MACHINE LEARNING APPROACHES IN PRACTICAL ANXIETY DETECTION

BY

ABDULRAHMAN E. ALKURDI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mechanical Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2024

Urbana, Illinois

Doctoral Committee:

Professor Elizabeth T. Hsiao-Wecksler, Chair and Director of Research
Teaching Associate Professor Manuel E. Hernandez, Co-Director of Research
Professor Richard Sowers
Associate Professor Chenhui Shao

ABSTRACT

This dissertation investigates the development and application of machine learning (ML) techniques for the robust detection of anxiety using wearable technology under diverse environmental conditions. The research encompasses three interlinked studies, each contributing uniquely towards advancing anxiety detection technologies in real-world settings. The first study conducts a thorough review of existing ML methodologies for anxiety detection, identifying the evolution from traditional feature-based (FB) models to advanced end-to-end (E2E) deep learning approaches. It evaluates their applicability across different scenarios, highlighting the challenges of integrating such technologies in practical applications due to issues like noise interference and model overfitting. In the second study, the focus shifts to the practical implementation of these models in noisy environments. It explores the resilience of both FB and E2E models by introducing artificial noise into the WESAD dataset and examining their performance. This part of the research emphasizes the critical impact of environmental disturbances on model accuracy, particularly in wearable technologies, and demonstrates the superior noise resistance of FB models compared to E2E models. The third study extends this analysis by applying transfer learning techniques to adapt these models to real-world datasets, specifically the RADWear and WEAR datasets, which reflect a broad spectrum of real-life conditions. The study assesses the efficacy of transfer learning in enhancing model robustness and addresses the challenges of deploying these technologies in dynamic and uncontrolled environments. Despite the high potential of transfer learning, the results reveal that E2E models consistently underperform in comparison to FB models, which display greater adaptability and reliability under varied environmental conditions. Overall, this dissertation highlights the complexities of developing effective ML-based anxiety detection systems that are capable of operating in real-world scenarios. It underscores the need for further research into optimizing model architectures, improving noise management strategies, and refining data collection techniques to enhance the practicality and effectiveness of anxiety detection tools. The insights gained from these studies pave the way for future advancements in wearable technology for mental health, offering a foundation for more personalized and responsive mental health care solutions.

ACKNOWLEDGMENTS

To all those without whom I would not be here. To co-pilot and partner-in-crime, my wife, Samour. To our derivatives, kids, Sulaiman and Aya. To my parents. To my brothers and sister: Makki the elder, Abdulelah the Gex, and Maria the second smartest one in the family. To the Alfaris brothers for their continued guidance.

To Liz, my strong advisor, for being a kind and accommodating soul. The lessons I learned from her extend beyond the academic and will last longer in my life and career than anything that appears in this dissertation. To Manuel Hernandez, my co-director and project PI, whose insights and guidance have been invaluable in shaping the direction and success of my research. To Rich Sowers, co-PI, for his expert advice and steadfast support through the complexities of my research. To Chenhui Shao, a distinguished member of my dissertation committee, whose expertise has significantly enriched my work.

It would not be complete without thanking my knowledgeable and supportive HDCL colleagues who have been patient to my long ramblings: Mahshid Mansouri, Maxine He, Nick Thompson, Kevin Gim, Prateek Garag, Chenzhang Xiao, Yinan Pei, Yu Chen, Nadja Marin, Keona Banks, Ezekiel Hsieh, and Yixiao Liu. I cannot thank you enough for the things I learned from you and for the help you gave me. I would also like to thank the numerous undergraduate students who helped me throughout my PhD projects.

To my teachers, colleagues, mentors, and everyone whose shoulders contributed to where I stand today and in the future. In the profound words of Dr. Cornel West, "I am who I am because somebody loved me, somebody attended to me, somebody focused on me."

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1. Background and Motivation for Studying Anxiety

1.1.1 Introduction to Affective Computing

Before delving into the nuances of stress and anxiety, it is imperative to introduce the field of affective computing. Affective computing is an interdisciplinary domain that merges insights from psychology, neuroscience, and computer science to study, understand, and respond to human emotions [1]. At its core, affective computing involves the development of systems and devices that can recognize, interpret, and process human affects, wherein ‘affect’ refers to the complex spectrum of subjective feelings and emotions [2]. In the context of this thesis, affective computing is pivotal as it provides the technological and methodological framework for detecting and analyzing emotional states, such as anxiety and stress, using computational models and algorithms.

1.1.2 Stress and Anxiety (Definitions)

In the realm of anxiety theory, the concepts of stress and anxiety are intricately linked, necessitating a clear delineation of their definitions. In the context of anxiety theory, stress and anxiety are interrelated, and it is appropriate to start with their definitions. Stress is typically characterized as an adverse stimulus that triggers physiological and psychological responses [3]. Delving into Spielberger's Trait-State anxiety theory, anxiety is factored into state and trait forms [3]–[5]. State anxiety is described as a temporary and transient negative emotional reaction to stress, marked by heightened activity in the sympathetic nervous system [5]. In contrast, trait anxiety denotes an individual's inherent tendency to experience state anxiety when encountering stress. For the purposes of this work, the focus is primarily on the detection of state anxiety, herein referred to simply as anxiety. This delineation is based on the use of standard, validated methods for evaluating or inducing anxiety in the literature; in their absence, we categorize the affective state as stress.

1.1.3 Prevalence and Impact of Anxiety

The pervasive impact of anxiety on long-term health and job performance underscores the urgency for developing diagnostic and therapeutic tools. Such tools not only promise to enhance access to diagnosis and treatment but also serve as invaluable resources for mental health professionals, whose expertise is increasingly sought after [6]–[9]. Additionally, there is a pressing need for systems capable of monitoring the workload in individuals engaged in high-stress occupations or activities [10], [11].

Prolonged exposure to anxiety has been demonstrably linked to a decline in quality of life and long-term health detriments, notably escalating the risk of cardiovascular diseases [12], [13] and weakening immune defenses [14]. The correlation between anxiety and reduced performance efficiency is well-established, with numerous studies correlating anxiety with impaired decision-making [15], [16], diminished situational awareness, and compromised cognitive performance [17]–[19]. These impairments are particularly critical in professions requiring heightened mental or physical alertness, such as driving, firefighting, and aviation.

In a broader context, mental health disorders, particularly anxiety, have emerged as one of the foremost global health challenges [20]. Data from the American Psychiatric Association reveal that about 18% of Americans grapple with anxiety disorders [21]. with similar prevalence rates observed in Europe [22], [23]. Globally, the World Health Organization highlights that depression and anxiety contribute to an estimated annual productivity loss of US$ 1 trillion [24]. Despite the substantial impact of anxiety on health and productivity, access to mental health services remains alarmingly limited. A 2009 study revealed that over three-quarters (77%) of U.S. counties face a severe shortage of mental health professionals, leaving more than half of the population's mental health needs unmet [7].

1.1.4 Measuring Anxiety

Anxiety assessment traditionally rely on methods like self-reported questionnaires, which, despite their widespread use, are constrained by recall bias and an inability to track real-time changes across diverse contexts [25]–[27]. Recall bias, a systematic error stemming from

inaccuracies in participants' recollection of past events, significantly limits the effectiveness of retrospective approaches in dynamically characterizing and understanding behavior [28].

There has been significant research aimed at enhancing the understanding of anxiety manifestations and developing quantitative tools for its assessment. This research predominantly has focused on the autonomic nervous system's role, hormonal changes [29]–[33], and distinct patterns of subjective and physiological responses [34]–[36]. The autonomic nervous system, which regulates involuntary physiological processes like respiration and digestion, consists of three components: the sympathetic, parasympathetic, and enteric systems [37]. The sympathetic nervous system, in particular, is crucial for the fight-or-flight response, triggering hormonal, physiological, and behavioral changes in response to adverse stimuli [25] [37]. Anxiety detection has been explored through various markers such as brain activation patterns [38]–[40], cardiac responses [41], [42], muscular activation [32], electrodermal activity (EDA) [43], and eye-related activity [44], all linked to the sympathetic nervous system's dynamic activation [34]–[36]. Additionally, behavioral changes, including alterations in speech, facial expressions, body motion, and head movement, have been observed as indicators of anxiety [45].

1.1.4.1. Traditional Questionnaires

Common tools like the State-Trait Anxiety Inventory (STAI) [46] or the Beck Anxiety Inventory (BAI) [47] involve standardized questionnaires that measure subjective experiences of anxiety [46][47]. While these tools are easy to administer and score, they are susceptible to biases, particularly in individuals uncomfortable with expressing their emotions [48].

Clinical interviews, conducted by trained professionals, offer a more structured approach to diagnosing anxiety disorders. These interviews, often structured or semi-structured, provide a more accurate assessment than self-reports but are time-intensive and costly [49].

Behavioral observations, another method, rely on analyzing observable indicators of anxiety, such as facial expressions and body language [50]. These observations, made by trained individuals or via video recordings, are particularly useful for individuals unable to complete questionnaires or uncomfortable discussing their emotions. However, the subjectivity of these

observations and their potential inaccuracy in individual's adept at masking anxiety present significant limitations [51].

Each method presents unique advantages and challenges: questionnaires are user-friendly but prone to bias; clinical interviews are accurate but resource-intensive; and behavioral observations are helpful for non-verbal assessments but are subjective and potentially misleading [51].

1.1.4.2. Biophysiological Measures

The multifaceted nature of anxiety necessitates employing a range of measurement techniques, each offering unique insights into an individual's affective state. This section focuses on biophysiological measures, examining the utilization of brain activation patterns and other physiological indicators to capture the nuanced manifestations of anxiety.

1.1.4.2.1. Brain Activity

Brain activity, as a biophysiological measure, plays a crucial role in anxiety detection. Methods such as electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), and Positron Emission Tomography (PET) are employed to sense and analyze brain activity, particularly focusing on the sympathetic nervous system's activation patterns.

a. Electroencephalogram (EEG)

EEG, known for its excellent temporal resolution, enables precise tracking of neuronal activity fluctuations, offering valuable insights into stress and anxiety. However, it is limited by its trade-off between temporal and spatial resolution, susceptibility to noise, and its predominantly lab-based application due to bulkiness [52]–[54]. Studies have utilized EEG in various capacities, from analyzing frontal asymmetry during high stress to applying machine learning (ML) techniques for stress detection with notable accuracy [55]–[57]. Nonetheless, challenges include difficulties in localizing brain sources responsible for induced activities and sensitivity to artifacts during data collection.

b. Functional Near-Infrared Spectroscopy (fNIRS)

fNIRS, leverages the absorption spectrum of hemoglobin, offers high spatial resolution and is more resistant to motion artifacts compared to EEG. It can be used alongside other neuroimaging techniques without electromagnetic interference. Despite these advantages, the lack of standardized data analysis guidelines for fNIRS can lead to inconsistent results [55], [56]. Research employing fNIRS has demonstrated its efficacy in stress detection, particularly in assessing hemodynamic responses in the prefrontal cortex during stress-inducing tasks [58], [59].

c. Positron emission tomography (PET)

PET, utilizing nuclear medicine, provides detailed functional imaging with high spatial resolution. It has been instrumental in examining neurochemical changes under psychological stress. Despite its advantages, PET's application in stress detection is limited due to high costs, lack of portability, and low temporal resolution, making it less suitable for real-world applications [59].

1.1.4.2.2. Cardiac Activity

Cardiac activity, as a non-invasive measure, plays a crucial role in detecting anxiety, reflecting the interaction between physiological states and the sympathetic nervous system. The concept of heart rate variability (HRV) has evolved significantly since its inception in the mid-20th century, establishing itself as a robust indicator of anxiety [60]–[62]. The cardiac cycle, sensitive to activations of the sympathetic nervous system, has provided intricate insights into anxiety manifestations [63], [64]. A notable relationship exists between HRV and heart rate (HR): an increase in anxiety-induced HR typically corresponds with a decrease in HRV [65]–[68]. HRV is commonly defined mathematically as the standard deviation of NN-intervals (SDNN) [69].

a. Electrocardiogram (ECG) and Photoplethysmography (PPG)

With technological advancements, cardiac activity analysis has progressed from conventional ECG to more practical, wearable technologies like PPG. PPG, using infrared light to measure blood volume pulse (BVP), has become increasingly accessible for real-life

applications, despite its susceptibility to motion artifacts. Innovations such as analyzing HRV in the frequency domain have enhanced the accuracy of anxiety detection [70]. Additionally, remote cardiac detection technologies like facial thermography and radio-wave estimations represent a significant shift towards versatile, non-contact methods in cardiac activity analysis.

b. Facial Thermography

Recent developments in non-contact telemetry methods, including facial thermography, radio-wave heart rate estimation, and ballistocardiography, have furthered the potential of remote cardiac activity detection [71]–[73]. Facial thermography, for instance, has been utilized to detect mental workload through changes in blood flow [71], [74].

1.1.4.2.3. Electromyography (EMG)

Anxiety detection can also be accomplished through various peripheral biophysiological measures. EMG is used to measure muscle tension and activation patterns, including those of the trapezius muscle, with a noted relationship between anxiety levels and EMG activity [32], [75], [84], [76]–[83]. Respiration patterns linked to autonomic nervous system activity have also been utilized in anxiety detection methods [85]–[90].

1.1.4.2.4. Respiration

Respiration parameters, closely linked with the autonomic nervous system, have become integral to stress detection. Changes in respiration patterns, such as respiratory rate and tidal volume, are indicative of variations in sympathetic and parasympathetic activity, essential components of the physiological stress response [85], [86]. Novel techniques, like using bioradar for respiratory signal capture, have shown promising results in stress detection. For example, the application of Recurrence Quantification Analysis (RQA) on respiratory patterns achieved a high precision in identifying mental stress [88]. Similarly, a framework utilizing a KINECT sensor for non-contact stress detection based solely on respiratory signals demonstrated significant accuracy in distinguishing psychological and physical stress states [89]. Research by Seo et al. highlighted the effectiveness of end-to-end neural networks over conventional ML models in analyzing ECG and respiration signals for stress detection [91].

1.1.4.2.5. Electrodermal Activity (EDA), Acceleration and other measures

Electrodermal Activity (EDA), also known as skin conductance, reflects the skin's ability to conduct electricity, which varies with its moisture level. This measure is highly sensitive to emotional arousal, particularly anxiety, because it is directly influenced by the sympathetic nervous system. EDA monitoring allows for the detection of subtle physiological changes associated with stress and anxiety responses, providing real-time insights into emotional states. The utility of EDA in machine learning applications for anxiety detection is well-documented, as it offers a quantifiable and responsive indicator of psychological arousal that can be continuously and non-invasively assessed [92]–[94].

Anxiety detection can also be accomplished through various peripheral biophysiological measures. Acceleration (ACC) signals derived from behavioral proxies to major biophysiological activities were also utilized for anxiety detection. Skin temperature changes, linked to physiological changes under anxiety, are used in ML algorithms for anxiety detection [95]–[99]. Lastly, eye pupil dilation, regulated by the autonomic system, serves as a potential biomarker for anxiety responses [95], [100]–[103]. Pupil diameter (PD) [100]–[104], voice intonation [105] and body pose [106]were observed, which, unlike measuring stress hormones, can be acquired through non-invasive means.

1.1.4.3. Behavioral measures

Recent advancements in anxiety detection have emphasized behavioral approaches due to their capacity to unveil hidden or subconscious cognitive and affective states through observable physical and interactive markers [107]–[110]). These methods tap into the intricate relationship between complex sensorimotor behavior and cognition, using an individual's movements and physical interactions as proxies for affective states like anxiety.

Researchers have explored a variety of behavioral methodologies to understand anxiety's manifestations. These include analyzing physical activity [111], body pose [105], keyboard strokes [112], voice characteristics [113], and head movements [114]. For instance, Schindler et al. identified characteristic body postures associated with anxiety in a virtual reality environment [106], while Vizer et al. observed changes in keyboard stroke patterns during anxiety-inducing

tasks [115]. Furthermore, Scherer et al. reported alterations in voice characteristics under anxiety [116], and Giannakakis et al. noted increased head movements and erratic gaze patterns in anxious individuals during a driving simulation [109].

Despite their potential, behavioral methods face challenges, particularly in adapting to individual differences and real-world settings. Nonetheless, they hold significant promise for early identification and intervention of anxiety, potentially preventing the progression of more severe mental health issues.

1.1.5 Anxiety-Inducing Experimental Methods

Inducing anxiety in experimental settings is vital for studying physiological and psychological responses. Ground truth metrics in these studies often rely on self-reported measures and expert evaluations.

1.1.5.1. Methods to Induce Anxiety in the Laboratory Environment

Anxiety can be induced in laboratory settings through cognitive tasks, emotional manipulation, and physiological manipulation. Cognitive tasks like the Trier Social Stress Test (TSST) [117] and the Stroop Color Word Test (SCWT) [118] are designed to elicit anxiety responses. Emotional manipulation, such as exposing participants to disturbing images or anticipatory anxiety through fear conditioning, is another method [118]. Physiological manipulation, like the cold pressor test, induces physical discomfort and stress [119]. The experimental environment for anxiety detection research is chosen based on the research goals and target population. Laboratory settings offer controlled conditions for manipulating variables and monitoring responses, ideal for testing new anxiety detection methods. In contrast, real-world settings capture anxiety in naturalistic environments like homes, workplaces, or public spaces, assessing the feasibility and effectiveness of detection methods in everyday contexts.

1.1.5.2. Experiment Environments

The experimental environment for anxiety detection research is chosen based on the research goals and target population. Laboratory settings offer controlled conditions for manipulating variables and monitoring responses, ideal for testing new anxiety detection

methods. In contrast, real-world settings capture anxiety in naturalistic environments like homes, workplaces, or public spaces, assessing the feasibility and effectiveness of detection methods in everyday contexts.

Hybrid environments blend laboratory and real-world settings, providing controlled yet realistic contexts. Virtual reality (VR) environments are increasingly utilized in such research, offering immersive situations that induce anxiety-like experiences while allowing detailed monitoring of participant responses. For instance, a study employed VR for biofeedback interventions in anxiety treatment [120].

Anxiety detection in laboratory conditions often achieve high accuracy (95% or greater) [54], [121]–[129]. However, performances typically reduce in real-world settings with environmental disturbances, leading to lower accuracy rates 47%-77% [130]–[137].

This variance in performance underscores a critical challenge in the field of anxiety detection: the translation of laboratory-based findings to real-world applications. The high accuracy rates achieved in the lab set a benchmark for the potential of anxiety detection methodologies. However, the drop in accuracy in naturalistic settings highlights the complexities and unpredictability inherent in real-world environments. This gap between lab and life emphasizes the need for developing and testing anxiety detection methods that are robust against environmental disturbances and adaptable to the varied conditions of everyday life. As such, the future of anxiety detection research lies not only in the development of sophisticated detection methods but also in ensuring their applicability and reliability in the dynamic and often unpredictable real-world scenarios where they are most needed.

1.1.6 Available Datasets for Anxiety Detection

1.1.6.1. Stress Recognition in Automobile Drivers (SRAD) in Physionet

The Stress Recognition in Automobile Drivers (SRAD) dataset, described in Healey and Picard in 2005 [62], and found in the PhysioNet database [138], is a seminal collection of biophysiological measures aimed at understanding stress levels of drivers in real-world conditions. As one of the earliest databases in this domain, SRAD offers invaluable insights into

human responses to driving-related stressors. It includes a comprehensive array of physiological signals, encompassing cardiovascular and respiratory patterns, among others.

The dataset is notable for its rich variety of signals, including cardiovascular patterns such as ECG, Blood Pressure (BP), and Heart Rate (HR), as well as EDA, which is sensitive to emotional arousal. Additionally, respiratory patterns were recorded, providing insights into breathing rates and signs of stress or relaxation. Ten participants, equipped with wearable sensors, contributed to this dataset. The ECG was captured using a multi-lead system, and a photoplethysmogram (PPG) sensor was employed to measure blood volume changes for deriving heart rate and blood pressure. EDA data were obtained using electrodes placed on the fingers, while respiratory patterns were monitored using chest and abdomen bands.

Participants' driving sessions varied in duration, ranging from 30 minutes to a couple of hours, ensuring a representative data set that included both short-duration stressors, like car horns, and long-duration stressors, such as traffic congestion. The dataset was collected under various driving conditions, including city and highway driving, as well as scenarios with intentional stressors like sudden braking or unexpected obstacles, to understand how these challenges impact physiological markers of stress.

In summary, the SRAD dataset for stress while driving offers a comprehensive set of physiological signals captured in real-world driving conditions. It provides critical insights into the impact of everyday driving challenges on the body's stress markers and holds potential for informing safer driving practices and interventions.

1.1.6.2. Wearable Stress and Affect Detection (WESAD)

The Wearable Stress and Affect Detection (WESAD) dataset is a more recent dataset published by Schmidt et al. in 2018 [139]. It is distinguished by its multimodal nature, encompassing data on three affective states: baseline, amusement, and stress. This coverage enhances the dataset's utility in developing models that are more accurate and generalizable in complex situations.

WESAD was collected from 15 participants wearing wrist and chest devices, capturing EDA, HR, BVP, ECG, respiration, and skin temperature. The study design included various conditions like baseline, amusement (watching humorous clips), and stress (exposed to the TSST), followed by guided meditation periods between test conditions to neutralize affective states. The comprehensive data, combined with self-reported stress levels, make WESAD a valuable resource for researchers developing and benchmarking anxiety detection methods using wearable devices.

Since its publication, WESAD has been widely cited and used in research, reflecting its significant contribution to the field. Notably, Dzieżyc et al. demonstrated the feasibility of using deep learning models for anxiety detection using the WESAD dataset [140], highlighting its practical applicability in advancing affective computing research. The dataset can be obtained from Ubiquitous Computing research group's website, part of the School for Science and Technology at the University of Siegen in Germany [139].

1.1.6.3. The Anxiety Dataset

The Anxiety Dataset, published by Elgendi et al. in 2022 [141], uniquely combines psychological self-reports with physiological markers, offering a holistic view of anxiety. It includes ECG and respiration signals, collected using a Biopac system (Goleta, CA, USA) to ensure a naturalistic data collection environment. Participants were exposed to anxiety-inducing video clips and completed the BAI [47], and HAM-A [142], questionnaires to provide a comprehensive perspective on their anxiety levels. The study's participant pool of 19 individuals, diverse in age and cultural background, enriches the dataset's applicability across different demographics. This integration of psychological and physiological data made the dataset a significant resource for researchers focusing on anxiety assessment and management. The dataset can be obtained here: https://figshare.com/articles/dataset/Anxiety_Dataset_2022/19875217.

1.1.6.4. University of Burgundy Franche-Comté Physiological studies dataset (UBFC-Phys)

The University of Burgundy Franche-Comté Physiological studies (UBFC-Phys) dataset by Sabour and colleagues in 2023 [143] offers a multimodal approach to psychophysiological

studies. It was collected from on 56 participants (46 females and 10 males) aged 19-38. Each participant underwent an experiment with a rigorous protocol inspired by the well-known TSST and was conducted in three stages: a rest task, a speech task, and an arithmetic task with different levels of difficulty, and each task taking 3 minutes for a totally around of 9 minutes per participant. During the experiment, participants were filmed and wore the E4 Empatica wristband that measured their BVP and EDA signals. The dataset includes the biophysiological measures as well as the video recording of the participants undergoing the three tasks. Self-reported anxiety scores were calculated before and after the experimental sessions. The dataset can be used to study the relationship between physiological signals and social stress, as well as to develop algorithms for detecting and classifying anxiety.

1.1.6.5. AffectiveROAD

The AffectiveROAD dataset by Haouij et al. in 2018 [144] is a publicly available dataset of physiological and environmental data collected from drivers in real-world driving conditions. It was collected using a driving protocol inspired by the SRAD dataset. The dataset contains 13 drives performed by 10 drivers. The total duration of the experiment was 86 minutes. Similar to the SRAD dataset, AffectiveROAD incorporated the biophysiological signals but also added external environmental sensors. The biophysiological signals collected for this study were: PPG, 3D hand acceleration (ACC), EDA, ECG, skin temperature and breathing rate. The external environmental sensors measured temperature, pressure, humidity, luminosity, audio signals, and sound amplitude, as well as video of the outside and inside of the vehicle. The dataset can be access from MIT Media Lab here https://www.media.mit.edu/tools/affectiveroad/.

1.1.6.6. CogLoad and Snake Dataset

The CogLoad and Snake datasets, described by Gjoreski et al. 2020 [145], represent valuable public resources for investigating cognitive load and user experience in game environments. These datasets offer researchers a wealth of physiological and cognitive data collected from participants performing various tasks.

The CogLoad dataset comprises physiological measures like heart rate, EDA, respiration rate, ACC, and skin temperature, collected from over 23 participants engaging in diverse

cognitive load tasks, including reading, problem-solving, and memory challenges. Additionally, the dataset incorporates subjective measures through the NASA Task Load Index, allowing for a comprehensive assessment of individual cognitive workload. This resource proves instrumental in exploring the impact of various factors such as task difficulty, time pressure, and fatigue on cognitive load.

The Snake dataset, in contrast, focuses specifically on the cognitive demands associated with playing a 2D snake game with adjustable difficulty levels. It features physiological data (heart rate, EDA, and respiration rate) from 146 participants, alongside eye-tracking data (gaze fixation) and performance data (accuracy, reaction time, collisions). Subjective data on perceived difficulty, frustration, and enjoyment further enhances the dataset's utility. This comprehensive data collection facilitates research on cognitive load, user experience, affective computing, and educational game design.

Together, the CogLoad and Snake datasets offer a valuable resource for researchers investigating the interplay between cognitive processes and game-based experiences. By examining these datasets, researchers can gain valuable insights into how cognitive load evolves under different conditions, how user experience is shaped by game design elements, and how physiological and behavioral data can be leveraged to understand and predict emotional states in real-time. This knowledge can be used to develop more effective learning tools, create more engaging game experiences, and advance the field of affective computing. The dataset can be accessed here https://gitlab.fri.uni-lj.si/lrk/mobile-cogload-dataset.

1.1.6.7. Monitoring stress with a wrist device using context

The dataset described by Gjoreski et al. offers a distinctive approach to stress detection in both laboratory and real-life settings [132]. It involved 21 participants for laboratory experiments and 5 participants in real-life scenarios, with data collection spanning 55 days. The study primarily utilized the Empatica E4 wrist device to gather comprehensive biophysiological data, including BVP, EDA, Heart Rate (HR), Inter-Beat Interval (IBI), and Skin Temperature (TEMP). Stressors ranged from cognitive tasks in the lab to various real-life stressors. Additionally, the study incorporated the (STAI) for subjective stress assessment, supplementing this with innovative data labeling techniques such as stress logs and Ecological Momentary

Assessment (EMA) via smartphones in real-life settings. This approach allowed for detailed and dynamic data annotation in naturalistic environments. The dataset's dual focus on laboratory and real-world contexts, combined with its extended duration of data collection, makes it a valuable resource for developing and testing robust ML models for stress detection. The dataset can be accessed here https://martingjoreski.github.io/.

1.1.6.8. The RADWear and WEAR datasets:

The RADWear and WEAR ongoing experimental studies, conducted by our collaborative group at the University of Illinois at Urbana-Champaign and supported by UIUC Jump ARCHES program grants (Hernandez, 2021; Hernandez,2022), aim to understand the impact of experimental environments on cognitive load and stress, using a multi-faceted approach. These studies extend beyond the confines of controlled laboratory settings, venturing into real-world environments to assess the efficacy of wearable technology in monitoring stress and anxiety in authentic contexts. This dual-environment approach maximizes the validity and applicability of the findings, paving the way for the development of personalized interventions tailored to individual stress response profiles.

The RADWear dataset specifically targets the medical field by collecting data from 20 medical students in situ during their hospital rotations. It captures the demanding nature of their environment, with initial testing during the calibration protocol that includes a meditation session and a cold pressor test to establish baseline and stressed states, respectively. This is followed by continuous data collection throughout their two-week rotations in the hospital, capturing real-time physiological responses to the intense medical training environment.

Conversely, the WEAR dataset focuses on a broader academic setting, involving 60 STEM-major university students, half of whom are from minority backgrounds, thereby representing a diverse cohort. The WEAR protocol includes a series of controlled lab experiments designed to induce stress and relaxation through methodologies such as the Trier Social Stress Test (TSST), meditation, cold pressor test, and Stroop test. After the initial laboratory testing, WEAR participants undergo ten days of data collection in real-world settings, allowing for an assessment of stress responses outside of the lab environment.

For this dissertation, data from nine RADWear and 27 WEAR test participants were examined, as presented in Chapter 4. Data collection utilized a variety of sensors, including EEG, EMG, E4 wristband, and Hexoskin shirt, to capture extensive physiological responses, such as cardiovascular, EDA, and respiratory patterns. This comprehensive range of data, collected in both controlled and real-world settings, underscores the datasets' value in examining the nuanced effects of environmental factors on stress and anxiety.

In summary, the RADWear and WEAR datasets represent significant strides in the domain of stress and anxiety detection. They offer detailed physiological data and insights into human responses under various stressors and will serve as valuable resources for researchers seeking to develop and validate algorithms for real-time stress detection. These datasets contribute to the advancement of mental health care through the design of personalized interventions and wearable technology solutions.

1.2. Objective and research significance

This dissertation embarked upon a detailed investigation of machine learning (ML) methodologies for anxiety detection, directly confronting both theoretical frameworks and practical obstacles in the discipline. The overarching aim was to bolster the precision, dependability, and real-world functionality of anxiety detection tools, with a specialized emphasis on their robustness in the face of environmental noise—a notable variable that significantly influences the performance and reliability of these systems. Through three synergistic studies, the research critically appraised the existing status of ML in anxiety detection, appraised the resilience of these models against environmental noise conditions, and scrutinized the potency of cutting-edge ML strategies in the midst of real-world, noisy backdrops. This holistic method was directed not just toward pushing forward the technological and methodological boundaries of anxiety detection, but also toward making a substantial contribution to the wider realm of mental health care. It strove to clear a path for the creation of more sophisticated, robust, and user-friendly diagnostic and monitoring tools, casting light on the nuances of ML's role in this emergent field. This section delineates the impetus, aims, and significance of each constituent study, collectively emphasizing the dissertation's role in enhancing our understanding and application of ML in anxiety detection.

The dissertation commenced with a thorough exploration of anxiety detection methodologies, spanning from nuanced biophysiological measurements to comprehensive behavioral analyses. This foundational work set a contextually rich stage for the dissertation's targeted research pursuits. The subsequent discourse shifted from a general treatise on the diverse methods of detecting anxiety to a concentrated critique on the instrumental role of ML within this domain. The quintessence of this dissertation's scholarly inquiry lies in its capacity to interrogate and surmount prevailing challenges intrinsic to anxiety detection, with particular attention given to the dynamic and often confounding influences of environmental noise and practical utility. The proposed studies were designed to critically assess the tenacity of distinct ML model variants across a spectrum of operational scenarios, aspiring to propel the development of robust and efficacious anxiety detection methodologies. By outlining the motivations, objectives, and consequential significance of these interrelated yet distinctive studies, this section aims to frame the narrative of the dissertation as one that not only traverses the frontiers of current scholarship but also sets forth new paradigms in the application of ML for the perceptive detection of anxiety.

Study 1: Review of ML Methods for Anxiety Detection

The emergence of ML in the realm of psychological research, particularly in anxiety detection, represents a transformative shift in our ability to understand and manage mental health conditions. The objective of Study 1 was to comprehensively review the application of ML methods in anxiety detection, a field that has seen significant growth but also faces unique challenges and opportunities for advancement.

Motivation for the Study:

The motivation behind this study stemmed from the increasing prevalence of anxiety disorders and the pressing need for more efficient, accurate, and accessible diagnostic tools. Traditional methods of anxiety detection, as discussed in earlier sections, rely on self-report questionnaires and clinical assessments, which, while valuable, have limitations in terms of subjectivity, scalability, and real-time applicability. ML offers a promising alternative, leveraging the vast amount of data available from the diverse datasets discussed previously.

Utilizing ML algorithms to analyze physiological, behavioral, and environmental data offers the opportunity to create more refined, objective, and real-time methods for detecting anxiety.

However, the application of ML in this context is not without its challenges. Issues such as data quality, model generalizability, ethical considerations, and the interpretability of ML models need to be critically examined. This study aimed to map the current landscape of ML applications in anxiety detection, identifying key trends, challenges, and areas for future research.

Significance of the Study:

The significance of this study lies in its potential to inform future research directions and technological developments in anxiety detection. By conducting a thorough review of existing ML methods, this study provides insights into the strengths and weaknesses of current approaches, highlighting gaps in the literature, and suggesting pathways for innovation.

Furthermore, understanding the application of ML in anxiety detection has broader implications for mental health research and practice. It can lead to the development of more effective treatment strategies, personalized interventions, and preventative measures. Additionally, this study contributes to the ongoing discourse on the ethical and practical considerations of using ML in mental health care, ensuring that these technologies are developed and implemented responsibly and inclusively.

In summary, Study 1 aimed to probe the convergence of ML and anxiety detection, evaluating how this emerging discipline might amplify our grasp and treatment of anxiety disorders. This review not only sketches the present landscape but also sets a trajectory for forthcoming explorations and deployments, aspiring to cultivate more sophisticated, efficient, and compassionate mental health care methodologies.

Study 2: Resilience of ML Models in Anxiety Detection: Assessing the Impact of Environmental Noise on Wearable Technology

Motivation for the Study:

The primary objective of Study 2 was to rigorously evaluate the effectiveness of feature-based (FB) and end-to-end (E2E) machine learning models for anxiety detection under varying conditions of synthetic environmental noise. Utilizing the enhanced WESAD dataset, this study aimed to explore how different levels of Gaussian noise affect the performance metrics of these models, such as accuracy and F1-score. The research sought to delineate the resilience of FB and E2E models to noise interference, thereby providing a clearer picture of their robustness and suitability for real-world applications.

Significance of the Study:

This study addresses a significant gap in the literature concerning the impact of environmental noise on the reliability of anxiety detection systems used in wearable technology. By simulating realistic noise conditions, the research aimed to validate and enhance the applicability of these models in actual usage scenarios where environmental noise is an unavoidable factor. This is particularly vital given the growing dependence on wearable devices for continuous mental health monitoring and the need for systems that can perform reliably in diverse conditions.

The findings from this study are expected to substantially progress the field of machine learning in anxiety detection by developing algorithms that maintain high accuracy even in less-than-ideal environmental conditions. This research not only adheres to but also benchmarks against established best practices in feature extraction and data processing methodologies. By doing so, it contributes to the refinement of anxiety detection technologies, setting new benchmarks for future research and development.

In conclusion, Study 2 enhances our understanding of the interplay between environmental noise and machine learning models in detecting anxiety. The insights derived

from this research are crucial for the advancement of next-generation anxiety detection tools, ensuring they are effective and robust enough for everyday use in mental health management.

Study 3: Advancing Anxiety Detection in Noisy Environments: Efficacy of ML Models on RADWear and Wear Datasets

Motivation for the Study:

The primary objective of this study was to rigorously evaluate the effectiveness of anxiety detection models, particularly in noisy environmental conditions. By introducing noise into the WESAD dataset and employing transfer learning techniques on our RADWear and WEAR datasets, this study sought to test the adaptability and accuracy of models initially trained on standard datasets in more complex, real-life scenarios. The research aimed to understand the impact of environmental noise on these models, observing changes in performance and their ability to accurately identify and classify anxiety-related patterns under varied noise conditions. This study critically assessed the resilience of feature-based (FB) models compared to the less consistent end-to-end (E2E) models in handling diverse environmental noises.

