In addition, after proper data preprocessing and synchronization, sensory data fusion should be used to achieve maximum accuracy.

1.3 Motivations

Literature highlights that physical and mental health states are naturally linked, and the presence of a physical health issue increases the likelihood of developing a mental health problem and vice versa. In the context of physical and mental health assessment, this thesis presents scientific work in two significant areas: no-invasive stomach hunger and satiety sensation detection using wearables and interruption-free human flow experience recognition, respectively. The following motivations (M_1 – M_7) for this thesis stem from the limitations of science and technology in the past:

(M_1) Stomach Sensations Detection Using Non-invasive Wearables:
According to the current literature, none of the available studies use non-invasive physiological measurements to detect stomach hunger and satiety sensations objectively, and no publicly available dataset exists. Most studies in this domain used blood glucose levels to predict these states; others employed wearables in different contexts. Therefore, new approaches investigating these states by utilizing non-invasive ubiquitous sensor technology and machine learning algorithms are necessary, which will help to understand this problem further and assist individuals in their daily lives.

(M_2) Identifying Adequate Wearables for Hunger and Satiety State Recognition:
Since no previous research is available on the most promising wearables for hunger and satiety detection, it is also well-established that multiple wearables can discomfort subjects. Therefore, identifying the appropriate wearables to achieve the best performances and reducing the discomfort level of the participants for future research is optimistic.

(M_3) Finding the Most Effective Features and Feature Extraction Approaches:
The adequate features of a practical feature extraction approach can lead to acceptable performance and vice versa. Therefore, in the context of limited or no research work on hunger and satiety state recognition, comparative analysis between feature extraction approaches and selecting the best features can provide additional insights into the topic and help achieve the most satisfactory performances.

(M_4) Human Flow Experience Recognition During Work Activities:
Human flow experiences can contribute to increased performance, well-being,

and productivity. The number of scientific methods for automatic human flow experience recognition is quite limited, and most of them focus on video game users' data, which can bring privacy concerns. In contrast, others performed poorly and faced interruptions. Therefore, new investigations for objective and interruption-free flow recognition using machine learning and wearables in the working context seem promising.

(M_5) **Finding the Most Effective Feature Extraction Approach for Flow Recognition:** Appropriate features of an optimum feature extraction approach facilitate achieving remarkable performance. In the context of flow experience recognition using machine learning, where there is not much literature available, determining the most compelling feature extraction approach to acquire the best performance is of significant interest. Moreover, a comparison of feature extraction approaches can also point in the direction of future studies.

(M_6) **Identifying the Best Wearables for Flow Recognition:** Multiple or inappropriate wearables can cause discomfort to participants during long-term data recordings, resulting in low-quality or erroneous information that can affect the system's performance. To address this issue in flow research, finding appropriate wearables to achieve the best results and limit participant discomfort levels seems promising.

(M_7) **Discovering the Relationship Between Emotions and Human Flow Experience:** Literature highlights that emotions can heavily influence the human flow experience. Finding the association between them is crucial, as it can aid in getting remarkable flow recognition performances by integrating emotion data using transfer learning-based approaches. In addition, it facilitates addressing the data scarcity issue and providing future research directions.

1.4 Own Contributions

The author's scientific research on this thesis's motivations, M_1, M_2, and M_3, has been published in MDPI's Sensors journal [22]. The scientific research related to the motivations M_4, M_5, M_6, and M_7 has also recently been published in Elsevier's Computers and Biology journal [59].

To summarize, the following points outline the scientific contributions (C_1 – C_7) of this thesis:

(C_1) **Stomach Sensations Detection Using Non-invasive Physiological Measurements:**
This contribution presents scientific methods carried out to recognize stomach hunger and satiety states using wearable physiological data and machine learning algorithms. First, physiological data is collected from subjects utilizing wearables such as the *Empatica E4* wristband, the *RespiBAN* professional device, and the *JINS meme* smart glasses. Then, a machine learning-based pipeline is designed to distinguish between hunger and satiety classes. The author's proposed method discriminated between both classes with an accuracy of 93.43% and an AF1 score of 87.86%. It is worth mentioning that the recognition of hunger and satiety using non-invasive physiological data is proposed for the first time in this thesis.

(C_2) **Proposing a Setup of Wearable Sensors for Hunger and Satiety Recognition:**
In the context of C_1, an RF algorithm was employed to perform a comparative analysis to identify the most suitable wearables based on accuracy and AF1 score. Experiments in this study revealed that the most discriminative features come from three distinct sensor modalities: *EDA*, *infrared Thermopile* (Tmp), and *Blood Volume Pulse* (BVP). These sensors are part of the *Empatica E4* wristband, which is the most influential device in this study and can be used as a standalone wearable for hunger and satiety state recognition.

(C_3) **Finding the Most Effective Features and Feature Extraction Approaches:**
Three scientific experiments are performed in the context of C_1: First, the deep feature learning and manual feature engineering approaches are compared to determine which is more effective. Second, the feature selection algorithms, such as *RF*, *XGBoost*, and *Boruta*, are compared. Third, identify the best features for detecting hunger and satiety based on the best feature selection approach. Experiments revealed that deep feature learning approaches do not always produce suitable recognition results, especially when the dataset is small. Results also showed that RF is the most compelling feature selection algorithm compared to others. Moreover, it was found that the 80th percentile of *EDA*, the average of *EDA*, the 80th percentile of *BVP*, the average of *Tmp*, and the average of *BVP* are the top five effective features for hunger and satiety recognition.

(C_4) **Human Flow Experience Recognition During Work Activities:**
This contribution proposes scientific methods for recognizing interruption-free

human flow experiences using non-invasive wearables and machine learning algorithms. At first, wearables like the *Empatica* E4 wristband, the *Emotiv Epoc* X EEG headset, and the Biosignalplux *RespiBAN* are used to collect physiological data from people while they do math and reading tasks. Then, a machine-learning pipeline is designed to automatically distinguish between flow and non-flow classes. The proposed transfer learning-based method distinguished between the two classes with a remarkable AF1 score of 74.92% and an accuracy of 75.10% compared to the literature.

(C_5) **Finding the Most Effective Feature Extraction Approach for Flow Recognition:** Three scientific experiments were conducted in this context. Specifically, flow recognition with eighteen (18) manually crafted features for each data segment of each sensor channel, feature learning using *DNNs* (i.e., MLP and CNN), and a transfer learning-based method using the *DEAP* emotion recognition dataset. Experiments revealed that the proposed CNN-based deep learning approach best suits flow recognition with flow data. However, the proposed CNN-based transfer learning technique outperforms the others when the emotion dataset is utilized as the source dataset for the high arousal vs. low arousal classification task. It is also worth mentioning that the proposed transfer learning-based feature extraction technique is introduced for the first time in this thesis.

(C_6) **Identifying the Best Wearables for Flow Recognition:** A comparative analysis was performed to identify suitable wearables based on accuracy and AF1 score, using data from each wearable device separately and by fusing all wearable devices' data. Experiments revealed that the *Emotiv Epoc X* — an EEG headset, outperforms the *Empatica E4* wristband and *RespiBAN* wearable in terms of flow vs. non-flow discrimination. However, higher performance was obtained when multimodal data (from all devices) was fused and used with the feature learning approach, indicating that data from all devices can be used to achieve high performance.

(C_7) **Discovering the Relationship Between Emotions and Human Flow Experience:** Literature hints that emotions (such as arousal and valence) can influence human flow experience. To find out how emotions and flow experiences are connected, experiments are carried out using the *DEAP* emotion recognition dataset and the *Physiological Sense Flow* (PhySF) dataset with CNN-based transfer learning. The results of these experiments revealed that low and excessive physiological arousal hinders the flow state, while moderate physiological arousal promotes it. In

contrast, physiological valence is not particularly crucial for flow state detection. Furthermore, with an accuracy of 75.10%, the proposed approach outperformed past results in this domain when the source task was high arousal vs. low arousal classification in a subject-independent manner.

1.5 Overview

There are two main Chapters in this thesis (Chapters 2 and 3), each of which describes the author's scientific research. The outline of the thesis is presented in Figure 1.3.

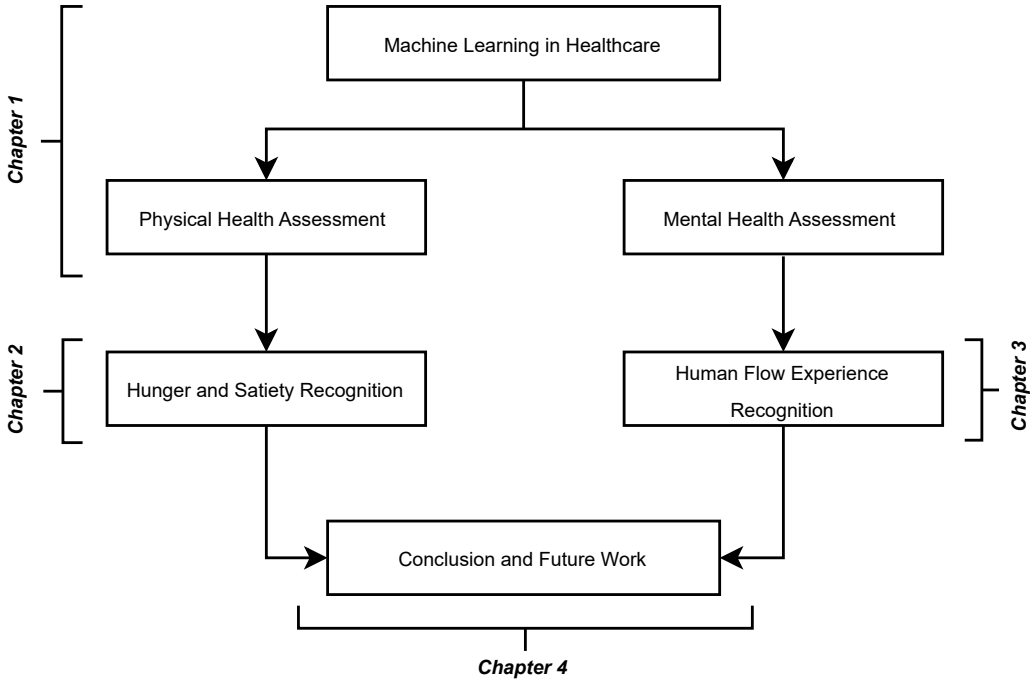

Figure 1.3: Outline of the thesis.

Chapter 2 presents the scientific work accomplished to recognize hunger and satiety states using non-invasive wearable sensors. It provides an overview of the related work in the context of hunger and satiety state assessment. Furthermore, it explains the data acquisition process of this study and presents a comparative analysis between different feature extraction approaches, either based on manual feature engineering or deep feature learning. Additionally, this chapter presents a comprehensive investigation conducted to identify the most suitable sensor channel, device, features, feature selection approach, and classifier for accurately distinguishing between hunger and satiety. Subsequently, it reveals the outcomes of the implemented methods.

Chapter 3 describes scientific work follow-through to address interruption-free human flow experience recognition during work activities (i.e., arithmetic and reading tasks). It presents an overview of related works in this field and a comprehensive comparative analysis of feature extraction approaches. In addition, this chapter also describes the author's proposed transfer learning-based technique implemented to enhance flow recognition results and address the data scarcity issue of this field using an emotion recognition dataset. Subsequently, it presents the results of the implemented approaches.

Finally, Chapter 4 highlights the findings described within the scope of this dissertation, focusing specifically on non-invasive hunger and satiety detection and interruption-free human flow experience recognition using wearable sensors. It concludes with an analysis of the work necessary to enhance and expand the proposed approaches, limitations of the current research, and a discussion of future research.

Chapter 2

Hunger and Satiety Recognition Using Wearable Sensors and Machine Learning

The perception of hunger and satiety is of great importance to maintaining a healthy body weight and avoiding chronic diseases such as obesity, underweight, or deficiency syndromes due to malnutrition. There are a number of disease patterns characterized by a chronic loss of this perception. Based on the literature, such stomach perceptions are challenging to recognize using non-invasive measurements. Aiming to develop an objective and effective hunger and satiety detection system — this chapter presents a hierarchical framework to recognize these states non-invasively using machine learning algorithms and multimodal wearable devices data such as the *Empatica E4* wristband, the *RespiBAN* wearable device, and the *JINS MEME* smart glasses. To accomplish this, physiological data is acquired from the five healthy, normal-weight subjects inactively sitting on a chair in a state of hunger and satiety since no public dataset related to this problem is available. After the data preprocessing, models are designed to recognize hunger and satiety states using two feature extraction approaches: *Feature Engineering* and deep *Feature Learning*. In addition, wearable devices, sensor channels, feature extraction methods, feature selection algorithms, and classifiers are compared to identify the most compelling of each, which facilitates recognizing hunger and satiety with remarkable accuracy. Parts of the content of this chapter are strongly inspired by my own publication [22].

2.1 Introduction

2.1.1 Background

The perception of hunger and satiety states occurs within the hypothalamic areas of the brain, which process several endocrine signals coming from peripheral organs such as the stomach, liver, pancreas, intestine, or fat [60]. During hunger, the blood glucose levels drop [61], and humans seek the food required for their nourishment. In contrast, satiety defines the feeling of fullness and satisfaction after eating. It can vary from person to person, meal to meal, and the composition and quantity of the food consumed. Nevertheless, differentiating between these perceptions is crucial to maintaining a stable body weight, preventing malnutrition, and developing healthy eating habits. Specifically, overweight and obesity are known to be associated with a gradually advanced loss of this perception, leading to overeating, underlying the disease [62]. According to the *World Health Organization* (WHO), 39% of adults aged 18 years and older were overweight, and 13% were obese in 2016 [63]. So far, standard methods to determine hunger and satiety are invasive, i.e., through hormonal analyses from blood samples or based on self-assessment, such as the *Visual Analog Scales* (VAS) [64, 65]. The latter records subjective sensations such as the desire to eat, hunger, satiety, and nausea [66, 67]. In contrast, invasive methods are primarily employed within controlled experimental contexts and are not viable for everyday application.

Furthermore, the recent progress in ubiquitous sensor technology, the computational capacities of machines, and the evolution of machine learning algorithms have ushered new research directions for researchers in diverse fields [21, 68–70]. With this progress, it now becomes feasible for researchers to acquire multimodal data from probands and design systems that can assist them in various application fields, including psychology, medicine, and biology [38, 45, 59, 71, 72]. Therefore, investigating ubiquitous sensor technology and machine learning in this domain can lead to more precise and timely diagnoses, reducing errors and improving patient outcomes. To solve a problem using machine learning involves a sequence of steps such as data acquisition, preprocessing, segmentation, feature extraction, and classification. Usually, these steps are optimized in parallel to yield the best classification performance. All of them are presented in detail in Section 1.1.3.

2.1.2 Current Challenges

As mentioned in Section 2.1.1, hunger and satiety are the fundamental perceptions that can play an essential role in maintaining a healthy body weight and preventing malnutrition. A fair distinction between them is essential for a healthy life. However, for patients facing the advanced loss of these perceptions, a non-invasive, objective, and easy-to-use method may assist them in training these perceptions [22].

In the highlights of the available literature, no state-of-the-art study reveals the feasibility of non-invasive hunger and satiety perception detection using wearable technology and machine learning algorithms. The current methods to analyze these states are either invasive, based on self-assessment tools, or not practicable for daily use. For example, the study by Krishnan et al. [64] focused on plasma satiety hormones to predict hunger and appetite-related VAS responses. Al-Zubaidi et al. [73] analyzed resting-state functional magnetic resonance imaging (rs-fMRI) signals to detect brain changes during hunger and satiety, and Maria et al. [74] classified audio signals into growling and burp sounds for hungry stomach detection.

The invasive methods are not without their disadvantages. For example, they require technical expertise, are costly, and have the potential for severe complications such as infection compared to non-invasive methods. Furthermore, invasive methods are mainly used in experimental settings because analyzing blood hormones daily related to hunger and satiety is not feasible. Moreover, imaging and audio signals can also bring privacy concerns.

On the other hand, the available self-assessment tools, specifically the VAS, are designed to quantify subjective stomach sensations, encompassing aspects such as the desire to eat, sensations of hunger and fullness, and the presence of nausea [66, 67]. Even though these tools can be helpful in some circumstances — they are not typically regarded as reliable. Because they are susceptible to external influences like the proband's stress, age, sitting environment, or temperature, which can affect the assessment's results.

In addition, there is currently no publicly accessible dataset on hunger and satiety perceptions. The data collection process is complex, demanding substantial resources and a high degree of expertise. In contrast, the availability of a public dataset can offer invaluable opportunities to scientists and researchers, facilitating their work and enabling more contributions to the field [75].

2.1.3 Research Motivation

Health is fundamental to human development. Nevertheless, most common health issues, particularly gastrointestinal symptoms and disorders, are linked with stomach perceptions or eating habits [76]. For instance, obesity and overweight are known to be associated with a gradually advanced loss of hunger and satiety perception, leading to overeating, underlying the disease [62]. Therefore, differentiating between hunger and satiety states is integral to overall health, maintaining a stable body weight, and stemming malnutrition.

Measuring hunger and satiety non-invasively using wearable sensors would allow a better understanding of this problem and the development of a system that may assist obese and overweight people in their daily lives [22]. However, there is currently no in-depth machine learning-based study in the related literature that aims to objectively recognize these states, and no public dataset related to this topic is available. Therefore, the current challenges mentioned in Section 2.1.2 and the importance of this topic in the context of health motivated the author to investigate this problem and fill the scientific gap by collecting multimodal sensor data and performing machine learning experiments in a subject-independent manner. In this research, it is hypothesized that modern non-invasive wearable sensors provide enough quality physiological data that can be used to distinguish hunger and satiety states objectively and non-invasively using supervised machine learning approaches.

The outline of this chapter is as follows: the most relevant state-of-the-art literature is presented in Section 2.2. Section 2.3 illustrates the author's contribution to recognizing hunger and satiety states using non-invasive physiological measurements. Section 2.4 describes the *Physiological Sense Hunger* (PhySH) dataset acquired for this study and the wearable devices used to collect this dataset. The data preprocessing steps for the *PhySH* dataset are presented in Section 2.5. Section 2.6 illustrates the implementation of two tested feature extraction approaches. Section 2.7 demonstrates the classification step. Section 2.8 explains the experiments and reveals the results. Section 2.9 provides a scientific discussion and the findings of this study. Finally, Section 2.10 concludes this chapter with a summary.

2.2 Related Work

In recent years, some hunger detection methods have been devised for clinical or behavioral assessments [64, 73, 74, 77–81]. Table 2.1 reveals the dataset information, feature extraction methods, detection (i.e., focused classes), and sensors or systems mainly used in the reviewed studies.

As mentioned earlier, non-invasive hunger and satiety state recognition employing multimodal physiological signals is challenging. There are few studies on this topic. Nevertheless, they are focusing on this in some other context, and their presented approaches cannot be applied to detect aforesaid states objectively and non-invasively. For example, Barajas-Montiel and Reyes-Garcia [81] applied traditional signal processing and pattern recognition methods to detect hunger vs. no-hunger cries and pain vs. no-pain cries from infant acoustic data. Their detection of hunger and no-hunger cries is based on acoustic features in the form of frequencies. Furthermore, the model proposed in this study [81] is specific to infants and cannot be generalized to the young and elderly population to recognize hunger and satiety states.

Interestingly, Maria and Jeyaseelan [74] considered growling sounds to identify the hungry stomach sensations. These sounds were synthetically collected using a mobile phone and compared with a growling sound obtained online. After the data preprocessing using smoothing methods and median filtering, features were extracted using manual feature engineering and deep feature learning approaches, mainly using an RNN. They claimed that an accuracy of 93.75% was achieved using RNN when classification was performed between the growling and burp sounds. However, in light of the literature, the validity of the online-obtained growling sound is questionable. Furthermore, the authors did not clearly mention how they split their dataset for the model training and testing phases.

Similarly, Krishnan et al. [64] performed a series of experiments to model sensations of hunger and fullness following food consumption by utilizing a multilayer feed-forward ANN — with a dataset relating concentration-time courses of plasma satiety hormones to VAS assessments. The authors asserted that their proposed model successfully predicted the VAS responses to different food compositions. Furthermore, the predicted VAS responses discriminated the satiety effects of high-satiating food types from less-satiating food types, both in orally fed and ileal-infused forms. However, obtaining plasma hormone levels frequently on a daily basis is not an easy task. It is

Table 2.1: Research publications found in the literature that attempted to assess hunger or related states.

Study	Sensors / System	Dataset Information	Features	Detection
[81]	Microphone	1627—samples of hunger and pain cries (acoustic data of infants)	Acoustic features by means of frequencies	Hunger cry, no-hunger cry, pain cry and no-pain cry
[74]	Microphone	Synthetically collected audio signals through mobile phones	SF, CDF and GCC	Growling vs. Burp sound
[64]	VAS	13—subjects plasma concentrations of satiety hormones from blood samples	Feature learning (ANN)	VAS responses from satiety hormone values
[77]	In vitro gastrointestinal model	Gastric viscosity and intestinal digestion from tiny-TIMagc	-	Appetite ratings of foods
[78]	Microsoft Band, Affectiva Q sensor, Microphone	8—subjects (3 female, 5 male) from 26 to 54 years	Statistical features	Time until the next eating event, and about-to-eat
[73]	fMRI	24—male subjects from 20 to 30 years (fMRI data)	3—features (DC, ReHo and fALFF)	Neuronal resting state alterations changes during hunger and satiety
[80]	EDA	35—patients (20 of them used as control group)	-	Hunger vs. Stress
[79]	EEG	EEG signals	-	Hunger, thirst, and rest-room sensations

VAS: Visual analog scales; ANN: Artificial neural network; fMRI: Functional magnetic resonance imaging; DC: Degree of centrality; ReHo: Regional homogeneity; fALFF: Fractional amplitude of low-frequency fluctuations; EEG: Electroencephalography; SF: Spectral features; CDF: Cepstral domain features; GCC: Gammatone cepstral coefficients; EDA: Electrodermal activity; tiny-TIMagc: In vitro gastrointestinal model.

time-consuming, and can cause infections compared to physiological signals collected by smart wearable sensor devices.

The study conducted by Bellmann et al. [77] confirmed that human clinical trials are time-consuming and costly. Therefore, a gastrointestinal model in conjunction with an ANN was considered to predict the sensations of hunger and satiety after ingestion of different meals. The proposed model is then trained with a series of training datasets to create a prediction set, and the model's measurements are linked to VAS scores for hunger and satiety prediction. The authors claimed that such gastrointestinal modeling using advanced machine learning algorithms has the potential to transform such models into powerful predictive tools that can predict physiological responses to food. However, miniaturized sensor-based acquisition of physiological responses is still state-of-the-art in this field.

With similar intentions, Rahman et al. [78] proposed that predicting eating events can enable users to adopt better eating behaviors. For this investigation, a set of sensor devices was used to record physical activity, location, HR, electrodermal response, skin temperature, and calories ingested while eight users ate. Their work uses 158 window-level engineered features followed by correlation-based feature selection (CFS) to train a classifier to predict the about-to-eat event. The time until the next eating event was predicted using a regression analysis. However, employing motion sensors such as accelerometers and gyroscopes is questionable for predicting the time until the next eating event. Additionally, the authors should have revealed the potential of the used sensor modalities, which can help determine the optimal sensor channel or device for predicting eating events.

The study by, Al-Zubaidi et al. [73] investigated the influence of hunger and satiety on resting-state functional magnetic resonance imaging (rs-fMRI) using connectivity models, that is, local connectivity, global connectivity, and amplitude rs-fMRI signals. The connectivity parameters of ninety brain regions are extracted and utilized in the sequential forward sliding window strategy — in conjunction with a linear SVM, to determine which connectivity model best discriminates between metabolic states (hunger and satiety). The authors asserted that the amplitude of the rs-fMRI signals is slightly more precise (compared to the other used models, such as local and global connectivity) to recognize resting brain changes during hunger and satiety sensations with an 81% classification accuracy.

In parallel, research conducted by Gogate and Bakal [80] emphasizes the importance of monitoring physical parameters such as hunger and stress in patients who are obstructed by muscle power loss diseases. To investigate this, a hunger and stress monitoring system (HS-MS) was developed by employing *Galvanic Skin Response* (GSR) sensors attached to the patient's index and middle fingers. Thirty-five patients, with 20 serving as a control group, were involved and found the HS-MS system to have an overall accuracy of 86.60% with a response time of 5 seconds. The authors claimed that their system (HS-MS) could be a beneficial tool, providing comfort to patients and applicable in hospitals or for elderly individuals living alone. However, they did not specify a method for data processing and feature extraction, nor did they use classical or modern classification methods.

Later, Lakshmi et al. [79] also conducted an investigation to detect hunger, specifically among physically disabled individuals. Their primary purpose was to establish a communication link based on the brain's thoughts, bypassing the need for muscle control. In this methodology, a single-channel EEG electrode was positioned on the person's scalp to detect sensations like hunger, thirst, and the need for the toilet by presenting images in front of them. The final outcome was determined by analyzing the individual's attention levels. Each image's attention level was compared using MATLAB, and the resulting attention value was determined.

In general, there are few studies [64, 73, 74, 77–81] which are exploring similar problems. However, each of the aforementioned studies has its limitations. For instance, Krishnan et al. [64] utilized an invasive data collection method, while the results of Bellmann et al. [77] relied on gastrointestinal models. Rahman et al. [78] employed motion sensors to determine the 'time until the next eating' event, which is questionable. Both Maria and Jeyaseelan [74] and Barajas-Montiel and Reyes-Garcia [81] utilized microphones to capture data, potentially posing privacy risks. Authors in [73, 79, 80] relied on manually crafted features, whereas feature learning can perform as well or better than state-of-the-art [82]. To date, no automated system for detecting hunger and satiety using multimodal physiological signals has been evaluated, and there is no publicly available dataset for this purpose.

2.3 Own Contribution

In the literature highlights, this is the first scientific study that focused on non-invasive physiological measurements to recognize hunger and satiety. This work has already been published as a journal article [22]. The contribution of this chapter (i.e., study) is to provide a hierarchical framework capable of recognizing hunger and satiety states objectively and non-invasively using ubiquitous wearables and machine learning algorithms. In addition to this, a comparison between feature extraction approaches, feature selection approaches, classification algorithms, sensor devices, and sensor channels is also presented.

To summarize, this chapter contributed the following:

1. Briefly review the related work and efforts in the context of hunger and satiety state recognition and summarize the progress of this domain.
2. Investigate the use of non-invasive multimodal sensor devices such as the *Empatica E4* wristband, the *RespiBAN* professional device, and the *JINS meme* smart glasses in the context of hunger and satiety recognition and develop a state-of-the-art machine learning model based on PRP, as shown in Figure 1.1, which learns hunger and satiety patterns from collected multimodal physiological signals and classifies them into their respective hunger and satiety classes.
3. Acquire, analyze, and compare wearable devices and their sensor channel data to select the most relevant physiological signals — to accurately classify hunger and satiety states from relevant non-invasive physiological signals.
4. Perform comparative analyses between feature extraction approaches, feature selection methods, and machine learning algorithms to identify the best feature extraction approach, the best feature selection method, the best features, and the best-suited algorithm to achieve highly accurate classification results.

2.4 Dataset Description

As highlighted in Section 2.1, no public dataset related to this problem is available. Thus, a dataset named *Physiological Sense Hunger* (PhySH) is acquired and investig-

ated in this study. The data collection process for the *PhySH* dataset[1] involved five healthy, normal-weight subjects. These individuals lack adverse lifestyle habits such as alcoholism or drug interventions. The detailed demographic information of the subjects who participated can be found in Table 2.2.

Table 2.2: Demographic data of each subject (S1–S5) who took part in the data collection process.

Subject Name	Gender	Age (in years)	Weight (in Kg)
S1	female	23	65
S2	male	29	71
S3	male	37	72
S4	male	26	81
S5	male	27	75

Kg: Kilograms; S1: Subject 1; S2: Subject 2; S3: Subject 3; S4: Subject 4; S5: Subject 5.

Subjects were instructed when enlisted — to refrain from eating for 16 hours before the data acquisition. However, drinking water was allowed. The subjects read and signed the informed consent form before the experiment. The wearable sensor devices used to acquire the physiological signals for hunger and satiety states are shown in Figures 2.1, 2.2, and 2.3.

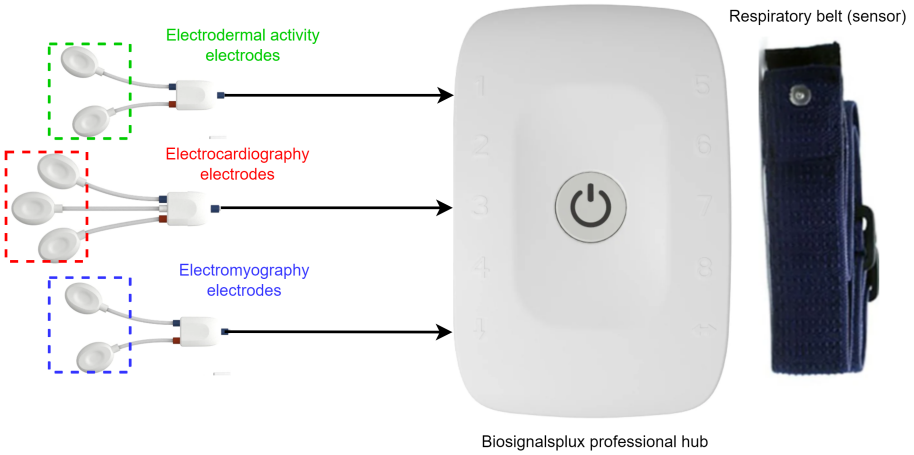

Figure 2.1: Overview of the RespiBAN (Plux Wireless Biosignals S.A., Lisboa, Portugal) [83] wearable device: the respiratory sensor is part of the adjustable belt, which can be worn around the chest or waist. The Biosignalsplux professional hub can be attached to the belt to include other sensor modalities such as Electrodermal activity (EDA), Electrocardiography (ECG), and Electromyography (EMG), which are highlighted in this figure.

[1] `https://osf.io/38qav/?view_only=bbeeb9b05a5f41e28fcac66a42240a10` (accessed: 20.11.2023).

Figure 2.2: Overview of the Empatica E4 wristband (Emaptica Inc., Cambridge MA, USA) [84] for the collection of various physiological signals, including highlighting different functional areas related to distinct sensor modalities. It is worth noting that this wearable also holds an accelerometer sensor, but its signals were not utilized due to their irrelevance to this study.

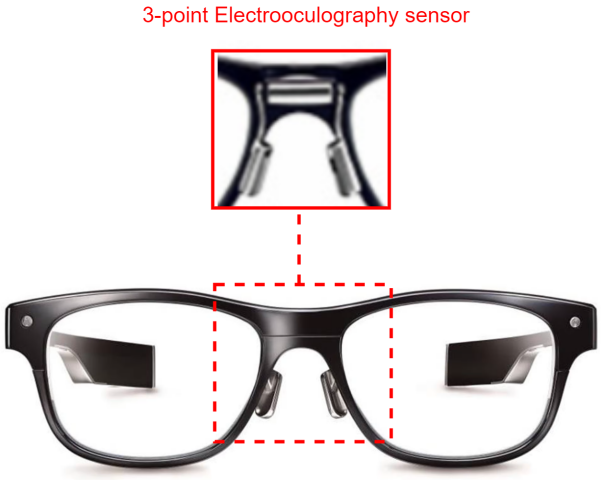

Figure 2.3: Overview of the JINS MEME smart glasses (Jins Inc., Tokyo, Japan) [85] for the collection of Electrooculography (EOG) physiological signals, including highlighting functional areas related to sensor modalities. It is worth mentioning that this wearable also contains an accelerometer and gyroscope sensor, but their signals were not utilized due to their irrelevance to this study.

The *PhySH* dataset was acquired while subjects were inactively sitting on a chair. The experiment began with a five-minute baseline recording, during which participants were advised to relax. The actual data acquisition step for each subject consisted of two phases, namely, the *hunger* and the *satiety* phase. In the hunger phase, data collection lasted five minutes. Subsequently, subjects were served with food so that they could eat and satisfy their hunger desire. After this step, the data acquisition process was resumed

for the satiety phase, which lasted for 30 minutes. The wearable devices utilized to gather the data and their corresponding sensor channels are described below:

1. **RespiBAN (Plux Wireless Biosignals S. A., Lisboa, Portugal)** [83]: Subjects wear the respiration belt with the Biosignalsplux hub on the chest, at thorax level, with the electrode connectors facing forward. The respiratory belt contains the respiration (Resp) sensor, and Biosignalsplux hub provides the possibility of connecting to other sensors such as EDA, ECG, and EMG, as shown in Figure 2.1. The description of these sensors is as follows:

    - **Resp**: The Resp sensor helps measure thoracic or abdominal circumference changes during respiration [86]. These measurements can indicate inhalation, expiration, and breathing resilience and can be used to extract breathing rates and determine breathing patterns. It is usually worn using a comfortable and flexible length-adjustable belt. It is sampled at 475 Hz.
    - **EDA** [87]: The Electrodermal activity, or EDA sensor, measures the electrodermal response, which is the change in the skin's electrical conductivity in response to sweat secretion. It is also sampled at 475 Hz. The EDA of RespiBAN (Eda_RB) consists of two electrodes placed on the front, in the middle of the index finger, and in the middle of the middle finger of the subject's non-dominant hand.
    - **ECG** [88]: The ECG sensor records the electrical impulses through the heart muscle and can also be used to provide information on the heart's response to physical exertion. The ECG sensor of the RespiBAN wearable device consists of three electrodes (as shown in Figure 2.1) placed on the subject's right upper pectoral, left upper pectoral, and left bottom thoracic cage. It is also sampled at 475 Hz.
    - **EMG** [89]: The EMG sensor measures the electrical activity associated with muscle contractions and the respective nerve cells that control them. The EMG sensor of the RespiBAN wearable device consists of two electrodes (as shown in Figure 2.1) placed on the subject's abdomen above the belly button. It is sampled at 475 Hz.

2. **Empatica E4 wristband (Emaptica Inc., Cambridge MA, USA)** [84]: The Empatica E4 is a comfortable wearable device designed for continuous, real-time, non-invasive, and unobtrusive monitoring of physiological parameters in daily

life. It contains Photoplethysmography (PPG), Tmp, and EDA sensors that allow the measurements of sympathetic nervous system activity and Heart Rate Variability (HRV). The description of these sensors is as follows:

- **PPG**: The PPG (as shown in the red area in Figure 2.2) is an optical sensor that emits light to measure the light absorption density of the blood vessels, which can be used to measure BVP, from which HR, HRV, inter-beat interval (IBI), and other cardiovascular features can be derived. It is sampled at 1 Hz.
- **Tmp**: The infrared thermopile sensor (i.e., Tmp) measures the peripheral skin temperature. As the blue color in Figure 2.2 indicates, it is a part of the Empatica E4 wristband. Metabolic activity, body temperature, and ambient temperatures can have a significant impact on this measure [90]. Nevertheless, it is important for clinical electrodiagnostic evaluations. In light of the literature, changes in skin temperature may be indications of various health conditions [91]. It is sampled at 5 Hz.
- **EDA**: The Electrodermal activity sensor of the Empatica E4 (Eda_E4) wristband (as shown with the green color in Figure 2.2) measures the electrodermal response, which is the change in the skin's electrical conductivity in response to sweat secretion. In light of the literature, strong emotional reactions, hunger, or stress affect the change in skin conductance [80, 92]. It is sampled at 5 Hz.

3. **JINS MEME smart glasses (Jins Inc., Tokyo, Japan)** [85]: JINS MEME is state-of-the-art smart eyewear (as shown in Figure 2.3) that is equipped with miniaturized EOG and motion sensors (such as an accelerometer and gyeroscope) to detect a user's eye movements and head motions. The EOG electrodes are placed in three locations on the frame. These electrodes can track blink duration and eye movements in different directions. Its design is almost the same as that of Wellington-type eyeglasses, with the intention that people can use it in their daily lives [93]. It is sampled at 20 Hz.

2.5 Data Preprocessing

Data preprocessing is an essential step of PRP (see Figure 1.1). It is crucial to ensure data quality, suitability, and effectiveness for successful modeling and analysis. This

process includes dealing with missing values, scaling, transformation, and data segmentation [94]. Addressing these issues can lead to a well-structured and consistent dataset, which, in turn, improves the machine learning models' performance and interpretability.

The *PhySH* dataset is acquired using three distinct devices encompassing n=22 sensor channels. More precisely, it consists of four channels of the *RespiBAN* wearable, eight channels of the *Empatica E4* wristband (including three accelerometers), and ten channels of the *JINS meme* smart glasses (including three accelerometers and three gyroscopes). The preliminary analysis identified that nine sensor channels of the dataset belong to motion sensors (i.e., accelerometers and gyroscopes). In light of the literature, motion sensors are well-known for movement monitoring, gait analysis, and fall detection. They may also be appropriate for other health-related sensitive assessments but not stomach sensation detection. Thus, in the first step, nine irrelevant channels were removed from the dataset, which, in turn, reduced the dimensionality of the *PhySH* dataset.

To ensure that all the sensor channels share a common repetition — in the second step, the PhySH dataset is synchronized, resampled to a frequency of 100 Hz, and linearly interpolated. This is because three wearables were utilized in the data collection process, each containing sensors of varying sampling frequencies.

Subsequently, each sensor channel's data was segmented employing a *Sliding Window Segmentation* (SWS) technique [95]. The SWS is a well-established approach for segmenting wearable sensor signals for pattern recognition systems. It involves dividing the sensor signals into fixed-sized time windows, with two possible configurations: non-overlapping and overlapping windows. Time windows do not intersect in the non-overlapping window technique, whereas in the overlapping window technique, they do. There is a general idea of the positive impact of using overlapping sliding windows on the performance of time-series-based recognition systems [96].

Adequately selected segmentation parameters of the SWS technique can facilitate achieving remarkable performances. Nevertheless, for time-series data, the present method for determining the best values of these parameters, i.e., window size (T) and step size (ΔS), is empirical [96]. Researchers generally test various combinations of T and ΔS to select the one that adds to the system's precision [97]. Therefore, regarding the available literature and based on the author's preliminary experiments, window

length T of sizes 10, 30, and 60 seconds were tested with a step size ΔS of 50% of window length T.

2.6 Feature Extraction

Feature extraction refers to selecting or transforming raw sensor data into specific values relevant to the problem to be solved (i.e., *features*). Approaches to feature extraction derive new feature values in a linear or nonlinear manner that capture essential information while discarding redundant or less valuable data. This study employs two feature extraction approaches, namely feature engineering and deep feature learning, as presented in Section 2.6.1 and 2.6.2, respectively.

2.6.1 Feature Engineering

The feature engineering approach, sometimes called manual feature engineering — involves crafting features manually, typically using specialized knowledge or basic transformation functions on sensor signals. In this work, $F = 18$ features are manually crafted [98, 99], also referred to as HCFs — which consist of the statistical and frequency-related values of the input signals or their power spectrum. All HCFs are listed in Table 2.3.

Table 2.3: List of Hand-Crafted Features (HCFs) that are computed independently for each data segment of each sensor channel.

Hand-Crafted Features	
Maximum	Minimum
Average	Standard deviation
Zero-crossing	Percentile 20
Percentile 50	Percentile 80
Interquartile	Skewness
Kurtosis	Auto-correlation
First-order mean	Second-order mean
Norm of the first-order mean	Norm of the second-order mean
Spectral energy	Spectral entropy

Each of the features listed in Table 2.3 is precisely characterized as follows:

- Maximum: It signifies the highest value in a feature vector. Let X refer to a feature vector containing multiple values for a particular time window. The function $Max(X)$ finds and returns the maximum feature value $x_i \in X$.

- Minimum: It indicates the smallest within a given feature vector. Let X be the feature vector having multiple values for a particular time window. The function $Min(X)$ determines and returns the minimum feature vector value $x_i \in X$.

- Average: It denotes the mean (μ) of all the X feature vector values. Let N indicate the total number of values in X, then the μ of X can be computed as:

$$Average(X) = \mu = \frac{1}{N}\sum_{i=1}^{n} x_i \tag{2.1}$$

- Standard deviation: It quantifies the amount of variability in the feature vector $X = \{x_1, x_2, x_3 \dots x_N\}$. Usually, it gives a more robust measure — particularly in conditions where outliers can bias the mean (μ) value. It is generally denoted by a sigma (σ) sign and can be computed as:

$$Std(X) = \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2.2}$$

- Zero crossing : This measure indicates a count, i.e., the number of times a signal crosses its median. For a given feature vector X, it can be estimated as [98]:

$$Zc(X) = \mid x_i < Median(X) < x_{i+1} \mid + \mid x_i > Median(X) > x_{i+1} \mid \tag{2.3}$$

- Percentiles (20th, 50th, and 80th): A percentile value indicates a number below which a specified percentage (%) of the feature vector (X) values falls. It gives further insights into the X values, i.e., how values are distributed. For example, the 20th percentile means that 20% of the values of X are smaller than that value, and 80% of the values are higher than that value. The 50th percentile is usually the median, and the 80th percentile means that 80% of values are smaller than that value, and 20% are higher.

- Interquartile: It computes the difference between the third quartile (i.e., 75th percentile) and the first quartile (i.e., 25th percentile) and is known as the interquartile range.

- Skewness: It indicates the degree of asymmetry in X feature vector values for a particular time window. A high skewness value represents a lack of symmetry below and above the μ of X and vice versa. It can be calculated as [98]:

$$Skewness(X) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^{\frac{3}{2}}} \tag{2.4}$$

- Kurtosis: It indicates the shape of the feature value distribution, in contrast to skewness, which considers the symmetry of the distribution. It considers the amount of peakedness (or, conversely, the amount of flatness) of the distribution toward the mean. For a feature vector X, it can be computed as [98]:

$$Kurtosis(X) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^{3}} - 3 \tag{2.5}$$

- Auto-correlation: It measures the degree of similarity between a given time series data and its delayed (lagged) version across consecutive time intervals [100]. The k-lag auto-correlation (r_k) can be formally defined as follows:

$$r_k = \frac{\sum_{i=1}^{N-k}(x_i - \mu)(x_{i+k} - \mu)}{\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2.6}$$

- First-order mean: It denotes the smallest sample value, x_1, within a sorted or arranged feature set X.

- Second-order mean: It represents the second smallest sample value, x_2, within an arranged feature set X.

- Norm values: The norm values typically denote the distance of a given feature vector from its origin. Two common norm measures, i.e., the L_1-norm, also known as the Manhattan distance, and the L_2-norm, generally referred to as the Euclidean

distance [101], are utilized to compute the norm of the first-order mean and the norm of second-order mean, respectively.

- Spectral energy: In essence, a signal is a function of varying amplitude over time. The frequency spectrum provides a second view of it, which indicates how much a signal lies within each given frequency band over a range of frequencies. The *Fourier Transform* method is a common approach to transform a time-varying signal into its frequency spectra. It can be implemented easily through the *Fast Fourier Transform* (FFT). The spectral energy measure quantifies the signal energy across different frequencies in a frequency domain representation. It can be calculated as the sum of the squares of the magnitude of the frequency content $F(n)$ [98]. That is,

$$S_E = \sum_{i=1}^{N} F(n)^2 \tag{2.7}$$

Here, S_E means spectral energy. It can alternatively be calculated using the normalized frequency spectrum, as depicted in Equation 2.8.

$$NS_E = \sum_{i=1}^{N} \hat{F}(n)^2 \tag{2.8}$$

Here, NS_E means the normalized version of S_E and $\hat{F}(n) = \frac{F(n)}{\sum_{i=1}^{N} \hat{F}(n)}$ represents the normalized frequency spectrum.

- Spectral entropy: It quantifies the signal's entropy within the frequency domain. Typically, prior to spectral entropy computation, the signal's frequency content is normalized. Then, it can be calculated by using Equation 2.9.

$$S_{EN} = -\sum_{i=1}^{N} \hat{F}(n) \times \log(\hat{F}(n)) \tag{2.9}$$

Here, S_{EN} denotes spectral entropy.

All the 18 hand-crafted features (HCFs) mentioned above are computed independently for each axis of each sensor channel and then concatenated to obtain a feature vector (X) of size $F\times$ sensor (S) channels (18 $\times$ 13 = 234). To remove the effects of discrepancies between the values of each feature, min-max normalization was performed for each feature to project its values into the interval [0, 1]. The normalization constants calculated on the training set were again used to calculate the features in the test set.

After the preliminary experiments, feature selection is performed on the HCFs to remove useless or redundant features and to decrease the classification model's complexity. In light of the literature, feature selection can improve the performance of a model and determine the interdependence between features and class labels [98]. A common approach for feature selection is feature ranking, which quantifies the ability of a feature to predict the desired class.

Three algorithms, namely the RF, the XGBoost, and the Bruta are used separately to select the best features and provide a comparative analysis between the feature selection algorithms on hunger and satiety data. The RF is a commonly used algorithm for feature importance ranking and selection. It has outperformed other known methods, such as *Recursive Feature Elimination* (RFE) and *Bruta* in the literature [102]. The RF is a tree-based learner (as mentioned in Section 1.1.3, Chapter 1) that generally grows by applying the classification or regression tree method (CART) [103], where binary splits recursively partition the tree into homogeneous or nearly homogeneous terminal nodes. After a fair split, the data moves from the root tree node to the child nodes, improving the homogeneity of the child nodes relative to the parent node [104]. Typically, RF consists of a set of hundreds of trees, where each tree grows using a dataset sample. However, RF can be slow in training when used with a large number of trees and is sometimes unsuitable for many sparse features [27, 105–107].

In the RF, trees are generally grown non-deterministically following a two-step randomization procedure. Apart from the randomization applied by growing the tree using a sample of the primary data, a subsequent level of randomization is set at the node level as the tree grows. This two-step randomization aims to decorate the trees so that the RF ensemble has low variance. Features ranked by RF are based on the quality of the purity improvement (the fraction of data items that belong to the class) of the node. Given a node n and the estimated class probabilities $p(k|n) \quad k = 1, ...C$. The Gini index can be defined by using the following equation [108].

$$G(n) = 1 - \sum_{k=1}^{C} p(k|n)^2 \tag{2.10}$$

In Equation (2.10), C is the total number of classes. In order to obtain the Gini index-based measure at each node, the Gini index decline is calculated for the variable used for partitioning. The Gini index-based measure of variable importance is then obtained by the average drop in the Gini index.

Similar to RF, XGBoost is an ensemble learning algorithm capable of classification, regression, and feature selection tasks. It is based on a gradient-boosting framework, which sequentially builds an ensemble of learners (i.e., DTs), where each learner attempts to correct the errors of the learners that came before it. XGBoost is less prone to overfitting, can handle missing values, and has limited sensitivity to outliers [30, 109].

Boruta is a wrapper feature selection approach that is based on RF. It selects or eliminates features after computing the feature importance score. The quality of its feature selection depends on the DTs of the RF algorithm. However, using an RF with a large number of DTs can increase Boruta's sensitivity. Increasing the number of trees in the RF may increase its computation time, which limits its use for analyzing very large datasets [110, 111].

To make a fair comparison between the above-mentioned feature selection algorithms in this study — the author selected the best 18, 54, 72, 90, and 108 features (from a total of 234 features) with Boruta, XGB, and RF and classified them with XGB and RF classifiers. The results of these experiments, including each class's accuracy and AF1 score, are presented in the experiments and results section. Nevertheless, the scientific discussion section illustrates the best features enabling remarkable performances.

2.6.2 Feature Learning

Feature learning involves learning features from labeled input data in an automated way without any human input. It has become increasingly popular over the past years with the popularization of ANNs and DNNs. During training, they are fed with raw input data to learn a mapping against each class in an end-to-end fashion. The ANN and DNN models

have been shown to perform well on various tasks (e.g., image classification [112], activity recognition [45, 99], and sleep stage classification [72]). However, training such models can be challenging as it is computationally more expensive than training traditional models. Moreover, finding optimal architectures is a non-trivial process.

In the past, Multi-Layer Perceptrons (MLPs) [113] and Convolutional Neural Networks (CNNs) [114] have been used for various tasks. The MLPs represent the most primitive type of ANN. In order to process 2D sensor data with its sensor axis (S) and time (T), the input data are first normalized using the batch-normalization layer [115] and then passed to fully connected layers that expect 1D input. A syntactic example of the MLP architecture can be seen in Figure 2.4.

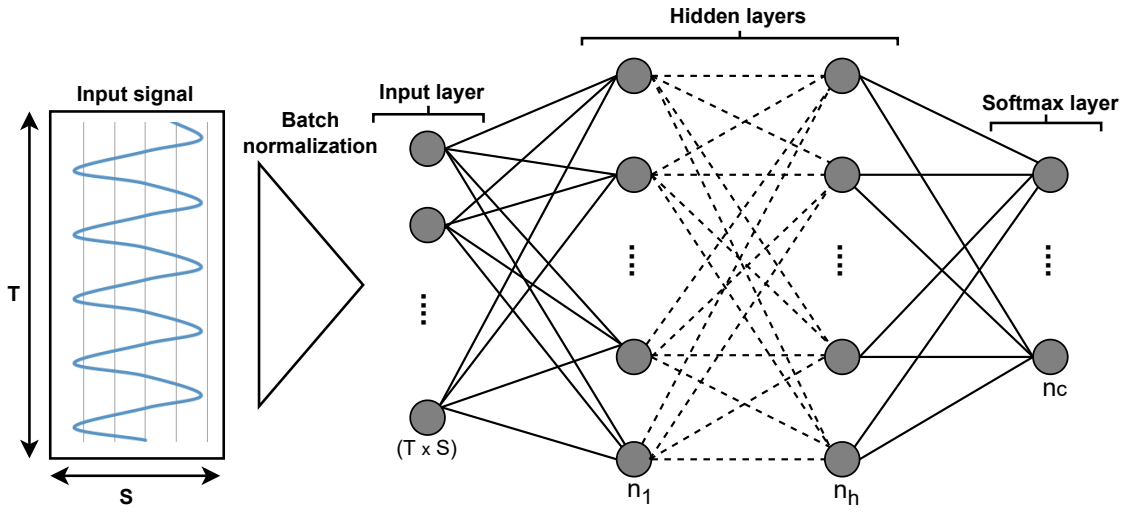

Figure 2.4: Illustration of the Multilayer Perceptron (MLP) model with input layer, hidden layers, and output classes represented by the softmax layer. Physiological signals of various sensor channels (S) of a window length (T) are converted into a (T $\times$ S) dimensional vector, which is passed from the input layer to different queued hidden layers (h) and the softmax layer in the last, for the classification of the features learned by the hidden layers. Here, n_1, n_h, and n_c represent the neurons of the 1^{st} hidden layer, neurons of the h^{th} hidden layer, and neurons of the softmax layer, respectively.

In CNN architectures, the convolutional layers are the main building blocks usually used to perform convolutional operations between one or several convolutional filters (or kernels) learned during the training phase and the layer input. The convolution operation can be applied by sliding the convolution kernels over the input data. In this study, raw sensor data are given as 3D input (S$\times$ T$\times$ 1) to the CNN model for processing. After a series of convolutional and pooling layers, the output of the last convolutional layer is smoothed into a 1D vector and fed into the softmax layer. The *Rectified Linear Unit* (ReLU) is the most commonly used activation function for convolutional layers. It is also

common to add multiple dense layers of a multilayer perceptron to the CNN architecture for classification problems. In that case, a softmax activation function aids in connecting the aftermost dense layer to the output layer. An example of a CNN model can be seen in Figure 2.5.

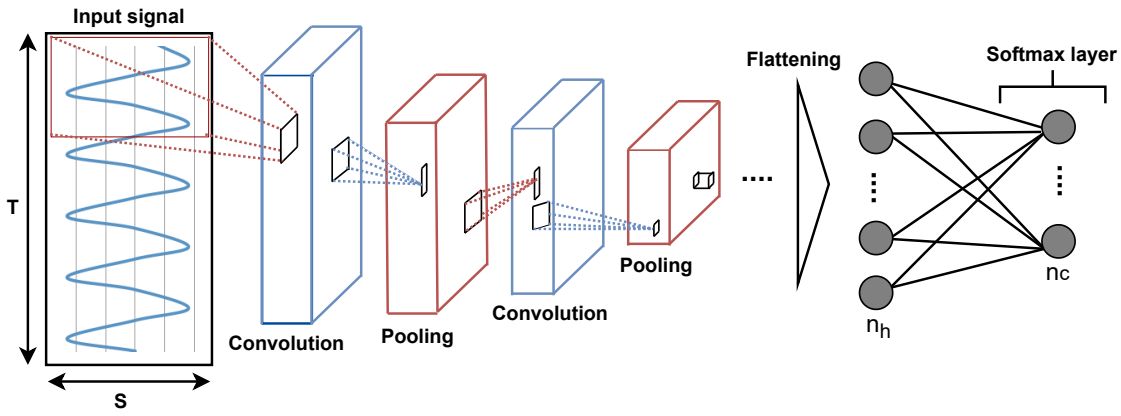

Figure 2.5: Illustration of the Convolutional Neural Network (CNN) model with convolutional layers, pooling layers, hidden (dense) layers, and output classes represented by a softmax layer. Here, S denotes the physiological signals of various sensor channels in a window length T, and n_h and n_c represent the neurons of the hidden and softmax layers, respectively. Convolutional and pooling layers process the input signals (data) to extract profound features. After the flattening step, learned features are passed to the hidden layer for further processing or classification through the softmax layer.

Since no automated method for the optimization of DNN hyper-parameters has been found so far, trial-and-error was used to obtain the best hyper-parameters for the DNNs that are tested in this study. The hyper-parameter values that were used in the feature-learning-based experiments (i.e., MLP and CNN) are provided in Tables 2.4 and 2.5, respectively.

Table 2.4: MLP architecture with learning rate set to 10^{-4}.

Layer Name	**Neurons / Dropout Rate**	**Activation**
Dense	64	ReLU
Batch Norm	-	-
Dense	16	ReLU
Dropout	0.5	-
Flatten	-	-
Dense	8	ReLU
Dropout	0.5	-
Dense	2	Softmax

Table 2.5: CNN architecture with a fixed dropout rate of 0.5 and learning rate of 10^{-4}.

Layer Name	No. Kernels (Units)	Kernel (Pool) Size	Stride	Activation
Convolutional	64	(1,1)	(1,1)	ReLU
Batch Norm	-	-	-	-
Convolutional	32	(1,1)	(1,1)	ReLU
Convolutional	16	(1,1)	(1,1)	ReLU
Flatten	-	-	-	-
Dense	2	-	-	Softmax

These experiments aimed to test feature-learning methods with a dual objective. The primary goal was to analyze the efficacy of MLP and CNN for automatically extracting features. The secondary objective was to examine and compare the results of human-crafted and machine-learned features. The results of classifying hunger and satiety using the approaches mentioned above are presented in the experiments and results section.

2.7 Classification

As mentioned earlier (see Section 1.1.2), classification is a supervised machine learning method that categorizes the data into distinct classes using a classification algorithm called classifier. The primary goal of a classifier is to learn patterns within the dataset during training and use that knowledge to predict the class to which new, unseen data belongs. Usually, it is common practice to extract features from the labeled dataset before the classification. Typically, machine learning algorithms work with two types of parameters, namely learnable parameters and hyper-parameters. Learnable parameters are those that the algorithms learn themselves during training, while hyper-parameters are specified by engineers or scientists prior to the training in order to regulate how algorithms learn and change the performance of the model.