Significance of the Study:

This investigation marks a significant shift from theoretical exploration to actionable application within the dissertation, emphasizing the development of models that function optimally under real-world conditions. The use of the RADWear and WEAR datasets, which encompass a variety of environmental settings, was pivotal for evaluating the robustness of anxiety detection models against environmental noise. This research addressed a significant gap in the anxiety detection domain: the influence of environmental noise, a variable that is frequently encountered but seldom studied in depth. By focusing on the practical application and resilience of these models in realistic settings, the study advanced the field of mental health monitoring substantially. The findings highlighted the superior performance and reliability of FB models in noisy environments and detailed the limitations of E2E models, suggesting that while transfer learning shows promise, the specific application and model architecture require careful consideration to improve efficacy. The ultimate aim of this study was to enhance the practicality

of anxiety detection tools across varied real-world environments, significantly expanding their applicability in mental health care.

CHAPTER 2: MACHINE LEARNING APPROACHES IN ANXIETY DETECTION - A COMPREHENSIVE REVIEW

2.1. Abstract

This review provides a detailed exploration of the current state of anxiety detection using machine learning, with a focus on both feature-based and end-to-end deep learning models. The field has experienced rapid growth, with a significant increase in academic output necessitating an updated review of methodologies, model architectures, experimental conditions, feature selections, and dataset utilization. Feature-based Machine Learning models, such as Support Vector Machines, Decision Trees, and Random Forests, have been extensively employed due to their interpretability and simplicity. However, these models require manual feature engineering, which can be labor-intensive and potentially biased. End-to-end Deep Learning models, including Convolutional Neural Networks and Recurrent Neural Networks, have emerged as powerful alternatives, capable of automatic feature extraction and handling large datasets effectively. This review also highlights the challenges in applying these models to real-world scenarios, where data noise and variability significantly affect performance. Strategies for noise reduction and robust detection methodologies are discussed, emphasizing the need for models resilient to real-world conditions. Additionally, the review underscores the potential of personalized healthcare approaches in anxiety detection, exploring the possibilities of individual-focused model tuning and transfer learning. Moving forward, the field aims not only to detect but also to predict anxiety onset, advancing towards intervention strategies. Real-time biofeedback applications, such as augmented virtual reality and heart rate variability biofeedback training, are promising areas for future research. This review underscores the need for further exploration into model architectures and their suitability for different types of data, advocating for a more nuanced and personalized approach to anxiety detection using machine learning.

2.2. Abbreviations

AB	Adaboost
AES	Apathy Evaluation Scale
BFS	Breadth-first search
Boost	Boosting algorithm, i.e., AdaBoost or XGBoost
CNN	Convolutional Neural Network
CPT	Cold Pressor Test
DASS	Depression, Anxiety and Stress Scale
DT	Decision Tree
E2E	End-to-end models without feature engineering
EM	Expectation Maximization
EMA	Ecological Momentary Assessment
FB	Feature-based models
FBCSP	Filter Bank Common Spatial Patterns
FTT	Feature-Tokenizer transformer
GHQ	General Health Questionnaire
GMM	Gaussian Mixture Model
GP	Gaussian Process
Ham-A	Hamilton Anxiety Rating Scale
Ham-D	Hamilton Depression Rating Scale
kNN	k-Nearest Neighbor
LDA	Linear discriminant analysis
LR	Linear Regression
LSAS	Liebowitz Social Anxiety Scale
MLP	Multilayer perceptron
NB	Naïve Bayes classifier
PANAS	Positive and Negative Affect Schedule
PSS	Perceived Stress Scale
QDA	Quadratic Discriminant Analysis
ResNet	Residual Network
RF	Random Forest
RNN	Recurrent Neural Network
SAM	Self-Assessment Manikins
SCWT	Stroop color word test
SGD	Stochastic gradient descent
SMGK	Spectral Mixture with Gaussian Kernels
SSSQ	Short Stress Scale Questionnaire
SVM	Support Vector Machine
TMCT	Trier Mental Challenge Test
TPOT	Tree-Based Pipeline Optimization

TSST	Trier Social Stress Test
VAD	Valence, Affect and Dominance Rating

2.3. Introduction

The detection of anxiety through machine learning (ML) has become an increasingly critical area of research, marked by significant advancements in both feature-based (FB) models and end-to-end (E2E) deep learning (DL) approaches. As the academic output in this field expands, there is a pressing need to reassess the methodologies, model architectures, feature selection processes, and dataset applications that underpin current practices. Traditional feature-based models like Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF) are popular for their interpretability and ease of implementation, yet they often require intensive manual feature engineering which may introduce bias. On the other hand, E2E models, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), offer a promising alternative with their capacity for automatic feature extraction and handling of complex datasets. Despite their computational intensity and less transparent operation, these models have shown superior performance in anxiety detection tasks. This review offers a comprehensive overview and summary that delves into the nuanced challenges of applying machine learning to anxiety detection in real-world scenarios, explores innovative solutions for enhancing model robustness, and considers the future of personalized healthcare through machine learning technologies.

2.3.1 Background on Stress and Anxiety

In the realm of anxiety theory, the concepts of stress and anxiety are intricately linked, necessitating a clear delineation of their definitions. In the context of anxiety theory, stress and anxiety are interrelated, and it is appropriate to start with their definitions. Stress is typically characterized as an adverse stimulus that triggers physiological and psychological responses [3]. Delving into Spielberger's Trait-State anxiety theory, anxiety is factored into state and trait forms [3]–[5]. State anxiety is described as a temporary and transient negative emotional reaction to stress, marked by heightened activity in the sympathetic nervous system [5]. In contrast, trait anxiety denotes an individual's inherent tendency to experience state anxiety when encountering stress. For the purposes of this work, the focus is primarily on the detection of state anxiety,

herein referred to simply as anxiety. This delineation is based on the use of standard, validated methods for evaluating or inducing anxiety in the literature; in their absence, we categorize the affective state as stress.

2.3.2 Prevalence and Impact of Anxiety

Anxiety significantly affects long-term health and job performance, necessitating the development of diagnostic and therapeutic tools to enhance treatment access and support mental health professionals [6]–[9]. High-stress jobs also demand workload monitoring systems [10], [11]. Prolonged anxiety exposure correlates with severe health risks like cardiovascular diseases and weakened immunity [12], [13], [14], impairing decision-making and cognitive performance, crucial in high-alert professions [15], [16], [17]–[19]. Globally, anxiety disorders are a major challenge, with significant prevalence and economic impact, yet access to mental health services remains limited, exacerbating the issue [20], [21], [22], [23].

2.3.3 Measuring Anxiety

The assessment of anxiety has traditionally depended on self-reported questionnaires. However, these methods are limited by recall bias and their inability to monitor changes in real-time across different contexts [25]–[27]. Recall bias refers to a systematic error due to inaccuracies in remembering past events, reducing the effectiveness of retrospective analyses in understanding behavior dynamically [28]. Recent research has focused on understanding anxiety through the autonomic nervous system, hormonal variations [29]–[33], and specific patterns of subjective and physiological responses [34]–[36].

The sympathetic nervous system, part of the autonomic system responsible for the fight-or-flight response, plays a key role in anxiety through hormonal and physiological changes [25] [37]. Anxiety detection methods include analyzing brain activation [38]–[40], cardiac responses [41], [42] muscular activity [32], electrodermal activity [43], eye movements [44], and behavioral changes like speech and facial expressions [45], all indicative of the sympathetic nervous system's activity.

2.3.3.1. Traditional Measures

Traditional anxiety assessment tools include the State-Trait Anxiety Inventory (STAI) [46] and the Beck Anxiety Inventory (BAI) [47], which measure anxiety through standardized questionnaires. These tools are straightforward to administer and score but may be biased, especially in individuals who are reluctant to express their emotions [48]. Clinical interviews, either structured or semi-structured, conducted by trained professionals, offer a more precise diagnosis of anxiety disorders than self-reports but are more costly and time-consuming [49]. Behavioral observations focus on visible signs of anxiety, such as facial expressions and body language, and are useful for those unable to complete questionnaires or discuss their emotions, though these observations can be subjective and may not accurately reflect anxiety in individuals who can mask their symptoms [50]. Each assessment method has its pros and cons: questionnaires are accessible but subject to bias, clinical interviews are thorough but require significant resources, and behavioral observations are valuable for non-verbal assessments but can be subjective and misleading [51].

2.3.3.2. Biophysiological and behavioral Measures

The exploration of anxiety detection has significantly benefited from biophysiological measures, notably through the analysis of brain and cardiac activities, electromyography (EMG), respiration, electrodermal activity (EDA), and behavioral indicators. Brain activity measurement techniques such as EEG, fNIRS, and PET offer insights into the sympathetic nervous system's response to anxiety, despite challenges such as spatial resolution, susceptibility to noise, and cost [52]–[59]. Cardiac activity, particularly heart rate variability (HRV), has been established as a robust indicator of anxiety, with advancements in wearable technologies like PPG enhancing real-time, non-invasive monitoring [60], [61], [70], [62]–[69].

Additionally, EMG and respiration analysis provide peripheral insights into anxiety levels through muscle tension and autonomic nervous system activity, [32], [77]–[81]. EDA and other measures, including pupil dilation and skin temperature changes, have been incorporated into machine learning algorithms for anxiety detection, demonstrating the potential of non-invasive techniques [95]–[103]. Behavioral measures, leveraging observable physical and interactive markers, offer a complementary approach to understanding anxiety through the

analysis of physical activity, voice characteristics, and other behavioral patterns, despite the need for adaptation to individual differences and real-world applicability [105], [106], [116], [146], [107]–[110], [112]–[115]. These multifaceted approaches underscore the complexity of anxiety detection and the necessity for diverse measurement techniques to capture its nuanced manifestations.

2.3.4 Experimental conditions

Experimental methods for inducing anxiety are crucial for studying its physiological and psychological impacts, utilizing both self-reported measures and expert evaluations for ground truth metrics. Anxiety can be triggered in laboratory environments through cognitive tasks, emotional manipulation, and physiological stressors. Cognitive tasks like the Trier Social Stress Test (TSST) and the Stroop Color Word Test (SCWT) are specifically designed to provoke anxiety responses [117], [118]. Emotional manipulations, such as exposure to disturbing images or fear conditioning, and physiological manipulations, like the cold pressor test, serve as other effective methods [118], [119].

The choice of experimental environment, whether it be laboratory settings, real-world contexts, or hybrid environments, is dictated by the research objectives and the demographic of interest. Laboratory settings offer a controlled environment for testing new detection methods, while real-world settings provide insights into the applicability and effectiveness of these methods in daily scenarios. Hybrid environments and virtual reality (VR) offer a blend of controlled conditions with realistic experiences [120].

While laboratory-based anxiety detection research often reports high accuracy rates of 95% or greater, these findings generally see a decrease in accuracy when applied to real-world settings due to environmental disturbances [54], [121]–[129]. This discrepancy highlights a significant challenge in anxiety detection research: the translation of lab-based findings to practical, real-world applications. The gap between laboratory precision and real-world variability underscores the necessity for developing detection methods that are not only sophisticated but also resilient to environmental disturbances and adaptable to the diverse conditions encountered in daily life. This adaptation is vital for ensuring the reliability and

applicability of anxiety detection technologies in the dynamic and unpredictable environments where they are most needed.

2.3.5 The Shift from FB to E2E methods in Anxiety Detection

The advent of ML in affective computing marks a significant shift from traditional methods to more data-driven approaches. Initial ML efforts in this field predominantly utilized FB models, requiring domain-specific knowledge for feature extraction. However, these models were potentially biased, limited by the information included for training. Furthermore, most FB models in early research were supervised learning models, with unsupervised models showing mixed results and thus being less reliable for anxiety detection [147], [148]. The rapid growth of anxiety detection research, particularly the significant uptick in scholarly contributions in 2023, highlights the evolving nature of this domain and the necessity for an updated examination.

2.3.6 Objective and Significance

This review summarizes the recent advancements and persistent challenges in applying ML techniques for anxiety detection, with a particular focus on their integration with wearable technologies. It investigates how current ML models, encompassing both FB and E2E approaches, are evolving to tackle the intricate challenges inherent in anxiety detection. Through a detailed examination of various methodologies used for detecting and measuring anxiety—ranging from biophysiological metrics such as Electrocardiogram (ECG) and Electromyography (EMG) to behavioral indicators like body pose and voice characteristics—the chapter aims to provide a comprehensive review of state-of-the-art ML techniques in this domain. The objective is to highlight the innovations, applications, and limitations. This analysis is pivotal for pushing the boundaries of anxiety detection using ML, aiming to improve the realism and applicability of research in this area. Ultimately, it contributes to the development of more effective, reliable predictive models for mental health monitoring, underlining the significance of advancing this field for enhanced mental health care and challenges in applicability in real-world scenarios.

2.4. State Of Art Of Anxiety Detection Using ML

2.4.1 Overview of Current Trends and Advancements

In this section, we explore the state-of-the-art in anxiety detection through ML, focusing on the interpretation of physiological and behavioral metrics. The field has seen a remarkable increase in academic output, with a significant 18% of relevant publications emerging in 2023, indicating a rapidly evolving landscape [149]. This necessitated an updated review of recent works, emphasizing ML architectures, experimental conditions, ground truth establishment, anxiety induction methods, feature selection, and the use of open-source datasets.

ML techniques, leveraging physiological and behavioral metrics, have demonstrated exceptional capabilities in detecting anxiety. These metrics, fraught with complexities and uncertainties, were navigated efficiently by various ML models. The review aimed to dissect these models, scrutinizing their architecture, efficacy, and application in diverse experimental conditions.

2.4.2 Methodology of Review

The literature review methodology encompassed a thorough search across Springer, IEEE Xplore, and PubMed databases, covering publications from 2010 to September 2023. The search was conducted using keywords combinations, such as “stress or anxiety or mental workload detection," and “AI or machine learning or deep learning” (Table 2.1).

Table 2.1: Search strings used for each of the databases

Database	Search String
PubMed	("machine learning") AND "anxiety" NOT ("depression" OR "Autism" OR Stroke OR "depressive" OR phobia)
IEEE Xplore	("machine learning") AND (("psychological stress" OR “mental stress” OR “emotional stress” OR “mental workload” OR “stressful) OR "anxiety")
Scopus	TITLE-ABS-KEY ("machine learning" AND ("psychological stress" OR "mental stress" OR "emotional stress" OR "mental workload" OR "cognitive workload" OR "Cognitive stress" OR "anxiety")) AND NOT TITLE-ABS (review OR survey OR scoping OR autism OR autistic OR diabetic) AND NOT TITLE (treatment OR suicide OR surgery OR depression OR depressed OR "anxiety disorders" OR vaccine OR child OR children OR cells OR glycemia OR tumor OR tremor OR sex OR wealth OR "mental illness" OR disorder OR "management system" OR "intelligence" OR disease) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp"))

A systematic screening process was implemented, involving title screening, abstract evaluation, and full-text analysis. Inclusion criteria focused on original articles that studied state anxiety detection using ML and evaluated model performance. Articles were required to employ validated methods for inducing or measuring anxiety, such as the Trier Social Stress Test (TSST) [117] or the State-Trait Anxiety Inventory (STAI) [46] Review papers were excluded from consideration.

The initial search yielded 2330 articles, i.e., 1963, 298 and 69 articles from Springer, IEEE Xplore, and PubMed, respectively (Figure 2.1). Title screening narrowed the field to 598. Further abstract evaluation reduced this number to 236, and a full-text review culminated in 71 relevant papers. Data extracted from these articles included ML architectures, ground truth questionnaires, environmental conditions, tasks for inducing anxiety, model performance, physiological signals, and datasets utilized. These 71 articles were further segregated between ML approaches that used FB models (57 articles) or E2E models (8 articles) (Figure 2.1).

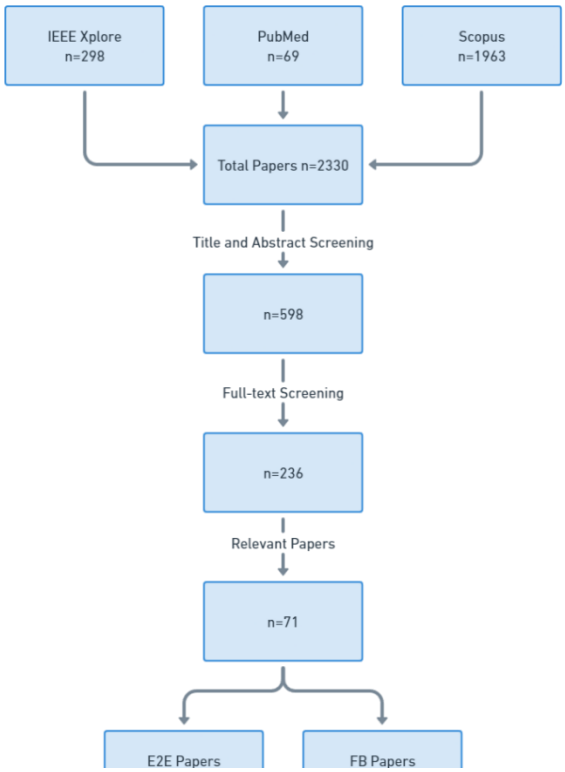


Figure 2.1: Flow chart of review method database results and screening methods for articles, segregated based on ML approach used: FB and E2E.

2.4.3 ML models and Architectures

In this section, the evolving landscape of machine learning methods for anxiety detection is explored, focusing on the comparative efficacy and architectural nuances of Feature-Based and End-to-End models as pivotal approaches in understanding and predicting anxiety states.

2.4.3.1. FB models

The early studies of ML in affective computing predominantly utilized FB (Table 2.2). These models (e.g., SVM, RF, and DT) depend on engineered features specifically chosen for their relevance to the state of interest, in this case, anxiety [61], [62]. Their advantages lie in interpretability, simplicity, and lower computational demands. They are particularly effective when dealing with limited data. However, a significant limitation is the necessity for manual feature engineering, which is both labor-intensive and requires domain expertise. The model's performance heavily relies on the quality of these selected features.

Table 2.2: Flow chart of review method database results and screening methods for articles, segregated based on ML approach used: FB and E2E.

FB category	Models	References
Traditional models	SVM	[150], [54], [151], [128], [152], [121], [153], [122], [154], [155], [109], [132], [124], [156], [157], [158], [159], [160], [161], [162], [163], [164], [125], [126], [165], [166], [167], [168], [169], [170], [139], [92], [171], [172], [173], [174], [175], [176], [177]
	NB	[150], [54], [151], [153], [122], [155], [109], [132], [124], [178], [179], [162], [163], [164], [180], [173], [174], [181], [177]
	RF	[150], [182], [183], [152], [184], [153], [122], [185], [136], [132], [156], [157], [160], [162], [163], [164], [125], [126], [166], [186], [180], [172], [173], [174], [175], [177]
	Boost	[128], [156], [165], [166], [168], [175]
	DT	[182], [151], [128], [153], [154], [155], [132], [157], [158], [168], [180], [172], [175]
	kNN	[150], [54], [182], [151], [128], [153], [155], [109], [132], [124], [96], [187], [157], [159], [161], [179], [165], [180], [172], [181], [177]
	GMM	[158]
	LDA	[131], [150], [188], [151], [128], [155], [159], [161], [125], [126], [180], [92], [175]
	LR	[182], [152], [93]
ANN	CNN	[188], [182], [189], [125], [126], [130]
	MLP	[182], [128], [153], [156], [179], [172], [173]
	LSTM	[182], [190], [191], [192]
	ResNet	[156], [159]

2.4.3.2. E2E models

Recent advancements in hardware and software have led to the rise of E2E or deep learning models (Table 2.3). These deep artificial neural network (ANN) models, including CNN, RNN, and Transformer-based models like BERT, can capture intricate details through

complex architectures. They operate directly on raw data, automatically learning features without the need for manual feature engineering. This approach often results in improved performance due to their ability to handle large data volumes and capture complex relationships. For instance, Dziezyc et al. demonstrated the use of a deep learning model for detecting human anxiety states based on raw physiological data [193]. However, these models are often criticized for being "black box" models due to a lack of interpretability and are computationally intensive, requiring substantial data to prevent overfitting [194].

Table 2.3: Summary tables of articles that used Feature based models

E2E models	Subcategory	References
CNN	-	[195], [196], [197], [198], [130]
FCN	-	[193]
InceptionTime	-	[193]
RNN	LSTM	[195]
	StresNet	[199]
ResNet	-	[193]
Encoder	-	[193]
Hybrid models	Time CNN	[193]
	CNN-LSTM	[193]

Table 2.4: Results of articles that utilized feature-based model approaches

Authors	Machine Learning Architecture or model Used	Questionnaires	Environment L: Lab, W: Wild	Task Used to Induce Stress or Anxiety	Performance (Accuracy)	Biophysical Signals Used	Dataset used
Akbas et al., 2011 [131]	LDA	Non-standard	Hybrid W	Driving task	77%	ECG, EMG, RESP, foot EDA, hand EDA	SRAD
Akella et al., 2021 [150]	SVM, AB, LDA, ridge classifier, Deep Belief Network, RF, kNN	None	L	TSST	91%	EEG	
AlShorman et al., 2022 [54]	kNN, SVM, NB	DASS	L	CPT	90%	EEG	
Appriou et al., 2020 [188]	LDA, FBCSP, & CNN	STAI	L	N-back tasks	60%	EEG	
Arya et al., 2023 [182]	CNN, LSTM, CNN-LSTM, SVM, GB, LR, MLP, RF, DT, kNN	Non-standard	Hybrid W	Driving task	96%	ECG, EMG, RESP, Foot EDA, Hand EDA	SRAD
Assaf et al., 2023 [189]	CNN	custom	L	Mental workload	96%	EEG	
Badr et al., 2023 [200]	kNN, LDA, NB, DT, SVM	None	L	SCWT	99%	EEG	
Beh et al., 2023 [201]	Boost	Not specified	L	SCWT, Mental workload	81%	PPG	CLAS
Benchekroun et al., 2022 [183]	RF	None	L	SCWT, mental arithmetic task	80%	ECG and PPG	
Bobade et al., 2020 [128]	DT, RF, Boost, kNN, LDA, SVM, and MLP.	PANAS, STAI, SAM, and SSSQ	L	TSST	95%	ECG, EDA, EMG, EDA, TEMP, and ACC	
Campanella et al., 2023 [152]	RF, LR, and SVM	None	L	MIST, Mental Test	76.5%	PPG, ACC, EDA and TEMP	

Table 2.4 (Cont.): Results of articles that utilized feature-based model approaches

Authors	Machine Learning Architecture or model Used	Questionnaires	Environment L: Lab, W: Wild	Task Used to Induce Stress or Anxiety	Performance (Accuracy)	Biophysical Signals Used	Dataset used
Choi et al., 2012 [93]	LR	Non-standard	L	Dual tracking, SCWT, memory search, Mirror tracing, and public speech	81%	HR, RESP, EDA, and EMG	
Cruz et al., 2020 [121]	SVM	Non-standard	Hybrid W	Driving task	96%	ECG	SRAD
Dahal et al., 2023 [184]	RF	PANAS, STAI, SAM, and SSSQ	W	TSST	97%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	WESAD + SWELL
Dalmeida et al., 2021 [153]	SVM, KNN, MLP, RF, Boost	Non-standard	Hybrid W	Driving task	80%	ECG	SRAD
Delmastro et al., 2020 [122]	RF, DT, AB, SVM and Adaptive Neuro Fuzzy Inference Systems (ANFISs)	AES and Ham-D, and PSS	L	Physical exercise, SCWT	96%	ECG and EDA	
Dhaouadi et al., 2020 [190]	LSTM, FC,	PANAS, STAI, SAM, and SSSQ	L	Gaming	95%	ECG, EMG, and EDA	
Ding et al., 2022 [154]	LR, SVM, LASSO, and DT	STAI- S, and VAD	L	Negative visual stimuli from the International Affective Picture System, stop-signal task	Not reported	ECG, and EDA	
Erkus et al., 2020 [155]	LDA, kNN, NB, SVM, DT	PANAS, BDI, STAI.	L	Watching video clips	90%	HR, Pupil dilation, EMG, and EDA	
Fukuda et al., 2014 [202], [203]	nonlinear Bayesian regression model	STAI	L	Baseline and relaxation	91%	NIRS	
Gazi et al., 2021 [185]	RF	Non-standard	Hybrid W	Spider exposure in virtual environment for arachnophobic individuals	88%	ECG, EDA, and RESP	SRAD
Giannakakis et al., 2017 [109]	kNN, NB, AB, SVM, generalized likelihood ratio	Self-reports. Details not mentioned	L	Stressful images and videos, SCWT	91%	Facial cues	
Gjoreski et al., 2016 [136]	RF	STAI	W and L	mental arithmetic	83% in L, 92% in W	EEG	
Gjoreski et al., 2017 [132]	DT, NB, kNN, SVM, Bagging, Boosting, RF, Ensemble Selection.	STAI in lab and EMA in wild	L and W	MIST in lab, and examinations in wild	73%	PPG, EDA, Temp, and ACC	
Han et al., 2020 [124]	kNN, SVM, NB	None	L and W	Memory game, image test, CPT, TSST, SCWT, and physical activity.	95%	ECG, PPG, and EDA	
Henry et al., 2023 [156]	SVM, RF, Boost, MLP, ResNet, and a Feature-Tokenizer transformer (FTT)	PANAS, STAI, SAM, and SSSQ	L	TSST	82%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	CASE + WESAD
Kader et al., 2023 [178]	NB	PSS	L	SCWT	71%	EEG	
Karthikeyan et al., 2011 [96], [187]	KNN	None	L	SCWT	88%	ECG, TEMP and EMG	
Kim et al., 2021 [157]	SVM, LF, RF, DT, kNN	None	L	SCWT and Arithmetic test	70%	ECG, EGG, RESP, and EEG	
Kurniawan et al., 2013 [158]	K-means, GMM, SVM and DT	None	L	SCWT, TSST, TMCT	70%	Speech and facial expression, and EDA	
Lingelbach et al., 2021 [159]	LR, LDA, SVM, kNN, RFC, GBC, TPOT, DeepResNet	None	L	N-back task while driving	72%	EDA	
Liu et al., 2023 [160]	SGD, SVM, RF, TPOT	Non-standard	Hybrid W	Driving task	0.781 - 0.934 AUC score	foot GSR, ECG	SRAD
Mamdouh et al., 2023 [161]	kNN, LDA, SVM	PSS	L	MIST, MAE.	87% acc	EEG	
Mazlan et al., 2023 [179]	kNN, NB, MLP, and GNB	PSS and DASS	L	SCWT	46.44% - 99.85% for 4 level classification (baseline, low, medium, high)	EEG	

Table 2.4 (Cont): Results of articles that utilized feature-based model approaches

Authors	Machine Learning Architecture or model Used	Questionnaires	Environment L: Lab, W: Wild	Task Used to Induce Stress or Anxiety	Performance (Accuracy)	Biophysical Signals Used	Dataset used
Meteier et al., 2021, 2022, and 2023 [162]–[164]	RF, Support Vector Classifier, NB	PANAS	L	Simulated Driving	83%	EDA, ECG, and RESP	
Mozafari et al., 2020 & 2021 [125], [126]	LDA, SVM, RF, CNN and information-theoretical learning, and transfer component analysis	GHQ	L	SCWT	96%	GSR, PPG and RESP	
Mozos et al., 2017 [165]	kNN, SVM, Boost	STAI	L	TSST	92%	EDA and PPG	
Naegelin et al., 2023 [166]	SVM, RF, Boost	Non-standard	L	Office tasks, TSST-G in simulated office environment	F1 score: 0.501 - 0.531 for HRV	HRV	
Pinge et al., 2022 [167]	SVM	None	L	Social test, mental arithmetic, and CPT	80%	ECG and PPG	
Plarre et al., 2011 [168]	SVM, Boost, DT	Non-standard	L	Mental workload, social stress, and CPT	72%	ECG, TEMP, ACC, EDA, and RESP	
Praveenkumar et al., 2022 [191]	Deep Neural Network and LSTM.	PANAS, STAI, SAM, and SSSQ	L	TSST	82%	ECG, ACC, RESP, TEMP	WESAD
Quadir et al., 2023 [186]	RF	PANAS, STAI, SAM, and SSSQ	L	TSST	98% for WESAD, 95% F1 score for real data collected from participants	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	WESAD + IoT
Ragav et al., 2023	EMI-LSTM, EMI-GRU, EMI-FastGRNN	PANAS, STAI, SAM, and SSSQ	L	TSST	99%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	WESAD
Rashid et al., 2023 [180]	DT, RF, AB, LDA, kNN	PANAS, STAI, SAM, and SSSQ	L	TSST	85%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	WESAD
Sandulescu et al., 2015 & 2016 [169], [170]	SVM and GMM	STAI	L	TSST; Physical stress.	86%	Humidity, TEMP, HR, ACC, and speech	
Schmidt et al., 2018 [139]	SVM	PANAS, STAI, SAM, and SSSQ	L	TSST	75%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	WESAD
Schmidt et al., 2019 [130]	CNN	STAI	W	None	47%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	
Setz et al., 2010 [92]	SVM and LDA	Non-standard	L	MIST, TSST, and mental workload	83%	ECG, EDA, RESP, and ACC	
Shaposhnyk et al., 2023 [171]	SVM	CSAI	Hybrid W	Driving task	68%	ECG, EDA, and TEMP	SRAD
Siam et al., 2023 [172]	KNN, SVM, DT, LR, RF, and MLP	Non-standard	Hybrid W	Driving task	97%	ECG, EMG, EDA, and RESP	SRAD
Silva et al., 2020 [173]	LR, MLP, NB, SVM, RF, and kNN	PSS	L	Computer-based exam	78%	PPG	
Subhani et al., 2017 [174]	LR, SVM, and NB	PSS	L	MIST	94%	EEG	
Vaz et al., 2023 [175]	LR, LDA, DT, SVM, Boost, RF	PANAS, STAI, SAM, and SSSQ	L	TSST	92%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	WESAD
Wijsman et al., 2011 [181]	NB, kNN, Fisher's Lease square linear classifier	PSS	L	Arithmetic task, logical puzzle task, and memory task	80%	ECG, RESP, EDA and EMG	
Xing et al., 2020 [176]	SVM	Clinical Interview for DSM-5, LSAS, Ham-A, and Ham-D	L	Emotional face matching task	72%	fMRI	
Zhu et al., 2023 [177]	kNN, SVM, NB, LR, RF	PANAS, STAI, SAM, and SSSQ	L	TSST	87%	EDA	WESAD

Table 2.5: Results of articles that utilized end-to-end model approaches

Authors	Machine Learning Architecture or model Used	Questionnaires	Environment L: Lab, W: Wild	Task Used to Induce Stress or Anxiety	Performance (Accuracy)	Biophysical Signals Used	Dataset used
Barki et al., 2023 [196]	CNN	none	L	SCWT	92%	PPG	
Chatterjee et al., 2022 [197]	CNN	PANAS, STAI, SAM, and SSSQ	L	TSST	91%	ECG, EDA, EMG, EDA, TEMP, and ACC	
Dziezyc et al., 2020 [193]	FCN, ResNet, MLP, Encoder, Time-CNN, MCDCNN, StresNet, CNN-LSTM, MLP-LSTM and InceptionTime	WESAD	L	TSST	79%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	WESAD
Fan et al., 2023 [204]	Deep CNN, CNN-CBAM	PANAS, STAI, SAM, and SSSQ	L	TSST	97% 4-class	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	WESAD
Huynh et al., 2021 [199]	StressNAS	PANAS, STAI, SAM, and SSSQ	L	TSST	93%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	WESAD
Sah et al., 2022 [198]	CNN	PANAS, STAI, SAM, and SSSQ	L	TSST	92%	EDA	WESAD
Amin et al., 2023 [195]	CNN, LSTM	non-standard	Hybrid W	Driving task	86.53% - 96.59% for 2-level classification, 76.5% - 87.95% for 3-level classification	ECG, HR, EDA, EMG, and RESP from SRAD dataset and BR, EDA, BVP, ECG, TEMP, ACC, posture, and activity from AffectiveROAD dataset	AffectiveROAD, SRAD (Healey & Picard)
Schmidt et al., 2019 [130]	CNN	STAI	W	None	47%	PPG, ACC, EDA, RESP, ECG, EMG, and TEMP	

2.4.3.3. Comparison of FB and E2E Models

Understanding the distinction between FB DL models and E2E models is crucial. FB models (e.g., [1], [61]) rely on pre-processed, manually engineered features, like traditional FB models. These models can be deep, using complex architectures, but they operate on pre-selected features rather than raw data. They offer more control and interpretability but depend heavily on the quality and relevance of chosen features. In contrast, E2E models learn directly from raw data. They utilize deep neural networks capable of handling high-dimensional data and extracting intricate patterns. E2E models offer less interpretability and require large datasets but are less prone to biases that can arise from human error in feature engineering. The choice between these approaches often depends on the specific goals of the study, the nature of the data available, and the computational resources at hand.

2.4.4 Key Findings and Insights

In examining the data (Tables 2.2 and 2.3), which encapsulate the utilization of FB and E2E ML models in anxiety detection, several critical insights emerge. There is a notable diversity in the models employed, ranging from SVM to Linear Discriminant Analysis (LDA), indicating a rich exploration of various ML methodologies within the field. The majority of studies (86%) were performed in controlled laboratory settings, suggesting that FB models exhibit considerable efficacy under such conditions. However, this efficacy might not extend as robustly to dynamic real-world scenarios, highlighted by the predominance of laboratory-based studies and a relative scarcity of real-world data applications.

Among FB models, SVM was the most prevalent, used in 32 studies, followed by RF in 28 studies, and kNN in 22 studies. DT were also featured in 12 studies. In terms of performance, the highest reported accuracies for FB models reached up to 99% by Badr et al., 2023 using SVM [200], while E2E models achieved accuracies as high as 97% by [204]. However, no single classical ML model has consistently outperformed others across all studies.

Furthermore, the studies predominantly utilized biophysical signals such as ECG, EEG, and EDA, emphasizing their perceived effectiveness in reflecting physiological changes associated with anxiety. Standardized questionnaires like the STAI and the Depression Anxiety Stress Scales (DASS) [205] are frequently used for ground truth establishment, which aids in validating the models against recognized psychological measures.

A key observation was the progression from traditional FB models toward more advanced E2E deep learning models, reflecting the evolution toward embracing more sophisticated, data-driven approaches in the affective computing field. Note inclusion of ANN models in summaries of both FB (Table 2.2) and E2E (Table 2.3) models. The frequent utilization and notable performance of CNNs in anxiety detection, particularly with time-series data, underscores an intriguing shift despite the more natural fit of RNNs and Long Short-Term Memory networks (LSTM) for such data.

The impact of testing environments on model selection and performance is particularly pronounced, with a prevailing focus on laboratory-based studies revealing a significant gap in the

application and testing of these models under real-world conditions. This disparity underscores the risk of overfitting models to highly controlled and predictable laboratory environments, which may not accurately reflect the complexities and unpredictability of everyday scenarios. The necessity for future research to pivot towards enhancing the real-world applicability of anxiety detection models involves adapting models to be more robust against the inherent noise and variability of real-life data and ensuring that these models are tested and validated in diverse, uncontrolled environments.

In summary, these key findings and insights from Tables 2.2 and 2.3 offer a comprehensive understanding of the current state of machine learning in anxiety detection, highlighting the diversity and evolution of methodologies, the critical role of biophysical signals and psychological questionnaires, the impact of testing environments, and the emerging preference for advanced architectures like CNNs in the field. As the field progresses, emphasizing real-world applicability, robustness to environmental factors, and standardization in experimental methodologies will be crucial in advancing towards more effective and reliable anxiety detection tools.

2.4.5 Efficacy and Application of CNNs in Time-Series Anxiety Detection

The transition towards E2E methods, particularly the rising adoption of Convolutional Neural Networks, marks a significant evolution in anxiety detection methodologies of adopting more complex, data-driven approaches. While traditionally, RNS and LSTM have been the go-to models for sequential data due to their ability to capture temporal dependencies, the effectiveness and frequent use of CNNs in this domain have become a noteworthy trend. Despite the challenges associated with the interpretability and computational demands of deep learning models, CNNs have emerged as a powerful tool, often preferred for their robust feature extraction capabilities and efficiency in handling complex patterns inherent in physiological signals [206].