Taking into account the significant impact of hyper-parameters on the overall model performance. The HCFs (see Section 2.6.1) are computed employing various settings of SWS, such as a window size (T) of 10, 30, and 60 seconds with a step size (ΔS) of 50% of the T, since no state-of-the-art method to search for these hyper-parameters exists [96]. Traditional classifiers such as SVM, DT, and RF are trained separately for each setting of T and ΔS in a *Leave One Subject Out* (LOSO) cross-validation configuration to select the best values for the aforementioned parameters. The

performance of each classifier in terms of accuracy for each tested value of T and ΔS can be seen in Figure 2.6.

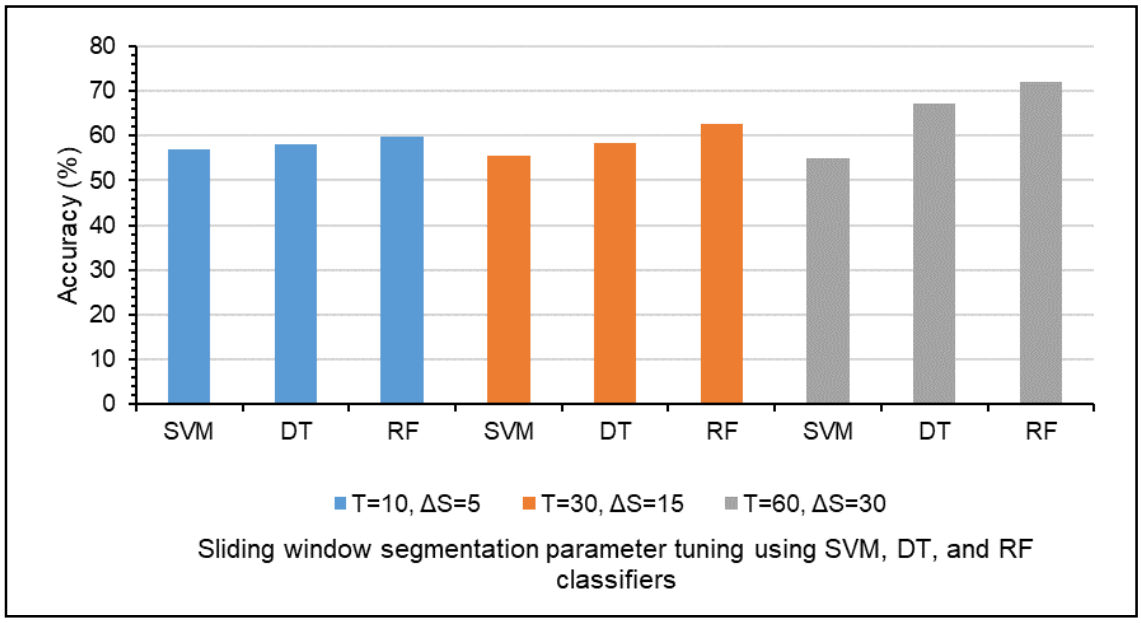

Figure 2.6: Performance of the SVM, DT, and RF classifiers for the selection of the best-performing classifiers and optimized values of the window size and step size.

As shown in Figure 2.6, RF outperformed in each tested setting compared to the other used classifiers in terms of accuracy. In particular, it showed good performance when T and ΔS were set to 60 and 30 seconds, respectively. In contrast, SVM and DT did not provide any significant information about these parameters. All the results of the feature engineering approach in terms of accuracy and the AF1 score of each class are presented in Section 2.8.

Since no automated method for optimizing the hyper-parameters of DNNs has proven its effectiveness in practice so far, the best values for these parameters in this study were determined through the trial-and-error method. The SWS parameters, such as T and ΔS, determine how long in time the input of the network is and how much time needs to pass between two consecutive windows of data. These parameters control the rate at which the learning algorithm picks up new information. The learning rate (lr) hyper-parameter is also crucial in addition to T and ΔS. The lr determines the rate at which the back-propagation algorithm updates the network weights during each training iteration. More specifically, each neural weight w_n at iteration $n \in \mathbf{N}^*$ is updated following the formula:

$$w_n = w_{n-1} - lr \times \frac{\partial L}{\partial w}(w_{n-1})$$

where L designates the loss function comparing the network outputs to the expected outputs.

In initial experiments, various configurations for the T, ΔS, and learning rate (lr) parameters were analyzed. The performance of these tested configurations is shown in Figure 2.7.

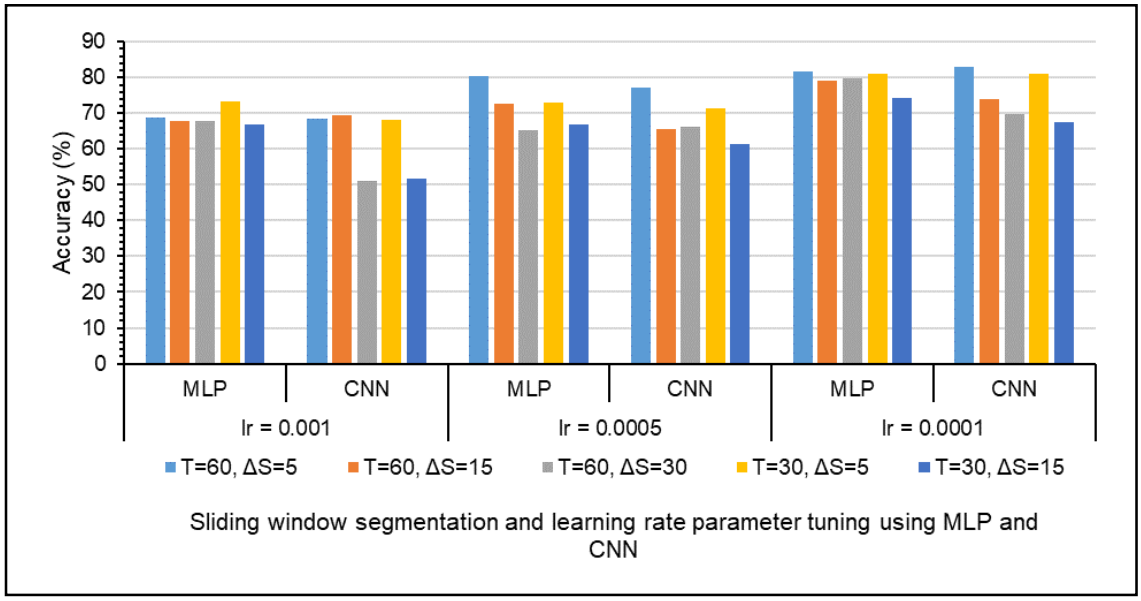

Figure 2.7: Illustrate the selection of hyper-parameters for the implemented feature learning approaches, MLP and CNN. lr: learning rate; T: window size; ΔS: step size.

The aforesaid probe revealed that the hyper-parameter values such as T = 60s, ΔS = 5s, and lr =10^{-4} (i.e., 0.0001) are significant for MLP and CNN models of this study. All hunger and satiety state classification results of the author's implemented feature learning approaches in terms of the accuracy of each class, overall accuracy, and AF1 score are shown in the experiments and results section.

2.8 Experiments and Results

In this study, all machine learning algorithms and deep learning models were implemented using Python 3.9. For the algorithms such as SVM, DT, and RF and the

deep learning models such as MLP and CNN, the libraries Sklearn and Keras with Tensorflow 2.2.0 backends were used. Adaptive Moment Estimation (ADAM) [116] was chosen as the optimizer for the deep learning model with an initial learning rate of 10^{-4}, and trained with 50 epochs at a batch size of 32. The categorical cross entropy was used as the loss function for the deep learning models. It is worth mentioning that it was decided not to report the result of a single *LOSO* cross-validation but the average results obtained after performing it five times.

2.8.1 Feature Engineering

Preliminary experiments with all HCFs (i.e., without feature selection) using SVM, DT, and RF classifiers were carried out to determine the best segmentation parameters. The results of these experiments are shown in Table 2.6. It can be seen that the best-performing configuration is obtained when using RF with T = 60s and ΔS = 30s, and it largely outperforms the others that were tested. Therefore, the author selected these segmentation parameters and the RF classifier for the rest of the experiments. However, the overall classification results remain mediocre, with an AF1 score of around 60%.

Table 2.6: Results of binary classification of hunger and satiety.

Classifier	Win Size (*T*)	Step Size (ΔS)	Acc. Hungry	Acc. Satiety	Acc	AF1 Score
SVM	10	05	20.90	70.37	56.89	45.63
DT	10	05	27.94	70.40	58.04	49.17
RF	10	05	30.97	71.75	59.90	51.36
SVM	30	15	21.61	68.86	55.43	45.24
DT	30	15	21.93	71.54	58.29	46.73
RF	30	15	**38.59**	73.23	62.71	55.91
SVM	60	30	13.19	69.50	55.00	41.34
DT	60	30	18.44	79.43	67.14	48.93
RF	**60**	**30**	36.36	**82.05**	**72.00**	**59.21**

DT: Decision tree classifier; RF: Random forest classifier; SVM: Support vector machine classifier; Acc: Accuracy; AF1 Score: Averaged macro F1 score.

To improve the initial classification results (see Table 2.6) and verify the potential of each sensor channel for hunger and satiety detection — experiments were also conducted with each sensor channel separately in the LOSO cross-validation manner. Figure 2.8 shows

the boxplot, which represents the mean and standard deviation (in dotted lines) of the obtained accuracies for each sensor using the RF classifier.

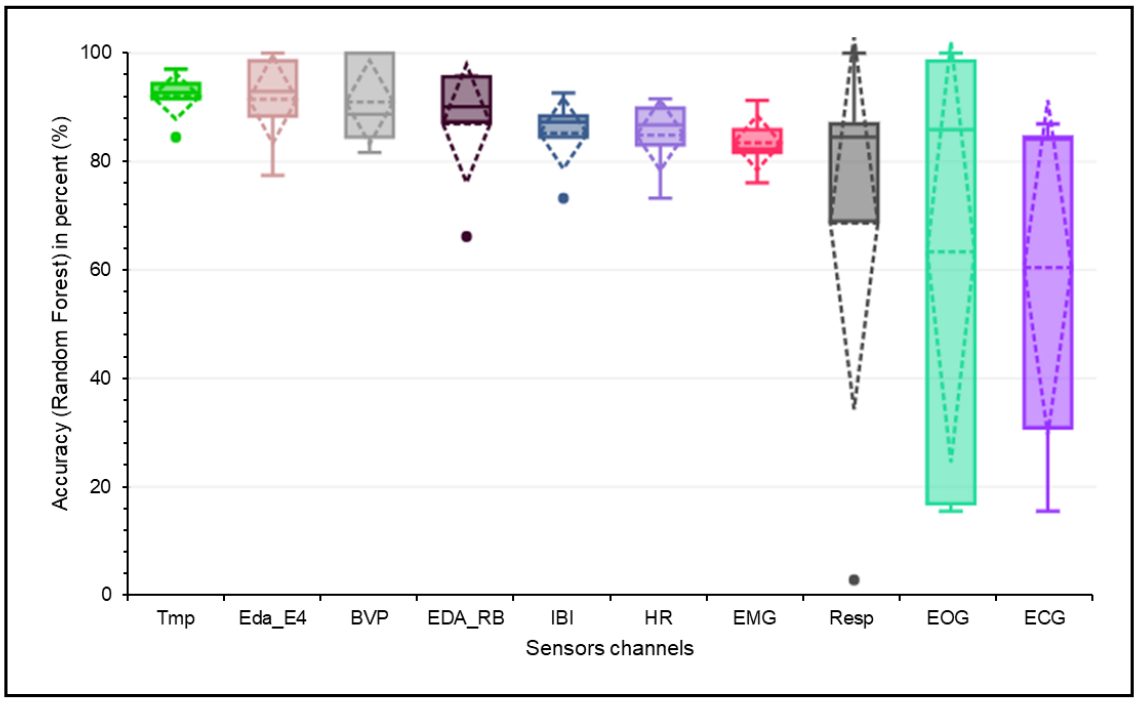

Figure 2.8: Importance of each sensor channel in recognizing hunger and satiety.

The standard deviations of Resp, ECG, and EOG are higher compared to the other sensors. The results in Table 2.7 also show that these sensors are the least significant because their accuracy is less than 70%, there seems to be a huge variance among the different subjects. Therefore, the author excluded the Resp, ECG, and EOG sensor data for further experiments. Moreover, the literature also confirms the importance of Tmp, BVP, and EDA (Eda_E4 and Eda_RB) signals in the detection of hunger. For example, the research of Mandryk and Klarkowski [117] reveals that BVP increases in response to hunger and decreases in response to relaxation. He et al. [118] identified changes in Tmp, EDA, and HR values following the ingestion of food. The authors in [80] had already used EDA for hunger detection. Furthermore, IBI and HR are directly related to BVP since they are derived from it.

To achieve highly precise results for hunger detection, further experiments were performed with the best 18, 54, 72, 90, and 108 features of the selected sensor channels (i.e., excluding Resp, ECG, and EOG), ranked by their increasing Gini impurity scores. With the best 18 features, an accuracy of 93.43% and an AF1 score of 87.86% were obtained, as shown in Table 2.8.

Table 2.7: Hunger and satiety classification results on each sensor channel using RF classifier.

Sensor	Acc. Hungry	Acc. Satiety	Acc	AF1 Score
Tmp	73.08	95.30	92.00	84.19
Eda_E4	70.59	94.98	91.43	82.79
BVP	67.35	94.68	90.86	81.02
Eda_RB	62.18	92.25	87.14	77.22
IBI	43.48	91.95	85.14	67.46
HR	40.45	91.33	84.86	65.89
EMG	30.95	90.58	83.43	60.77
Resp	29.30	79.56	68.29	54.43
EOG	39.25	73.25	62.86	56.25
ECG	21.59	73.66	60.57	47.63

Acc: Accuracy; BVP: Blood volume pulse; Eda_E4: Electrodermal activity sensor of empatica E4 wristband; Tmp: infrared Thermopile; IBI: Inter-beat interval; HR: Heart rate; Resp: Respiratory; Eda_RB: Electrodermal activity sensor of RespiBAN; ECG: Electrocardiography; EMG: Electromyography; EOG: Electrooculography. Note: For these experiments, a window size 60s and a step size of 30s were used to compute the 18 hand-crafted features for each axis of the sensor channel.

Table 2.8: Results of the classification of hunger and satiety using RF classifier based on the best features selected with feature importance ranking.

No. of Best Features	Acc. Hungry	Acc. Satiety	Acc	AF1 Score
18	**79.65**	**96.08**	**93.43**	**87.86**
54	66.02	94.14	90.00	80.08
72	68.18	95.42	92.00	81.80
90	68.00	94.67	90.86	81.33
108	67.33	94.49	90.57	80.91

Acc: Accuracy; AF1 score: The averaged macro F1 score is computed by taking the arithmetic mean of all the per-class F1 scores.

These results (as shown in Table 2.8) indicate that the best performances could be obtained with just 18 HCFs based on the FIR. Moreover, there is not much difference in the classification results of the best 54, 72, 90, and 108 features. Furthermore, the results with 18 HCFs are notably better than the results that were obtained using all sensors (see Table 2.6). It could be concluded that Resp, ECG, and EOG are the least informative sensors in this case, while BVP, Eda_E4, Tmp, HR, Eda_RB, and EMG are the most informative sensors and could be used to detect hunger and satiety.

In order to make a comparison between the manual feature selection approaches in this study, the author selected the best 18, 54, 72, 90, and 108 features with Boruta, XGBoost, and RF and classified them with XGBoost and RF classifiers. The best results of each classifier in each setting are shown in Table 2.9. The best configuration was

obtained using RF for feature selection and classification. This also confirms previous findings, as shown in Table 2.8.

Table 2.9: Results of the classification of hunger and satiety using RF and XGBoost classifier based on the best features selected with Boruta, XGBoost, and RF.

Classifier	FS Algo.	B. Features	Acc. Hungry	Acc. Satiety	Acc	AF1 Score
RF	**RF**	**18**	**79.65**	**96.08**	**93.43**	**87.86**
RF	Boruta	108	72.53	95.89	92.86	84.21
RF	XGBoost	54	73.12	95.88	92.86	84.50
XGBoost	RF	18	69.23	94.63	90.86	81.93
XGBoost	Boruta	54	53.33	93.11	88.00	73.22
XGBoost	XGBoost	18	63.92	94.20	90.00	79.06

B. Features: Selected best features; FS Algo.: Feature selection algorithm; RF: Random Forest; XGBoost: eXtreme Gradient Boosting; Acc: Accuracy; AF1 Score: Averaged macro F1 score.

Additionally, experiments were carried out to determine the relative relevance of each wearable device (i.e., the Empatica E4 wristband, the JINS MEME smart glasses, and the RespiBAN professional device, with ECG, EMG, and EDA sensors) in detecting hunger and satiety using the RF classifier. Figure 2.9 shows the results of each device using the best 18 features in each case. Experimental results show that Empatica appears to be the best wearable device, outperforms the other devices, and might be used as the only wearable device for monitoring hunger and satiety.

2.8.2 Feature Learning

To provide a comparative analysis between feature engineering and feature learning methods on the *PhySH* dataset, experiments were also performed using MLP and CNN (see Section 2.6.2) with the best hyper-parameter settings (see Section 2.7). The initial probes using feature learning were very disappointing, and models could not learn very well because of the class imbalance problem, as the PhySH dataset contains 5 minutes of data for hunger and 30 minutes for the satiety class. Therefore, further experiments were performed with 5 minutes of data from each class. With the CNN, an accuracy of 82.90% and an AF1 score of 82.54% were obtained, as shown in Table 2.10.

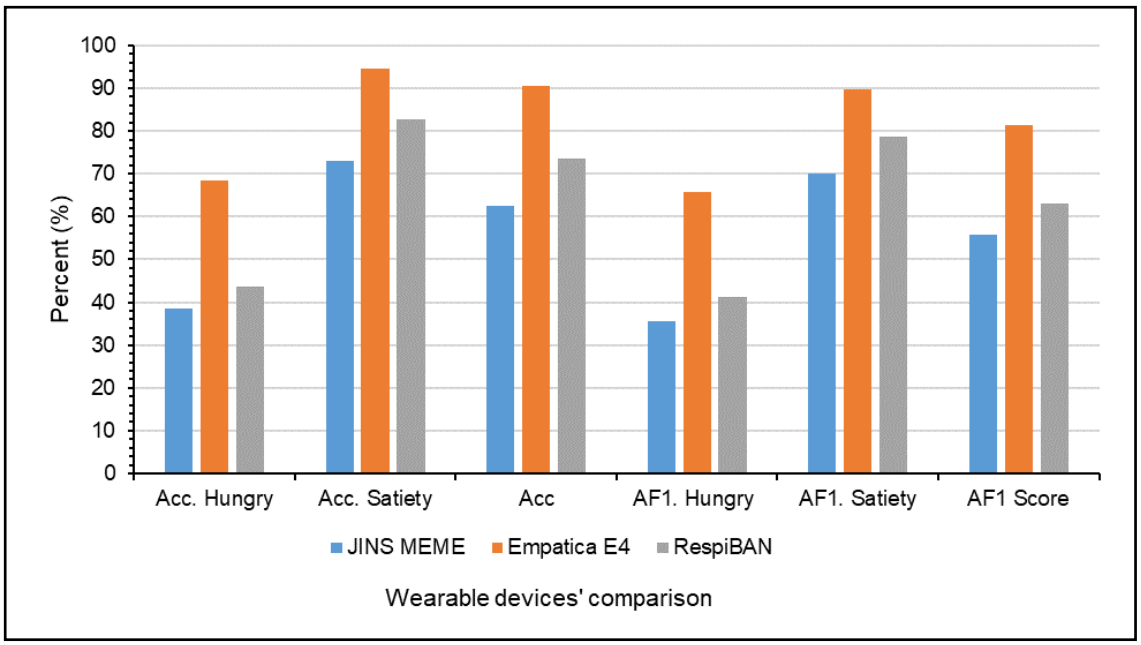

Figure 2.9: Comparison of sensor devices on the basis of accuracy (Acc) and macro averaged F1 score (AF1) for the detection of hunger and satiety non-invasively. Empatica: Empatica E4 wristband; JIMS MEME: JINS MEME smart glasses; RespiBAN: RespiBan professional wearable device, including ECG, EMG, and EDA sensors.

Table 2.10: Results of the classification of hunger and satiety using feature learning approaches.

Classifier	Acc. Hungry	Acc. Satiety	Acc	AF1 Score
MLP	77.79	81.35	80.14	79.57
CNN	**81.37**	**83.70**	**82.90**	**82.54**

Acc: Accuracy; CNN: Convolutional Neural Network; MLP: Multi-Layer Perceptron.

2.8.3 Comparison with the Literature

As far as is known, there is no publicly available dataset for hunger and satiety detection using physiological measurements, and most of the literature work in this field is focused on identifying something different. For example, Maria and Jeyaseelan [74] focused on identifying hungry stomachs using audio signals, and they classified the audio signals into growling and burp sounds and related them to hunger detection. Al-Zubaidi et al. [73] used resting-state functional magnetic resonance imaging (rs-fMRI) data to find changes in the brain's resting state that happen when a person is hungry or full. According to the available literature, there are only two studies that use physiological measurements to investigate something similar but in a different context [78, 80]. For example, Rahman et al. [78] predicted about-to-eat moments for

just-in-time eating interventions using similar types of physiological measurements, and Gogate and Bakal [80] investigated hunger and stress using GSRs. The results of these studies are shown in Table 2.11.

Compared to the literature[2] results reported in Table 2.11, this study achieved an improved accuracy of 93.43% and an AF1 score of 87.86% with HCFs using RF-based feature selection and classification. An accuracy of 82.90% and an AF1 score of 82.54% were achieved with the deep feature learning approach using the CNN architecture. The results of each of the implemented approaches (i.e., manual feature engineering and deep feature learning) in the form of a confusion matrix are shown in Figure 2.10 (a–c). It is also worth noting that this is the first scientific work to investigate hunger and satiety states using physiological measurements.

Table 2.11: Comparison of hunger and satiety recognition results obtained using the implemented approaches of this study with known results from the related peer-reviewed literature.

Approach	Acc. Hungry	Acc. Satiety	Acc	AF1 Score
Rahman et al. [78]	-	-	-	69.00
Gogate and Bakal [80]	-	-	86.00	-
HCFs	**79.65**	**96.08**	**93.43**	**87.86**
MLP	77.79	81.35	80.14	79.57
CNN	81.37	83.70	82.90	82.54

HCFs: It is the proposed feature engineering approach, in which hand-crafted features (HCFs) are computed to train an RF classifier on the best 18 selected features using the RF. MLP is the proposed feature learning approach using *Multilayer Perceptron* (MLP). CNN is the proposed feature learning approach using *Convolutional Neural Network* (CNN). Notably, promising results were only achieved with the proposed HCFs using the appropriate feature selection method.

[2] Note: A direct comparison between the results of this study and the results presented in the literature is not possible because of the differences in setups, datasets, etc., so this comparison is just provided for informative purposes to show the effectiveness (potential) of the results of this study.

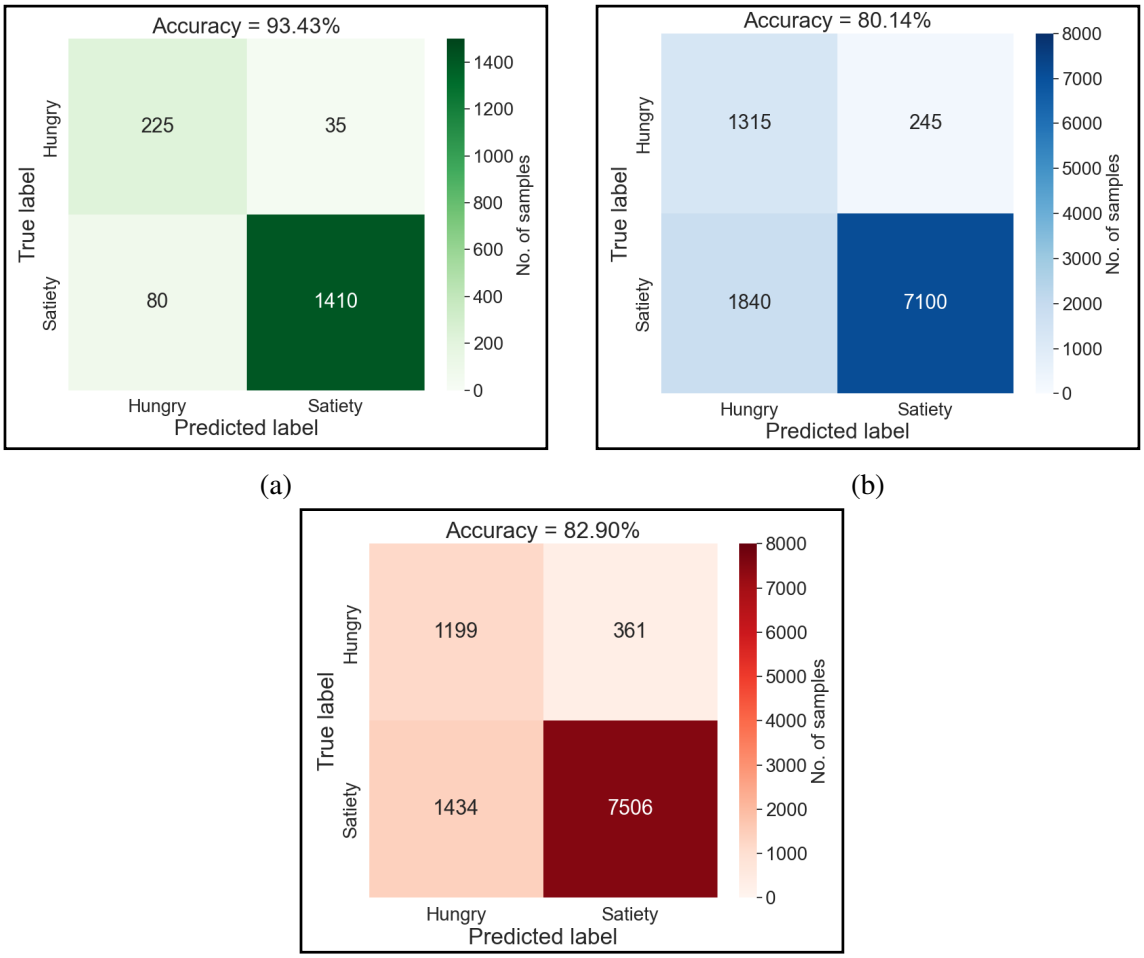

Figure 2.10: List of the confusion matrices (a–c) for hunger and satiety state recognition using three feature extraction approaches: (a) - feature engineering based on the computed HCFs, (b) – feature learning using MLP, and (c) – feature learning using CNN. In each figure (i.e., a–c), the top-left quadrant represents a true negative (t_n) value, the top-right quadrant represents a false positive (f_p) value, the bottom-left quadrant represents a false negative (f_n) value, and the bottom-right quadrant represents a true positive (t_p) value, respectively. All values (i.e., t_p, t_n, f_n, f_p) of the confusion matrices (a–c) are cumulative over five folds and the five runs of an experiment.

2.9 Scientific Discussion

The findings of the scientific experiment on recognizing hunger and satiety are discussed in detail in the following points:

- One of the main objectives of this research was to develop a machine-learning model to recognize hunger and satiety states using non-invasive wearable

physiological sensor signals. Therefore, wearable devices such as the *Empatica E4* wristband, *JINS MEME* smart glasses, and *RespiBAN* professional with ECG, EMG, and EDA sensors are utilized to acquire physiological data related to hunger and satiety states. The results of this study confirm that these devices provide sufficient quality data required to determine human health states. These devices can be employed to capture physiological signals related to the perception of hunger and satiety in patients or people with occupational constraints, as opposed to invasive [64], gastrointestinal models [77], fMRI-based data [73], and gastric tone signals [74]. Furthermore, the author's proposed non-invasive multimodal system with carefully selected sensor channels outperformed previous approaches with an accuracy of 93.43% and an AF1 score of 87.86%.

- Each classification algorithm is based on different mathematical models [119] and may produce different results for the same dataset. In order to obtain highly accurate results and select the best classifier for further experiments, the experiments are conducted with different classifiers, window sizes, and step sizes. It was found that the RF classifier was best suited for hunger and satiety detection using the manual feature engineering approach, and it outperformed the DT and SVM classifiers in each scenario. It was also observed that the window size of 60s and the step size of 30s were significant for each classifier.

- In the past, deep learning-based approaches have delivered promising results in a variety of application domains such as biology, medicine, and psychology [45, 72, 120–126]. However, they are computationally expensive and also require a large number of training samples [127] to build successful models compared to traditional approaches that use manually crafted features. To compare the results of feature learning and feature engineering approaches of this study — 18 features were computed independently for each axis of each sensor channel. They were subsequently concatenated to obtain a feature vector of 18 × sensor (S) axis size. It was found that well-engineered features can perform better than deep learning approaches in the case of a limited number of training samples.

- *Feature Importance Ranking* (FIR) was exploited to determine each feature's importance in this study. It measures each input feature's contribution to the model's performance. It turned out that the most accurate results can be obtained only with the best 18 HCFs (as shown in Table 2.8), and adding other irrelevant and redundant features can introduce noise into the data, reducing the classifier's performance. It can be pointed out that the top five features come exclusively

from three different sensor channels (Eda_E4, BVP, and Tmp) and are either computing the mean or the 80th percentile of the data values. The 80th percentile approximates the maximum value in a data segment that is less sensitive to noise or outliers than the actual maximum computation. This would indicate that the average and upper data values in Eda_E4, BVP, and Tmp are critical to distinguish between hunger and satiety. This feature selection also validates previous results achieved while identifying each sensor channel's importance (Table 2.7). It seems to confirm findings from the literature that ushered these sensor channels to be relevant in detecting hunger and satiety [80, 117, 118], as shown in Figure 2.8. The overall selected best 18 features that facilitate achieving highly precise results in this study can be seen in Figure 2.11.

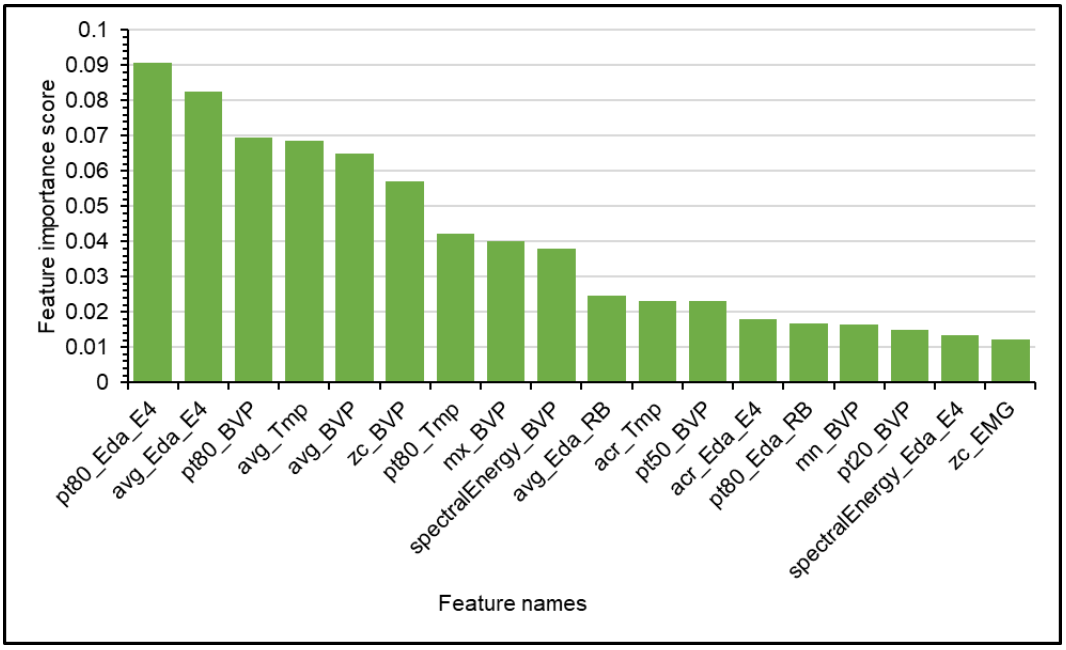

Figure 2.11: The overall 18 best features. Note: pt80: 80th percentile; avg: average; zc: zero crossings; mx: maximum; acr: auto-correlation; pt50: 50th percentile; mn: minimum; pt20: 20th percentile; BVP: Blood Volume Pulse; HR: Heart Rate; Tmp: Temperature; Eda: Electrodermal activity; RB: Respiration belt; E4: Empatica E4.

- Long-term monitoring with a large number of wearable sensors may be uncomfortable for users [128]. Therefore, eliminating irrelevant sensors can decrease the degree of discomfort and improve the robustness of the classification system by reducing the dimensionality. This can also save a lot of money [129]. Therefore, In this work, all sensor channels and wearable devices are compared to determine the most suitable sensor channel and wearable device for hunger and satiety detection. It was found that PPG (BVP, IBI, and HR), EDA (Empatica E4

and RespiBAN), Tmp, and EMG were the appropriate sensor modalities for this study, and Resp, ECG, and EOG were the least appropriate. Additionally, experiments revealed that compared to other used devices (see Figure 2.9), the Empatica E4 wristband performed better and stood out as the most suitable device for this study.

2.10 Summary

In this chapter, an objective and non-invasive machine learning model is introduced to recognize stomach hunger and satiety perceptions using multimodal physiological measurements. The proposed system enables the detection of hunger and satiety with an accuracy of 93.43% and an average AF1 score of 87.86% in the *Leave One Subject Out* (LOSO) cross-validation configuration. The results of this study lead to the following conclusions: Firstly, state-of-the-art wearable sensors provide good-quality physiological data on hunger and satiety and could be used to build a non-invasive and objective system. Secondly, deep learning models do not necessarily perform well, especially when there are a limited number of training samples. In addition, feature selection could help remove unnecessary and redundant features that lead to noise, leading to better results. Finally, the experiments of this study indicated that the most discriminative features come from three specific sensor modalities: *Electrodermal activity* (EDA), *infrared Thermopile* (Tmp), and *Blood Volume Pulse* (BVP). These sensors are part of the Empatica E4 wristband, which is the most influential device in this study and can be used as a standalone device. To acquire a deeper understanding of hunger and satiety perception, additional experiments with long-term measurements are necessary. These experiments will not only enhance the training of deep learning models but also enable a more detailed categorization of hunger and satiety, thereby providing valuable insights for future research.

Chapter 3

Human Flow Experience Recognition Enhanced by Transfer Learning Methods Using Emotion Data

Human flow experience, sometimes called flow experience, is a specific positive and affective state of mind that occurs when humans are completely absorbed in an activity and forget everything else. It can be distinguished from the rest of routine activities by some characteristics, such as a balance between the challenges of an activity and the skills required to face those challenges, regardless of whether an individual is an apprentice or an elite performer. This state can lead to high performance, well-being, and productivity at work. In the past, few studies have been conducted to determine the human flow experience using physiological wearable sensor devices. In contrast, other studies in this domain rely on retrospective self-reported data, which requires interrupting participants in their tasks. However, there is currently no in-depth study in the related literature that aims to recognize flow experience during working activities objectively, and no public dataset related to this topic is available. In addition, it is still unclear to the researcher how to circumvent the data scarcity issue in this domain.

Aiming to develop an objective flow recognition system and address the data scarcity issue in this field — this chapter presents a wearable-based flow experience recognition framework to address the said limitations. Parts of the content of this chapter are strongly inspired by the author's own publication [59] in Elsevier's Computers in Biology and Medicine journal.

3.1 Introduction

3.1.1 Background

Flow experience is a distinct positive and affective state of mind that occurs when a person is completely immersed in an activity [130]. It is characterized as a state of self-forgetting and absorption during a task with demands that seem to match one's own skills exactly [131]. In addition, enjoyment is described as a key component of flow [132], as flow is usually perceived as a positive and rewarding experience [131]. In the past, psychologists have highlighted the impact of flow on psychological well-being [133, 134], and flow has been shown to lead to positive outcomes in the workplace, such as high task performance and increased job satisfaction [135], which can bring benefits for both employees and employers [136]. The positive consequences of flow underline the relevance of measuring and recognizing it. However, the current standard for measuring flow consists of retrospective self-report questionnaires, which require interrupting participants in their tasks. This leads to an interruption of the flow experience and can restrict the potential positive consequences [137]. Therefore, psychological research has an ongoing debate about new methods to measure flow experience unobtrusively [137]. Measuring flow in an interruption-free manner would allow researchers to better understand the experience without interfering with the mechanisms themselves. In this way, flow situations can be better understood, anticipated, and actively created in order to benefit from the positive consequences of the experience. For instance, some questionnaire studies have already identified work tasks in which flow is experienced, such as planning, problem-solving, or evaluation tasks [138]. By measuring flow more precisely, flow situations could be better distinguished from non-flow situations. Then, work tasks could be distributed and planned in a way that helps to increase the flow.

Furthermore, advances in wearable sensor technology combined with supervised machine learning models allow researchers to align self-reported data with objectively measurable physiological signals to develop algorithms that automatically recognize emotional and motivational states. Therefore, the automatic recognition of human flow experience in real-time from body-worn physiological sensor signals using supervised machine learning is a worthwhile investigation since machine learning techniques [139] have already shown reliable performances on such sensor signals in a variety of

applications in different fields such as psychology, medicine, and biology [21, 22, 68, 71, 72, 99].

3.1.2 Current Challenges

The central challenge flow researchers are confronted with is that measuring flow with questionnaires interrupts the experience itself, and this interruption can lead to flow being gone for some time afterward [137]. In the worst case, flow research, as it is conducted in most cases, can result in less flow. To assess flow interruption-free and in real-time, new approaches to recognize it using physiological sensor data have to be applied and investigated [140].

A natural attempt at analyzing flow in an interruption-free manner is to use unobtrusive wearable sensors to collect physiological data and then apply machine learning to the collected data to train flow recognition models. However, wearable-based flow experience recognition is generally a complex task because there is no state-of-the-art method to obtain accurate class label information from experts, such as in the video or image analysis domain. In contrast, in flow research, participants typically provide class label information during or after the end of each task via a questionnaire [137]. In light of the literature, there is no public dataset available, and there are very few studies on this topic. The availability of a public dataset provides golden opportunities for scientists and researchers [75]. For example, it helps to reproduce the research results, enables scientists to build on others' work, and provides an opportunity to investigate the problem further, which can help improve accuracy or performance [141].

Most of the past flow literature studies focused on the game users' data to recognize flow when users played a video game, while a limited set of studies investigated flow in the working context. However, the literature suggests that playing games is a voluntary task that belongs to leisure or fun activities [142], and it is more probable for participants (or game users) to feel a flow state while playing compared to work activities that are not necessarily voluntary. Therefore, it is important to consider work-related activities and determine how flow is experienced at work.

From the machine learning point of view, only two studies so far have implemented deep feature learning approaches [50, 143]. However, they also have some fundamental limitations, such as not implementing the cross-validation strategy in a way that the

data of test participants could not be used for the training purpose and vice versa, and their presented results are mediocre ($\approx$ 70% accuracy), which could be improved. Moreover, studies have yet to explore — how to circumvent the data scarcity problem in the flow recognition domain and how to use emotional data to enhance flow recognition performance.

3.1.3 Research Motivation

The current challenges mentioned in Section 3.1.2 inspired the author to investigate the question of interruption-free and objective flow experience recognition in the working context. As mentioned in Section 3.1.1, measuring flow in an interruption-free manner using unobtrusive wearable sensors would allow researchers to understand the experience better — in this way, flow situations can be better anticipated and actively created in order to benefit from the positive consequences of the experience. However, there is currently no in-depth machine learning-based study in the related literature that aims to objectively recognize flow experience during working activities, and no public dataset related to this topic is available. This motivated the current study [59] to fill this gap by collecting multimodal sensor data and performing investigative experiments in a subject-independent manner.

Machine learning has gained much popularity in recent years due to the increasing availability of datasets and the impressive performances of its models in some application domains, sometimes beating human performance. Deep learning, notably, has become prevalent since deep feature learning approaches have been shown to outperform feature engineering methods in several application fields, particularly those related to image processing [144]. However, machine learning approaches, specifically deep learning ones — rely on large quantities of data to train robust models, which has caused slow progress in application fields relying on time-series data due to data scarcity problems. This is the case in particular for sensor-based flow experience recognition. To circumvent these issues, a subset of machine learning called *transfer learning* has received a lot of attention in the past decade. Transfer learning techniques attempt to extract knowledge from solving one task, referred to as the *source task*, and use it to improve the performance of a different but related task, referred to as the *target task* [145]. The idea behind such a transfer of information is that scarcity of data for the target task could potentially be mitigated by data available in large enough

quantities for the source task. For this reason, transfer learning has become standard in applications using images due to the existence of powerful models trained using very large-scale datasets such as *ImageNet* as the source dataset. But it remains significantly less explored for applications involving time-series data due to finding a suitable source dataset not being as obvious.

Moreover, flow literature revealed that researchers had associated some emotion states with flow [131, 146–149]. For instance, Csikszentmihalyi [131], the pioneer of flow theory, described flow as an optimal experience that can arise from a balance between the demands of an activity and one's own skills. If the demands are too high, a state of anxiety occurs instead; if, on the other hand, one's own skills exceed the demands, boredom arises [131]. This suggests moderate physiological arousal should promote flow, whereas low or excessive physiological arousal should impede it [147, 149]. This association between flow and moderate arousal has already been observed in an experiment exposing participants to stressors and applying physiological and questionnaire measurements. An underlying u-shaped relationship between physiological arousal and flow is assumed, in which flow occurs when arousal is neither too low nor too high [146]. The aforementioned associations that researchers have found so far, combined with the availability of relatively large benchmark emotion datasets, such as the *DEAP* dataset [150], motivated the author to conduct transfer learning experiments and try to leverage emotion data to improve the flow recognition performances that currently remain subpar, as mentioned in Section 3.1.2.

The outline of this chapter is as follows: the most relevant state-of-the-art literature to recognize human flow experience is presented in Section 3.2. Section 3.3 explains the author's contribution to recognizing wearable-based interruption-free human flow experiences and implementing a transfer learning approach to circumvent the data scarcity issue in this field. Section 3.4 describes the datasets utilized in the experiments. The data preprocessing steps for both datasets are presented in Section 3.5. Section 3.6 describes the author's implementation of the three tested feature extraction approaches. Section 3.7 presents the classification steps for the flow vs. non-flow state recognition. Section 3.8 describes the experiments and presents the results. Section 3.9 provides a scientific discussion. Finally, Section 3.10 concludes this chapter with a summary.

3.2 Related Work

Developing an unbiased system for predicting human flow experience when working with wearable multimodal sensor signals is difficult because flow experience is highly based on emotion states, which are not easy to decode [151]. In emotion analysis and flow detection, data are usually labeled by participants after each experiment, which may not be reliable and result in noisy labeling [152]. For subjective experiences such as flow, there is, in particular, no possibility to acquire accurate labels using unbiased external observers. Thus, there are few studies on this topic. Most rely on self-reports, focus on game users, use limited sensor modalities, or investigate a related but different problem from human flow experience recognition. Table 3.1 presents the dataset information, feature extraction methods, detection (i.e., focused classes), and sensors employed in recent past studies.

As mentioned above, there are few studies on this topic [50, 71, 143, 153–158], and all of them have some fundamental limitations (i.e., mediocre accuracy or not implementing the train and test strategy in a way that training data could not be used in the testing phase) and no public dataset available. In addition, studies still have to investigate how to circumvent the data scarcity issue in this domain. For example, Berta et al. [71] used a commercially available 4-channel EEG headset to access flow in game users and collect data from 22 participants while they played the video game. They manually computed 36 features and reported an accuracy of 50.1% with a user-independent classification approach and 66.4% with a user-dependent classification approach using an SVM classifier. However, they mentioned that new feature extraction and classification methods could improve their results.

Knierim et al. [153] collected self-reported and ECG data from an office worker over two weeks and evaluated it to investigate the possibility of predicting flow in the field through personalized models. The RF classifier with feature selection based on an *ANOVA* was employed to distinguish between low vs. high flow classes. In addition, experiments were also performed using *LASSO* regression with 10-fold cross-valuation to determine the flow intensities. They reported a mean absolute error (MAE) of 1.18, an R^2 of 0.17 for *LASSO* regression, and an F1 score of 0.65 for binary RF classification. However, for a good generalized model, feature learning techniques like MLP or CNN should be used to test the model using data from various subjects in a subject-independent way, where data from test subjects should not be utilized to train the model.

Table 3.1: Research publications found in the literature that attempted to assess human flow experience using wearable sensor data with machine learning approaches.

Study	Wearables / Sensors	Dataset Information	Features	Detection
[71]	EEG	22— subjects data when they were palying a video game	Frequency-based features	Boredom, flow and frustration
[153]	ECG	66—samples of a clerical worker over the course of 2 weeks	24—HRV related time and frequency domain features	Flow vs. non-Flow
[154]	ECG, Respiration and fNIRS	77— subjects data when they were playing a video game	Time and frequency domain features	Concentration, flow, arousal, affective valence and attentional effort
[155]	ECG-chest band	158/9— subjects data, during lab and field work respectively	Time and frequency domain cardiac features	Low level flow vs. High level flow
[50]	Empatica E4 wristband (BVP, EDA)	13—subjects data of 390 unique work activities, corresponding to 284 hours	99—Context, time and frequency domain features, feature learning	Low level flow vs. High level flow
[143]	Empatica E4 wristband (BVP, EDA, Tmp)	72—subjects data when they were playing a video game	Feature learning	Low flow vs. High flow
[156]	Kyto ear clip, Microsoft wristband, Tobii X120 eye tracker	12—subjects workplace activities data	Time and frequency domain features	Not-flow vs. Flow
[157]	BVP, EDA, ECG	175—subjects data related to lab work activities	Time and frequency domain features	Low level flow vs. High level Flow
[158]	Ear-EEG sensor	6—subjects data related to arithmetic, typing or puzzle solving activities	Frequency-based features	Low level flow vs. High level Flow

EEG: Electroencephalography; fNIRS: Functional near-infrared spectroscopy; ECG: Electrocardiography; BVP: Blood volume pulse; EDA: Electrodermal activity; Tmp: infrared Thermopile; HRV: Heart rate variability.

The study conducted by Harmat et al. [154] examined how psychological states change under different experimental conditions and what associations exist between self-reported flow and physiological measures. For the experiments, they collected data from 77 participants who played *Tetris* under three experimental conditions (i.e., easy, optimal, and difficult). The physiological data was recorded continuously during all experimental conditions using ECG, respiration, and functional near-infrared spectroscopy (fNIRS). *Statistica* 12, an analytics software package initially developed by *StatSoft* [159], was used to measure psychological state (flow, concentration, attentional performance, arousal, and valence) under all experimental conditions. The associations between self-reported psychological flow and physiological measures were examined using a series of repeated measures in linear mixed model analyses. The authors found that the subjective flow is positively related to the respiratory depth, and higher respiratory depth during high flow indicates a more relaxed state.

Rissler et al. [155] performed experiments to identify the flow experience in various laboratory (lab) and field work-related tasks, utilizing data recorded with an ECG chest band. Python's *HRV* package was used to compute heart rate variability linked features from the data collected during both activities (i.e., lab and field work). To distinguish between low-level and high-level flow with increased precision, the researchers employed a set of state-of-the-art classification algorithms, including C4.5, RF, AdaBoost, SVM, NB, and DT, training each separately. They reported an accuracy of 70% in field work and 68% for lab work in the binary low-level vs. high-level classification.

The study conducted by Di Lascio et al. [50] also performed a physiological flow classification for work activities by collecting data from 13 participants using the *Empatica E4* wristband [84]. The authors employed manual feature engineering and deep feature learning-based approaches to classify the physiological data into binary classes (i.e., low-level flow vs. high-level flow). The *SMOTE* algorithm was utilized for the class imbalance problem, and with a 5-fold cross-validation, they achieved an accuracy of 70.93%. The effect of contextual information related to the type of activity, time of day, and day of the week on perceived flow was also analyzed. It was found that the type of activity is relevant contextual information that should be considered when detecting flow during work activities.

Interestingly, Maier et al. [143] conducted experiments to determine the optimal user experience based on the physiological responses captured by the *Empatica E4*

wristband [84]. The features were computed manually and with deep feature learning from data collected from 72 participants playing the *Tetris* game with different difficulty levels. Experiments were conducted with 5-fold cross-validation for binary (low flow, high flow) and ternary (boredom, stress, and flow) classification. They reported the best accuracy of 67.50% for binary classification using a CNN-based feature learning approach and 49.23% accuracy for ternary classification using an RF classifier.

Lee et al. [156] claimed that using EEG and ECG sensors and chest straps is either too uncomfortable or not socially acceptable. Moreover, they even said that using video cameras is also too invasive for the workplace. Therefore, they performed experiments using a *Kyto* ear clip sensor, a *Microsoft Band* 2 wristband, and a *Tobii X12* eye tracker. These devices provided the HR, HRV, Tmp, and diameters of the right and left pupils of the participants. The data was collected from 12 subjects while performing their work activities. After the data preprocessing and time and frequency domain feature extraction, they trained SVM, DT, and RF classifiers to classify flow and not-flow states in a *LOSO* cross-validation manner. They claimed an Area Under the Receiver Operator Curve (AUC) of 0.889 and a standard deviation (SD) of 0.087. However, their work is not peer-reviewed.

Similarly, Rissler et al. [157] performed a laboratory experiment with 175 participants using BVP, EDA, and ECG wearable sensor data in another study. They created a simulated environment where participants had to perform an invoice-matching task derived from a real-world *Enterprise Resource Planning System* procedure. They computed different time and frequency-based features, including the cardiac features (HRV) on frequency bands such as low frequency (LF) and high frequency (HF). The data was then randomly split between training and testing sets with an 80/20 percent ratio, and SVM and RF classifiers were trained to classify between high and low flow states. The authors claimed that cardiac features (HRV-LF, HRV-HF, and LF/HF ratio) played an important role in flow intensity classification. They reported their best accuracy of 72.30% with the RF classifier. However, they should have mentioned whether the train- ing and testing data split was subject-dependent or independent.

Interestingly, Bartholomeyczik et al. [158] conducted experiments in the context of work activities to improve the automatic detection of flow in work-related situations — by observing flow across three controlled tasks: arithmetic, typing, and puzzle solving. An ear EEG sensor collected flow experience-related measurements from six knowledge

workers (i.e., PhD students). Afterward, they computed *Welch's* [160] power spectral density-based features using a one-second window with 50% overlap for each frequency band. The experiments were performed to compare EEG power across tasks and the difficulty of conditions, in addition to finding the correlation between self-reported data and power-bands. However, the sample size used in this study is too small, and their analysis needs to include current state-of-the-art machine-learning approaches.

In the work presented in this chapter (study) [59], the author hypothesized that modern multimodal wearable sensors enable us to distinguish between flow and non-flow states during work activities and do so with higher accuracy than reported in the literature. Therefore, this study not only compared different state-of-the-art approaches to feature extraction and classification but also investigated how to transfer emotional information to flow recognition, as previous work [146, 147] has shown a connection between these two aspects.

3.3 Own Contribution

This section presents the authors' main contributions to this chapter's scientific work. It is worth mentioning that parts of the content of this work have already been published [59] as a journal article in an internationally well-recognized Q1 journal. The author's main contributions are summarized below:

1. Investigate the use of multimodal wearable sensors in the context of human flow experience recognition and develop a state-of-the-art machine learning model that learns flow and non-flow patterns from physiological responses and classifies them into their respective classes.

2. Analyze and compare multimodal wearable device data to determine the most meaningful one for flow recognition.

3. Objectively investigate feature extraction approaches and machine learning algorithms for flow recognition by performing a comparative analysis between them in a subject-independent manner.

4. Propose a deep transfer learning model for human flow state recognition using emotion classification as a source task and investigate which emotion-based source

task helps to achieve enhanced flow recognition results and helps to circumvent the data scarcity problem in the flow recognition domain.

3.4 Datasets Description

Two datasets are used in this scientific research, namely the *PhySF* and *DEAP* [150]. The PhySF dataset was collected at the *Assessment of Physical and Psychological Signals Laboratory* (APPS Lab) at the University of Lübeck [161] with the help of psychologists from the *Department of Psychology at the University of Lübeck*. All primary experiments were conducted using the PhySF dataset. To get around the problem of not having enough data in the flow recognition field — the DEAP dataset was used as the source dataset, and PhySF as the target dataset in the experiments based on the transfer learning approach. Sections 3.4.1 and 3.4.2 describe each dataset in detail, respectively.

3.4.1 Emotion Recognition Dataset: DEAP

DEAP is a dataset for emotion analysis in which thirty-two (32) participants (16 males and 16 females aged between 19 and 37 years, with an average age of 26.9 years) participated in the data collection process. Each participant was asked to watch 40 video clips. Each video clip contained a one-minute music extract to elicit various emotional states. Physiological signals comprising EEG channels (n=32) and peripheral channels (n=15) were recorded while the participants watched the videos. In addition, before the actual recording, baseline signals were recorded for two minutes from each subject while they relaxed and looked at a fixation cross on a screen. After watching them, each participant had to rate each video in terms of arousal, valence, dominance, and liking. The level of arousal, valence, dominance, and liking was assessed using the *Manikin Self-Assessment Scale* (SAM) [162], yielding numerical values between 1 (very low) and 9 (very high) for each emotional dimension.

This study used a preprocessed version of the data sampled at 128 Hz, made available online by the authors of the DEAP dataset[1]. The dataset is available upon request (signed EULA) for educational purposes. For more information about the dataset, the sensor devices, and the data collection process, see [150, 163].

[1] `https://www.eecs.qmul.ac.uk/mmv/datasets/deap/` (accessed: 05.05.2023).

3.4.2 Flow Recognition Dataset: PhySF

Twenty-six (26) healthy individuals (8 males and 18 females aged between 18 – 40 years) participated in the PhySF dataset[2] collection, whose demographic information is presented in Table 3.2.

Table 3.2: Demographic information of all the subjects (S1–S25) who took part in the data collection process of the PhySF dataset.

Subject Name	Gender	Age (in years)	Subject Name	Gender	Age (in years)
S1	male	35	S2	female	25
S3	female	23	S4	male	37
S5	female	27	S6	female	24
S7	male	25	S8	male	27
S9	male	28	S10	male	24
S11	female	27	S12	female	20
S13	female	19	S14	female	18
S15	male	28	S16	female	24
S17	female	40	S18	female	20
S19	female	21	S20	female	22
S21	female	24	S22	female	23
S23	male	19	S24	female	19
S25	female	19	-	-	-

The recruited participants consisted of students and employees of the University of Lübeck. They were required to be at least 18 years old, speak German fluently, and not have any cardiovascular disease, non-correctable vision, or movement impairment. The measurements were conducted under regional sanitary regulations during the COVID-19 pandemic between June and December 2022. The data of one volunteer was excluded from the dataset due to missing values. The students of the psychology department received 1.5-course credits for their participation. The study was conducted following the *Declaration of Helsinki* and approved by the *Institutional Review Board of the University of Lübeck (April 14, 2022; No. 22-112)*.

Subjects were asked to read and sign the informed consent form before the experiment. The experiment began with a 5-minute baseline recording, during which participants were shown a fixation cross sign (they were advised to relax during this time). They were then instructed to perform mathematical and reading tasks using a web application [164]. The mathematical task consisted of 150 simple arithmetic questions, for example, $7765+$

[2] `https://osf.io/hgj6p/?view_only=1a23513a3ff24eae9afb47bcaba9f4f2` (accessed: 03.08.2023).