The efficacy of CNNs in anxiety detection can be attributed to several key factors. Their ability to perform automatic feature extraction and adaptability in learning spatial hierarchies of features make them particularly suited for deciphering complex patterns in physiological signals. Advancements in CNN architecture, especially the development of 1D

variants for time-series data, have further broadened their applicability [207]. These networks are known for their robustness, often demonstrating resilience against overfitting [208], and providing computational efficiency, which is particularly advantageous given the extensive datasets typically involved in anxiety detection [208], [209]. Furthermore, the pooling layers inherent in CNNs act as a form of noise reduction, making them suitable for real-world applications where data may be noisy or inconsistent.

The notable efficacy of CNNs extends beyond their architectural benefits. They are adept at identifying local temporal patterns within specific data windows, crucial in recognizing significant local features in physiological data related to anxiety. While they might not inherently capture long-term dependencies as effectively as RNNs or LSTMs, the preprocessing and transformation techniques applied to time-series data can reveal key features that CNNs are well-equipped to learn complex and abstract features from data. The integration of hybrid models, wherein CNNs are used alongside RNNs or LSTMs, showcases an innovative strategy that leverages the strengths of both architectures, combining CNN's feature extraction prowess with the temporal modeling strengths of RNNs/LSTMs (e.g., [133], [193], [199], [210]). Having said that, the one implementation that has been observed did not perform as well as basic models [193].

Despite CNN's limited capability in capturing long-term dependencies, their success in anxiety detection illustrates the importance of model selection being closely aligned with the specifics of the task and data at hand. The dynamic nature of machine learning necessitates ongoing research and refinement of best practices. The proven success of CNNs in related domains of signal processing and pattern recognition, such as in image and speech recognition, has likely influenced their adoption for anxiety detection using physiological data, suggesting a transfer of confidence and methodology from these areas [211], [212]. As the field continues to evolve, the collective understanding of when and how to utilize CNNs, either alone or in conjunction with other models, will expand, guiding future advancements in anxiety detection and the broader realm of time-series data analysis. The convergence of these insights emphasizes the multifaceted nature of model selection and effectiveness, underscoring the evolving landscape of machine learning techniques in mental health applications.

2.4.6 Environmental Impact on Anxiety Detection Models: Laboratory Precision vs. Real-World Applicability

Recent analyses of ML approaches for anxiety detection have underscored the critical role of the testing environment in influencing model selection and performance. While the predominant focus has been on controlled laboratory environments, with 71 studies analyzed, only 4 were conducted in the wild where participant's motion was not limited. It is imperative to discuss the insights gained from these limited but valuable 'wild' studies. The small number of studies conducted in real-world settings highlights a significant research gap, indicating a serious lack of insight into how these existing state-of-the-art models perform outside controlled environments.

For instance, the study by Han et al. in 2020 [124], which is limited to 3 participants, yielded a 81% accuracy rate. While the paper claims a 100% accuracy rate with activity recognition, the small sample size casts doubt on the generalizability of these results. Similarly, 2016 and 2017 studies by Gjoreski et al. [132], [136], though limited to 5 participants each, incorporated physical activity-based context information that seems to boost the model's performance, highlighting the potential for improvement with more complex and representative data.

Schmidt et al., 2019 study [130], is one of the few that attempts to capture anxiety detection in a truly uncontrolled environment with 12 participants. However, the performance reported is relatively low, at 47% F1-score for binary stress detection while 31% for the 3-class case. Further emphasizing the challenge of applying these models effectively in real-world scenarios.

2.4.7 Addressing the Limitations in Affective State Variability in Anxiety Detection

Another critical issue, not directly related to environmental conditions, is the common limitation in the variety of affective states considered in the datasets and experiments. The majority of studies (82%) focused on distinguishing between baseline and stress states, which while effective in controlled lab settings, may not translate well to more nuanced real-world scenarios where multiple affective states are present. In lab conditions, the models may perform well in differentiating between baseline and stress, but when faced with various non-baseline

states in the wild, their performance could diminish significantly. This aspect of model robustness and differentiation in the presence of diverse affective states has not been thoroughly tested, representing a considerable gap in the research.

The limited studies conducted in natural, uncontrolled environments provide crucial insights into the challenges and potential strategies for improving real-world applicability. Future research must address the serious lack of in-depth, comprehensive studies in wild settings, consider a wider variety of affective states, and focus on increasing the size and diversity of participant groups to enhance the generalizability and reliability of anxiety detection models.

By emphasizing the specific challenges and gaps identified through the limited in-the-wild studies and the issues related to the range of affective states, this review chapter aims to provide a deeper understanding of the current landscape of machine learning in anxiety detection and the critical areas for future research and development.

2.5. Challenges and future work

2.5.1 Robustness to real-life data with noise

One of the most pressing challenges in the field of anxiety detection using wearables is dealing with real-world data that may contain various types of noise. These can include motion-related noise, electrical interference, or device-specific noise for example. Addressing this issue involves two primary strategies: (1) Noise Reduction**:** Developing methods to either remove or minimize noise components from the data. (2) Noise-Robust Detection**:** Creating detection methodologies that are inherently robust to the presence of noise.

Efforts have been made to discriminate between psychological stress and physical stress, for instance, by using accelerometer data to categorize the physical activity state during anxiety detection [132], [213]. However, this approach is limited as it hinders detection during physical activity.

The need for analysis and quantification of real-world effects on detection performance is evident. This encompasses evaluating how different model architectures withstand real-world conditions, the sensitivity of certain signals and features, and the development of more robust

detection modes that are validated with real-world data. The impact of device fidelity, such as sampling rates, is also a crucial factor in performance.

2.5.2 Architectural exploration

The black box nature of deep learning models and their computational intensity, along with the variability in performance metrics and occasional lack of reporting, suggest potential inconsistencies in the field, making cross-comparison of models challenging. The dynamic nature of machine learning, where best practices are continually refined and evolved through ongoing research and experimentation, is evident in these shifting trends and preferences in model architectures. Given the notable performance of CNNs in anxiety detection with time-series data, further exploration into these models could provide better explainability of performance. This could assist in developing more tailored models for anxiety detection, addressing both the feature extraction capabilities and the challenges in interpretability and computational intensity.

2.5.3 Towards anxiety prediction and intervention

The future of anxiety detection lies in moving beyond mere detection towards predicting anxiety onset and developing intervention strategies. There is a notable gap in methodologies for predicting anxiety onset. However, promising applications in real-time biofeedback, such as augmented virtual reality, HRV biofeedback training, and virtual reality visualization [214]–[216], offer avenues to reduce anxiety levels. Leveraging existing anxiety detection frameworks could significantly enhance the evaluation and improvement of anxiety therapies, facilitating coping mechanisms for negative emotions.

2.6. Conclusions

This chapter critically examined the current state of machine learning methods in anxiety detection, revealing the significant advancements as well as the ongoing challenges in the field. The extensive review highlights a diverse range of methodologies, from traditional Feature-Based models to advanced End-to-End approaches, each with their own set of strengths and limitations. The analysis of various studies has shown a high efficacy of these models in

controlled laboratory settings, while also underscoring the need for expanded research in real-world environments to assess and enhance their practical applicability.

The insights from studies conducted in 'wild' settings, though limited, have revealed a crucial gap in the research and a pressing need for models that can effectively operate amidst the complexities of real-life scenarios, such as increased noise. Moreover, the challenge of differentiating between various affective states in more dynamic environments has been recognized as an area requiring substantial advancement.

As the field progresses, future research must focus on addressing these gaps, improving the robustness, versatility, and generalizability of anxiety detection models. Emphasizing real-world applicability, enhancing noise resilience, exploring innovative architectures, and expanding the understanding of affective state variability will be key in advancing the field. The ultimate goal is to develop reliable, effective, and universally applicable machine learning tools for anxiety detection, contributing significantly to the well-being and mental health care of individuals across diverse settings.

By continuing to bridge the divide between laboratory precision and real-world complexity, the field can move closer to realizing the full potential of machine learning in providing meaningful insights and interventions for anxiety and stress-related conditions. The journey is complex and challenging, but with continued research, collaboration, and innovation, the future holds promising prospects for anxiety detection and mental health care.

CHAPTER 3: RESILIENCE OF MACHINE LEARNING MODELS IN ANXIETY DETECTION: ASSESSING THE IMPACT OF ENVIRONMENTAL NOISE ON WEARABLE TECHNOLOGY

3.1. Abstract

The importance of mental health monitoring systems is well understood, though the development of systems robust to real-life environmental conditions is yet to mature. In this study, the resilience of machine learning models for anxiety detection through wearable technology was explored to better inform the further development of mental health monitoring systems. This work informed us on the best models, modalities, and features for use in real-world applications. The effectiveness of feature-based (FB) and end-to-end (E2E) machine learning models for anxiety detection was rigorously evaluated under varying conditions of environmental noise. By adding synthetic Gaussian noise to a well-known open access affective states dataset collected with commercially available wearable devices (WESAD), a performance baseline was established using the original dataset. This was followed by an examination of the impact of noise on model accuracy to better understand model performance (F_1-score and Accuracy) changes as a function of noise. That was done by adding ten varying magnitudes of Gaussian noise. The results of the analysis revealed that with the increase in noise, the performance of FB models dropped from a high of 90% F1-score and 92% accuracy to 65% and 70%, respectively; while E2E models showed a decrease from an 85% F1-score and 87% accuracy to below 60% and 65%, respectively. This indicated a proportional decline in performance across both FB and E2E models as noise levels increased, challenging initial assumptions about model resilience. This analysis highlights the need for more robust algorithms capable of maintaining accuracy in noisy, real-world environments and emphasizes the importance of considering environmental factors in the development of wearable anxiety detection systems.

3.2. Introduction

We previously conducted a comprehensive review of the state-of-the-art in anxiety detection using wearables and machine learning (Chapter 2). The review highlighted a significant impediment to the success of anxiety detection systems, which is the impact of noise. There also remains a gap in the literature regarding an in-depth analysis of how and why noise adversely affects these algorithms.

The interest in developing robust anxiety detection systems stems from the substantial public health impact of anxiety disorders, which are among the most common mental health issues globally [20]–[23]. We believe that effective detection systems can facilitate early intervention, enhance objective measurement, personalize healthcare, increase accessibility, and provide valuable research insights.

The primary objective of this study was to rigorously evaluate the effectiveness and accuracy of machine learning models to detect anxiety under various types and levels of environmental noise. Specifically, the study sought to understand the extent to which these models can maintain accuracy in identifying and classifying anxiety-related patterns amidst noise interference.

Furthermore, this research aimed to explore the potential benefits of employing deep learning end-to-end models in improving the robustness of anxiety detection systems against noisy conditions. To this end, this study utilized the Wearable Stress and Affect Detection (WESAD) dataset [139], a publicly available resource, as the basis for experimentation. This dataset was augmented with synthetic noise to simulate real-world noisy environments, thereby providing a platform to assess model performance as a function of noise interference.

In pursuit of replicating and benchmarking against the state-of-the-art, this study adhered closely to established literature [62], [139], [193] in terms of filtering parameters, data processing techniques, and methodologies for feature selection and extraction. This approach ensured that the findings were grounded in current best practices [35], [95], [217], [218], allowing for a meaningful contribution to the field of anxiety detection.

Studying noise is essential for several reasons. First, it allows for a more accurate simulation of real-world conditions, as noise is an inherent part of data collected through wearable technology. Understanding the different types of noise and their effects on data helps in creating models that are resilient and reflective of real-life scenarios. Second, this study aids in enhancing the robustness of machine learning algorithms. By analyzing how noise influences data and model performance, researchers can develop algorithms that maintain high accuracy levels, even in suboptimal conditions. Last, a thorough analysis of noise contributes to the reduction of potential errors, leading to more reliable and trustworthy applications of machine learning.

3.2.1 Background of Stress and Anxiety

3.2.1.1. Stress and Anxiety: Definitions and Context

In the realm of anxiety theory, stress and anxiety are closely intertwined. It is essential to define these terms to establish a clear understanding. Stress is typically characterized as an adverse stimulus [3]. According to Spielberger's Trait-State anxiety theory, anxiety is bifurcated into state and trait anxiety [3]–[5]. State anxiety represents a temporary, transient emotional response to stress, marked by increased sympathetic nervous system activity. In contrast, trait anxiety denotes an individual's inherent tendency to experience state anxiety under stress. This study primarily focused on state anxiety, hereafter referred to as 'anxiety', especially in the context of detection methodologies. The distinction between anxiety and stress in this study hinges on the use of standardized methods for evaluating or inducing it; the former is considered anxiety, while the latter is categorized as stress.

3.2.1.2. Prevalence and Impact of Anxiety

The pervasive impact of anxiety on long-term health and job performance underscores the urgency for developing effective anxiety detection tools [6]–[9]. These tools could significantly enhance accessibility to diagnosis and treatment and serve as valuable resources for mental health professionals. Anxiety's detrimental effects extend to increased cardiovascular disease risks, weakened immune responses, and reduced performance efficiency [12], [14], [219], [220] . The prevalence of anxiety is alarmingly high, with significant portions of the world

populations affected, contributing to substantial economic losses [20]–[23]. This prevalence is juxtaposed against the stark reality of inadequate mental health services, highlighting a critical area of need.

3.2.1.3. Measuring Anxiety: Traditional and Biophysiological Approaches

The limitations of traditional anxiety monitoring methods, such as self-reported questionnaires, necessitate more dynamic and quantitative measures [25]. These measures should be capable of capturing real-time changes across various contexts [26], [27]. The focus has shifted towards understanding the autonomic nervous system's role in anxiety, leading to the development of diverse tools for anxiety assessment. These tools range from neuroimaging techniques like Electroencephalography (EEG), and Functional Near-Infrared Spectroscopy (fNIRS) [38]–[40], each with its unique advantages and limitations, to cardiac activity analysis using heart rate variability (HRV) and innovative non-contact telemetry methods [41], [42]. Peripheral biophysiological measures, including Electromyography (EMG) [32], respiration patterns [37], and Electrodermal Activity (EDA) [43], offer additional insights into anxiety states. Behavioral measures have also emerged as crucial tools in detecting anxiety, utilizing physical and interactive markers to reveal cognitive and affective states related to anxiety [27].

a. ECG and BVP

Cardiac activity, particularly heart rate variability (HRV), has been a key non-invasive indicator for anxiety detection since last century [60], reflecting the sympathetic nervous system's activation. HRV negatively correlates with heart rate, i.e., variability decreases while heart rate increases with anxiety. While early studies used ECG for cardiac monitoring, photoplethysmography (PPG) has emerged as a practical alternative, despite its susceptibility to motion artifacts. PPG measures blood volume pulse, offering portability for real-life heart rate and HRV monitoring. Research by Sayers et al. [64], and others [67], [68] highlights the inverse relationship between HRV and heart rate, using it as an anxiety measure. HRV, often represented as the standard deviation of NN-intervals (SDNN), diminishes as heart rate increases under anxiety, establishing its utility in anxiety detection [61], [62], [64].

b. EDA and TEMP

EDA signal is known to contain two major components: Phasic components (Skin Conductance Response: SCR) and Tonic components (Skin Conductance Level: SCL). The SCR component captures the reaction to a response in a shorter time-scale, in the sub-10 seconds scale, while the SCL is the slow changing component measuring the psychological activation with a timescale ranging between tens of seconds to minutes [221], [222]. Change in skin temperature (TEMP) has been shown to closely relate with EDA response and activation (e.g., [97], [108], [223], [224]).

c. RESP, EMG, and ACC

Exploring the impact of physiological signals on anxiety detection, this section delves into the significance of respiration (RESP), electromyography (EMG), and acceleration (ACC) metrics. RESP has been found to be closely relate with nervous system activation, and used for detection of stress and anxiety (e.g., [85], [88], [91], [225]). Similarly, EMG has been shown to be a useful signal for affective detection. An added advantage of EMG is that it can also be used to differentiate between different kinds of stresses, between physical stimulation, and cognitive stimulation [32], [79]. ACC been shown in multiple studies to be successfully used for affective computing and anxiety detection [226].

3.2.1.4. Anxiety Experimental Methods: Inducing Anxiety for Research

Various methods have been employed in laboratory settings to induce anxiety, each tailored to specific research objectives. These methods range from cognitive tasks like the Trier Social Stress Test [117], evoking emotions through experienced stimuli [227], to physiological manipulation like the Cold Pressor Test [119]. The choice of method is contingent upon the research question, whether it is examining anxiety's effects on cognitive performance or other aspects.

3.2.2 WESAD Dataset

The WESAD (Wearable Stress and Affect Detection) dataset, developed by Schmidt et al. [139], represents a significant contribution to the field of wearable stress detection. It offers

a comprehensive and multimodal dataset that included three affective states: baseline, amusement, and stress, providing a more nuanced understanding of affective responses. This dataset has been widely recognized in academic circles, as evidenced by its numerous citations across various scholarly databases (e.g., [128], [228]–[233]).

While the WESAD dataset provides a valuable resource for studying anxiety detection using wearable sensors, it is important to acknowledge its potential limitations. The dataset was collected from a relatively small sample of 15 participants, which may limit the generalizability of the findings to larger and more diverse populations. Additionally, the data were collected in a controlled laboratory setting, which may not fully capture the complexity and variability of real-world environments. These factors should be considered when interpreting the results and applying the findings to real-world scenarios.

3.2.2.1. Hardware, Protocol, and Ground Truth in WESAD

The WESAD study involved 15 participants, with a gender distribution of 20% female. Two wearable devices were used:

- RespiBAN Professional (PLUX Wireless Biosignal S.A., Portugal): Worn around the chest (Figure 3.1), it measured ECG, EDA, EMG, RESP, TEMP, and 3-axis acceleration (ACC) at a sampling rate of 700 Hz. Specific sensor placements and methodologies were adopted for accurate data collection. EDA was recorded from the rectus abdominis due to the high density of sweat glands [234]. EMG was collected from Trapezius muscles on both sides of the spine, an ideal location [32], [235].
- Empatica E4 Wristband (Empatica, Inc., Boston, MA, USA): Worn on the nondominant hand, it recorded BVP, EDA, TEMP, and ACC at varying sampling rates.

The study's protocol included baseline, amusement, and stress conditions, interspersed with meditation and recovery segments. The sample distribution of these affects is shown in Table 3.1. Participant self-reports after each condition provided the ground truth, incorporating scales like Positive and Negative Affect Schedule (PANAS) [236], State-Trait Anxiety Inventory (STAI) [46], and Short Stress State Questionnaire (SSSQ) [237]. The baseline condition involved participants spending the first 20 minutes with sensors, engaging in neutral activities like reading

magazines, to establish a neutral affective state. For amusement, they watched a mix of humorous and neutral video clips, selected from the affective film library of Samson et al. [238]. The anxious state was induced using the Trier Social Stress Test (TSST) [117]. Meditation and recovery periods were interspersed between conditions to neutralize the participants' affective state. The protocol varied to prevent order effects. Participants' self-reports after each condition, using scales from PANAS, STAI, and SSSQ, provided ground truth data for the study.

Table 3.1: Samples distribution among available affect classes

Class	Number of samples	%
Baseline	70,490	53%
Stress	39,900	30%
Amusement	22,610	17%
Total	133,000	100%

Figure 3.1: Devices used for the WESAD experiment are shown here. Left: RespiBAN wearable chest device. Right: E4 Empatica Wristband [139]. *Left image reference https://www.techtruster.dk/wearable-tech-bliver-terapi/ right image ref: https://www.empatica.com/research/e4/ .*

3.2.2.2. Review of WESAD Methodology

The WESAD study adopted a comprehensive approach to data collection and preprocessing, as described in the original paper. It involved the collection of various physiological signals under controlled conditions, which were then segmented for analysis. Following established methodologies in the field, such as those described by Siirtola et al. [233], the data were segmented using a specific window size and stride to ensure consistency and comparability with other studies in wearable anxiety detection.

The WESAD researchers utilized a window size of 60 seconds with a stride of 30 seconds, a standard practice that balances temporal resolution and computational efficiency, allowing for detailed analysis of physiological responses over time. This segmentation approach is critical for capturing relevant features from physiological data, which are then used to train machine learning models to detect states of anxiety, amusement, and baseline conditions effectively.

The preprocessing steps outlined by the WESAD team aimed to prepare the data for robust machine learning applications, ensuring that the models developed could reliably identify and differentiate between the emotional states under investigation. By aligning with techniques recommended by Siirtola and colleagues, the WESAD study ensured that its findings were grounded in sound methodological practices, enhancing the reliability and applicability of its conclusions in the broader context of anxiety detection research.

3.2.3 Exploring Machine Learning Models for Anxiety Detection

The evolution of machine learning in affective computing has transitioned from relying on classical models requiring extensive domain expertise for feature extraction to the adoption of sophisticated Deep Learning (DL) approaches that promise more detailed data capture without explicit feature engineering. These observations are shown in Chapter 2 Table 2.5. This can potentially cause some bias, since models can only gain insight from information that the researchers hand pick for training or that they may disregard other information, due to preconceived notions or limited understanding [239]. These models are referred to as feature-based (FB) models. Recent breakthroughs in hardware and software have led to the domination of Deep Learning (DL) models in machine learning applications, which can capture more detail due to their high dimensional complex architectures [240], [241]. The end-to-end approach to DL models does not require feature engineering. These models are referred to as End-to-End models (E2E). It is worth noting that DL models can be implemented with a feature-based approach, but the added complexity does not yield better performance, as described in chapter 2 indicating performance of E2E models achieving 79% Accuracy, compared to 97% of FB models [242].

3.2.3.1. FB models

Traditional models like Decision Tree [243], Random Forest, Support Vector Machine [244], Adaboost, Linear Discriminant Analysis [245], k-nearest neighbor [245], and XGBoost [246] have been employed for anxiety detection. Despite significant advancements in the field of machine learning, these six models remain dominant and effective in the realm of machine learning-based anxiety detection, continuing to deliver high performance in identifying anxiety [247], [248], while providing high detection performance. For example, Schmidt et al. [139] was able to achieve F_1-score of 68.85% and accuracy of 79.57% using all modalities. The enduring relevance of these FB models lies in their robustness and the depth of interpretability that they offer, making them indispensable for rigorous anxiety detection analyses and ensuring their inclusion in this study's comprehensive model evaluation.

3.2.3.2. E2E models

Following the foundational work of Schmidt et al. [139], there has been a surge in applying advanced machine learning models to the WESAD dataset, notably by Dziezyc et al. [193] who implemented ten E2E models for anxiety detection. Models incorporated in their study were: Fully Convolutional Network (FCN), Residual Network (Resnet), Multilayer Perceptron (MLP), Encoder, Time convolutional neural network (Time-CNN), Multichannel deep convolutional neural network (MC-DCNN), Spectrotemporal residual network (Stresnet), Convolutional neural network with long-short term memory (CNN-LSTM), Multilayer perceptron with long-short term memory (MLP-LSTM), and InceptionTime (Inception). Descriptions of these architectures are presented in Table 3.2.

A key revelation from these studies is the surprising efficacy of nonrecurrent architectures such as CNNs over recurrent ones like LSTMs, challenging the traditional reliance on feature engineering for anxiety detection and confirmed by subsequent studies [195], [196], [204]. These E2E models not only simplified the detection process by eliminating the need for manual feature selection but also highlighted the evolution of machine learning techniques capable of achieving high accuracies, such as a 92% F1-score in binary anxiety detection settings. The exploration of E2E models is critical, as it demonstrates their potential to

revolutionize anxiety detection, offering insights into developing more adaptive and efficient systems.

Table 3.2: Summary of DL architectures presented in [193]*, describing their layer sequences and configurations.*

Architecture	Description
FCN	N[1] × [CL[2] – CL – CL] – FC[3]
Resnet	N x [ResBloc[4] - …- ResBloc] – FC
Stresnet	N x [ResBlock-Time + ResBlock-Freq] – FC
MLP	N x [FC - …- FC] - FC
Encoder	N x [CL – CL – CL - Att[5]] - FC
Time-CNN	N x [CL – CL] - FC
CNN-LSTM	N x [CL – CL - LSTM[6]] - FC
MLP-LSTM	N x [FC - …- FC - LSTM] - FC
MC-DCNN	N x [CL – CL] - FC - FC
Inception	N x [Inc[7]] - FC

3.2.4 Impact of Environmental Noise on Model Performance

Environmental noise presents a significant challenge to the performance of machine learning models, particularly in the domain of anxiety detection using wearable devices. This subsection outlines our approach to evaluating how simulated environmental noise impacts the effectiveness of various machine learning models, specifically feature-based and end-to-end architectures. This investigation is critical, as the existing literature lacks comprehensive insights into how such noise affects these models, highlighting a significant gap in our understanding of model robustness under real-world conditions.

In this study, we utilized a range of FB and E2E models, including those detailed in Table 3.2 and previously characterized by Dziezyc's group, to assess their resilience to noise. Our

[1] N: number of signals
[2] CL: convolutional layer
[3] FC: fully connected layer
[4] ResBloc: residual block with three CLs
[5] Att: Attention Module
[6] LSTM: long short term memory module
[7] Inc: Inception Module

analysis was geared towards identifying the models that maintain the highest accuracy in noisy environments, which is pivotal for developing reliable anxiety detection systems. Our research questions are driven by the need to determine whether existing state-of-the-art (SOTA) models can effectively perform under challenging noise conditions and which specific features and model architectures are most adept at handling environmental disturbances.

This initial section introduced the framework for a detailed analysis that will be further elaborated in the Methods section. Here, the focus is on assessing the effects of simulated environmental noise on the performance metrics of various machine learning models. This examination aims to clarify how noise impacts model accuracy, thereby informing the selection and optimization of machine learning models for real-world applications where environmental noise is a constant factor.

3.3. Methods

The analysis and evaluation employed in this study used the following process. The steps were: preprocessing, feature extraction (for FB models), model creation, noise addition, and training and testing. Each of these steps are expanded upon below, aiming to allow as much reproducibility as possible. Another principle that guided our choices was the ability to benchmark with the existing SOTA approaches, including which models to implement, which features to include, and which models to test. A general overview of how our complete processing and modeling pipeline is presented in Figure 3.2. The custom Python code used to perform all the work mentioned here are available in: https://github.com/AbdulAlkurdi/anxietyFB and https://github.com/AbdulAlkurdi/anxietyE2E.

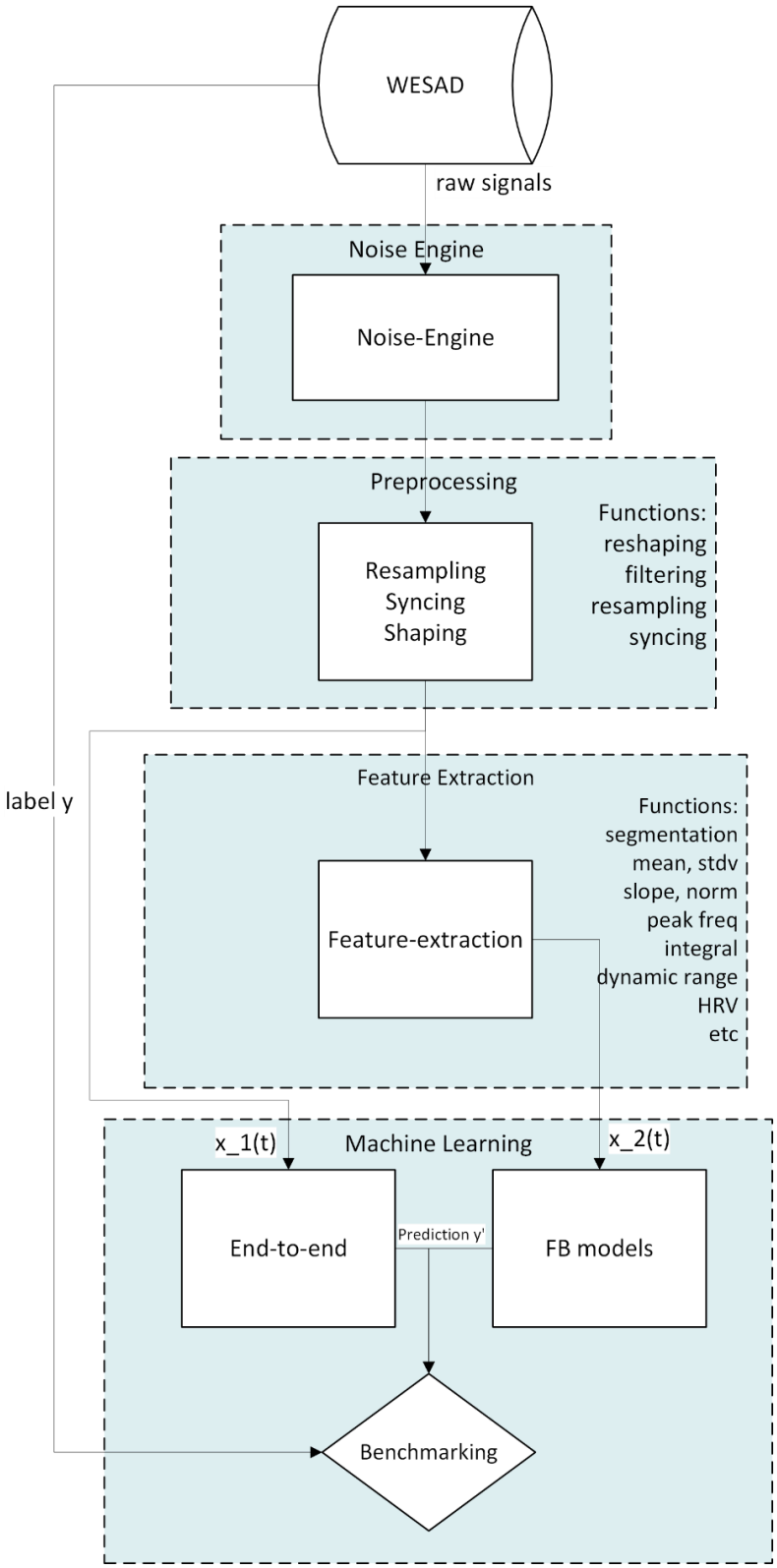


Figure 3.2: This flowchart outlines a pipeline for processing and analyzing wearable sensor data, from raw data collection through noise filtering, preprocessing, and feature extraction, to machine learning.

3.3.1 Data Preprocessing

In preparing the data for machine learning analysis, particular attention was paid to the selection and treatment of input modalities. In this study, some modalities were dropped to reduce the number of inputs to the models, aiming to reduce redundancy and improve model performance. This adjustment was also better suited to align with the RADWear and Wear datasets, which feature a more limited set of modalities compared to the comprehensive WESAD dataset. This approach ensured that the models were optimized for performance while being applicable to real-world conditions reflected in the datasets used. The preprocessing steps included normalization, feature extraction, and the handling of missing data, which were crucial for preparing the dataset for effective machine learning analysis.

The preprocessing stage was crucial in preparing the raw sensor data for analysis and modeling. The first step involved defining the sampling rates for various sensors, such as the accelerometer, blood volume pulse, and electrodermal activity. This ensured that the data from different sensors were properly aligned and synchronized. One step that differed in our preprocessing implementation, compared to Schmidt et al., was that the raw signals were preprocessed without converting raw signals from raw values into their respective units. For example, EDA from the RespiBAN was provided in voltage and conversion was required to obtain siemens (S) units.

For FB models, EDA and ACC signals were filtered using a lowpass Butterworth filter with cut-off frequencies of 5 Hz and 13 Hz, respectively. For ECG peak detection, the Automatic Multiscale-based Peak Detection (AMPD) algorithm was based on [249].

The preprocessing steps for E2E models differed from those for FB models. The steps applied to the signals for E2E models were winsorization, filtering, downsampling, and min-max normalization. For the winsorization, 3%-97% was used to remove extreme values. A Butterworth low-pass filter with a 10 Hz cutoff was applied. Signals were then downsampled (Table 3.3) to reduce dimensionality and decrease number of learning parameters, which helped reduce computational resource consumption. Finally, signals were normalized using a min-max normalization function. To prepare the data for temporal modeling, 60 second sliding windows were created for each signal with 30 second stride. Although, choice for stride length differs

from FB models, and goes against Kreibig's findings [35], it was validated by comparing performance metrics under different stride lengths, showing no significant impact on the effectiveness of the E2E models.

Table 3.3: Original and Downsampled Sampling Frequency F_s for each modality

Modality	Original F_s	Downsampled F_s
ECG	700 Hz	70 Hz
ACC (RespiBAN)	700 Hz	70 Hz
RESP	700 Hz	70 Hz
BVP	64 Hz	64 Hz
EDA	4 Hz	4 Hz
TEMP	4 Hz	4 Hz
ACC (E4)	32 Hz	8 Hz

3.3.1.1. Feature-extraction

For FB models, feature extraction was used to reduce the initial set of 92 features from the seven biophysiological signal modalities (Table 3.4) to a reduced feature set that excluded less relevant features. This feature selection process aimed to identify the most informative and discriminative variables that could effectively predict anxiety levels.

First, after the features were calculated for each of the modalities, a Spearman correlation matrix was computed, to understand the relationship between the features and label. A heatmap of the correlation matrix was then created to help with quick visualization of the results (Figure 3.3). Based on the correlation analysis, important features were selected and prioritized for model training.

Table 3.4: Features calculated for each biophysiological signal.