$19 = \_\_\_$; $5583 - 7 = \_\_\_$, etc. The reading task consisted of a short story of about 3000 words ("Die verborgene Seite der Medaille" [The hidden side of the coin] by Scavezzon [165]). Participants kept wearing the sensor devices during the whole data collection. After each task, the participants indicated if and at what point of the task they experienced flow.

In this study, three (03) wearable devices, namely the *Emotiv Epoc X* - an EEG headset, the *Empatica E4* wristband, and the *Biosignalplux RespiBAN* wearable, are utilized to acquire the physiological measures related to the flow and non-flow states of the study participants. All of these devices are shown in Figure 3.1.

Figure 3.1: List of wearable devices utilized to acquire the physiological measurement of the participants: (a) - Headband style Emotiv Epoc X with 14 electrodes on flexible plastic armbands, (b) - Empatica E4 wristband, and (c) - Biosignalplux RespiBAN wearable device.

The *Emotiv Epoc X* [166], as shown in Figure 3.1(a) is a 14-channel EEG headset for scalable and context-aware exploration of human brain activity. It is placed on the scalp, registering the bioelectrical activity of specific brain regions. The electrodes were hydrated using a saline solution. The EEG data were acquired during the experiments at a sampling rate of 128 Hz.

The description of the *Empatica E4* wearable wristband, i.e., Figure 3.1(b), and the *Biosignalplux RespiBAN*, i.e., Figure 3.1(c), devices and their sensor channels such as EDA, BVP, ECG, EMG and EOG are provided in Section 2.4. Moreover, in contrast to Study 1, detailed in Chapter 2, the EDA or *Electrodermal activity* electrodes mentioned in Figure 2.1 were substituted with *EOG* electrodes [167]. This is because the EDA sensor is available in the Empatica E4 wristband. The description of the utilized EOG sensor (electrodes) is as follows:

The *Biosignalsplux EOG* electrodes are designed for seamless data acquisition of the subject's gaze patterns. They record electrical potentials in a specific facial region relative to a reference electrode. During the data recordings, the EOG electrodes were adhered on the right and left sides of the outer canthi, respectively. In addition, a reference electrode was also adhered to the back of the left ear.

After performing each task (such as arithmetic and reading), while the wearable devices mentioned above recorded participants' physiological data (as shown in Figure 3.2). The subjects were asked to put themselves back to the task they had just completed and label their physiological measures by answering simple questions using a web interface (i.e., Unipark [164]). The questions that were asked to the participants are listed in Table 3.3.

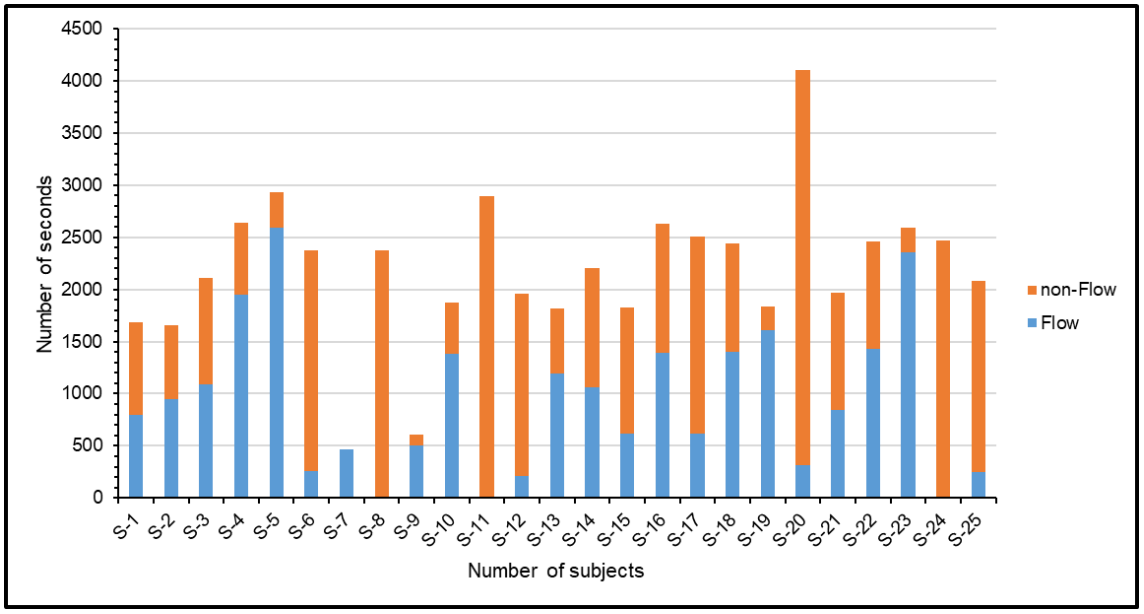

Figure 3.2: The distribution of each subject's flow and non-flow class data in the PhySF dataset. *Y-axis:* represents the number of seconds of data for each class label. *X-axis:* represents the number of subjects in the PhySF dataset.

Table 3.3: List of questions that were asked to the study participants after each task in order to determine the labels of the wearable's physiological measurements.

No.	Questions	Options
1	Were you in the "Flow" during the task?	1. Yes 2. No
2	When were you in the "Flow" during the task?	1. First third of the task 2. The second third of the task 3. Last third of the task

For the first question (see Table 3.3), only a single selection or choice of 'Yes' or 'No' was possible. Which was asked to check if the participants experienced flow or not. If they answered 'No', then the data was labeled as a non-flow state. If they answered 'Yes' then the second question was to check at which part of the task they experienced flow state (i.e., first third of the task, second third of the task, or last third of the task). For the second question, multiple selections were possible. In addition, the second question was only asked if the participants answered 'Yes' for the first question. The proportion of self-reported flow and non-flow state physiological data for each subject is shown in Figure 3.2.

3.5 Data Preprocessing

Machine learning algorithms' performances generally depend on the data quality used to train them. In cases of insufficient, unnecessary, and irrelevant data, they may provide inaccurate or less understandable results. Therefore, data preprocessing is an essential step in the machine learning pipeline to obtain highly precise results. The steps performed to preprocess the *DEAP* and *PhySF* datasets to recognize human flow experiences accurately are described in Sections 3.5.1 and 3.5.2.

3.5.1 DEAP

The version of the DEAP dataset [163] down-sampled to 128 Hz was used for this study. As a first step, similar sensor channel data (also present in the PhySF dataset) were carefully separated for use in the transfer learning-based experiments. This is because the shape of the input data must be the same for both tasks (i.e., source and target) to successfully transfer the learned weights from the source domain to the target domain. The selected sensor channels of both datasets are shown in Table 3.4. It is worth noting that experiments were also performed with randomly selected sensor channels from both datasets, but the results were below the baseline. Subsequently, the separated data were segmented using a *Sliding Window Segmentation* (SWS) technique. The current method for selecting the optimal window size is empirical [96]. People usually test different window sizes and choose the one that maximizes the recognition system's performance [97]. Therefore, window sizes of 10, 30, and 60 seconds (s) were

tested with a stride (i.e., step size) of 50% of the window sizes. The best settings were found with a window length T and a step size ΔS of 10s and 5s, respectively.

Table 3.4: Selected sensor channels in both datasets for the proposed transfer learning-based approach.

DEAP dataset		**PhySF dataset**	
Channel No.	**Channel Name**	**Channel No.**	**Channel Name**
2	AF3	1	AF3
3	F7	2	F7
4	F3	3	F3
6	FC5	4	FC5
7	T7	5	T7
11	P7	6	P7
15	O1	7	O1
17	O2	8	O2
20	P8	9	P8
24	T8	10	T8
25	FC6	11	FC6
27	F4	12	F4
28	F8	13	F8
29	AF4	14	AF4
41	GSR-1	16	EDA
45	Resp	20	Resp
47	Temp	17	Tmp

Note: Galvanic skin response (GSR) and Electrodermal activity (EDA) both represent a similar sensor; Temp and Tmp, both represent skin temperature recorded with infrared Thermopile sensor.

3.5.2 PhySF

The PhySF dataset comprises n=23 sensor channels (i.e., 14 Emotiv Epoc X, 5 Empatica E4, and 4 RespiBAN channels). Figure 3.1 presents the wearable devices used for this study. The details of the Emotiv Epoc X wearable are presented in Section 3.4.2, and the details of Empatica E4 and RespiBAN wearables are presented in Section 2.4. The data from each device had a different sampling rate. In the first step, data from all sensor channels were synchronized at a target frequency of 128 Hz using linear interpolation to infer the missing data values at the target timestamps. In the second step, the resampled data were segmented using the SWS technique. Values for the segmented window length T were tested for $T \in \{10s, 30s, 60s\}$, with a segmentation step size ΔS always set to 50% of T. The best performances were obtained when T and

ΔS were set to 10s and 5s, respectively. The results with these parameters are reported in Section 3.8.

3.6 Feature Extraction

As mentioned earlier, there are very few studies on this topic, and no public dataset is available. Therefore, a dataset, namely *PhySF*, is acquired and made online available. To provide future research direction for this field, a comparative analysis of feature extraction approaches is conducted. The following are the three feature extraction approaches that are tested to recognize flow vs. non-flow states:

- One feature engineering baseline that follows the traditional approach of manually computing simple statistical and frequency-related features from the input time-series data.
- One feature learning baseline that follows training a DNN model end-to-end to learn features from raw input time-series data.
- The approach proposed in this chapter follows transfer learning to enhance feature learning by transferring knowledge from emotion-related time-series data.

Each of the three feature extraction approaches mentioned above, for instance, feature engineering, feature learning, and the author's proposed one, i.e., feature learning enhanced by transfer learning, is described in more detail in the following subsections.

3.6.1 Feature Engineering

As mentioned earlier, feature engineering is selecting, manipulating, and transforming raw data into meaningful features that can be used in a machine learning-based prediction model. In this work, $F = 18$ features were manually crafted on the PhySF dataset, also called HCFs. These HCFs are simple statistical and frequency values on the physiological input signals or their power spectrum. All HCFs used in this study are listed in Table 2.3. They were computed independently on each sensor channel for each data segment as suggested in [22, 99] and then concatenated to form one (1) D feature vector of size $F \times S = 18 \times 23 = 414$, representing a data segment.

3.6.2 Feature Learning

Following the feature engineering, experiments were accomplished to investigate the use of deep learning approaches. Specifically, the MLP and CNN models with a softmax classification layer were tested as feature learners. Since the automatic optimization of DNN hyper-parameters is still a research topic under investigation so far [168, 169], the trial-and-error method was used to find the best values for these parameters. Tables 3.5 and 3.6, respectively, reveal these parameter values.

Table 3.5: MLP architecture and hyper-parameter values. The model was trained with the ADAM optimizer [116] using an initial learning rate of 5×10^{-4}.

Layer Name	Neurons / Dropout Rate	Activation
Batch Norm [115]	-	-
Dense	64	LeakyReLU (alpha=0.2)
Dropout	0.50	-
Dense	32	LeakyReLU (alpha=0.2)
Dropout	0.50	-
Dense	16	LeakyReLU (alpha=0.2)
Dropout	0.50	-
Flatten	-	-
Dense	16	LeakyReLU (alpha=0.2)
Dense	2	Softmax

Batch Norm — also known as batch normalization, is a method used to make the training of ANNs faster and more stable by normalizing the layers' inputs. Notably, it was applied before the activation function [115].

Table 3.6: CNN architecture and hyper-parameter values. The model was trained with the ADAM optimizer [116] using an initial learning rate of 5×10^{-4} and a fixed dropout rate of 0.50.

Layer Name	No. Kernels (Units)	Kernel / Pool Size	Stride	Activation
Batch Norm [115]	-	-	-	-
Convolutional	64	(2,1)	(1,1)	LeakyReLU (alpha=0.2)
MaxPooling	-	(3,1)	-	-
Dropout	-	-	-	-
Convolutional	32	(2,1)	(1,1)	LeakyReLU (alpha=0.2)
MaxPooling	-	(3,1)	-	-
Dropout	-	-	-	-
Convolutional	16	(2,1)	(1,1)	LeakyReLU (alpha=0.2)
MaxPooling	-	(3,1)	-	-
Dropout	-	-	-	-
Flatten	-	-	-	-
Dense	16	-	-	LeakyReLU (alpha=0.2)
Dense	2	-	-	Softmax

3.6.3 Transfer Learning

To enhance the performances obtained by feature learning approaches, a method that consists of transferring DNN weights learned while training for a problem related to emotion recognition was tested. The DEAP dataset was selected as the source dataset because of its importance in emotion recognition research and since its data were acquired using similar sensor modalities as the PhySF dataset. More specifically, regarding the second point, the DEAP dataset consists of 48 channels (32 EEG and 16 others) that include 17 common channels with PhySF (i.e., AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4, GSR, Tmp, and Resp). The subsets of the DEAP and PhySF datasets containing only the aforementioned sensor channels are used as source and target datasets to maximize the chances of the transfer being successful. Figure 3.3 illustrates the principle of the author's implemented transfer learning approach.

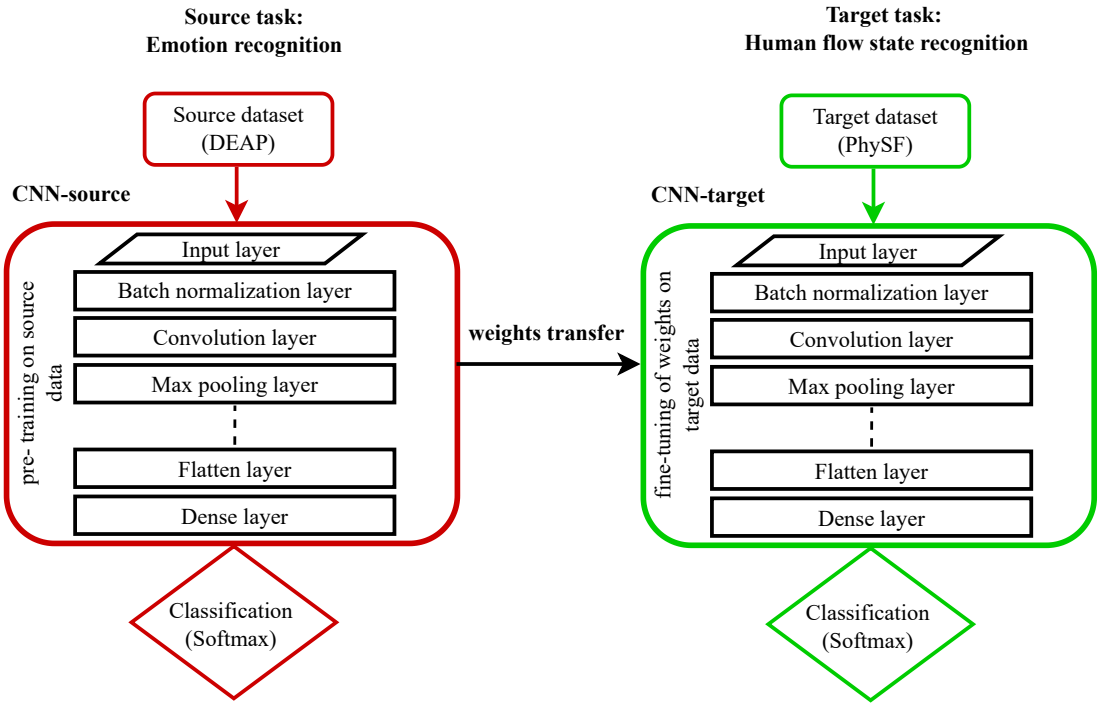

Figure 3.3: The principle of the author's implemented CNN-based transfer learning approach. In the first step, the CNN model is trained on the source dataset (i.e., DEAP) to solve the source task (emotion classification). In the second step, the weights learned in the first step are utilized to initialize the weights of the target model (CNN), which is then fine-tuned on the target dataset (i.e., PhySF) to recognize the human flow state.

The DEAP dataset contains emotion-related annotations, including arousal, valence, boredom, and liking ratings. However, following the standard practice of the literature

[170–172], arousal and valence recognition are considered two separate binary classification problems between low and high arousal and valence for the source task. In the source task classification problem, the data was split between training and testing, following a ratio of 70/30%. The DEAP dataset numerical ratings provided by the subjects were used to split the data into two classes using a value of 5 as the cutoff between the two classes (≤ 5 for low, > 5 for high). Each classification problem, such as arousal (high vs. low) or valence (high vs. low), was used once separately as a source task to check which emotional dimension could lead to learning better flow detection features. A third classification problem with four classes (i.e., quadrants) involving both arousal and valence simultaneously (i.e., low arousal/low valence, low arousal/high valence, high arousal/high valence, and high arousal/low valence) was also tested as a source task.

To perform the knowledge transfer in order to enhance the feature learning — the CNN model described in Section 3.6.2 was selected as it revealed higher performance in classifying flow and non-flow than the MLP baseline. The CNN model was appended to a softmax classification layer to obtain class estimations for the source task; it is specified as CNN-source, as shown in Figure 3.3. In the first step, the subset of the DEAP dataset that has 17 sensor channels in common with the PhySF dataset was used as the source dataset. The CNN-source was trained end-to-end, and the weights and biases it learned were saved. Subsequently, the subset of the PhySF dataset that has 17 sensor channels in common with the DEAP dataset was used as the target dataset. The target task was to classify flow and non-flow into one of the two categories. The weights and biases of the CNN-source were transferred to a CNN model trained to solve the target problem called CNN-target. The CNN-source and CNN-target share the same architecture, except for the softmax classification layer, which may vary in terms of the number of neurons that match the number of classes of the source or target task considered. The CNN-target was then fine-tuned in a supervised manner on the PhySF dataset. It is worth noting that in this experiment, no layer was frozen from the input layer to the flatten layer.

3.7 Classification

To obtain the best classification performance possible, experiments are performed with four different state-of-the-art classifiers. These include AdaBoost, RF, SVM, and

XGBoost [22, 173, 174]. All classifiers were trained with the features extracted using the methods presented in Section 3.6.1. In the case of feature learning approaches, for instance, deep feature learning and transfer learning, the DNNs were trained following the standard procedure using a softmax classification layer. All the classification results with each implemented approach are presented in Section 3.8.

As illustrated in Figure 3.2, the PhySF dataset is highly imbalanced since some subjects did not report any occurrence of one of the two classes. For example, the flow class is missing for subjects 8, 11, and 24, and the non-flow class is missing for subject 7. To consider this imbalance while obtaining features with a generalization capacity as large as possible — all classifiers were trained with a *Stratified-K-Fold - Cross Validation* (SKF-CV) for the target task. In these experiments, the subjects of the PhySF dataset were evenly split across $K = 5$ folds while ensuring that the number of subjects who predominantly reported being in flow and those not being in flow was mostly balanced within each fold. Since the SWS parameters, such as sliding window length T and step size ΔS, have shown a high impact on classification performances using time-series data [97]. Therefore, to find the best values for these parameters, AdaBoost and SVM classifiers with a window length T of 10, 30, and 60 seconds were tested with a step size ΔS always set to 50% of T, as shown in Figure 3.4.

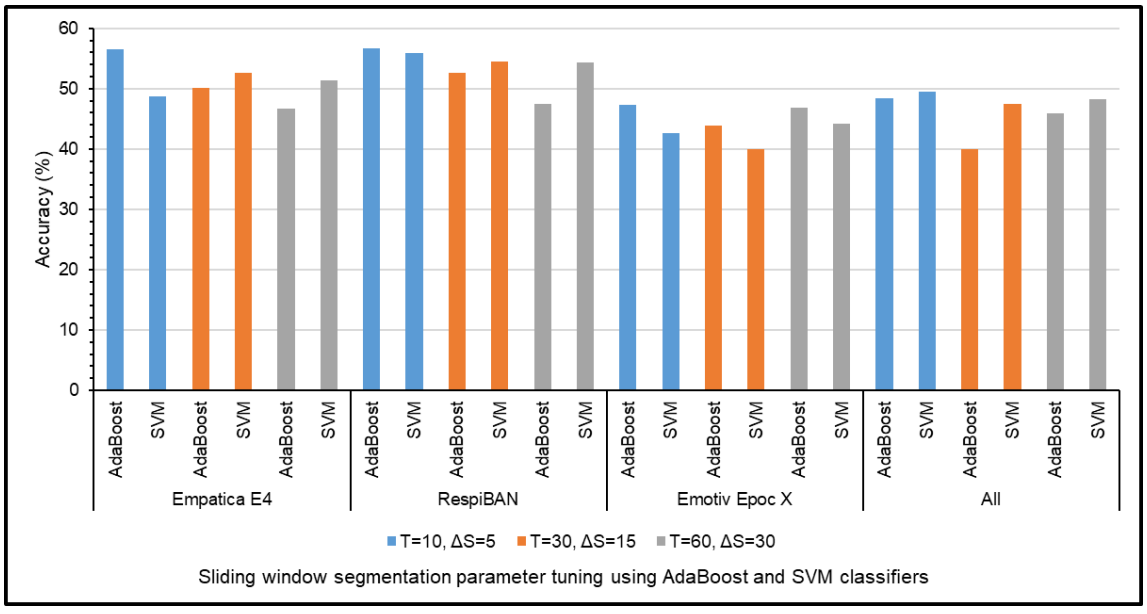

Figure 3.4: Results of the experiments performed to get the optimized values of *SWS* parameters such as window length T and step size ΔS. The AdaBoost and SVM classifiers are trained using HCFs with $T \in \{10, 30, 60\}$ seconds and $\Delta S \in \{5, 15, 30\}$ seconds data to find the best T and ΔS values.

The distribution of subjects in each fold for SKF-CV and the flow and non-flow state data of each subject in seconds are presented in Table 3.7. The SKF-CV ensures that the following two conditions are verified during the training of the classifiers: both training and testing sets are balanced between the two classes, and the classifiers are trained in a subject-independent manner (i.e., evaluated on subjects who were not seen during the training process).

Table 3.7: Distribution of 25 subjects to 5 different folds (K) for SKF-CV and the measurement of flow and non-flow data (in seconds) of each subject of the PhySF dataset.

Subject No.	Flow and non-Flow data in seconds		Splitting of subjects into various folds				
	Flow	**non-Flow**	**K1**	**K2**	**K3**	**K4**	**K5**
S1	792	896	X	-	-	-	-
S2	952	709	-	X	-	-	-
S3	1086	1027	-	-	X	-	-
S4	1951	687	-	-	-	X	-
S5	2594	340	-	-	-	-	X
S6	254	2119	X	-	-	-	-
S7	467	0	-	X	-	-	-
S8	0	2374	-	-	X	-	-
S9	502	106	-	-	-	X	-
S10	1382	490	-	-	-	-	X
S11	0	2893	X	-	-	-	-
S12	210	1745	-	X	-	-	-
S13	1189	627	-	-	X	-	-
S14	1058	1151	-	-	-	X	-
S15	619	1210	-	-	-	-	X
S16	1394	1235	X	-	-	-	-
S17	615	1889	-	X	-	-	-
S18	1397	1040	-	-	X	-	-
S19	1609	227	-	-	-	X	-
S20	316	3792	-	-	-	-	X
S21	846	1119	X	-	-	-	-
S22	1429	1031	-	X	-	-	-
S23	2359	237	-	-	X	-	-
S24	0	2470	-	-	-	X	-
S25	246	1833	-	-	-	-	X

K (1–5): Represents the fold index; S1–S25: Represents 25 subjects who participated in the collection of the PhySF dataset. Values in the parenthesis represent the proportion of flow and non-flow class data in seconds.

The SWS experiments (illustrated in Figure 3.4) revealed that the AdaBoost classifier works better with small values of T and ΔS, such as 10 and 5 seconds, respectively than with $T \in \{30, 60\}$, and $\Delta S \in \{15, 30\}$ for both: the data from each single device

and data from all devices together. However, the SVM classifier did not provide any significant information on the best parameter values. Therefore, all the flow experience recognition experiments were performed with $T = 10$ and $\Delta S = 5$ with the three feature extraction approaches mentioned in Section 3.6. The results of these experiments are presented in Section 3.8.

3.8 Experiments and Results

In this study, all models were implemented using Python 3.9.13. More specifically, for the machine learning classifiers (i.e., AdaBoost, RF, SVM, and XGBoost) and deep learning models (i.e., MLP and CNN), the libraries scikit-learn, XGBoost, and Keras with a Tensorflow 2.10.0 backend was used. An *Adaptive Moment Estimation* (ADAM) with an initial learning rate of 5×10^{-4} was chosen as the optimizer for models based on DNN and trained with 50 epochs at a batch size of 16. The categorical cross entropy was used as the loss function for the DNN models. The results of the three implemented approaches, feature engineering, feature learning, and transfer learning, are respectively shown in Sections 3.8.1,3.8.2, and 3.8.3.

3.8.1 Feature Engineering

The 18 HCFs listed in Table 2.3 were computed independently and used to train four classifiers (i.e., AdaBoost, RF, SVM, and XGBoost). They were computed for each data segment of each sensor channel — which leads to a total of 414 features (for instance, Empatica E4: 18 × 5, RespiBAN: 18 × 4, and Emotiv Epoc: 18 × 14).

To determine the best-performing wearable devices and classifier for the flow experience recognition problem — all classifiers such as AdaBoost, RF, SVM, and XGBoost, were trained once separately on the computed HCFs of each wearable device data and once with all combined data. The performances of the implemented classifiers using the above criteria are shown in Table 3.8.

Table 3.8: Flow and non-flow state recognition results with four different classifiers using feature engineering (HCFs) for each wearable device data and the combination of all device's data with a window size T of 10s, a step size ΔS of 5s, and a sampling frequency of 128 Hz.

Classifier	Wearable Device	Accuracy	AF1 Score	Sensitivity	Specificity
AdaBoost	Emotiv	47.29	47.21	53.35	43.12
RF		45.07	40.68	21.92	60.99
SVM		42.66	38.45	20.27	58.05
XGBoost		49.74	47.53	35.84	59.30
AdaBoost	Empatica	56.59	**55.28**	48.51	62.14
RF		47.60	47.53	**62.87**	37.13
SVM		48.75	48.72	62.72	39.17
XGBoost		46.96	46.86	52.46	43.19
AdaBoost	RespiBAN	**56.70**	53.39	36.94	70.25
RF		55.76	45.81	15.86	83.13
SVM		55.92	38.86	03.80	**91.67**
XGBoost		51.20	45.89	24.45	69.54
AdaBoost	All	48.42	47.30	41.59	53.11
RF		48.78	46.29	33.51	59.25
SVM		49.58	47.27	35.20	59.44
XGBoost		42.54	42.20	42.88	42.31

AF1 Score: macro Averaged F1 score; Empatica: Empatica E4 wristband; RespiBAN: Biosignalsplux wearable device (including Resp, ECG, EOG, and EMG sensors); Emotiv: Emotiv Epoc X - a 14 channel EEG headset.

As shown in Table 3.8, the results of the feature engineering baseline were mediocre for each device and all device data using various classifiers. This formed the basis for testing the feature learning approach since feature learning approaches have performed better than feature engineering approaches in the past in various application domains [99].

3.8.2 Feature Learning

Similarly to the feature engineering baselines, experiments with a feature learning approach were performed to compare the performances obtained with each device separately and by using all wearable data. The MLP and CNN baseline results are presented in Tables 3.9 and 3.10, respectively.

The results presented in Tables 3.9 and 3.10 reveal that the CNN-based feature learning approach is more effective than MLP-based feature learning and HCFs-based feature engineering. In addition, when data from Emotiv, Empatica, and RespiBAN are used

Table 3.9: Flow and non-flow state recognition results using the feature learning approach with MLP for each device data individually and for the data of all combined devices.

Wearable Devices	Accuracy	AF1 Score	Sensitivity	Specificity
Emotiv	64.39	63.83	60.87	67.01
Empatica	54.49	52.80	41.74	63.97
RespiBAN	57.09	56.50	53.35	59.86
All	**72.72**	**71.68**	**62.86**	**80.07**

MLP: Multi-Layer Perceptron (a fully connected feed-forward Artificial Neural Network); AF1 Score: Macro-Averaged F1 Score.

Table 3.10: Flow and non-flow state recognition results using the feature learning approach with CNN for each device data individually and for the data of all combined devices.

Wearable Devices	Accuracy	AF1 Score	Sensitivity	Specificity
Emotiv	64.97	64.95	78.29	55.07
Empatica	60.08	57.51	41.82	73.67
RespiBAN	59.60	57.84	46.05	69.62
All	**73.63**	**72.70**	**64.02**	**80.78**

CNN: Convolutional Neural Network; AF1 Score: Macro-Averaged F1 Score.

together, the CNN-based approach acquires an accuracy of 73.63%, which is notable. However, when data from all devices is used separately, the Emotiv wearable is more effective than Empatica and RespiBAN wearables.

3.8.3 Transfer Learning

Since no public dataset related to human flow experience is available, DNNs have performed well on large datasets. Therefore, experiments were performed with a deep transfer learning approach to circumvent the data scarcity issue in this field by using the *DEAP* emotion recognition dataset as a source dataset and PhySF as a target dataset. This is because researchers have previously found a correlation between human emotions (such as arousal and valence) and the human flow experience [146, 147]. In addition, the DEAP dataset is available for academic research purposes and can be obtained with a reasonable request. Furthermore, both of the mentioned datasets have used similar sensor channels (i.e., at least 17 sensor channels are common in PhySF and DEAP). This research hypothesizes that transferring the learned weights of the emotion source dataset could improve flow recognition performance on the PhySF dataset. Table 3.11 illustrates the results of the implemented transfer learning approach.

Table 3.11: Flow and non-flow state recognition results using transfer learning approach with CNN. Two datasets were used, DEAP(17) and PhySF(17), which are the subsets of 17 common sensor channel data, respectively taken from the DEAP and PhySF datasets. The tested source tasks consist of the classification of arousal (high arousal vs. low arousal), valence (high valence vs. low valence), or quadrants (4-class problem with low arousal/low valence, low arousal/high valence, high arousal/high valence, and high arousal/low valence) on the DEAP(17) dataset. The target task is the classification of flow vs. non-flow states on the PhySF(17) dataset.

Source Task	**Target Task**			
	Flow vs. non-flow classification on PhySF(17)			
	Accuracy	**AF1 Score**	**Sensitivity**	**Specificity**
None	70.09	70.01	89.36	56.13
Arousal classification on DEAP(17)	**75.10**	**74.92**	**79.33**	**72.04**
Valence classification on DEAP(17)	70.30	70.24	88.80	56.88
Quadrants classification on DEAP(17)	71.53	71.47	90.34	57.89

PhySF(17): Physiological Sense Flow dataset of 17 common sensor channels (14-EEG + EDA + Tmp + Resp); DEAP(17): DEAP dataset of 17 common sensor channels (14-EEG + GSR-1 + Temp + Resp); Temp and/or Tmp: infrared Thermopile; EDA: Electrodermal activity; GSR: Galvanic skin response; Resp: Respiratory rate.

Improved results, an accuracy of 75.10%, and an AF1 score of 74.92% were obtained with the implemented transfer learning-based approach when the DEAP dataset was classified into high vs. low arousal, as shown in Table 3.11. These results are about 5% higher than those without transfer learning using 17 sensor channels and also satisfy this research hypothesis and confirm that better results can be obtained with the transfer learning approach when a related dataset is used as a source. Furthermore, the results also reveal that arousal is more critical than valence for flow human the flow experience.

3.8.4 Comparison with the Literature

In the highlight of literature, there is no publicly available dataset for human flow state detection, and most of the literature in this field is based on game user or self-reported data and utilizes only the traditional feature engineering approaches such as time and frequency-based features extraction. As far as is known, only two studies implemented the feature learning approach in this domain [50, 143]. The results of these studies are shown in Table 3.12.

Table 3.12: Comparison of flow and non-flow state recognition results obtained by employing the implemented approaches of this work, such as feature engineering, feature learning, and transfer learning-based feature learning, with state-of-the-art results of literature studies.

Approach	Accuracy	AF1 Score	Sensitivity	Specificity
Maier et al. [143]	67.50	-	-	-
Di Lascio et al. [50]	70.93	-	-	-
HCFs	56.70	53.39	36.94	70.25
CNN	73.63	72.70	64.02	80.87
Deep TL	**75.10**	**74.92**	**79.33**	**72.04**

HCFs: Represents the hand-crafted features (HCFs) based feature engineering. The CNN is the proposed feature learning approach using a Convolutional Neural Network (CNN). Deep TL is the proposed transfer learning approach with CNN using DEAP as a source dataset for high vs. low arousal classification tasks. Notably, promising results were achieved in this study using the proposed Deep TL approach.

Compared to the literature[3] results revealed in Table 3.12 — this study achieved an improved accuracy of 73.63% and an AF1 score of 72.70% with the feature learning approach using CNN, an accuracy of 75.10%, and an AF1 score of 74.92% with the transfer learning approach using DEAP as a source dataset. The results of each of the implemented approaches, such as feature engineering, feature learning, and transfer learning in the form of a confusion matrix, are shown in Figure 3.5 (a–c). Moreover, the implemented models of this study outperformed the results from the literature when trained in a subject-independent cross-validation manner, while the work of Maier et al. [143] and Di Lascio et al. [50] was based on a subject-dependent strategy. It is also worth noting that this is the first scientific work introducing transfer learning-based human flow recognition.

[3] Note: A direct comparison between the results of this study and the results presented in the literature is not possible because of the differences in setups, datasets, etc., so this comparison is just provided for informative purposes to show the effectiveness (potential) of the results of this study.

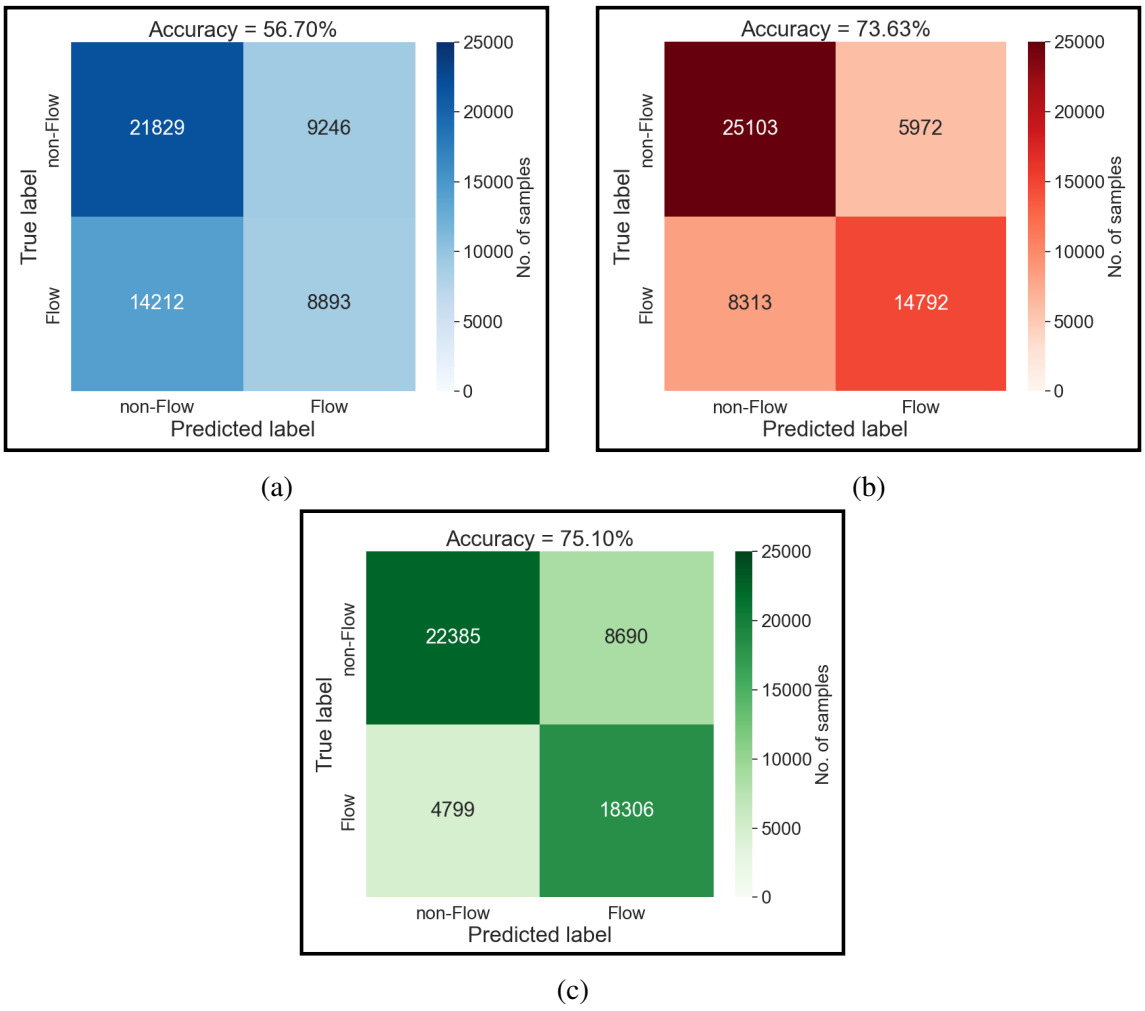

Figure 3.5: List of the confusion matrices (a–c) for flow and non-flow recognition using three feature extraction approaches: (a) - feature engineering based on computed hand-crafted features (HCFs), (b) - feature learning using CNN, and (c) - transfer learning. In each figure (i.e., a–c), the top-left quadrant represents a true negative (t_n) value, the top-right quadrant represents a false positive (f_p) value, the bottom-left quadrant represents a false negative (f_n) value, and the bottom-right quadrant represents a true positive (t_p) value, respectively. All values (i.e., t_p, t_n, f_n, f_p) of the confusion matrices (a–c) are cumulative over five folds and the five runs of an experiment.

3.9 Scientific Discussion

The following points provide a detailed discussion of the aforementioned results:

- **Physiological signals for flow detection:**

  There are few studies on this topic in the literature, and most of them employed a limited number of sensor channels compared to this study [50, 71, 143, 158,

175]. For example, Berta et al. [71] and Bartholomeyczik et al. [158] only analyzed EEG signals, Passalacqua et al. [175] performed experiments with only EDA data, Di Lascio et al. [50] and Maier et al. [143] considered only the investigation with Empatica E4 data for their experiments. Furthermore, they also reported mediocre performances of their models. Compared to the literature, this study utilized three sensor devices comprising physiological signals of 23 sensor channels (see Section 3.4.2 and Figure 3.1) to detect human flow patterns. The effectiveness of each wearable device in terms of flow recognition with feature learning using CNN is illustrated in Figure 3.6, which evidences correlations between flow and physiological variables contributed by each device (such as HR, EDA, and EEG [64–69]). The multimodal approaches of the current study achieved notable performances (as shown in Tables 3.9, 3.10, and 3.11). They showed that it is feasible to measure the human flow states physiologically using machine learning with high accuracy using multimodal physiological sensor signals, and it is relevant because the flow has positive consequences [133–135].

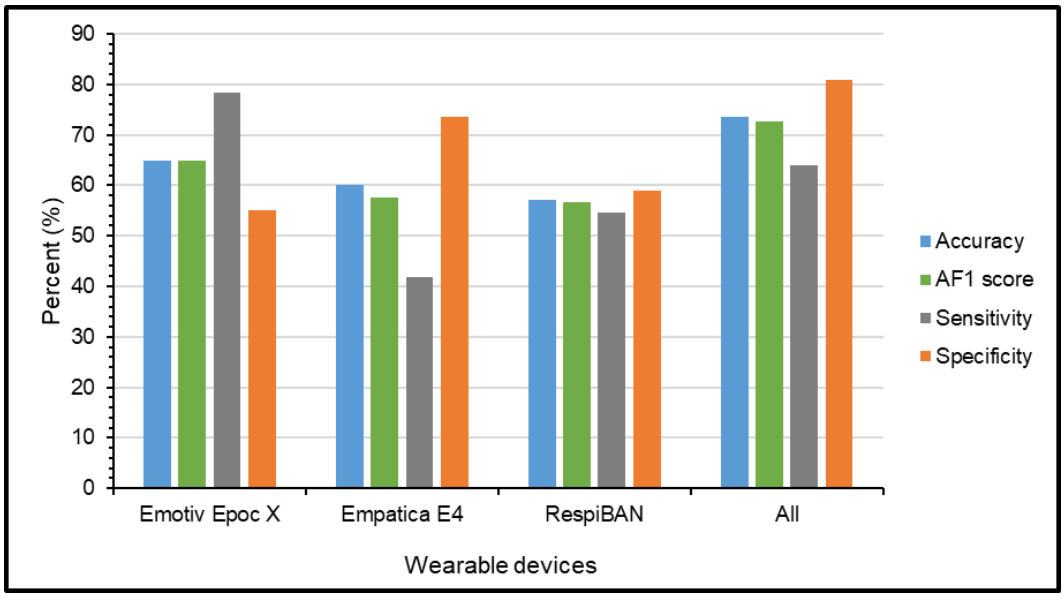

Figure 3.6: Comparison of sensor devices based on accuracy, averaged macro F1 (AF1) score, sensitivity, and specificity using feature learning with CNN. Emotiv: Emotiv Epoc X 14 channel EEG headset; Empatica: Empatica E4 wristband; RespiBAN: RespiBan professional wearable device, including ECG, EMG, and EOG sensors; All: Using all devices data (PhySF).

- **Comparison to the state-of-the-art:** It is worth noting that most previous work in this domain is based on manual feature engineering approaches. In light of the literature, only two past studies have used deep feature learning [50, 143], but

they also have some limitations. For example, Di Lascio et al. [50] mentioned that their work is based on something other than subject-independent cross-validation, which should be improved in future work to address the generalization capacity of the model. The work of Maier et al. [143] is also based on a subject-dependent cross-validation approach. Additionally, they also consider game users' data rather than work activities. In this study, the author focused on and implemented a subject-independent cross-validation approach, which is mandatory to obtain a good generalized model. Thus, experiments were performed with subject-independent SKF-CV, and this study achieved notable performances that outperformed the previous results.

- **Feature engineering against feature learning:** To extract meaningful features from the raw physiological measurements, both manual feature engineering and deep feature learning approaches were implemented and tested in this study. All manually created features, or HCFs, are listed in Table 2.3 and described in Section 2.6.1. However, the results of the manual feature engineering approaches were not convincing, as shown in Table 3.8. Thus, two deep feature learning-based approaches, such as MLP (i.e., a fully connected neural network) and CNN (i.e., a Convolutional Neural Network), were proposed. It is also worth noting that the deep feature learning approaches surpassed the manual feature engineering in the literature [99] in some application fields where enough data is available for training the models. In this study, optimal results were also obtained with a CNN-based deep feature learning approach, as shown in Table 3.10.

- **Relation between flow and emotion:** Based on relevant literature that showed a relationship between flow and emotions [130, 146], this study investigated a method to transfer emotion-related information to flow recognition. More specifically, a CNN-based transfer learning approach was implemented with DEAP as the source dataset and PhySF as the target dataset, which provided very promising results compared to the literature and proposed feature learning approaches such as MLP and CNN of this study when arousal classification was used as a source task, as shown in Table 3.11. Surprisingly, the transfer approach yielded slight improvement compared to the case without transfer whenever valence information was included in the target task. However, these results follow the literature findings as some researchers [131, 147] suggested that flow is associated with affect and arousal. In a similar order of ideas, research by [146, 154, 176] suggested that moderate physiological arousal promotes flow, while

boredom and stress (i.e., low and excessive physiological arousal) hinder it, as shown in Figure 3.7.

Regarding the relationship between flow and arousal, an experiment by Peifer et al. [146] provides evidence that even in the presence of a potentially negative stressor, flow can be experienced. The valence of emotional arousal, thus, can be subjectively reinterpreted when experiencing flow [177]. In line with this, the *Transactional Model of Stress and Flow* [137] proposes that a stressor that leads to a state of arousal can be interpreted as a challenge instead of a threat. A manageable challenge can result in the experience of flow [137]. This leads to the assumption that arousal is a reliable indicator for flow detection. In contrast, studies investigating emotional valence sensor data concerning flow showed inconsistent results, for example [137, 178–180], and thus not yet be seen as reliable indicators for flow. Future research, however, should focus not only on arousal but also on valence.

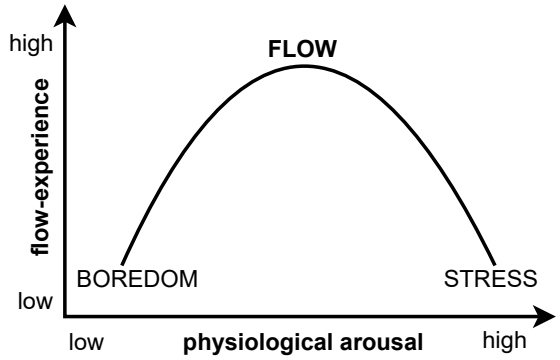

Figure 3.7: Relationship between flow experience and physiological arousal. *Y-axis:* represents flow experience (low-high). *X-axis:* represents physiological arousal (low-high). Boredom and stress, which represent low and excessive physiological arousal, hinder the flow state, while moderate physiological arousal promotes it [146].

3.10 Summary

This chapter demonstrates the feasibility of using multimodal wearable devices for human flow experience recognition and the possibility of employing emotion-related data to enhance human flow recognition. The author's proposed multimodal system based on deep transfer learning between the *DEAP* and the *PhySF* datasets discriminates between flow and non-flow states with an accuracy of 75.10% and an AF1 score of 74.92% in a subject-independent *SKF-CV* configuration. The results of

this study lead to the following conclusions: Firstly, it is possible to achieve good flow recognition performance with multimodal physiological signals. Secondly, extracting emotion-based information related to flow and using it to enhance flow recognition performances is also feasible with transfer learning approaches. Which can also help to circumvent the data scarcity issue in this domain. Lastly, feature learning approaches could perform better in the context of flow recognition than feature engineering approaches. However, a limitation of this study is the limited dataset, consisting of only 25 participants. Future studies conducted with more subjects are necessary to validate the results of this study further. In the future, the author would like to collect more data related to the collaborative work environment [124, 181] (i.e., team flow) and test the feasibility of the proposed approaches of this study to identify team flow. The author also wanted to investigate further research using only the *Empatica E4*, as other devices, especially the *Emotiv Epoc X* data, are intrusive and not easy to use in practice.

Chapter 4

Conclusion and Future Work

4.1 Summary

Adequate nutrition, satisfaction at work, proper hours of sleep, and physical recreation help promote good health [182]. Focusing on the impact of nutrition and work satisfaction on healthy living, including physical and mental health, this thesis presents two studies that analyze non-invasive multimodal time-series data using state-of-the-art machine learning algorithms to recognize stomach sensations as hunger and satiety and human flow experiences as flow and non-flow, respectively.

Chapter 2 details the research on non-invasive detection of stomach hunger and satiety sensations through physiological measurements. Literature highlights the importance of understanding these states for maintaining a healthy body weight and preventing chronic conditions like obesity, malnutrition, and deficiency syndromes. Historically, these states have been challenging to measure due to their subjective nature and the invasive nature of accurate measurement techniques. To address the aforementioned challenges, this study provides innovative approaches to detecting stomach sensations using non-invasive physiological measurements and machine learning. In the first step, a dataset named *Physiological Sense Hunger* (PhySH) is collected employing wearables such as the *Empatica E4* wristband, the *RespiBAN* wearable device, and *JINS MEME* smart glasses from healthy, normal-weight subjects inactively sitting on a chair in a state of hunger and satiety. After the necessary preprocessing steps, for example, synchronization, re-sampling, and interpolation, multimodal time-series data is

segmented using *Sliding Window Segmentation* (SWS). Following the *SWS*, features are extracted using two approaches, namely, feature engineering and deep feature learning. In the last phase, models are trained in a *Leave One Subject Out* (LOSO) cross-validation manner to identify these states through classification objectively. These steps are performed by using data from each device, each sensor channel, all the devices' data together, and certain (selected) features to identify the aforementioned states — to compare wearable devices, sensor channels, feature extraction approaches, and feature selection methods because there is no state-of-the-art research on this topic. This is the first study that recognizes hunger and satiety both objectively and non-invasively. The experiments revealed that:

- It is feasible to accurately recognize hunger and satiety through non-invasive physiological measurements.
- The *Empatica E4* wristband is the more influential device than the others and can be used as a stand-alone wearable for future data collection related to this probe.
- Feature learning approaches do not necessarily perform well in the case of a limited dataset compared to manually engineered features with proper feature selection.
- The *Random Forest* (RF) algorithm performed better as a classifier and feature selector than others. In the future, it can be utilized as a stand-alone for related probes.

With the successful application of sensor-based, non-invasive experimental strategies, the author expected that it could contribute to the identification of human flow experiences as well. The flow experience is a specific positive and affective state of mind when humans are completely absorbed in an activity and forget everything else [130]. The literature has revealed a positive relationship between the flow experience and mental health [183] and emphasized the use of it in psychotherapy and mental health rehabilitation, showing that finding flow in everyday life is connected to well-being and reduced symptomatology [184]. In addition to this, flow has been shown to lead to positive outcomes in the workplace, such as high task performance and increased job satisfaction [135], which can bring benefits for both employees and employers [136]. The positive consequences of flow underline the relevance of measuring and recognizing it. In this context, Chapter 3 introduces the hierarchical

models designed to objectively recognize human flow experiences, such as flow and non-flow, without interruption, using unobtrusive wearable sensor technology and machine learning algorithms.

Since most of the previous work is based on self-reports (which are problematic because once people are interrupted, they feel out of the flow), or experiments are conducted in the context of users playing games, having mediocre recognition performance, and not performing the experiments in an objective way. Therefore, this chapter focuses on the development of a wearable-based, objective, and interruption-free system in the work context. As there is no public dataset available, in the first step, a dataset named *Physiological Sense Flow* (PhySF) is collected using multimodal wearables such as the *Empatica E4* wristband, the *RespiBAN* wearable, and the *Emotiv Epoc X* from 25 participants when they were doing arithmetic and reading task activities.

In literature highlights, feature extraction is a decisive step in the *Pattern Recognition Pipeline* (PRP). Therefore, experiments are performed to recognize human flow experiences using three feature extraction methods: feature engineering, feature learning, and the proposed feature learning using transfer learning. Compared to the literature, remarkable performance (75.10%) is achieved by employing feature learning when an emotion recognition-based (DEAP) dataset is utilized as the source dataset to classify arousal data into high vs. low arousal using a deep source CNN model. During the classification of the *PhySF* dataset into flow and non-flow states, the weights learned from the previous step are used to initialize the deep target CNN model. The results of this study revealed that:

- The emotions, particularly arousal and flow experience, are connected, and an emotion-based dataset can be used as a latent task to enhance flow recognition performances compared to the emotional valence data concerning flow, which showed inconsistent results and thus not yet be seen as reliable indicators for flow [137, 178–180].
- The use of the source dataset using the transfer learning technique can also help to solve the data scarcity issue of the flow recognition domain.
- The multimodal wearable devices provide good quality physiological data that can be utilized in the psychological context, such as to identify flow experiences interruption-free, objectively, and in real-time. Furthermore, measuring flow in

an interruption-free manner would allow the researcher to better understand the experience without interfering with the mechanisms themselves. In this way, flow situations can be better understood, anticipated, and actively created in order to benefit from the positive consequences of the experience.

To summarize the aforementioned brief description of the work, the findings of this thesis highlight the encouraging potential of using ubiquitous wearable sensors to recognize stomach sensations such as hunger and satiety and human flow experience states such as flow and non-flow states using machine learning. Particularly, the scientific work presented to detect non-invasive and objective hunger and satiety is of great importance to maintaining a healthy body weight and avoiding chronic diseases such as obesity, underweight, or deficiency syndromes due to malnutrition. Similarly, work presented to recognize human flow experience using wearables can help measure flow in an interruption-free manner, which will allow for a better understanding of experience without interfering with the mechanisms themselves. This would also lead to positive outcomes at work, such as high task performance and increased job satisfaction.

4.2 Scientific Findings

As far as is known, the research work presented in *Chapter 2* is the first in which hunger and satiety states are recognized for the first time from multimodal physiological measurements. After extensive investigative experiments, this study scientifically contributed the following findings:

- It has been found that well-engineered and selected features on hunger and satiety time-series data can perform better than deep learning approaches in the case of a limited number of training samples. This finding is investigated by using the *Feature Importance Ranking* (FIR), which measures each input feature's contribution to the model's performance. It turned out that the most accurate results can be obtained only with the best 18 HCFs (as illustrated in Figure 2.11, Chapter 2), and the addition of other irrelevant and redundant features can introduce noise into the data, which can reduce the performance of a classifier. It can be pointed out that the top five features come exclusively from three different

sensor modalities: *EDA*, *Tmp*, and *BVP*, and are either computing the mean or the $80th$ percentile of the data values. Percentile 80 provides an approximation of the maximum value in a data segment that is less sensitive to noise or outliers than the actual maximum computation. This indicates that the average and upper data values in these three sensor modalities are of high importance to distinguish between hunger and satiety.

- Eliminating irrelevant sensors can decrease the degree of discomfort for subjects, improve the robustness of the hunger recognition system by reducing its dimensionality, and also save a lot of money [129]. Therefore, all sensor channels and wearable devices (see Figures 2.1, 2.2, and 2.3) are compared to determine the most suitable sensor channel and wearable device for hunger and satiety detection. It was discovered that the *Photoplethysmography*, also known as PPG (including BVP, IBI, and HR measures), EDA (Empatica E4 and RespiBAN), Tmp, and EMG were the appropriate sensor modalities for hunger detection, and Resp, ECG, and EOG were the least appropriate (as shown in Figure 2.8). It was also found that the Empatica E4 wristband is the most suitable device compared to the other devices (see Figure 2.9), and in the future, it can be used as a standalone device to acquire hunger and satiety-state relevant data.

The research work related to human flow experience recognition is investigated and presented in *Chapter 3*, in which flow and non-flow states are recognized objectively from the multimodal physiological measurements. Extensive investigative experiments are performed to answer the question of *how to circumvent the data scarcity issue* in this domain and *how to objectively increase flow recognition performances*. In addition, feature extraction techniques and wearable sensor channels are also compared to find the most suitable one for this domain. In general, this study contributed the following scientific findings:

- In light of the literature that shows a relationship between flow and emotions [130, 146], a transfer learning-based method to transfer emotion-related information to flow recognition is investigated in order to circumvent the data scarcity issue and achieve improved flow recognition performances. For this investigation, a CNN-based transfer learning approach with *DEAP* as the source dataset and *PhySF* as the target dataset was utilized, which outperformed the results of the author's CNN-based flow recognition approach and previous results

in the literature when arousal classification was used as a source task (as shown in Table 3.11). Nevertheless, this transfer approach showed a slight improvement compared to the case without transfer whenever valence information was included in the target task. However, these results follow the literature findings, as some researchers [131, 147] suggested that flow is associated with affect and arousal. This leads to the finding that arousal is a reliable indicator for flow detection. In contrast, studies investigating emotional valence sensor data concerning flow showed inconsistent results, for example [137, 178–180], and thus are not yet seen as reliable indicators for flow.

- Emotiv Epoc X — an EEG headband, is the most effective wearable for human flow experience recognition compared to others such as Empatica E4 and RespiBAN. This was investigated using comparative analysis (see Tables 3.9, 3.10, and Figure 3.6). Nevertheless, it is worth mentioning that high recognition performances were obtained using combined multimodal data from all wearable devices.

- Experiments also revealed that *feature learning* is the more potential feature extraction approach than manual *feature engineering* (see Table 3.12) for human flow experience-based time-series data. Nevertheless, enhanced recognition performances were achieved using the CNN-based transfer learning approach (which is also a type of feature learning), when the *DEAP* dataset was used as the source dataset and *PhySF* as the target dataset.