Signal	Feature	
ACC	Mean, stdv, min, max, abs integral of each x, y, z-axis and of norm. Peak freq of each axis	Stdv: Standard deviation, Min: minimum value, Max: maximum value, Net: Magnitude or length. Peak freq: highest freq domain component
BVP	Mean, stdv, min, max, Peak freq	
ECG	Mean, stdv, min, max. bpm, ibi, sdnn, sdsd, rmssd, pnn20, pnn50	bpm: beat per min; ibi: interbeat interval, sdnn: stdv of ibi, sdsd: stdv of ibi diff, rmssd: rms of ibi, pnn20: % of successive beats with more than 20ms diff,
EDA	Mean, stdv, min, max of each signal, SCL, and SLR. Slope, and drange	SCL: Skin Conductance Level, SLR: Skin Conductance Responses, drange: dynamic range.
EMG	Mean, stdv, min, max, drange, abs integral	
RESP	Mean, stdv, of each signal, inhalation, and exhalation. i/e ratio, and resp rate	i/e: inhalation to exhalation ratio. Resp rate: respiration rate.
TEMP	Mean, std, min, max, drange, slope	

Second, the selection criteria involved setting a threshold for correlation coefficients. This step aimed to streamline the analysis and focus on the most impactful variables. The specific features to be excluded were determined based on their relevance to the research questions and their potential contribution to the models' predictive power. This was done in a 2-step process. The first was to look at the features of the same modality that were too closely correlated with each other. Features of the same signal with correlation higher than 0.95 were eliminated, where the feature with higher correlation to anxiety, was kept. The second step was looking at the feature correlation with anxiety label and eliminating the ones with less than

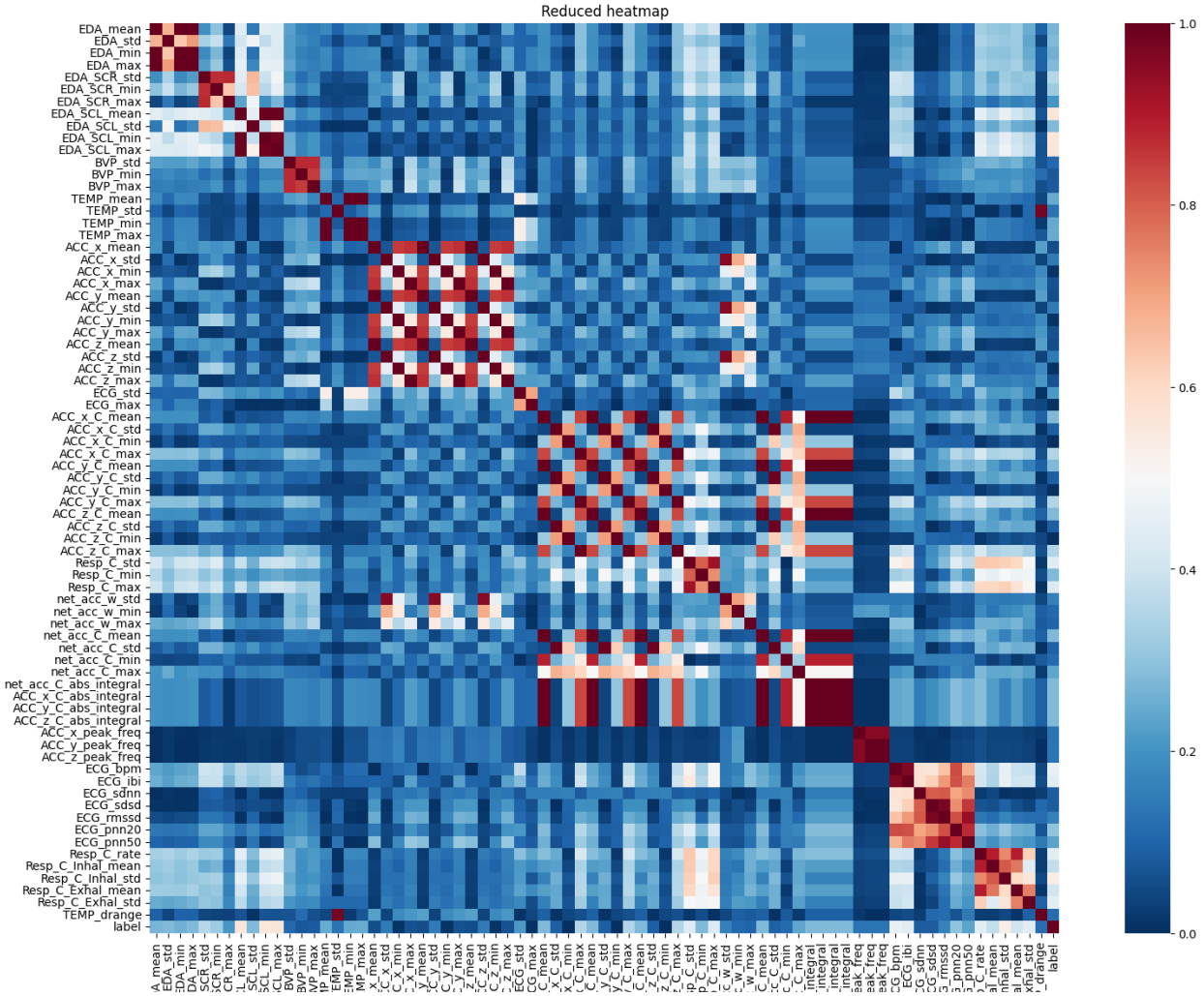


Figure 3.3: Spearman correlation between all features, including labels. Features of some modalities are too closely correlated with each other to warrant inclusion. For example, for EDA features, EDA mean, min and max are almost identical, unlike EDA std. In that case, only one of the 3 features was kept.

0.00009 difference as can be seen in Figure 3.4. This reduced the feature space from 92 features down to 54.

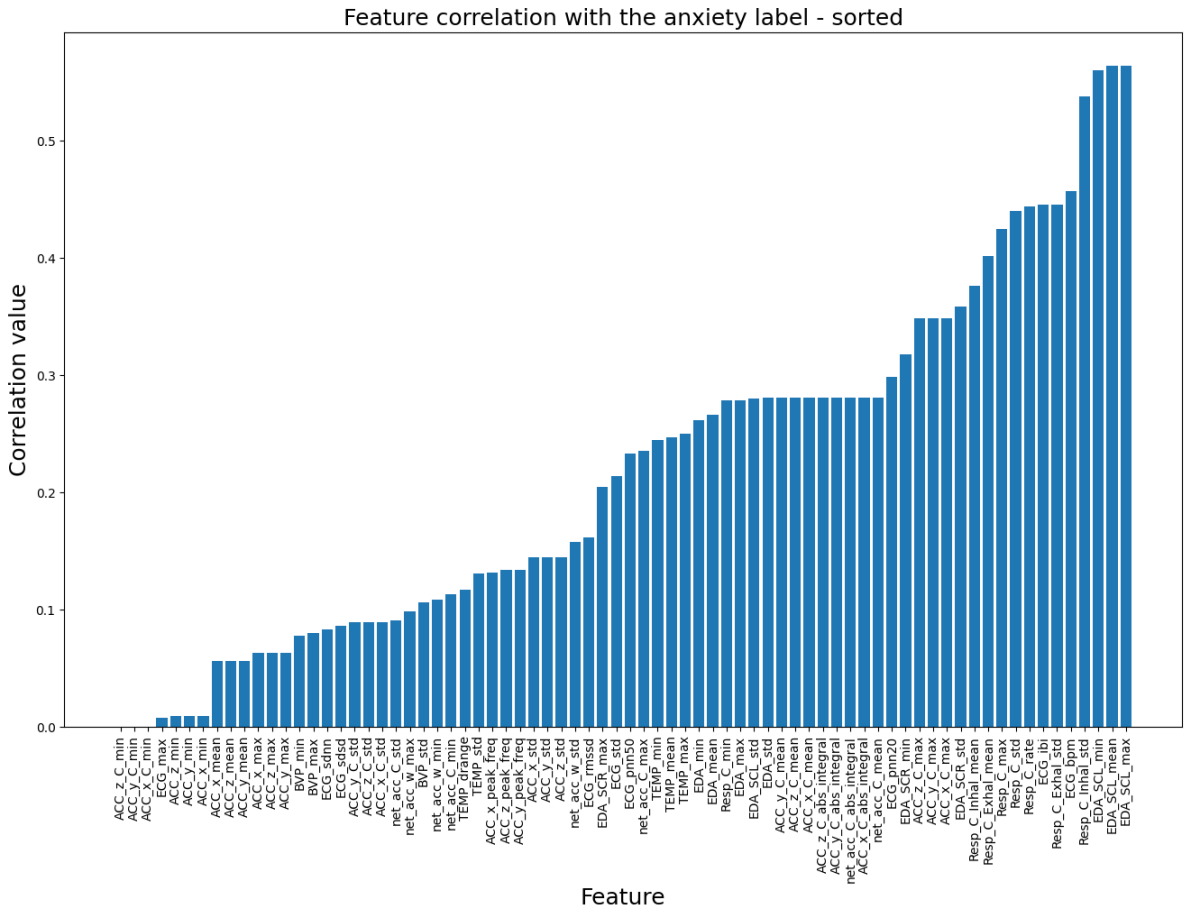


Figure 3.4: All 92 features and their correlation with the anxiety label. This is used to reduce the number of features to eliminate redundancy. For example, $ACC_{x\ mean}$ is too similar to $ACC_{Y\ mean}$ and $ACC_{Z\ mean}$.

3.3.2 Feature and Modality Analysis

The primary goal of this step was to determine the most effective features and modalities for anxiety detection in environments with variable noise levels. The analysis commenced by evaluating the performance of individual features and modalities to determine their contribution under different noise conditions. This involved a detailed examination of feature importance scores derived from feature-based models to pinpoint which modalities captured the most relevant information. Subsequently, the analysis identified which features and modalities maintained their discriminative power even in the presence of high noise levels, suggesting their suitability for real-world applications. Additionally, we investigated how the distribution of

feature values changed with varying noise levels, which provided insights into the stability and robustness of each feature, helping to identify those less sensitive to noise-induced variations.

3.3.3 Comparative and Conclusive Model Analysis

This section evaluated both FB and end-to-end models under various noise conditions to determine which approach was more suitable for deployment in noisy, real-world environments. The performance of each model was rigorously tested across a spectrum of simulated noise conditions, with metrics such as accuracy, precision, recall, and F1-score calculated to assess efficacy. The comparative analysis highlighted models that not only performed well across different noise levels but also demonstrated significant resilience to environmental disturbances. Based on these findings, specific models were recommended for practical application in anxiety detection systems, paving the way for robust, real-world deployments.

3.3.4 Machine Learning Models

For our implementation, we combined the traditional FB approaches as well as the state-of-the-art E2E models in this study.

3.3.4.1. Feature-based Models

Seven FB models were examined. Following the literature and Schmidt's [139] choice for FB models, we included Decision Tree (DT), Random Forest (RF), Adaboost (AB), Linear Discriminant Analysis (LDA), and k-nearest neighbor (kNN). Support Vector Machine (SVM) was included as it is one of the most commonly used algorithms in the literature (Chapter 2). XGBoost (XGB) was also added as an improvement over existing implementation of boosting algorithms, such as Adaboost, as it has been shown to perform better [246]. It is worth noting that our execution of feature calculation differs from what the Schmidt group did in their paper; this was due to the closed-source nature of their code. They did mention which features they calculated, but do not specify the mathematical implementation to extract said features. All the FB models were created using scikit-learn 1.2.1 ML library [250].

a. FB Model Training and Testing

To train and evaluate each FB model, models were first initialized for each of the chosen algorithms, calculating performance metrics and feature importances directly during model evaluation. The training process involved setting up specific models, such as SVM, using the scikit-learn's fit method. These models were then trained on the designated training data and employed to make predictions on the testing set. Comparisons of these predictions with the actual labels facilitated the calculation of key performance metrics including accuracy, precision, recall, and F1 score.

Further, we generated a feature importance list for each model, which ranked the features by their influence on the model's predictions. This provided critical insights into which variables significantly impacted the detection of anxiety. Detailed results for each model's performance, along with the top seven influential features, were systematically documented to underscore the predictive capabilities and strengths of each model within the study.

3.3.4.2. End-to-End models

To maintain the goal of producing a comparative study and benchmark with the field, we used the same ten E2E models utilized by Dziezyc et al. [193] (Table 3.2). Although these models do not require feature-engineering, they are computationally intensive and require significantly longer to train and store, compared to classical FB models.

3.3.5 Model Training and Testing

The training was performed on the three classes contained in WESAD (baseline, amusement, stress), aiming to provide the models with a comprehensive understanding of the data's nuances. This approach enriches the model's capability to discern subtle differences, enhancing its performance when simplified to a binary classification of non-stress (baseline + amusement) versus stress for testing [240], [251]–[253]. A 5-fold cross-validation method was utilized (Figure 3.5). The dataset was randomly divided into equal five test sets. For each of the test sets, the rest of the data were split into training and validation set. For each of the architectures, 5-fold cross validation, and 5 training iterations were performed. Performance metrics were then averaged and reported. Finally, using the train_test_split function in scikit-

learn [250] , the preprocessed data were split into training (80%) and testing (20%) sets, with the training set used to train the machine learning models, and the testing set used to evaluate model performance on unseen data. Finally, the preprocessed data were split into training and testing sets using an 80/20 ratio using scikit-learn's train_test_split function, where the 80% of the WESAD dataset was used to create the training set that was used to train the machine learning models, while 20% of the WESAD dataset was used to create the testing set was used to evaluate the models' performance on unseen data.

Figure 3.5: Visualization of the 5-fold cross validation method implemented. The dataset was split into 5 folds. Each of the folds was then chosen as the test set for that fold, and the rest of the data were then split into training and validation set.

F_1-score and accuracy were calculated based on the following equations:

$$Accuracy = \frac{True\ Positive + True\ negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall\ = \frac{True\ Positive}{True\ Posive\ + False\ Negative}$$

$$F_1 - score = \ 2\ \times \frac{Precision\ \times Recall}{Precision + Recall}$$

3.3.6 Simulating real-world data using noise addition

In the quest to simulate real-world conditions for anxiety detection using wearable technology, this study explored the application of synthesized noise to the WESAD dataset. This approach aimed to enhance the robustness and applicability of machine learning models by evaluating their performance under simulated conditions that reflect the complexity of real-world scenarios.

Recognizing the ubiquity and impact of environmental noise on data quality, the decision was made to employ Gaussian noise as the primary method for this simulation. The Gaussian noise model was chosen for its ability to represent a wide range of common noise types encountered in everyday settings, thereby ensuring that the generated synthetic dataset possessed a realistic degree of variability and complexity.

We implemented a Gaussian noise function to generate noise. To not introduce bias to our models, we used a zero average Gaussian function with a fixed standard deviation, meaning that we used a zero-mean homoscedastic Gaussian function. The driver for this function was the signal to noise ratio (SNR), that drove our choice for standard deviation. We ran the models for the following ten pre-defined values, SNRs = {0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6}. For a given SNR, the standard deviation σ was:

$$\sigma = \frac{E(S^2)}{SNR},$$

where *E(S)* is the expectation of signal S. Thus, the modified Gaussian noise WESAD value for a given signal S can be expressed as:

$$S_{GN} = S + N(0, \sigma^2)$$

where $N(0, \sigma^2)$ is the Gaussian function.

For each SNR, noise was added to the WESAD dataset, generating a noisy set, SNR_i. Each SNR value was generated five times, to reduce results variance for each of the SNRs. This resulted in 100 datasets, all of which were then tested on the FB models. Because of the significant limitations associated with running the large E2E models (high computation time and cost), the results of the FB models were analyzed to assess if the number of datasets could be

reduced. Results will be discussed in the following section, but based on the FB model results, the SNRs for E2E models were reduced to only four cases, SNRs = {0.01, 0.1, 0.15, 0.4}. All of the cases are shown in Table 3.5.

Table 3.5: Signal-to-noise values used to generate noisy WESAD datasets for FB and E2E models.

Approach	SNR	Samples per SNR
FB	0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6	10
E2E	0.01, 0.1, 0.15, 0.4	1

To provide further insight into the performance characteristics of each model as a function of noise, violin plots [254] were generated to investigate behaviors across various conditions (e.g. Figure 3.9). Violin plots are particularly useful for this analysis because they combine elements of box plots and kernel density plots, presenting a comprehensive view of the data distribution. These plots illustrate not only the median and interquartile ranges—commonly depicted in box plots—but also the probability density of the data at different values, represented through the thickness of the violin shape.

Violin plots were created to explore the robustness of feature-based and end-to-end models under different levels of synthetic noise. This visualization method enables a clear comparison of how performance metrics such as F1-score and accuracy are distributed, highlighting potential skews or multiple modes in the data that are not immediately apparent in more traditional plots. By examining these plots, readers can discern which models maintain a narrower distribution of performance metrics—indicative of stability—and which models show wider distributions, suggesting variability in robustness to noise.

3.4. Results

3.4.1 Baseline

The initial phase involved replicating the results of Feature-Based and End-to-End machine learning models on the original, unaltered WESAD dataset to establish a baseline. For reference, using FB models, Schmidt et al. detected anxiety with an F_1-score of 0.91 and an accuracy of 0.92 for binary classification (stress vs. non-stress) using all modalities with the LDA model [139]. Using E2E models, Dziezyc et al. detected anxiety with an F_1-score of 0.73 and an accuracy of 0.79 using the FCN architecture [193]. For context, a random guess would

yield a 0.50 accuracy and 0.48 F_1-score, while a weighted guess would result in 0.30 accuracy and 0.42 F_1-score, as shown in class distribution in Table 3.1.

3.4.1.1. FB models

For FB models, XGB and DT performed the best with accuracy of 0.99 and F_1-score 0.99 for both models (Table 3.6).

Table 3.6: Accuracies from the original WESAD dataset for the 7 FB models used in this study.

Model	Accuracy
DT	0.99
RF	0.91
LDA	0.93
KNN	0.94
AB	0.81
SVM	0.95
XGB	0.99

After evaluating the FB model performance results, features were explored for each of the models to determine importance of features, as well as importance of modalities. As can be seen in Table 3.7, the Skin Conductance Level (SCL) max feature from EDA was the feature of the highest importance for six of the seven models. This result led to EDA being defined as the most important modality for anxiety detection, and more specifically SCL max. The second most important feature was BVP max and ranked number 2. Both of these findings are somewhat surprising, since they do not agree with previous SOTA results (e.g., [130], [139], [193]). In Table 3.7, the columns categorize and quantify the contributions of different physiological signals and their specific features to the model's performance. The Weighted average is a crucial metric, calculated by squaring each feature's importance score and dividing this total by the sum of the top 7 feature importance scores, expressed as:

$$W = \frac{\sum_{i=1}^{n} {X_i}^2}{\sum_{i=1}^{n} X_i}$$

$$W = weighted\ average$$

$$n = number\ of\ features$$

$$X_i = Feature\ importance$$

where X_i represents the *i*th feature importance value listed in Table 3.7 for a given modality, and n represents the total number of feature importance scores listed in Table 3.7.

This method emphasizes the impact of more influential features by assigning greater weight to higher scores, providing a nuanced view of how each feature impacts the model's ability to detect anxiety. This table format effectively showcases which physiological measurements are most impactful under varying conditions, highlighting the features that consistently influence anxiety detection outcomes.

Table 3.7: Feature importance for each of the FB models trained on the original WESAD dataset. Table lists top seven features per FB model and each feature's contribution to unity.

Modality	Feature	Weighted average	DT	kNN	LDA	XGB	SVM	RF	AB
ACC	$ACC_{x\ C\ mean}$	0.06	0.03						
	$ACC_{x\ min}$		0.03				0.07		0.12
	$ACC_{x\ std}$					0.05			
	$ACC_{net\ w\ min}$		0.04				0.06		
	$ACC_{net\ w\ max}$						0.05		
BVP	BVP_{max}	0.15		0.39					
	BVP_{min}			0.38					
	BVP_{std}			0.14					
ECG	ECG_{bpm}	0.10	0.09	0.14		0.07	0.05		0.26
	ECG_{pnn50}					0.05			
	ECG_{rmssd}			0.10					
	ECG_{sdnn}			0.11					
	ECG_{sdsd}			0.03					
	ECG_{std}		0.13		0.06	0.04		0.06	0.10
EDA	EDA_{SCL_max}	0.18	0.43		0.15	0.15	0.21	0.15	0.30
	EDA_{SCR_max}				0.06			0.06	0.02
	EDA_{SCR_min}				0.06			0.06	
	EDA_{SCR_std}				0.08			0.08	
	EDA_{std}					0.04			0.02
RESP	$Resp_{C\_Exhal\_std}$	0.08			0.08	0.10		0.08	
	$Resp_{C\_Inhal\_std}$				0.06		0.09	0.06	
TEMP	$TEMP_{mean}$	0.08	0.11				0.05		0.04

Figure 3.6 illustrates the feature importance for both XGB and DT, highlighting how certain features significantly influence model outcomes. This visual representation aids in understanding why these models outperformed others, and it substantiated our focus on refining these specific models to enhance their predictive power and practical utility in clinical and everyday applications.

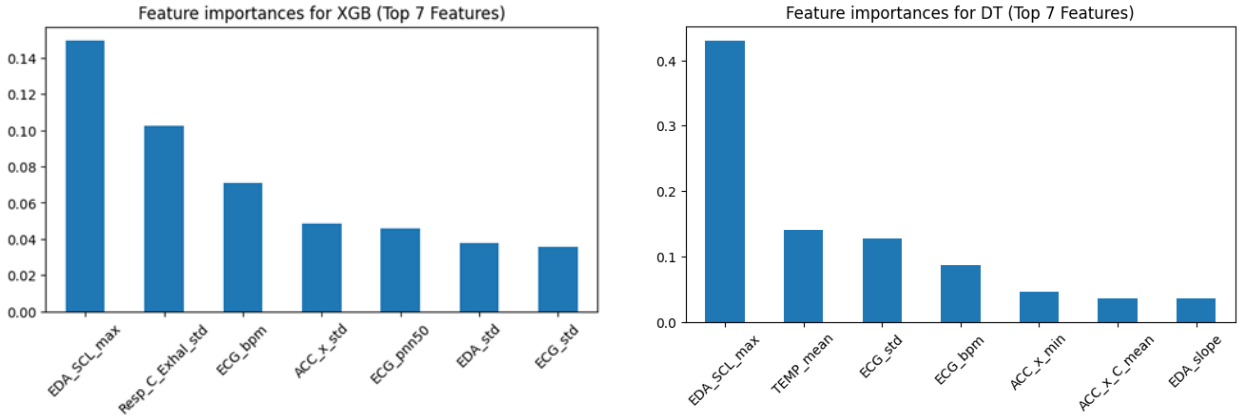


Figure 3.6: Top 7 features for the two highest performing classic FB models on original WESAD dataset: XGB and DT. Y-axis is Feature Importance, where summation for all features is unity

3.4.1.2. E2E Models

Results for the E2E models is shown in Table 3.8. Performance of models was consistent with that observed in Dziezyc et al. [193]. The top performing models were found to be FCN and Resnet. Performance for E2E models was reported as an average between all 5-fold cases, which was reported as Average Accuracy and Average F_1-score (Table 3.8). We also examined the maximum performance of each architecture and present the Accuracy for that case (Table 3.8, last column).

Table 3.8: E2E model performance results for original WESAD dataset by averaged Accuracy and F1-score and Accuracy over the number of models ran, for best performing dataset, according to Dziezyc's method

E2E Model	Average Accuracy (std)	Average F_1-score (std)	Accuracy (max)
FCN	0.79 (0.03)	0.75 (0.04)	0.95
Resnet	0.80 (0.05)	0.74 (0.07)	0.96
Time-CNN	0.76 (0.03)	0.67 (0.04)	0.89
MCDCNN	0.74 (0.03)	0.65 (0.05)	0.89
MLP-LSTM	0.72 (0.02)	0.60 (0.03)	0.89
Encoder	0.69 (0.04)	0.59 (0.05)	0.89
MLP	0.69 (0.01)	0.59 (0.02)	0.93
CNN-LSTM	0.69 (0.02)	0.54 (0.02)	0.85
Inception	0.65 (0.07)	0.52 (0.07)	0.91
Random guess	0.50	0.50	
Majority class	0.53	0.23	

3.4.2 GN Results

3.4.2.1. FB Models

FB models have demonstrated a robust ability to detect anxiety in controlled settings, leveraging well-defined physiological signals and engineered features for effective performance assessment.

The influence of signal-to-noise ratio on the performance of FB models reveals a nuanced impact on their ability to discern stress-related patterns. As SNR decreases, indicating higher noise levels, there was a gradual reduction in model accuracy (Figure 3.7). Although XGB and DT upheld their superior performance observed when using original WESAD dataset, the AB model showed a notable decline in efficacy under higher noise conditions, diverging from the other FB models. This variation in performance can be expected due to the different ways FB models process and interpret data. For instance, models like XGB are designed to handle various types of data irregularities and have mechanisms that can effectively deal with noise to some extent. In contrast, models such as AB might be more sensitive to noise, particularly if the noise disrupts the patterns they rely on for decision-making. Consequently, as the SNR decreases, it is expected that the performance of models would vary, reflecting their individual capacities to filter out noise and maintain accuracy.

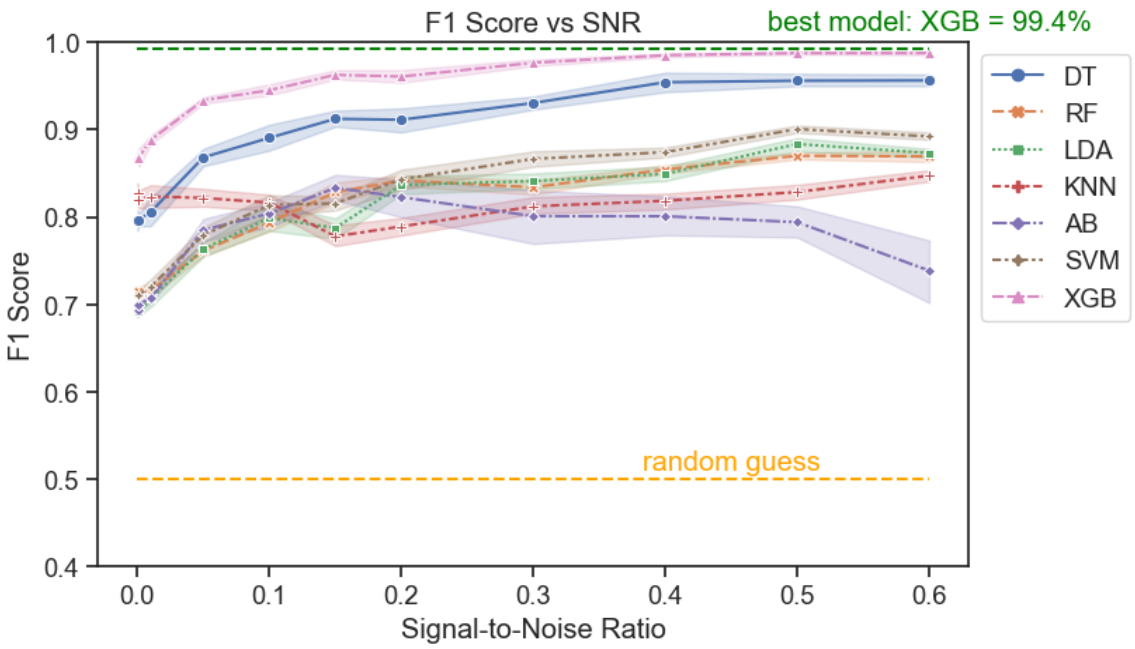


Figure 3.7: Performance of FB models across various noise levels. Dashed orange line is the random guess baseline performance, and the dashed green line is the best performing model with no noise. Performance degraded with increased noise (smaller SNR) but with varying sensitivity to noise, as expected. Error bands display 5% confidence interval for around the plotted mean.

This trend underscores the importance of optimizing noise handling capabilities within FB models to ensure their effectiveness in real-world applications where noise variability is common With noise, performance degraded proportionally with increasing noise (smaller SNR value). These results were consistent for all models, except for AB Another finding was that XGB outperformed DT and maintained the performance advantage across the range of SNR tested, solidifying its position as the best performing model across conditions tested.

With respect to feature importance, it was observed that feature importance ranking was also relatively consistent across SNR. For example, looking at the feature importance ranking for XGB across different SNR in Figure 3.8, EDA_{SCL_mean} continued to be the top feature across all SNR.

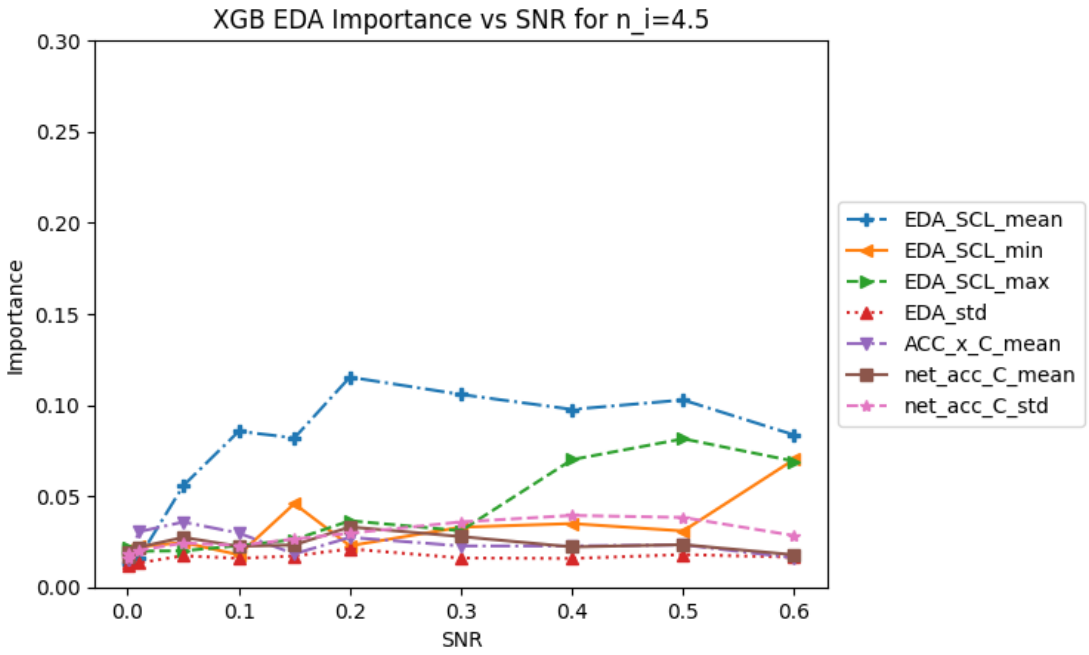


Figure 3.8 Feature importance of the top 7 features for the XGB model across SNR. EDASCL mean maintained its dominance.

Violin plots in Figure 3.9 offer a visual exploration of the FB models' performance distribution across different levels of SNR, illustrating the variability and robustness of each model amidst varying noise conditions. These plots combine the features of box plots and density plots, showing not only the median of the F1 scores but also the density and spread of scores around the median. This visual analysis is particularly insightful as it highlights the sensitivity of models like XGB and DT to different noise levels, which is not immediately apparent from median F1 scores alone. The plots emphasize that, while performance generally declines with increased noise, some models show a tighter distribution of F1 scores, suggesting a more consistent performance regardless of noise level. Conversely, models exhibiting wider distributions imply greater variability in performance, which could indicate less reliability in noisy conditions.

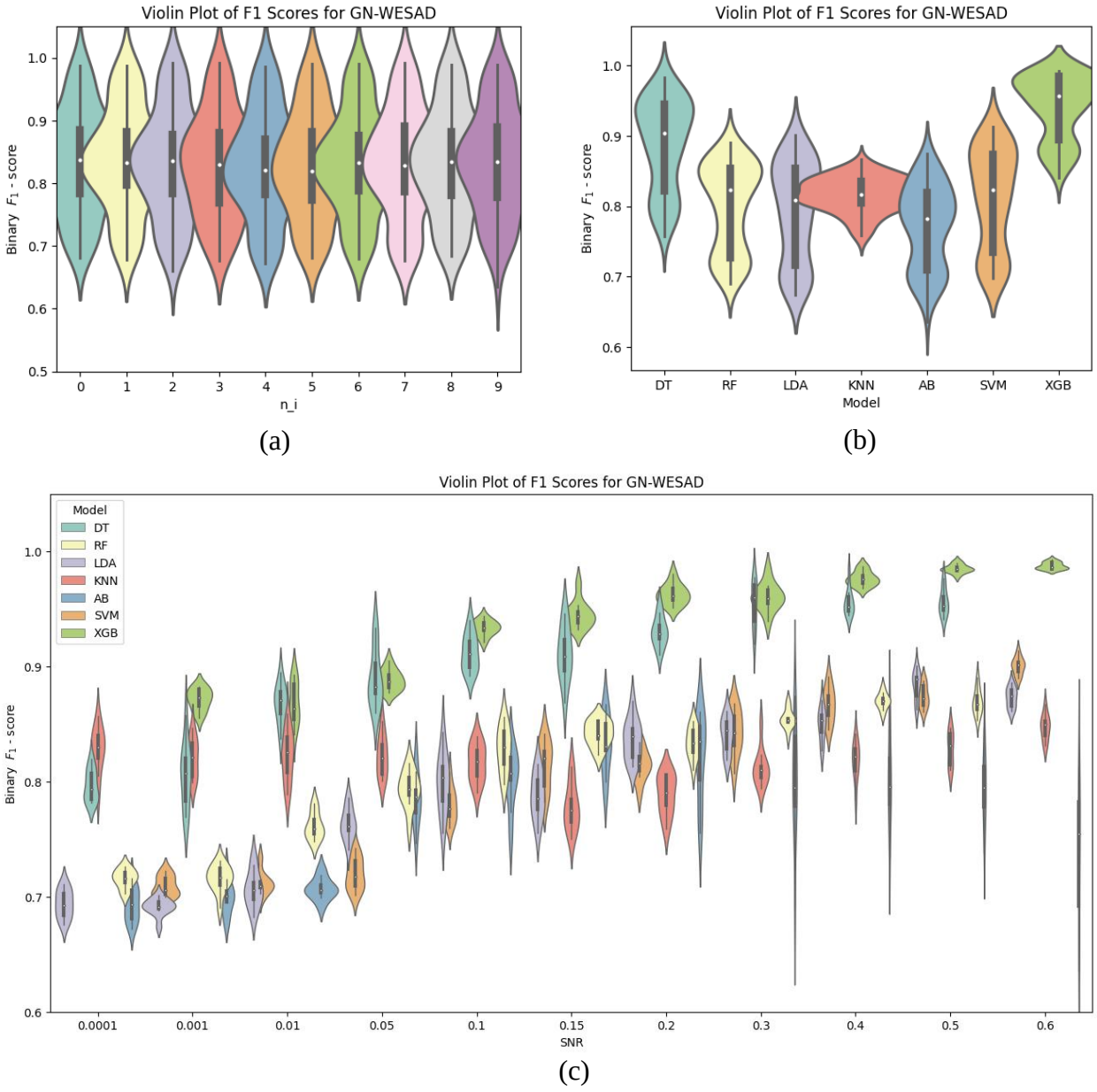
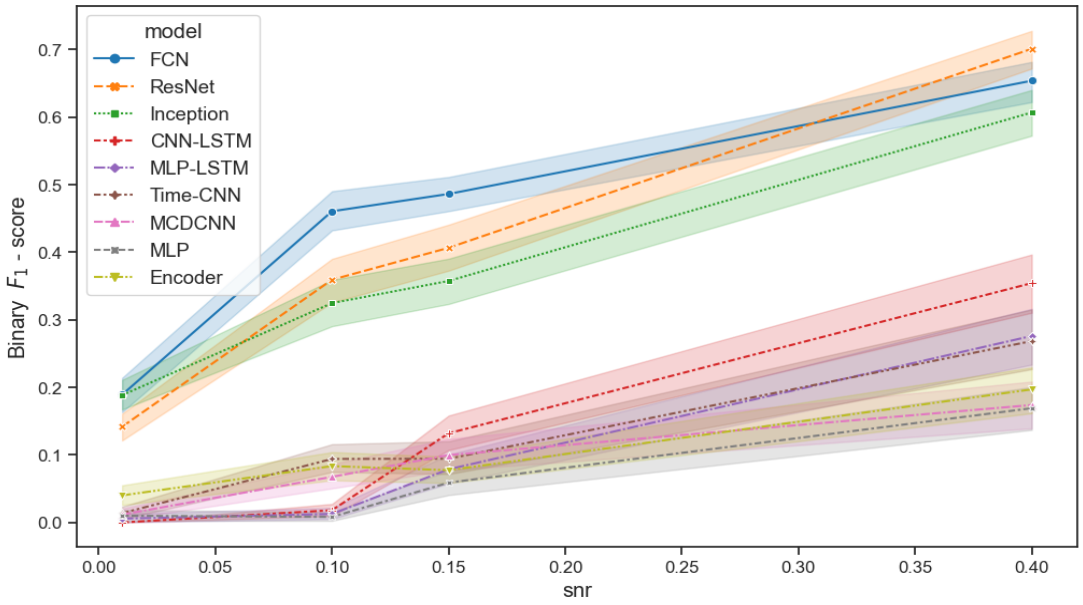


Figure 3.9: Violin plots of mean F_1-score for FB models for the binary case using WESAD dataset with Gaussian noise. Performance distribution across (a) sampled noise, (b) FB models, and (c) SNR by FB model.

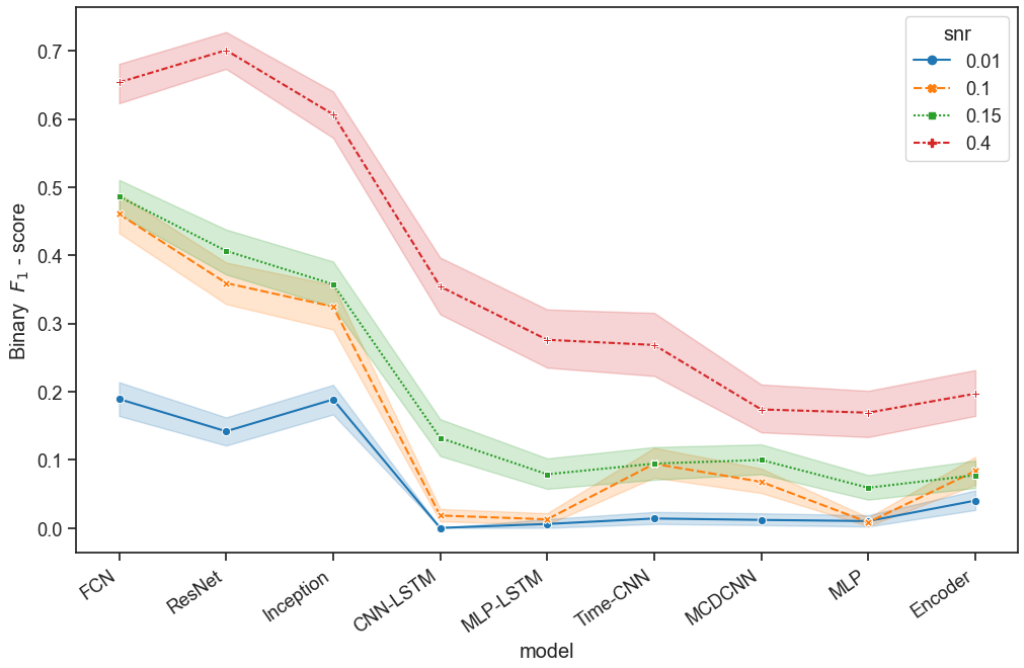
3.4.2.2. E2E Models

With the noisy dataset, performance of all E2E models significantly degraded (Figure 3.10(a), Table 3.9). FCN and Resnet achieved mean F_1-scores of 0.75 (standard deviation, 0.24) and 0.74 (0.22). Noteworthy changes were that Resnet outperformed FCN in the SNR=0.4 case,

but gets over taken again at higher noise levels (lower SNR) (Figure 3.10(b)). At higher noise levels, all models underperform random guess for the binary case (i.e., F_1-score of 0.5).



(a)



(b)

Figure 3.10: F_1-score for E2E models by (a) SNR and (b) model architectures. Error bands display 5% confidence interval for around the plotted mean.

From the violin plots in Figure 3.11, the detailed view of the distribution of each of the model's performance as a function of noise, provides insight. The top tail of the violin plots indicates that a significant portion of each of the model population performing quite well. So, perhaps, it would be adequate to look at the top 5% percentile, for a few reasons. The first is that even though reporting mean performance is indicative of how each architecture performs overall, in this case, a strategy similar to that of evolutionary algorithms is chosen, where the top models are chosen and evaluated and kept. At the highest noise level of 0.01, the only model that had better performance than random guess was the FCN model.

Table 3.9: Mean F_1-score of E2E models with different levels of noise

Architecture	*SNR = 0.01*	*SNR = 0.1*	*SNR = 0.15*	*SNR = 0.4*	*baseline*
FCN	0.19	0.46	0.49	0.65	0.75
Resnet	0.14	0.36	0.41	0.70	0.74
Time-CNN	0.01	0.09	0.09	0.27	0.67
MCDCNN	0.01	0.07	0.10	0.17	0.65
MLP-LSTM	0.01	0.01	0.08	0.28	0.60
Encoder	0.04	0.08	0.08	0.20	0.59
MLP	0.01	0.01	0.06	0.17	0.59
CNN-LSTM	0.00	0.02	0.13	0.35	0.54
Inception	0.19	0.32	0.36	0.61	0.52

Moreover, looking at Figure 3.12, provides some insights into the how the top 5% percentile paints a clearer picture of performance as function of noise. Both MLP variants failed to perform better than a random guess below SNR of 0.1. Another interesting behavior was that ResNet, above 0.1, maintained a specific level even with a significant increase of noise. The two best performing models maintained their edge over random guess. ResNet broke down at SNR 0.01, while FCN maintained its edge over random guess.

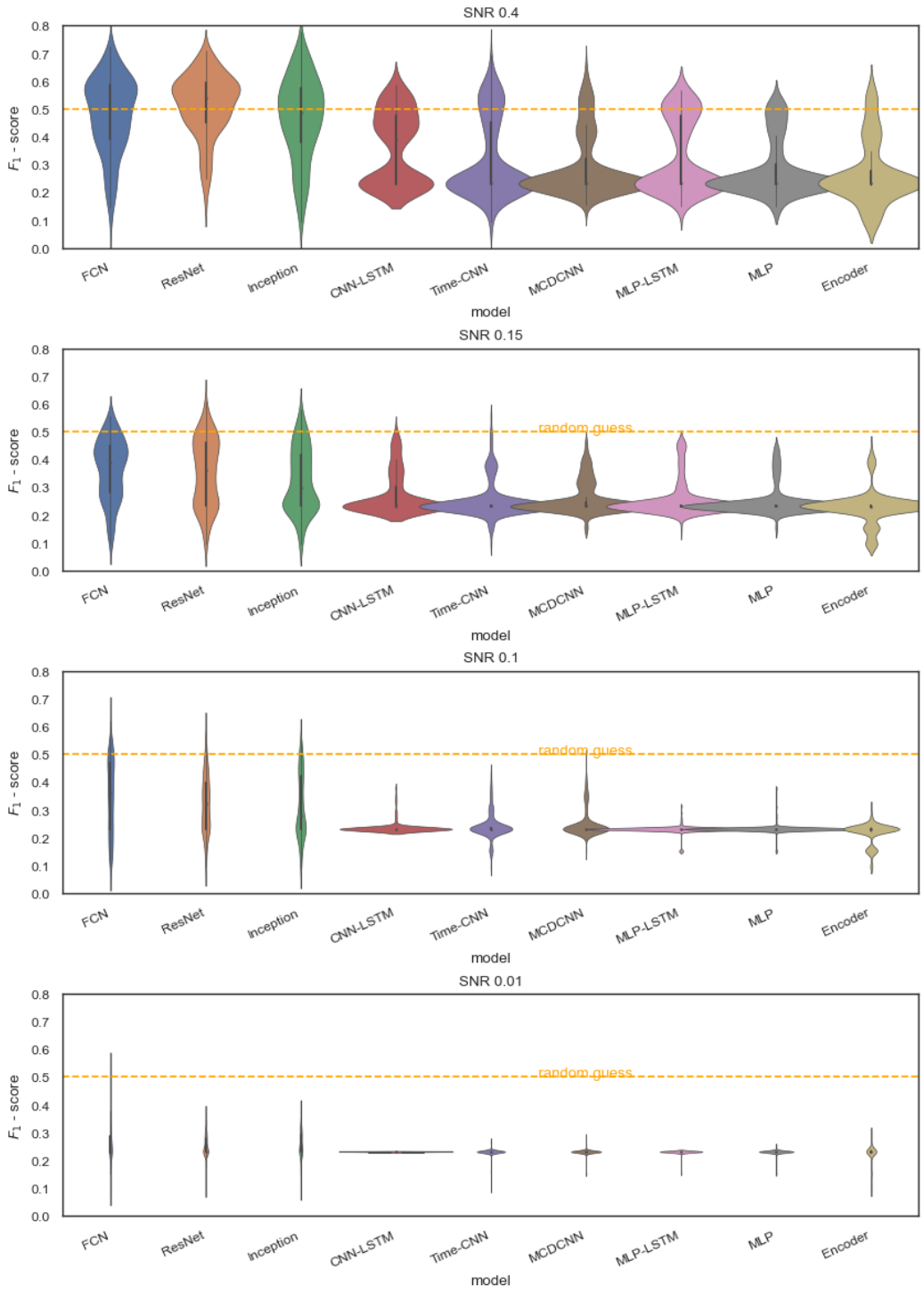


Figure 3.11: Violin plots of mean F_1-score for E2E models for the binary case. Results are separated by SNR for each of the subplots.

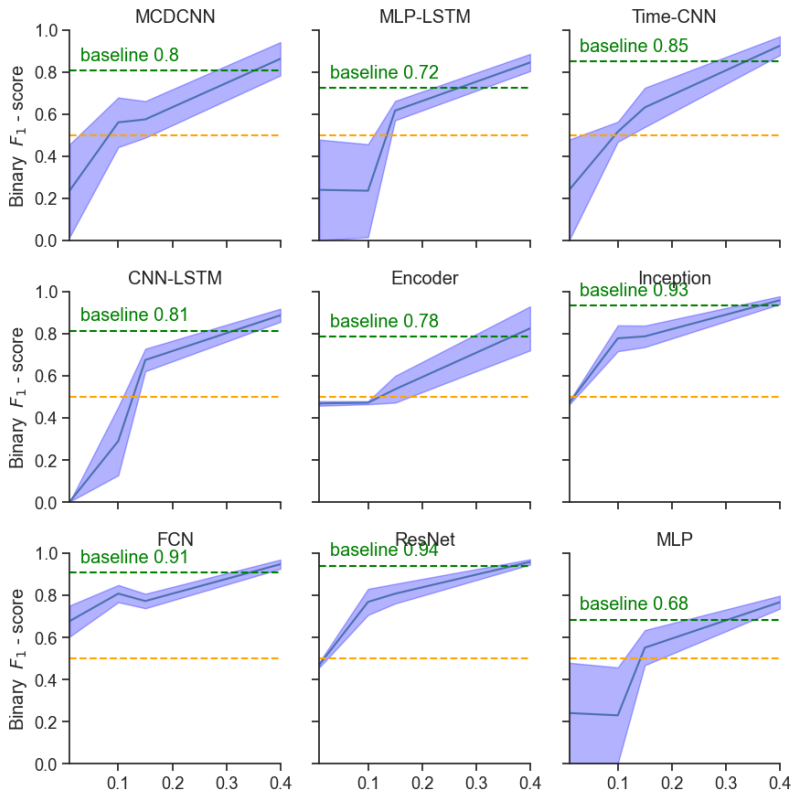


Figure 3.12: F_1-score for E2E by SNR. To evaluate model performance more closely, top 5% percentile F_1-scores were evaluated for each model across SNR to assess how performance changed. The solid line is the midpoint and the cloud is the upper and lower bands of the 95%-100% percentile range.

3.5. Discussion

This study provides a comprehensive analysis of the resilience of both feature-based and end-to-end models to environmental noise, revealing significant performance disparities. FB models, particularly XGB, demonstrated superior robustness across all noise conditions, maintaining higher accuracy levels compared to E2E models (Figure 3.6 and Figure 3.7). This finding supports our initial hypothesis that some models would exhibit enhanced resistance to noise, highlighting the effectiveness of FB models in noisy real-world scenarios.

To provide further insight into performance characteristics of each model as a function of noise, several violin plots were generated to investigate behaviors for FB models (Figure 3.9). There are several noteworthy observations. First, all FB models performed consistently across different SNR samples, yielding a uniform distribution across the samples (Figure 3.9(a)), except for AB. Even though kNN was not the highest performer, it maintained a consistent performance level, and even increased in performance at higher noise levels, which was a unique behavior

compared to the rest of the models. Finally, even though XGB and DT performed closely, performance distribution for XGB was tighter than that of DT, solidifying it as the best choice under given conditions (Figure 3.9(b)).

EDA appeared to be the most important modality with $EDA_{scl\ max}$ being the most important feature (Figure 3.8). Even though EDA was the most important modality, it may be of concern, since skin conductance can easily be disrupted in the presence of physical activity, or sweat-inducing warm weather. BVP can be used as a supplementary signal to support model detection in the case where EDA quality is reduced.

For E2E models, it was initially surprising to see that non-recurrent models outperformed recurrent models, but then it was realized that the relatively short-time frame associated with anxiety can be captured well enough with the convolution layer of the CNN models that it was able to preserve some contextual information from the recent past, providing an equivalence to the memory cell in RNN networks. Conversely, E2E models exhibited a significant reduction in performance with the introduction of noise, with FCN outperforming other models under these conditions yet still falling short of the robustness displayed by FB models (Figure 3.10). ResNet, while initially performing well, saw a decline in effectiveness at lower SNR levels, underscoring the limitations of E2E models in handling noisy data.

The consistency of dominant features in maintaining their predictive power, even in noisy environments, was evident across both model types (Figure 3.8). This consistency challenges the expectation that E2E models would inherently handle noise better due to their capacity to learn complex patterns directly from raw data. The proportional performance decline of E2E models with increased noise necessitates a reevaluation of their advantages over FB models in noisy settings.

These observations underscore the importance of feature engineering and model selection in developing robust anxiety detection systems for real-world applications. They prompt further investigation into the fundamental characteristics of machine learning models and the nature of the data they process, shifting focus towards a deeper understanding of how different models and features interact with environmental noise. As we advance our research, exploring these

dynamics will be crucial for enhancing the reliability and efficacy of wearable technology in monitoring anxiety in diverse and challenging conditions.

3.6. Future Work

Building on the findings of this study, several avenues for future research have been identified. These areas not only promise to extend the knowledge base but also aim to address the challenges and limitations encountered in the current study. The future work should focus on:

Advanced Noise Modeling: While this study incorporated Gaussian noise, future research could explore different noise types such as noise modeled after motion artefacts, sensor-specific or device-specific noise. These could better mimic the complex and unpredictable nature of real-world environments, providing a more rigorous testbed for anxiety detection algorithms. This provides insights into the generalizability of the models to various real-world noise scenarios.

Model Optimization and Enhancement: Investigating methods to enhance the robustness of existing models against noise is a critical next step. This could involve developing new algorithms or refining existing ones to better handle noisy data, potentially through advanced signal processing techniques or novel machine learning approaches.

Cross-Dataset Validation: Testing the models on different datasets, especially those collected in real-world settings, would help validate the generalizability of the findings and the robustness of the models across diverse populations and environments.

By addressing these areas, future research can significantly advance the field of anxiety detection using wearable technology, leading to more effective, user-friendly, and widely applicable solutions.

3.7. Conclusions

This study embarked on an exploratory journey to understand the impact of noise on the performance of machine learning models in the context of anxiety detection using wearable

technology. Through experimentation and analysis, several key insights emerged, reshaping our understanding of model robustness in noisy environments.

First, the study revealed that both FB and E2E models decline in performance as noise levels increase. Some models are more negatively affected by increased noise than others, which confirmed the hypothesis that some models will be better suited under noisy conditions than others.

Second, the consistency in the performance of dominant features across both noise-free and noisy conditions highlights the robust nature of these features. This observation is crucial for future research and development in the field, as it underscores the importance of identifying and leveraging such robust features in the design of anxiety detection systems.

Furthermore, the study's findings on end-to-end models provide a critical perspective on their perceived advantages. While these models are celebrated for their ability to capture complex patterns beyond engineered features, their performance in the face of noise was not better than feature-based models and were actually worse. This insight calls for a reevaluation of the strategies employed in developing anxiety detection systems, particularly in terms of model selection and feature engineering.

In conclusion, this research contributes to the field of anxiety detection using wearable technology by offering a nuanced understanding of how environmental noise impacts model performance. It paves the way for future studies to delve deeper into the dynamics of noise and model robustness, ultimately leading to the development of more effective and reliable anxiety detection systems. As wearable technology continues to evolve, the insights gained from this study will be invaluable in guiding the creation of solutions that are not only technologically advanced but also resilient in the face of real-world challenges.

CHAPTER 4: ADVANCING ANXIETY DETECTION IN NOISY ENVIRONMENTS: EFFICACY OF MACHINE LEARNING MODELS ON REAL-WORLD DATASETS

4.1. Abstract

This chapter evaluated the performance of machine learning models for anxiety detection, focusing on feature-based (FB) and end-to-end (E2E) models using wearable technology in real-world conditions. The study utilized the RADWear and WEAR datasets, providing insights into the models' robustness and the specific challenges posed by diverse environmental noise. Despite the relatively poor performance of E2E models, FB models, particularly XGBoost and Decision Trees, demonstrated considerable resilience, maintaining higher accuracy and reliability across different noise levels. This investigation included an in-depth analysis of transfer learning, highlighting its potential and limitations in adapting models developed on standard datasets, like WESAD, to complex real-life scenarios. Moreover, the study analyzed the distributed feature importance across various physiological signals, such as electrodermal activity (EDA) and electrocardiogram (ECG), considering their vulnerability to environmental factors. It was found that integrating multiple physiological data types could significantly enhance model robustness. The results underscored the need for a nuanced understanding of signal contributions to model efficacy, suggesting that while FB models showed promise, the architecture and training of E2E models require optimization for better performance in practical applications.

4.2. Introduction

Anxiety disorders are among the most prevalent mental health issues worldwide, affecting millions and significantly impacting quality of life [12], [43], [255]. While traditional methods like self-reports and clinical interviews offer insights into anxiety levels, they are limited by subjectivity and intermittency [26]. The advent of wearable technology presents new possibilities for continuous, objective, and non-invasive monitoring of anxiety in real-world settings.

Prior research has demonstrated the potential of wearable devices for anxiety detection using various physiological signals, such as electrodermal activity (EDA), heart rate variability (HRV), and accelerometer data, as explored in Chapter 1. However, these studies are

predominantly conducted in controlled laboratory settings, which may not accurately reflect real-world conditions. Furthermore, the impact of environmental noise on model performance has been underexplored, as demonstrated in Table 2.5.

This study addresses these gaps by evaluating both feature-based (FB) and end-to-end (E2E) machine learning models on the RADWear and WEAR datasets under a variety of real-world conditions. We placed significant emphasis on distributed feature importance across different physiological signals, considering their specific failure modes which may affect detection accuracy. For example, EDA signals can be disrupted in wet or humid environments, acceleration data may be unreliable during physical activity, and ECG signals may be affected by conductivity issues, making them challenging to integrate into wearable devices [256]–[258]. BVP signals are also susceptible to disruptions caused by gaps between the device and the skin [70], [259], [260].

By investigating the robustness and reliability of these models in challenging environments, we aim to identify the most effective approaches for real-world application. This includes optimizing the selection of modalities to ensure accurate and efficient monitoring of anxiety.

The insights gained from this study could significantly advance the field of wearable technology for mental health monitoring, potentially revolutionizing how we approach anxiety detection and management, and paving the way for personalized interventions that enhance the quality of life for individuals globally.

The overarching goal of this study was to develop robust machine learning models for anxiety detection using wearable devices, focusing particularly on their performance in noisy, real-world environments. By leveraging the RADWear and WEAR datasets, which reflect a wide range of environmental conditions and participant demographics, we aimed to identify the most informative features and modalities for anxiety detection. Ultimately, we sought to streamline the number of modalities required for effective monitoring, ideally reducing them to a single, user-friendly device such as a smart wristband.

The study extends beyond the controlled laboratory setting, delving into real-world environments to assess the practicality and effectiveness of wearable technology for monitoring stress and anxiety. This strategy enhances both the validity and applicability of the findings, leading to potential personalized interventions based on individual stress response profiles. Chapter 1 introduced the RADWear and WEAR datasets, and Chapter 3 focused on several machine learning models that were trained, validated, and tested on a noise-inoculated WESAD dataset. To align with the RADWear and WEAR wearable hardware, the WESAD dataset was adapted by omitting EDA, TEMP, and EMG modalities from the chest device.

Self-learned models refer to models that have been trained and evaluated using the same dataset, without the influence of additional external data through transfer learning. This method provides a pure evaluation of model performance, establishing a foundational understanding of each model's effectiveness under controlled conditions.

Transfer learning is an approach that leverages knowledge from a pre-trained model to address a related problem with limited labeled data [261]. By adapting the pre-trained model to the new task, transfer learning reduces the need for extensive training data and computational resources [262]. It is particularly beneficial when the target task has limited labeled data, is similar to the pre-trained model's task, or when training from scratch is expensive or time-consuming [263]. Transfer learning has been successfully applied in various domains, including computer vision, natural language processing, and speech recognition [264], [265].

4.3. Methods

To preserve the benefits of utilizing a comparable dataset for transfer learning, data processing and feature extraction on the RADWear and WEAR datasets were performed using the same script that was used on WESAD. The RespiBAN data were downsampled to 70Hz from 700Hz, for all modalities. For Hexoskin, the modalities did not have the same sampling rate. For example, ECG was sampled at 256 Hz, while RESP and ACC were sampled at 128 Hz and 64 Hz, respectively. From the understanding of the dynamics of these measurements and our testing, these differences were negligible with respect to final system performance. More detail had been presented in Chapter 3.

4.3.1 The RADWear and WEAR datasets

The RADWear and WEAR studies were designed to analyze the impact of anxiety on two higher education student populations. The RADWear dataset captures the demanding nature of a medical school setting through data from medical students. For this study, the RADWear dataset included 9 participants, 4 males and 5 females with an average height of 167 ± 7.5 cm, weight of 79.4 ± 27 kg, and age of 27.3 ± 2.3 years. The WEAR dataset captures data from STEM-major university students. The WEAR dataset included 27 participants, 12 majority males, 0 minority males, 10 majority females, and 5 minority females with an average height of 161 ± 5 cm, weight of 67.5 ± 20 kg, and age of 23 ± 2.3 years. Ethnic and racial backgrounds were defined as Majority (Caucasian, Asian, mixed Asian/Caucasian) or Minority (Hispanic/Latin, African American, mixed Asian/African American, and Native American identities). The academic disciplines were broad, with fields of study ranging from Engineering to Animal Science.

For both studies, the E4 wristband and Hexoskin smart shirt were used to collect data. The E4 wristband was utilized to collect various biophysiological signals including Blood Volume Pulse (BVP), Electrodermal Activity (EDA), 3-axis acceleration (ACC), and skin temperature (TEMP). In parallel, the Hexoskin smart shirt recorded electrocardiogram (ECG), respiration (RESP), and 3-axis acceleration (ACC) data.

For the WEAR study, in-lab test data were also collected using high-grade research equipment. The actiCHamp system (Brain Products GmbH, Gilching, Germany) was used for capturing Electroencephalogram (EEG) and further EDA data, while ECG, Electromyography (EMG), and additional ACC data were gathered using the Delsys Trigno system (Delsys, Inc., Nantick, MA, USA).

The RADWear and WEAR datasets involved a calibration protocol designed to establish baseline states for meditation and excitement/anxiety, captured during a cold pressor test in a lab setting. This initial session last about 30 minutes and were aimed at marking reference points for evaluating anxiety levels using the State-Trait Anxiety Inventory (STAI) X-2 questionnaire at the beginning and the STAI Y6 survey after each segment to measure changes in anxiety levels.

For the RADWear participants, following this calibration, the participants engaged in their medical rotations, during which data were collected in real-world conditions, termed "in-the-wild." These in-the-wild sessions were extensive, consisting of two, two-week sessions (during work hours at least 5 days per week lasting 6-8 hours each day). For this study, nine RADWear participants were included in the Calibration sessions, two of which have yet to complete their In-the-wild sessions.

The WEAR participants underwent a comprehensive in-lab testing session lasting approximately 4 hours including the Calibration session. After completing the Calibration session, the participants performed the In-lab session that included a series of stress and anxiety-inducing protocols: the Trier Social Stress Test (TSST), seated Stroop test, and walking Stroop test. These activities were selected to induce anxiety and serve as validated methods to establish ground truth for anxiety levels. The in-the-wild data collection for these test participant had not yet been completed by any participant. Similar to the RADWear study, WEAR participants began with an intake survey (STAI X2) and complete follow-up questionnaires (STAI Y6) after each test to assess anxiety responses.

For this study, the datasets were organized into 3 subsets based on similarity of the test conditions: 1) RADWear and WEAR calibration sessions, 2) WEAR in-lab sessions, and 3) RADWear in-the-wild datasets. This selection of data subsets inform how models perform under varying levels of environmental conditions that restrict participants' motion.

4.3.2 Addressing noise

Because of the RADWear and WEAR datasets' experimental protocols containing physical activity in both lab and wild settings, feature extraction presented more of a challenge. This was due to significant amounts of motion-artifact noise, which negatively affected the signal and the ability to extract features. Having tested publicly available feature extraction methods for the available biophysiological measures, the performance was determined to be inadequate. This was especially true for heart rate signals as noise can muddy the signal's key markers that were used for feature extraction (Figure 4.1). Consequently, a feature extraction algorithm, with more of a focus on overcoming this issue was developed to outperform existing available solutions, based on concepts from Malik et al. [266], [267] and methods from

Scholkmann et al. [249]. In the one of the worst recorded ECG sessions observed during RADWear data collections, our in-house algorithm was able to detect peaks with a 98% accuracy, while the next-best algorithm achieved a 60% peak detection accuracy, which is not acceptable for anxiety detection algorithms.

For the RADWear in-the-wild data, due to the sparsity of the daily intake survey that served as a ground truth, it was reasonable to remove these corrupted segments from being processed further and dropped. Another challenge was faced was that, due to the quality of some of the segments, it was not possible to compute some features. This resulted in either 1) one or more features partially missing some data points, or 2) one or more features completely missing. Imputation was used to resolve this issue. Some features were unable to be computed for the whole segment, while others had values missing at some intervals. For the first case, a simple mean imputation was employed to replace the missing points. If a significant portion of a segment was missing or too much noise existed, that segment was dropped.

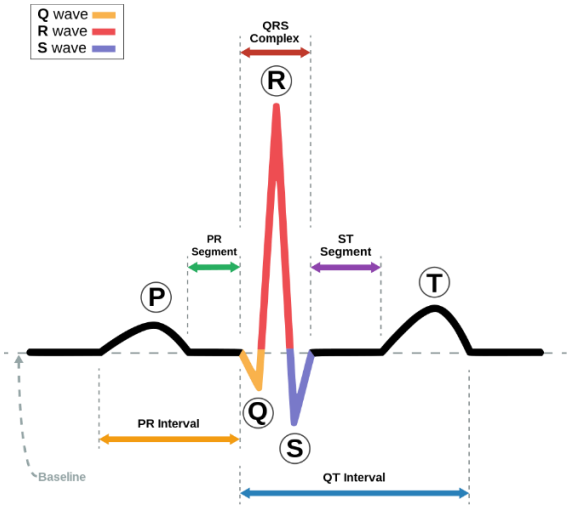


Figure 4.1: Key markers in ECG signals that were used to extract features to represent heart rate variability. For example, RR-interval was generated by calculating distance between each R peak and subsquent R peak. Figure taken from Rao et al. [274]

4.3.3 Label generation

Within the study protocol, after each of the defined test conditions, the participants filled out questionnaires to assess their state. The questionnaires contained questions from STAI and were evaluated following the STAI standard evaluation methods for evaluating anxiety level of an individual [26], [46], resulting in a label that was then used as a ground truth for the ML model. The STAI contains two types of questions: anxiety-present and anxiety-absent questions. The anxiety-present questions were rated on a 4-point scale, where the lowest response is scored 1 and the highest response scored 4. Anxiety-absent questions were scored inversely. The scores were summed and assessed. It is commonly for the scores to be classified as no or low anxiety below 37, moderate anxiety above 37 and below 50, and high anxiety above 50 and below 80 [46]. For this study, only the presence or absence of anxiety was considered, with the threshold for scoring is set at 37, which is equivalent to approximately a score of 11 for STAI Y6, which consists of 6 questions, instead of 20 like STAI X-2 (see Figure 4.2 for a visual representation of the STAI scores for one participant across different test conditions).

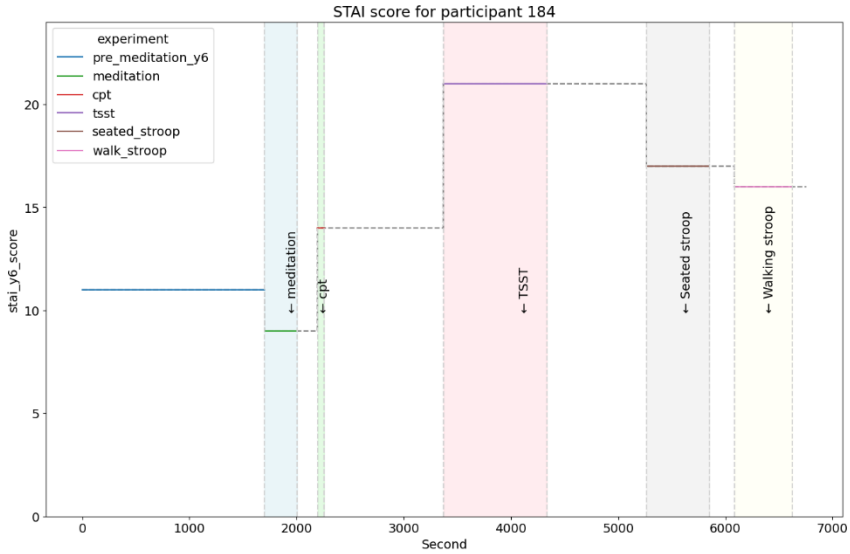


Figure 4.2: STAI score for one of the participants at the different test conditions experienced. The threshold considered to be indicative of observation of anxiety is 11 for the STAI Y6. It can be observed that the meditation session reduced anxiety level, while CPT increased it. For this participant, the TSST seemed to be highest cause for anxiety.

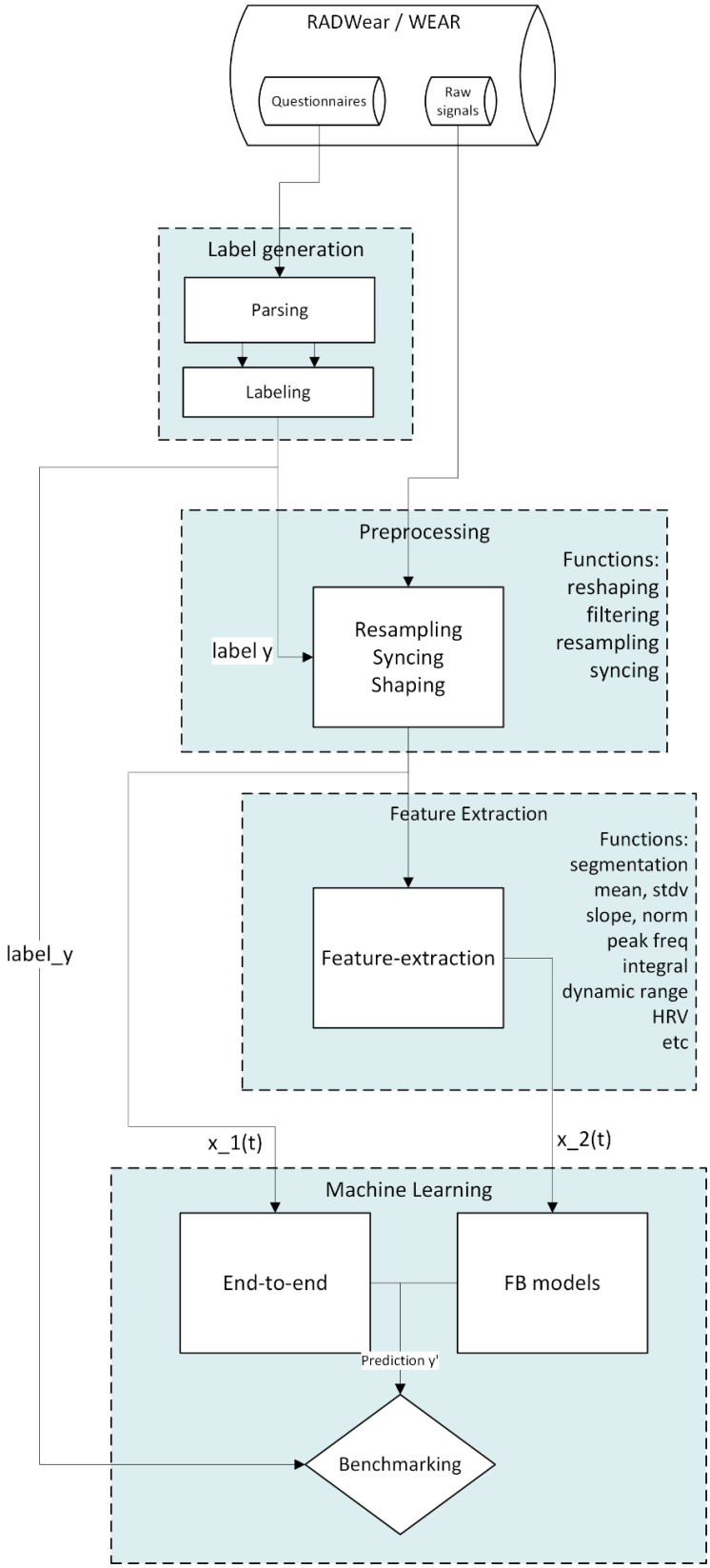


Figure 4.3: This flowchart outlines the pipeline for processing and analyzing wearable sensor data, from raw data collection through noise filtering, preprocessing, and feature extraction, to machine learning.

4.3.4 Class balancing

Generally, to have effective classification outcomes from ML algorithms, datasets should be balanced between classes. Imbalanced datasets can lead to biased model performance, misleading performance metrics, limited generalization and robustness, and the neglect of important minority classes [268]–[271], necessitating the use of balancing techniques to ensure fair and accurate machine learning models. Undersampling the dominant class is utilized to reduce the imbalance [272]. The laboratory-controlled datasets (RADWear + WEAR Calibration, WEAR in-lab) had good balance (Table 4.1). Although the binary dataset for the WESAD had a skewed balance (70/3), this ratio was considered acceptable. Due to the nature of the RADWear in-the-wild dataset, the anxious and non-anxious classes were imbalanced, with 80% of the data being labeled as not anxious. Undersampling was performed to the RADWear in-the-wild dataset to match the that of the Calibration and In-lab segments. Undersampling was performed by reducing the number of datapoints collected from the dominant class, which preserves the number of points of the minority class.

Table 4.1: Class distribution for each subset of the datasets.

Dataset	Not Anxious	Anxiety	Adjusted
WESAD	70%	30%	×
RADWear + WEAR Calibration	58%	42%	×
WEAR in-lab	59%	41%	×
RADWear in-the-wild	80.66%	19.34%	to 30% and 41%

4.3.5 Self-learned Machine learning models and Transfer Learning

Afterwards, models that were trained previously in Chapter 3 were tested on the three subsets to test the validity of transfer learning for this application. Both FB and E2E models were employed. FB models used were: DT, RF, LDA, KNN, AB, SVM, and XGB. Only XGB and DT were utilized to test the performance for transfer learning. FCN and ResNet were the architectures used for E2E models, as these were previously observed to perform the best compared to eight other E2E models (Chapter 3).

In Chapter 3, the study introduced noise-augmented data derived from the WESAD dataset to assess the resilience of machine learning models to environmental disturbances. This approach was crucial for understanding how these models performed under simulated conditions

that mimicked real-world noise, which is often encountered in daily activities and can significantly affect the accuracy of anxiety detection systems.

By training models on both the original and noise-augmented WESAD data, the study established a performance benchmark, examining how noise impacts model effectiveness and identifying which models maintain their predictive power despite increased noise levels. This process not only highlighted the robustness of certain models but also allowed for the optimization of these models to withstand typical real-world disruptions.

In Chapter 4, the findings from Chapter 3 become particularly valuable. The transfer learning techniques applied here involved testing the models—initially trained on WESAD and its noise-augmented version—on the RADWear and WEAR datasets. The transition to testing on these datasets, which feature real-life conditions and more complex environments than those simulated by noise augmentation, was facilitated by the preliminary insights gained from the noise impact analysis. This step ensured that the models not only generalized well across different types of input data but were also applicable in practical settings where environmental variability is the norm.

Overall, analyzing models with noise-augmented data in Chapter 3 enhances the relevance and applicability of these models for Chapter 4, where real-world conditions play a crucial role. It provided a solid foundation for understanding model performance in the face of unpredictable environmental factors, ensuring that the anxiety detection systems developed are both effective and reliable in varied real-world scenarios.