4.3 Limitations

In this thesis, various hierarchical frameworks are proposed that are based on multimodal physiological measurements (*PhySH* and *PhySF* datasets) using either manual feature engineering or deep feature learning approaches to recognize hunger and satiety — and human flow experiences (see Chapters 2 and 3, respectively). Despite the overall performance being relatively promising, there are a few limitations that hinder the practical use of the proposed frameworks:

- The PhySH dataset, which is acquired and investigated (as defined in Chapter 2), is limited in size not only by the number of participants but also by the number

of samples of each class (particularly the hunger class). This limited dataset size hinders the proposed models' practical use.

- The proposed models for hunger and state detection (see Chapter 2) faced the data scarcity issue. As, there is no public dataset on this problem, and the dataset acquired in this study (*PhySH*) is also small. However, deep transfer learning has been used to solve this problem in many application domains, which should be investigated to solve the scarcity issue of this study.

- The *PhySF* dataset, which is acquired and investigated (as described in Chapter 3) to recognize the human flow experiences during work activities, contains the physiological measurements in the context of arithmetic and reading tasks, respectively. To generalize the proposed frameworks so that they can be used in real life, physiological recordings in other working contexts should be included as well.

- Compared to the most recent research in this field, the author's proposed deep transfer learning-based model (as illustrated in Table 3.11, Chapter 3) is able to recognize about 75.10% of flow experiences. This is significant. However, there are still some misclassifications between the flow and non-flow states. The main reason for this is the noisy labels in the *PhySF* dataset. This is because the participants or subjects of the study labeled their data after each task via an online questionnaire. These labels are not yet investigated with the help of field experts such as psychologists or by some flow measurement scale.

- The DEAP emotion analysis dataset is utilized as a source dataset (see Section 3.4.1, Chapter 3) to improve the performance of the proposed transfer learning-based flow experience recognition model, which restricts the actual use of this model in real. To check the true effectiveness of this model, it should also be tested with some other related source datasets.

4.4 Future Work

Regarding the limitations outlined in Section 4.3, the following points provide specific guidance for potential future research endeavors:

- As mentioned earlier, the *PhySH* dataset, which contained 5 minutes of hunger and 30 minutes of satiety physiological measurements, was used to recognize the hunger and satiety classes [22]. In the future, long-term measurements with different food compositions [185] should be acquired from actual patients. It can facilitate a more in-depth exploration of sub-classes within the dataset or test food hypersensitivity [186].

- To address the class imbalance and data scarcity problems for hunger and satiety recognition using the *PhySH* dataset, data augmentation techniques can be investigated in the future. Data augmentation techniques increase the sample size of a particular or all classes. They create modified copies of the given samples, which leads to an increase in sample size [187].

- Another approach that can support circumventing the data scarcity issue and improving hunger and satiety recognition performances is transfer learning. It has previously demonstrated its utility on time-series data in various domains, including human flow experience recognition and human activity recognition [59, 170, 188]. Given this, it appears promising for improving hunger and satiety recognition while mitigating data scarcity concerns [187].

- The current human flow experience recognition performance is about 75.10% [59], likely due to the noisy flow and non-flow class labels. The noisy data can be identified using *unsupervised machine learning* techniques such as cluster analysis [189]. A cluster analysis can help identify natural groupings or patterns in the dataset, ushering towards the identification of outliers or noise. Removing or correcting these can enhance the quality of training classifiers, which can lead to better classification accuracy. This can be investigated to improve flow recognition performances.

- Alternatively, a meta-learning algorithm known as *contrastive learning* could be tested to enhance flow recognition performance. It is based on the principle of contrasting samples against each other to learn features that seem common between the classes and features that set apart a class from another [190]. This approach has previously demonstrated success on video and time-series data for other domains, such as human activity recognition and sleep stage classification [190–192]. Therefore, investigating the potential of contrastive learning for enhancing flow recognition performance seems promising for future work.

- To mitigate data discrepancies [193] in data acquisition caused by diverse sensor devices, it is also advisable to test various wearables with the characteristics of being portable, user-friendly, non-invasive, and affordable, for example, Oura ring [194], Fitbit sense [195], Interaxon Muse [196], Masimo SedLine® Brain [197], etc.
- The proposed models can be tested on other publicly available, profound datasets to show their efficacy and generalizability.

Bibliography

[1] David E Bloom and David Canning. “The health and wealth of nations”. In: *Science* 287.5456 (2000), pp. 1207–1209.

[2] Carol Graham, Lucas Higuera, and Eduardo Lora. “Which health conditions cause the most unhappiness?” In: *Health economics* 20.12 (2011), pp. 1431–1447.

[3] Paul Dolan and Robert Metcalfe. “Valuing Health: A Brief Report on Subjective Well-Being versus Preferences”. In: *Medical Decision Making* 32.4 (2012), pp. 578–582.

[4] Chris Naylor, Preety Das, Shilpa Ross, Matthew Honeyman, James Thompson, and Helen Gilburt. “Bringing together physical and mental health”. In: *King’s Fund* 109.10 (2016), pp. 364–366.

[5] Barret Rush, Leo Anthony Celi, and David J Stone. “Applying machine learning to continuously monitored physiological data”. In: *Journal of clinical monitoring and computing* 33 (2019), pp. 887–893.

[6] Erick Martinez-Ríos, Luis Montesinos, Mariel Alfaro-Ponce, and Leandro Pecchia. “A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data”. In: *Biomedical Signal Processing and Control* 68 (2021), p. 102813.

[7] David C Mohr, Mi Zhang, and Stephen M Schueller. “Personal sensing: understanding mental health using ubiquitous sensors and machine learning”. In: *Annual review of clinical psychology* 13 (2017), pp. 23–47.

[8] Irvin Hussein Lopez-Nava and Angelica Munoz-Melendez. “Wearable inertial sensors for human motion analysis: A review”. In: *IEEE Sensors Journal* 16.22 (2016), pp. 7821–7834.

[9] Tung Khuc and F Necati Catbas. “Completely contactless structural health monitoring of real-life structures using cameras and computer vision”. In: *Structural Control and Health Monitoring* 24.1 (2017), e1852.

[10] Yiding Gu, Ting Zhang, Hao Chen, Feng Wang, Yueming Pu, Chunming Gao, and Shibin Li. “Mini review on flexible and wearable electronics for monitoring human health information”. In: *Nanoscale research letters* 14 (2019), pp. 1–15.

[11] Tyler R Ray, Maja Ivanovic, Paul M Curtis, Daniel Franklin, Kerem Guventurk, William J Jeang, Joseph Chafetz, Hannah Gaertner, Grace Young, Steve Rebollo, et al. “Soft, skin-interfaced sweat stickers for cystic fibrosis diagnosis and management”. In: *Science translational medicine* 13.587 (2021), p. 8109.

[12] Guanglei Li and Dan Wen. “Wearable biochemical sensors for human health monitoring: sensing materials and manufacturing technologies”. In: *Journal of materials chemistry B* 8.16 (2020), pp. 3423–3436.

[13] Zhongyu Li, Mengmeng Xiao, Chuanhong Jin, and Zhiyong Zhang. “Toward the Commercialization of Carbon Nanotube Field Effect Transistor Biosensors”. In: *Biosensors* 13.3 (2023), p. 326.

[14] Niloy Chatterjee, Krishnendu Manna, Niladri Mukherjee, and Krishna Das Saha. “Challenges and future prospects and commercial viability of biosensor-based devices for disease diagnosis”. In: *Biosensor Based Advanced Cancer Diagnostics* (2022), pp. 333–352.

[15] Zhenghui Li, Julien Le Kernec, Qammer Abbasi, Francesco Fioranelli, Shufan Yang, and Olivier Romain. “Radar-based human activity recognition with adaptive thresholding towards resource constrained platforms”. In: *Scientific Reports* 13.1 (2023), p. 3473.

[16] Giovanni Diraco, Alessandro Leone, and Pietro Siciliano. “A radar-based smart sensor for unobtrusive elderly monitoring in ambient assisted living applications”. In: *Biosensors* 7.4 (2017), p. 55.

[17] Ruth Ravichandran, Sang-Wha Sien, Shwetak N Patel, Julie A Kientz, and Laura R Pina. “Making sense of sleep sensors: How sleep sensing technologies support and undermine sleep health”. In: *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2017, pp. 6864–6875.

[18] Sharon Keenan and Max Hirshkowitz. “Monitoring and staging human sleep”. In: *Principles and practice of sleep medicine* 5 (2011), pp. 1602–1609.

[19] Lina Zhou, Shimei Pan, Jianwu Wang, and Athanasios V Vasilakos. “Machine learning on big data: Opportunities and challenges”. In: *Neurocomputing* 237 (2017), pp. 350–361.

[20] Gabriella Kazai. “In search of quality in crowdsourcing for search engine evaluation”. In: *Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings 33*. Springer. 2011, pp. 165–176.

[21] Muhammad Tausif Irshad, Muhammad Adeel Nisar, Philip Gouverneur, Marion Rapp, and Marcin Grzegorzek. “Ai approaches towards Prechtl’s assessment of general movements: A systematic literature review”. In: *Sensors* 20.18 (2020), p. 5321.

[22] Muhammad Tausif Irshad, Muhammad Adeel Nisar, Xinyu Huang, Jana Hartz, Olaf Flak, Frédéric Li, Philip Gouverneur, Artur Piet, Kerstin M Oltmanns, and Marcin Grzegorzek. “SenseHunger: Machine Learning Approach to Hunger Detection Using Wearable Sensors”. In: *Sensors* 22.20 (2022), p. 7711.

[23] Irina Rish et al. “An empirical study of the naive Bayes classifier”. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22. 2001, pp. 41–46.

[24] William S Noble. “What is a support vector machine?” In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

[25] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. “Support vector machine classification and validation of cancer tissue samples using microarray expression data”. In: *Bioinformatics* 16.10 (2000), pp. 906–914.

[26] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. “An introduction to decision tree modeling”. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 18.6 (2004), pp. 275–285.

[27] Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R Sveinsson. “Random forests for land cover classification”. In: *Pattern recognition letters* 27.4 (2006), pp. 294–300.

[28] Robert E Schapire. “Explaining adaboost”. In: *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (2013), pp. 37–52.

[29] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[30] Yung-Chia Chang, Kuei-Hu Chang, and Guan-Jhih Wu. "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions". In: *Applied Soft Computing* 73 (2018), pp. 914–920.

[31] Bruce G Buchanan. "A (very) brief history of artificial intelligence". In: *Ai Magazine* 26.4 (2005), pp. 53–53.

[32] Frédéric Li. *Deep Earning for Time-series Classification Enhanced by Transfer Learning Based on Sensor Modality Discrimination*. Logos Verlag, 2021.

[33] Roland Memisevic, Christopher Zach, Marc Pollefeys, and Geoffrey E Hinton. "Gated softmax classification". In: *Advances in neural information processing systems* 23 (2010).

[34] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[35] Katarzyna Janocha and Wojciech Marian Czarnecki. "On loss functions for deep neural networks in classification". In: *arXiv preprint arXiv:1702.05659* (2017).

[36] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. *Learning internal representations by back-propagating errors in Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Eds*. 1986.

[37] Paul J Werbos. "Backpropagation through time: what it does and how to do it". In: *Proceedings of the IEEE* 78.10 (1990), pp. 1550–1560.

[38] Philip Gouverneur, Frédéric Li, Wacław M Adamczyk, Tibor M Szikszay, Kerstin Luedtke, and Marcin Grzegorzek. "Comparison of feature extraction methods for physiological signals for heat-based pain recognition". In: *Sensors* 21.14 (2021), p. 4838.

[39] Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, and Giacomo Boracchi. "Online anomaly detection for long-term ECG monitoring using wearable devices". In: *Pattern Recognition* 88 (2019), pp. 482–492.

[40] Preeti Khera and Neelesh Kumar. "Role of machine learning in gait analysis: a review". In: *Journal of Medical Engineering & Technology* 44.8 (2020), pp. 441–467.

[41] M Umair Bin Altaf, Taras Butko, and Biing-Hwang Juang. “Acoustic gaits: Gait analysis with footstep sounds”. In: *IEEE Transactions on Biomedical Engineering* 62.8 (2015), pp. 2001–2011.

[42] Da Ma, Vincent Chow, Karteek Popuri, and Mirza Faisal Beg. “Comprehensive validation of automated whole body skeletal muscle, adipose tissue, and bone segmentation from 3D CT images for body composition analysis: Towards extended body composition”. In: *arXiv preprint arXiv:2106.00652* (2021).

[43] Jiyeon Ha, Taeyong Park, Hong-Kyu Kim, Youngbin Shin, Yousun Ko, Dong Wook Kim, Yu Sub Sung, Jiwoo Lee, Su Jung Ham, Seungwoo Khang, et al. “Development of a fully automatic deep learning system for L3 selection and body composition assessment on computed tomography”. In: *Scientific reports* 11.1 (2021), p. 21656.

[44] Adil Mehmood Khan, Young-Koo Lee, Sungyoung Y Lee, and Tae-Seong Kim. “A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer”. In: *IEEE transactions on information technology in biomedicine* 14.5 (2010), pp. 1166–1172.

[45] Muhammad Adeel Nisar, Kimiaki Shirahama, Frédéric Li, Xinyu Huang, and Marcin Grzegorzek. “Rank pooling approach for wearable sensor-based ADLs recognition”. In: *Sensors* 20.12 (2020), p. 3463.

[46] Igor Bisio, Fabio Lavagetto, Mario Marchese, and Andrea Sciarrone. “Smartphone-based user Activity Recognition Method for Health Remote Monitoring Applications.” In: *PECCS*. 2012, pp. 200–205.

[47] MAX Hamilton. “Development of a rating scale for primary depressive illness”. In: *British journal of social and clinical psychology* 6.4 (1967), pp. 278–296.

[48] Hodjat Rahmati, Harald Martens, Ole Morten Aamo, Øyvind Stavdahl, Ragnhild Støen, and Lars Adde. “Frequency analysis and feature reduction method for prediction of cerebral palsy in young infants”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 24.11 (2016), pp. 1225–1234.

[49] Isaac Moshe, Yannik Terhorst, Kennedy Opoku Asare, Lasse Bosse Sander, Denzil Ferreira, Harald Baumeister, David C Mohr, and Laura Pulkki-Råback. “Predicting symptoms of depression and anxiety using smartphone and wearable data”. In: *Frontiers in psychiatry* 12 (2021), p. 625247.

[50] Elena Di Lascio, Shkurta Gashi, Maike E Debus, and Silvia Santini. "Automatic Recognition of Flow During Work Activities Using Context and Physiological Signals". In: *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2021, pp. 1–8.

[51] Mahnoosh Sadeghi, Anthony D McDonald, and Farzan Sasangohar. "Posttraumatic stress disorder hyperarousal event detection using smartwatch physiological and activity data". In: *Plos one* 17.5 (2022), e0267749.

[52] Jing Zhang, J Don Richardson, and Benjamin T Dunkley. "Classifying post-traumatic stress disorder using the magnetoencephalographic connectome and machine learning". In: *Scientific reports* 10.1 (2020), p. 5937.

[53] Ricardo Buettner, David Beil, Stefanie Scholtz, and Aadel Djemai. "Development of a machine learning based algorithm to accurately detect schizophrenia based on one-minute EEG recordings". In: (2020).

[54] Donald J Chmielewski, Tasha Palmer, and Vasilios Manousiouthakis. "On the theory of optimal sensor placement". In: *AIChE journal* 48.5 (2002), pp. 1001–1012.

[55] Suranga Seneviratne, Yining Hu, Tham Nguyen, Guohao Lan, Sara Khalifa, Kanchana Thilakarathna, Mahbub Hassan, and Aruna Seneviratne. "A survey of wearable devices and challenges". In: *IEEE Communications Surveys & Tutorials* 19.4 (2017), pp. 2573–2620.

[56] HY Guo, L Zhang, LL Zhang, and JX Zhou. "Optimal placement of sensors for structural health monitoring using improved genetic algorithms". In: *Smart materials and structures* 13.3 (2004), p. 528.

[57] Wei Tong, Bingbing Jiang, Fengyuan Xu, Qun Li, and Sheng Zhong. "Privacy-preserving data integrity verification in mobile edge computing". In: *2019 IEEE 39th international conference on distributed computing systems (ICDCS)*. IEEE. 2019, pp. 1007–1018.

[58] Michael Rosenblum, Arkady Pikovsky, Jurgen Kurths, Carsten Schäfer, and Peter A Tass. "Phase synchronization: from theory to data analysis". In: *Handbook of biological physics*. Vol. 4. Elsevier, 2001, pp. 279–321.

[59] Muhammad Tausif Irshad, Frédéric Li, Muhammad Adeel Nisar, Xinyu Huang, Martje Buss, Leonie Kloep, Corinna Peifer, Barbara Kozusznik, Anita Pollak, Adrian Pyszka, et al. "Wearable-based human flow experience recognition

enhanced by transfer learning methods using emotion data". In: *Computers in Biology and Medicine* (2023), p. 107489.

[60] Kamila Jauch-Chara and Kerstin M Oltmanns. "Obesity–a neuropsychological disease? Systematic review and neuropsychological model". In: *Progress in neurobiology* 114 (2014), pp. 84–101.

[61] Mary C Gannon, Frank Q Nuttall, James T Lane, Sean Fang, Vinendra Gupta, and Charles R Sandhofer. "Effect of 24 hours of starvation on plasma glucose and insulin concentrations in subjects with untreated non—insulin-dependent diabetes mellitus". In: *Metabolism* 45.4 (1996), pp. 492–497.

[62] Ann E Macpherson-Sánchez. "Integrating fundamental concepts of obesity and eating disorders: implications for the obesity epidemic". In: *American journal of public health* 105.4 (2015), e71–e85.

[63] WHO. *Obesity and overweight*. `https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight`. [Online; accessed 27-June-2021].

[64] Shaji Krishnan, Henk FJ Hendriks, Merete L Hartvigsen, and Albert A de Graaf. "Feed-forward neural network model for hunger and satiety related VAS score prediction". In: *Theoretical Biology and Medical Modelling* 13.1 (2016), pp. 1–12.

[65] Barbara A Parker, K Sturm, CG MacIntosh, C Feinle, M Horowitz, and IM Chapman. "Relation between food intake and visual analogue scale ratings of appetite and other sensations in healthy older and young subjects". In: *European journal of clinical nutrition* 58.2 (2004), pp. 212–218.

[66] CP Sepple and NW Read. "Gastrointestinal correlates of the development of hunger in man". In: *Appetite* 13.3 (1989), pp. 183–191.

[67] Peter J Rogers and John E Blundell. "Effect of anorexic drugs on food intake and the micro-structure of eating in human subjects". In: *Psychopharmacology* 66.2 (1979), pp. 159–165.

[68] Xinyu Huang, Kimiaki Shirahama, Muhammad Tausif Irshad, Muhammad Adeel Nisar, Artur Piet, and Marcin Grzegorzek. "Sleep Stage Classification in Children Using Self-Attention and Gaussian Noise Data Augmentation". In: *Sensors* 23.7 (2023), p. 3446.

[69] Muhammad Adeel Nisar, Kimiaki Shirahama, Muhammad Tausif Irshad, Xinyu Huang, and Marcin Grzegorzek. "A Hierarchical Multitask Learning Approach for the Recognition of Activities of Daily Living Using Data from Wearable Sensors". In: *Sensors* 23.19 (2023), p. 8234.

[70] Rafał Doniec, Justyna Konior, Szymon Sieciński, Artur Piet, Muhammad Tausif Irshad, Natalia Piaseczna, Md Abid Hasan, Frédéric Li, Muhammad Adeel Nisar, and Marcin Grzegorzek. "Sensor-Based Classification of Primary and Secondary Car Driver Activities Using Convolutional Neural Networks". In: *Sensors* 23.12 (2023), p. 5551.

[71] Riccardo Berta, Francesco Bellotti, Alessandro De Gloria, Danu Pranantha, and Carlotta Schatten. "Electroencephalogram and physiological signal analysis for assessing flow in games". In: *IEEE Transactions on Computational Intelligence and AI in Games* 5.2 (2013), pp. 164–175.

[72] Xinyu Huang, Kimiaki Shirahama, Frederic Li, and Marcin Grzegorzek. "Sleep stage classification for child patients using DeConvolutional Neural Network". In: *Artificial Intelligence in Medicine* 110 (2020), p. 101981.

[73] Arkan Al-Zubaidi, Alfred Mertins, Marcus Heldmann, Kamila Jauch-Chara, and Thomas F Münte. "Machine learning based classification of resting-state fMRI features exemplified by metabolic state (hunger/satiety)". In: *Frontiers in human neuroscience* 13 (2019), p. 164.

[74] A Maria and A Sengol Jeyaseelan. "Development of Optimal Feature Selection and Deep Learning Toward Hungry Stomach Detection Using Audio Signals". In: *Journal of Control, Automation and Electrical Systems* 32.4 (2021), pp. 853–874.

[75] Ahmed S BaHammam and Michael WL Chee. *Publicly available health research datasets: opportunities and responsibilities*. 2022.

[76] Yasuhiro Sato and Shin Fukudo. "Gastrointestinal symptoms and disorders in patients with eating disorders". In: *Clinical journal of gastroenterology* 8 (2015), pp. 255–263.

[77] Susann Bellmann, Shaji Krishnan, Albert de Graaf, Rianne A de Ligt, Wilrike J Pasman, Mans Minekus, and Robert Havenaar. "Appetite ratings of foods are predictable with an in vitro advanced gastrointestinal model in combination with an in silico artificial neural network". In: *Food research international* 122 (2019), pp. 77–86.

[78] Tauhidur Rahman, Mary Czerwinski, Ran Gilad-Bachrach, and Paul Johns. “Predicting" about-to-eat" moments for just-in-time eating intervention”. In: *Proceedings of the 6th International Conference on Digital Health Conference*. 2016, pp. 141–150.

[79] S Lakshmi, P Kavipriya, MR Ebenezar Jebarani, and T Vino. “A Novel Approach of Human Hunger Detection especially for physically challenged people”. In: *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE. 2021, pp. 921–927.

[80] Uttara Gogate and Jagdish Bakal. “Hunger and stress monitoring system using galvanic skin”. In: *Indonesian Journal of Electrical Engineering and Computer Science* 13.3 (2019), pp. 861–865.

[81] Sandra E Barajas-Montiel and Carlos A Reyes-Garcia. “Identifying pain and hunger in infant cry with classifiers ensembles”. In: *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06)*. Vol. 2. IEEE. 2005, pp. 770–775.

[82] Dong Yu, Michael L Seltzer, Jinyu Li, Jui-Ting Huang, and Frank Seide. “Feature learning in deep neural networks-studies on speech recognition tasks”. In: *arXiv preprint arXiv:1301.3605* (2013).

[83] *respiBAN*. https://plux.info/biosignalsplux-wearables/313-respiban-professional-820202407.html. [Online; accessed 08-Aug-2021].

[84] *Empatica E4 Wristband*. https://www.empatica.com/research/e4/. [Online; accessed 25-Jan-2023].

[85] *JINS MEME: Eyewear that Sees Your EVERYDAY*. https://jins-meme.com/en/. [Online; accessed 08-Aug-2021].

[86] Marc Hesse, Peter Christ, Timm Hörmann, and Ulrich Rückert. “A respiration sensor for a chest-strap based wireless body sensor”. In: *SENSORS, 2014 IEEE*. IEEE. 2014, pp. 490–493.

[87] *Electrodermal activity (EDA)*. https://plux.info/sensors/280-electrodermal-activity-eda-820201202.html. [Online; accessed 18-Aug-2021].

[88] *Electrocardiography*. https://plux.info/sensors/277-electrocardiogram-ecg-820201203.html. [Online; accessed 25-Jan-2023].

[89] *Electromyography (EMG)*. `https://plux.info/sensors/283-electromyography-emg-820201201.html`. [Online; accessed 18-Aug-2021].

[90] John R Murlin. "Skin temperature, its measurement and significance for energy metabolism". In: *Ergebnisse der Physiologie, biologischen Chemie und experimentellen Pharmakologie* 42 (1939), pp. 153–227.

[91] Lawrence M Baker and William M Taylor. "The relationship under stress between changes in skin temperature, electrical skin resistance, and pulse rate." In: *Journal of experimental psychology* 48.5 (1954), p. 361.

[92] Stéphanie Khalfa, Peretz Isabelle, Blondin Jean-Pierre, and Robert Manon. "Event-related skin conductance responses to musical emotions in humans". In: *Neuroscience letters* 328.2 (2002), pp. 145–149.

[93] Yuji Uema and Kazutaka Inoue. "JINS MEME algorithm for estimation and tracking of concentration of users". In: *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 2017, pp. 297–300.

[94] Sotiris B Kotsiantis, Dimitris Kanellopoulos, and Panagiotis E Pintelas. "Data preprocessing for supervised leaning". In: *International journal of computer science* 1.2 (2006), pp. 111–117.

[95] Chia-Shang James Chu. "Time series segmentation: A sliding window approach". In: *Information Sciences* 85.1-3 (1995), pp. 147–173.

[96] Akbar Dehghani, Omid Sarbishei, Tristan Glatard, and Emad Shihab. "A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors". In: *Sensors* 19.22 (2019), p. 5026.

[97] Gaojing Wang, Qingquan Li, Lei Wang, Wei Wang, Mengqi Wu, and Tao Liu. "Impact of sliding window length in indoor human motion modes and pose pattern recognition based on smartphone sensors". In: *Sensors* 18.6 (2018), p. 1965.

[98] Diane J Cook and Narayanan C Krishnan. *Activity learning: discovering, recognizing, and predicting human behavior from sensor data*. John Wiley & Sons, 2015.

[99] Frédéric Li, Kimiaki Shirahama, Muhammad Adeel Nisar, Lukas Köping, and Marcin Grzegorzek. "Comparison of feature learning methods for human activity recognition using wearable sensors". In: *Sensors* 18.2 (2018), p. 679.

[100] EP Box George, M Jenkins Gwilym, C Reinsel Gregory, and M Ljung Greta. "Time series analysis: forecasting and control". In: *San Francisco: Holden Bay* (1976).

[101] *How Statistical Norms Improve Modeling. Available online:* https://towardsdatascience.com/norms-penalties-and-multitask-learning-2f1db5f97c1f. [Online; accessed 05-Sep-2023].

[102] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. "Variable selection using random forests". In: *Pattern Recognition Letters* 31.14 (2010), pp. 2225–2236. DOI: 10.1016/j.patrec.2010.03.014. URL: https://doi.org/10.1016/j.patrec.2010.03.014.

[103] Xindong Wu and Vipin Kumar. *The top ten algorithms in data mining*. CRC press, 2009.

[104] Cuong Nguyen, Yong Wang, and Ha-Nam Nguyen. "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic". In: *Journal of Biomedical Science and Engineering* 06 (Jan. 2013), pp. 551–560. DOI: 10.4236/jbise.2013.65070.

[105] Rung-Ching Chen, Christine Dewi, Su-Wen Huang, and Rezzy Eko Caraka. "Selecting critical features for data classification based on machine learning methods". In: *Journal of Big Data* 7.1 (2020), pp. 1–26.

[106] D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. "Random forests for classification in ecology". In: *Ecology* 88.11 (2007), pp. 2783–2792.

[107] Bardan Ghimire, John Rogan, and Jennifer Miller. "Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic". In: *Remote Sensing Letters* 1.1 (2010), pp. 45–54.

[108] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

[109] Xiuzhi Sang, Wanyue Xiao, Huiwen Zheng, Yang Yang, and Taigang Liu. "HMMPred: accurate prediction of DNA-binding proteins based on HMM profiles and XGBoost feature selection". In: *Computational and mathematical methods in medicine* 2020 (2020).

[110] Witold R Rudnicki, Mariusz Wrzesień, and Wiesław Paja. "All relevant feature selection methods and applications". In: *Feature Selection for Data and Pattern Recognition*. Springer, 2015, pp. 11–28.