For transfer learning, models that have been tested on WESAD, and noise-augmented WESAD, were tested on the three subsets. The use of WESAD serves to establish a performance benchmark as a reference point. Since the models were initially trained and validated on this dataset in Chapter 3, showcasing their performance on WESAD helps in understanding their efficiency and accuracy before they were tested under more variable conditions such as those provided by the RADWear and WEAR datasets. It is important to articulate that the WESAD dataset acts as a foundational dataset for training the models, whose performance metrics were crucial for establishing a comparative analysis.

The incorporation of noise-augmented data in Chapter 3 was a strategic decision to bolster the resilience of machine learning models against environmental noise, which is a common and disruptive element in real-world settings. The process of augmenting the WESAD dataset with noise simulates these realistic disturbances, thus enabling us to stress-test the models and prepare them for the unpredictable variances they would encounter in practical scenarios. By doing so, we aimed to ensure that the models retain their predictive accuracy even when faced with data that has been compromised or distorted by external factors.

Furthermore, employing transfer learning with models pre-trained on both the original and noise-augmented WESAD datasets serves a dual purpose. Firstly, it leverages the noise resilience built during the augmentation process, allowing us to examine how well these models can adapt to and perform on new, noisier datasets like RADWear and WEAR. Secondly, it extends the models' applicability beyond controlled laboratory conditions, facilitating their deployment in diverse and challenging environments where noise is a given. This approach underscores our commitment to developing robust anxiety detection systems that are not just theoretically sound but also practically viable and reliable across various real-life applications.

There were 11 noise augmented datasets, each with a different amount of noise, dictated by a specified signal-to-noise (SNR) ratio, which were 0.0001, 0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5 and 0.6 for FB models, and 0.01, 0.1, 0.15 and 0.4 for E2E models.

To assess the performance of each model evaluated at each subset, F_1-score and Accuracy were used to evaluate the models. These metrics were calculated using the following equations:

$$Accuracy = \frac{True\ Positive + True\ negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Posive + False\ Negative}$$

$$F_1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.4. Results

4.4.1 Self-Learned Models

The self-learned models refer to those tested and trained exclusively on each of the three data subsets, including the original WESAD dataset as a reference point for benchmarking. Analysis of the model performances on the RADWear and WEAR datasets (Table 4.2) highlighted the superiority of the XGB and DT models. Notably, the XGB model achieved an impressive 0.99 accuracy and F1-score on the WESAD dataset, aligning with findings from Chapter 3. This performance was maintained across the RADWear and WEAR calibration subsets, where XGB recorded 0.95 accuracy and a 0.94 F1-score. Furthermore, even in the challenging conditions of the WEAR in-lab settings, XGB showed robust performance with 0.92 accuracy and 0.90 F1-score. These results validate the effectiveness of self-learned models in real-world scenarios, substantially outperforming the 50% accuracy expected of random guessing in binary classification.

Table 4.2: F_1-score performance results of models trained and tested on the same subsets of the datasets.

Data	XGB		DT		LDA		RF	
	ACC	F_1	ACC	F_1	ACC	F_1	ACC	F_1
WESAD	0.99	0.99	0.99	0. 99	0.93	0.92	0.91	0.89
RADWear + WEAR Calibration	0.95	0.94	0.91	0.90	0.82	0.80	0.88	0.86
WEAR in-lab	0.92	0.90	0.93	0.94	0.69	0.65	0.80	0.76
RADWear in-the-wild (balanced to 30%)	0.87	0.77	0.97	0.95	0.79	0.52	0.82	0.72
RADWear in-the-wild (balanced to 41%)	0.88	0.58	0.99	0.66	0.73	0.61	0.78	070
Random guess	0.50							

Interestingly, the most predictive features varied across models and datasets and yielded a different picture than previously observed (Figure 3.3 and Table 4.3). In Table 3.3, the performance metrics of various features across different models were presented, offering a detailed comparison of how different data types influence the effectiveness of the models. The "Weighted average" represents the aggregated influence of each feature across different models. This calculation was performed using a weighted mean, where each feature's importance is

squared within a model, summed across models, and then divided by the total sum of the feature importances. This calculation method emphasizes features that were consistently significant across various models, highlighting their predictive power in detecting anxiety. This approach particularly underscores the adaptability and critical role of specific features under conditions such as motion, where accelerometer data becomes increasingly significant.

For the XGB model, ECG-related features, particularly ECG_{max} and ECG_{min}, consistently ranked among the top features. However, in the presence of motion, the wristband's accelerometer-derived features, such as $ACC_{net\ w\ std}$, gained prominence. This suggests that while ECG signals were highly informative for anxiety detection, accelerometer data can provide complementary information, especially in scenarios involving movement.

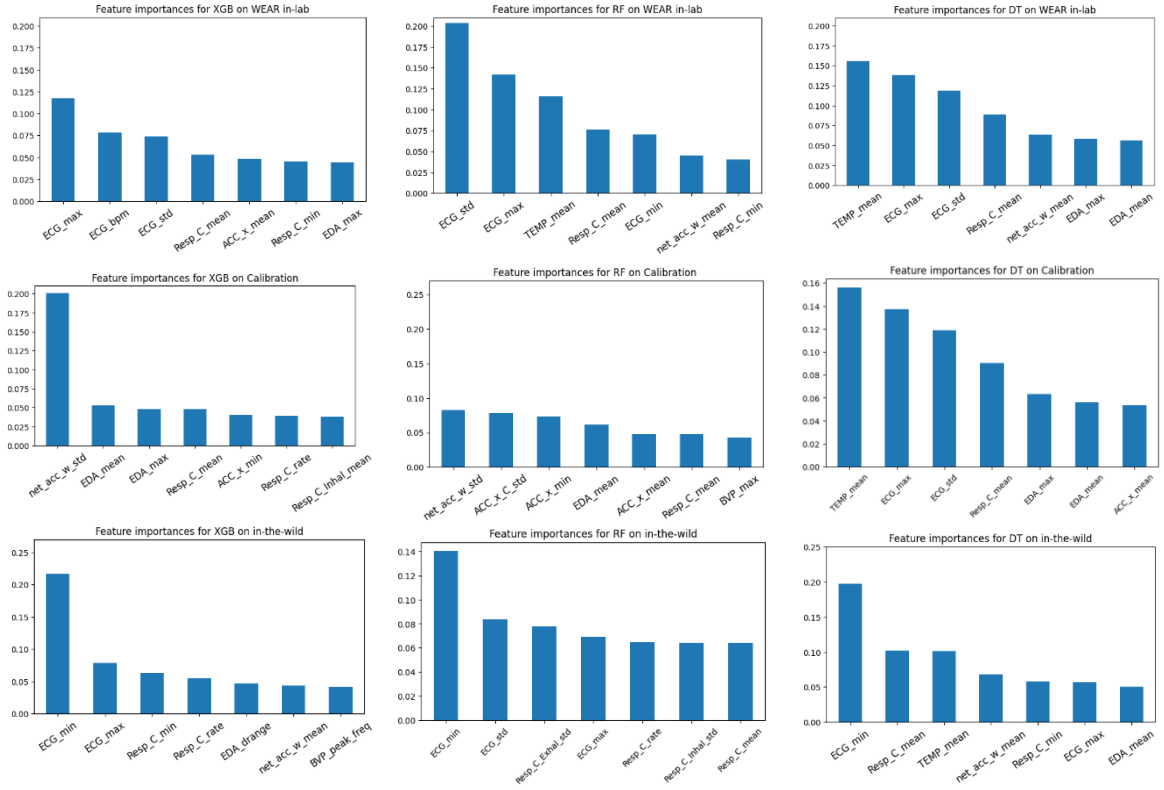


Figure 4.4: Feature importance for the 3 top performing models (XBG, RF, DT). There is a wider range of features than previously observed in Chapter 3. The x-axis lists top 7 features for the given model, while the y-axis is Feature Importance, where summation for all features is unity.

Table 4.3: Feature importance for self-learned models on RADWear + WEAR Calibration, WEAR in-lab and RADWear in-the-wild subsets.

	Modality	Modality Weighted Average	Feature	Feature Weighted Average	RF	DT	XGB
RADWear in-the-wild	ECG	0.13	ECG_{min}	0.17	0.14	0.18	0.19
			ECG_{std}	0.08	0.08		
			ECG_{max}	0.06	0.06	0.04	0.07
	Resp	0.07	$Resp_{Exhal\ std}$	0.07	0.07		
			$Resp_{mean}$	0.07	0.07	0.09	0.04
			$Resp_{rate}$	0.06	0.07		0.06
			$Resp_{I/E}$	0.06	0.06		
			$Resp_{min}$	0.08		0.08	0.08
	EDA	0.06	EDA_{mean}	0.07		0.07	
			EDA_{drange}	0.06			0.06
	BVP	0.04	$BVP_{peak\ freq}$	0.04			0.04
	TEMP	0.11	$TEMP_{mean}$	0.11		0.11	
	ACC	0.06	$ACC_{net\ w\ mean}$	0.06		0.06	
RADWear + WEAR Calibration	EDA	0.10	EDA_{mean}	0.12	0.06	0.17	0.05
			EDA_{max}	0.05			0.05
	RESP	0.05	$Resp_{min}$	0.07		0.07	
			$Resp_{mean}$	0.05	0.05		0.05
			$Resp_{rate}$	0.05		0.05	0.04
			$Resp_{Inhalmean}$	0.04		0.03	0.04
	BVP	0.04	BVP_{max}	0.04	0.04		
	ACC	0.13	$ACC_{net\ w\ std}$	0.21	0.08	0.26	0.20
			$ACC_{net\ w\ mean}$	0.04		0.04	
			$ACC_{x\ min}$	0.06		0.06	
			$ACC_{x\ C\ std}$	0.08	0.08		
			$ACC_{x\ min}$	0.06	0.07		0.04
			$ACC_{x\ mean}$	0.05	0.05		
WEAR in-lab	ECG	0.12	ECG_{bpm}	0.08			0.08
			ECG_{max}	0.13	0.14	0.14	0.12
			ECG_{min}	0.07	0.07		
			ECG_{std}	0.15	0.20	0.12	0.07
	EDA	0.05	EDA_{max}	0.05		0.06	0.04
			EDA_{mean}	0.05		0.05	
	RESP	0.06	$Resp_{mean}$	0.07	0.08	0.09	0.05
			$Resp_{min}$	0.04	0.04		0.05
	ACC	0.05	$ACC_{x\ mean}$	0.05		0.06	0.05
			$ACC_{net\ w\ mean}$	0.05	0.05		
	TEMP	0.14	$TEMP_{mean}$	0.14	0.12	0.16	

When evaluating the transfer learning (TL) performance (Figure 4.5 and Table 4.4), we discovered that the RF model excelled in the calibration and in-lab settings, while SVM achieved the best results for in-the-wild data. Intriguingly, models like XGB and DT, which previously outperformed in the baseline case, did not maintain their superiority in the TL scenario. This finding highlights the importance of carefully selecting models based on the specific characteristics of the target dataset and the potential limitations of directly transferring models trained on one dataset to another.

In exploring the performance of transfer learning across our datasets, we identified a significant trend related to the distribution of feature importance weights in the models. Models exhibiting more evenly distributed feature importance weights demonstrated more successful adaptation when applied to new datasets, particularly when transitioning from controlled to more variable real-world conditions. For instance, models like LDA, which generally maintained a balanced importance across features, adapted more effectively compared to those heavily reliant on one or two specific features, such as XGB. The latter showed reduced performance during transfer learning tasks, especially when applied to the in-the-wild subset of the WEAR dataset. This suggests that a more uniform distribution of feature weights might enhance a model's adaptability, supporting better generalization across different experimental conditions and datasets. This finding underscores the importance of considering feature balance during model training for effective transfer learning applications.

Table 4.4: Performance results for Transfer Learning models on calibration, in-lab, and in-the-wild subsets of the data.

Model	DT		RF		XGB	
	Accuracy	F_1-Score	Accuracy	F_1-Score	Accuracy	F_1-Score
RADWear + WEAR Calibration	0.12	0.12	0.76	0.76	0.07	0.07
WEAR in-lab	0.18	0.39	0.93	0.94	0.44	0.61
RADWear in-the-wild anxiety class balanced to 41%	0.18	0.18	0.93	0.93	0.44	0.44

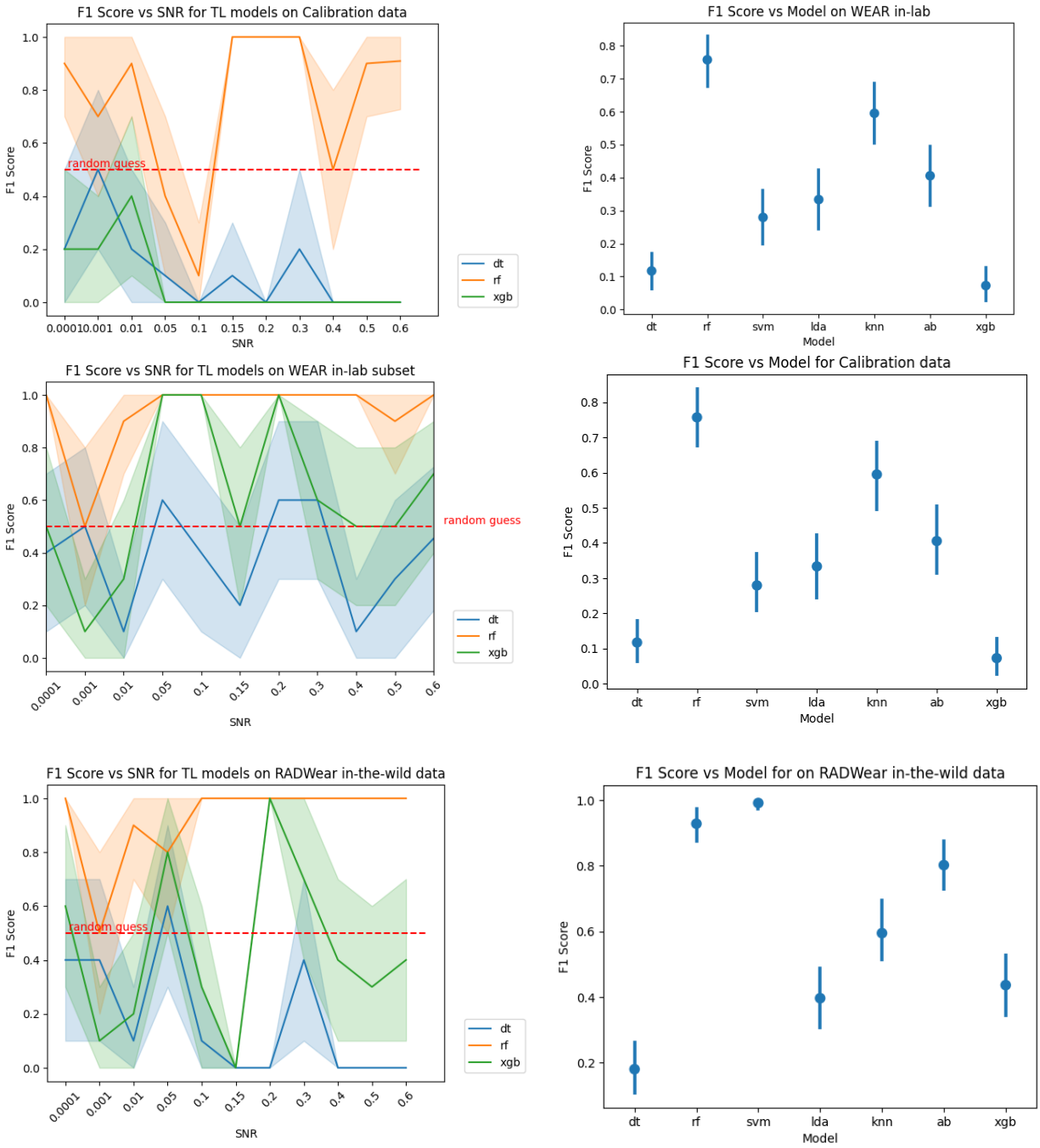


Figure 4.5: F1-score Transfer learned models are shown with the left column showing F_1-score for models trained at different noise levels at different SNR, shown in the x-axis. The right column shows F_1-score for each of the tested models at for all SNRs. F_1-score, as tested on a) on Calibration data for WEAR and RADWear in the top row, b) WEAR in-lab sessions in the middle row and, c) on RADWear in-the-wild in the bottom row.

In addition to the FB models, we also evaluated the performance of E2E models (Table 4.5). When trained on the WESAD dataset and tested on the RADWear and WEAR calibration subset, the FCN model achieved an F1-score of 0.29 and an accuracy of 0.39, with the ResNet

model doing even more poorly. Similar trends were observed in the WEAR in-lab and RADWear in-the-wild settings, with FCN consistently surpassing ResNet in terms of both F1-score and accuracy. Notably, the performance of E2E models was substantially lower compared to the FB models, suggesting that the feature-based approach may be more suitable for the specific characteristics of the RADWear and WEAR datasets.

Table 4.5: Performance of TL models from WESAD and Noise-augmented WESAD onto subsets of the RADWear and WEAR datsts. Noise-augmented WESAD model was trained on the set with SNR = 0.2.

MODEL	TRAINING SET	TESTING SET	F_1-SCORE	ACCURACY
FCN	WESAD	WEAR and RADWear Calibration	0.29	0.39
FCN	WESAD	WEAR in-lab	0.25	0.37
FCN	WESAD	RADWear in-the-wild	0.18	0.21
ResNet	WESAD	WEAR and RADWear Calibration	0.15	0.30
ResNet	WESAD	WEAR in-lab	0.15	0.30
ResNet	WESAD	RADWear in-the-wild	0.09	0.15
FCN	Noise-augmented WESAD	WEAR and RADWear Calibration	0.23	0.36
FCN	Noise-augmented WESAD	WEAR in-lab	0.15	0.30
FCN	Noise-augmented WESAD	RADWear in-the-wild	0.11	0.21
ResNet	Noise-augmented WESAD	WEAR and RADWear Calibration	0.12	0.17
ResNet	Noise-augmented WESAD	WEAR in-lab	0.20	0.22
ResNet	Noise-augmented WESAD	RADWear in-the-wild	0.07	0.10

Furthermore, we explored the impact of noise augmentation on the performance of E2E models (Table 4.5). When trained on the noise-augmented WESAD dataset (SNR = 0.2) and tested on the RADWear and WEAR subsets, both FCN and ResNet models exhibited a decrease in performance compared to their counterparts trained on the original WESAD dataset.

While the FB models showed varied performance across different datasets, the E2E models, specifically the Fully Convolutional Network and Residual Network, exhibited distinct performance characteristics under transfer learning conditions. The FCN and ResNet models, originally trained on the WESAD dataset and then tested on RADWear and WEAR datasets, demonstrated varying degrees of effectiveness when adapting to noisy real-world data. As depicted in Table 4.5, the FCN model achieved an F1-score ranging from 0.18 to 0.29 and accuracy from 0.21 to 0.39 across different testing scenarios. Notably, its performance was somewhat better on the WEAR + RADWear Calibration data with an F1-score of 0.29 and accuracy of 0.39. ResNet models, however, performed slightly lower under similar conditions with F1-scores and accuracy consistently below those of FCN, peaking at an F1-score of 0.15 and accuracy of 0.30 in both the WEAR + RADWear Calibration settings. This indicates a challenge in the model's ability to generalize from the laboratory to more dynamic real-world

conditions, especially in noise-augmented datasets where the highest F1-score was only 0.12 and accuracy reached 0.17.

4.5. Discussion, Limitations, and Future Work

Our study makes significant strides in advancing the field of anxiety detection using wearables and machine learning. By leveraging the RADWear and WEAR datasets, which capture real-world scenarios, we demonstrated the feasibility of developing robust models that can accurately detect anxiety in noisy environments. The strong performance of our models, particularly XGB and DT, in both controlled and real-world settings, underscores the potential for deploying such techniques in practical mental health monitoring applications.

Compared to previous research, our work stands out in several aspects. First, we introduce the RADWear and WEAR datasets, which offer a diverse range of participants and experimental conditions, including both in-lab and in-the-wild settings. This rich data collection enabled us to explore our models' performance across various contexts, enhancing the generalizability of our findings. Second, we conducted an analysis of feature importance, shedding light on the relative contributions of different physiological signals and derived features in anxiety detection. Third, we explored the applicability of transfer learning, highlighting the challenges and opportunities in adapting models trained on one dataset to another. Notably, it seems that models with feature importance close in weight perform better for transfer learning, while models with feature importance heavily weighted on one or two features, like XGB for calibration and in the wild, do not transfer learning well.

Our exploration of noise augmentation in the context of E2E models raises important considerations. While noise augmentation has been shown to enhance the robustness of FB models (Chapter 3), its effectiveness for E2E models in transfer learning scenarios appears to be limited. This finding highlights the need for further research into techniques specifically designed to improve the transferability and generalization of E2E models across different datasets and environments. In summary, while noise augmentation has shown to increase the robustness of FB models, its effectiveness for E2E is not evident.

The study's findings emphasize the importance of selecting features that were robust to various environmental factors and failure modes. For instance, EDA signals may be disrupted in wet, rainy, or humid environments, or in the presence of physical activity. Similarly, acceleration-based features may not be reliable when physical activity is present. Similarly, ECG signals can be affected by changes in conductivity and were more challenging to incorporate into wearable devices. BVP signals were also susceptible to disruptions caused by gaps between the device and the skin. Consequently, models that rely on a distributed feature importance, where multiple signals contribute to the detection of anxiety, were more likely to be resilient to individual signal failures. Future research should focus on identifying and engineering features that are robust to these challenges, enhancing the reliability of anxiety detection models in real-world settings.

However, our study is not without limitations. The sample sizes of the RADWear and WEAR datasets used in this study, while diverse, may not fully capture the entire spectrum of individual variability in anxiety responses. As those studies are ongoing, these datasets will expand, incorporating a broader range of participant demographics and clinical profiles. Additionally, while our FB models demonstrate promising performance, further validation in longitudinal studies and real-world deployments would be necessary to assess their long-term reliability and usefulness in clinical practice.

Developing accurate and robust anxiety detection models can pave the way for personalized mental health interventions that are delivered in real-time using wearable devices. This could revolutionize the way we monitor and manage anxiety disorders, enabling early detection, timely support, and targeted treatments. However, translating these models into practical tools will require addressing challenges such as data privacy, user acceptance, and seamless integration with existing healthcare systems.

The performance of E2E models in our study provides additional insights into the challenges and opportunities of applying deep learning techniques for anxiety detection in real-world settings. Performances of the E2E models were much lower compared to FB models. This suggests that the feature-based approach, which leverages domain knowledge and carefully

crafted features, may be more suitable for the specific characteristics of the RADWear and WEAR datasets.

The underperformance of E2E models in transfer learning scenarios, particularly in the RADWear in-the-wild dataset, suggests limitations in the model's architecture or training regimen when confronted with highly variable real-world data. Additionally, using transfer learning based on noise-augmented WESAD models did not show significant improvements, indicating that further optimization or a different approach might be required to handle real-world variability effectively.

While the feature-based models such as XGB showed a higher resilience to noise and adaptability to various data conditions, the E2E models struggled, particularly in uncontrolled environments. This difference might be attributed to the E2E models requiring more nuanced feature representations which were not as effectively captured in noisy environments compared to more controlled settings where feature-based models can leverage specific, well-defined features.

It is worth noting that while E2E model performance could be improved through different methods (i.e., architectural optimization), the performance of FB models in all observed conditions provides sufficient performance that does not justify pursuit of E2E models further.

4.6. Conclusions

This study underscores the efficacy of feature-based models, particularly XGBoost and Decision Trees, in accurately detecting anxiety under both controlled conditions and real-world environments. The introduction of the RADWear and WEAR datasets enhances our understanding of the robustness of these models amidst environmental noise and varied conditions. The analysis of feature importance and the implementation of transfer learning have significantly contributed to advancing the field of anxiety detection using wearable technology. Key findings emphasize the resilience of feature-based models and the critical role of precise feature selection in maintaining model accuracy across diverse settings. These insights not only validate the effectiveness of current methodologies but also underscore the potential of these models in practical mental health monitoring applications.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

5.1. Contributions

The dissertation investigated machine learning-based anxiety detection methods applied to wearable data. Each of the studies comprised within the dissertation contributes uniquely to the field of anxiety detection using machine learning.

Chapter 1 introduced the topic, setting the stage for the subsequent in-depth exploration.

Chapter 2, titled "Machine Learning Approaches in Anxiety Detection - A Comprehensive Review," reviewed various machine learning methods for anxiety detection, highlighting the state-of-the-art techniques and their applications. It offered a detailed exploration of the application of machine learning (ML) methods in anxiety detection. This chapter distinguished itself by providing an extensive analysis of both feature-based and end-to-end deep learning approaches in the context of anxiety detection. It comprehensively covered the growth of the field, particularly in 2023, and discussed the methodologies, model architectures, experimental conditions, feature selections, and dataset utilization within this domain.

A key focus of this chapter was the contrast between feature-based ML models, like Support Vector Machines, Decision Trees, and Random Forests, and end-to-end deep learning models, including Convolutional Neural Networks and Recurrent Neural Networks. The chapter highlighted the performance strengths of each approach and architecture, noting the interpretability and simplicity of feature-based models against the automatic feature extraction capabilities of end-to-end models. The review also delved into the challenges these models face in real-world scenarios, emphasizing the need for models resilient to real-world conditions, including data noise and variability.

Furthermore, the chapter underscored the potential of personalized healthcare approaches in anxiety detection. It presented the study’s critical examination of the current landscape of ML applications in anxiety detection, setting the stage for future advancements in the field.

Chapter 3, titled "Resilience of Machine Learning Models in Anxiety Detection: Assessing the Impact of Environmental Noise on Wearable Technology," significantly

contributed to the field of anxiety detection through its focus on the resilience of machine learning models against environmental noise. This chapter thoroughly investigated the performance of various anxiety detection models under conditions simulating real-world disturbances, using enhanced datasets like WESAD with added synthetic noise. It assesses how different levels of Gaussian noise impacted the accuracy and reliability of both feature-based and end-to-end machine learning models.

A notable contribution of this chapter was its exploration of the robustness and adaptability of these models to noisy data. It delved into the challenges of maintaining high accuracy levels amidst environmental noise, a critical factor in real-world applications. By conducting comprehensive evaluations, the chapter provided valuable insights into the effectiveness of different model architectures in noisy conditions, setting new benchmarks for model resilience.

This chapter's unique findings, which have never been studied before, will be pivotal in guiding the development of more reliable and efficient machine learning models for anxiety detection, enhancing their practicality in everyday scenarios. This comprehensive analysis not only advances the field but also provided deeper insights into architectural performance, crucial for future architectural enhancements in anxiety detection technology. It underscored the importance of creating models that were not just theoretically sound but also practically applicable in diverse and unpredictable real-life environments.

Chapter 4, titled "Advancing Anxiety Detection in Noisy Environments: Efficacy of Machine Learning Models on Real-world Datasets," built upon the foundation laid in the previous chapters by applying the developed models to real-world scenarios. This chapter is pivotal in transitioning from theoretical aspects and controlled environments to practical, everyday applications. It explored the adaptability and effectiveness of machine learning models in naturalistic settings, where environmental noise and variability are inherent.

A key contribution of this chapter was its emphasis on real-world applicability. It demonstrated how models trained on laboratory data can be adapted and optimized for use in everyday environments, overcoming challenges posed by uncontrolled settings. This involved

the implementation of advanced techniques like transfer learning to enhance model robustness against real-world disturbances.

Furthermore, the chapter provided insights into the performance of anxiety detection models in diverse real-life situations, assessing their reliability and efficiency outside the laboratory. This exploration is crucial for the advancement of wearable technology for mental health monitoring, offering valuable guidance for the development of practical, user-centric anxiety detection tools. The findings in this chapter significantly contribute to bridging the gap between academic research and real-world application, paving the way for more accessible and effective mental health technologies.

This final chapter, Chapter 5, concludes the study with a summary of findings, implications for future research, and potential applications in mental health monitoring. The dissertation's overall contribution lies in its comprehensive examination of machine learning in anxiety detection, particularly in the context of environmental noise and real-world applicability, thereby advancing the field significantly.

5.1.1 Redefining Time-Series Analysis: The Rising Dominance of CNNs in Anxiety Detection

A key observation from the dissertation is the progression from traditional feature-based models towards more advanced end-to-end deep learning models. This shift reflects the field's evolution towards embracing more sophisticated, data-driven approaches. However, this transition also brings forth challenges such as the black box nature of deep learning models and their computational intensity. The variability in performance metrics and occasional lack of detail in reporting that enables replicating authors' work in some studies suggests potential inconsistencies in the field, making cross-comparison of models challenging.

Although there are few studies to draw concrete conclusions from, it is an intriguing observation that Convolutional Neural Networks (CNN) were frequently used and perform well in anxiety detection using time-series data, despite that Recurrent Neural Networks (RNN) and Long Short-Term Memory networks (LSTM) were often considered more natural fits for such data [273]. Below are a few reasons why it is believed that CNNs might be prevalent and

effective in this context. The frequent utilization and notable performance of CNNs in the realm of anxiety detection, particularly with time-series data, present an intriguing scenario in the landscape of machine learning. This preference for CNNs over more traditional choices like RNNs and LSTMs can be attributed to several key factors.

First, CNNs **excel in automatic feature extraction** [206]. Their ability to effectively identify and learn spatial hierarchies of features makes them adept at deciphering complex patterns in physiological signals, which are commonly analyzed in anxiety detection. This capability becomes particularly beneficial in extracting intricate patterns from time-series data.

Second, the field has witnessed significant **advancements in CNN architectures**, notably the development of 1D CNNs specifically designed for time-series data [207]. These advancements have broadened the scope of CNNs, enabling them to handle sequential data while retaining their robust feature extraction capabilities.

In terms of robustness and generalization, CNNs have an edge. Known for their **resilience**, these networks tend to be less susceptible to overfitting, particularly in scenarios involving large datasets, compared to some RNN variants [208]. Additionally, the pooling layers in CNNs serve as an inherent **noise rejection mechanism**, acting like a low-pass filter. This feature is particularly advantageous in 'wild' applications of anxiety detection, where data are often marred with noise, thus enhancing the suitability of CNNs for real-world deployments.

From a **computational** standpoint, CNNs offer **efficiency**. They generally require less training time and are more straightforward to optimize, an advantage particularly pronounced when dealing with extensive datasets [208], [209].

The proven success of CNNs in related domains of signal processing and pattern recognition, such as in image and speech recognition, has likely influenced their adoption for anxiety detection using physiological data, suggesting a transfer of confidence and methodology from these areas [211], [212].

Moreover, the integration of hybrid approaches, combining the strengths of CNNs with RNNs or LSTMs, showcases an innovative strategy. These hybrid models leverage CNNs for

their feature extraction prowess and RNNs/LSTMs for their sequence modeling strengths, thereby addressing a broader spectrum of data characteristics. Having said that, the one implementation that has been observed did not perform as well as basic models [193].

However, it is important to note that the choice of a model architecture should align with the specific nuances of the data and the task at hand. While CNNs have demonstrated promising results, RNNs and LSTMs hold their unique advantages, particularly in capturing temporal dependencies, and may be more suited for certain types of time-series data. This underlines the dynamic nature of machine learning, where best practices are continually refined and evolved through ongoing research and experimentation.

The notable efficacy of Convolutional Neural Networks (CNNs) in anxiety detection using time-series data, despite their inherent limitations in capturing long-term temporal dependencies, can be elucidated through multiple factors. Primarily, CNNs, especially when adapted as 1D CNNs for time-series analysis, are adept at identifying local temporal patterns within specific data windows. This capability, though not extending to long-term dependencies, is crucial in recognizing significant local features in physiological data related to anxiety. Additionally, the preprocessing and transformation of time-series data play a pivotal role; techniques like windowing and segmenting, or applying Fourier Transforms, can reveal key features that CNNs effectively learn. The depth and architectural variations of modern CNNs also contribute, as these allow the learning of complex and abstract features from data. Hybrid models, wherein CNNs are used in tandem with Recurrent Neural Networks (RNNs) or Long Short-Term Memory networks (LSTMs), further bolster their effectiveness by combining CNNs' feature extraction capabilities with the temporal modeling strengths of RNNs/LSTMs.

Moreover, the task-specific efficacy of CNNs should not be underestimated. For certain aspects of anxiety detection, such as identifying specific physiological markers, the importance of temporal context might be secondary compared to detecting critical features. Enhanced training techniques, including advanced regularization, optimization algorithms, and activation functions, have also improved CNNs' generalization capabilities, even with the complexities of time-series data. Finally, the training of CNNs on large, diverse datasets can sometimes compensate for their architectural limitations, as the richness of the data helps in generalizing

well, even in scenarios demanding an understanding of temporal relationships. Thus, while CNNs may lack inherent recurrence, a combination of these factors contributes to their successful application in anxiety detection from time-series data, highlighting the multifaceted nature of model selection and effectiveness in machine learning applications.

5.1.2 Impact of Testing Environments on Model Selection and Performance

In the context of anxiety detection using machine learning, the prevailing focus on laboratory-based studies reveals a significant gap in the application and testing of these models under real-world conditions. This disparity underscores the risk of overfitting models to highly controlled and predictable laboratory environments, which may not accurately reflect the complexities and unpredictability of everyday scenarios. The consequence is a potential reduction in the effectiveness and reliability of these models when deployed in natural settings, where factors such as environmental noise, uncontrolled stimuli, and user variability come into play.

Moreover, the diversity in methods used to induce stress or anxiety in these studies, while showcasing the field's efforts to mimic a range of stress-inducing scenarios, also points to a lack of standardization in testing conditions. This variability in experimental setups can lead to challenges in comparing and generalizing the results across different studies. It also raises questions about the reproducibility of findings and the ability of models to adapt to various types of stressors and anxiety-inducing conditions.

These observations highlight the necessity for future research to pivot towards enhancing the real-world applicability of anxiety detection models. This involves not only adapting models to be more robust against the inherent noise and variability of real-life data but also ensuring that these models are tested and validated in diverse, uncontrolled environments that more accurately represent the settings in which they will be ultimately deployed.

In summary, these insights into the environmental impact on model choices in the field of machine learning for anxiety detection serve to guide future research directions. Emphasizing real-world applicability, robustness to environmental factors, and standardization in experimental

methodologies will be crucial in advancing the field towards more effective and reliable anxiety detection tools.

5.2. Future work

5.2.1 Robustness to real-life data with noise:

One of the most pressing challenges in the field of anxiety detection using wearables is dealing with real-world data that may contain various types of noise. These can range from motion-related noise, electrical interference, to device-specific noise. Addressing this issue involves two primary strategies:

- Noise Reduction: Developing methods to either remove or minimize noise components from the data.
- Noise-Robust Detection: Creating detection methodologies that are inherently robust to the presence of noise.

Efforts have been made to discriminate between psychological stress and physical stress, for instance, by using accelerometer data to categorize the physical activity state during anxiety detection [132], [213]. However, this approach is limited as it hinders detection during physical activity.

The need for analysis and quantification of real-world effects on detection performance is evident. This encompasses evaluating how different model architectures withstand real-world conditions, the sensitivity of certain signals and features, and the development of more robust detection modes that are validated with real-world data. The impact of device fidelity, such as sampling rates, is also a crucial factor in performance.

5.2.2 Architectural exploration

Given the notable performance of CNNs in anxiety detection with time-series data, further exploration into these models could provide better explainability of performance. This could assist in developing more tailored models for anxiety detection, addressing both the feature extraction capabilities and the challenges in interpretability and computational intensity.

5.2.3 Personalized health

Personalized healthcare in anxiety detection presents an untapped potential. Investigating the reasons behind model failures at an individual level is crucial. This includes identifying detrimental features for specific individuals, exploring alternative signals or features that may be more effective, and applying model-tuning or other techniques to improve detection accuracy. Furthermore, leveraging transfer learning could enhance performance on an individual basis, addressing the subjectivity of anxiety detection.

5.2.4 Towards anxiety prediction and intervention

The future of anxiety detection lies in moving beyond mere detection towards predicting anxiety onset and developing intervention strategies. There is a notable gap in methodologies for predicting anxiety onset. However, promising applications in real-time biofeedback, such as augmented virtual reality, HRV biofeedback training, and virtual reality visualization [214]–[216], offer avenues to reduce anxiety levels. Leveraging existing anxiety detection frameworks could significantly enhance the evaluation and improvement of anxiety therapies, facilitating coping mechanisms for negative emotions.

5.2.5 Longitudinal Studies

Conducting long-term studies with participants in naturalistic settings would provide valuable insights into how these models perform over time and in the face of real-world variability and noise.

5.2.6 Personalization of Models

Future research could focus on personalizing anxiety detection models to individual users. This personalization could account for individual differences in physiological responses to stress and anxiety, potentially improving the accuracy and reliability of detection.

5.2.7 Cross-Dataset Validation

Testing the models on different datasets, especially those collected in real-world settings, would help validate the generalizability of the findings and the robustness of the models across diverse populations and environments.

5.2.8 Ethical and Privacy Considerations

As wearable technology for anxiety detection advances, it is crucial to address the ethical implications and privacy concerns associated with continuous monitoring of individuals' physiological data.

5.2.9 Integration with Intervention Strategies

Linking anxiety detection with timely and appropriate intervention strategies could be a significant step forward. This would involve developing systems that not only detect anxiety but also trigger interventions or provide feedback to the user.

5.2.10 Economic and Scalability Analysis

Assessing the economic feasibility and scalability of deploying these systems in real-world settings is essential. This includes evaluating the cost-effectiveness and practicality of widespread implementation.

5.2.11 User Experience and Acceptance Studies

Understanding user perceptions, acceptance, and the overall experience of using wearable anxiety detection systems is crucial. This involves gathering feedback from potential users to refine the design and functionality of these systems.

By addressing these areas, future research can significantly advance the field of anxiety detection using wearable technology, leading to more effective, user-friendly, and widely applicable solutions.

5.3. Concluding remarks.

In conclusion, this body of work delved into the domain of machine learning-based anxiety detection using wearable data, highlighting significant advancements, and identifying areas for future exploration. Across its chapters, it has illuminated the transition from feature-based to advanced end-to-end deep learning models, emphasizing the evolution towards sophisticated, data-driven approaches. A consistent theme has been the necessity for models to perform reliably amidst the noise and variability of real-world settings, underlining the challenges of 'black box' models and computational demands.

This research underscores the pressing need for robust, adaptable, and personalized anxiety detection systems that consider a wide range of affective states. It sets the stage for future work focusing on enhancing real-world applicability, model personalization, and the integration of intervention strategies. The findings and methodologies presented in this dissertation contribute to advancing the field, paving the way for more effective and nuanced anxiety detection and mental health monitoring tools.

REFERENCES

[1] R. W. Picard, *Affective computing*. 2000. Accessed: Dec. 08, 2023. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=GaVncRTcb1gC&oi=fnd&pg=PR9&dq=affective+computing&ots=F6h6omxbdc&sig=UWpB3dqrRcBAZ1kSPqAd0jX1DQY

[2] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980, doi: 10.1037/H0077714.

[3] C. D. Spielberger, *Theory and Research on Anxiety*. Oxford, England: Academic Press Inc., 1966. doi: 10.1016/b978-1-4832-3131-0.50006-8.

[4] C. D. Spielberger, "Notes and Comments Trait-State Anxiety and Motor Behavior," *J. Mot. Behav.*, vol. 3, no. 3, pp. 265–279, Sep. 1971, doi: 10.1080/00222895.1971.10734907.

[5] D. Hackfort and C. D. Spielberger, *Sport-Related anxiety: Current trends in theory and research*. Academic Press Inc., 2021. doi: 10.4324/9781315781594-21.

[6] C. R. Thomas and C. E. Holzer, "The Continuing Shortage of Child and Adolescent Psychiatrists," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 45, no. 9, pp. 1023–1031, Sep. 2006, doi: 10.1097/01.CHI.0000225353.16831.5D.

[7] K. C. Thomas, A. R. Ellis, T. R. Konrad, C. E. Holzer, and J. P. Morrissey, "County-level estimates of mental health professional shortage in the United States," *Psychiatr. Serv.*, vol. 60, no. 10, pp. 1323–1328, 2009, doi: 10.1176/PS.2009.60.10.1323.

[8] W. J. Kim, "Child and Adolescent Psychiatry Workforce: A Critical Shortage and National Challenge," *Acad. Psychiatry*, vol. 27, no. 4, pp. 277–282, Dec. 2003, doi: 10.1176/appi.ap.27.4.277.

[9] A. Satiani, J. Niedermier, B. Satiani, and D. P. Svendsen, "Projected workforce of psychiatrists in the United States: A population analysis," *Psychiatr. Serv.*, vol. 69, no. 6, pp. 710–713, Jun. 2018, doi: 10.1176/appi.ps.201700344.

[10] H. Ayaz, P. A. Shewokis, S. Bunce, K. Izzetoglu, B. Willems, and B. Onaral, "Optical brain monitoring for operator training and mental workload assessment," *Neuroimage*, vol. 59, no. 1, pp. 36–47, Jan. 2012, doi: 10.1016/J.NEUROIMAGE.2011.06.023.

[11] S. Sibi, H. Ayaz, D. P. Kuhns, D. M. Sirkin, and W. Ju, "Monitoring driver cognitive load using functional near infrared spectroscopy in partially autonomous cars," *IEEE Intell. Veh. Symp. Proc.*, vol. 2016-August, pp. 419–425, Aug. 2016, doi: 10.1109/IVS.2016.7535420.

[12] C. M. Celano, D. J. Daunis, H. N. Lokko, K. A. Campbell, and J. C. Huffman, "Anxiety Disorders and Cardiovascular Disease," *Curr. Psychiatry Rep.*, vol. 18, no. 11, 2016, doi: 10.1007/s11920-016-0739-5.

[13] E. W. De Heer *et al.*, "The Association of depression and anxiety with pain: A study from NESDA," *PLoS One*, vol. 9, no. 10, pp. 1–11, 2014, doi: 10.1371/journal.pone.0106907.

[14] S. C. Segerstrom and G. E. Miller, "Psychological stress and the human immune system: A meta-analytic study of 30 years of inquiry," *Psychol. Bull.*, vol. 130, no. 4, pp. 601–630, 2004, doi: 10.1037/0033-2909.130.4.601.

[15] R. Martens, "Anxiety and motor behavior: A review," *J. Mot. Behav.*, vol. 3, no. 2, pp. 151–179, 1971, doi: 10.1080/00222895.1971.10734899.

[16] A. D. Baddeley, "Selective attention and performance in dangerous environments," *Br. J.*

Psychol., vol. 63, no. 4, pp. 537–546, 1972, doi: 10.1111/j.2044-8295.1972.tb01304.x.
[17] N. Derakshan and M. W. Eysenck, "Anxiety, processing efficiency, and cognitive performance: New developments from attentional control theory," *Eur. Psychol.*, vol. 14, no. 2, pp. 168–176, 2009, doi: 10.1027/1016-9040.14.2.168.
[18] R. L. Helmreich, T. R. Chidester, H. C. Foushee, S. Gregorich, and J. A. Wilhelm, "How effective is cockpit resource management training? Exploring issues in evaluating the impact of programs to enhance crew coordination.," *Flight Saf. Dig.*, vol. 9, no. 5, pp. 1–17, May 1990.
[19] B. Hoffman, "Cognitive efficiency: A conceptual and methodological comparison," *Learn. Instr.*, vol. 22, no. 2, pp. 133–144, 2012, doi: 10.1016/j.learninstruc.2011.09.001.
[20] P. Y. Correction Collins *et al.*, "Grand challenges in global mental health," *Nature*, vol. 475, no. 7354, pp. 27–30, Jul. 2011, doi: 10.1038/475027a.
[21] R. C. Kessler *et al.*, "Design and field procedures in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A)," *Int. J. Methods Psychiatr. Res.*, vol. 18, no. 2, pp. 69–83, 2009, doi: 10.1002/MPR.279.
[22] J. Canals, N. Voltas, C. Hernández-Martínez, S. Cosi, and V. Arija, "Prevalence of DSM-5 anxiety disorders, comorbidity, and persistence of symptoms in Spanish early adolescents," *Eur. Child Adolesc. Psychiatry*, vol. 28, no. 1, pp. 131–143, Jan. 2019, doi: 10.1007/s00787-018-1207-z.
[23] H. U. Wittchen *et al.*, "The size and burden of mental disorders and other disorders of the brain in Europe 2010," *Eur. Neuropsychopharmacol.*, vol. 21, no. 9, pp. 655–679, Sep. 2011, doi: 10.1016/j.euroneuro.2011.07.018.
[24] etal. Bloom, D.E., Cafiero, E.T., Jané-Llopis, E., Abrahams-Gessel, S., "The Global Economic Burden of Non-communicable," *Harvard Sch. public Heal.*, vol. 10, no. 8, pp. 910–915, 2011.
[25] A. Althubaiti, "Information bias in health research: Definition, pitfalls, and adjustment methods," *J. Multidiscip. Healthc.*, vol. 9, pp. 211–217, May 2016, doi: 10.2147/JMDH.S104807.
[26] L. J. Julian, "Measures of Anxiety," *Arthritis Care*, vol. 63, pp. 1–11, 2011, doi: 10.1002/acr.20561.Measures.
[27] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annual Review of Clinical Psychology*, vol. 4. pp. 1–32, 2008. doi: 10.1146/annurev.clinpsy.3.022806.091415.
[28] H. Johal, J. Moro, and M. Bhandari, "Principles of Evidence-Based Management of Scaphoid Fractures," *Scaphoid Fract. Evidence-Based Manag.*, pp. 7–20, Jan. 2018, doi: 10.1016/B978-0-323-48564-7.00002-2.
[29] M. Boudarene, J. J. Legros, and M. Timsit-Berthier, "Étude de la réponse de stress: Rôle de l'anxiété, du cortisol et du DHEAs," *Encephale*, vol. 28, no. 2, pp. 139–146, 2002.
[30] J. S. Lerner, R. M. Gonzalez, R. E. Dahl, A. R. Hariri, and S. E. Taylor, "Facial expressions of emotion reveal neuroendocrine and cardiovascular stress responses," *Biol. Psychiatry*, vol. 58, no. 9, pp. 743–750, Nov. 2005, doi: 10.1016/j.biopsych.2005.08.011.
[31] F. E. Ritter, R. Ceballos, A. L. Reifers, L. C. Klein, and A. Reifers, "Measuring the Effect of Dental Work as a Stressor on Cognition," University Part, PA, 2005.
[32] U. Lundberg *et al.*, "Psychophysiological stress and emg activity of the trapezius muscle," *Int. J. Behav. Med.*, vol. 1, no. 4, pp. 354–370, Dec. 1994, doi: 10.1207/s15327558ijbm0104_5.

[33] P. Jin, “Efficacy of Tai Chi, brisk walking, meditation, and reading in reducing mental and emotional stress,” *J. Psychosom. Res.*, vol. 36, no. 4, pp. 361–370, 1992, doi: 10.1016/0022-3999(92)90072-A.
[34] R. Hoehn-Saric and D. R. McLeod, “The peripheral sympathetic nervous system. Its role in normal and pathologic anxiety,” *Psychiatric Clinics of North America*, vol. 11, no. 2. pp. 375–386, 1988. doi: 10.1016/s0193-953x(18)30504-5.
[35] S. D. Kreibig, “Autonomic nervous system activity in emotion: A review,” *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, Jul. 2010, doi: 10.1016/J.BIOPSYCHO.2010.03.010.
[36] E. Lambert *et al.*, “Association between the sympathetic firing pattern and anxiety level in patients with the metabolic syndrome and elevated blood pressure,” *J. Hypertens.*, vol. 28, no. 3, pp. 543–550, 2010, doi: 10.1097/HJH.0b013e3283350ea4.
[37] J. A. Waxenbaum and M. Varacallo, *Anatomy, Autonomic Nervous System*. 2019.
[38] A. G. Guggisberg, C. W. Hess, and J. Mathis, “The significance of the sympathetic nervous system in the pathophysiology of periodic leg movements in sleep,” *Sleep*, vol. 30, no. 6, pp. 755–766, 2007, doi: 10.1093/sleep/30.6.755.
[39] S.-H. Seo and J.-T. Lee, “Stress and EEG,” in *Convergence and Hybrid Information Technologies*, 2010. doi: 10.5772/9651.
[40] M. Tanida, K. Sakatani, R. Takano, and K. Tagai, “Relation between asymmetry of prefrontal cortex activities and the autonomic nervous system during a mental arithmetic task: Near infrared spectroscopy study,” *Neurosci. Lett.*, vol. 369, no. 1, pp. 69–74, Oct. 2004, doi: 10.1016/j.neulet.2004.07.076.
[41] G. Glick and E. Braunwald, “Relative roles of the sympathetic and parasympathetic nervous systems in the reflex control of heart rate,” *Circ. Res.*, vol. 16, pp. 363–375, 1965, doi: 10.1161/01.RES.16.4.363.
[42] A. Steptoe and M. Marmot, “Impaired cardiovascular recovery following stress predicts 3-year increases in blood pressure,” *J. Hypertens.*, vol. 23, no. 3, pp. 529–536, 2005, doi: 10.1097/01.hjh.0000160208.66405.a8.
[43] H. D. Critchley, “Study of the stress response: role of anxiety, cortisol and DHEAs,” *Neurosci.*, vol. 8, no. 2, pp. 132–142, Jun. 2002, doi: 10.1177/107385840200800209.
[44] B. Weber, T. Fischer, and R. Riedl, “Brain and autonomic nervous system activity measurement in software engineering: A systematic literature review,” *J. Syst. Softw.*, vol. 178, p. 110946, Aug. 2021, doi: 10.1016/J.JSS.2021.110946.
[45] D. Lamb, “Use of behavioral measures in anxiety research.,” *Psychol. Rep.*, vol. 43, no. 3 Pt 2, pp. 1079–1085, 1978, doi: 10.2466/pr0.1978.43.3f.1079.
[46] C. D. Spielberger, F. Gonzalez-Reigosa, A. Martinez-Urrutia, L. F. S. Natalicio, and D. S. Natalicio, “The State-Trait Anxiety Inventory,” *Rev. Interam. Psicol. J. Psychol.*, vol. 5, no. 3 & 4, pp. 3–4, 1971, doi: 10.30849/RIP/IJP.V5I3.
[47] A. T. Beck, N. Epstein, G. Brown, and R. A. Steer, “An inventory for measuring clinical anxiety: Psychometric properties.,” *J. Consult. Clin. Psychol.*, vol. 56, no. 6, pp. 893–897, 1988, doi: 10.1037/0022-006X.56.6.893.
[48] C. Demetriou, B. U. Ozer, and C. A. Essau, “Self-Report Questionnaires,” *Encycl. Clin. Psychol.*, pp. 1–6, Jan. 2015, doi: 10.1002/9781118625392.WBECP507.
[49] S. R. Arikian and J. M. German, “A review of the diagnosis, pharmacologie treatment, and economic aspects of anxiety disorders,” *Prim. Care Companion J. Clin. Psychiatry*, vol. 3, no. 3, pp. 110–117, 2001, doi: 10.4088/pcc.v03n0302.
[50] H. K. Walker, W. D. Hall, and J. W. Hurst, *Clinical Methods: The History, Physical and*

Laboratory Examinations, 3rd editio. Butterworth-Heinemann, 1990.
[51] R. Kaplan and D. Saccuzzo, *Psychological testing: Principles, applications, and issues*. Boston, MA: Cengage Learning, 1982.
[52] J. A. Coan and J. J. B. Allen, "Frontal EEG asymmetry as a moderator and mediator of emotion," *Biological Psychology*, vol. 67, no. 1–2. Elsevier, pp. 7–50, 2004. doi: 10.1016/j.biopsycho.2004.03.002.
[53] R. S. Lewis, N. Y. Weekes, and T. H. Wang, "The effect of a naturalistic stressor on frontal EEG asymmetry, stress, and health," *Biol. Psychol.*, vol. 75, no. 3, pp. 239–247, 2007, doi: 10.1016/j.biopsycho.2007.03.004.
[54] O. AlShorman *et al.*, "Frontal lobe real-time EEG analysis using machine learning techniques for mental stress detection," *J. Integr. Neurosci.*, vol. 21, no. 1, p. 020, Jan. 2022, doi: 10.31083/j.jin2101020.
[55] M. Ferrari and V. Quaresima, "A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application," *NeuroImage*, vol. 63, no. 2. pp. 921–935, Nov. 2012. doi: 10.1016/j.neuroimage.2012.03.049.
[56] P. Pinti *et al.*, "The present and future use of functional near-infrared spectroscopy (Fnirs) for cognitive neuroscience," *Ann. N. Y. Acad. Sci.*, vol. 1464, no. 1, pp. 5–29, 2020, doi: 10.1111/nyas.13948.
[57] J. M. Hooker and R. E. Carson, "Human Positron Emission Tomography Neuroimaging," *Annual Review of Biomedical Engineering*, vol. 21. pp. 551–581, 2019. doi: 10.1146/annurev-bioeng-062117-121056.
[58] F. Al-Shargie, T. B. Tang, and M. Kiguchi, "Mental stress grading based on fNIRS signals," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., Oct. 2016, pp. 5140–5143. doi: 10.1109/EMBC.2016.7591884.
[59] F. Al-Shargie, T. B. Tang, and M. Kiguchi, "Stress Assessment Based on Decision Fusion of EEG and fNIRS Signals," *IEEE Access*, vol. 5, pp. 19889–19896, Sep. 2017, doi: 10.1109/ACCESS.2017.2754325.
[60] R. B. Malmo and C. Shagass, "Heart rate variability during a stress situation in psychiatric patients with and without frontal lobe operation," *Rev. Can. Biol.*, vol. 7, no. 1, p. 188, 1948.
[61] R. W. Picard, E. Vyzas, and J. A. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, 2001, doi: 10.1109/34.954607.
[62] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, 2005, doi: 10.1109/TITS.2005.848368.
[63] H. D. Critchley, J. Eccles, and S. N. Garfinkel, "Interaction between cognition, emotion, and the autonomic nervous system," in *Handbook of Clinical Neurology*, vol. 117, 2013, pp. 59–77. doi: 10.1016/B978-0-444-53491-0.00006-7.
[64] B. M. Sayers, "Analysis of Heart Rate Variability," *Ergonomics*, vol. 16, no. 1, pp. 17–32, 1973, doi: 10.1080/00140137308924479.
[65] M. M. Pulopulos, M. A. Vanderhasselt, and R. De Raedt, "Association between changes in heart rate variability during the anticipation of a stressful situation and the stress-induced cortisol response," *Psychoneuroendocrinology*, vol. 94, pp. 63–71, Aug. 2018, doi: 10.1016/J.PSYNEUEN.2018.05.004.

[66] F. Lombardi, “Physiological Understanding of HRV Components,” *Dyn. Electrocardiogr.*, pp. 40–47, 2007, doi: 10.1002/9780470987483.ch5.
[67] J. A. Burdick and J. T. Scarbrough, “Heart rate and heart-rate variability: an attempt to clarify.,” *Percept. Mot. Skills*, vol. 26, no. 3, pp. 1047–1048, Aug. 1968, doi: 10.2466/pms.1968.26.3c.1047.
[68] P. S. Blitz, J. Hoogstraten, and G. Mulder, “Mental load, heart rate and heart rate variability,” *Psychol. Forsch.*, vol. 33, no. 4, pp. 277–288, 1970, doi: 10.1007/BF00424555.
[69] R. K. Dishman, Y. Nakamura, M. E. Garcia, R. W. Thompson, A. L. Dunn, and S. N. Blair, “Heart rate variability, trait anxiety, and perceived stress among physically fit men and women,” *Int. J. Psychophysiol.*, vol. 37, no. 2, pp. 121–133, 2000, doi: 10.1016/S0167-8760(00)00085-4.
[70] G. Lu *et al.*, “A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects,” *J. Med. Eng. Technol.*, vol. 33, no. 8, pp. 634–641, Nov. 2009, doi: 10.3109/03091900903150998.
[71] M. L. Reyes, J. D. Lee, Y. Liang, J. D. Hoffman, and R. W. Huang, “Capturing Driver Response to In-Vehicle Human-Machine Interface Technologies Using Facial Thermography,” *Driv. Assesment Conf.*, vol. 5, no. 2009, pp. 536–542, Jun. 2009, doi: 10.17077/DRIVINGASSESSMENT.1368.
[72] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, “Smart homes that monitor breathing and heart rate,” in *Conference on Human Factors in Computing Systems - Proceedings*, 2015, pp. 837–846. doi: 10.1145/2702123.2702200.
[73] L. Giovangrandi, O. T. Inan, R. M. Wiard, M. Etemadi, and G. T. A. Kovacs, “Ballistocardiography - A method worth revisiting,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, NIH Public Access, 2011, pp. 4279–4282. doi: 10.1109/IEMBS.2011.6091062.
[74] A. Moliné *et al.*, “The mental nose and the Pinocchio effect: Thermography, planning, anxiety, and lies,” *J. Investig. Psychol. Offender Profiling*, vol. 15, no. 2, pp. 234–248, Jun. 2018, doi: 10.1002/jip.1505.
[75] G. Krantz, M. Forsman, and U. Lundberg, “Consistency in physiological stress responses and electromyographic activity during induced stress exposure in women and men,” *Integr. Physiol. Behav. Sci.*, vol. 39, no. 2, pp. 105–118, Apr. 2004, doi: 10.1007/BF02734276.
[76] R. Merletti, M. Aventaggiato, A. Botter, A. Holobar, H. Marateb, and T. M. M. Vieira, “Advances in Surface EMG: Recent Progress in Detection and Processing Techniques,” *Crit. Rev. Biomed. Eng.*, vol. 38, no. 4, pp. 305–345, 2010, doi: 10.1615/CritRevBiomedEng.v38.i4.10.
[77] R. Merletti, A. Botter, C. Cescon, M. A. Minetto, and T. M. M. Vieira, “Advances in Surface EMG: Recent Progress in Clinical Research Applications,” *Crit. Rev. Biomed. Eng.*, vol. 38, no. 4, pp. 347–379, 2010, doi: 10.1615/CritRevBiomedEng.v38.i4.20.
[78] D. Bansevicius, R. H. Westgaard, and C. Jensen, “Mental Stress of Long Duration: EMG Activity, Perceived Tension, Fatigue, and Pain Development in Pain-Free Subjects,” *Headache*, vol. 39, no. 8, pp. 499–510, 1997.
[79] D. Rissén, B. Melin, L. Sandsjö, I. Dohns, and U. Lundberg, “Surface EMG and psychophysiological stress reactions in women during repetitive work,” *Eur. J. Appl. Physiol.*, vol. 83, no. 2–3, pp. 215–222, 2000, doi: 10.1007/s004210000281.

[80] L. M. Schleifer, T. W. Spalding, S. E. Kerick, J. R. Cram, R. Ley, and B. D. Hatfield, "Mental stress and trapezius muscle activation under psychomotor challenge: A focus on EMG gaps during computer work," *Psychophysiology*, vol. 45, no. 3, pp. 356–365, 2008, doi: 10.1111/j.1469-8986.2008.00645.x.
[81] J. Wijsman, B. Grundlehner, J. Penders, and H. Hermens, "Trapezius muscle EMG as predictor of mental stress," *Trans. Embed. Comput. Syst.*, vol. 12, no. 4, Jun. 2013, doi: 10.1145/2485984.2485987.
[82] O. Hidaka, M. Yanagi, and K. Takada, "Mental stress-induced physiological changes in the human masseter muscle," *J. Dent. Res.*, vol. 83, no. 3, pp. 227–231, 2004, doi: 10.1177/154405910408300308.
[83] O. Hidaka, M. Yanagi, and K. Takada, "Changes in masseteric hemodynamics time-related to mental stress," *J. Dent. Res.*, vol. 83, no. 2, pp. 185–190, Feb. 2004, doi: 10.1177/154405910408300220.
[84] C. M. Tsai, S. L. Chou, E. N. Gale, and W. D. Mccall, "Human masticatory muscle activity and jaw position under experimental stress," *J. Oral Rehabil.*, vol. 29, no. 1, pp. 44–51, 2002, doi: 10.1046/j.1365-2842.2002.00810.x.
[85] P. Grossman, "Respiration, Stress, and Cardiovascular Function," *Psychophysiology*, vol. 20, no. 3, pp. 284–300, 1983, doi: 10.1111/j.1469-8986.1983.tb02156.x.
[86] W. M. Suess, A. B. Alexander, D. D. Smith, H. W. Sweeney, and R. J. Marion, "The Effects of Psychological Stress on Respiration: A Preliminary Study of Anxiety and Hyperventilation," *Psychophysiology*, vol. 17, no. 6, pp. 535–540, Nov. 1980, doi: 10.1111/j.1469-8986.1980.tb02293.x.
[87] L. Anishchenko and A. Turetzkaya, "Improved Non-Contact Mental Stress Detection via Bioradar," in *Proceedings of the International Conference on Biomedical Innovations and Applications, BIA 2020*, 2020, pp. 21–24. doi: 10.1109/BIA50171.2020.9244492.
[88] J. R. Machado Fernández and L. Anishchenko, "Mental stress detection using bioradar respiratory signals," *Biomed. Signal Process. Control*, vol. 43, pp. 244–249, May 2018, doi: 10.1016/j.bspc.2018.03.006.
[89] Y. Shan, S. Li, and T. Chen, "Respiratory signal and human stress: non-contact detection of stress with a low-cost depth sensing camera," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 8, pp. 1825–1837, Aug. 2020, doi: 10.1007/S13042-020-01074-X.
[90] W. Seo, N. Kim, S. Kim, C. Lee, and S. M. Park, "Deep ECG-respiration network (DeepER net) for recognizing mental stress," *Sensors (Switzerland)*, vol. 19, no. 13, 2019, doi: 10.3390/s19133021.
[91] W. Seo, N. Kim, S. Kim, C. Lee, and S. M. Park, "Deep ECG-respiration network (DeepER net) for recognizing mental stress," *Sensors (Switzerland)*, vol. 19, no. 13, Jul. 2019, doi: 10.3390/s19133021.
[92] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tr, and U. Ehlert, "Discriminating Stress From Cognitive Load Using a Wearable EDA Device," *Technology*, vol. 14, no. 2, pp. 410–417, 2010.
[93] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 2, pp. 279–286, Mar. 2012, doi: 10.1109/TITB.2011.2169804.
[94] S. Amalan *et al.*, "Electrodermal Activity based Classification of Induced Stress in a Controlled Setting," in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, IEEE, Jun. 2018, pp. 1–6. doi:

10.1109/MeMeA.2018.8438703.
[95] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, “Review on Psychological Stress Detection Using Biosignals,” *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 440–460, Jan. 2022, doi: 10.1109/TAFFC.2019.2927337.
[96] K. Palanisamy, M. Murugappan, and S. Yaacob, “Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress,” *J. Phys. Ther. Sci.*, vol. 24, no. 12, pp. 1341–1344, 2012, doi: 10.1589/jpts.24.1341.
[97] A. Barreto, J. Zhai, and M. Adjouadi, “Non-intrusive physiological monitoring for automated stress detection in human-computer interaction,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4796 LNCS. pp. 29–38, 2007. doi: 10.1007/978-3-540-75773-3_4.
[98] T. Yamakoshi *et al.*, “Feasibility study on driver’s stress detection from differential skin temperature measurement,” *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS’08 - "Personalized Healthc. through Technol.*, pp. 1076–1079, 2008, doi: 10.1109/iembs.2008.4649346.
[99] A. Masaki, K. Nagumo, B. Lamsal, K. Oiwa, and A. Nozawa, “Anomaly detection in facial skin temperature using variational autoencoder,” *Artif. Life Robot.*, vol. 26, no. 1, pp. 122–128, 2021, doi: 10.1007/s10015-020-00634-2.
[100] T. Partala and V. Surakka, “Pupil size variation as an indication of affective processing,” *Int. J. Hum. Comput. Stud.*, vol. 59, no. 1–2, pp. 185–198, 2003, doi: 10.1016/S1071-5819(03)00017-X.
[101] M. Pedrotti *et al.*, “Automatic Stress Classification With Pupil Diameter Analysis,” *Int. J. Hum. Comput. Interact.*, vol. 30, no. 3, pp. 220–236, 2014, doi: 10.1080/10447318.2013.848320.
[102] S. Baltaci and D. Gokcay, “Stress Detection in Human–Computer Interaction: Fusion of Pupil Dilation and Facial Temperature Features,” *Int. J. Hum. Comput. Interact.*, vol. 32, no. 12, pp. 956–966, 2016, doi: 10.1080/10447318.2016.1220069.
[103] P. Ren, A. Barreto, J. Huang, Y. Gao, F. R. Ortega, and M. Adjouadi, “Off-line and on-line stress detection through processing of the pupil diameter signal,” *Ann. Biomed. Eng.*, vol. 42, no. 1, pp. 162–176, 2014, doi: 10.1007/s10439-013-0880-9.
[104] S. Baltaci and D. Gokcay, “Role of pupil dilation and facial temperature features in stress detection,” in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, IEEE, Apr. 2014, pp. 1259–1262. doi: 10.1109/SIU.2014.6830465.
[105] I. Lefter, G. J. Burghouts, and L. J. M. Rothkrantz, “Recognizing Stress Using Semantics and Modulation of Speech and Gestures,” *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 162–175, Apr. 2016, doi: 10.1109/TAFFC.2015.2451622.
[106] K. Schindler, L. Van Gool, and B. de Gelder, “Recognizing emotions expressed by body pose: A biologically inspired neural model,” *Neural Networks*, vol. 21, no. 9, pp. 1238–1246, 2008, doi: 10.1016/j.neunet.2008.05.003.
[107] S. Gedam and S. Paul, “A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques,” *IEEE Access*, vol. 9, pp. 84045–84066, 2021, doi: 10.1109/ACCESS.2021.3085502.
[108] K. H. Kim, S. W. Bang, and S. R. Kim, “Emotion recognition system using short-term monitoring of physiological signals,” *Med. Biol. Eng. Comput.*, vol. 42, no. 3, pp. 419–427, May 2004, doi: 10.1007/BF02344719.

[109] G. Giannakakis *et al.*, “Stress and anxiety detection using facial cues from videos,” *Biomed. Signal Process. Control*, vol. 31, pp. 89–101, 2017, doi: 10.1016/j.bspc.2016.06.020.
[110] L. J. M. Rothkrantz, P. Wiggers, J. W. A. Van Wees, and R. J. Van Vark, “Voice stress analysis,” in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, Springer Verlag, 2004, pp. 449–456. doi: 10.1007/978-3-540-30120-2_57.
[111] M. Obuchi *et al.*, “Predicting brain functional connectivity using mobile sensing,” *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–22, 2020, doi: 10.1145/3381001.
[112] L. Pepa, A. Sabatelli, L. Ciabattoni, A. Monteriù, F. Lamberti, and L. Morra, “Stress Detection in Computer Users from Keyboard and Mouse Dynamics,” *IEEE Trans. Consum. Electron.*, vol. 67, no. 1, pp. 12–19, Feb. 2021, doi: 10.1109/TCE.2020.3045228.
[113] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006, doi: 10.1016/J.SPECOM.2006.04.003.
[114] B. Liao, Q. Zhao, X. Dong, L. Zhao, and Y. Zhang, “The two-stage approximation of coupled-mode theory for optical fiber and its application to an all-fiber acousto-optic modulator,” in *Optoelectronic Devices and Integration*, H. Ming, X. Zhang, and M. Y. Chen, Eds., SPIE, Jan. 2005, p. 70. doi: 10.1117/12.576897.
[115] L. M. Vizer, L. Zhou, and A. Sears, “Automated stress detection using keystroke and linguistic features: An exploratory study,” *Int. J. Hum. Comput. Stud.*, vol. 67, no. 10, pp. 870–886, Oct. 2009, doi: 10.1016/J.IJHCS.2009.07.005.
[116] S. Scherer, G. Stratou, J. Gratch, and L. P. Morency, “Investigating voice quality as a speaker-independent indicator of depression and PTSD,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 847–851, 2013, doi: 10.21437/interspeech.2013-240.
[117] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer, “The ‘Trier social stress test’ - A tool for investigating psychobiological stress responses in a laboratory setting,” in *Neuropsychobiology*, 1993, pp. 76–81. doi: 10.1159/000119004.
[118] M. R. Milad and G. J. Quirk, “Fear extinction as a model for translational neuroscience: Ten years of progress,” *Annu. Rev. Psychol.*, vol. 63, pp. 129–151, 2012, doi: 10.1146/annurev.psych.121208.131631.
[119] W. Lovallo, “The Cold Pressor Test and Autonomic Function: A Review and Integration,” *Psychophysiology*, vol. 12, no. 3, pp. 268–282, 1975, doi: 10.1111/j.1469-8986.1975.tb01289.x.
[120] O. D. Kothgassner, A. Goreis, I. Bauda, A. Ziegenaus, L. M. Glenk, and A. Felnhofer, “Virtual reality biofeedback interventions for treating anxiety: A systematic review, meta-analysis and future perspective,” *Wien. Klin. Wochenschr.*, vol. 134, no. Suppl 1, p. 49, Jan. 2022, doi: 10.1007/S00508-021-01991-Z.
[121] A. P. Cruz, A. Pradeep, K. R. Sivasankar, and K. . Krishnaveni, “A Decision Tree Optimised SVM Model for Stress Detection using Biosignals,” in *2020 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, Jul. 2020, pp. 0841–0845. doi: 10.1109/ICCSP48568.2020.9182043.
[122] F. Delmastro, F. Di Martino, and C. Dolciotti, “Cognitive Training and Stress Detection in MCI Frail Older People through Wearable Sensors and Machine Learning,” *IEEE Access*,

vol. 8, pp. 65573–65590, 2020, doi: 10.1109/ACCESS.2020.2985301.
[123] C. Goumopoulos and N. Potha, “Mental fatigue detection using a wearable commodity device and machine learning,” *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789, 2022, doi: 10.1007/s12652-021-03674-z.
[124] H. J. Han, S. Labbaf, J. L. Borelli, N. Dutt, and A. M. Rahmani, “Objective stress monitoring based on wearable sensors in everyday settings,” *J. Med. Eng. Technol.*, vol. 44, no. 4, pp. 177–189, 2020, doi: 10.1080/03091902.2020.1759707.
[125] M. Mozafari, F. Firouzi, and B. Farahani, “Towards IoT-enabled Multimodal Mental Stress Monitoring,” in *2020 International Conference on Omni-Layer Intelligent Systems, COINS 2020*, 2020, pp. 1–8. doi: 10.1109/COINS49042.2020.9191392.
[126] M. Mozafari, R. Goubran, and J. R. Green, “A fusion model for cross-subject stress level detection based on transfer learning,” in *2021 IEEE Sensors Applications Symposium, SAS 2021 - Proceedings*, 2021, pp. 1–6. doi: 10.1109/SAS51076.2021.9530085.
[127] S. Sultana, M. A. Rahman, and M. Zavid Parvez, “Detection of stress for visually impaired people using EEG signals based on time-frequency domain analysis,” in *Proceedings - International Conference on Machine Learning and Cybernetics*, 2020, pp. 118–123. doi: 10.1109/ICMLC51923.2020.9469562.
[128] P. Bobade and M. Vani, “Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data,” *Proc. 2nd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2020*, pp. 51–57, Jul. 2020, doi: 10.1109/ICIRCA48905.2020.9183244.
[129] E. Alyan, N. M. Saad, N. Kamel, and M. A. Rahman, “Investigating Frontal Neurovascular Coupling in Response to Workplace Design-Related Stress,” *IEEE Access*, vol. 8, pp. 218911–218923, 2020, doi: 10.1109/ACCESS.2020.3040540.
[130] P. Schmidt, R. Dürichen, A. Reiss, K. Van Laerhoven, and T. Plötz, “Multi-target Affect Detection in the Wild: An exploratory study,” in *Proceedings - International Symposium on Wearable Computers, ISWC*, 2019, pp. 211–219. doi: 10.1145/3341163.3347741.
[131] A. Akbas, “Evaluation of the physiological data indicating the dynamic stress level of drivers,” *Sci. Res. Essays*, vol. 6, no. 2, pp. 430–439, 2011, doi: 10.5897/SRE10.943.
[132] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, “Monitoring stress with a wrist device using context,” *J. Biomed. Inform.*, vol. 73, pp. 159–170, 2017, doi: 10.1016/j.jbi.2017.08.006.
[133] M. Dziezyc *et al.*, “How to catch them all? Enhanced data collection for emotion recognition in the field,” in *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 348–351. doi: 10.1109/PerComWorkshops51409.2021.9431143.
[134] S. Saganowski, “Bringing Emotion Recognition out of the Lab into Real Life: Recent Advances in Sensors and Machine Learning,” *Electronics (Switzerland)*, vol. 11, no. 3. Multidisciplinary Digital Publishing Institute, p. 496, Feb. 08, 2022. doi: 10.3390/electronics11030496.
[135] E. E. Kaczor, B. Chapman, S. Carreiro, P. Indic, and J. Stapp, “Objective measurement of physician stress in the emergency department using a wearable sensor,” in *Proceedings of the Annual Hawaii International Conference on System Sciences*, NIH Public Access, 2020, pp. 3729–3738. doi: 10.24251/hicss.2020.456.
[136] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, “Continuous stress detection using a wrist device - in laboratory and real life,” in *UbiComp 2016 Adjunct - Proceedings of the*

2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2016, pp. 1185–1193. doi: 10.1145/2968219.2968306.
[137] F. Larradet, R. Niewiadomski, G. Barresi, D. G. Caldwell, and L. S. Mattos, "Toward Emotion Recognition From Physiological Signals in the Wild: Approaching the Methodological Issues in Real-Life Data Collection," *Frontiers in Psychology*, vol. 11. Frontiers Media S.A., p. 1111, Jul. 15, 2020. doi: 10.3389/fpsyg.2020.01111.
[138] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals.," *Circulation*, vol. 101, no. 23, Jun. 2000, doi: 10.1161/01.cir.101.23.e215.
[139] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," *ICMI 2018 - Proc. 2018 Int. Conf. Multimodal Interact.*, pp. 400–408, Oct. 2018, doi: 10.1145/3242969.3242985.
[140] M. Dzieżyc, M. Gjoreski, P. Kazienko, S. Saganowski, and M. Gams, "Can we ditch feature engineering? End-to-end deep learning for affect recognition from physiological sensor data," *Sensors (Switzerland)*, vol. 20, no. 22, pp. 1–21, 2020, doi: 10.3390/s20226535.
[141] M. Elgendi, V. Galli, C. Ahmadizadeh, and C. Menon, "Dataset of Psychological Scales and Physiological Signals Collected for Anxiety Assessment Using a Portable Device," *Data*, vol. 7, no. 9, 2022, doi: 10.3390/data7090132.
[142] E. Thompson, "Hamilton Rating Scale for Anxiety (HAM-A)," *Occup. Med. (Chic. Ill).*, vol. 65, no. 7, p. 601, Oct. 2015, doi: 10.1093/occmed/kqv054.
[143] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang, "UBFC-Phys: A Multimodal Database For Psychophysiological Studies of Social Stress," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 622–636, 2023, doi: 10.1109/TAFFC.2021.3056960.
[144] N. El Haouij, J. M. Poggi, S. Sevestre-Ghalila, R. Ghozi, and M. Jadane, "AffectiveROAD system and database to assess driver's attention," *Proc. ACM Symp. Appl. Comput.*, pp. 800–803, Apr. 2018, doi: 10.1145/3167132.3167395.
[145] M. Gjoreski *et al.*, "Datasets for Cognitive Load Inference Using Wearable Sensors and Psychological Traits," *Appl. Sci. 2020, Vol. 10, Page 3843*, vol. 10, no. 11, p. 3843, May 2020, doi: 10.3390/APP10113843.
[146] M. Obuchi *et al.*, "Predicting brain functional connectivity using mobile sensing," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–22, 2020, doi: 10.1145/3381001.
[147] B. Shickel, S. Siegel, M. Heesacker, S. Benton, and P. Rashidi, "Automatic Detection and Classification of Cognitive Distortions in Mental Health Text," in *Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020*, 2020, pp. 275–280. doi: 10.1109/BIBE50027.2020.00052.
[148] R. Sánchez-Reolid, M. T. López, and A. Fernández-Caballero, "Machine Learning for Stress Detection from Electrodermal Activity: A Scoping Review," no. November, pp. 1–29, 2020, doi: 10.20944/preprints202011.0043.v1.
[149] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy 2021, Vol. 23, Page 18*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/E23010018.
[150] A. Akella *et al.*, "Classifying Multi-Level Stress Responses from Brain Cortical EEG in Nurses and Non-Health Professionals Using Machine Learning Auto Encoder," *IEEE J. Transl. Eng. Heal. Med.*, vol. 9, pp. 1–9, 2021, doi: 10.1109/JTEHM.2021.3077760.