[111] Miron B. Kursa and Witold R. Rudnicki. "Feature Selection with the Boruta Package". In: *Journal of Statistical Software* 36.11 (2010), pp. 1–13.

[112] Abdullah-Al Nahid, Mohamad Ali Mehrabi, and Yinan Kong. "Histopathological breast cancer image classification by deep neural network techniques guided by local clustering". In: *BioMed research international* 2018 (2018).

[113] Umut Orhan, Mahmut Hekim, and Mahmut Ozer. "EEG signals classification using the K-means clustering and a multilayer perceptron neural network model". In: *Expert Systems with Applications* 38.10 (2011), pp. 13475–13481.

[114] Zhuoran Chen, Gege Ma, Yandan Jiang, Baoliang Wang, and Manuchehr Soleimani. "Application of deep neural network to the reconstruction of two-phase material imaging by capacitively coupled electrical resistance tomography". In: *Electronics* 10.9 (2021), p. 1058.

[115] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[116] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[117] Regan Mandryk and Madison Klarkowski. "Physiological measures for game evaluation". In: *Game Usability*. CRC Press, 2008, pp. 161–187.

[118] Wei He, Sanne Boesveldt, Sylvain Delplanque, Cees de Graaf, and René A De Wijk. "Sensory-specific satiety: Added insights from autonomic nervous system responses and facial expressions". In: *Physiology & behavior* 170 (2017), pp. 12–18.

[119] Nabanita Dutta, Umashankar Subramaniam, and Sanjeevikumar Padmanaban. "Mathematical models of classification algorithm of Machine learning". In: *International Meeting on Advanced Technologies in Energy and Electrical Engineering*. Hamad bin Khalifa University Press (HBKU Press). 2020, p. 3.

[120] Soumya Deep Roy, Soham Das, Devroop Kar, Friedhelm Schwenker, and Ram Sarkar. “Computer Aided Breast Cancer Detection Using Ensembling of Texture and Statistical Image Features”. In: *Sensors* 21.11 (2021), p. 3628.

[121] Helge Malmgren and Magnus Borga. *Artificial Neural Networks in Medicine and Biology: Proceedings of the ANNIMAB-1 Conference, Göteborg, Sweden, 13-16 May 2000*. Springer Science & Business Media, 2000.

[122] Stephen A Bustin. “Nucleic acid quantification and disease outcome prediction in colorectal cancer”. In: *Personalized medicine* (2006).

[123] Jigneshkumar L Patel and Ramesh K Goyal. “Applications of artificial neural networks in medical science”. In: *Current clinical pharmacology* 2.3 (2007), pp. 217–226.

[124] Corinna Peifer, Anita Pollak, Olaf Flak, Adrian Pyszka, Muhammad Adeel Nisar, Muhammad Tausif Irshad, Marcin Grzegorzek, Bastian Kordyaka, and Barbara Kożusznik. “The Symphony of Team Flow in Virtual Teams. Using Artificial Intelligence for Its Recognition and Promotion”. In: *Frontiers in Psychology* (2021), p. 3538.

[125] Weiming Hu, Haoyuan Chen, Wanli Liu, Xiaoyan Li, Hongzan Sun, Xinyu Huang, Marcin Grzegorzek, and Chen Li. “A comparative study of gastric histopathology sub-size image classification: From linear regression to visual transformer”. In: *Frontiers in Medicine* 9 (2022).

[126] Weiming Hu, Xintong Li, Chen Li, Rui Li, Tao Jiang, Hongzan Sun, Xinyu Huang, Marcin Grzegorzek, and Xiaoyan Li. “A state-of-the-art survey of artificial neural networks for Whole-slide Image analysis: From popular Convolutional Neural Networks to potential visual transformers”. In: *Computers in Biology and Medicine* 161 (2023), p. 107034.

[127] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. “Continual lifelong learning with neural networks: A review”. In: *Neural Networks* 113 (2019), pp. 54–71.

[128] Kevin T Sweeney, Tomás E Ward, and Seán F McLoone. “Artifact removal in physiological signals—Practices and possibilities”. In: *IEEE transactions on information technology in biomedicine* 16.3 (2012), pp. 488–500.

[129] Tian Lan, Deniz Erdogmus, Andre Adami, Misha Pavel, and Santosh Mathan. "Salient EEG channel selection in brain computer interfaces by mutual information maximization". In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE. 2006, pp. 7064–7067.

[130] Mihaly Csikszentmihalhi. *Finding flow: The psychology of engagement with everyday life*. Hachette UK, 2020.

[131] M Csikszentmihalyi. "Beyond boredom and anxiety. san francisco: Josseybass". In: *Well-being: Thefoundations of hedonic psychology* (1975), pp. 134–154.

[132] Stefan Engeser. "Theoretical integration and future lines of flow research". In: *Advances in flow research* (2012), pp. 187–199.

[133] Marta Bassi, Patrizia Steca, Dario Monzani, Andrea Greco, and Antonella Delle Fave. "Personality and optimal experience in adolescence: Implications for well-being and development". In: *Journal of Happiness Studies* 15 (2014), pp. 829–843.

[134] Corinna Peifer, Christine Syrek, Vivian Ostwald, Eva Schuh, and Conny H Antoni. "Thieves of flow: how unfinished tasks at work are related to flow experience and wellbeing". In: *Journal of Happiness Studies* 21 (2020), pp. 1641–1660.

[135] Roberta Maeran and Francesco Cangiano. "Flow experience and job characteristics: Analyzing the role of flow in job satisfaction". In: *TPM-Testing, Psychometrics, Methodology in Applied Psychology* 20.1 (2013), pp. 13–26.

[136] Corinna Peifer and Gina Wolters. "Flow in the Context of Work". In: *Advances in flow research*. Springer, 2021, pp. 287–321.

[137] Corinna Peifer and Jasmine Tan. "The psychophysiology of flow experience". In: *Advances in flow research*. Springer, 2021, pp. 191–230.

[138] Karina Nielsen and Bryan Cleal. "Predicting flow at work: Investigating the activities and job characteristics that predict flow states at work." In: *Journal of occupational health psychology* 15.2 (2010), p. 180.

[139] Yosuke Tezuka, Naho Murayama, Yosuke Morioka, and Naoto Suzuki. "The influence of answer to the self-report scale on cardiovascular recovery". In: *International Journal of Psychophysiology* 2.94 (2014), p. 246.

[140] Corinna Peifer, Annette Kluge, Nikol Rummel, and Dorothea Kolossa. “Fostering flow experience in HCI to enhance and allocate human energy”. In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 204–220.

[141] Dan Brickley, Matthew Burgess, and Natasha Noy. “Google Dataset Search: Building a search engine for datasets in an open Web ecosystem”. In: *The World Wide Web Conference*. 2019, pp. 1365–1375.

[142] Yuichi Fujiki, Konstantinos Kazakos, Colin Puri, Pradeep Buddharaju, Ioannis Pavlidis, and James Levine. “NEAT-o-Games: blending physical activity and fun in the daily routine”. In: *Computers in Entertainment (CIE)* 6.2 (2008), pp. 1–22.

[143] Marco Maier, Daniel Elsner, Chadly Marouane, Meike Zehnle, and Christoph Fuchs. “DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks.” In: *AAMAS*. 2019, pp. 2108–2110.

[144] Ling Shao, Ziyun Cai, Li Liu, and Ke Lu. “Performance evaluation of deep feature learning for RGB-D image/video classification”. In: *Information Sciences* 385 (2017), pp. 266–283.

[145] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2010), pp. 1345–1359.

[146] Corinna Peifer, André Schulz, Hartmut Schächinger, Nicola Baumann, and Conny H Antoni. “The relation of flow-experience and physiological arousal under stress—can u shape it?” In: *Journal of Experimental Social Psychology* 53 (2014), pp. 62–69.

[147] Yu Tian, Yulong Bian, Piguo Han, Peng Wang, Fengqiang Gao, and Yingmin Chen. “Physiological signal analysis for evaluating flow during playing of computer games of varying difficulty”. In: *Frontiers in psychology* 8 (2017), p. 1121.

[148] M Csikszentmihalyi, F Massimini, and M Carli. “The monitoring of optimal experience: A tool for psychiatric rehabilitation”. In: *Journal of Nervous and Mental Disease* 175 (1987), pp. 545–549.

[149] Carroll E Izard and Carroll E Izard. “Differential emotions theory”. In: *Human emotions* (1977), pp. 43–66.

[150] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. "Deap: A database for emotion analysis; using physiological signals". In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.

[151] Philip A Kragel and Kevin S LaBar. "Decoding the nature of emotion in the brain". In: *Trends in cognitive sciences* 20.6 (2016), pp. 444–455.

[152] Chang Li, Yimeng Hou, Rencheng Song, Juan Cheng, Yu Liu, and Xun Chen. "Multi-channel EEG-based emotion recognition in the presence of noisy labels". In: *Science China Information Sciences* 65.4 (2022), p. 140405.

[153] Michael T Knierim, Victor Pieper, Max Schemmer, Nico Loewe, and Pierluigi Reali. "Predicting In-Field Flow Experiences Over Two Weeks from ECG Data: A Case Study". In: *NeuroIS Retreat*. Springer. 2021, pp. 96–102.

[154] László Harmat, Örjan de Manzano, Töres Theorell, Lennart Högman, Håkan Fischer, and Fredrik Ullén. "Physiological correlates of the flow experience during computer game playing". In: *International Journal of Psychophysiology* 97.1 (2015), pp. 1–7.

[155] Raphael Rissler, Mario Nadj, Maximilian Xiling Li, Nico Loewe, Michael T Knierim, and Alexander Maedche. "To be or not to be in flow at work: physiological classification of flow using machine learning". In: *IEEE transactions on affective computing* (2020).

[156] Matthew Lee. "Detecting affective flow states of knowledge workers using physiological sensors". In: *arXiv preprint arXiv:2006.10635* (2020).

[157] Raphael Rissler, Mario Nadj, Maximilian Xiling Li, Michael Thomas Knierim, and Alexander Maedche. "Got flow? Using machine learning on physiological data to classify flow". In: *Extended abstracts of the 2018 CHI conference on human factors in computing systems*. 2018, pp. 1–6.

[158] Karen Bartholomeyczik, Michael Thomas Knierim, Petra Nieken, Julia Seitz, Fabio Stano, and Christof Weinhardt. "Flow in Knowledge Work: An Initial Evaluation of Flow Psychophysiology Across Three Cognitive Tasks". In: *Information Systems and Neuroscience: NeuroIS Retreat 2022*. Springer, 2022, pp. 23–33.

[159] *Statistica 12*. `https://www.statsoft.de/de/home`. [Online; accessed 31-Mar-2023].

[160] P Kumar Rahi, Rajesh Mehra, et al. “Analysis of power spectrum estimation using welch method for various window techniques”. In: *International Journal of Emerging Technologies and Engineering* 2.6 (2014), pp. 106–109.

[161] *Apps Lab*. https://www.imi.uni-luebeck.de/forschung/p44-apps-lab.html. [Online; accessed 13-Mar-2023].

[162] LK McEvoy, ME Smith, and A Gevins. “Test–retest reliability of cognitive EEG”. In: *Clinical Neurophysiology* 111.3 (2000), pp. 457–463.

[163] *DEAP: A dataset for emotion analysis using EEG, physiological, and video signals*. https://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html. [Online; accessed 22-Feb-2023].

[164] *Unipark*. https://www.unipark.com/. [Online; accessed 31-Mar-2023]. 2023.

[165] Boris Scavezzon. *Ein Album voller Kurzgeschichten*. Frankfurter Literaturverlag, 2010, pp. 49–58.

[166] *Emotiv Epoc X: Scalable and contextual human brain wear — providing access to professional-grade brain data with an improved and easy-to-use design*. https://www.emotiv.com/epoc-x/. [Online; accessed 25-Jan-2023]. 2023.

[167] *Electrooculography*. https://www.pluxbiosignals.com/collections/biosignalsplux/products/electrooculography-eog-sensor-1. [Online; accessed 25-Jan-2023].

[168] Tong Yu and Hong Zhu. “Hyper-parameter optimization: A review of algorithms and applications”. In: *arXiv preprint arXiv:2003.05689* (2020).

[169] James Bergstra and Yoshua Bengio. “Random search for hyper-parameter optimization.” In: *Journal of machine learning research* 13.2 (2012).

[170] Frédéric Li, Kimiaki Shirahama, Muhammad Adeel Nisar, Xinyu Huang, and Marcin Grzegorzek. “Deep transfer learning for time series data based on sensor modality classification”. In: *Sensors* 20.15 (2020), p. 4271.

[171] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. “Emotion recognition using multimodal residual LSTM network”. In: *Proceedings of the 27th ACM international conference on multimedia*. 2019, pp. 176–183.

[172] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. “A systematic survey on multimodal emotion recognition using learning algorithms”. In: *Intelligent Systems with Applications* 17 (2023), p. 200171.

[173] Tuan Nguyen-Sy, Jad Wakim, Quy-Dong To, Minh-Ngoc Vu, The-Duong Nguyen, and Thoi-Trung Nguyen. "Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method". In: *Construction and Building Materials* 260 (2020), p. 119757.

[174] Yan Zhao, Xing Chen, and Jun Yin. "Adaptive boosting-based computational model for predicting potential miRNA-disease associations". In: *Bioinformatics* 35.22 (2019), pp. 4730–4738.

[175] Mario Passalacqua, Raphaël Morin, Sylvain Sénécal, Lennart E Nacke, and Pierre-Majorique Léger. "Demystifying the First-Time Experience of Mobile Games: The Presence of a Tutorial Has a Positive Impact on Non-Expert Players' Flow and Continuous-Use Intentions". In: *Multimodal Technologies and Interaction* 4.3 (2020), p. 41.

[176] Denise M Bressler and Alec M Bodzin. "A mixed methods assessment of students' flow experiences during a mobile augmented reality science game". In: *Journal of computer assisted learning* 29.6 (2013), pp. 505–517.

[177] Edward J Donner and Mihaly Csikszentmihalyi. "Transforming stress to flow". In: *Executive Excellence* 9 (1992), pp. 16–16.

[178] Örjan De Manzano, Töres Theorell, László Harmat, and Fredrik Ullén. "The psychophysiology of flow during piano playing." In: *Emotion* 10.3 (2010), p. 301.

[179] J Matias Kivikangas et al. "Psychophysiology of flow experience: An explorative study". In: (2006).

[180] Lennart E Nacke and Craig A Lindley. "Affective ludology, flow and immersion in a first-person shooter: Measurement of player experience". In: *arXiv preprint arXiv:1004.0248* (2010).

[181] Fabian Pels and Jens Kleinert. "Perspectives on group flow: Existing theoretical approaches and the development of the integrative group flow theory." In: *Group Dynamics: Theory, Research, and Practice* (2022).

[182] RG Brackenridge. "Health and the hospital". In: *Essential Medicine*. Springer, 1971, pp. 1–4.

[183] Michael William Ainscoe. *Flow: the concept and implications for mental well-being and health*. Bangor University (United Kingdom), 1989.

[184] Eleonora Riva, Teresa Freire, and Marta Bassi. “The flow experience in clinical settings: Applications in psychotherapy and mental health rehabilitation”. In: *Flow experience: Empirical research and applications* (2016), pp. 309–326.

[185] Barbara Burlingame. “Fostering quality data in food composition databases: visions for the future”. In: *Journal of Food Composition and Analysis* 17.3-4 (2004), pp. 251–258.

[186] Lennart Jablonski, Torge Jensen, Greta M Ahlemann, Xinyu Huang, Vivian V Tetzlaff-Lelleck, Artur Piet, Franziska Schmelter, Valerie S Dinkler, Christian Sina, and Marcin Grzegorzek. “Sensor-Based Detection of Food Hypersensitivity Using Machine Learning”. In: *Proceedings of the 8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*. 2023, pp. 1–8.

[187] Ms Aayushi Bansal, Dr Rewa Sharma, and Dr Mamta Kathuria. “A systematic review on data scarcity problem in deep learning: solution and applications”. In: *ACM Computing Surveys (CSUR)* 54.10s (2022), pp. 1–29.

[188] Muhammad Adeel Nisar. *Sensor-based human activity recognition for assistive health technologies*. Logos Verlag Berlin, Nov. 2023.

[189] Pawel Poryzala and Andrzej Materka. “Cluster analysis of CCA coefficients for robust detection of the asynchronous SSVEPs in brain–computer interfaces”. In: *Biomedical Signal Processing and Control* 10 (2014), pp. 201–208.

[190] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[191] Xinyu Huang, Franziska Schmelter, Muhammad Tausif Irshad, Artur Piet, Muhammad Adeel Nisar, Christian Sina, and Marcin Grzegorzek. “Optimizing sleep staging on multimodal time series: Leveraging borderline synthetic minority oversampling technique and supervised convolutional contrastive learning”. In: *Computers in Biology and Medicine* 166 (2023), p. 107501.

[192] Huang Xinyu. *Sensor-Based Sleep Stage Classification Using Deep Learning*. Logos Verlag Berlin, 2023.

[193] Reza Rawassizadeh, Elaheh Momeni, Chelsea Dobbins, Pejman Mirza-Babaei, and Ramin Rahnamoun. “Lesson learned from collecting quantified self information via mobile and wearable devices”. In: *Journal of Sensor and Actuator Networks* 4.4 (2015), pp. 315–335.

[194] Hannah R Nolasco, Andrew Vargo, Niklas Bohley, Christian Brinkhaus, and Koichi Kise. “Examining Participant Adherence with Wearables in an In-the-Wild Setting”. In: *Sensors* 23.14 (2023), p. 6479.

[195] Vincenzo Ronca, Ana C Martinez-Levy, Alessia Vozzi, Andrea Giorgi, Pietro Aricò, Rossella Capotorto, Gianluca Borghini, Fabio Babiloni, and Gianluca Di Flumeri. “Wearable Technologies for Electrodermal and Cardiac Activity measurements: A Comparison between Fitbit Sense, Empatica E4 and Shimmer GSR3+”. In: *Sensors* 23.13 (2023), p. 5847.

[196] Sulaiman Girivirya. “Analysis of Mindfulness Practices Using Electroencephalogram (EEG) Interaxon Muse™ Headband Against the Concept of Self-Acceptance of Poststroke Clients”. In: *INFLUENCE: International Journal of Science Review* 5.2 (Mar. 2023), pp. 11–19.

[197] Lichy Han, David R. Drover, Marianne C. Chen, Amit R. Saxena, Sarah L. Eagleman, Vladimir Nekhendzy, Angelica Pritchard, and Robson Capasso. “EEG response of dexmedetomidine during drug induced sleep endoscopy”. In: *Frontiers in Neuroscience* 17 (2023).

List of Own Publications

[1] **M. T. Irshad**, M. A. Nisar, P. Gouverneur, M. Rapp, and M. Grzegorzek, “Ai approaches towards prechtl’s assessment of general movements: A systematic literature review,” Sensors, vol. 20, no. 18, p. 5321, 2020.

[2] C. Peifer, A. Pollak, O. Flak, A. Pyszka, M. A. Nisar, **M. T. Irshad**, M. Grzegorzek, B. Kordyaka, and B. Kożusznik, “The symphony of team flow in virtual teams. Using artificial intelligence for its recognition and promotion,” Frontiers in Psychology, p. 3538, 2021.

[3] **M. T. Irshad**, M. A. Nisar, X. Huang, J. Hartz, O. Flak, F. Li, P. Gouverneur, A. Piet, K. M. Oltmanns, and M. Grzegorzek, “Sensehunger: Machine learning approach to hunger detection using wearable sensors,” Sensors, vol. 22, no. 20, p. 7711, 2022.

[4] E. Bösemann, **M. T. Irshad**, H. Fischer, and M. Grzegorzek, "Evaluation of the usability of a ventilation test framework," in Proceedings of the Student Conference of the Hanse Innovation Campus Lübeck, Germany, March 01-03, 2023.

[5] X. Huang, K. Shirahama, **M. T. Irshad**, M. A. Nisar, A. Piet, and M. Grzegorzek, “Sleep stage classification in children using self-attention and gaussian noise data augmentation,” Sensors, vol. 23, no. 7, p. 3446, 2023.

[6] L. Kloep, M. Buss, M. Grzegorzek, **M. T. Irshad**, P. Gouverneur, B. Kożusznik, A. Pollak and C. Peifer, "A computational approach to understand social flow and its role in interpersonal relationships in virtual teams – project outline and first results from a pilot study," in 21th Annual Congress of the European Association of Work and Organizational Psychology (EAWOP), Katowice, Poland, May 24-27, 2023.

[7] R. Doniec, J. Konior, S. Siecinski, A. Piet, **M. T. Irshad**, N. Piaseczna, M. A. Hasan, F. Li, M. A. Nisar, and M. Grzegorzek, “Sensor-based classification of primary and

secondary car driver activities using convolutional neural networks," Sensors, vol. 23, no. 12, p. 5551, 2023.

[8] S. Siecinski, **M. T. Irshad**, M. A. Hasan, E. J. Tkacz, P. S. Kostka, and M. Grzegorzek, "Symmetric projection attractor reconstruction analysis as a method to assess seismocardiogram quality in a healthy population," in 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, July 24-27, 2023.

[9] **M. T. Irshad**, F. Li, M. A. Nisar, X. Huang, M. Buss, L. Kloep, C. Peifer, B. Kożusznik, and M. Grzegorzek, "Wearable-based human flow experience recognition enhanced by transfer learning methods using emotion data," Computers in Biology and Medicine, p. 107 489, 2023.

[10] X. Huang, F. Schmelter, **M. T. Irshad**, A. Piet, M. A. Nisar, C. Sina, and M. Grzegorzek, "Optimizing sleep staging on multimodal time series: Leveraging borderline synthetic minority oversampling technique and supervised convolutional contrastive learning," Computers in Biology and Medicine, p. 107 501, 2023.

[11] M. A. Hasan, F. Li, A. Piet, P. Gouverneur, **M. T. Irshad**, and M. Grzegorzek, "Exploring the benefits of time series data augmentation for wearable human activity recognition," in 8th International Workshop on Sensor-based Activity Recognition and Artificial Intelligence (iWOAR), Lübeck, Germany, September 21-21, 2023.

[12] S. Sieciński, **M. T. Irshad**, M. A. Hasan, E. J. Tkacz, and M. Grzegorzek, "Assessment of quality of gyrocardiograms based on features derived from symmetric projection attractor reconstruction," in 8th International Workshop on Sensor-based Activity Recognition and Artificial Intelligence (iWOAR), Lübeck, Germany, September 21-21, 2023.

[13] M. A. Nisar, K. Shirahama, **M. T. Irshad**, X. Huang, and M. Grzegorzek, "A hierarchical multitask learning approach for the recognition of activities of daily living using data from wearable sensors," Sensors, vol. 23, no. 19, p. 8234, 2023.

List of Abbreviations

PRP ..... Pattern Recognition Pipeline
HCFs ..... Hand-Crafted Features
ANNs ..... Artificial Neural Networks
MLPs ..... Multilayer Perceptrons
CNNs ..... Convolutional Neural Networks
RNNs ..... Recurrent Neural Networks
ADAM ..... Adaptive Moment Estimation
NB ..... Naïve Bayes
SVM ..... Support Vector Machine
DT ..... Decision Tree
RF ..... Random Forest
XGBoost ..... eXtreme Gradient Boosting
AdaBoost ..... Adaptive Boosting
SWS ..... Sliding Window Segmentation
LOSO ..... Leave One Subject Out
SKF-CV ..... Stratified-K-Fold - Cross Validation
FIR ..... Feature Importance Ranking
EEG ..... Electroencephalography
EMG ..... Electromyography
EOG ..... Electrooculography
ECG ..... Electrocardiography
BVP ..... Blood Volume Pulse
EDA ..... Electrodermal activity
GSR ..... Galvanic Skin Response
HRV ..... Heart Rate Variability
RB ..... Respiration belt
ReLU ..... Rectified Linear Units
LeakyReLU ..... Leaky Rectified Linear Units

GMs . . . . . General Movements
PhySH . . . . . Physiological Sense Hunger (dataset)
PhySF . . . . . Physiological Sense Flow (dataset)
DEAP . . . . . Database for Emotion Analysis using Physiological Signals
VAS . . . . . Visual Analog Scale
CT . . . . . Computed Tomography
HAR . . . . . Human Activity Recognition
PTSD . . . . . Post-Traumatic Stress Disorder

List of Figures

1.1 Pattern Recognition Pipeline (PRP). 7

1.2 Graph of an artificial neuron. 12

1.3 Outline of the thesis. 24

2.1 Overview of the RespiBAN wearable, including highlighting functional areas. 35

2.2 Overview of the Empatica E4 wristband, including highlighting functional areas. 36

2.3 Overview of the JINS MEME smart glasses, including highlighting functional areas. 36

2.4 Illustration of the Multilayer Perceptron (MLP). 46

2.5 Illustration of the Convolutional Neural Network (CNN). 47

2.6 Performance comparison of the SVM, DT, and RF classifiers using different window sizes and step sizes. 49

2.7 Hyper-parameters selection for the implemented feature learning approaches (MLP and CNN). 50

2.8 Performance comparison of each sensor channel in recognizing hunger and satiety states. 52

2.9 Performance comparison of wearable devices in recognizing hunger and satiety states. 55

2.10 List of confusion matrices. 57

2.11 List the 18 best features selected to classify hunger and satiety sensations accurately. 59

3.1 List of wearable devices utilized to acquire the physiological data for human flow experience recognition. 73

3.2 The distribution of each subject's flow and non-flow class data in the PhySF dataset. 74
3.3 The principle of the author's implemented CNN-based transfer learning approach. 79
3.4 Performance comparison for the selection of optimized Sliding Window Segmentation (SWS) parameters. 81
3.5 Confusion matrices of the approaches carried out for the flow and non-flow state recognition. 88
3.6 Performance comparison of wearable sensor devices in recognizing flow and non-flow states. 89
3.7 Relationship between flow experience and physiological arousal. 91

List of Tables

2.1 List the related state-of-the-art literature on hunger and satiety state detection. . . . 31
2.2 Demographic information of the subjects of the PhySH dataset 35
2.3 List of Hand-Crafted Features (HCFs) used for hunger and satiety state recognition. . . . 40
2.4 The MLP architecture and hyper-parameter values for hunger and satiety state detection. . . . 47
2.5 The CNN architecture and hyper-parameter values for hunger and satiety state detection. . . . 48
2.6 Results of hunger and satiety state classification with different sliding window sizes and step sizes. . . . 51
2.7 Results of each wearable sensor separately for hunger and satiety state classification using an RF classifier. . . . 53
2.8 Results of the classification of hunger and satiety states based on the best-selected features using the RF feature importance ranking. 53
2.9 Comparison of feature selection algorithms (Boruta, XGBoost, and RF) for the classification of hunger and satiety states. . . . 54
2.10 Results of hunger and satiety state recognition using feature learning (MLP and CNN). . . . 55
2.11 Comparison of hunger and satiety state recognition results with related peer-reviewed literature. . . . 56

3.1 List the related state-of-the-art literature on human flow experience recognition. . . . 67
3.2 Demographic information of all the subjects of the PhySF dataset. 72
3.3 List of questions asked to the study participants to determine the labels of the PhySF dataset. . . . 74

3.4 Selected sensor channels in both datasets for the proposed transfer learning-based approach. 76
3.5 The MLP architecture and hyper-parameter values for flow and non-flow state classification. 78
3.6 The CNN architecture and hyper-parameter values for flow and non-flow state classification. 78
3.7 Distribution of PhySF dataset subjects into 5 different folds of SKF-CV and flow and non-flow state data of each subject in seconds. 82
3.8 Results of flow and non-flow state recognition employing manual feature engineering. 84
3.9 Results of flow and non-flow state recognition using MLP-based feature learning. 85
3.10 Results of flow and non-flow state recognition employing CNN-based feature learning. 85
3.11 Results of flow and non-flow state recognition using CNN-based transfer learning. 86
3.12 Comparison of flow and non-flow state recognition results with state-of-the-art literature studies. 87