[151] Y. Badr, F. Al-Shargie, U. Tariq, F. Babiloni, F. Al Mughairbi, and H. Al-Nashash, "Classification of Mental Stress using Dry EEG Electrodes and Machine Learning," in *2023 Advances in Science and Engineering Technology International Conferences, ASET 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ASET56582.2023.10180884.
[152] S. Campanella, A. Altaleb, A. Belli, P. Pierleoni, and L. Palma, "A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques," *Sensors 2023, Vol. 23, Page 3565*, vol. 23, no. 7, p. 3565, Mar. 2023, doi: 10.3390/S23073565.
[153] K. M. Dalmeida and G. L. Masala, "Hrv features as viable physiological markers for stress detection using wearable devices," *Sensors*, vol. 21, no. 8, p. 2873, Apr. 2021, doi: 10.3390/s21082873.
[154] Y. Ding, J. Liu, X. Zhang, and Z. Yang, "Dynamic Tracking of State Anxiety via Multi-Modal Data and Machine Learning," *Front. Psychiatry*, vol. 13, Mar. 2022, doi: 10.3389/fpsyt.2022.757961.
[155] E. C. Erkus, V. Purutcuoglu, F. Ari, and D. Gokcay, "Comparison of Several Machine Learning Classifiers for Arousal Classification: A Preliminary study," in *2020 Medical Technologies Congress (TIPTEKNO)*, IEEE, Nov. 2020, pp. 1–7. doi: 10.1109/TIPTEKNO50054.2020.9299316.
[156] J. Henry, H. Lloyd, M. Turner, and C. Kendrick, "On the robustness of machine learning models for stress and anxiety recognition from heart activity signals," *IEEE Sens. J.*, Jul. 2023, doi: 10.1109/JSEN.2023.3276413.
[157] N. Kim, W. Seo, S. Kim, and S. M. Park, "Electrogastrogram: Demonstrating Feasibility in Mental Stress Assessment Using Sensor Fusion," *IEEE Sens. J.*, vol. 21, no. 13, pp. 14503–14514, 2021, doi: 10.1109/JSEN.2020.3026717.
[158] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection from speech and Galvanic Skin Response signals," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, IEEE, Jun. 2013, pp. 209–214. doi: 10.1109/CBMS.2013.6627790.
[159] K. Lingelbach, M. Bui, F. DIederichs, and M. Vukelic, "Exploring Conventional, Automated and Deep Machine Learning for Electrodermal Activity-Based Drivers' Stress Recognition," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2021, pp. 1339–1344. doi: 10.1109/SMC52423.2021.9658662.
[160] Y. Liu, H. Li, J. Wang, H. Zhang, and X. Zheng, "Psychological stress detection based on heart rate variability," in *International Conference on Electronic Information Engineering and Computer Science (EIECS 2022)*, Y. Yue, Ed., SPIE, Apr. 2023, p. 150. doi: 10.1117/12.2668856.
[161] M. Mamdouh, R. Mahmoud, O. Attallah, and A. Al-Kabbany, "Stress Detection in the Wild: On the Impact of Cross-Training on Mental State Detection," pp. 150–158, Jun. 2023, doi: 10.1109/NRSC58893.2023.10153050.
[162] Q. Meteier *et al.*, "Classification of Drivers' Workload Using Physiological Signals in Conditional Automation," *Front. Psychol.*, vol. 12, 2021, doi: 10.3389/fpsyg.2021.596038.
[163] Q. Meteier *et al.*, "Relevant Physiological Indicators for Assessing Workload in Conditionally Automated Driving, Through Three-Class Classification and Regression," *Front. Comput. Sci.*, vol. 3, 2022, doi: 10.3389/fcomp.2021.775282.
[164] Q. Meteier *et al.*, "A dataset on the physiological state and behavior of drivers in

conditionally automated driving," *Data Br.*, vol. 47, p. 109027, Apr. 2023, doi: 10.1016/J.DIB.2023.109027.
[165] O. M. Mozos *et al.*, "Stress detection using wearable physiological and sociometric sensors," *Int. J. Neural Syst.*, vol. 27, no. 2, Mar. 2017, doi: 10.1142/S0129065716500416.
[166] M. Naegelin *et al.*, "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," *J. Biomed. Inform.*, vol. 139, p. 104299, Mar. 2023, doi: 10.1016/J.JBI.2023.104299.
[167] A. Pinge, S. Bandyopadhyay, S. Ghosh, and S. Sen, "A Comparative Study between ECG-based and PPG-based Heart Rate Monitors for Stress Detection," *2022 14th Int. Conf. Commun. Syst. NETworkS, COMSNETS 2022*, pp. 84–89, 2022, doi: 10.1109/COMSNETS53615.2022.9668342.
[168] K. Plarre *et al.*, "Continuous inference of psychological stress from sensory measurements collected in the natural environment," in *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks, IPSN'11*, 2011, pp. 97–108.
[169] V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, and O. M. Mozos, "Stress detection using wearable physiological sensors," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9107. pp. 526–532, 2015. doi: 10.1007/978-3-319-18914-7_55.
[170] V. Sandulescu, S. Andrews, D. Ellis, R. Dobrescu, and O. Martinez-Mozos, "Mobile app for stress monitoring using voice features," in *2015 E-Health and Bioengineering Conference, EHB 2015*, IEEE, Nov. 2016, pp. 1–4. doi: 10.1109/EHB.2015.7391411.
[171] O. Shaposhnyk, S. Yanushkevich, V. Babenko, M. Chernykh, and I. Nastenko, "Inferring Cognitive Load Level from Physiological and Personality Traits," *Int. Conf. Inf. Digit. Technol. 2023, IDT 2023*, pp. 233–242, 2023, doi: 10.1109/IDT59031.2023.10194430.
[172] A. I. Siam, S. A. Gamel, and F. M. Talaat, "Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques," *Neural Comput. Appl.*, vol. 35, no. 17, pp. 12891–12904, Jun. 2023, doi: 10.1007/S00521-023-08428-W/TABLES/9.
[173] E. Silva, J. Aguiar, L. P. Reis, J. O. e. Sá, J. Gonçalves, and V. Carvalho, "Stress among Portuguese Medical Students: the EuStress Solution," *J. Med. Syst.*, vol. 44, no. 2, 2020, doi: 10.1007/s10916-019-1520-1.
[174] A. R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel, and A. S. Malik, "Machine learning framework for the detection of mental stress at multiple levels," *IEEE Access*, vol. 5, pp. 13545–13556, Jul. 2017, doi: 10.1109/ACCESS.2017.2723622.
[175] M. Vaz, T. Summavielle, R. Sebastião, and R. P. Ribeiro, "Multimodal Classification of Anxiety Based on Physiological Signals," *Appl. Sci.*, vol. 13, no. 11, 2023, doi: 10.3390/app13116368.
[176] M. Xing, J. M. Fitzgerald, and H. Klumpp, "Classification of Social Anxiety Disorder With Support Vector Machine Analysis Using Neural Correlates of Social Signals of Threat," *Front. Psychiatry*, vol. 11, Mar. 2020, doi: 10.3389/fpsyt.2020.00144.
[177] L. Zhu *et al.*, "Stress Detection Through Wrist-Based Electrodermal Activity Monitoring and Machine Learning," *IEEE J. Biomed. Heal. Informatics*, vol. 27, no. 5, pp. 2155–2165, May 2023, doi: 10.1109/JBHI.2023.3239305.
[178] L. A. Kader, F. Yahya, U. Tariq, and H. Al-Nashash, "Mental Stress Assessment Using Low in Cost Single Channel EEG System," *2023 Adv. Sci. Eng. Technol. Int. Conf. ASET*

2023, 2023, doi: 10.1109/ASET56582.2023.10180651.
[179] M. R. Bin Mazlan, A. S. B. A. Sukor, A. H. Bin Adom, R. B. Jamaluddin, and S. A. B. Awang, "Investigation of Different Classifiers for Stress Level Classification using PCA-Based Machine Learning Method," *2023 19th IEEE Int. Colloq. Signal Process. Its Appl. CSPA 2023 - Conf. Proc.*, pp. 168–173, 2023, doi: 10.1109/CSPA57446.2023.10087367.
[180] N. Rashid, T. Mortlock, and M. A. Al Faruque, "Stress Detection using Context-Aware Sensor Fusion from Wearable Devices," *IEEE Internet Things J.*, Aug. 2023, doi: 10.1109/JIOT.2023.3265768.
[181] J. Wijsman, B. Grundlehner, H. Liu, H. Hermens, and J. Penders, "Towards mental stress detection using wearable physiological sensors," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, IEEE, Aug. 2011, pp. 1798–1801. doi: 10.1109/IEMBS.2011.6090512.
[182] L. Arya, H. Chowdhary, I. Agrawal, and I. Sreedevi, "Towards Accurate Stress Classification: Combining Advanced Feature Selection and Deep Learning," *2023 3rd IEEE Int. Conf. Softw. Eng. Artif. Intell. SEAI 2023*, pp. 47–52, 2023, doi: 10.1109/SEAI59139.2023.10217367.
[183] M. Benchekroun *et al.*, "Comparison of Stress Detection through ECG and PPG signals using a Random Forest-based Algorithm," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2022-July, pp. 3150–3153, 2022, doi: 10.1109/EMBC48229.2022.9870984.
[184] K. Dahal, B. Bogue-Jimenez, and A. Doblas, "Global Stress Detection Framework Combining a Reduced Set of HRV Features and Random Forest Model," *Sensors*, vol. 23, no. 11, p. 5220, Jun. 2023, doi: 10.3390/S23115220/S1.
[185] A. H. Gazi *et al.*, "Respiratory markers significantly enhance anxiety detection using multimodal physiological sensing," in *BHI 2021 - 2021 IEEE EMBS International Conference on Biomedical and Health Informatics, Proceedings*, 2021, pp. 1–4. doi: 10.1109/BHI50953.2021.9508589.
[186] M. A. Quadir, S. Bhardwaj, N. Verma, A. K. Sivaraman, and K. F. Tee, "IoT-Based Mental Health Monitoring System Using Machine Learning Stress Prediction Algorithm in Real-Time Application," *Lect. Notes Electr. Eng.*, vol. 1021 LNEE, pp. 249–263, 2023, doi: 10.1007/978-981-99-1051-9_16.
[187] P. Karthikeyan, M. Murugappan, and S. Yaacob, "EMG Signal Based Human Stress Level Classification Using Wavelet Packet Transform," in *Communications in Computer and Information Science*, vol. 330 CCIS, 2012, pp. 236–243. doi: 10.1007/978-3-642-35197-6_26.
[188] A. Appriou, A. Cichocki, and F. Lotte, "Modern Machine-Learning Algorithms: For Classifying Cognitive and Affective States From Electroencephalography Signals," *IEEE Syst. Man, Cybern. Mag.*, vol. 6, no. 3, pp. 29–38, 2020, doi: 10.1109/msmc.2020.2968638.
[189] A. H. Assaf, H. Ben Abdessalem, and C. Frasson, "Detecting Mental Fatigue in Intelligent Tutoring Systems," pp. 66–74, 2023, doi: 10.1007/978-3-031-32883-1_6/FIGURES/2.
[190] S. Dhaouadi and M. M. Ben Khelifa, "A multimodal Physiological-Based Stress Recognition: Deep Learning Models' Evaluation in Gamers' Monitoring Application," in *2020 International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2020*, IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/ATSIP49331.2020.9231666.

[191] S. Praveenkumar. and T. Karthick, “Automatic Stress Recognition System with Deep Learning using Multimodal Psychological Data,” in *Proceedings of the 2022 International Conference on Electronic Systems and Intelligent Computing, ICESIC 2022*, IEEE, Apr. 2022, pp. 122–127. doi: 10.1109/ICESIC53714.2022.9783595.
[192] A. Ragav, N. H. Krishna, N. Narayanan, K. Thelly, and V. Vijayaraghavan, “Scalable deep learning for stress and affect detection on resource-constrained devices,” in *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, 2019, pp. 1585–1592. doi: 10.1109/ICMLA.2019.00261.
[193] M. Dziezyc, M. Gjoreski, P. Kazienko, S. Saganowski, and M. Gams, “Can we ditch feature engineering? End-to-end deep learning for affect recognition from physiological sensor data,” *Sensors (Switzerland)*, vol. 20, no. 22, pp. 1–21, Nov. 2020, doi: 10.3390/s20226535.
[194] X. Ying, “An Overview of Overfitting and its Solutions,” *J. Phys. Conf. Ser.*, vol. 1168, no. 2, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.
[195] M. Amin, K. Ullah, M. Asif, H. Shah, A. Mehmood, and M. A. Khan, “Real-World Driver Stress Recognition and Diagnosis Based on Multimodal Deep Learning and Fuzzy EDAS Approaches,” *Diagnostics*, vol. 13, no. 11, p. 1897, May 2023, doi: 10.3390/diagnostics13111897.
[196] H. Barki and W. Y. Chung, “Mental Stress Detection Using a Wearable In-Ear Plethysmography,” *Biosens. 2023, Vol. 13, Page 397*, vol. 13, no. 3, p. 397, Mar. 2023, doi: 10.3390/BIOS13030397.
[197] D. Chatterjee, S. Dutta, R. Shaikh, and S. K. Saha, “A lightweight deep neural network for detection of mental states from physiological signals,” *Innov. Syst. Softw. Eng.*, pp. 1–8, Jul. 2022, doi: 10.1007/s11334-022-00470-6.
[198] R. K. Sah, M. J. Cleveland, A. Habibi, and H. Ghasemzadeh, “Stressalyzer: Convolutional Neural Network Framework for Personalized Stress Classification,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2022-July, pp. 4658–4663, 2022, doi: 10.1109/EMBC48229.2022.9871842.
[199] L. Huynh, T. Nguyen, T. Nguyen, S. Pirttikangas, and P. Siirtola, “StressNAS: Affect State and Stress Detection Using Neural Architecture Search,” in *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, New York, NY, USA: ACM, Sep. 2021, pp. 121–125. doi: 10.1145/3460418.3479320.
[200] Y. Badr, F. Al-Shargie, U. Tariq, F. Babiloni, F. Al Mughairbi, and H. Al-Nashash, “Classification of Mental Stress using Dry EEG Electrodes and Machine Learning,” in *2023 Advances in Science and Engineering Technology International Conferences (ASET)*, IEEE, Feb. 2023, pp. 1–5. doi: 10.1109/ASET56582.2023.10180884.
[201] W.-K. Beh, Y.-H. Wu, and A.-Y. Wu, “Robust PPG-Based Mental Workload Assessment System Using Wearable Devices,” *IEEE J. Biomed. Heal. Informatics*, vol. 27, no. 5, pp. 2323–2333, May 2023, doi: 10.1109/JBHI.2021.3138639.
[202] Y. Fukuda, W. Ishikawa, R. Kanayama, T. Matsumoto, N. Takemura, and K. Sakatani, “Bayesian prediction of anxiety level in aged people at rest using 2-channel NIRS data from prefrontal cortex,” *Adv. Exp. Med. Biol.*, vol. 812, pp. 303–308, 2014, doi: 10.1007/978-1-4939-0620-8_40.
[203] Y. Fukuda, Y. Ida, T. Matsumoto, N. Takemura, and K. Sakatani, “A bayesian algorithm for anxiety index prediction based on cerebral blood oxygenation in the prefrontal cortex

measured by near infrared spectroscopy," *IEEE J. Transl. Eng. Heal. Med.*, vol. 2, 2014, doi: 10.1109/JTEHM.2014.2361757.
[204] T. Fan *et al.*, "A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition," *Comput. Biol. Med.*, vol. 159, p. 106938, Jun. 2023, doi: 10.1016/J.COMPBIOMED.2023.106938.
[205] I. P. Clara, B. J. Cox, and M. W. Enns, "Confirmatory Factor Analysis of the Depression-Anxiety-Stress Scales in Depressed and Anxious Patients," *J. Psychopathol. Behav. Assess.*, vol. 23, no. 1, pp. 61–67, 2001, doi: 10.1023/A:1011095624717.
[206] B. Rim, N. J. Sung, S. Min, and M. Hong, "Deep Learning in Physiological Signal Data: A Survey," *Sensors 2020, Vol. 20, Page 969*, vol. 20, no. 4, p. 969, Feb. 2020, doi: 10.3390/S20040969.
[207] W. Tang, G. Long, L. Liu, T. Zhou, M. Blumenstein, and J. Jiang, "Omni-Scale Cnns: a Simple and Effective Kernel Size Configuration for Time Series Classification," *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, 2022, Accessed: Nov. 18, 2023. [Online]. Available: https://www.researchgate.net/profile/Lu-Liu-139/publication/339471768_Rethinking_1D-CNN_for_Time_Series_Classification_A_Stronger_Baseline/links/5e5d2d9492851cefa1d60aaf/Rethinking-1D-CNN-for-Time-Series-Classification-A-Stronger-Baseline.pdf
[208] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data 2021 81*, vol. 8, no. 1, pp. 1–74, Mar. 2021, doi: 10.1186/S40537-021-00444-8.
[209] F. Tu, S. Yin, P. Ouyang, S. Tang, L. Liu, and S. Wei, "Deep Convolutional Neural Network Architecture with Reconfigurable Computation Patterns," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 25, no. 8, pp. 2220–2233, Aug. 2017, doi: 10.1109/TVLSI.2017.2688340.
[210] M. Gjoreski, M. Z. Gams, M. Luštrek, P. Genc, J. U. Garbas, and T. Hassan, "Machine Learning and End-to-End Deep Learning for Monitoring Driver Distractions from Physiological and Visual Signals," *IEEE Access*, vol. 8, pp. 70590–70603, 2020, doi: 10.1109/ACCESS.2020.2986810.
[211] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev. 2020 538*, vol. 53, no. 8, pp. 5455–5516, Apr. 2020, doi: 10.1007/S10462-020-09825-6.
[212] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/J.PATCOG.2017.10.013.
[213] M. R. Askari, M. Abdel-Latif, M. Rashid, M. Sevil, and A. Cinar, "Detection and Classification of Unannounced Physical Activities and Acute Psychological Stress Events for Interventions in Diabetes Treatment," *Algorithms*, vol. 15, no. 10, p. 352, Sep. 2022, doi: 10.3390/a15100352.
[214] F. R. Ihmig, H. Antonio Gogeascoechea, F. Neurohr-Parakenings, S. K. Schäfer, J. Lass-Hennemann, and T. Michael, "On-line anxiety level detection from biosignals: Machine learning based on a randomized controlled trial with spider-fearful individuals," *PLoS One*, vol. 15, no. 6, pp. 1–20, 2020, doi: 10.1371/journal.pone.0231517.
[215] V. C. Goessl, J. E. Curtiss, and S. G. Hofmann, "The effect of heart rate variability biofeedback training on stress and anxiety: A meta-analysis," *Psychol. Med.*, vol. 47, no. 15, pp. 2578–2586, Nov. 2017, doi: 10.1017/S0033291717001003.
[216] S. Gradl, M. Wirth, T. Zillig, and B. M. Eskofier, "Visualization of heart activity in virtual

reality: A biofeedback application using wearable sensors," in *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks, BSN 2018*, IEEE, Mar. 2018, pp. 152–155. doi: 10.1109/BSN.2018.8329681.
[217] L. Ancillon, M. Elgendi, and C. Menon, "Machine Learning for Anxiety Detection Using Biosignals: A Review," *Diagnostics*, vol. 12, no. 8, p. 1794, Jul. 2022, doi: 10.3390/diagnostics12081794.
[218] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychological Medicine*, vol. 49, no. 9. pp. 1426–1448, 2019. doi: 10.1017/S0033291719000151.
[219] T. G. M. Vrijkotte, L. J. P. Van Doornen, and E. J. C. De Geus, "Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability," *Hypertension*, vol. 35, no. 4, pp. 880–886, 2000, doi: 10.1161/01.HYP.35.4.880.
[220] G. F. Wilson, "An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures," *Int. J. Aviat. Psychol.*, vol. 12, no. 1 SPEC, pp. 3–18, 2002, doi: 10.1207/s15327108ijap1201_2.
[221] G. G. Berntson and J. T. Cacioppo, "Integrative Physiology: Homeostasis, Allostasis, and the Orchestration of Systemic Physiology," in *Handbook of Psychophysiology*, Cambridge University Press, 2007, pp. 433–452. doi: 10.1017/cbo9780511546396.019.
[222] G. Regalia, D. Resnati, and S. Tognetti, "Sensors on the Wrist," *Encycl. Sensors Biosens. Vol. 1-4, First Ed.*, vol. 1–4, pp. 1–20, Jan. 2022, doi: 10.1016/B978-0-12-822548-6.00130-8.
[223] A. Sano *et al.*, "Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones," in *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks, BSN 2015*, 2015, pp. 1–6. doi: 10.1109/BSN.2015.7299420.
[224] A. Sano *et al.*, "Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study," *J. Med. Internet Res.*, vol. 20, no. 6, 2018, doi: 10.2196/jmir.9410.
[225] W. M. Suess, A. B. Alexander, D. D. Smith, H. W. Sweeney, and R. J. Marion, "The Effects of Psychological Stress on Respiration: A Preliminary Study of Anxiety and Hyperventilation," *Psychophysiology*, vol. 17, no. 6, pp. 535–540, 1980, doi: 10.1111/j.1469-8986.1980.tb02293.x.
[226] F. Gonçalves, D. Carneiro, J. Pêgo, and P. Novais, "Monitoring mental stress through mouse behaviour and decision-making patterns," *Advances in Intelligent Systems and Computing*, vol. 806. pp. 40–47, 2019. doi: 10.1007/978-3-030-01746-0_5.
[227] S. Koelstra *et al.*, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.
[228] P. Sarkar and A. Etemad, "Self-Supervised ECG Representation Learning for Emotion Recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1541–1554, 2022, doi: 10.1109/TAFFC.2020.3014842.
[229] R. Li and Z. Liu, "Stress detection using deep neural networks," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 11, pp. 1–10, Dec. 2020, doi: 10.1186/s12911-020-01299-4.
[230] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep PPG: Large-scale heart rate estimation with convolutional neural networks," *Sensors (Switzerland)*, vol. 19, no. 14, p. 3079, Jul. 2019, doi: 10.3390/s19143079.

[231] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, “An explainable deep fusion network for affect recognition using physiological signals,” *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 2069–2072, Nov. 2019, doi: 10.1145/3357384.3358160.
[232] P. H. Charlton, P. A. Kyriacou, J. Mant, V. Marozas, P. Chowienczyk, and J. Alastruey, “Wearable Photoplethysmography for Cardiovascular Monitoring,” *Proc. IEEE*, vol. 110, no. 3, pp. 355–381, Mar. 2022, doi: 10.1109/JPROC.2022.3149785.
[233] P. Siirtola, “Continuous stress detection using the sensors of commercial smartwatch,” in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, New York, NY, USA: ACM, Sep. 2019, pp. 1198–1201. doi: 10.1145/3341162.3344831.
[234] N. A. S. Taylor and C. A. Machado-Moreira, “Regional variations in transepidermal water loss, eccrine sweat gland density, sweat secretion rates and electrolyte composition in resting and exercising humans,” *Extrem. Physiol. Med.*, vol. 2, no. 1, Feb. 2013, doi: 10.1186/2046-7648-2-4.
[235] T. Öberg, L. Sandsjö, and R. Kadefors, “Electromyogram mean power frequency in non-fatigued trapezius muscle,” *Eur. J. Appl. Physiol. Occup. Physiol.*, vol. 61, no. 5–6, pp. 362–369, Dec. 1990, doi: 10.1007/BF00236054.
[236] D. Watson, L. A. Clark, and A. Tellegen, “Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales,” *J. Pers. Soc. Psychol.*, vol. 54, no. 6, pp. 1063–1070, 1988, doi: 10.1037/0022-3514.54.6.1063.
[237] W. S. Helton, “Validation of a Short Stress State Questionnaire,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 48, no. 11, pp. 1238–1242, Sep. 2004, doi: 10.1177/154193120404801107.
[238] A. C. Samson, S. D. Kreibig, B. Soderstrom, A. A. Wade, and J. J. Gross, “Eliciting positive, negative and mixed emotional states: A film library for affective scientists,” *Cogn. Emot.*, vol. 30, no. 5, pp. 827–856, Jul. 2016, doi: 10.1080/02699931.2015.1031089.
[239] D. Roselli, J. Matthews, and N. Talagala, “Managing bias in AI,” *Web Conf. 2019 - Companion World Wide Web Conf. WWW 2019*, pp. 539–544, May 2019, doi: 10.1145/3308560.3317590.
[240] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.
[241] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, May 27, 2015. doi: 10.1038/nature14539.
[242] A. E. Alkurdi, M. He, M. E. Hernandez, and E. T. Hsiao-wecksler, “Machine Learning Approaches In Anxiety Detection - A Comprehensive Review,” *IEEE Affect. Comput.*.
[243] Y. Zhang, Y. OwecHko, and J. Zhang, “Driver cognitive workload estimation: A data-driven perspective,” in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2004, pp. 642–647. doi: 10.1109/itsc.2004.1398976.
[244] C. D. Katsis, N. Katertsidis, G. Ganiatsas, and D. I. Fotiadis, “Toward emotion recognition in car-racing drivers: A biosignal processing approach,” *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, vol. 38, no. 3, pp. 502–512, 2008, doi: 10.1109/TSMCA.2008.918624.
[245] E. Fix and J. L. Hodges, “Discriminatory Analysis. Nonparametric Discrimination:

Consistency Properties," *Int. Stat. Rev. / Rev. Int. Stat.*, vol. 57, no. 3, p. 238, 1989, doi: 10.2307/1403797.
[246] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[247] N. Long, Y. Lei, L. Peng, P. Xu, and P. Mao, "A scoping review on monitoring mental health using smart wearable devices," *Math. Biosci. Eng.*, vol. 19, no. 8, pp. 7899–7919, 2022, doi: 10.3934/mbe.2022369.
[248] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on Psychological Stress Detection Using Biosignals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 440–460, 2022, doi: 10.1109/TAFFC.2019.2927337.
[249] F. Scholkmann, J. Boss, and M. Wolf, "An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals," *Algorithms*, vol. 5, no. 4, pp. 588–603, Nov. 2012, doi: 10.3390/a5040588.
[250] Fabian Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, [Online]. Available: http://scikit-learn.sourceforge.net.
[251] G. Ou and Y. L. Murphey, "Multi-class pattern classification using neural networks," *Pattern Recognit.*, vol. 40, no. 1, pp. 4–18, Jan. 2007, doi: 10.1016/j.patcog.2006.04.041.
[252] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, 2004.
[253] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002, doi: 10.1109/72.991427.
[254] J. L. Hintze and R. D. Nelson, "Violin plots: A box plot-density trace synergism," *Am. Stat.*, vol. 52, no. 2, pp. 181–184, 1998, doi: 10.1080/00031305.1998.10480559.
[255] K. J. Bär and H. Critchley, "Autonomic Control," *Brain Mapp. An Encycl. Ref.*, vol. 2, pp. 635–642, Feb. 2015, doi: 10.1016/B978-0-12-397025-1.00058-0.
[256] S. Seneviratne *et al.*, "A Survey of Wearable Devices and Challenges," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 4, pp. 2573–2620, 2017, doi: 10.1109/COMST.2017.2731979.
[257] M. Elgendi and C. Menon, "Assessing anxiety disorders using wearable devices: Challenges and future directions," *Brain Sci.*, vol. 9, no. 3, 2019, doi: 10.3390/brainsci9030050.
[258] M. A. Serhani, H. T. El Kassabi, H. Ismail, and A. N. Navaz, "ECG monitoring systems: Review, architecture, processes, and key challenges," *Sensors (Switzerland)*, vol. 20, no. 6, 2020, doi: 10.3390/s20061796.
[259] M. Elgendi, "Optimal signal quality index for photoplethysmogram signals," *Bioengineering*, vol. 3, no. 4, p. 21, Sep. 2016, doi: 10.3390/bioengineering3040021.
[260] R. Couceiro, P. Carvalho, R. P. Paiva, J. Henriques, I. Quintal, and J. Muehlsteff, "Detection of Motion Artifacts in Photoplethysmographic Signals: Algorithms Comparison," in *IFMBE Proceedings*, vol. 42, 2014, pp. 327–330. doi: 10.1007/978-3-319-03005-0_83.
[261] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.
[262] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, p. 9, Dec. 2016, doi: 10.1186/s40537-016-0043-6.

[263] F. Zhuang *et al.*, “A Comprehensive Survey on Transfer Learning,” *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.
[264] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11141 LNCS, 2018, pp. 270–279. doi: 10.1007/978-3-030-01424-7_27.
[265] L. Shao, F. Zhu, and X. Li, “Transfer learning for visual categorization: A survey,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015, doi: 10.1109/TNNLS.2014.2330900.
[266] M. Malik *et al.*, “Heart rate variability: Standards of measurement, physiological interpretation, and clinical use,” *Circulation*, vol. 93, no. 5, pp. 1043–1065, Mar. 1996, doi: 10.1161/01.cir.93.5.1043.
[267] G. G. Berntson *et al.*, “Heart rate variability: Origins methods, and interpretive caveats,” *Psychophysiology*, vol. 34, no. 6, pp. 623–648, 1997, doi: 10.1111/J.1469-8986.1997.TB02140.X.
[268] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
[269] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
[270] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.
[271] G. M. Weiss, “Mining with rarity,” *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, Jun. 2004, doi: 10.1145/1007730.1007734.
[272] X. Y. Liu, J. Wu, and Z. H. Zhou, “Exploratory undersampling for class-imbalance learning,” *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 39, no. 2, pp. 539–550, 2009, doi: 10.1109/TSMCB.2008.2007853.
[273] P. Lara-Benítez, M. Carranza-García, and J. C. Riquelme, “An Experimental Review on Deep Learning Architectures for Time Series Forecasting,” *https://doi.org/10.1142/S0129065721300011*, vol. 31, no. 3, Feb. 2021, doi: 10.1142/S0129065721300011.
[274] P. Tirumala Rao, S. Koteswarao Rao, G. Manikanta, and S. Ravi Kumar, “Distinguishing normal and abnormal ECG signal,” *Indian J. Sci. Technol.*, vol. 9, no. 10, Mar. 2016, doi: 10.17485/IJST/2016/V9I10/85449.

APPENDIX A: CODE REPOSITORIES

Code for the Feature-based work is available at https://github.com/AbdulAlkurdi/anxietyFB
Code for the End-to-End work is available at https://github.com/AbdulAlkurdi/anxietyE2E